

PHYSICS

HSC COURSE
THIRD EDITION

2

MICHAEL ANDRIESSEN • PETER PENTLAND
RICHARD GAUT • BRUCE MCKAY • JILL TACON

jacaranda plus

JACARANDA HSC SCIENCE

Third edition published 2008 by
John Wiley & Sons Australia, Ltd
42 McDougall Street, Milton, Qld 4064

Typeset in 10.5/12pt New Baskerville

© Michael Andriessen, Peter Pentland, Richard Gaut, Bruce McKay,
Jillian Tacon and Upgrade Business Systems (Ric Morante) 2008

First edition published 2001

© Michael Andriessen, Peter Pentland, Richard Gaut and
Bruce McKay 2001

Second edition published 2003

© Michael Andriessen, Peter Pentland, Richard Gaut, Bruce McKay
and Jillian Tacon 2003

The moral rights of the authors have been asserted.

National Library of Australia
Cataloguing-in-Publication data

Title: Physics 2 HSC course/Michael Andriessen ... [et al.].

Edition: 3rd ed.

ISBN: 978 0 7314 0823 8 (pbk.)

Notes: Includes index.

Target audience: For secondary school age.

Subjects: Physics — Textbooks.

Other authors/contributors: Andriessen, Michael.

Dewey number: 530

Reproduction and communication for educational purposes

The Australian *Copyright Act 1968* allows a maximum of one chapter or 10% of the pages of this work, whichever is the greater, to be reproduced and/or communicated by any educational institution for its educational purposes provided that the educational institution (or the body that administers it) has given a remuneration notice to Copyright Agency Limited (CAL).

Reproduction and communication for other purposes

Except as permitted under the Act (for example, a fair dealing for the purposes of study, research, criticism or review), no part of this book may be reproduced, stored in a retrieval system, communicated or transmitted in any form or by any means without prior written permission. All inquiries should be made to the publisher.

All activities have been written with the safety of both teacher and student in mind. Some, however, involve physical activity or the use of equipment or tools. All due care should be taken when performing such activities. Neither the publisher nor the authors can accept responsibility for any injury that may be sustained when completing activities described in this textbook.

Front and back cover images: © Photodisc

Illustrated by the Wiley Art Studio

Printed in Singapore by
Craft Print

10 9 8 7 6 5 4 3 2

CONTENTS

Preface viii
About eBookPLUS ix
Syllabus grid x
Acknowledgements xvi

HSC CORE MODULE

Space

Chapter 1: Earth's gravitational field 2

1.1 The Earth's gravity 3
1.2 Weight 6
1.3 Gravitational potential energy 7
Summary 10
Questions 10
Practical activities 11

Chapter 2: Launching into space 13

2.1 Projectile motion 14
2.2 Escape velocity 23
2.3 Lift-off 24
Summary 33
Questions 33
Practical activities 35

Chapter 3: Orbiting and re-entry 38

3.1 In orbit 39
3.2 Re-entry 50
Summary 56
Questions 56
Practical activities 58

Chapter 4: Gravity in the solar system 60

4.1 The Law of Universal Gravitation 61
4.2 Gravitational fields 65
4.3 The slingshot effect 66
Summary 70
Questions 70

Chapter 5: Space and time 71

5.1 The aether model 72
5.2 Special relativity 74
5.3 Consequences of special relativity 77
Summary 93
Questions 93
Practical activities 96

HSC CORE MODULE

Motors and generators

Chapter 6: The motor effect and DC electric motors 100

6.1 The motor effect 103
6.2 Forces between two parallel conductors 105
6.3 Torque 107
6.4 DC electric motors 109
Summary 116
Questions 116
Practical activities 120

Chapter 7: Generating electricity 122

- 7.1 The discoveries of Michael Faraday 123
- 7.2 Electromagnetic induction 126
- 7.3 Generating a potential difference 127
- 7.4 Lenz's law 128
- 7.5 Eddy currents 131

Summary 134

Questions 134

Practical activities 137

Chapter 8: Generators and power distribution 139

- 8.1 Generators 140
- 8.2 Electric power generating stations 146
- 8.3 Transformers 148
- 8.4 Power distribution 151
- 8.5 Electricity and society 156

Summary 157

Questions 157

Practical activities 160

Chapter 9: AC electric motors 163

- 9.1 Main features of an AC motor 164
- 9.2 Energy transformations and transfers 169

Summary 171

Questions 171

Practical activities 172

HSC CORE MODULE

**From ideas to
implementation**

Chapter 10: Cathode rays and the development of television 174

- 10.1 The discovery of cathode rays 175
- 10.2 Effect of electric fields on cathode rays 177
- 10.3 Effect of magnetic fields on cathode rays 182
- 10.4 Determining the charge-to-mass ratio of cathode rays 183
- 10.5 Cathode rays — waves or particles? 184
- 10.6 Applications of cathode rays 186

Summary 189

Questions 189

Practical activities 191

Chapter 11: The photoelectric effect and black body radiation 193

- 11.1 Maxwell's theory of electromagnetic waves 194
- 11.2 Heinrich Hertz and experiments with radio waves 196
- 11.3 The black body problem and the ultraviolet catastrophe 199
- 11.4 What do we mean by 'classical physics' and 'quantum theory'? 202
- 11.5 The photoelectric effect 202

Summary 209

Questions 209

Practical activities 211

HSC OPTION MODULE

Astrophysics

Chapter 12: The development and application of transistors 212

- 12.1 Conductors, insulators and semiconductors 213
- 12.2 Band structures in semiconductors 216
- 12.3 Doping and band structure 219
- 12.4 Thermionic devices 220
- 12.5 Solid state devices 222
- 12.6 Thermionic versus solid state devices 224
- 12.7 Invention of the transistor 225
- 12.8 Integrated circuits 227

Summary 230

Questions 230

Practical activities 231

Chapter 13: Superconductivity 232

- 13.1 Interference 233
- 13.2 Diffraction 235
- 13.3 X-ray diffraction 235
- 13.4 Bragg's experiment 238
- 13.5 The crystal lattice structure of metals 239
- 13.6 Superconductivity 240
- 13.7 How is superconductivity explained? 243

Summary 251

Questions 251

Practical activities 253

Chapter 14: Looking and seeing 256

- 14.1 Galileo's telescopes 257
- 14.2 Atmospheric absorption of the electromagnetic spectrum 258
- 14.3 Telescopes 261
- 14.4 Seeing 265
- 14.5 Modern methods to improve telescope performance 265

Summary 270

Questions 270

Practical activities 272

Chapter 15: Astronomical measurement 274

- 15.1 Astrometry 275
- 15.2 Spectroscopy 279
- 15.3 Photometry 289

Summary 299

Questions 299

Practical activities 302

Chapter 16: Binaries and variables 305

- 16.1 Binaries 306
- 16.2 Variables 312

Summary 316

Questions 316

Practical activities 318

HSC OPTION MODULE

Medical physics

Chapter 17: Star lives 320

- 17.1 Star birth 321
- 17.2 Main sequence star life 324
- 17.3 Star life after the main sequence 327
- 17.4 Star death 332

Summary 336

Questions 336

Practical activities 338

Chapter 18: The use of ultrasound in medicine 340

- 18.1 What type of sound is ultrasound? 341
- 18.2 Using ultrasound to detect structure inside the body 343
- 18.3 Producing and detecting ultrasound: the piezoelectric effect 347
- 18.4 Gathering and using information in an ultrasound scan 348
- 18.5 Using ultrasound to examine blood flow 352

Summary 358

Questions 358

Chapter 19: Electromagnetic radiation as a diagnostic tool 361

- 19.1 X-rays in medical diagnosis 362
- 19.2 CT scans in medical diagnosis 368
- 19.3 Endoscopes in medical diagnosis 373

Summary 378

Questions 378

Practical activities 380

Chapter 20: Radioactivity as a diagnostic tool 381

- 20.1 Radioactivity and the use of radioisotopes 382
- 20.2 Positron emission tomography (PET) 392
- 20.3 Imaging methods working together 394

Summary 396

Questions 396

Chapter 21: Magnetic resonance imaging as a diagnostic tool 398

- 21.1 The patient and the image using MRI 399
- 21.2 The MRI machine: effect on atoms in the patient 402
- 21.3 Medical uses of MRI 410
- 21.4 Comparison of the main imaging techniques 412

Summary 415

Questions 415

Chapter 22: The atomic models of Rutherford and Bohr 418

- 22.1 The Rutherford model of the atom 419
- 22.2 Bohr's model of the atom 423
- 22.3 Bohr's postulates 427
- 22.4 Mathematics of the Rutherford and Bohr models 429
- 22.5 Limitations of the Bohr model of the atom 434

Summary 435

Questions 435

Practical activities 437

HSC OPTION MODULE

From quanta to quarks

Chapter 23: Development of quantum mechanics 440

- 23.1 Diffraction 441
23.2 Steps towards a complete quantum theory model of the atom 444
Summary 452
Questions 452

Chapter 24: Probing the nucleus 453

- 24.1 Discoveries pre-dating the nucleus 454
24.2 Discovery of the neutron 458
24.3 Discovery of the neutrino 461
24.4 The strong nuclear force 466
24.5 Mass defect and binding energy of the nucleus 468

Summary 472
Questions 472

Chapter 25: Nuclear fission and other uses of nuclear physics 474

- 25.1 Energy from the nucleus 475
25.2 The discovery of nuclear fission 476
25.3 The development of the atom bomb 480
25.4 Nuclear fission reactors 484
25.5 Medical and industrial applications of radioisotopes 489
25.6 Neutron scattering 492

Summary 493
Questions 493

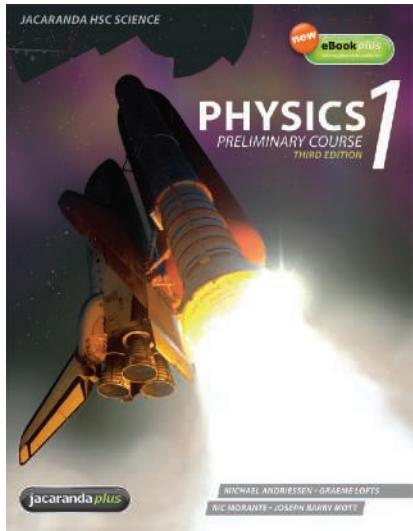
Chapter 26: Quarks and the Standard Model of particle physics 495

- 26.1 Instruments used by particle physicists 496
26.2 The Standard Model of particle physics 503

Summary 516
Questions 516
Practical activities 518

- Glossary 521
Appendix 1: Formulae and data sheet 526
Appendix 2: Periodic table 528
Appendix 3: Key words for examination questions 529
Answers to numerical questions 531
Index 536

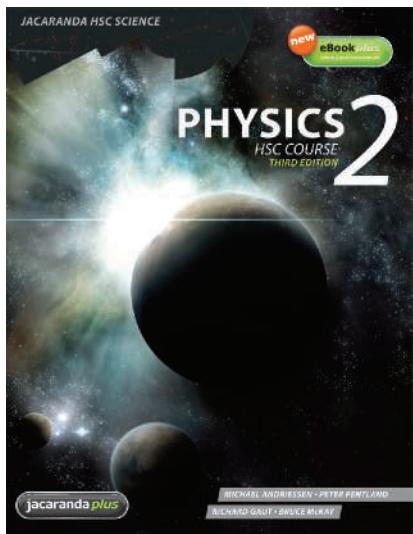
PREFACE



This third edition of *Physics 2: HSC Course* is revised and updated to meet all the requirements of the amended Stage 6 Physics Syllabus for Year 12 students in New South Wales. Written by a team of experienced Physics teachers, *Physics 2* offers a complete resource with coverage of the three core modules as well as three option modules: Quanta to Quarks, Astrophysics and Medical Physics. An additional option topic, The Age of Silicon, is available online.

Physics 2 features:

- full-colour, high-quality, detailed illustrations to enhance students' understanding of Physics concepts
- clearly written explanations and sample problems
- interest boxes focusing on up-to-date information, current research and new discoveries
- practical activities at the end of each chapter to support the syllabus investigations
- key terms highlighted and defined in the context of the chapters and in a complete glossary
- chapter reviews that provide a summary and a range of problem-solving and descriptive questions.



eBook plus

Next generation teaching and learning

This title features eBookPLUS: an electronic version of the textbook and a complementary set of targeted digital resources. These flexible and engaging ICT activities are available online at the JacarandaPLUS website (www.jacplus.com.au).

eBookPLUS icons within the text direct students to the online resources, which include:

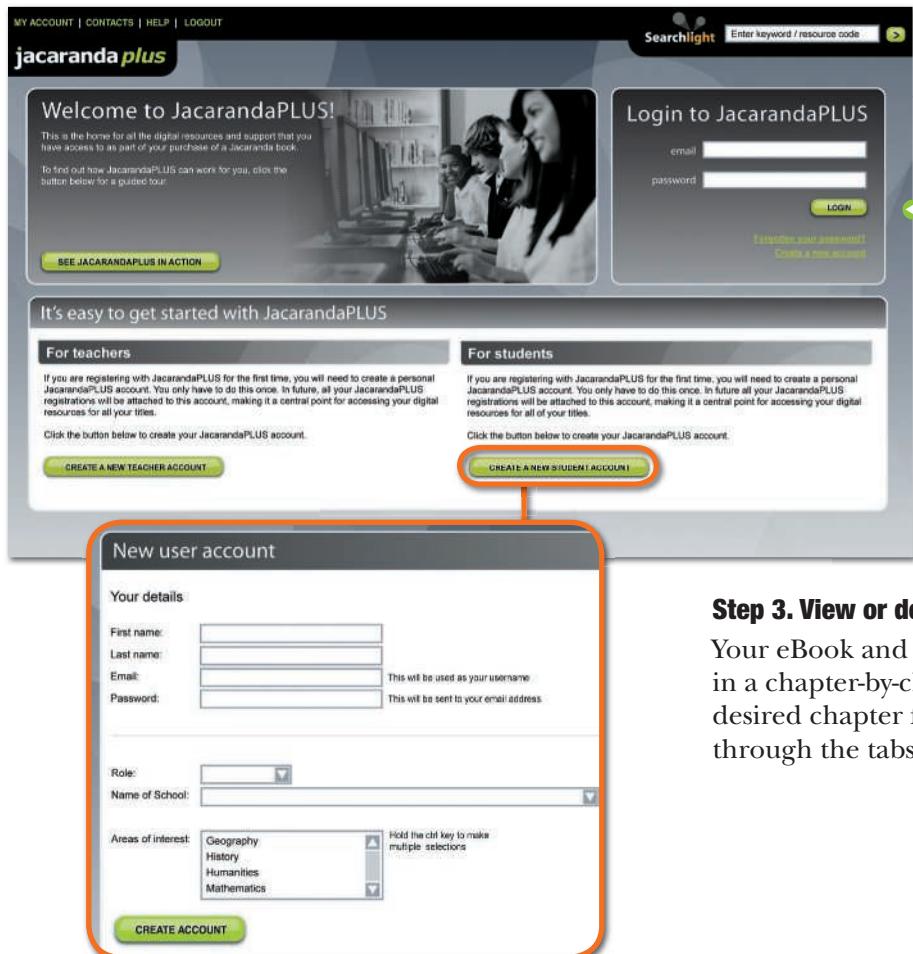
- *eModelling*: Excel spreadsheets that provide examples of numerical and algebraic modelling
- *eLessons*: Video and animations that reinforce study by bringing key concepts to life
- *Interactivities*: Interactive study activities that enhance student understanding of key concepts through hands-on experience
- *Weblinks*: HTML links to other useful support material on the internet.

About eBookPLUS

Physics 2: HSC Course, 3rd edition features eBookPLUS: an electronic version of the entire textbook and supporting multimedia resources. It is available for you online at the JacarandaPLUS website (www.jacplus.com.au).

Using the JacarandaPLUS website

To access your eBookPLUS resources, simply log on to www.jacplus.com.au. There are three easy steps for using the JacarandaPLUS system.



Step 1. Create a user account

The first time you use the JacarandaPLUS system, you will need to create a user account. Go to the JacarandaPLUS home page (www.jacplus.com.au) and follow the instructions on screen.

Step 2. Enter your registration code

Once you have created a new account and logged in, you will be prompted to enter your unique registration code for this book, which is printed on the inside front cover of your textbook.

LOGIN

Once you have created your account, you can use the same email address and password in the future to register any JacarandaPLUS books.

Step 3. View or download eBookPLUS resources

Your eBook and supporting resources are provided in a chapter-by-chapter format. Simply select the desired chapter from the drop-down list and navigate through the tabs to locate the appropriate resource.

Minimum requirements

- Internet Explorer 7, Mozilla Firefox 1.5 or Safari 1.3
- Adobe Flash Player 9
- Javascript must be enabled (most browsers are enabled by default).

Troubleshooting

- Go to the JacarandaPLUS help page at www.jacplus.com.au
- Contact John Wiley & Sons Australia, Ltd.
Email: support@jacplus.com.au
Phone: 1800 JAC PLUS (1800 522 7587)

SYLLABUS GRID

Core module: SPACE (chapters 1–5, pages 1–98)

1. The Earth has a gravitational field that exerts a force on objects both on it and around it

<p><i>Students learn to:</i></p> <ul style="list-style-type: none"> define weight as the force on an object due to a gravitational field explain that a change in gravitational potential energy is related to work done define gravitational potential energy as the work done to move an object from a very large distance away to a point in a gravitational field $E_p = -G \frac{m_1 m_2}{r}$	page 8 7 7–9	<p><i>Students:</i></p> <ul style="list-style-type: none"> perform an investigation and gather information to determine a value for acceleration due to gravity using pendulum motion or computer-assisted technology and identify reasons for possible variations from the value 9.8 m s^{-2} gather secondary information to predict the value of acceleration due to gravity on other planets analyse information using the expression $F = mg$ to determine the weight force for a body on Earth and for the same body on other planets 	page 11 5, 10, 12 6, 10, 12
---	-----------------------	--	--

2. Many factors have to be taken into account to achieve a successful rocket launch, maintain a stable orbit and return to Earth

<p><i>Students learn to:</i></p> <ul style="list-style-type: none"> describe the trajectory of an object undergoing projectile motion within the Earth's gravitational field in terms of horizontal and vertical components describe Galileo's analysis of projectile motion explain the concept of escape velocity in terms of the: <ul style="list-style-type: none"> gravitational constant mass and radius of the planet outline Newton's concept of escape velocity identify why the term 'g forces' is used to explain the forces acting on an astronaut during launch discuss the effect of the Earth's orbital motion and its rotational motion on the launch of a rocket analyse the changing acceleration of a rocket during launch in terms of the: <ul style="list-style-type: none"> Law of Conservation of Momentum forces experienced by astronauts analyse the forces involved in uniform circular motion for a range of objects, including satellites orbiting the Earth compare qualitatively low Earth and geo-stationary orbits define the term orbital velocity and the quantitative and qualitative relationship between orbital velocity, the gravitational constant, mass of the central body, mass of the satellite and the radius of the orbit using Kepler's Law of Periods account for the orbital decay of satellites in low Earth orbit discuss issues associated with safe re-entry into the Earth's atmosphere and landing on the Earth's surface identify that there is an optimum angle for safe re-entry for a manned spacecraft into the Earth's atmosphere and the consequences of failing to achieve this angle 	page 14–23 14 23–4 23 26–31 31–2 25, 26–7 (see also 36–7) 39–41 47–8 41–4 49–50 50–5 51	<p><i>Students:</i></p> <ul style="list-style-type: none"> solve problems and analyse information to calculate the actual velocity of a projectile from its horizontal and vertical components using: $v_x^2 = u_x^2$ $v = \sqrt{u_x^2 + at}$ $v_y^2 = u_y^2 + 2a_y \Delta y$ $\Delta x = u_x t$ $\Delta y = u_y t + \frac{1}{2} a_y t^2$ perform a first-hand investigation, gather information and analyse data to calculate initial and final velocity, maximum height reached, range and time of flight of a projectile for a range of situations by using simulations, data loggers and computer analysis identify data sources, gather, analyse and present information on the contribution of one of the following to the development of space exploration: Tsiolkovsky, Oberth, Goddard, Esnault-Pelterie, O'Neill or von Braun solve problems and analyse information to calculate the centripetal force acting on a satellite undergoing uniform circular motion about the Earth using $F = \frac{mv^2}{r}$ solve problems and analyse information using: $\frac{r^3}{T^2} = \frac{GM}{4\pi^2}$ 	page 19–22, 33–4 35 32 40–1, 56, 58–9 42–4, 56–7
--	--	--	---

3. The Solar System is held together by gravity

<p><i>Students learn to:</i></p> <ul style="list-style-type: none"> describe a gravitational field in the region surrounding a massive object in terms of its effects on other masses in it define Newton's Law of Universal Gravitation $F = G \frac{m_1 m_2}{d^2}$ <ul style="list-style-type: none"> discuss the importance of Newton's Law of Universal Gravitation in understanding and calculating the motion of satellites identify that a slingshot effect can be provided by planets for space probes 	page 65–6 61–2 62–5 66–9	<p><i>Students:</i></p> <ul style="list-style-type: none"> present information and use available evidence to discuss the factors affecting the strength of the gravitational force solve problems and analyse information using $F = G \frac{m_1 m_2}{d^2}$ 	page 61–4, 70 61–2, 70
--	--	---	----------------------------------

4. Current and emerging understanding about time and space has been dependent upon earlier models of the transmission of light

<p><i>Students learn to:</i></p> <ul style="list-style-type: none"> outline the features of the aether model for the transmission of light describe and evaluate the Michelson-Morley attempt to measure the relative velocity of the Earth through the aether discuss the role of the Michelson-Morley experiments in making determinations about competing theories outline the nature of inertial frames of reference discuss the principle of relativity describe the significance of Einstein's assumption of the constancy of the speed of light identify that if c is constant then space and time become relative discuss the concept that length standards are defined in terms of time in contrast to the original metre standard 	page 72 72–4 74 74–5 74–5 75–6 76 77	<p><i>Students:</i></p> <ul style="list-style-type: none"> gather and process information to interpret the results of the Michelson-Morley experiment perform an investigation to help distinguish between non-inertial and inertial frames of reference analyse and interpret some of Einstein's thought experiments involving mirrors and trains and discuss the relationship between thought and reality analyse information to discuss the relationship between theory and the evidence supporting it, using Einstein's predictions based on relativity that were made many years before evidence was available to support it 	page 96–7 97–8 75–6, 77–8 80–1
---	--	---	--

(continued)

<ul style="list-style-type: none"> explain qualitatively and quantitatively the consequence of special relativity in relation to: <ul style="list-style-type: none"> the relativity of simultaneity the equivalence between mass and energy length contraction time dilation mass dilation discuss the implications of mass increase, time dilation and length contraction for space travel 	<p>77–8 88–9 81–4 78–9 85–8 89–92</p>	<ul style="list-style-type: none"> solve problems and analyse information using: $E = mc^2$ $l_v = l_0 \sqrt{1 - \frac{v^2}{c^2}}$ $t_v = \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}}$ $m_v = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$	<p>89, 95 84, 94–5 81, 94–5 88, 95</p>
---	---	---	--

Core module: MOTORS AND GENERATORS (chapters 6–9, pages 101–72)

1. Motors use the effect of forces on current-carrying conductors in magnetic fields

<p><i>Students learn to:</i></p> <ul style="list-style-type: none"> discuss the effect on the magnitude of the force on a current-carrying conductor of variations in: <ul style="list-style-type: none"> the strength of the magnetic field in which it is located the magnitude of the current in the conductor the length of the conductor in the external magnetic field the angle between the direction of the external magnetic field and the direction of the length of the conductor describe qualitatively and quantitatively the force between long parallel current-carrying conductors: $\frac{F}{l} = k \frac{I_1 I_2}{d}$ define torque as the turning moment of a force using: $\tau = Fd$ identify that the motor effect is due to the force acting on a current-carrying conductor in a magnetic field describe the forces experienced by a current-carrying loop in a magnetic field and describe the net result of the forces describe the main features of a DC electric motor and the role of each feature identify that the required magnetic fields in DC motors can be produced either by current-carrying coils or permanent magnets 	<p>page 104–5 105–7 107–8 104–5 102–3 109–11 109–11, 112</p>	<p><i>Students:</i></p> <ul style="list-style-type: none"> solve problems using: $\frac{F}{l} = k \frac{I_1 I_2}{d}$ <ul style="list-style-type: none"> perform a first-hand investigation to demonstrate the motor effect solve problems and analyse information about the force on current-carrying conductors in magnetic fields using $F = BIl \sin \theta$ <ul style="list-style-type: none"> solve problems and analyse information about simple motors using: $\tau = nBIA \cos \theta$ <ul style="list-style-type: none"> identify data sources, gather and process information to qualitatively describe the application of the motor effect in: <ul style="list-style-type: none"> the galvanometer the loudspeaker 	<p>page 107, 117–19 120 105, 117–19 114, 118–19, 121 114–15</p>
---	--	---	---

2. The relative motion between a conductor and magnetic field is used to generate an electrical voltage

<p><i>Students learn to:</i></p> <ul style="list-style-type: none"> outline Michael Faraday's discovery of the generation of an electric current by a moving magnet define magnetic field strength B as magnetic flux density describe the concept of magnetic flux in terms of magnetic flux density and surface area describe generated potential difference as the rate of change of magnetic flux through a circuit account for Lenz's Law in terms of conservation of energy and relate it to the production of back emf in motors explain that, in electric motors, back emf opposes the supply emf explain the production of eddy currents in terms of Lenz's Law 	<p>page 123–6 126 126–7 127–8 128–30 129–30 131–2</p>	<p><i>Students:</i></p> <ul style="list-style-type: none"> perform an investigation to model the generation of an electric current by moving a magnet in a coil or a coil near a magnet plan, choose equipment or resources for, and perform a first-hand investigation to predict and verify the effect on a generated electric current when: <ul style="list-style-type: none"> the distance between the coil and the magnet is varied the strength of the magnet is varied the relative motion between the coil and the magnet is varied gather, analyse and present information to explain how induction is used in cooktops in electric ranges gather secondary information to identify how eddy currents have been utilised in electromagnetic braking 	<p>page 137–8 137–8 137–8 133 132</p>
--	---	--	---

3. Generators are used to provide large-scale power production

<p><i>Students learn to:</i></p> <ul style="list-style-type: none"> describe the main components of a generator compare the structure and function of a generator to an electric motor describe the differences between AC and DC generators discuss the energy losses that occur as energy is fed through transmission lines from the generator to the consumer assess the effects of the development of AC generators on society and the environment 	<p>page 140–1 141, 109, 164–5 142–5 151–4 147–8, 156</p>	<p><i>Students:</i></p> <ul style="list-style-type: none"> plan, choose equipment or resources for, and perform a first-hand investigation to demonstrate the production of an alternating current gather secondary information to discuss advantages and disadvantages of AC and DC generators and relate these to their use analyse secondary information on the competition between Westinghouse and Edison to supply electricity to cities gather and analyse information to identify how transmission lines are: <ul style="list-style-type: none"> insulated from supporting structures protected from lightning strikes 	<p>page 160–1 160 147–8 154–5</p>
---	--	---	---

4. Transformers allow generated power to be either increased or decreased before it is used

<p><i>Students learn to:</i></p> <ul style="list-style-type: none"> describe the purpose of transformers in electrical circuits compare step-up and step-down transformers identify the relationship between the ratio of the number of turns in the primary and secondary coils and the ratio of primary to secondary voltage explain why voltage transformations are related to conservation of energy explain the role of transformers in electricity sub-stations discuss why some electrical appliances in the home that are connected to the mains domestic power supply use a transformer discuss the impact of the development of transformers on society 	<p>page 148–9 149 149–50 149–50 154 148–9, 153 152–3, 156</p>	<p><i>Students:</i></p> <ul style="list-style-type: none"> perform an investigation to model the structure of a transformer to demonstrate how secondary voltage is produced solve problems and analyse information about transformers using: $\frac{V_p}{V_s} = \frac{n_p}{n_s}$ <ul style="list-style-type: none"> gather, analyse and use available evidence to discuss how difficulties of heating caused by eddy currents in transformers may be overcome gather and analyse secondary information to discuss the need for transformers in the transfer of electrical energy from a power station to its point of use 	<p>page 160–1 150–1 151 152–4, 162</p>
--	---	--	--

5. Motors are used in industries and the home usually to convert electrical energy into more useful forms of energy

<i>Students learn to:</i>	page 164–9	<i>Students:</i>	page 172
<ul style="list-style-type: none"> describe the main features of an AC electric motor 		<ul style="list-style-type: none"> perform an investigation to demonstrate the principle of an AC induction motor gather, process and analyse information to identify some of the energy transfers and transformations involving the conversion of electrical energy into more useful forms in the home and industry 	169–70

Core module: FROM IDEAS TO IMPLEMENTATION (chapters 10–13, pages 173–254)

1. Increased understanding of cathode rays led to the development of television

<i>Students learn to:</i>	page 184–5	<i>Students:</i>	page 191
<ul style="list-style-type: none"> explain why the apparent inconsistent behaviour of cathode rays caused debate as to whether they were charged particles or electromagnetic waves 	175–6	<ul style="list-style-type: none"> perform an investigation and gather first-hand information to observe the occurrence of different striation patterns for different pressures in discharge tubes 	192
<ul style="list-style-type: none"> explain that cathode ray tubes allowed the manipulation of a stream of charged particles 	182	<ul style="list-style-type: none"> perform an investigation to demonstrate and identify properties of cathode rays using discharge tubes: <ul style="list-style-type: none"> containing a Maltese cross containing electric plates with a fluorescent display screen containing a glass wheel analyse the information gathered to determine the sign of the charge on cathode rays 	
<ul style="list-style-type: none"> identify that moving charged particles in a magnetic field experience a force 	177–9	<ul style="list-style-type: none"> solve problems and analyse information using: $F = qvB\sin \theta$ $F = qE$ and $E = \frac{V}{d}$ 	178, 181, 182, 189–90
<ul style="list-style-type: none"> identify that charged plates produce an electric field describe quantitatively the force acting on a charge moving through a magnetic field $F = qvB\sin \theta$ discuss qualitatively the electric field strength due to a point charge, positive and negative charges and oppositely charged parallel plates describe quantitatively the electric field due to oppositely charged parallel plates outline Thomson's experiment to measure the charge : mass ratio of an electron outline the role of: <ul style="list-style-type: none"> electrodes in the electron gun the deflection plates or coils the fluorescent screen in the cathode ray tube of conventional TV displays and oscilloscopes 	182		

2. The reconceptualisation of the model of light led to an understanding of the photoelectric effect and black body radiation

<i>Students learn to:</i>	page 196–8, 202–3	<i>Students:</i>	page 211
<ul style="list-style-type: none"> describe Hertz's observation of the effect of a radio wave on a receiver and the photoelectric effect he produced but failed to investigate outline qualitatively Hertz's experiments in measuring the speed of radio waves and how they relate to light waves identify Planck's hypothesis that radiation emitted and absorbed by the walls of a black body cavity is quantised identify Einstein's contribution to quantum theory and its relation to black body radiation explain the particle model of light in terms of photons with particular energy and frequency identify the relationships between photon energy, frequency, speed of light and wavelength: $E = hf$ and $c = f\lambda$ 	196–8	<ul style="list-style-type: none"> perform an investigation to demonstrate the production and reception of radio waves 	205–6
	199–201	<ul style="list-style-type: none"> identify data sources, gather, process and analyse information and use available evidence to assess Einstein's contribution to quantum theory and its relation to black body radiation 	207–8
	205	<ul style="list-style-type: none"> identify data sources and gather, process and present information to summarise the use of the photoelectric effect in: <ul style="list-style-type: none"> solar cells photocells 	
	204	<ul style="list-style-type: none"> solve problems and analyse information using: $E = hf$ and $c = f\lambda$ 	201, 209–10
	201–5	<ul style="list-style-type: none"> process information to discuss Einstein and Planck's differing views about whether science research is removed from social and political forces 	206

3. Limitations of past technologies and increased research into the structure of the atom resulted in the invention of transistors

<i>Students learn to:</i>	page 213–14	<i>Students:</i>	page 230, 231
<ul style="list-style-type: none"> identify that some electrons in solids are shared between atoms and move freely describe the difference between conductors, insulators and semiconductors in terms of band structures and relative electrical resistance identify absences of electrons in a nearly full band as holes, and recognise that both electrons and holes help to carry current compare qualitatively the relative number of free electrons that can drift from atom to atom in conductors, semiconductors and insulators identify that the use of germanium in early transistors is related to lack of ability to produce other materials of suitable purity describe how 'doping' a semiconductor can change its electrical properties identify differences in p- and n-type semiconductors in terms of the relative number of negative charge carriers and positive holes describe differences between solid state and thermionic devices and discuss why solid state devices replaced thermionic devices 	213–20	<ul style="list-style-type: none"> perform an investigation to model the behaviour of semiconductors, including the creation of a hole or positive charge on the atom that has lost the electron and the movement of electrons and holes in opposite directions when an electric field is applied across the semiconductor 	225–6, 231
	216–17,		
	219–20		
	213–15		
	218–19		
	217, 219–20		
	219–20		
	220–5		

4. Investigations into the electrical properties of particular metals at different temperatures led to the identification of superconductivity and the exploration of possible applications

<i>Students learn to:</i>	page 236–9	<i>Students:</i>	page 241–2
<ul style="list-style-type: none"> outline the methods used by the Braggs to determine crystal structure identify that metals possess a crystal lattice structure describe conduction in metals as a free movement of electrons unimpeded by the lattice identify that resistance in metals is increased by the presence of impurities and scattering of electrons by lattice vibrations describe the occurrence in superconductors below their critical temperature of a population of electron pairs unaffected by electrical resistance discuss the BCS theory discuss the advantages of using superconductors and identify limitations to their use 	239	<ul style="list-style-type: none"> process information to identify some of the metals, metal alloys and compounds that have been identified as exhibiting the property of superconductivity and their critical temperatures 	253–4
	240	<ul style="list-style-type: none"> perform an investigation to demonstrate magnetic levitation 	240–1, 245
	240	<ul style="list-style-type: none"> analyse information to explain why a magnet is able to hover above a superconducting material that has reached the temperature at which it is superconducting 	248–9
	243–6	<ul style="list-style-type: none"> gather and process information to describe how superconductors and the effects of magnetic fields have been applied to develop a maglev train 	
	243–4	<ul style="list-style-type: none"> process information to discuss possible applications of superconductivity and the effects of those applications on computers, generators and motors and transmission of electricity through power grids 	246–8
	240–2,		
	246–50		

Option module: ASTROPHYSICS (chapters 14–17, pages 255–338)

1. Our understanding of celestial objects depends upon observations made from Earth or space near the Earth

<i>Students learn to:</i>		<i>Students:</i>	
<ul style="list-style-type: none"> discuss Galileo's use of the telescope to identify features of the Moon discuss why some wavebands can be more easily detected from space define the terms 'resolution' and 'sensitivity' of telescopes discuss the problems associated with ground-based astronomy in terms of resolution and absorption of radiation and atmospheric distortion outline methods by which the resolution and/or sensitivity of ground-based systems can be improved, including: <ul style="list-style-type: none"> adaptive optics interferometry active optics 	page 257–8 258–60 262–4 265 265–8	<ul style="list-style-type: none"> identify data sources, plan, choose equipment or resources for, and perform an investigation to demonstrate why it is desirable for telescopes to have a large diameter objective lens or mirror in terms of both sensitivity and resolution 	page 272

2. Careful measurement of a celestial object's position, in the sky, (astrometry) may be used to determine its distance

<i>Students learn to:</i>		<i>Students:</i>	
<ul style="list-style-type: none"> define the terms parallax, parsec, light-year explain how trigonometric parallax can be used to determine the distance to stars discuss the limitations of trigonometric parallax measurements 	page 275–6 275–7 277–8	<ul style="list-style-type: none"> solve problems and analyse information to calculate the distance to a star given its trigonometric parallax using: $d = \frac{1}{p}$ gather and process information to determine the relative limits to trigonometric parallax distance determinations using recent ground-based and space-based telescopes 	page 277, 299–300 302–3

3. Spectroscopy is a vital tool for astronomers and provides a wealth of information

<i>Students learn to:</i>		<i>Students:</i>	
<ul style="list-style-type: none"> account for the production of emission and absorption spectra and compare these with a continuous black body spectrum describe the technology needed to measure astronomical spectra identify the general types of spectra produced by stars, emission nebulae, galaxies and quasars describe the key features of stellar spectra and describe how these are used to classify stars describe how spectra can provide information on surface temperature, rotational and translational velocity, density and chemical composition of stars 	page 280–2 279–80 285 286–7 288–9	<ul style="list-style-type: none"> perform a first-hand investigation to examine a variety of spectra produced by discharge tubes, reflected sunlight, or incandescent filaments analyse information to predict the surface temperature of a star from its intensity/wavelength graph 	page 303 300, refer 281–2

4. Photometric measurements can be used for determining distance and comparing objects

<i>Students learn to:</i>		<i>Students:</i>	
<ul style="list-style-type: none"> define absolute and apparent magnitude explain how the concept of magnitude can be used to determine the distance to a celestial object outline spectroscopic parallax explain how two-colour values (i.e. colour index, B-V) are obtained and why they are useful describe the advantages of photoelectric technologies over photographic methods for photometry 	page 291 291–2 293–5 293–7 298	<ul style="list-style-type: none"> solve problems and analyse information using: $M = m - 5 \log\left(\frac{d}{10}\right)$ and $\frac{I_A}{I_B} = 100 \frac{m_B - m_A}{5}$ to calculate the absolute or apparent magnitude of stars using data and a reference star perform an investigation to demonstrate the use of filters for photometric measurements identify data sources, gather, process and present information to assess the impact of improvements in measurement technologies on our understanding of celestial objects 	page 291, 292, 293, 294–5, 300–1 303–4 289, 298

5. The study of binary and variable stars reveals vital information about stars

<i>Students learn to:</i>		<i>Students:</i>	
<ul style="list-style-type: none"> describe binary stars in terms of the means of their detection: visual, eclipsing, spectroscopic and astrometric explain the importance of binary stars in determining stellar masses classify variable stars as either intrinsic or extrinsic and periodic or non-periodic explain the importance of the period-luminosity relationship for determining the distance of cepheids 	page 306–10 306–8 312–314 315	<ul style="list-style-type: none"> perform an investigation to model the light curves of eclipsing binaries using computer simulation solve problems and analyse information by applying: $m_1 + m_2 = \frac{4\pi^2 r^3}{G\pi^2}$ 	page 318–19 308, 318

6. Stars evolve and eventually 'die'

<i>Students learn to:</i>		<i>Students:</i>	
<ul style="list-style-type: none"> describe the processes involved in stellar formation outline the key stages in a star's life in terms of the physical processes involved describe the types of nuclear reactions involved in Main Sequence and post-Main Sequence stars discuss the synthesis of elements in stars by fusion explain how the age of a globular cluster can be determined from its zero-age main sequence plot for an H–R diagram explain the concept of star death in relation to: <ul style="list-style-type: none"> planetary nebula supernovae white dwarfs neutron stars/pulsars black holes 	page 321–4 321–34, 335 324–8, 331 332 329–30 332–3	<ul style="list-style-type: none"> present information by plotting Hertzsprung–Russell diagrams for: nearby or brightest stars; stars in a young open cluster; stars in a globular cluster analyse information from an H–R diagram and use available evidence to determine the characteristics of a star and its evolutionary stage present information by plotting on an H–R diagram the pathways of stars of 1, 5 and 10 solar masses during their life cycle 	page 330 335, 338 335

Option module: MEDICAL PHYSICS (chapters 18–21, pages 339–416)

1. The properties of ultrasound waves can be used as diagnostic tools

<i>Students learn to:</i>	page	<i>Students:</i>	page
<ul style="list-style-type: none"> identify the differences between ultrasound and sound in normal hearing range describe the piezoelectric effect and the effect of using an alternating potential difference with a piezoelectric crystal define acoustic impedance: $Z = \rho v$ and identify that different materials have different acoustic impedances describe how the principles of acoustic impedance and reflection and refraction are applied to ultrasound define the ratio of reflected to initial intensity as: $\frac{I_r}{I_0} = \frac{[Z_2 - Z_1]^2}{[Z_2 + Z_1]^2}$	341–2 347 344–5 344–6 345–6 345–6 348–51 352–5 354–5	<ul style="list-style-type: none"> solve problems and analyse information to calculate the acoustic impedance of a range of materials, including bone, muscle, soft tissue, fat, blood and air and explain the types of tissues that ultrasound can be used to examine gather secondary information to observe at least two ultrasound images of body organs identify data sources and gather information to observe the flow of blood through the heart from a Doppler ultrasound video image identify data sources, gather, process and analyse information to describe how ultrasound is used to measure bone density solve problems and analyse information using: $Z = \rho v$ $\frac{I_r}{I_0} = \frac{[Z_2 - Z_1]^2}{[Z_2 + Z_1]^2}$	343, 358 344, 350, 355–6 355, 360 351–2 344–6, 358–9
<ul style="list-style-type: none"> identify that the greater the difference in acoustic impedance between two materials, the greater is the reflected proportion of the incident pulse describe the situations in which A scans, B scans, and phase and sector scans would be used and the reasons for the use of each describe the Doppler effect in sound waves and how it is used in ultrasonics to obtain flow characteristics of blood moving through the heart outline some cardiac problems that can be detected through the use of the Doppler effect 			

2. The physical properties of electromagnetic radiation can be used as diagnostic tools

<i>Students learn to:</i>	page	<i>Students:</i>	page
<ul style="list-style-type: none"> describe how X-rays are currently produced compare the differences between 'soft' and 'hard' X-rays explain how a computed axial tomography (CT) scan is produced describe circumstances where a CT scan would be a superior diagnostic tool compared to either X-rays or ultrasound explain how an endoscope works in relation to total internal reflection discuss differences between the role of coherent and incoherent bundles of fibres in an endoscope explain how an endoscope is used in: <ul style="list-style-type: none"> observing internal organs obtaining tissue samples of internal organs for further testing 	362–3 366 368–70 371–7 373–7 375 376–7	<ul style="list-style-type: none"> gather information to observe at least one image of a fracture on an X-ray film and X-ray images of other body parts gather secondary information to observe a CT scan image and compare the information provided by CT scans to that provided by an X-ray image for the same body part perform a first-hand investigation to demonstrate the transfer of light by optical fibres gather secondary information to observe internal organs from images produced by an endoscope 	361, 367, 368 371–2, 377 380 377, 379

3. Radioactivity can be used as a diagnostic tool

<i>Students learn to:</i>	page	<i>Students:</i>	page
<ul style="list-style-type: none"> outline properties of radioactive isotopes and their half lives that are used to obtain scans of organs describe how radioactive isotopes may be metabolised by the body to bind or accumulate in the target organ identify that, during decay of specific radioactive nuclei, positrons are given off discuss the interaction of electrons and positrons resulting in the production of gamma rays describe how the positron emission tomography (PET) technique is used for diagnosis 	382–4 384, 385 392–3 392 393–5	<ul style="list-style-type: none"> perform an investigation to compare an image of bone scan with an X-ray image gather and process secondary information to compare a scanned image of at least one healthy body part or organ with a scanned image of its diseased counterpart 	390, 397 388, 390, 391

4. The magnetic field produced by nuclear particles can be used as a diagnostic tool

<i>Students learn to:</i>	page	<i>Students:</i>	page
<ul style="list-style-type: none"> identify that the nuclei of certain atoms and molecules behave as small magnets identify that protons and neutrons in the nucleus have properties of spin and describe how net spin is obtained explain that the behaviour of nuclei with a net spin, particularly hydrogen, is related to the magnetic field they produce describe the changes that occur in the orientation of the magnetic axis of nuclei before and after the application of a strong magnetic field define precessing and relate the frequency of the precessing to the composition of the nuclei and the strength of the applied external magnetic field discuss the effect of subjecting precessing nuclei to pulses of radio waves explain that the amplitude of the signal given out when precessing nuclei relax is related to the number of nuclei present explain that large differences would occur in the relaxation time between tissue containing hydrogen-bound water molecules and tissues containing other molecules 	399–400 400–1 400–1 403–4 404–5 405–8 408–9 409–10	<ul style="list-style-type: none"> perform an investigation to observe images from magnetic resonance image (MRI) scans, including a comparison of healthy and damaged tissue identify data sources, gather, process and present information using available evidence to explain why MRI scans can be used to: <ul style="list-style-type: none"> detect cancerous tissues identify areas of high blood flow distinguish between grey and white matter in the brain gather and process secondary information to identify the function of the electromagnet, radio frequency oscillator, radio receiver and computer in the MRI equipment identify data sources, gather and process information to compare the advantages and disadvantages of X-rays, CT scans, PET scans and MRI scans gather, analyse information and use available evidence to assess the impact of medical applications of physics on society 	408, 410, 411, 414 410–13 402 412–13, 415 415

Option module: From QUANTA TO QUARKS (chapters 22–26, pages 417–520)

1. Problems with the Rutherford model of the atom led to the search for a model that would better explain the observed phenomena

<i>Students learn to:</i>	page	<i>Students:</i>	page
<ul style="list-style-type: none"> discuss the structure of the Rutherford model of the atom, the existence of the nucleus and electron orbits analyse the significance of the hydrogen spectrum in the development of Bohr's model of the atom define Bohr's postulates discuss Planck's contribution to the concept of quantised energy describe how Bohr's postulates led to the development of a mathematical model to account for the existence of the hydrogen spectrum: $\frac{1}{\lambda} = R \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$ <ul style="list-style-type: none"> discuss the limitations of the Bohr model of the hydrogen atom 	421–3 424 427–8 423 429–33 434	<ul style="list-style-type: none"> perform a first-hand investigation to observe the visible components of the hydrogen spectrum process and present diagrammatic information to illustrate Bohr's explanation of the Balmer series solve problems and analyse information using: $\frac{1}{\lambda} = R \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$ <ul style="list-style-type: none"> analyse secondary information to identify the difficulties with the Rutherford-Bohr model, including its inability to completely explain: <ul style="list-style-type: none"> the spectra of larger atoms the relative intensity of spectral lines the existence of hyperfine spectral lines the Zeeman effect 	437–9 433–4 425, 435 434

2. The limitations of classical physics gave birth to quantum physics

<i>Students learn to:</i>	page	<i>Students:</i>	page
<ul style="list-style-type: none"> describe the impact of de Broglie's proposal that any kind of particle has both wave and particle properties define diffraction and identify that interference occurs between waves that have been diffracted describe the confirmation of de Broglie's proposal by Davisson and Germer explain the stability of the electron orbits in the Bohr atom using de Broglie's hypothesis 	444–6 441–3 446 446–7	<ul style="list-style-type: none"> solve problems and analyse information using: $\lambda = \frac{h}{mv}$ <ul style="list-style-type: none"> gather, process, analyse and present information and use available evidence to assess the contributions made by Heisenberg and Pauli to the development of atomic theory 	445, 452 448–51

3. The work of Chadwick and Fermi in producing artificial transmutations led to practical applications of nuclear physics

<i>Students learn to:</i>	page	<i>Students:</i>	page
<ul style="list-style-type: none"> define the components of the nucleus (protons and neutrons) as nucleons and contrast their properties discuss the importance of conservation laws to Chadwick's discovery of the neutron define the term 'transmutation' describe nuclear transmutations due to natural radioactivity describe Fermi's initial experimental observation of nuclear fission discuss Pauli's suggestion of the existence of the neutrino and relate it to the need to account for the energy distribution of electrons emitted in β-decay evaluate the relative contributions of electrostatic and gravitational forces between nucleons account for the need for the strong nuclear force and describe its properties explain the concept of a mass defect using Einstein's equivalence between mass and energy describe Fermi's demonstration of a controlled nuclear chain reaction in 1942 compare requirements for controlled and uncontrolled nuclear chain reactions 	457 460 456 456 476–8 461–5 467 467–8 468–70 481–2 482–7	<ul style="list-style-type: none"> perform a first-hand investigation or gather secondary information to observe radiation emitted from a nucleus using a Wilson cloud chamber or similar detection device solve problems and analyse information to calculate the mass defect and energy released in natural transmutation and fission reactions 	518–20 470–1, 473, 494

4. An understanding of the nucleus has led to large science projects and many applications

<i>Students learn to:</i>	page	<i>Students:</i>	page
<ul style="list-style-type: none"> explain the basic principles of a fission reactor describe some medical and industrial applications of radioisotopes describe how neutron scattering is used as a probe by referring to the properties of neutrons identify ways by which physicists continue to develop their understanding of matter, using accelerators as a probe to investigate the structure of matter discuss the key features and components of the standard model of matter, including quarks and leptons 	484–8 489–91 492 496–502 503–13	<ul style="list-style-type: none"> gather, process and analyse information to assess the significance of the Manhattan Project to society identify data sources, and gather, process and analyse information to describe the use of: <ul style="list-style-type: none"> a named isotope in medicine a named isotope in agriculture a named isotope in engineering 	480–4 489–91

ACKNOWLEDGEMENTS

The authors would like to thank the following people for their support during the writing of this book: Michael Andriessen gives special thanks to Christine, Sam and Luke for their understanding and patience; Peter Pentland wishes especially to thank his wife Helen Kennedy for her continuing support; Bruce McKay is indebted to close friend and former colleague Barry Mott for his valuable advice; Richard Gaut thanks Stephen and Greta for their helpful suggestions and feedback; Jill Tacon wishes to thank Dr Manjula Sharma and Dr Joe Khachan from the University of Sydney for their valuable advice and Lee Collins from Westmead Hospital for his expert assistance. Yoka McCallum's contribution and encouragement is also appreciated, and thanks must go to Dean Bunn for his permission to adapt some practical activities and other material from *Physics for a Modern World*.

The authors and publisher are grateful to the following individuals and organisations for their permission to reproduce photographs and other copyright material.

Images

- © Alan Bean/Novaspace Galleries: **2.2** • AIP Emilio Segrè Visual Archives, W.F. Meggers Collection: **22.8** • © ANSTO: **20.3, 20.4**/Gentech® generator courtesy of ANSTO • Otto Rogge/ANTPhoto.com.au: **11.21** • Bhathal, R., *Astronomy for the HSC*, Kangaroo Press 1993, p. 47. Reproduced by permission of Ragbir Bhathal: **17.12(a–c)** • © Black & Decker: **9.1**
- © Boeing: **3.9** • The Royal Institution, London, UK/Bridgeman Art Library: **7.1** • Courtesy of Brookhaven National Laboratory: **26.8(a, b)** • Cavendish Laboratory, Cambridge University: **10.1** • © University of Queensland, Centre for Magnetic Resonance: **21.15(a–c)** • Chicago Historical Society, *Birth of the Atomic Age*, Painting by Gary Sheahan, 1957: **25.5**
- © Corbis Australia/Australian Picture Library: **13.1; 18.16/Belt/Corbis/Annie Griffiths; 8.17/Corbis; 8.15, 8.16, 11.2, 11.5, 11.12, 23.11, 23.12, 23.13, 23.14, 23.15/Corbis/Bettman; 21.14(d)/Corbis/Howard Sochurek; 7.4/Corbis/Hulton-Deutsch Collection** • © CERN: **26.2** • © CFHT, 1996 Used with permission: **14.17** • Adapted from *Physics in Medical Diagnosis*, Chapman & Hall 1997, p. 213 fig. 5.23. Reproduced with permission of the author Dr Trevor A Delchar: **18.14(a)**
- © Digital Stock/Corbis Corporation: page **99; 2.4, 13.18**
- © Digital Vision: **2.1** • © Dorling Kindersley: **6.1** • Image courtesy of Elaine Collin: **19.19** • © Corbis/Greg Smith: **11.22** • Image courtesy of the European Space Agency © ESA: **14.13, 15.6** • Courtesy of EMI Archives: **19.10** • Fermilab National Accelerator Laboratory: **26.1, 26.7, 26.10** • Dave Finley, National Radio Astronomy Observatory, courtesy Associated Universities, Inc., and the National Science Foundation: **14.14**
- © Fundamental Photographs, New York: **15.8(a); 15.14, 22.9, 23.6/Wabash Instrument Corp.; 23.2(a)/Ken Kay** • © Getty Images/ Hulton Archive: **14.2, 22.2, 23.8** • David Grabham: **19.7, 21.14(c), 21.18(a, b)** • *Medical Physics* by J Pope, p. 81, fig 4.13 Reprinted by permission of Harcourt Education Ltd: **21.13**
- Reprinted by permission of HarperCollins Publishers Ltd. © Illingworth 1994: **15.21, 16.10** • Terry Herter: **16.15, 16.16**
- Adapted and reproduced from Martin Hollins, *Medical Physics*, 2nd edition, Nelson Thornes 2001, pp. 114, 122, 127, 100, 101, 186, 197: **18.6, 18.17, 19.6, 19.16, 19.17, 20.7(b), 21.6** • 'The Fermion Particles' and 'The Boson Force Carriers', *From Quarks to the Cosmos: Tools of Discovery* by Leon M Lederman and David N Schramm. © 1989, 1995 Scientific American Library. Reprinted by permission of Henry Holt and Company, LLC: **table 26.4**
- Courtesy of Hologic: **18.11** • © www.imageaddict.com.au: **14.8, 14.10** • ImageState: **10.5** • Kamioka Observatory, ICRR (Institute for Cosmic Ray Research), The University of Tokyo: **24.1** • *IEEE Review* March 2001, Vol 47, 102, p 42. Reproduced with permission of IET. www.theiet.org: **13.25** • Jared Schneidman Design: **12.24** • JAS Photography/John Sowden: **2.3** • Kenneth Krane,

- Modern Physics 2nd edition*, 1996, p. 24. Used by permission of John Wiley & Sons, Inc.: **5.4** • Resnick, R, *Introduction to Special Relativity*, figures 1.4 and 1.6, John Wiley & Sons Inc., © 1968: **5.5** • Halliday et al., *Fundamentals of Physics Extended*, 5th edition, John Wiley & Sons, Inc. 1997, figures 37.27 (b, c and d), 43.3, 37.22, 43.6. Used by permission of John Wiley & Sons, Inc.: **13.9, 13.12, 22.5, 22.12, 24.15** • Webster, J G (ed), *Medical Instrumentation*, 3rd edition, 1998, p. 559, adapted and used by permission of John Wiley & Sons, Inc.: **20.7(a)** • Adapted from 'The Particle Adventure', produced by the Particle Data Group, Lawrence Berkeley National Laboratory: **26.3, 26.4** • Bruce McKay: **13.4, 22.1, 23.2(b, c), 26.11** • David Malin Images: **15.1** • Mary Evans Picture Library: **14.3** • *Cambridge Encyclopedia of Astronomy*, ed. by Dr Simon Mitton, Jonathon Cape, 1977, © Cambridge University. Reproduced with permission of Simon Mitton: **15.8(b), 17.13** • Barbara Mochejska (Copernicus Astronomical Center), Andrew Szentgyorgyi (Harvard-Smithsonian Center for Astrophysics), F. L. Whipple Observatory: **17.11(b)** • NASA: **2.28, 3.1/MSFC, 3.6, 3.8, 4.8/NSSDC, 14.1/STSCI, 14.7, 17.3/NOAO, 17.16/The Hubble Heritage Team, 17.18(a)/ESA/ASU/J Hester and A Loll, 17.18(b)/CXC/ASU/J Hester et al.** • © Newspix: **20.13/Jody D'Arcy, 25.2** • © Moriel NessAiver: **21.2** • © Department of Nuclear Medicine, The Queen Elizabeth Hospital, South Australia: **20.7(d)** • De Jong et al., *Heinmann Physics Two*, p. 237. Reproduced with permission from Pearson Education Australia: **13.2** • Adapted from David Heffernan, *Physics Contexts 2*, Pearson Education Australia 2002, p. 315. Reproduced with permission of Pearson Education Australia: **18.8** • Peter Pentland: **9.9** • © Philips Medical Systems: **18.3(b), 19.11** • © Photodisc, Inc.: pages **1, 173, 255, 339, 417; 1.1, 4.1, 8.1, 8.24(a, b), 10.22, 11.19, 17.1, 18.1, 18.3(c), 19.9(a), 20.15(a), 21.17, 25.1** • photolibrary.com: **9.4/Tom Marexchal, 13.26(b)/Photo Researchers Inc., 17.11(a)/John Chumack; 18.15(a, b), 19.14, 20.9, 21.1, 21.3, 21.14(a, b), 23.1/Phototake; 19.8/Phototake Science; 19.9(b), 20.14/Phototake Science/Science Ltd; 20.1/Phototake Science/ISM**- photolibrary.com/Science Photo Library: **4.2, 5.1, 7.2, 14.5, 15.22, 19.1, 22.7; 3.11/NASA; 10.26(b)/Peter Apraham; 11.1, 14.6/David Nunuk; 12.1/John Walsh; 12.23/Martin Dohrn; 12.25(b)/Andrew Syred; 18.7/P Saada/Eurelios; 20.5/Philippe Plailly; 20.6(a)/Dr P Marazzi; 20.6(b, c)/CNRI; 20.11/Zephyr; 20.15(b)/Hank Morgan; 21.16/Dept of Cognitive Neurology;**
- © The Picture Source/Terry Oakley: **6.24, 12.12(b), 12.22, 12.25(a)** • © Greg Pitt: **20.7(c)** • Science Museum/Science & Society Picture Library: **24.8** • Reprinted from *Medical Physics*, 1978, John Wiley & Sons Inc. with permission from the authors, John R Cameron and James G Skofronick © 1992: **19.5(a, b)**
- Peter Storer: **8.22** • Sudbury Neutrino Observatory/R Chambers: **24.13** • © SP-AusNet: **8.21** • Courtesy of Transrapid International-USA, Inc.: **13.26(a)** • Reproduced with kind permission of TransGrid: **8.20** • Photo © UC Regents/Lick Observatory: **16.1** • © Richard Wainscoat: **14.19** • Jack Washburn: **13.11(b)** • Westmead Hospital Medical Physics Department: **18.14(b), 20.10(a, b), 20.12(a, b), 20.16** • Justine Wong: **19.13(a-d)** • Courtesy of Xerox Corporation: **10.9**

Text

- Extracts from Physics Stage 6 Syllabus © Board of Studies NSW, 2002 (pages **x–xv**) • Data and formulae sheets from Physics Higher School Certificate Examination © Board of Studies NSW (pages **526–7**) • Higher School Certificate Assessment Support Document © Board of Studies NSW, 1999 (key words, pages **529–30**)

Every effort has been made to trace the ownership of copyright material. Information that will enable the publisher to rectify any error or omission in subsequent editions will be welcome. In such cases, please contact the Permissions Section of John Wiley & Sons Australia, Ltd, who will arrange for the payment of the usual fee.



Chapter 1

Earth's gravitational field

Chapter 2

Launching into space

Chapter 3

Orbiting and re-entry

Chapter 4

Gravity in the solar system

Chapter 5

Space and time

SPACE

CHAPTER

1 EARTH'S GRAVITATIONAL FIELD



Remember

Before beginning this chapter, you should be able to:

- recall and apply Newton's Second Law of Motion:
 $F = ma$.

Key content

At the end of this chapter you should be able to:

- make a comparison between the acceleration due to gravity at various places over the Earth's surface as well as at other locations throughout the solar system
- define weight as the force on an object due to a gravitational field
- explain that work done to raise or lower a mass in a gravitational field is directly related to a change in the gravitational potential energy of the mass
- calculate the weight of a body on Earth, above the Earth or on other planets
- define gravitational potential energy as the work done in moving an object from a very large distance away to a point in a gravitational field.

Figure 1.1 The Earth rising as seen from the Moon

Gravity is a force of attraction that exists between any two masses. Usually this is a very small, if not negligible, force. However, when one or both of the masses is as large as a planet, then the force becomes very significant indeed. The force of attraction between the Earth and our own bodies is the force we call our *weight*. This force exists wherever we are on or near the Earth's surface (although, as we shall see, with some variation). We can say that a gravitational field exists around the Earth and we live within that field.

1.1 THE EARTH'S GRAVITY

The Earth is surrounded by a gravitational field. This type of field, discussed in more general terms in chapter 4, is a vector field within which a mass will experience a force. (Other vector fields include electric and magnetic fields.) The gravitational field around the Earth can be drawn as shown in figure 1.2. Note that the direction of a field line at any point is the direction of the force experienced by a mass placed at that point.

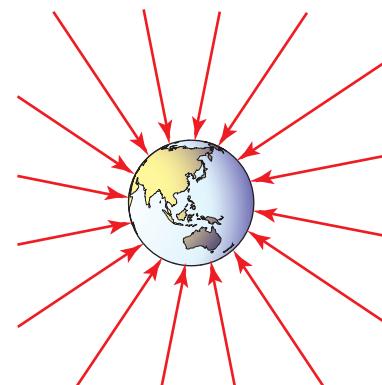


Figure 1.2 The gravitational field around the Earth

The field vector: \mathbf{g}

A **field vector** is a single vector that describes the strength and direction of a uniform vector field. For a gravitational field, the field vector is \mathbf{g} .

A **field vector** is a single vector that describes the strength and direction of a uniform vector field. For a gravitational field the field vector is \mathbf{g} , which is defined in this way:

$$\mathbf{g} = \frac{\mathbf{F}}{m}$$

where

\mathbf{F} = force exerted (N) on mass m

m = mass (kg) in the field

\mathbf{g} = the field vector (N kg^{-1}).

Vector symbols are indicated here in bold italics. The direction of the vector \mathbf{g} is the same as the direction of the associated force.

Note that a net force applied to a mass will cause it to accelerate. Newton's Second Law describes this relationship:

$$\mathbf{a} = \frac{\mathbf{F}}{m}$$

where

\mathbf{a} = acceleration (m s^{-2}).

Hence, we can say that the field vector \mathbf{g} also represents the acceleration due to gravity and we can calculate its value at the Earth's surface as described below.

The Law of Universal Gravitation (discussed in more detail in chapter 4) says that the magnitude of the force of attraction between the Earth and an object on the Earth's surface is given by:

$$F = G \frac{m_E m_O}{r_E^2}$$



1.1

Using a pendulum to determine \mathbf{g}

where

$$\begin{aligned}m_E &= \text{the mass of the Earth} \\&= 5.97 \times 10^{24} \text{ kg} \\m_O &= \text{mass of the object (kg)} \\r_E &= \text{radius of the Earth} \\&= 6.38 \times 10^6 \text{ m.}\end{aligned}$$

From the defining equation for g , on the previous page, we can see that the force experienced by the mass can also be described by:

$$F = m_O g.$$

$$\text{Equating these two we get: } m_O g = G \frac{m_E m_O}{r_E^2}.$$

$$\begin{aligned}\text{This simplifies to give: } g &= G \frac{m_E}{r_E^2} \\&= \frac{(6.672 \times 10^{-11})(5.974 \times 10^{24})}{(6.378 \times 10^6)^2}.\end{aligned}$$

$$\text{Hence, } g \approx 9.80 \text{ m s}^{-2}.$$

The value of the Earth's radius used here, 6378 km (at the Equator), is an average value so the value of g calculated, 9.80 m s^{-2} , also represents an average value.

Variations in the value of g

Variation with geographical location

The actual value of the acceleration due to gravity, g , that will apply in a given situation will depend upon geographical location. Minor variations in the value of g around the Earth's surface occur because:

- the Earth's crust or lithosphere shows variations in thickness and structure due to factors such as tectonic plate boundaries and dense mineral deposits. These variations can alter local values of g .
- the Earth is not a perfect sphere, but is flattened at the poles. This means that the value of g will be greater at the poles, since they are closer to the centre of the Earth.
- the spin of the Earth creates a centrifuge effect that reduces the effective value of g . The effect is greatest at the Equator and there is no effect at the poles.

As a result of these factors, the rate of acceleration due to gravity at the surface of the Earth varies from a minimum value at the Equator of 9.782 m s^{-2} to a maximum value of 9.832 m s^{-2} at the poles. The usual value used in equations requiring g is 9.8 m s^{-2} .

Variation with altitude

The formula for g shows that the value of g will also vary with altitude above the Earth's surface. By using a value of r equal to the radius of the Earth plus altitude, the following values can easily be calculated. It is clear from table 1.1 that the effect of the Earth's gravitational field is felt quite some distance out into space.

$$\text{The formula used is: } g = G \frac{m_E}{(r_E + \text{altitude})^2}.$$

Note that as altitude increases the value of g decreases, dropping to zero only when the altitude has an infinite value.

Table 1.1 The variation of g with altitude above Earth's surface

ALTITUDE (km)	g (m s^{-2})	COMMENT
0	9.80	Earth's surface
8.8	9.77	Mt Everest Summit
80	9.54	Arbitrary beginning of space
200	9.21	Mercury capsule orbit altitude
250	9.07	Space shuttle minimum orbit altitude
400	8.68	Space shuttle maximum orbit altitude
1 000	7.32	Upper limit for low Earth orbit
40 000	0.19	Communications satellite orbit altitude

Variation with planetary body



1.2

Weight values in the solar system and g

The formula for g also shows that the value of g depends upon the mass and radius of the central body which, in examples so far, has been the Earth. Other planets and natural satellites (moons) have a variety of masses and radii, so that the value of g elsewhere in our solar system can be quite different from that on Earth. The following table presents a few examples.

$$\text{The formula used here is: } g = G \frac{m_{\text{planet}}}{r_{\text{planet}}^2}.$$

Table 1.2 A comparison of g on the surface of other planetary bodies

BODY	MASS (kg)	RADIUS (km)	g (m s^{-2})
Moon	7.35×10^{22}	1 738	1.6
Mars	6.42×10^{23}	3 397	3.7
Jupiter	1.90×10^{27}	71 492	24.8
Pluto	1.31×10^{22}	1 151	0.66

SAMPLE PROBLEM

1.1

Determining acceleration due to gravity above the Moon

For each of the *Apollo* lunar landings, the command module continued orbiting the Moon at an altitude of about 110 km, awaiting the return of the Moon walkers. Determine the value of acceleration due to gravity at that altitude above the surface of the Moon (the radius of the Moon is 1738 km).

SOLUTION

$$\begin{aligned}
 g &= G \frac{m_M}{(r_M + \text{altitude})^2} \\
 &= \frac{(6.67 \times 10^{-11})(7.35 \times 10^{22})}{(1.738 \times 10^6 + 1.1 \times 10^5)^2} \\
 &= 1.4 \text{ m s}^{-2}
 \end{aligned}$$

That is, the acceleration due to gravity operating on the orbiting command module was approximately 1.4 m s^{-2} .

1.2 WEIGHT

Weight is defined as the force on a mass due to the gravitational field of a large celestial body, such as the Earth.

Weight is defined as the force on a mass due to the gravitational field of a large celestial body, such as the Earth. Since it is a force, it is measured in newtons. We can use Newton's Second Law to define a simple formula for weight:

Newton's Second Law states:

$$F = ma$$

and hence:

$$W = mg$$

where

W = weight (N)

m = mass (kg)

g = acceleration due to gravity at that place (m s^{-2}).

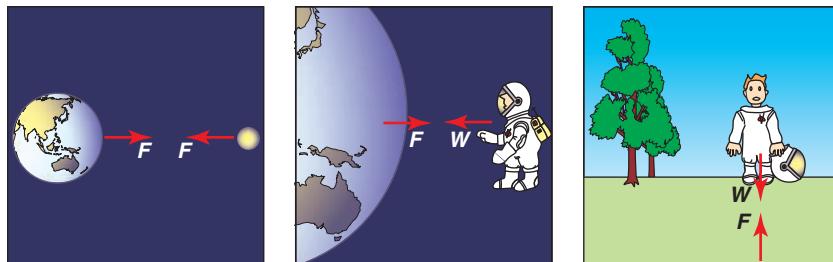


Figure 1.3 There is always a gravitational force between any two masses. When one of the masses is as large as a planet, the force on a small mass is called weight.

SAMPLE PROBLEM

1.2

Determining the weight of an astronaut

What would be the weight of a 100 kg astronaut (a) on the Earth, (b) on the Moon and (c) in an orbiting space shuttle?

SOLUTION

Use the values of g shown in tables 1.1 and 1.2. Note that the astronaut's mass does not change with position but weight does.

(a) On the Earth $W_E = mg_E$
 $= 100 \times 9.8$
 $= 980 \text{ N}$

(b) On the Moon $W_M = mg_M$
 $= 100 \times 1.6$
 $= 160 \text{ N}$

(c) In orbit $W_O = mg_O$
 $= 100 \times 8.68$ (at maximum altitude)
 $= 868 \text{ N}$

There is an apparent contradiction in this last answer. The astronaut in orbit still has a considerable weight, rather than being weightless. However, the answer is correct because, as we have already seen, the Earth's gravitational field extends quite some distance out into space. Why then do space shuttle astronauts experience weightlessness? As we shall see later, the weightlessness they feel is not real but only apparent, and is a consequence of their orbital motion around the Earth.

1.3 GRAVITATIONAL POTENTIAL ENERGY

Gravitational potential energy, E_p , is the energy of a mass due to its position within a gravitational field. On a large scale, gravitational potential energy is defined as the work done to move an object from infinity (or some point very far away) to a point within a gravitational field.

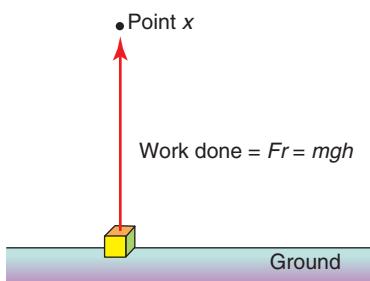


Figure 1.4 Gravitational potential energy, E_p = work done to move up to the point from the ground (the zero level)

Gravitational potential energy, E_p , is the energy of a mass due to its position within a gravitational field. Here on Earth, the E_p of an object at some point, x , above the ground is easily found as it is equal to the work done to move the object from the ground up to the point, x , as shown in figure 1.4.

$$\begin{aligned} \text{Gravitational potential energy } E_p &= \text{work done to move to the point} \\ &= \text{force required} \times \text{distance moved} \\ &\quad (\text{since work } W = Fr) \\ &= (mg) \times h = mgh \end{aligned}$$

Hence, in this case $E_p = mgh$. We chose the ground as our starting point because this is our defined zero level; that is, the place where $E_p = 0$. Note that since work must be done on the object to lift it, it acquires energy. Hence, at point x , E_p is greater than zero.

On a larger, planetary scale we need to rethink our approach. Due to the inverse square relationship in the Law of Universal Gravitation, the force of attraction between a planet and an object will drop to zero only at an infinite distance from the planet. For this reason we will now choose infinity (or some point a very large distance away) as our level of zero potential energy.

There is a strange side effect of our choice of zero level. Because gravitation is a force of attraction, work must be done on the object to move it from a point, x , to infinity; that is, against the field so that it gains energy, E_p .

Therefore, E_p at infinity $> E_p$ at point x
 but E_p at infinity $= 0$
 so that E_p at point $x < 0$
 that is, E_p at point x has a negative value! (see figure 1.5)

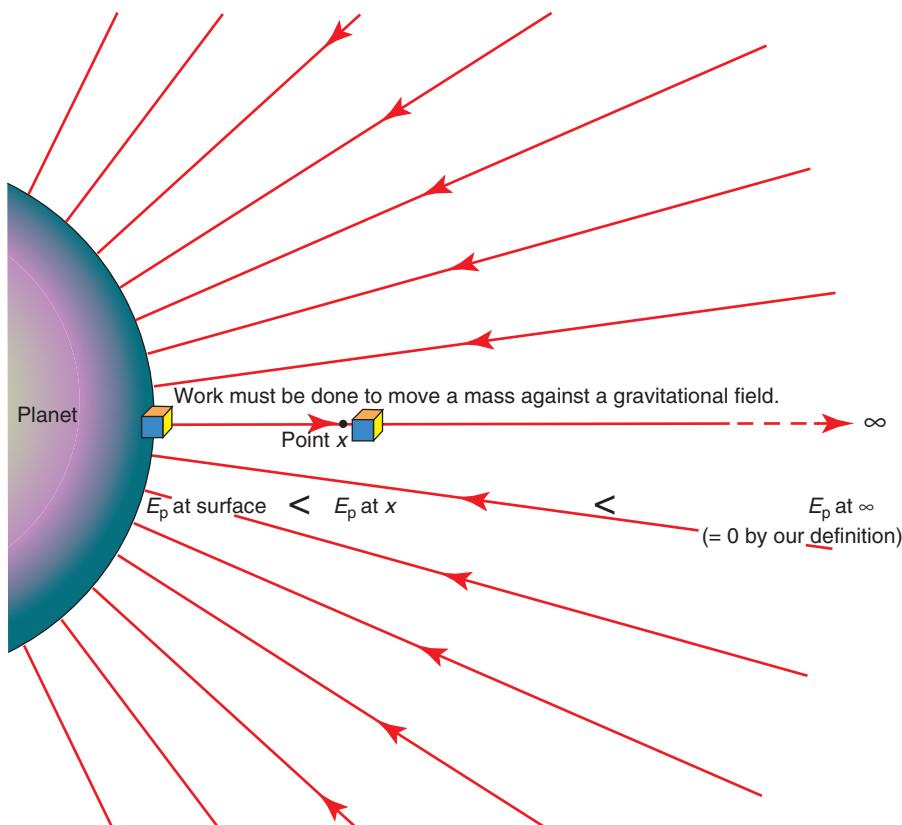


Figure 1.5 Different levels of E_p . If we choose a planet's surface as the zero level, E_p at x has a positive value. If infinity is chosen as the zero level, E_p has a negative value.

Using the same approach as earlier, the gravitational potential energy, E_p , of an object at a point, x , in a gravitational field is equal to the work done to move the object from the zero energy level at infinity (or some point very far away) to point x . It can be shown mathematically that:

$$E_p = -G \frac{m_1 m_2}{r}$$

where

m_1 = mass of planet (kg)

m_2 = mass of object (kg)

r = distance separating masses (m).

Figure 1.6 graphs this relationship to show how E_p varies in value in the space around a planet.

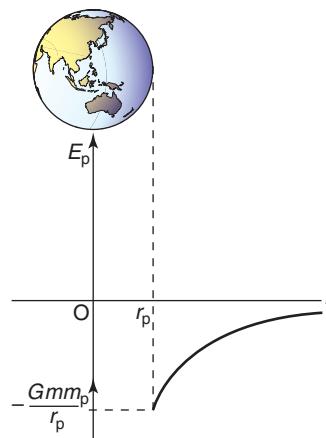


Figure 1.6 A graph showing how the negative value for gravitational potential energy, E_p , increases with distance up to a maximum value of zero

SAMPLE PROBLEM

1.3

Gravitational potential energy in the Sun–Earth–Moon system

Given the following data, determine the gravitational potential energy of:

- (a) the Moon within the Earth's gravitational field
- (b) the Earth within the Sun's gravitational field.

$$\text{Mass of the Earth} = 5.97 \times 10^{24} \text{ kg}$$

$$\text{Mass of the Moon} = 7.35 \times 10^{22} \text{ kg}$$

$$\text{Mass of the Sun} = 1.99 \times 10^{30} \text{ kg}$$

$$\text{Earth–Moon distance} = 3.84 \times 10^8 \text{ m on average}$$

$$\text{Earth–Sun distance} = 1.50 \times 10^{11} \text{ m on average (one astronomical unit, AU)}$$

SOLUTION

(a)

$$E_p = -G \frac{m_E m_M}{r}$$

$$= -\frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})(7.35 \times 10^{22})}{(3.84 \times 10^8)}$$

$$= -7.62 \times 10^{28} \text{ J}$$

That is, the gravitational potential energy of the Moon is approximately -7.62×10^{28} J. Put another way, the work that would be done in

moving the Moon from a very large distance away from Earth to its current distance would be -7.62×10^{28} J.

$$\begin{aligned}(b) \quad E_p &= -G \frac{m_E m_S}{r} \\ &= -\frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})(1.99 \times 10^{30})}{(1.50 \times 10^{11})} \\ &= -5.28 \times 10^{33} \text{ J}\end{aligned}$$

That is, the gravitational potential energy of the Earth is approximately -5.28×10^{33} J. The negative sign indicates that this would be work done by the system (not on the system) in moving the Earth from a very large distance away from the Sun to its present orbital distance. This negative work represents potential energy *lost* by the system as the Earth and the Sun are brought together (converted into other forms of energy, most probably kinetic). Since the E_p is reduced below the zero level (see figure 1.6), it is quite appropriate that it should appear as a negative value.

This missing energy actually lends stability to a system, since the Earth would need to get this amount of energy back from somewhere if ever it were to separate from the Sun. It can be thought of as binding energy, since the lack of this energy binds a system together.

SUMMARY

- The field vector g describes the strength and direction of the gravitational field at any point. At the surface of the Earth it has an average value of 9.8 N kg^{-1} or m s^{-2} .
 - The value of g at any specific point on the Earth's surface can vary from the average figure due to a number of factors. It will also vary with altitude.
 - Weight is the force on an object due to a significant gravitational field ($W = mg$).
 - The gravitational potential energy of an object at some point within a gravitational field is equivalent to the work done in moving the object from an infinite distance to that point.

$$E_p = -G \frac{m_1 m_2}{r}$$

QUESTIONS

- Define weight.
 - In general terms only, describe the variation in g that would be experienced in a spacecraft travelling directly from the planet Mars to its moon, Phobos, 9380 km away.
 - The gravitational field vector \mathbf{g} has an average value, on the surface of the Earth, of 9.8 N kg^{-1} or m s^{-2} . Show that the two alternative units quoted are equivalent.
 - Complete the following table to calculate the acceleration due to gravity and weight force experienced by an 80 kg person standing on the surface of each of the planets or moons indicated.
 - 6378 km at the Equator.
 - Define gravitational potential energy E_p .
 - Explain the reason for the selection of infinity as the place of zero gravitational potential energy.
 - Explain how this selection of zero level results in any point within a gravitational field having a negative gravitational potential energy.
 - Calculate the gravitational potential energy of a 1000 kg communications satellite orbiting the Earth at an altitude of 40 000 km. Use the data provided in question 7.
 - Use the following data to calculate the gravi-

BODY	MASS (kg)	RADIUS (km)	g ON SURFACE ($m\ s^{-2}$)	WEIGHT OF 80 kg PERSON THERE (N)
Mercury	3.30×10^{23}	2440		
Venus	4.87×10^{24}	6052		
Io	8.94×10^{22}	1821		
Callisto	1.08×10^{23}	2410		

5. The moon of Pluto, Charon (pronounced Kair-on), discovered in 1978, is one of the largest moons, in proportion to its planet or dwarf planet (as Pluto is), in the solar system.
 - (a) The mass of Charon is 1.62×10^{21} kg while the mass of Pluto is 1.31×10^{22} kg. Calculate the ratio of the mass of Charon to the mass of Pluto.
 - (b) The radius of Charon is 593 km and the radius of Pluto is 1151 km. Calculate the ratio of the radius of Charon to the radius of Pluto.
 - (c) Calculate the ratio of the density of Charon to the density of Pluto.
 - (d) Calculate the ratio of g on Charon to g on Pluto.
 6. Identify four different factors that cause the value of g to vary around the Earth.
 7. Construct a graph that shows the value of g each 5000 km above the surface of the Earth up to an altitude of 40 000 km (which corresponds to the altitude of communications satellites). The mass of the Earth is 5.97×10^{24} kg and the radius of the Earth is 6378 km at the Equator.
 8. Define gravitational potential energy E_p .
 9. Explain the reason for the selection of infinity as the place of zero gravitational potential energy.
 10. Explain how this selection of zero level results in any point within a gravitational field having a negative gravitational potential energy.
 11. Calculate the gravitational potential energy of a 1000 kg communications satellite orbiting the Earth at an altitude of 40 000 km. Use the data provided in question 7.
 12. Use the following data to calculate the gravitational potential energy of (a) Callisto as it orbits within Jupiter's gravitational field, and (b) Jupiter as it orbits within the Sun's gravitational field.
Mass of Jupiter = 1.90×10^{27} kg
Mass of Callisto = 1.08×10^{23} kg
Mass of the Sun = 1.99×10^{30} kg
Jupiter–Callisto distance = 1.88×10^9 m
on average
Jupiter–Sun distance = 7.78×10^{11} m
on average



1.1 USING A PENDULUM TO DETERMINE g

Aim

To determine the rate of acceleration due to gravity using the motion of a pendulum.

Apparatus

retort stand
bosshead and clamp
approximately 1 metre of string
50 g mass carrier or pendulum bob
stopwatch
metre rule

Theory

When a simple pendulum swings with a small angle, the mass on the end performs a good approximation of the back-and-forth motion called *simple harmonic motion*. The period of the pendulum, that is, the time taken to complete a single full back-and-forth swing, depends upon just two variables: the length of the string and the rate of acceleration due to gravity. The formula for the period is as shown below:

$$T = 2\pi \sqrt{\frac{l}{g}}$$

where

T = period of the pendulum (s)

l = length of the pendulum (m)

g = rate of acceleration due to gravity (m s^{-2}).

Method

- Set up the retort stand and clamp on the edge of a desk as shown in figure 1.7. Tie on the string and adjust its length to about 90 cm before attaching the 50 g mass carrier or pendulum bob to its end.

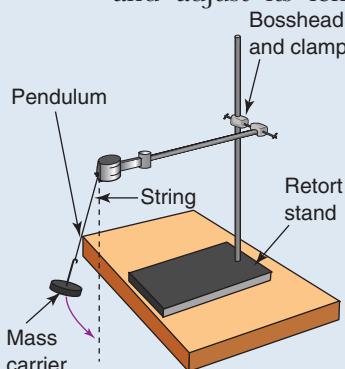


Figure 1.7 Apparatus for practical activity 1.1

- Using the metre rule, carefully measure the length of the pendulum from the knot at its top to the base of the mass carrier. Enter this length in your results table.
- Set the pendulum swinging gently (30° maximum deviation from vertical) and use

the stopwatch to time 10 complete back-and-forth swings. Be sure to start and stop the stopwatch at an extreme of the motion rather than somewhere in the middle. Enter your time for 10 swings in the results table.

- Repeat steps 2 and 3 at least five times, after shortening the string by 5 cm each time.

Results

Copy the table below into your practical book to record your results, and then complete the other columns of information.

TRIAL	TIME FOR 10 OSCILLATIONS (s)	PERIOD T (s)	PERIOD SQUARED T^2 (s ²)	LENGTH OF PENDULUM (m)
1				
2				
3				
4				
5				

Draw a graph of period squared versus length of the pendulum. Plot T^2 on the vertical axis and length on the horizontal axis.

Analysis

- Your graph should display a straight-line relationship. Draw a line of best fit and evaluate the gradient.
- Rearrange the pendulum equation given earlier to the form, $T^2 = kl$, where k is a combination of constants.
- Compare this formula with the general equation for a straight line: $y = kx$. This comparison shows that if T^2 forms the y -axis and length, l , forms the x -axis, the expression you derived for k in step 2 should correspond to the gradient of the graph you have drawn. Write down your expression:
gradient = _____ (complete).
- Use your expression to calculate a value for g , the acceleration due to gravity.

Questions

- This method usually produces very accurate results. Can you suggest a reason why it should be so reliable?
- What are the sources of error in this experiment?
- What could you do to improve the method of this experiment to make it even more accurate?



1.2 WEIGHT VALUES IN THE SOLAR SYSTEM AND g

Aim

To research g and weight values throughout the solar system.

Theory

The value of g on the surface of a planet depends upon the mass of the planet and its radius. The equation relating these variables is:

$$g = G \frac{m_{\text{planet}}}{r_{\text{planet}}^2}$$

Method

The table below lists the 16 most massive objects in our solar system, excluding the Sun, in descending

Results

A comparison of gravity throughout the solar system

BODY	CENTRE OF ORBIT	MASS (kg)	RADIUS (km)	g ON SURFACE (m s^{-2})	WEIGHT OF 100 kg PERSON ON SURFACE (N)
Jupiter	Sun				
Saturn	Sun				
Neptune	Sun				
Uranus	Sun				
Earth	Sun				
Venus	Sun				
Mars	Sun				
Mercury	Sun				
Ganymede	Jupiter				
Titan	Saturn				
Callisto	Jupiter				
Io	Jupiter				
Moon	Earth				
Europa	Jupiter				
Triton	Neptune				
Pluto	Sun				

order of mass. However, the mass values and radii have not been provided. Conduct research to determine these figures and then perform the calculations to complete the table as shown.

Analysis

Draw a bar graph of your results, with the bodies in their mass order along the horizontal axis, and acceleration due to gravity on the vertical axis. You may be surprised at some of the results.

Questions

- How does g on Jupiter compare with the rest of the plotted results?
- How does g on Saturn, Neptune, Uranus and Venus compare with g on Earth?
- How does g on Uranus compare with g on Venus?
- How does g on Mars compare with g on Mercury?
- How does g on all of the natural satellites (moons) listed compare with g on Pluto?
- There is some debate over whether Pluto should be downgraded in official status from a ‘planet’. Can you provide one good argument for each side of this debate?

CHAPTER 2

LAUNCHING INTO SPACE



Figure 2.1 The space shuttle launching from Cape Canaveral. It uses two solid-fuel rocket engines to supplement the thrust of its own liquid-fuel rocket engines.

Remember

Before beginning this chapter, you should be able to:

- describe the nature of velocity and calculate values using: $v = \frac{\Delta r}{t}$
- describe the nature of acceleration and calculate values using: $a = \frac{\Delta v}{\Delta t} = \frac{v - u}{t}$
- describe the nature of kinetic energy and calculate values using: $E_k = \frac{1}{2} mv^2$
- describe the nature of gravitational potential energy using: $E_p = -G \frac{m_1 m_2}{r}$
- describe the nature of momentum and calculate values using: $p = mv$
- apply Newton's Second Law of Motion: $\Sigma F = ma$.

Key content

At the end of this chapter you should be able to:

- describe the trajectory of a projectile in terms of horizontal and vertical components
- solve projectile motion problems that require you to analyse the motion and determine the velocity of the projectile at any time, the maximum height reached, the time of flight or the range of the motion
- describe Galileo's contribution to our understanding of projectile motion
- explain the concept of escape velocity
- outline Isaac Newton's concept of escape velocity
- identify why the term 'g forces' is used to describe forces acting upon an astronaut during a typical launch or re-entry
- discuss the effect of the Earth's orbital and rotational motion on the launch of a rocket
- analyse a rocket's launch acceleration in terms of forces and conservation of momentum
- present information on one notable figure in the history of rocket development and space exploration.

In chapter 1 we looked at the nature of gravitational fields. In this chapter we discuss the issue of escaping the Earth's gravitational field, at least as far as reaching an orbit around the Earth. This is still well within the reach of the Earth's gravitational field even though orbiting astronauts do not apparently feel its effects. This field is all that holds them in an orbit around the Earth and stops them from heading off further into space.

We will begin by considering simple projectiles, then move from them to projectiles launched into space, and then on to rockets launched into space.

2.1

PROJECTILE MOTION

A **projectile** is any object launched into the air.

A **projectile** is any object that is thrown, dropped or otherwise launched into the air. This includes such things as a box dropped from a plane, a thrown ball, a struck golf ball, a kicked football, or even a fired bullet or cannonball. For our purposes, however, it does not include a rocket, because the thrust of a rocket continues well into its flight. Projectiles are projected into the air and then left to complete their unpowered flight.

Throughout the flight the projectile is subject to just one force — the force of gravity, and just one acceleration — acceleration due to gravity, \mathbf{g} , or 9.8 m s^{-2} downwards near the surface of the Earth. This rate of acceleration applies to all objects, large or small. It is natural to think of heavier objects falling faster than lighter ones, but this is not what happens. All objects are accelerated towards the Earth at the same rate. This was first realised by Galileo Galilei.

Galileo postulated that all masses, whether large or small, fall at the same rate, and he conducted experiments to try to prove just that. However, air resistance gets in the way of such experiments and made the job quite difficult for him. He eventually overcame this difficulty by rolling balls down highly-polished inclines instead of simply dropping them, thereby reducing the effective acceleration. This lower rate of acceleration was less affected by air resistance and was easier to measure.

When astronauts went to the Moon, one of the experiments they performed was to drop a hammer and a feather together. On the Moon there is no air to get in the way, and the rate of acceleration due to gravity is much less than here on Earth, so that things fall more slowly. As figure 2.2 shows, these two factors made the experiment much easier to perform and the result was clear — the feather and the hammer hit the Moon's surface at the same time.

In order to simplify our analysis of a projectile's motion we will ignore the effect of air resistance.

Figure 2.2 An artist's impression of astronaut David Scott dropping a hammer and a feather on the Moon to demonstrate that all objects fall at the same rate



The trajectory

The **trajectory** of a projectile is the path that it follows during its flight.

A **stroboscope** is a light that produces quick flashes at regular (usually small) time periods.

The **trajectory** of a projectile is the path that it follows during its flight. In the absence of air resistance, the path of the flight of a projectile will trace out the shape of a parabola as shown by the photograph in figure 2.3, taken with the aid of a stroboscope. A **stroboscope** is a light that produces quick flashes at regular (usually small) time periods. If used with a camera, instead of a regular flashgun, a stroboscopic photograph is produced which shows multiple images of a moving object.

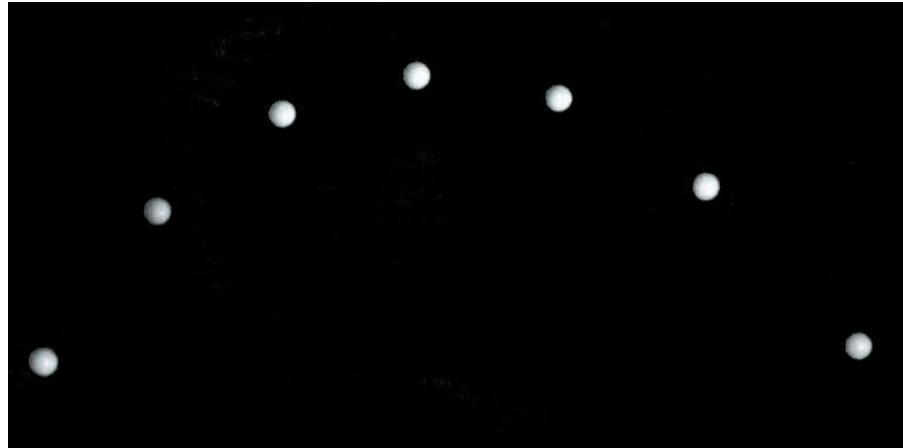


Figure 2.3 A stroboscopic photograph of a ball undergoing projectile motion

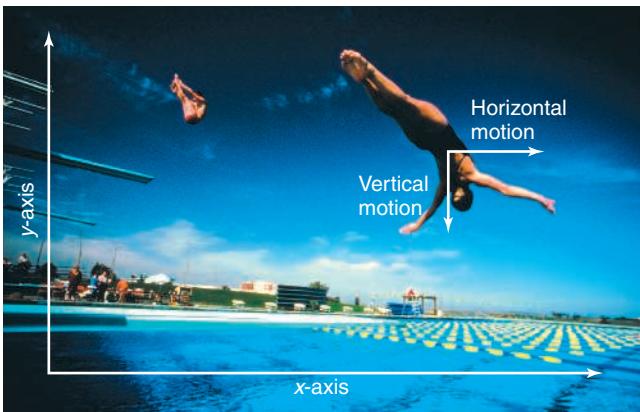


Figure 2.4 A frame of reference for the vertical and horizontal component motions of a projectile

To understand and analyse this motion we must note an observation first made by Galileo: the motion of a projectile can be regarded as two separate and independent motions superimposed upon each other. The first is a vertical motion, which is subject to acceleration due to gravity, and the second is a horizontal motion, which experiences no acceleration. Figure 2.4 places these two motions within a frame of reference, using the y -axis for the vertical motion and the x -axis for the horizontal motion.

Because the two motions are perpendicular, and therefore independent, we can treat them separately and analyse them separately.

Acceleration equations

You will recall from the Preliminary course topic ‘Moving about’ that whenever a moving object changes its velocity, such as a thrown ball, then it has accelerated. This acceleration is defined by the following equation:

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

where

a = acceleration (m s^{-2})

Δv = change in velocity (m s^{-1})

Δt = change in time (s)

v = final velocity (m s^{-1})

u = initial velocity (m s^{-1})

t = time taken (s).

Using this equation, a further set of equations describing accelerated motion can be derived. These equations are shown together here as a set:

$$\begin{aligned} v &= u + at \\ v^2 &= u^2 + 2ar \\ r &= ut + \frac{1}{2}at^2 \end{aligned}$$

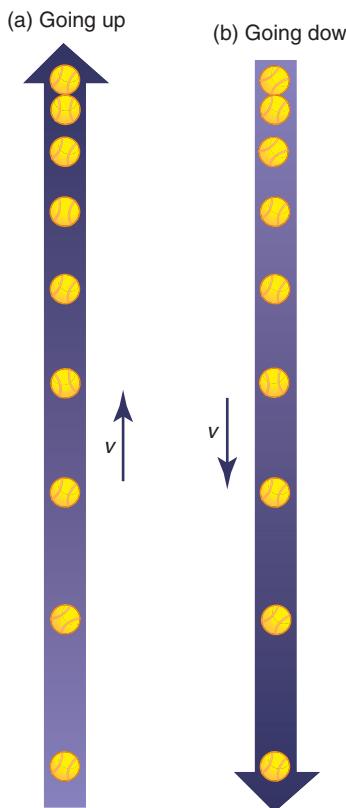


Figure 2.5 (a and b) The motion of a ball thrown vertically upwards

where

r = displacement (m).

We will use this set to derive equations specific to the vertical and horizontal motions.

The vertical motion

When a ball is thrown directly up, it is accelerated due to gravity directly down. As a result it will rise up, slow to a halt in the air and then fall back to Earth. As it falls it will speed up until, when back at its starting point, it is going as fast as it was when thrown. Furthermore, the time taken to fall from its peak height to the ground exactly equals the time taken to rise to the peak height. Figure 2.5 shows this motion broken up into equal time segments. Notice that we have taken *up* to be the positive direction, so that acceleration is always in the negative direction.

In adapting the acceleration equations for the vertical motion we need to note the following variables:

$$a = a_y = 9.8 \text{ m s}^{-2} \text{ down (as shown in figure 2.5(b))}$$

$$v = v_y$$

$$u = u_y$$

$$r = \Delta y \text{ (since displacement = change of position on the } y\text{-axis).}$$

Hence, our three equations become:

$$v_y = u_y + a_y t$$

$$v_y^2 = u_y^2 + 2a_y \Delta y$$

$$\Delta y = u_y t + \frac{1}{2} a_y t^2.$$

SAMPLE PROBLEM

2.1

Calculating the height, time and velocity of a ball thrown vertically

A ball is thrown directly upwards with a velocity of 45 km h^{-1} . Ignoring air resistance, determine:

- (a) its peak height
- (b) its time of flight
- (c) its velocity after 0.5 s
- (d) its velocity after 1.5 s.

SOLUTION

We shall take *up* to be the positive direction. Note that the velocity has not been given in a standard SI unit and must first be converted into m s^{-1} .

$$45 \text{ km h}^{-1} = \left(\frac{45}{3.6} \right) \text{ m s}^{-1} = 12.5 \text{ m s}^{-1}$$

- (a) Let us consider just the first half of the motion; that is, the rise up to the peak. We can say that for this segment:

$$u_y = 12.5 \text{ m s}^{-1}, v_y = 0 \text{ m s}^{-1}, a_y = -9.8 \text{ m s}^{-2}, \Delta y = ?$$

The equation to use has these four variables:

$$\begin{aligned}\therefore v_y^2 &= u_y^2 + 2a_y \Delta y \\ 0^2 &= 12.5^2 + 2 \times (-9.8) \Delta y \\ \therefore \Delta y &= 7.97 \text{ m} \approx 8.0 \text{ m.}\end{aligned}$$

That is, the maximum height reached by the ball is 8.0 m.

- (b) We must still focus on the ball's rise up to its peak height. We can now say that:

$$u_y = 12.5 \text{ m s}^{-1}, v_y = 0 \text{ m s}^{-1}, a_y = -9.8 \text{ m s}^{-2}, \Delta y = 7.97 \text{ m}, t = ?$$

The equation to use is:

$$\begin{aligned}v_y &= u_y + a_y t \\ 0 &= 12.5 + (-9.8)t \\ \therefore t &= 1.28 \text{ s.}\end{aligned}$$

This is the time to rise to the peak height. By symmetry, it will take the ball just as long to fall, so that the total trip time is:

$$2 \times 1.28 = 2.56 \text{ s} \approx 2.6 \text{ s}$$

- (c) We can now consider the entire up and down motion as a whole, and we can list the following data:

$$u_y = 12.5 \text{ m s}^{-1}, a_y = -9.8 \text{ m s}^{-2}, t = 0.5 \text{ s}, v_y = ?$$

The right equation to use has these four variables:

$$\begin{aligned}v_y &= u_y + a_y t \\ &= 12.5 + (-9.8 \times 0.5) \\ &= 7.6 \text{ m s}^{-1}.\end{aligned}$$

That is, the velocity of the ball after 0.5 s is 7.6 m s^{-1} upward.

- (d) We continue to consider the entire motion as a whole, and can list the following data:

$$u_y = 12.5 \text{ m s}^{-1}, a_y = -9.8 \text{ m s}^{-2}, t = 1.5 \text{ s}, v_y = ?$$

The right equation to use has these four variables:

$$\begin{aligned}v_y &= u_y + a_y t \\ &= 12.5 + (-9.8 \times 1.5) \\ &= -2.2 \text{ m s}^{-1}.\end{aligned}$$

That is, after 1.5 s the ball is falling and its velocity is 2.2 m s^{-1} downwards.

The horizontal motion

If a ball is pushed horizontally, ideally, once it is under way, it experiences no acceleration at all in its direction of motion. This is hard to visualise only because we are used to the force of friction, which resists any motion. The effect of this friction can be minimised by considering a flat disc sliding across a horizontal air table, such as an air hockey table.

If no acceleration is experienced, the disc will travel with a uniform, unchanging velocity. If we were to mark the position of the disc at regular time intervals, it would look like figure 2.6.

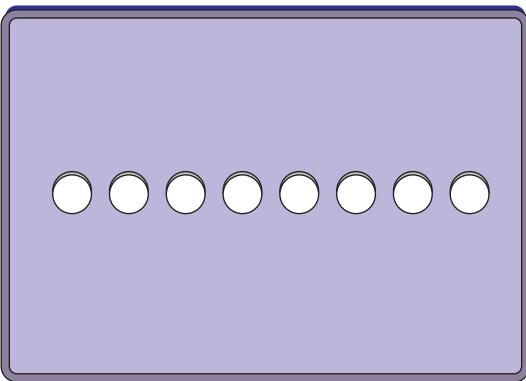


Figure 2.6 The motion of a disc sliding over an air table with uniform velocity

This is also the nature of the horizontal portion of a projectile's motion. Once free of the ground, there is no friction for a thrown ball, other than air resistance which we are currently ignoring, and so the ball will travel sideways above the ground in the same manner as the motion shown in figure 2.6.

In adapting the acceleration equations for the horizontal motion we need to note the following variables:

$$a = 0 \text{ m s}^{-2}$$

$$v = v_x$$

$$u = u_x$$

$r = \Delta x$ (since displacement = change of position on the x -axis).

Hence, our three equations become:

$$v_x = u_x \text{ (that is, horizontal velocity is uniform)}$$

$$v_x^2 = u_x^2$$

$$\Delta x = u_x t.$$

Any object travelling with a velocity will eventually bump into something and bring the motion to an end. The disc on the air table will, sooner or later, collide with the wall of the table, and a thrown ball will eventually collide with the ground. The point at which the ball hits the ground and stops moving defines the maximum horizontal displacement, or range. We can modify the third equation above to show this:

$$\text{Range} = \text{maximum } \Delta x = u_x \times \text{trip time}$$

where

$$u_x = \text{horizontal velocity of a projectile } (\text{m s}^{-1}).$$

SAMPLE PROBLEM

2.2

Calculating the velocity of a flat disc on an air table

A flat disc slides across a 1.5 m wide air table in 0.5 s. What was its velocity?

SOLUTION

$$\Delta x = 1.5 \text{ m}, t = 0.5 \text{ s}, u_x = ?$$

$$\Delta x = u_x t$$

$$1.5 = u_x \times 0.5$$

$$\therefore u_x = 3.0 \text{ m s}^{-1}$$

That is, the disc slid across the air table with a velocity of 3.0 m s^{-1} .

SAMPLE PROBLEM

2.3

Calculating the velocity of a bullet fired at a target

An air gun is fired horizontally at a target 81 m away and the bullet takes just 0.35 s to strike it. What was the velocity of the bullet?

SOLUTION

$$\Delta x = 81 \text{ m}, t = 0.35 \text{ s}, u_x = ?$$

$$\Delta x = u_x t$$

$$81 = u_x \times 0.35$$

$$\therefore u_x = 231 \text{ m s}^{-1} \approx 230 \text{ m s}^{-1}$$

That is, the bullet travelled to the target at a velocity of approximately 230 m s^{-1} .

SAMPLE PROBLEM**2.4****Calculating the range of the bullet**

The gun is now fired into the distance, in a direction that ensures that it won't hit anything during its flight. If it takes 0.5 s to fall to the ground and stop, what was its range?

SOLUTION

$$u_x = 230 \text{ m s}^{-1}, \text{ trip time } t = 0.5 \text{ s, range} = ?$$

$$\begin{aligned}\Delta x &= u_x t \\ &= 230 \times 0.5 \\ &= 115 \text{ m} \approx 120 \text{ m}\end{aligned}$$

That is, the bullet managed to travel a total of approximately 120 m before it stopped because it hit the ground.

Putting the two parts together

We now have a set of equations that describe each of the vertical and horizontal components of the projectile motion. They are summarised in table 2.1.

Table 2.1 The equation set for projectile motion

GENERAL FORM ACCELERATION EQUATION	x-DIRECTION (HORIZONTAL) <i>Note: a = 0</i>	y-DIRECTION (VERTICAL) <i>Note: a = 9.8 m s⁻² down</i>
$v = u + at$	$v_x = u_x$	$v_y = u_y + a_y t$
$v^2 = u^2 + 2ar$	$v_x^2 = u_x^2$	$v_y^2 = u_y^2 + 2a_y \Delta y$
$r = ut + \frac{1}{2} at^2$	$\Delta x = u_x t$	$\Delta y = u_y t + \frac{1}{2} a_y t^2$

**2.1****Modelling projectile motion****eBook plus**

Weblink:
Projectile motion
eModelling:

Freethrow shooter

Use a spreadsheet to predict the conditions necessary to shoot a basketball into a hoop.
doc-0006

Let us now look at how the accelerated vertical motion and the non-accelerated horizontal motion superimpose to give the parabolic trajectory of a projectile. Figure 2.7 shows the vertical motion on the left and the horizontal motion along the base. Each successive image in both motions occurs after the same periods of time. We now regard the images as time-matched pairs and use them as coordinates to plot the combined motion of the projectile. Figure 2.7 shows how this is done.

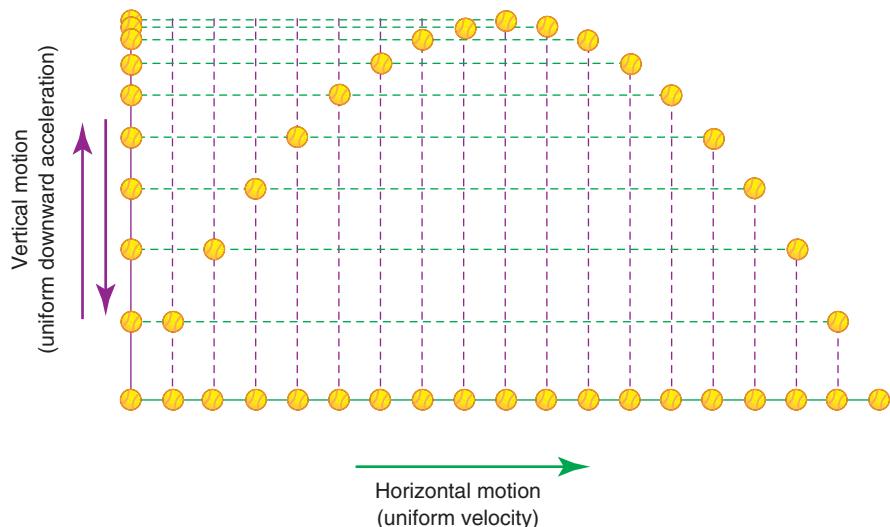


Figure 2.7 Combining the independent vertical and horizontal motions to produce the more complex parabolic trajectory of a projectile

The velocity of the projectile

Projectiles are most commonly sent out at some angle to the horizontal. This initial velocity can quite easily be resolved into vertical and horizontal components using trigonometry, as shown in figure 2.8. Performing this calculation determines the initial velocity in the vertical direction and the initial velocity in the horizontal direction.

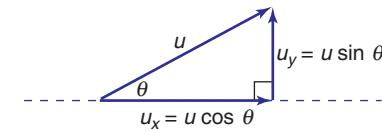


Figure 2.8 The initial velocity at some angle to the horizontal can be resolved into vertical and horizontal components, u_y and u_x .

SAMPLE PROBLEM

2.5a

Calculating the vertical and horizontal components of a projectile

A cannon is fired at a velocity of 400.0 m s^{-1} 30.0° above horizontal. Determine the vertical and horizontal components of this initial velocity.

SOLUTION

Using the method shown in figure 2.8, the following expressions can be deduced:

$$\begin{aligned} \text{The vertical component, } u_y &= 400.0 \sin 30.0^\circ = 200.0 \text{ m s}^{-1} \\ \text{The horizontal component, } u_x &= 400.0 \cos 30.0^\circ = 346.4 \text{ m s}^{-1}. \end{aligned}$$

The velocity of the projectile at other times during the motion can be found by combining the vertical and horizontal velocities together in a vector addition. Figure 2.9 shows how this vector addition is performed at a position where the projectile has almost risen to a peak. Notice that the real velocity of the projectile is directed at a tangent to the trajectory.

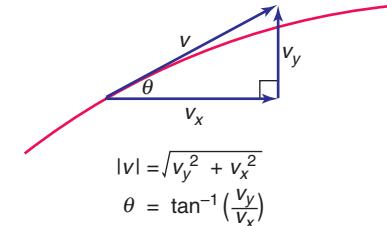


Figure 2.9 Determining the velocity of a projectile at any point in the motion

SAMPLE PROBLEM

2.5b

Calculating the velocity of the cannonball

Determine the velocity of the cannonball from sample problem 2.5a, 30.0 s after firing.

SOLUTION

We must first consider the vertical motion to deduce the vertical velocity after 30.0 s. We shall take upwards to be the positive direction.

$$u_y = 200.0 \text{ m s}^{-1}, a_y = -9.8 \text{ m s}^{-2}, t = 30.0 \text{ s}, v_y = ?$$

$$\begin{aligned} v_y &= u_y + a_y t \\ &= 200.0 + (-9.8 \times 30.0) \\ &= -94.0 \text{ m s}^{-1} \end{aligned}$$

That is, the vertical velocity is 94.0 m s^{-1} downwards.

We already know that the horizontal velocity is constant, so that after 30.0 s

$$v_x = u_x = 346 \text{ m s}^{-1}.$$

Finally, we need to add v_y and v_x together in a vector triangle as shown in figure 2.10.

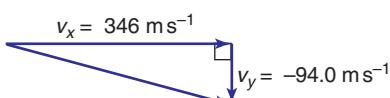


Figure 2.10 The vector addition of v_x and v_y

$$\begin{aligned}
 |v| &= \sqrt{v_y + v_x} \\
 &= \sqrt{(-94.0)^2 + 346^2} = 359 \text{ m s}^{-1} \\
 \theta &= \tan^{-1} \left(\frac{94.0}{346} \right) = 15.2^\circ
 \end{aligned}$$

That is, the velocity of the cannonball 30.0 s after firing is 359 m s^{-1} at 15° below horizontal.

eBookplus

eModelling:
Modelling
a stunt driver
A spreadsheet for a
powerful general model
of projectile motion
doc-0007

In figure 2.11 this calculation has been performed for several points along the trajectory of a fired bullet to show how the velocity varies throughout the motion. You can see how the velocity reduces to a minimum at the peak, because at this point the vertical velocity is zero although the horizontal velocity remains. As the projectile falls from its peak, its velocity increases again until, at the end of the trajectory, it has the same value as the initial velocity and even the same angle to the horizontal, although now it is directed below the horizontal.

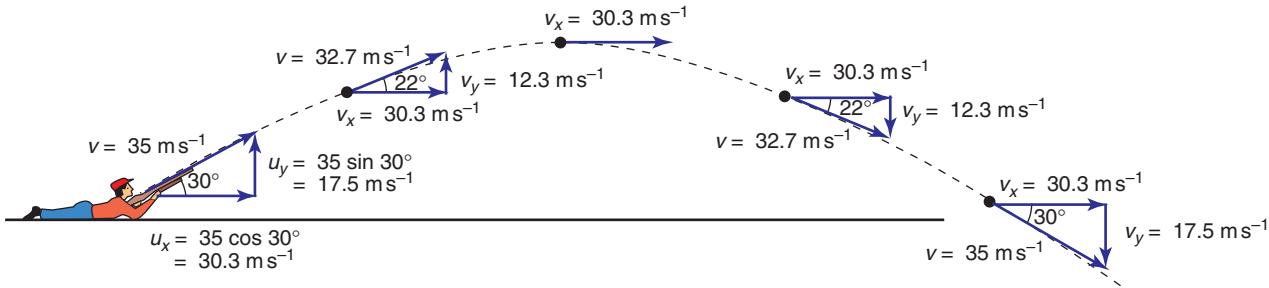


Figure 2.11 Velocity determined at many points along the trajectory of a bullet fired from an air gun

Determining other quantities

We are now in a position to outline a strategy for calculating other dimensions of a projectile's path, such as the height, range and trip time.

To determine maximum height, follow these steps:

1. Resolve initial velocity, u , into component u_y .
2. Consider the vertical motion up to the peak.
3. Note that $v_y = 0$ in this case.
4. Select an acceleration equation to suit the available data.
5. Calculate Δy , which will be maximum height.

SAMPLE PROBLEM

2.6a

SOLUTION

Calculating the maximum height of a tennis ball

A tennis ball is struck at a velocity of 25 m s^{-1} 15° above horizontal. Calculate the maximum height reached by this ball.

First, determine u_y and u_x :

$$\begin{aligned}
 u_y &= 25 \sin 15^\circ = 6.47 \text{ m s}^{-1} \\
 u_x &= 25 \cos 15^\circ = 24.2 \text{ m s}^{-1}.
 \end{aligned}$$

Next, consider the vertical motion up to the peak:

$$u_y = 6.47 \text{ m s}^{-1}, v_y = 0 \text{ m s}^{-1}, a_y = -9.8 \text{ m s}^{-2}, \Delta y = ?$$

$$\begin{aligned}
 v_y^2 &= u_y^2 + 2a_y \Delta y \\
 0^2 &= 6.47^2 + 2(-9.8)\Delta y \\
 \Delta y &= 2.1 \text{ m}.
 \end{aligned}$$

That is, the maximum height reached by the tennis ball is 2.1 m above the ground.

SAMPLE PROBLEM**2.6b**

2

To determine trip time, follow these steps:

1. Resolve initial velocity, u , into component u_y .
2. Consider the vertical motion up to the peak.
3. Note that $v_y = 0$ in this case.
4. Select an acceleration equation to suit the available data.
5. Calculate t , time to rise to the peak.
6. Double this time to find the trip time, since it takes just as long to fall as to rise.

Calculating the time for the struck tennis ball to return to the ground

Referring back to the tennis ball in sample problem 2.6a, determine the time it takes to return to the ground.

SOLUTION

Once again, consider the vertical motion up to the peak:

$$u_y = 6.47 \text{ m s}^{-1}, v_y = 0 \text{ m s}^{-1}, a_y = -9.8 \text{ m s}^{-2}, t = ?$$

$$\begin{aligned} v_y &= u_y + a_y t \\ 0 &= 6.47 + (-9.8)t \\ \therefore t &= 0.66 \text{ s} \end{aligned}$$

and hence,

$$\text{trip time} = 2t = 2 \times 0.66 = 1.32 \text{ s.}$$

That is, the time taken for the tennis ball to complete its flight and strike the ground is 1.32 s.

To determine the range, follow these steps:

1. Resolve initial velocity, u , into components u_y and u_x .
2. Analyse the vertical motion to find the trip time as shown above.
3. Now consider the horizontal motion and calculate the range using $\Delta x = u_x t$.

SAMPLE PROBLEM**2.6c**

Calculating the range of the tennis ball's trajectory

Referring once again back to the tennis ball in sample problem 2.6a, calculate the range of its trajectory.

SOLUTION

This time, consider just the horizontal portion of its motion:

$$u_x = 24.2 \text{ m s}^{-1}, \text{ trip time } t = 1.3 \text{ s, range} = ?$$

$$\begin{aligned} \Delta x &= u_x t \\ &= 24.2 \times 1.3 \\ &= 31.5 \text{ m} \approx 32 \text{ m.} \end{aligned}$$

That is, the tennis ball covered 32 m before striking the ground.

Note: Tennis players are able to strike a ball at this sort of velocity and higher, and still keep it within the court, because they impart spin to the ball, which can dramatically alter the trajectory. This course does not go into the effects of spin.

Air resistance

In all of our work on projectile motion we have ignored the effect of air resistance on the motion of the projectile. The reason for this is that it is simply too difficult for us to account for, since it depends on many factors such as the shape, surface area and texture of the projectile, as well as its

velocity through the air. In the real world, air resistance acts as a retarding force in both the vertical and horizontal directions. As a result, the path of the projectile is distorted away from a perfect parabola to the shape shown in figure 2.12.

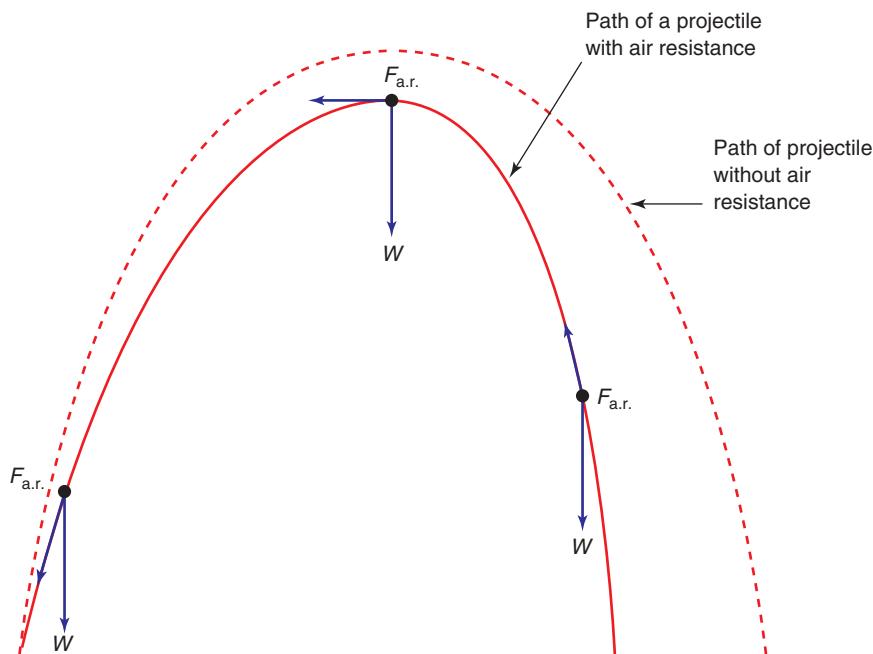


Figure 2.12 Air resistance opposes the velocity of a projectile at any given moment and distorts the trajectory away from a parabolic shape.

2.2 ESCAPE VELOCITY

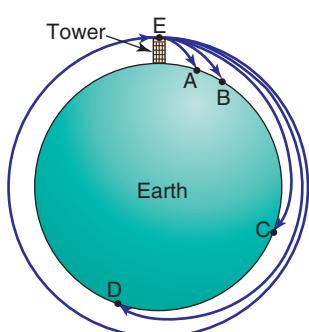


Figure 2.13 Newton's suggestion for achieving an orbit

Isaac Newton wrote that it should be possible to launch a projectile fast enough so that it achieves an orbit around the Earth. As shown in figure 2.13, his reason was that a stone thrown from a tall tower will cover a considerable range before striking the ground. If it is thrown faster, it will travel further before stopping. If thrown faster still, it will have an even greater range. If thrown fast enough then, as the stone falls, the Earth's surface curves away, so that the falling stone never actually lands on the ground and orbits the Earth. It was only a thought experiment, of course. He had no way of testing this idea but it does hit upon one important fact — that for any given altitude, there is a specific velocity required for any object to achieve a stable circular orbit.

If this specific velocity is exceeded slightly, then the object will follow an elliptical orbit around the Earth. If the specific velocity is exceeded further still, then the object will follow a parabolic or hyperbolic path away from the Earth. This is the manner in which space probes depart the Earth and head off into space.

We will now consider a situation similar to Newton's. Imagine throwing a stone directly up. When thrown, the stone will rise to a certain height before falling back to Earth. If thrown faster, it will rise higher. If thrown fast enough, it should rise up and continue to rise, slowing down but never falling back to Earth, and finally coming to rest only when it has completely escaped the Earth's gravitational field. The initial velocity required to achieve this is known as **escape velocity**.

Escape velocity is the initial velocity required by a projectile to rise vertically and just escape the gravitational field of a planet.

By considering the kinetic and gravitational potential energy of a projectile, it can be shown mathematically that the escape velocity of a planet depends only upon the universal gravitation constant, the mass and the radius of the planet. Their relationship is shown in the equation that follows.

$$\text{Escape velocity} = \sqrt{\frac{2Gm_{\text{planet}}}{r_{\text{planet}}}}$$

where

$$\begin{aligned} G &= \text{universal gravitational constant} \\ &= 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2} \\ m_{\text{planet}} &= \text{mass of the planet (kg)} \\ r_{\text{planet}} &= \text{radius of the planet (m).} \end{aligned}$$

We are now in a position to calculate the escape velocity for Earth:

$$\begin{aligned} \text{Escape velocity} &= \sqrt{\frac{2Gm_{\text{Earth}}}{r_{\text{Earth}}}} \\ &= \sqrt{\frac{2(6.67 \times 10^{-11})(5.97 \times 10^{24})}{6.38 \times 10^6}} \\ &= 11200 \text{ m s}^{-1} \approx 40000 \text{ km h}^{-1}. \end{aligned}$$

That is, the escape velocity on Earth is about 40000 km h^{-1} . This is a considerable velocity, but remember that this is the velocity with which a projectile must be launched directly up in order to completely escape the Earth's gravitational field. It does not apply to a rocket, which continues its thrust well after launch.

SAMPLE PROBLEM

2.7

Calculating escape velocity

Determine the escape velocity of the planet Venus, given that its mass is $4.87 \times 10^{24} \text{ kg}$ and its radius is 6052 km.

SOLUTION

$$\begin{aligned} \text{Escape velocity} &= \sqrt{\frac{2Gm_{\text{Venus}}}{r_{\text{Venus}}}} \\ &= \sqrt{\frac{2(6.67 \times 10^{-11})(4.87 \times 10^{24})}{6.052 \times 10^6}} \\ &= 10360 \text{ m s}^{-1} \approx 37300 \text{ km h}^{-1} \end{aligned}$$

That is, escape velocity on the planet Venus is approximately 37300 km h^{-1} .

2.3

LIFT-OFF

Let us now turn our attention to powered projectiles, that is, rockets. Whereas projectiles receive an initial velocity and are then left to fall through a trajectory, rockets receive a force called **thrust** from their engine(s) for a significant portion of their upwards flight, and become more conventional projectiles only after their engines are exhausted.

Thrust is the force delivered to a rocket by its engines.

Rockets

A rocket engine is different from most other engines in that it carries with it both its fuel and oxygen supply. Any fuel needs oxygen to burn and most engines, such as jet engines or internal combustion engines,

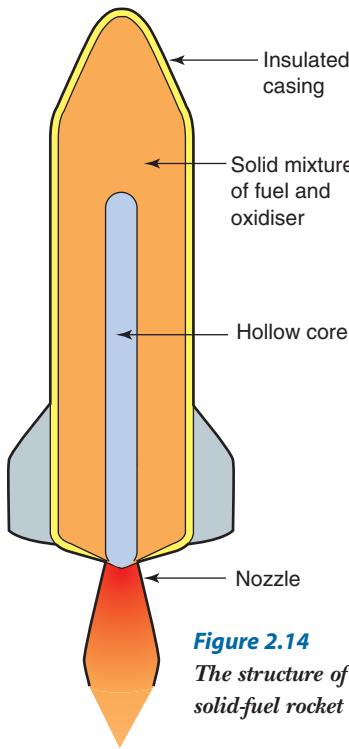


Figure 2.14
The structure of a solid-fuel rocket

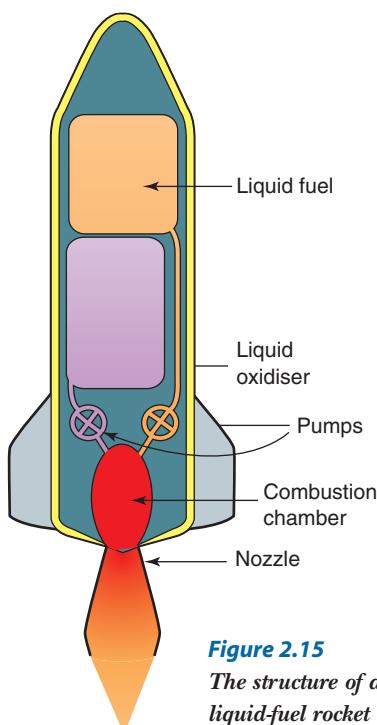


Figure 2.15
The structure of a liquid-fuel rocket

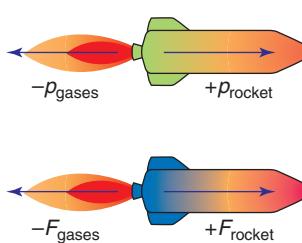


Figure 2.16 Momentum and force acting on a rocket

obtain the necessary oxygen from the air around them. However, in space there is no air or other atmosphere, which makes a rocket engine the natural choice.

Modern rockets can use either solid or liquid propellants. Solid rocket propellant is a manufactured mixture of a fuel, such as a mixture of hydrogen compounds and carbon, with an oxidiser, or oxygen supply, being a mixture of oxygen compounds. The dry, solid propellant is packed into an insulated cylindrical vessel, usually with a hollow core through its middle. The hollow core is not necessary, but it increases the surface area available for burning, and therefore the thrust. The end of the cylinder is fitted with a nozzle. Finally, an igniter built into the cylinder sparks off the rapid burning of the propellant. Hot gases are produced at an extreme rate and are forced out through the nozzle.

Liquid-propellant rockets keep both the liquid fuel, such as kerosene or liquid hydrogen, and the oxidiser, usually liquid oxygen, in separate storage tanks. Pumps force each liquid from their tanks and spray them into a combustion chamber where they mix as they burn, producing the hot gases that are expelled out through a nozzle (see figure 2.15).

The forward motion of the rocket can be understood by recalling the Law of Conservation of Momentum. This law states that during any interaction in a closed system the total momentum of the system remains unchanged. Stated another way, this means that during a launch, the momentum of the gases shooting out of the rear of the rocket must be equal to the forward momentum of the rocket itself, as shown in figure 2.16. This means that during any one-second time interval:

$$\text{Total change in momentum} = 0$$

$$\therefore -\Delta p_{\text{gases}} = \Delta p_{\text{rocket}}$$

$$-\Delta(mv)_{\text{gases}} = \Delta(mv)_{\text{rocket}}$$

where

$$\Delta p = \text{change in momentum } (\text{kg m s}^{-1})$$

$$m = \text{mass } (\text{kg})$$

$$v = \text{velocity } (\text{m s}^{-1}).$$

This means that the backward momentum of the gases ($-\Delta p$) is exactly equal in magnitude to the forward momentum of the rocket ($+\Delta p$), endowing the rocket with forward velocity. It is important to note that, while the mass of the gases during any given second is less than the mass of the rocket, their velocity is much greater, so that their momenta are equal but opposite. You should also recall that:

$$\Delta p = \text{impulse} = Ft$$

where

$$F = \text{force } (\text{N})$$

$$t = \text{time } (\text{s})$$

so that

$$-(Ft)_{\text{gases}} = (Ft)_{\text{rocket}}$$

or, for any one second interval,

$$-F_{\text{gases}} = F_{\text{rocket}}.$$

This is Newton's Third Law of Motion. This law says that for every force there is an equal but opposite force, and this is also the case here. The rocket is forcing a large volume of gases backward behind it, and the gases, in turn, force the rocket forward as shown in figure 2.16. Although the two forces are equal and opposite, the rocket experiences just one of them — the forward push that we call thrust.

PHYSICS IN FOCUS

The engines of the space shuttle

The thrust of a solid-fuel rocket engine cannot be varied once started, since the fuel is ignited and burns at the maximum possible rate until it is exhausted. The thrust of a liquid-fuel engine can be throttled to some extent, by varying the amount of fuel and oxidiser that enter the combustion chamber, allowing some control over the thrust delivered by the engine and the resultant rate of acceleration.

In figure 2.1 you can clearly see the solid rocket boosters either side of the space shuttle. After about two minutes firing, they are exhausted and

then separate from the shuttle — they fall into the ocean and are recovered for reuse. The large brown tank in the middle contains the external liquid-fuel tanks for the shuttle's three engines on the upward journey. Once in orbit, this tank is released — it eventually falls into the ocean but is not recovered. The liquid-fuel engines can be throttled and this gives the shuttle the ability to vary the launch thrust between 50% and 100% — an important feature that minimises the forces experienced by the crew.

Thrust and acceleration

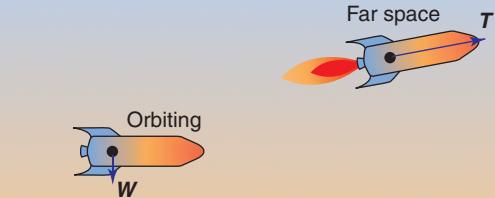
A rocket at various points in its lift-off and flight is shown in figure 2.17. As it is a mass subject to several forces, it will accelerate according to Newton's Second Law:

$$\Sigma F = ma$$

$$\therefore a = \frac{\Sigma F}{m}.$$

Figure 2.17 The forces and acceleration a rocket is subjected to during a launch. Also shown is the g force experienced by the astronauts within.

W = Weight
R = Reaction
T = Thrust



$R =$	W	$\frac{W}{2}$	0	0	0	0	0
$T =$	0	$\frac{W}{2}$	W	$1.5W$	$3W$	0	>0
$a = \frac{\Sigma F}{m}$	0	0	0	$0.5g$	$2g$	$-g$	$\frac{T}{m}$
$g\text{ force} = \frac{g + a}{9.8}$	1	1	1	1.5	3	0	$\frac{T}{9.8m}$

As shown in figure 2.17, the rocket is subject to the following forces:

- its weight force directed downward
- its thrust (the force delivered by the engines) directed upward
- the reaction force of the ground on the rocket (equal to the difference between the weight and the thrust while the rocket is on the ground) directed upward
- air resistance directed downward against the motion of the rocket once it has left the ground. This air resistance force can become significant as the speed of the rocket builds, but at the relatively low speeds of early lift-off its effect can be ignored.

SAMPLE PROBLEM

2.8

Calculating a model rocket's lift-off acceleration

A model rocket has a mass of 100.0 g and is able to produce a thrust of 4.50 N. Determine its initial rate of acceleration upon lift-off.

SOLUTION

$$\begin{aligned}a &= \frac{\sum F}{m} = \frac{(T - mg)}{m} \\&= \frac{(4.50 - 0.100 \times 9.8)}{0.100} \\&= 35 \text{ m s}^{-2}\end{aligned}$$

That is, the rocket's initial rate of acceleration will be 35 m s^{-2} .

A rocket's acceleration will not be constant, however, because fuel constitutes up to 90% of the mass of a typical rocket. As the fuel is burnt, the mass of the rocket decreases although the thrust remains essentially constant. Additionally, the gravitational field vector, \mathbf{g} , reduces slightly with increasing altitude, as seen in chapter 1. The result is that a rocket's rate of acceleration will increase as its flight progresses, and its velocity will increase logarithmically.

Consequently, the acceleration equation above can apply only at an instant in time, provided the mass and thrust are known at that instant. More detailed rocket equations exist for the enthusiast which allow the calculation of a rocket's velocity, maximum height and required fuel load; however, this theory is beyond the scope of the HSC physics course.

***g* forces**

Your body is a mass lying somewhere within a gravitational field, and therefore experiences a true weight, $W = mg$. The sensation of weight that you feel, however, derives from your apparent weight, which is equal to the sum of the contact forces resisting your true weight. This includes the normal reaction force of the floor on your body, or the thrust of a rocket engine.

The term 'g force' is used to express a person's apparent weight as a multiple of his/her normal true weight (that is, weight when standing on the surface of the Earth).

Hence,

$$\text{g force} = \frac{\text{apparent weight}}{\text{normal true weight}}.$$

Figure 2.18 shows the forces acting upon an astronaut during a launch. The astronaut's body is exerting a downward weight force on the floor, and the floor meets this with an upward reaction force equal to $m \times g$. In addition, the floor is exerting an upward accelerating force equal to $m \times a$.

The astronaut feels an apparent weight = $mg + ma$.

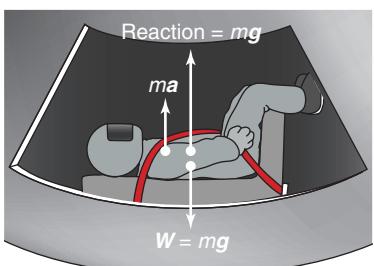


Figure 2.18 The forces acting on an astronaut during a launch

Therefore g force = $\frac{\text{apparent weight}}{\text{normal true weight}}$

$$= \frac{mg + ma}{9.8m}$$

and hence, g force = $\frac{g + a}{9.8}$

where

g = acceleration due to gravity at altitude (m s^{-2})
 m = mass of astronaut (kg).

Note that g force is closely related to acceleration.

It is common to experience variations in g force when riding up or down in an elevator, so let us compare this situation with that for an astronaut during launch. This is shown in figure 2.19. When the rocket

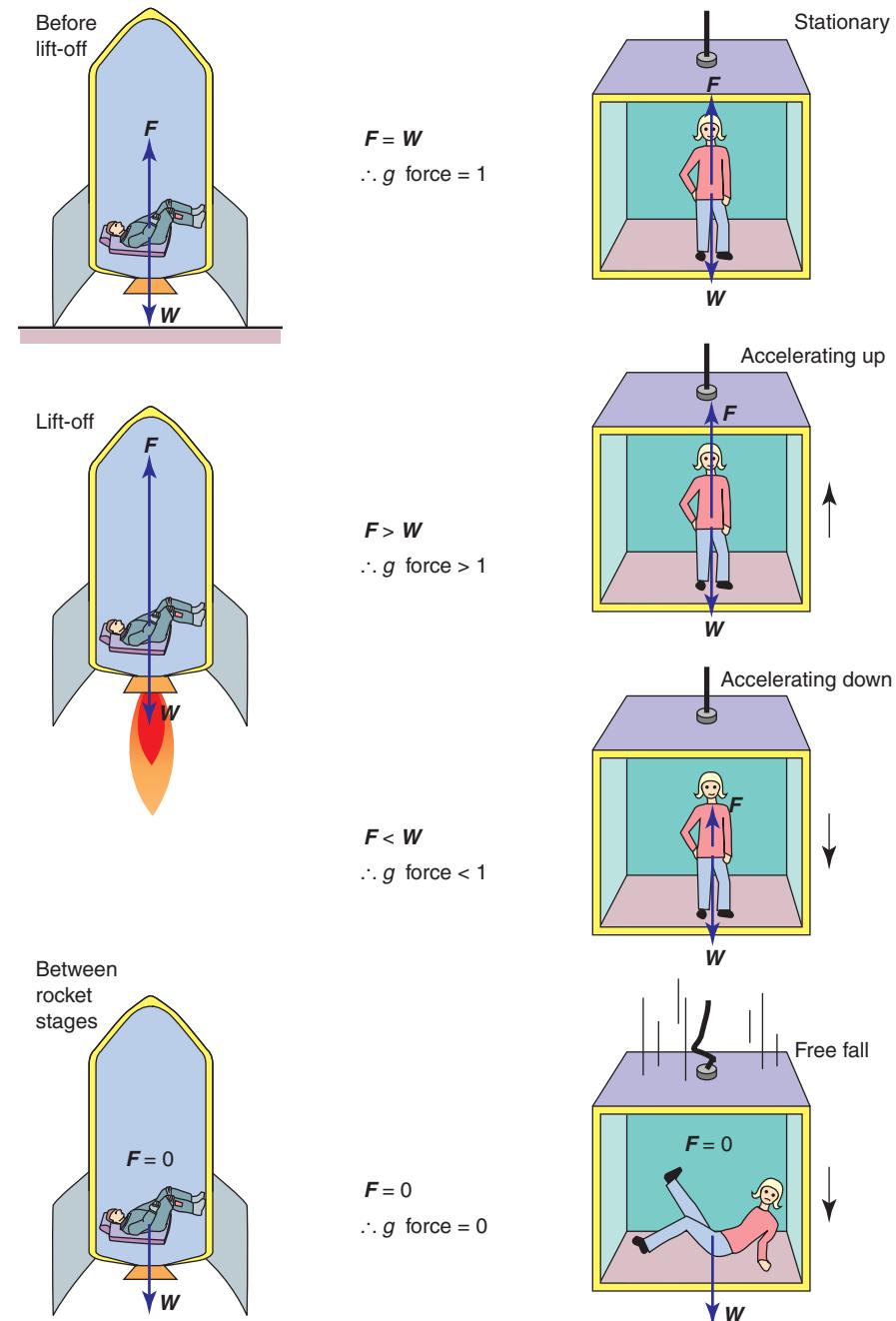


Figure 2.19 An occupant of an elevator experiences types of forces that are similar to those experienced by an astronaut during a launch.

and elevator are both stationary, the only forces acting are the weight and reaction force, which are equal in magnitude but opposite in direction. In this case, the apparent weight equals the true weight and the occupant experiences a g force of one (that is, a *one g load*).

When the elevator begins to accelerate upwards, it is analogous to the rocket lifting off. The floor will exert an upwards force on the occupant of $(mg + ma)$ so that the occupant experiences a g force of $\frac{(g+a)}{9.8}$, which is a value greater than one.

When the elevator accelerates downwards, the floor exerts an upward force less than the occupant's weight, so that the g force experienced is $\frac{(g-a)}{9.8}$, which is less than one. If the elevator were in free fall, the downward rate of acceleration would be g , so that the g force would have a value of zero. In other words, the floor would exert no force on the occupant, and the occupant would experience a zero apparent weight, that is, weightlessness within the accelerating frame of reference of the elevator.

This situation is analogous to a multi-stage rocket after it has jettisoned a spent stage but before it has ignited the next. During those few seconds there is only the downward acceleration due to gravity, so that the astronauts experience a zero g load (weightlessness).

SAMPLE PROBLEM

2.9

Calculating the g force on a model rocket

The model rocket has a pre-launch mass of 94.2 g, of which 6.24 g is solid propellant. It is able to deliver a thrust of 4.15 N for a period of 1.2 s. Assuming that the rocket is fired directly up, determine:

- the initial rate of acceleration and g force
- the final rate of acceleration and g force just prior to exhaustion of the fuel.

SOLUTION

We shall assume up to be the positive direction, and that $g = 9.8 \text{ m s}^{-2}$ at the relatively low height achieved by this rocket.

- Determine initial acceleration as follows:

$$\begin{aligned} a &= \frac{\sum F}{m} = \frac{(T - mg)}{m} \\ &= \frac{4.15 - (0.0942 \times 9.8)}{0.0942} \\ &= 34 \text{ m s}^{-2}. \end{aligned}$$

Also, g force can be determined as follows:

$$\begin{aligned} g \text{ force} &= \frac{g+a}{9.8} \\ &= \frac{9.8 + 34.3}{9.8} \\ &= 4.5. \end{aligned}$$

- Determine final acceleration as follows:

$$\text{Final mass} = 94.2 - 6.24 = 88 \text{ g}$$

$$\begin{aligned} \text{Hence, } a &= \frac{\sum F}{m} = \frac{(T - mg)}{m} \\ &= \frac{4.15 - (0.0880 \times 9.8)}{0.0880} \\ &= 37 \text{ m s}^{-2}. \end{aligned}$$

Also, final g force can be determined as follows:

$$\begin{aligned} g \text{ force} &= \frac{g+a}{9.8} \\ &= \frac{9.8+37.4}{9.8} \\ &= 4.8. \end{aligned}$$

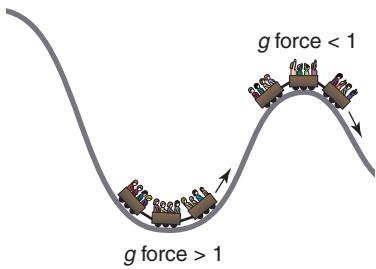


Figure 2.20 Riders on a roller-coaster experience changing g forces.

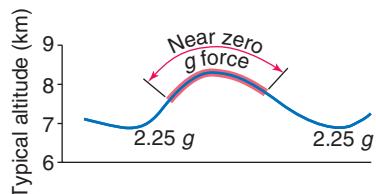


Figure 2.21 Aircraft trajectory to simulate near-weightlessness

As shown in figure 2.20, riders on a roller-coaster also experience variations in g force. When a roller-coaster zooms down through a dip in its track and turns upward, the riders experience an upward acceleration and, hence, a g force greater than one. However, when rolling over the top of a crest in the track and accelerating downhill, the riders will experience a downward acceleration and a g force less than 1.

This idea has been used to provide training for astronauts with a simulated weightless environment. The subjects sit within an aircraft, which flies a trajectory very similar to that of a roller-coaster ride, as shown in figure 2.21. At first the plane flies down in a shallow dive before turning hard up into a 50° climb. This turn will create a g load of approximately 2.25, but soon after the pilot pushes the nose over and throttles the engines down, turning the aircraft into a free-falling projectile following a parabolic path. During this phase the subjects within experience about half a minute of near-weightlessness before the pilot needs to throttle up the engines again to recover the dive and repeat the process.

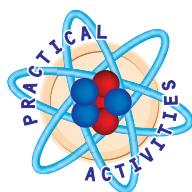
Variations in acceleration and g forces during a typical launch

As shown in figure 2.17, prior to lift-off a rocket has zero acceleration because of the balance that exists between the weight force and the reaction force plus thrust. The astronaut within is experiencing a one g load. This initial condition will not change until the building thrust exceeds the weight of the rocket, at which point the rocket will lift off.

Since the thrust now exceeds the weight, there is a net force upwards on the rocket, which begins to accelerate upwards. The g force experienced by the rocket will have a value slightly greater than one. From this point onwards, the mass of the rocket begins to decrease as fuel is consumed and, hence, the rate of acceleration and subsequent g force steadily climbs, reaching maximum values just before the rocket has exhausted its fuel.

At this point a single-stage rocket becomes a projectile, eventually falling to Earth. A multi-stage rocket, however, drops the spent stage away, momentarily experiencing zero g conditions as it coasts. The second-stage rocket fires and quickly develops the necessary thrust to exceed the effective weight at its altitude, and then starts to accelerate again. The g force experienced by the rocket and astronaut begins again at a value marginally greater than one and gradually builds to its maximum value just as the second-stage fuel supply is exhausted. If there is a third stage, the process is repeated.

The variation in g forces varied during the launch of *Saturn V*, a large three-stage rocket used to launch the *Apollo* spacecraft, is shown in figure 2.22. Note that the jagged peaks on the graph are due to the sequential shutdown of the multiple rocket engines of each stage — a technique designed specifically to avoid extreme g forces.



2.2

Acceleration and load during the Apollo 10 launch

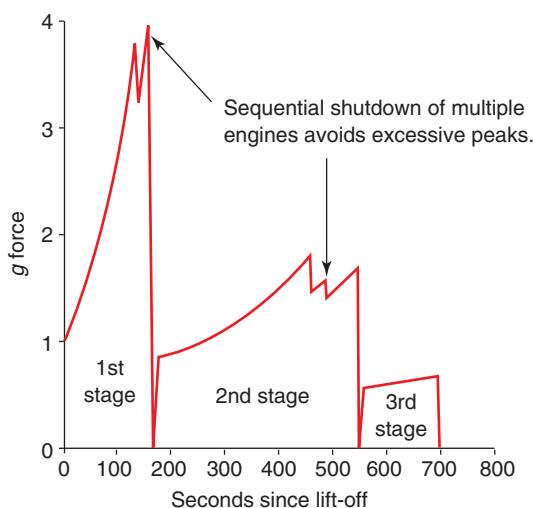


Figure 2.22 Variations in g forces during an Apollo–Saturn V launch



Figure 2.23 A rocket heading into orbit is launched to the east to receive a velocity boost from the Earth's rotational motion.

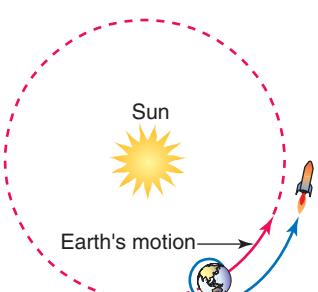


Figure 2.24 The flight of a rocket heading into space is timed so that it can head out in the direction of the Earth's motion and thereby receive an extra boost.

This figure shows that *Apollo* astronauts experienced a peak load of four g during lift-off. This is a significant force — at four g a person begins to lose their colour vision and peripheral vision. Soon after this the person will black out; although each individual has their own threshold level. In the first manned US space flight, astronaut Alan Shepard had to tolerate a peak g force of 6.3 g during launch. Rocket design has improved since then — space shuttle astronauts never experience loads greater than three g due to the shuttle's ability to throttle back its liquid fuel engines. This is discussed in more detail in chapter 3.

The effect of the Earth's motion on a launch

Why do cricket fast bowlers run up to the wicket before bowling the ball? The answer is that the velocity at which the ball is bowled is greater than it would have been if the bowler had not run up. This is because the velocity of the ball relative to the ground is equal to the velocity of the ball relative to the bowler, plus the velocity of the bowler relative to the ground. Algebraically, this is expressed as:

$$\text{ball } v_{\text{ground}} = \text{ball } v_{\text{bowler}} + \text{bowler } v_{\text{ground}}$$

In other words, a moving platform (the bowler) offers a boost to the velocity of a projectile (the ball) launched from it, if launched in the direction of motion of the platform.

The same principle applies to a rocket launched from the Earth. Consider that the Earth is revolving around the Sun at approximately 107 000 km h⁻¹ relative to the Sun. In addition, the Earth rotates once on its axis per day so that a point on the Equator has a rotational velocity of approximately 1700 km h⁻¹ relative to the Sun. Hence, the Earth is itself a moving platform with two different motions which can be exploited in a rocket launch to gain a boost in velocity.

Engineers planning to launch a rocket into orbit can exploit the Earth's rotation in order to achieve the velocity needed for a stable orbit. This is done by launching in the direction of the Earth's rotation; that is, by launching toward the east, as shown in figure 2.23. In this way, the rotational velocity of the launch site relative to the Sun will add to the orbital velocity of the rocket relative to the Earth, to produce a higher orbital velocity achieved by the rocket relative to the Sun.

In a similar way, engineers planning a rocket mission heading further into space can exploit the Earth's revolution around the Sun by planning the launch for a time of year when the direction of the Earth's orbital velocity corresponds to the desired heading. Only then is the rocket launched up into orbit. The rocket is allowed to proceed around its orbit until the direction of its orbital velocity corresponds with the Earth's, and then its engines are fired to push it out of orbit and further into space, as shown in figure 2.24. In this way the Earth's orbital velocity relative to the Sun adds to the rocket's orbital velocity relative to the Earth, to produce a higher velocity achieved by the rocket relative to the Sun.

Planning of this sort clearly favours certain times of the year over others, or even certain times of day depending upon the flight mission. These favourable periods are referred to as *launch windows*.

PHYSICS IN FOCUS

Space exploration and rocket science pioneers

The Chinese discovered gunpowder and used it to create fireworks. By the eleventh century, the Chinese were using simple rockets called *fire arrows* as weapons. In the late 1700s a British artillery officer, William Congreve, developed simple rockets for use by the British army. The Hale rockets followed 50 years later. Despite all of this, modern rocket science didn't begin in earnest until the late 1800s and early 1900s. Listed here are some of the most notable pioneers.

Konstantin Tsiolkovsky (1857–1935) was a Russian mathematics teacher who took an interest in rocketry, being inspired by Jules Verne's book *From the Earth to the Moon*. Working entirely on his own, he developed precise calculations for space flight and the details of many aspects of rocket design and space exploration. His work was purely theoretical as he performed no experiments, but his published work influenced rocket development around the world, especially in Russia. His ideas were wide ranging — from the very pragmatic, such as the design of a liquid-fuel rocket engine featuring throttling capability and multi-staging, to the (then) fanciful, such as space stations and artificial gravity, terraforming of other planets and extraterrestrial life. He was the inspiration for men such as Sergei Korolev (1906–1966) who was the Russian Chief Constructor responsible for *Sputnik I* and the Vostok rocket. *Sputnik I* was the world's first artificial satellite, while the Vostok rocket was used to send Yuri Gagarin into a single orbit of the Earth on 12 April 1961. This was the first time that a person had entered space.

Robert H. Goddard (1882–1945) was an American college professor of physics with a passion for rocketry. Also inspired by Jules Verne as a boy, Goddard decided early to dedicate his life to rocketry. Unlike Tsiolkovsky, Goddard was an engineer and an experimentalist. He conceived ideas then tested them, patenting those that were successful. He built and tested the world's first liquid-fuel rocket, which solved many technical problems such as fuel valving for throttle, start and stop, fuel injection, engine cooling and ignition. He was the first to use

gyroscopes and vanes for guidance, and to separate the payload from the rocket in flight and return it to Earth.

Herman Oberth (1894–1992) was born in Romania but lived in Germany. Purely a theorist, he was yet another inspired by Jules Verne. He wrote a doctoral thesis titled *By Rocketry to Space*. Although the University of Heidelberg rejected the thesis, he had the work published privately as a book. It promptly sold out. The subject of rocketry captured the public's imagination and Oberth was himself inspiring a new generation of rocket scientists. He was an early member of the VfR, or Society for Space Travel, and published another book titled *The Road to Space Travel*. This work won an award and Oberth used the prize money to purchase rocket motors for the VfR, assisting its development efforts. One of those inspired by Oberth was Wernher von Braun (1912–1977) who became the rocket engineer responsible for the development of the V2 rocket, which was used to bomb London during World War II, and later the Mercury-Redstone rocket which put the first Americans into space.

Roberts Esnault-Pelterie (1881–1957) was a French rocket pioneer. He published two important books — *Astronautics* in 1930 and *Astronautics Complement* in 1934. He suggested the idea that rockets be used as long-range ballistic missiles, and the French Army employed him to develop these rockets. He experimented with various liquid fuels in rocket motors of his design, starting with liquid oxygen and gasoline, then nitrogen peroxide and benzene, before attempting liquid oxygen and tetranitromethane. This last combination caused him a major hand injury.

Theodore von Karman (1881–1963) was born in Hungary but later settled in America. In the 1930s he became a professor of aeronautics at Caltech. There he established the 'Jet Propulsion Laboratory' dedicated to rocket work. The JPL still exists today, working closely with NASA and specialising in exploration of the solar system by space probe.

SUMMARY

- A projectile is any object that is launched into the air.
- The path of a projectile, called its trajectory, has a parabolic shape if air resistance is ignored. The trajectory can be analysed mathematically by regarding the vertical and horizontal components of the motion separately.
- The vertical motion of a projectile is uniformly accelerated motion and can be analysed using these equations:

$$\begin{aligned}v_y &= u_y + a_y t \\ \Delta y &= u_y t + \frac{1}{2} a_y t^2 \\ v_y^2 &= u_y^2 + 2a_y \Delta y.\end{aligned}$$

- The horizontal motion of a projectile is constant velocity and can be analysed using these equations:

$$\begin{aligned}v_x &= u_x \\ v_x^2 &= u_x^2 \\ \Delta x &= u_x t.\end{aligned}$$

- Escape velocity is the vertical velocity that a projectile would need to just escape the gravitational field of a planet. It is given by the equation:

$$\text{Escape velocity} = \sqrt{\frac{2Gm_{\text{planet}}}{r_{\text{planet}}}}.$$

- A rocket is different from a projectile because it continues to be propelled after it is launched, accelerating throughout most of its upward journey. Rockets differ from other engines such as jets because they carry with them the oxygen required to burn their fuel.
- The forward progress of a rocket can be explained using Newton's Third Law (equal and opposite forces) as well as by the conservation of momentum (total change in momentum equals zero).
- The acceleration of a rocket can be determined using Newton's Second Law: $\Sigma F = ma$.
- The term 'g force' is used to describe apparent weight as a multiple of normal weight, and is closely related to acceleration.
- During a rocket launch the g force experienced by an astronaut increases because the mass of fuel is reducing, even though the thrust remains essentially the same.
- The revolution and rotation of the Earth can be used to provide a launched rocket with an additional boost, allowing it to save fuel in achieving its target velocity.

QUESTIONS

- Explain why it is that the vertical and horizontal components of a projectile's motion are independent of each other. Identify any common variables.
- Describe the trajectory of a projectile.
- List any assumptions we are making in our treatment of projectile motion.
- Describe Galileo's contribution to our knowledge of projectile motion.
- What is the mathematical significance of vertical and horizontal motions being perpendicular?
- Describe the strategy you can employ to determine a projectile's:
 - (a) velocity
 - (c) trip time
 - (b) maximum height
 - (d) range.
- Describe the effect of air resistance on the trajectory of a projectile.
- A volleyball player sets the ball for a team mate. In doing so she taps the ball up at 5.0 m s^{-1} at an angle of 80.0° above the horizontal. If her fingers tapped the ball at a height of 1.9 m above the floor, calculate the maximum height to which the ball rises?
- An 'extreme' cyclist wants to perform a stunt in which he rides up a ramp, launching himself into the air, then flies through a hoop and lands on another ramp. The angle of each ramp is 30.0° and the cyclist is able to reach the launch height of 1.50 m with a launching speed of 30.0 km h^{-1} . Calculate:
 - (a) the maximum height above the ground that the lower edge of the hoop could be placed
 - (b) how far away the landing ramp should be placed.
- A football is kicked with a velocity of 35.0 m s^{-1} at an angle of 60.0° . Calculate:
 - (a) the 'hang time' of the ball (time in the air)
 - (b) the length of the kick.
- A basketball player stands 2.50 m from the ring. He faces the backboard, jumps up so that his hands are level with the ring and launches the ball at 5.00 m s^{-1} at an angle of 50.0° above the horizontal. Calculate whether he will score.
- A cannon's maximum range is achieved with a firing angle of 45.0° above the horizontal. If its muzzle velocity is 750.0 m s^{-1} , calculate the range achieved with a firing angle of:
 - (a) 40.0°
 - (b) 45.0°
 - (c) 50.0° .

13. A coastal defence cannon fires a shell horizontally from the top of a 50.0 m high cliff, directed out to sea as shown in figure 2.25, with a velocity of 1060.0 m s^{-1} . Calculate the range of the shell's trajectory.

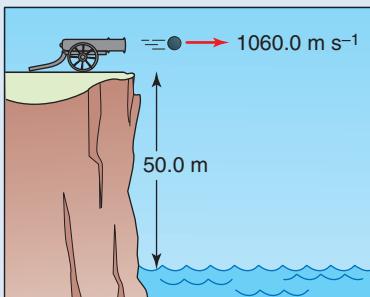


Figure 2.25

14. To increase the range of the shell in question 13, the cannon is lifted, so that it now points up at an angle of 45.0° as shown in figure 2.26. Calculate the new range.

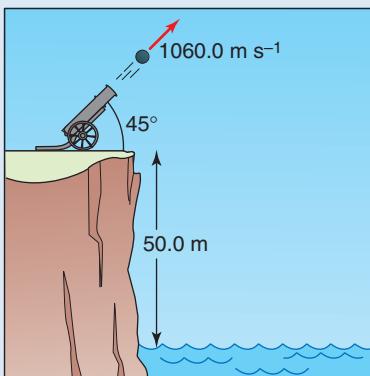


Figure 2.26

15. Identify the variables upon which the escape velocity of the Earth depends. If the mass of the Earth were somehow changed to four times its real value, state how the value of the escape velocity would change.
 16. Outline Newton's concept of escape velocity.
 17. Calculate the escape velocity of the following planets, using the data shown in the following table.

BODY	MASS (kg)	RADIUS (km)	ESCAPE VELOCITY (m s^{-1})
Mercury	3.3×10^{23}	2410	
Venus	4.9×10^{24}	6052	
Io	8.9×10^{22}	1821	
Callisto	1.1×10^{23}	2400	

18. A certain model rocket has a pre-launch mass of 87.3 g, of which 10.5 g is propellant. It is able to deliver a thrust of 6.10 N. Assuming that the rocket is fired directly up, calculate:
 (a) the initial rate of acceleration and g force
 (b) the final rate of acceleration and g force just prior to exhaustion of the fuel.
19. If a rocket had a mass of 32 000 kg, of which 85% was fuel, and a thrust of 400 000 N, calculate:
 (a) the rate of acceleration and g force at lift-off
 (b) the rate of acceleration and g force just prior to exhaustion of the fuel. Assume it is travelling horizontally and accelerating up to orbital velocity.
20. Identify the stage of a space mission during which an astronaut experiences the greatest g forces. Describe strategies that spacecraft designers can employ to ensure the survival of living occupants as well as delicate payloads.
21. Discuss the manner in which the rotation of the Earth and the revolution of the Earth around the Sun can be utilised by rocket designers.
22. (a) Explain rocket propulsion in terms of the Law of Conservation of Momentum.
 (b) Construct a diagram of a rocket to show the force pair that must exist due to Newton's Third Law.



2.1 MODELLING PROJECTILE MOTION

Aim

To model projectile motion by studying the motion of a ball bearing projected onto an inclined plane.

Apparatus

30 cm × 30 cm board	ball bearing
retort stand and clamp	graph paper
carbon paper	30 cm ruler (the ramp)

Theory

Galileo found that he could slow down the action of acceleration due to gravity by rolling a ball down a slope. In this way things happened slow enough for him to observe them. We are going to use that same strategy to slow down a projectile motion by projecting a ball bearing across an inclined plane. Recall that projectile motion can be considered as the addition of two linear motions at right angles to each other—the horizontal, constant velocity motion and the vertical, constant acceleration motion.

In the horizontal motion: $\Delta x = u_x t$

In the vertical motion: $\Delta y = u_y t + \frac{1}{2} a_y t^2$,

but we will let

$$u_y = 0$$

Thus $\Delta y = \frac{1}{2} a_y t^2$

We will assume that frictional forces can be neglected.

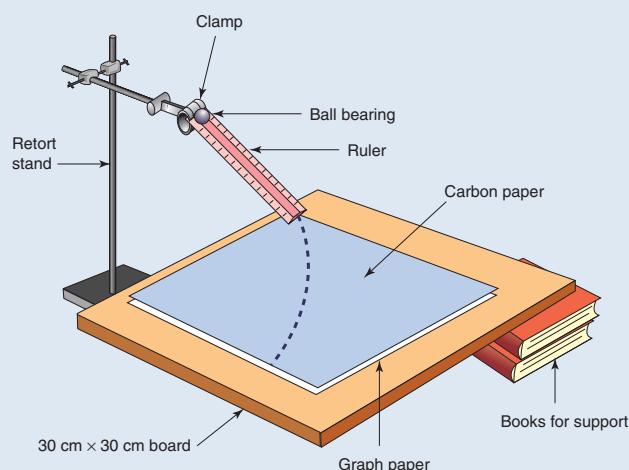


Figure 2.27 The path of the projectile (ball bearing) is marked as it rolls down the ramp on the carbon paper.

Method

- Set up the apparatus as shown in figure 2.27.
- Set up the inclined plane at an angle of approximately 20° and place the graph paper on it so that the ball will enter onto the inclined plane at a major division on the paper.
- Clamp the ruler so that the ball bearing rolling from it onto the inclined plane will be projected horizontally. Adjust the angle of the ruler so the path of the ball bearing will fit on the graph paper.
- Having adjusted the apparatus, place a piece of carbon paper on the graph paper and record the motion of the ball bearing projected onto the inclined plane.
- Remove the carbon paper and highlight the path for easier analysis.
- We will assume that the horizontal velocity of the ball bearing's motion remained constant. Therefore, the ball bearing took equal times to travel horizontally between the major divisions on the graph paper. Thus we can arbitrarily call one of these major divisions a unit of time. Beginning at the point where the ball entered the graph paper, label these major divisions 0, 1, 2, 3... time intervals.

Analysis

- Record and tabulate the distance down the slope that the ball bearing travelled during each time interval.
- Determine the average speed of the ball bearing down the slope during each time interval. Your answers should be in cm per time unit.
- Plot a graph of average speed down the slope versus time and determine a value for the acceleration of the ball down the slope. Your answer will be in cm per (time unit)².

Questions

- What do these graphs indicate about the motion of the ball down the plane?
- What assumptions have been made in order to obtain these results?
- How would the path of the ball bearing differ if:
 - the inclined plane was raised to a steeper angle while keeping the ramp as it was?
 - the angle of the ramp was raised and the inclined plane was kept as it was?
- The ball moves faster across the bottom of the paper than across the top, which represents an increase in kinetic energy. What is the source of this extra energy? Try to find out why the rolling mass of the ball introduces a problem into this energy conversion.



2.2 ACCELERATION AND LOAD DURING THE APOLLO 10 LAUNCH

Aim

To determine acceleration and load conditions applicable during the launch of *Apollo 10*.

Theory

Apollo 10 was launched on 18 May 1969, carrying a crew of three — Stafford, Young and Cernan. The rocket used for the Apollo missions was the *Saturn V*, the largest rocket ever built. Figure 2.28 is a press release diagram from 1969 and we will use it to extract some performance figures.

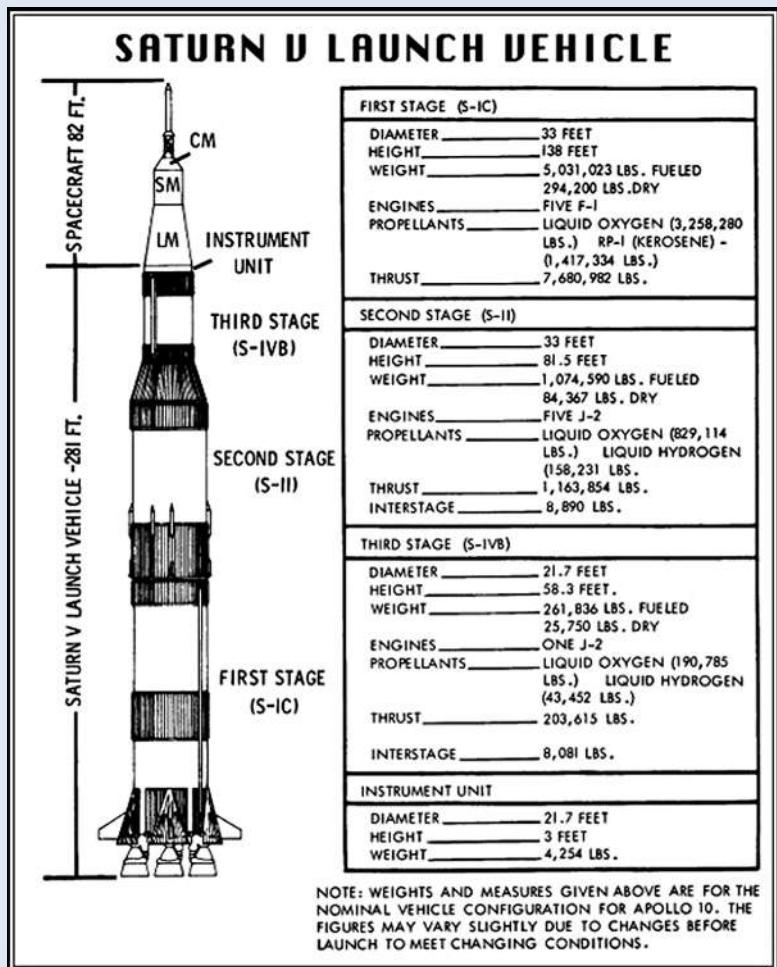


Figure 2.28 Specifications of an Apollo rocket

Note: The diagram distinguishes between the launch vehicle and the spacecraft. The spacecraft differed between missions, but on *Apollo 10* it consisted of the Lunar Module (LM) with a mass of 13 941 kg, and the Service Module (SM) and Command Module (CM) with a combined mass of 28 834 kg. The CM was the only part of the entire rocket to return to Earth.

The launch vehicle, the *Saturn V* rocket itself, was made up of three stages. The first stage consisted of five Rocketdyne F-1 engines, which burned liquid oxygen and kerosene. Upon launch it would burn for approximately 150 s before it was exhausted at an altitude of about 70 km. It then separated from the rocket and tumbled back down to the ocean.

The second stage used five Rocketdyne J-2 engines, which burned liquid oxygen and liquid hydrogen. After separation of the first stage, this second stage would ignite and burn for 365 s before it, too, separated from what remained of the rocket. Separation of stages was always accomplished by several small ‘interstage’ retro-rockets, which literally pulled the stage off. By now the rocket was at an altitude of 185 km with a speed of over 25 000 km h⁻¹.

The third stage, consisting of just one Rocketdyne J-2 engine, was then fired for 142 s before being shut down. The purpose of this burn was to insert the rocket into a 190 km high orbit with an orbital velocity of 28 100 km h⁻¹. The third stage rocket was not yet exhausted — it would be needed later to propel the craft out of orbit and toward the Moon — so let’s assume that half its fuel load was consumed in this burn.

Method

1. Note that the specifications in figure 2.28 are not SI units. The first task is to extract the information in part 1 of Results, and convert it to SI units. You will need the following conversion factors:

$$\begin{aligned} \text{Height: } 1 \text{ ft} &= 0.3048 \text{ m} \\ &1 \text{ mile} = 1.609 \text{ km} \end{aligned}$$

$$\text{Mass: } 1 \text{ lb} = 0.4536 \text{ kg}$$

$$\text{Thrust: } 1 \text{ lb} = 4.448 \text{ N}$$

2. Now use the information just extracted to fill in and complete the table at the bottom of this page. You will need the following formulas:

$$\text{Acceleration due to gravity } g = G \frac{m_E}{(r_E + \text{altitude})^2}$$

$$\text{Acceleration of a rocket } a = \frac{(T - mg \cos \theta)}{m}$$

(Allows for angle other than vertical.)

$$\text{Acceleration load or } g \text{ force} = \frac{T}{9.8m}$$

(To be consistent with above equation.)

where

$$G = \text{universal gravitation constant} \\ = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$$

$$m_E = \text{mass of Earth} \\ = 5.97 \times 10^{24} \text{ kg}$$

$$r_E = \text{radius of Earth} \\ = 6.38 \times 10^6 \text{ m}$$

$$T = \text{thrust (N)}$$

$$\theta = \text{angle of thrust from vertical (°).}$$

Results

1. Extract the following information from the theory above and from figure 2.28:

Spacecraft

$$\text{Height} = \text{_____ ft} = \text{_____ m} \\ \text{Mass} = \text{_____ lb} = \text{_____ kg}$$

Instrument unit

$$\text{Height} = \text{_____ ft} = \text{_____ m} \\ \text{Mass} = \text{_____ lb} = \text{_____ kg}$$

Third stage

$$\text{Height} = \text{_____ ft} = \text{_____ m} \\ \text{Mass fuelled} = \text{_____ lb} = \text{_____ kg} \\ \text{Mass dry} = \text{_____ lb} = \text{_____ kg} \\ \text{Thrust} = \text{_____ lb} = \text{_____ N}$$

Second stage

$$\text{Height} = \text{_____ ft} = \text{_____ m} \\ \text{Mass fuelled} = \text{_____ lb} = \text{_____ kg} \\ \text{Mass dry} = \text{_____ lb} = \text{_____ kg} \\ \text{Thrust} = \text{_____ lb} = \text{_____ N}$$

First stage

$$\text{Height} = \text{_____ ft} = \text{_____ m} \\ \text{Mass fuelled} = \text{_____ lb} = \text{_____ kg} \\ \text{Mass dry} = \text{_____ lb} = \text{_____ kg} \\ \text{Thrust} = \text{_____ lb} = \text{_____ N}$$

Entire Apollo 10 rocket

$$\text{Launch height} = \text{_____ m} \\ \text{Launch mass} = \text{_____ N}$$

2. Complete the results table below.

Questions

1. According to your table, what was the minimum and maximum g load experienced?

The actual maximum g loads experienced by Apollo astronauts at each stage were never quite as high as this, because they would turn the rocket engines off sequentially which would remove approximately $0.5g$ from the peak. In addition, air resistance would reduce the acceleration and resulting g force. Also, the minimum g loads were lower than that calculated, because between stages the rocket would coast for a few seconds, essentially in free fall, which placed the astronauts temporarily under a zero g load.

2. When do the greatest g loads occur during such a mission? See if you can find out the maximum loads experienced by an Apollo crew.

STAGE	TOTAL MASS OF ROCKET (kg)	AVAILABLE THRUST (N)	ALTITUDE (km)	g (m s^{-2})	ASSUMED ANGLE OF THRUST θ (°)	a (m s^{-2})	g FORCE LOAD
Launch Start 1st stage					0		
End 1st stage					45		
Start 2nd stage					45		
End 2nd stage					87		
Start 3rd stage					87		
End 3rd stage					90		

CHAPTER 3

ORBITING AND RE-ENTRY

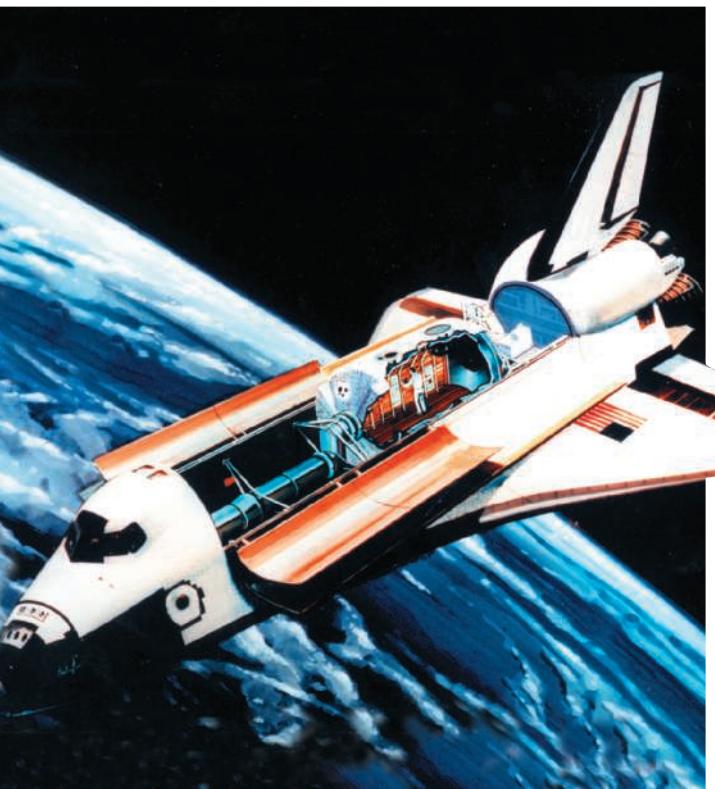


Figure 3.1 An artist's impression of the space shuttle in orbit. It occupies a low Earth orbit with an altitude between 250 km and 400 km.

Upon re-entry the space shuttle uses a unique flight pattern to minimise the load on its occupants.

Remember

Before beginning this chapter, you should be able to:

- describe and apply Newton's Second Law of Motion: $\sum F = ma$
- state the definition of momentum: $p = mv$
- state the definition of kinetic energy:
$$E_k = \frac{1}{2}mv^2$$

Key content

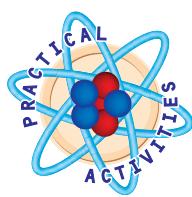
At the end of this chapter you should be able to:

- analyse the forces involved in a range of uniform circular motions, including the motion of a satellite orbiting the Earth
- be able to calculate the centripetal force acting on a satellite
- compare low Earth orbits to geostationary orbits
- define, describe and apply Kepler's Law of Periods
- define orbital velocity
- solve problems using Kepler's Law of Periods
- account for the orbital decay of satellites in low Earth orbit
- discuss the problems associated with a safe re-entry into the Earth's atmosphere and return to the Earth's surface
- identify the need for an optimum angle of re-entry and the consequences of failing to achieve it.

After a successful launch the next challenge is to sufficiently accelerate a spacecraft in order to place it into an orbit around the Earth. There are several different types and shapes of orbit, although our focus is upon low and geostationary circular orbits. In order to discover the orbital velocities required by these different types of orbit we will first look at circular motion and then apply this theory to orbits.

Most orbiting spacecraft do not need to be returned to Earth, although they will eventually fall back of their own accord. However, those with passengers do need to be returned, and in such a way as to keep the occupants alive. This means dealing with the potentially fatal problems of re-entering the Earth's atmosphere — the re-entry angle, the heat of re-entry and high g forces.

3.1 IN ORBIT



3.1

Investigating circular motion

Uniform circular motion is circular motion with a uniform orbital speed.

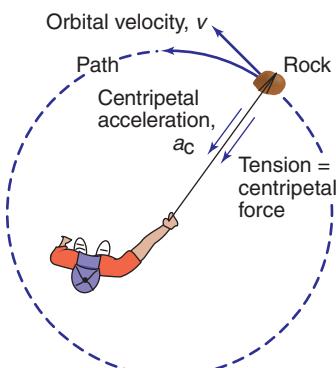


Figure 3.2 The string keeps the rock travelling in a circle, and gravity keeps a satellite in a circular orbit.

Centripetal force is the force that acts to maintain circular motion and is directed towards the centre of the circle.

Once a launched rocket has achieved a sufficient altitude above the surface of the Earth, it can be accelerated into the desired orbit. It must attain a specific speed that is dependent upon the mass of the Earth and the geometry of the orbit. If that speed is not reached, the spacecraft will follow a shortened elliptical orbit that dips back down toward the atmosphere, possibly causing immediate re-entry; if the speed is exceeded, the spacecraft will follow an elongated elliptical orbit that takes it away from the Earth. To see why this speed is so crucial we first need to study the simplest orbital motion — a uniform speed along a circular path around the Earth.

Uniform circular motion

Uniform circular motion is circular motion with a uniform orbital speed. As an example of circular motion, imagine you have a rock tied to a string and are whirling it around your head in a horizontal plane. Because the path of the rock is in a horizontal plane, as shown in figure 3.2, gravity plays no part in its motion. The Greek, Aristotle, considered that circular motion was a perfect and natural motion, but it is far from it. If you were to let go of the string, the rock would fly off at a tangent to the circle — a demonstration of Newton's First Law of Motion. He said that an object would continue in uniform motion in a straight line unless acted upon by a force. In the case of our rock, the force keeping it within a circular path is the tension in the string, and it is always directed back towards the hand at the centre of the circle. Without that force the rock will travel in a straight line.

The same is true of a spacecraft in orbit around the Earth, or any object in circular motion — some force is needed to keep it there, and that force is directed back towards the centre of the circle. In the case of the spacecraft, it is the gravitational attraction between the Earth and the spacecraft that acts to maintain the circular motion that is the orbit.

The force required to maintain circular motion, known as **centripetal force**, can be determined using the following equation.

$$\text{Centripetal force, } F_C = \frac{mv^2}{r}$$

where

F_C = centripetal force (N)

m = mass of object in motion (kg)

v = instantaneous or orbital velocity of the mass (m s^{-1})

r = radius of circular motion (m).

Newton's Second Law states that wherever there is a net force acting on an object there is an associated acceleration. Since this centripetal force is the only force acting on the motion, we can say that:

$$\text{Centripetal force, } \mathbf{F}_C = \text{mass} \times \text{centripetal acceleration, } \mathbf{a}_C$$

and therefore,

$$\text{centripetal acceleration, } \mathbf{a}_C = \frac{\mathbf{v}^2}{r}.$$

It may not be immediately apparent to you that the whirling rock in circular motion is accelerating. Consider that at any instant the velocity of the rock is at a tangent to the circle. As it progresses around the circle, the direction of its velocity is constantly changing, even though its magnitude (its speed) remains unchanged. Now recall that velocity is a vector quantity — to change it you need only change its direction. Hence, the velocity vector of the rock is constantly changing with time. This is the **centripetal acceleration** and it is also directed towards the centre of the circle, as shown in figure 3.2.

Several common circular motions and their centripetal forces are shown in table 3.1. In each case the centripetal force is directed back towards the centre of the circle.

Centripetal acceleration is always present in uniform circular motion. It is associated with centripetal force and is also directed towards the centre of the circle.

Table 3.1 A comparison of common circular motions

MOTION	F_C PROVIDED BY ...
Whirling rock on a string	The string
Electron orbiting atomic nucleus	Electron–nucleus electrical attraction
Car cornering	Friction between tyres and road
Moon revolving around Earth	Moon–Earth gravitational attraction
Satellite revolving around Earth	Satellite–Earth gravitational attraction

SAMPLE PROBLEM

3.1

Centripetal force and acceleration on a whirled rock

A rock of mass 250 g is attached to the end of a 1.5 m long string and whirled in a horizontal circle at 15 m s⁻¹. Calculate the centripetal force and acceleration of the rock.

SOLUTION

Centripetal force,

$$\begin{aligned}\mathbf{F}_C &= \frac{mv^2}{r} \\ &= \frac{0.250 \times 15^2}{1.5} \\ &= 37.5 \text{ N}\end{aligned}$$

Centripetal acceleration can be found using its formula or, more simply, using Newton's Second Law.

$$\begin{aligned}\mathbf{a}_C &= \frac{\mathbf{F}_C}{m} \\ &= \frac{37.5}{0.25} \\ &= 150 \text{ m s}^{-2}\end{aligned}$$

Calculating frictional force on a turning car

A car of mass 1450 kg is driven around a bend of radius 70.0 m. Determine the frictional force required between the tyres and the road in order to allow the car to travel at 70.0 km h⁻¹.

SOLUTION

The frictional force between the tyres and the road must provide sufficient centripetal force for the circular motion involved.

Firstly, note that $70.0 \text{ km h}^{-1} = \frac{70.0}{3.6} \text{ m s}^{-1} = 19.4 \text{ m s}^{-1}$.

$$\begin{aligned}\text{Centripetal force, } F_C &= \frac{mv^2}{r} \\ &= \frac{1450 \times 19.4^2}{70} \\ &= 7800 \text{ N.}\end{aligned}$$

That is, the total frictional force provided by the tyres must be at least 7800 N, or an average force of 1950 N per tyre.

From the previous chapters you will recall that an astronaut in orbit around the Earth still experiences an acceleration due to gravity of about 8.8 m s⁻². This, in turn, means that the astronaut still has significant true weight. It should now be clear that this acceleration due to gravity acts as the centripetal acceleration of the orbital motion and the astronaut's weight forms the centripetal force. Why, then, does the astronaut feel weightless? It is for the same reason that a person in a falling elevator also experiences weightlessness during the fall.

You should also recall that apparent weight is the sensation of weight created by those forces resisting a body's true weight. In the case of an astronaut in an orbiting spacecraft, and of a person in a falling elevator, there are no resisting forces acting on the person so that there is no apparent weight. Referring back to figures 2.17 and 2.19, we can see that the acceleration in both cases is $-g$ (taking 'up' to be the positive direction). Therefore, g force experienced is:

$$\begin{aligned}g \text{ force} &= \frac{g + a}{9.8 \text{ m}} \\ &= \frac{g + (-g)}{9.8 \text{ m}} \\ &= 0, \text{ that is, zero apparent weight.}\end{aligned}$$

eBook plus

Weblink:
Kepler's third law

Period, T , is the time taken to complete one orbit.

Kepler's third law — the Law of Periods

Johannes Kepler (1571–1642) discovered his third law, the Law of **Periods**, through trial and error in the very early 1600s. He had access to extraordinarily detailed observations of the motions of the planets made by Tycho Brahe, and in attempting to analyse them he came upon this relationship, among others. He expressed this law in this form:

$$\left(\frac{r^3}{T^2}\right) \text{ for planet 1} = \left(\frac{r^3}{T^2}\right) \text{ for planet 2}$$

This relation can be used to compare any two bodies orbiting the same object, for example, any two moons orbiting Jupiter or any two planets orbiting the Sun. An alternative expression of this law is:

$$\frac{r^3}{T^2} = k \text{ for any satellites orbiting a common central mass,}$$

Students of astrophysics will learn that the variable M in Kepler's Law of Periods actually refers to the combined mass of the system. The derivation can be found in chapter 16 on page 307. In the case of a satellite, however, M is approximately equal to the mass of the planet being orbited.

SAMPLE PROBLEM

3.3

where

- r = the radius of the orbit of any given satellite
- T = the period of that satellite's orbit
- k = a constant.

In chapter 4 we will see how the Law of Universal Gravitation can be used to derive an expression for the constant k, so that Kepler's Law of Periods becomes:

$$\frac{r^3}{T^2} = \frac{GM}{4\pi^2}$$

where

$$\begin{aligned} G &= \text{the universal gravitation constant} \\ &= 6.676 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2} \end{aligned}$$

M = the mass of the central body.

Calculating the periods of satellites

Calculate the periods of three different satellites orbiting the Earth at altitudes of (a) 250 km, (b) 400 km and (c) 40 000 km. The radius of the Earth is 6.38×10^6 m and the Earth's mass is 5.97×10^{24} kg.

SOLUTION

(a) At an altitude of 250 km:

$$\begin{aligned} \frac{r^3}{T^2} &= \frac{GM_E}{4\pi^2} \\ \frac{(6.38 \times 10^6 + 250 \times 10^3)^3}{T^2} &= \frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})}{4\pi^2} \\ \therefore T &= 5375 \text{ seconds} = 89.6 \text{ minutes} \end{aligned}$$

Note that the radius of the orbit equals the radius of the Earth plus the altitude.

(b) At an altitude of 400 km:

$$\begin{aligned} \frac{r^3}{T^2} &= \frac{GM_E}{4\pi^2} \\ \frac{(6.38 \times 10^6 + 400 \times 10^3)^3}{T^2} &= \frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})}{4\pi^2} \\ \therefore T &= 5560 \text{ seconds} = 92.7 \text{ minutes} \end{aligned}$$

(c) At an altitude of 40 000 km:

$$\begin{aligned} \frac{r^3}{T^2} &= \frac{GM_E}{4\pi^2} \\ \frac{(6.38 \times 10^6 + 40000 \times 10^3)^3}{T^2} &= \frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})}{4\pi^2} \\ \therefore T &= 99450 \text{ seconds} = 27.6 \text{ hours} \end{aligned}$$

Orbital velocity

Orbital velocity is the instantaneous direction and speed of an object in circular motion along its path. For uniform circular motion, its magnitude is constant and inversely proportional to the period of the orbit, as follows:

$$\text{orbital velocity } v = \frac{\text{circumference of the circle}}{\text{period } T}$$

$$v = \frac{2\pi r}{T}$$

This general formula can be applied to any object in circular motion.

SAMPLE PROBLEM**3.4****Calculating the orbital velocity of a turning car**

Calculate the orbital velocity of a car that travels completely around a 20.0 m radius roundabout in 8.00 s.

SOLUTION

$$\begin{aligned} v &= \frac{2\pi r}{T} \\ &= \frac{2\pi \times 20.0}{8.00} \\ &= 15.7 \text{ m s}^{-1} \end{aligned}$$

SAMPLE PROBLEM**3.5****Calculating the orbital velocity of satellites**

Determine the orbital velocities of the three satellites in sample problem 3.3, that is, with orbits of altitude (a) 250 km, (b) 400 km, and (c) 40 000 km.

SOLUTION

(a) At an altitude of 250 km:

$$\begin{aligned} v &= \frac{2\pi r}{T} \\ &= \frac{2\pi(6.38 \times 10^6 + 250 \times 10^3)}{5375} \\ &= 7750 \text{ ms}^{-1} \approx 27900 \text{ km h}^{-1} \end{aligned}$$

(b) At an altitude of 400 km:

$$\begin{aligned} v &= \frac{2\pi r}{T} \\ &= \frac{2\pi(6.38 \times 10^6 + 400 \times 10^3)}{5560} \\ &= 7660 \text{ m s}^{-1} \approx 27600 \text{ km h}^{-1} \end{aligned}$$

(c) At an altitude of 40 000 km:

$$\begin{aligned} v &= \frac{2\pi r}{T} \\ &= \frac{2\pi(6.38 \times 10^6 + 40000 \times 10^3)}{99450} \\ &= 2930 \text{ m s}^{-1} \approx 10550 \text{ km h}^{-1} \end{aligned}$$

Note the decrease in orbital velocity as the radius of the orbit increases.

If this expression for orbital velocity is substituted into Kepler's Law of Periods, then a formula for orbital velocity emerges that is specific to a satellite.

$$\begin{aligned} v &= \frac{2\pi r}{T} \text{ so } T = \frac{2\pi r}{v} \\ \therefore \frac{r^3}{\left(\frac{2\pi r}{v}\right)^2} &= \frac{GM}{4\pi^2} \end{aligned}$$

Rearranging gives

$$v = \sqrt{\frac{GM}{r}}$$

where

v = orbital velocity (m s^{-1})

G = universal gravitation constant

$= 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$

M = central mass (kg)

r = radius of the orbit (m).

Note that the value of a satellite's orbit depends on:

- the mass of the planet being orbited
- the radius of the orbit. For a satellite orbiting a planet, this is equal to the radius of the planet plus the altitude of the orbit.

Hence, for the case of a satellite orbiting the Earth, the formula becomes:

$$v = \sqrt{\frac{G m_E}{r_E + \text{altitude}}}$$

where

$$v = \text{orbital velocity (m s}^{-1}\text{)}$$

$$m_E = \text{mass of the Earth}$$

$$= 5.97 \times 10^{24} \text{ kg}$$

$$r_E = \text{radius of the Earth}$$

$$= 6.38 \times 10^6 \text{ m}$$

$$\text{altitude} = \text{height of orbit above the ground (m)}.$$

It is clear from this formula that altitude is the only variable that determines the orbital velocity required for a specific orbit. Further, the greater the radius of the orbit, the lower that velocity is.

SAMPLE PROBLEM

3.6

Calculating orbital velocity from altitude

Verify the results of sample problem 3.5 by calculating the orbital velocities directly from the altitudes.

SOLUTION

(a) At an altitude of 250 km:

$$\begin{aligned} v &= \sqrt{\frac{G m_E}{r_E + \text{altitude}}} \\ &= \sqrt{\frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})}{(6.38 \times 10^6 + 250 \times 10^3)}} \\ &= 7750 \text{ ms}^{-1} \approx 27900 \text{ km h}^{-1} \end{aligned}$$

(b) At an altitude of 400 km:

$$\begin{aligned} v &= \sqrt{\frac{G m_E}{r_E + \text{altitude}}} \\ &= \sqrt{\frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})}{(6.38 \times 10^6 + 400 \times 10^3)}} \\ &= 7660 \text{ m s}^{-1} \approx 27600 \text{ km h}^{-1} \end{aligned}$$

(c) At an altitude of 40 000 km:

$$\begin{aligned} v &= \sqrt{\frac{G m_E}{r_E + \text{altitude}}} \\ &= \sqrt{\frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})}{(6.38 \times 10^6 + 40000 \times 10^3)}} \\ &= 2930 \text{ m s}^{-1} \approx 10550 \text{ km h}^{-1} \end{aligned}$$

Orbital energy

Any satellite travelling in a stable circular orbit at a given orbital radius has a characteristic total mechanical energy E . This is the sum of its kinetic energy E_k (due to its orbital velocity) and its gravitational potential energy E_p (due to its height). The kinetic energy equation is:

$$E_k = \frac{1}{2} mv^2, \text{ and we have seen that } v = \sqrt{\frac{GM}{r}}.$$

Combining these equations gives a new equation for the kinetic energy of an orbiting satellite:

$$E_k = \frac{GMm}{2r}$$

where

M = mass of the central body being orbited (kg)

m = mass of the satellite (kg)

r = radius of the orbit (m).

We know from chapter one that $E_p = \frac{GMm}{r}$. Note that for a stable circular orbit the value of E_k is always half that of the E_p but positive in value. An expression for total mechanical energy can now be determined.

$$\text{Mechanical energy } E = E_k + E_p$$

$$= \frac{1}{2} \left(\frac{GMm}{r} \right) - \left(\frac{GMm}{r} \right)$$

$$\therefore E = -\frac{1}{2} \left(\frac{GMm}{r} \right)$$

This equation looks very similar to the equation for E_p and also represents a negative energy well. The value of the mechanical energy of a satellite orbiting a planet depends only on the masses involved and the radius of the orbit. A lower orbit produces a more negative value of E and, therefore, less energy, while a higher orbit corresponds to more energy.

A useful concept for comparing orbits is the specific orbital energy of a satellite, which is the mechanical energy per kilogram.

$$\text{Specific orbital energy } \varepsilon = \frac{E}{m} = -\frac{GM}{2r} \text{ for circular orbits.}$$

Refer to table 3.2, which lists orbital data for several different types of satellites. The first four rows list satellites with near-circular orbits but with increasing radii. Note that the specific orbital energy also increases as radius increases.

Elliptical orbits

The preceding theory assumes that we are dealing with circular orbits, but that is usually not the case. The most common orbital shape is an ellipse, or oval shape. Kepler also realised this and stated it as his first law. Ellipses can be round or elongated — the degree of stretch is known as eccentricity. Referring to figure 3.3, we can see that eccentricity is

defined as the ratio $\frac{c}{2a}$ where c is the distance between the two foci

of the ellipse and a is the semi-major axis. In fact, a circle is an ellipse with an eccentricity of zero. Most satellites are placed into near-circular orbits, as shown in table 3.2, but there are a few notable exceptions. Each of these types of orbits is discussed over the next few pages.

Much of the information shown in table 3.2 can be calculated if the semi-major axis a is known, and this can be determined from the *apogee* and *perigee* distances.

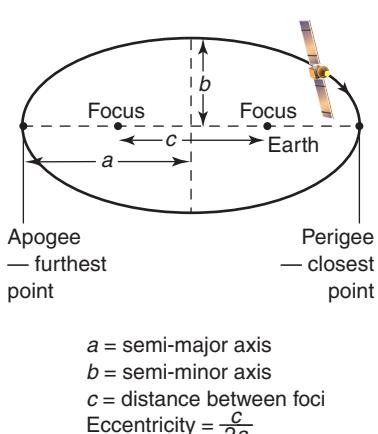


Figure 3.3 Various dimensions of an ellipse

Table 3.2 A sample of various Earth satellites as at February 2008, sorted according to the specific orbital energy of their orbits.

Satellite name	Purpose	Orbit description	Specific orbital energy (MJ kg^{-1})	Period (min)	Inclination (degrees)	Perigee altitude (km)	Velocity at perigee (km h^{-1})	Apogee altitude (km)	Velocity at apogee (km h^{-1})	Eccentricity
GENESAT	Biological research and amateur radio beacon	Low Earth orbit	-29.4	93	40	397	27 590	401	27 580	0.0050
USA 197	Military eye-in-the-sky	Polar low Earth orbit	-28.4	97	97.8	627	27 140	630	27 120	0.0024
IRIDIUM 95	Satellite phone communication	Low Earth orbit	-28.2	98	86.6	670	27 050	674	27 040	0.0030
NOAA 18	Weather	Polar low Earth orbit	-27.5	102	98.8	845	26 740	866	26 660	0.0123
RASCOM 1	African communications	Transfer orbit to geostationary position	-8.1	638	5.4	587	35 650	35 745	5 900	0.9677
MOLNIYA 3–53	TV and military communications	Elliptical Molniya orbit	-7.5	718	64.9	1 047	34 570	39 308	5 620	0.9481
NAVSTAR 59	Global Positioning System	High altitude GPS orbit	-7.5	718	55.2	20 092	13 980	20 273	13 890	0.0045
OPTUS D2	Australian and NZ television communications	Geostationary orbit	-4.7	1436	0	35 776	11 060	35 798	11 060	0.0003
SKYNET 5B	Military communications	Geostationary orbit	-4.7	1436	0.1	35 773	11 060	35 810	11 060	0.0005

Perigee or periapsis?

In orbital mechanics, the general term for the point of closest approach to the central body is periapsis, and the furthest point is apoapsis. However, these terms adapt to the body being orbited. When considering satellites orbiting the Earth, the terms become perigee and apogee. If orbiting the Moon, the terms become perilune and apolune; and if the Sun is orbited, then they are perihelion and aphelion. There are a range of other similar terms for other celestial objects that can be orbited.

Some relevant equations are:

$$\text{Semi-major axis } a = \frac{r_A + r_P}{2} = \frac{\text{apoee altitude} + \text{perigee altitude}}{2} + r_E$$

$$\text{Eccentricity } e = \frac{c}{2a} = \frac{r_A - r_P}{r_A + r_P}$$

$$\text{Specific orbital energy } \varepsilon = -\frac{GM}{2a}$$

The specific orbital energy equation is a more general form than that given earlier, and has been used to order the satellites in the table. You should note that as the size of the near-circular orbits increases so too does the energy. Note also that the specific orbital energy of an elliptical orbit lies between that of circular orbits that correspond to the ellipse's perigee and apogee altitudes.

The velocity of a satellite at any point along an elliptical path can be calculated using the following general equation:

$$v = \sqrt{GM\left(\frac{2}{r} - \frac{1}{a}\right)}$$

where

r = the orbital radius at the point being considered.

The satellite velocities at apogee and perigee in table 3.2 were calculated using this formula. (You should confirm that for a circular orbit $a = r$ and that this formula simplifies to the orbital velocity equation given on page 44.) Notice how the velocities of each satellite are least at the apogee and greatest at the perigee, that is, the satellites move quickly when closest to the Earth and slow down as they move further away. This, of course, is just what is described in Kepler's second law.

The Molniya orbit was developed specifically for this reason. Devised for Russian communications (as most of Russia lies too far north to be satisfactorily covered by a geostationary satellite), this very eccentric orbit places a high apogee over the desired location. A Molniya satellite will cruise slowly through this apogee before zipping around through the low perigee and returning quickly to the coverage area.

Types of orbit

Spacecraft or satellites placed into orbit will generally be placed into one of two altitudes — either a low Earth orbit or a geostationary orbit.

A **low Earth orbit** is generally an orbit higher than approximately 250 km, in order to avoid atmospheric drag, and lower than approximately 1000 km, which is the altitude at which the Van Allen radiation belts start to appear. These belts are regions of high radiation trapped by the Earth's magnetic field and pose significant risk to live space travellers as well as to electronic equipment. The space shuttle utilises a low Earth orbit somewhere between 250 km and 400 km depending upon the mission. At 250 km, an orbiting spacecraft has a velocity of 27900 km h^{-1} and takes just 90 minutes to complete an orbit of the Earth.

A **geostationary orbit** is at an altitude at which the period of the orbit precisely matches that of the Earth. If over the Equator, such an orbit would allow a satellite to remain 'parked' over a fixed point on the surface of the Earth throughout the day and night. From the Earth such a satellite appears to be stationary in the sky, always located in the same direction regardless of the time of day. This is particularly useful for communications satellites because a receiving dish need only point to a fixed spot in the sky in order to remain in contact with the satellite.

The altitude of such an orbit can be calculated from Kepler's Law of Periods. Firstly, the period of the orbit must equal the length of one sidereal day; that is, the time it takes the Earth to rotate once on its orbit, relative to the stars. This is 3 minutes and 56 seconds less than a 24-hour solar day, so that T is set to be 86 164 s. The radius of the orbit then works out to be 42 168 km, or 6.61 Earth radii. Subtracting the radius of the Earth gives the altitude as approximately 35 800 km. This places the satellite at the upper limits of the Van Allen radiation belts and near the edge of the magnetosphere, making them useful for scientific purposes as well. Australia has the AUSSAT and OPTUS satellites in geostationary orbits.

If a satellite at this height is not positioned over the Equator but at some other latitude, it will not remain fixed at one point in the sky. Instead, from the Earth the satellite will appear to trace out a 'figure of eight' path each 24 hours. It still has a period equal to Earth's, however, so this orbit is referred to as geosynchronous.

A **low Earth orbit** is an orbit higher than 250 km and lower than 1000 km.

A **geostationary orbit** is at an altitude at which the period of the orbit precisely matches that of the Earth. This corresponds to an altitude of approximately 35 800 km.

A **transfer orbit** is an orbit used to manoeuvre a satellite from one orbit to another.

A **transfer orbit** is a path used to manoeuvre a satellite from one orbit to another. Satellites headed for a geostationary orbit are first placed into a low Earth orbit and then boosted up from there using a transfer orbit, which has a specific orbital energy that lies between that of the lower and higher circular orbits. Orbital manoeuvres utilise Keplerian motion, which is not always intuitive. In order to move a satellite into a different orbit, the satellite's energy must be changed; this is achieved by rapidly altering the kinetic energy. Rockets are fired to change the satellite's velocity by a certain amount, referred to as 'delta-v' (Δv), which will increase or decrease the kinetic energy (and therefore the total energy) to alter the orbit as desired. However, as soon as the satellite begins to change altitude, transformations between the E_p and the E_k occur, so its speed is continually changing.

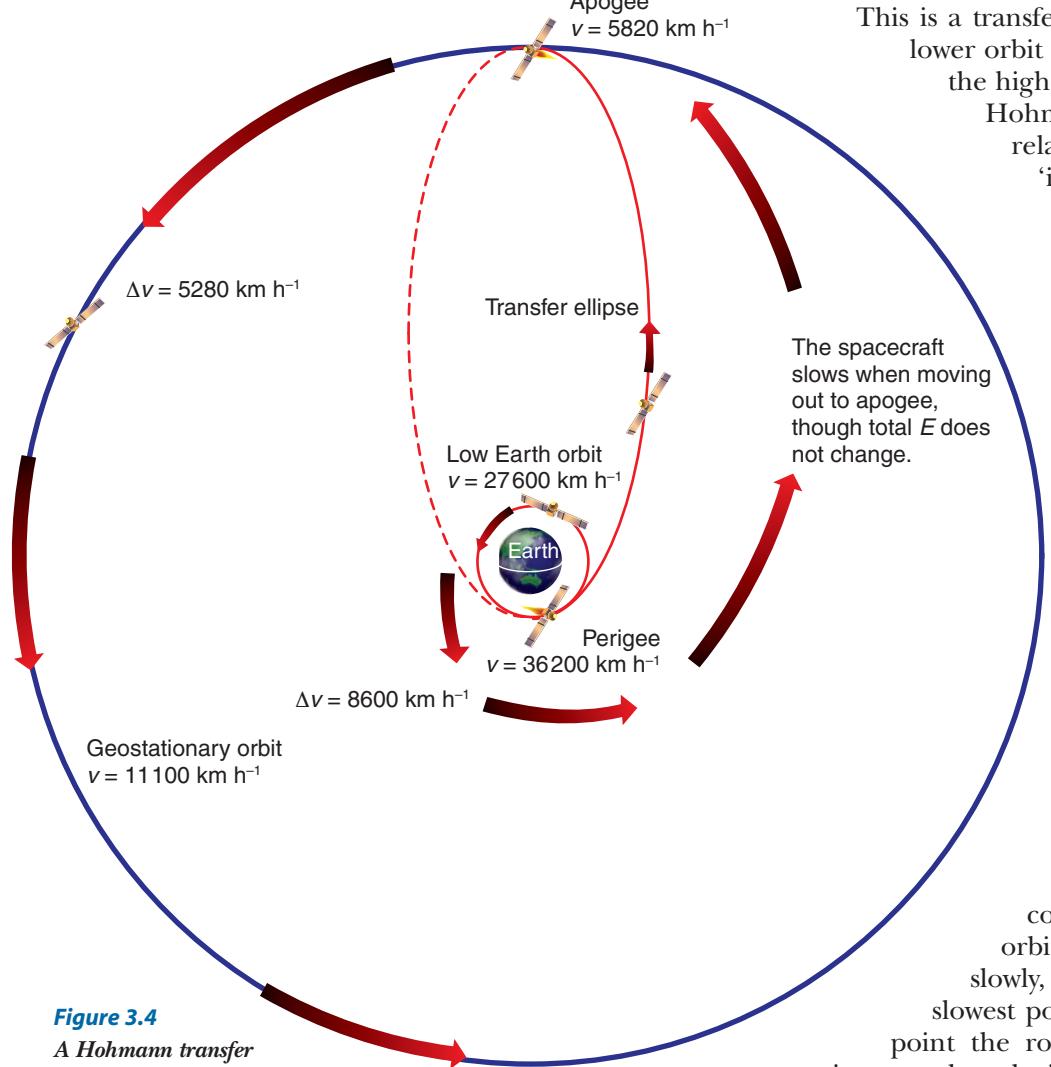


Figure 3.4
A Hohmann transfer orbit used to raise a satellite from a low Earth orbit of altitude 400 km up to a geostationary orbit of altitude 35 800 km

The simplest and most fuel-efficient path is a Hohmann transfer orbit, as shown in figure 3.4. This is a transfer ellipse that touches the lower orbit at its perigee and touches the higher orbit at its apogee. The Hohmann transfer involves two relatively quick (called 'impulsive') rocket boosts. (In orbital mechanics, the word 'impulsive' describes a quick change in velocity and energy.) In order to move to a higher orbit, the first boost increases the satellite's velocity, stretching the circular low Earth orbit out into a transfer ellipse. The perigee is the fastest point on this transfer orbit, and as the satellite moves along the ellipse it slows again. When it finally reaches the apogee, it will be at the correct altitude for its new orbit, but it will be moving too slowly, the apogee being the slowest point on the ellipse. At this point the rockets are fired again, to increase the velocity to that required for the new higher, stable and circular orbit.

In order to move down from a higher to a lower orbit, the process is reversed, requiring two negative delta-v rocket boosts, that is, retro-firing of the rockets. These two boosts will slow the satellite, hence changing its orbit, first from the higher circular orbit into a transfer ellipse that reaches down to lower altitudes, then from the perigee of the transfer ellipse into a circular low Earth orbit.

PHYSICS IN FOCUS

Other types of orbit

There are several, more unusual types of orbit. One is a low altitude polar orbit. A satellite flying 1000 km over the North Pole and then the South Pole, orbiting the Earth once every 100 minutes, will, over the course of 24 hours, be able to survey the entire globe as it spins beneath it.

If the plane of the orbit is about 8 degrees off the north-south plane, the mass of the Earth's equatorial bulge causes the plane of the orbit to slowly rotate in time with the Earth's rotation so that it maintains its attitude to the Sun. Such a sun-synchronous orbit allows a satellite to always orbit along a path over the twilight between day and night, known as the terminator.

An even stranger orbit is an orbit at a Langrangian point. A Langrangian point is a position in space at which a satellite can maintain a stable orbit despite the gravitational influence of two significant masses, in this case the Sun and the Earth.

Any spacecraft launched toward the Sun will, according to Kepler's Law of Periods, begin to speed up as its orbital radius (its distance from the Sun) decreases and the gravitational pull of the Sun increases, thereby increasing the centripetal

force. This means that it will begin to speed ahead of the Earth and eventually lose contact. However, the Earth's gravity is pulling in the opposite direction and, at a particular distance, it will reduce the pull of the Sun just enough to allow the spacecraft to slow to a speed that matches the Earth's progress. This point is known as the L1 point and is approximately four times the distance to the Moon, or one-hundredth of the distance to the Sun.

A spacecraft placed at the L1 point will orbit the Sun along with the Earth, maintaining its relative position between the Sun and the Earth throughout the year. This strategy is ideal for studying the solar wind, and has been used for the Advanced Composition Explorer (ACE) and the Solar and Heliospheric Observatory (SOHO). Rather than residing at the point, these spacecraft move in small orbits around it.

There are several other Langrangian points, such as the L2 directly behind the Earth, at which point the Earth's gravity adds to the Sun's to speed up a satellite so that it can accompany the Earth, as well as more points around the Moon.

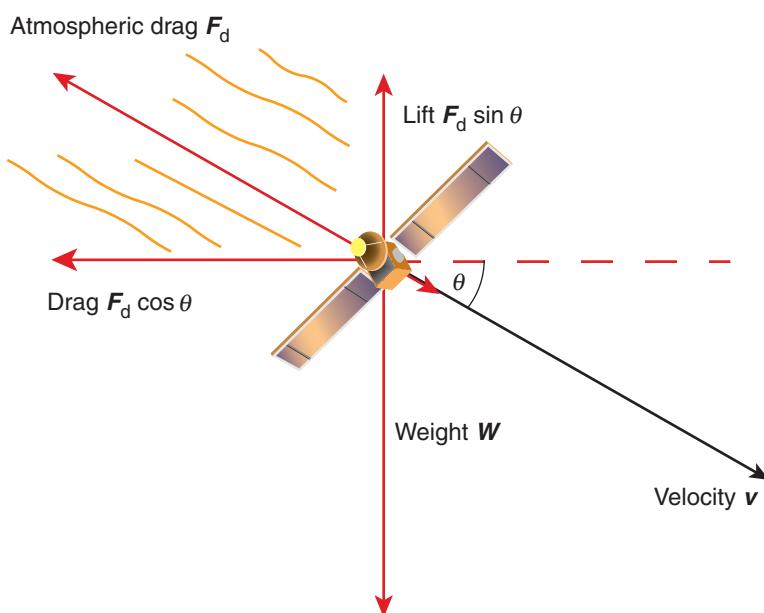


Figure 3.5 Forces acting on a satellite

Orbital decay

All satellites in low Earth orbit are subject to some degree of atmospheric drag that will eventually decay their orbit and limit their lifetimes. Although the atmosphere is very thin 1000 km above the surface of the Earth, it is still sufficient to cause some friction with a satellite. This friction causes a gradual ('non-impulsive') loss of energy to heat. Referring back to the equation for orbital energy, you can see that a loss of orbital energy necessarily means a loss of altitude, so that this gradual loss of energy causes a low-orbiting satellite to slowly spiral back towards the Earth.

A number of factors combine to make this an accelerating process. Referring to figure 3.5, we see that the two forces acting on a low-orbiting satellite are its weight and atmospheric drag. The equation for the atmospheric drag is as follows:

$$F_d = -\frac{1}{2} \rho v^2 C_d A$$

where

$$\rho = \text{air density } (\text{kg m}^{-3})$$

$$v = \text{velocity } (\text{m s}^{-1})$$

C_d = coefficient of drag (a ratio expressing how streamlined the shape is)

A = cross-sectional area (m^2).

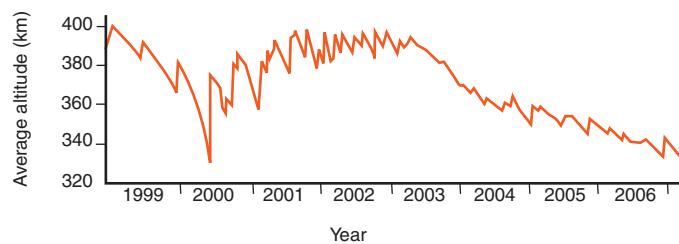
As the satellite descends this force increases for two reasons:

- The density of the atmosphere increases by a factor of almost 10^9 from an altitude of 150 km down to ground level.
- As the satellite loses altitude some of its E_p is transformed into E_k and it speeds up — it is literally starting to fall back to Earth. The drag force is proportional to the square of the velocity, so as the speed increases the drag increases much more sharply.

At an altitude of about 80 km the atmospheric drag increases sufficiently to start slowing the descending satellite; however, by now the increasing air density means that the braking effect is building quite rapidly. At some point, usually around an altitude of 60 km, the atmospheric drag increases sharply, leading to a catastrophe of heat and g force. Braking occurs so suddenly that the heat generated usually becomes sufficient to vaporise all but the largest of satellites, in addition to the generation of extreme g forces.

Designers plan for an expected satellite lifetime by building in small rocket boosters so that the satellite can be lifted periodically back up to its intended orbital altitude. Figure 3.6 shows the changes in the altitude of the International Space Station over several years. Being a very large, unstreamlined shape, it experiences more drag than most other satellites; it loses about 90 metres of altitude per day due to atmospheric drag. It is lifted regularly — the larger lifts were done by visiting space shuttles (which ceased for some time after the *Columbia* disaster in 2003) and the smaller ones by Russian rockets.

Figure 3.6 Altitude changes of the International Space Station. The drops in altitude represent orbital decay due to atmospheric drag. The increases in altitude are lifts performed by visiting spacecraft.



However, the actual service life of a satellite can be unpredictable as the atmosphere itself can change. For example, an increase in solar radiation can cause the atmosphere to expand and rise up to meet a satellite, increasing the atmospheric density at that altitude. Many satellites have been prematurely lost this way during a solar cycle maximum, including the first US space station, Skylab, in 1979.

3.2 RE-ENTRY

The process of deliberately leaving a stable Earth orbit and re-entering the atmosphere in order to return to the surface of the Earth is known as ‘de-orbiting’. De-orbiting has some significant differences from orbital decay. Orbital decay is an unintended, gradual (‘non-impulsive’) process resulting in a spiral path downward, whereas de-orbiting is a deliberate, impulsive orbital manoeuvre resulting in an elliptical path down to the atmosphere.

The de-orbit manoeuvre

The first phase of the de-orbit manoeuvre is to alter the spacecraft's orbit into a transfer ellipse that intersects the atmosphere at the desired angle. A shallow angle is selected to minimise the extreme heating and g forces that can destroy a spacecraft on an uncontrolled re-entry. However, if the angle is too shallow, the spacecraft may skip off the atmosphere instead of penetrating it. For example, as shown in figure 3.7, the optimum angle for the Apollo missions was between 5.2° and 7.2° .

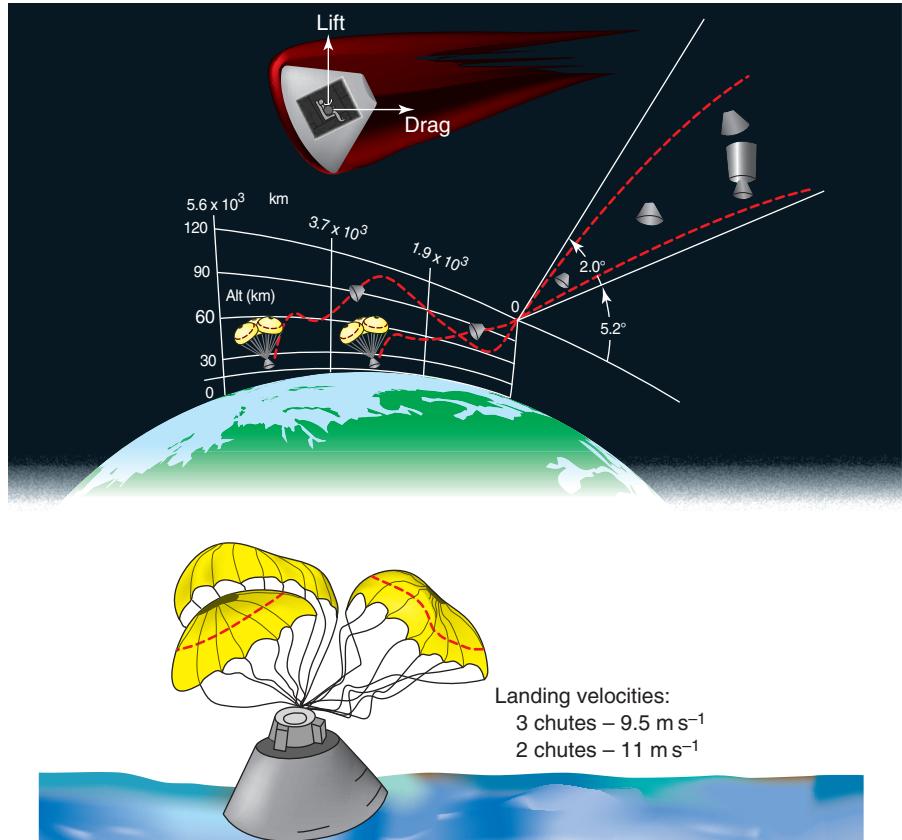


Figure 3.7 Re-entry of an Apollo capsule

Just as for other transfer ellipses, in order to collapse the shape of the stable circular orbit into a smaller transfer ellipse, some energy must be lost. This is achieved by a retroburn of the spacecraft's rockets — that is, pointing them ahead of the spacecraft and executing a short burn to quickly reduce the velocity and thus the kinetic energy. The required transfer ellipse must be calculated in advance, as this will determine when and for how long the retroburn must occur in order to achieve the required delta-v.

The next phase is the glide phase, in which the spacecraft's orbit changes and carries it down to meet the atmosphere at the 're-entry interface', around 120 km altitude. While the spacecraft is gliding down, no further manoeuvring is required. However, the spacecraft is speeding up again because gravitational potential energy is transforming into kinetic energy as it falls.

The third phase of the process is the actual re-entry into and through the atmosphere, during which the issues of heat, g forces, ionisation blackout and reaching the surface must be dealt with.

An extreme heat

Why is it that re-entry produces so much heat? Consider that the spacecraft has a velocity, even after retrofiring, of tens of thousands of

kilometres per hour. This velocity means that the spacecraft has significant kinetic energy. Additionally, the altitude of the spacecraft's orbit means that it also has considerable gravitational potential energy, which is also lost as the spacecraft's altitude decreases during re-entry.

As the spacecraft re-enters, it experiences friction with the molecules of the atmosphere. This friction is a force directed against the motion of the spacecraft and causes it to decelerate; that is, to slow down. The enormous kinetic energy the spacecraft possesses is converted into heat, and that heat can cause the spacecraft to reach extreme temperatures.

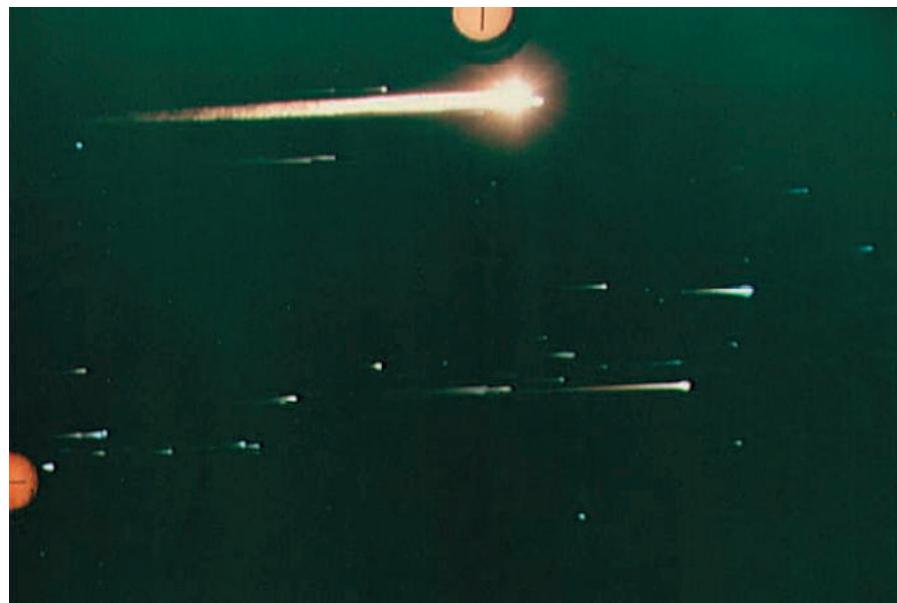


Figure 3.8 The Apollo 8 capsule re-entering the Earth's atmosphere

Research into the heat of re-entry was first conducted in the early 1950s in the USA, not for space flight but to build a durable warhead for intercontinental ballistic missiles (ICBMs). A typical ICBM reaches an altitude of 1400 km and has a range of 10 000 km. Computer design had produced streamlined missiles with long needle-shaped nose cones, but these designs reached temperatures of 7500° during re-entry — high enough to vaporise the nose cone.

In 1952, aeronautical engineer Harry Julian Allen, working with only pen and paper, calculated that the best shape for re-entry was a blunt one. When a blunt shape collides with the upper atmosphere at re-entry speeds, it produces a shockwave of compressed air in front of itself, much like the bow wave of a boat. Most of the heat is then generated in the compressing air, and significantly less heat is caused by friction of the air against the object itself.

Allen's discovery led to a new design of warhead — one that would detach from the rocket at altitude and re-enter the atmosphere backwards, presenting its blunt rear end as it fell. This was also the same basic design used for early space capsules such as the Mercury, Gemini and Apollo missions (see figure 3.7), as well as the planned Orion spacecraft. The first Mercury rocket, in particular, was simply a slightly modified ICBM, with the detachable nose cone adapted for occupation by a single short person. The space shuttle still uses this idea — by keeping its nose well up during re-entry, it presents a flat underbelly to the atmosphere to create the shockwave.

However, the blunt shape would still experience high temperatures, and a protective ‘heat shield’ was needed. After considerable research a technique called ablation was settled on. In this technique the nose cone is covered with a ceramic material, such as fibreglass, which is vaporised or ‘ablated’ during re-entry heating. The vaporising of the surface dissipates the heat and carries it away. This technique was used with success on all of the Mercury, Gemini and Apollo capsules during the 1950s, 1960s and 1970s; and will be used again on the Orion capsules. The Orion spacecraft is planned to take astronauts back to the Moon by 2020. As it returns it will re-enter Earth’s atmosphere with greater velocity than the space shuttle — about $40\,000 \text{ km h}^{-1}$. The ablative heat shield developed for this purpose is shown in figure 3.9. Called the TPS (Thermal Protection System), it is constructed of a modern material called Phenolic Impregnated Carbon Ablator (PICA).

The space shuttle’s slower orbital speed allows it to use a different approach. Each shuttle has a covering of insulating tiles. These tiles are made of glass fibre but are approximately 90% air. This gives them excellent thermal insulation properties and also conserves mass. Unfortunately they are also porous, so they absorb water, meaning that the surface must be waterproofed between each flight. Damage to the space shuttle *Columbia*’s heat shield is thought to have caused its disintegration and the loss of seven astronauts on 1 February 2003. Investigators believe the scorching air of re-entry penetrated a cracked panel on the left wing and melted the metal support structures inside. Events such as these have prompted the planned retirement of the space shuttle fleet in 2010.



Figure 3.9 Boeing have developed this prototype PICA heat shield for the new Orion spacecraft, which will return astronauts to Earth in a capsule much like the Apollo capsules, although on a larger scale.
Copyright © Boeing

Decelerating *g* forces

Even with a spacecraft designed to cope with re-entry heating, there is still another major consideration — the survival of any living occupants. Greater angles of re-entry mean greater rates of deceleration. This means a faster rate of heat build-up as kinetic energy is converted, but it also means greater *g* forces experienced by the occupants of the spacecraft. In the 1950s this was a very real concern, because designers were aware that the re-entry angles required by their spacecraft would generate loads up to 20 g . Unfortunately, early studies using large centrifuges suggested that *g* forces on astronauts should be restricted to 3 g if possible, and that 8 g represented a maximum safe load although symptoms such as chest pain and loss of consciousness could be experienced at this level.

Research was conducted into ways to increase a human’s tolerance of *g* forces. The findings included those below:

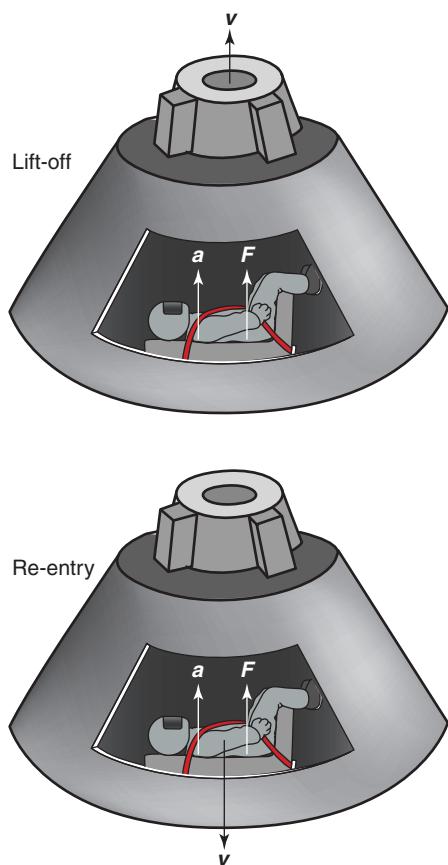


Figure 3.10 The application of g force upon an astronaut during lift-off and re-entry

- A transverse application of g load is easiest to cope with because blood is not forced away from the brain. This meant that the astronaut should be lying down at take-off, not standing or sitting vertically.
- An ‘eyeballs-in’ application of g loads is easier to tolerate than ‘eyeballs-out’. This meant that the astronaut should lift-off forwards (facing up) but re-enter backward (facing up) since the g forces are always directed upwards. This was consistent with the idea of the nose-cone presenting its rear face as it re-entered, as in figure 3.6.
- Supporting the body in as many places as possible increases tolerance. Supporting suits, webbing, netting and plaster casts were all tried, with designers eventually settling on a contoured couch, built of fibreglass and moulded to suit the body of a specific astronaut. Using this couch volunteers were successfully subjected to loads of up to 20 g , though this represented the limit of their tolerance.

In the *Mercury* rocket program that followed soon after this work, all of the above strategies were employed to the benefit of the astronauts. During his flight, Alan Shepard, the first American in space, was subjected to a maximum lift-off g force of 6.3 and a maximum re-entry g force of 11.6.

PHYSICS FACT

In 1959, a remarkable experiment was conducted by R. Flanagan Gray, a physician working at the US Navy’s centrifuge laboratory in Pennsylvania. The Navy was interested in water immersion as a means of body support to increase tolerance of high g forces. Gray designed a large aluminium capsule that could be fitted to the centrifuge and filled with water. It was nicknamed the ‘Iron Maiden’, which sounds ominous, but it was fitted with an emergency flushing mechanism. Gray tested it himself. Completely submerged inside the capsule, he held his breath as the centrifuge wound up, subjected him to a load of 31 g for 5 seconds, then wound down again. This established a new record for tolerance of g forces.

Ionisation blackout

An additional problem was discovered early in the development of space flight. As heat builds up around a spacecraft during re-entry, atoms in the air around it become ionised, forming a layer around the spacecraft. Radio signals cannot penetrate this layer of ionised particles, preventing communication between the ground and the spacecraft. All telemetry and verbal communication by radio is cut off for the duration of this **ionisation blackout**, the length of which depends upon the re-entry profile.

Apollo capsules would experience an ionisation blackout of three to four minutes, whereas the space shuttle suffers a somewhat longer period of 16 minutes.

Reaching the surface

Figure 3.7 (page 51) shows how a typical *Apollo* capsule, which would contain three astronauts, would reach an altitude of 400 000 ft or 120 km, considered the ‘entry interface’, at a re-entry angle between 5.2° and 7.2° . It would then descend from this altitude over a range of 2800 to 4600 km, continually slowing down and, at some point, suffering ionisation blackout. In the last portion of its descent, parachutes would be

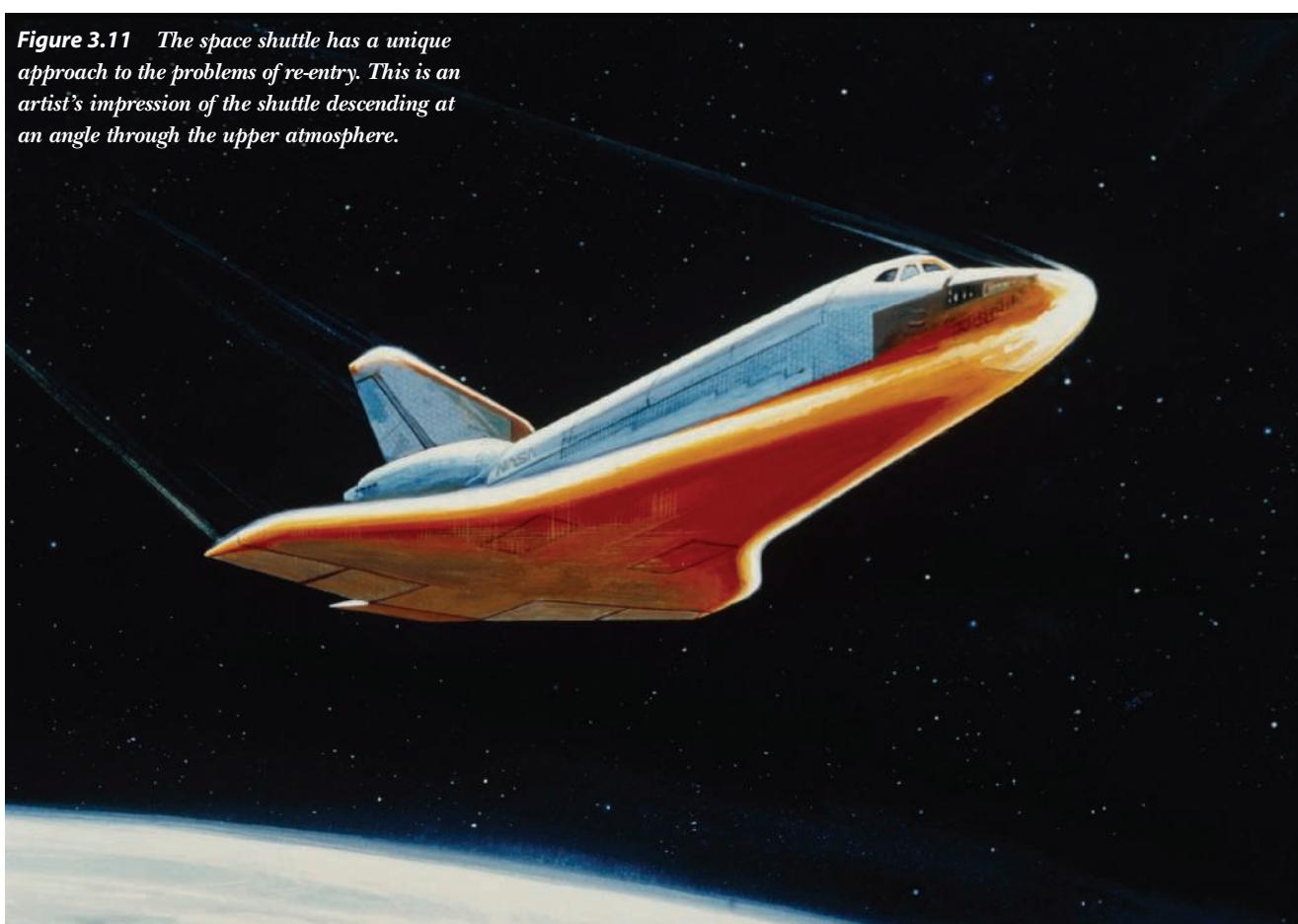
Ionisation blackout is a period of no communication with a spacecraft due to a surrounding layer of ionised atoms forming in the heat of re-entry.

released to slow it to about 33 km h^{-1} . Finally, it would splash down into the ocean to await recovery by a naval vessel. This was essentially the same strategy as that used by the earlier Mercury and Gemini spacecraft.

Soviet missions often ended a little differently, as they descended over land. The early Vostok capsules included an ejection seat, and the cosmonaut would eject at a suitable altitude, descending to the ground by parachute. Later Soyuz spacecraft, and the Chinese Shenzhou spacecraft that were based upon them, improved on this somewhat. With the occupants remaining inside, the capsules descended by a series of parachutes. In the last few moments the heat shield would be jettisoned, revealing a set of retrorockets. These were fired at a height of just 2 metres to provide a soft landing onto the ground.

As figure 3.11 shows, the space shuttle displays a unique solution to the problem of reaching the surface of the Earth without subjecting its occupants to loads greater than $3 g$. As it has wings, the pilot is able to control the attitude of the space shuttle and direct its descent. During the period of maximum deceleration and heat, its nose is held up at an angle of 40° , which slows its progress and presents the underbelly as a protected blunt surface. Past this stage, it is flown in a series of sharp-banking S-turns in order to control its descent, much like a snow skier descending a steep mountain. When it is just 1.5 km from the runway, it is gliding down an 18° gradient, much steeper than the 3° approach of a large airliner. When 500 m above the ground, speed brakes are applied (special flaps that increase drag) so that it settles in to a 1.5° final approach. The crew deploys the landing gear and within seconds the space shuttle touches down on its runway.

Figure 3.11 The space shuttle has a unique approach to the problems of re-entry. This is an artist's impression of the shuttle descending at an angle through the upper atmosphere.



SUMMARY

- Orbital motion is an example of uniform circular motion. Centripetal force is the force required to maintain circular motion and is given by the equation:

$$\text{Centripetal force, } F_C = \frac{mv^2}{r}.$$

- The period of a satellite's orbit is related to its radius by this expression of Kepler's Law of Periods:

$$\frac{r^3}{T^2} = \frac{GM}{4\pi^2}.$$

- The period of a satellite's orbit is related to its velocity by this expression:

$$T = \frac{2\pi r}{v}.$$

- Combining the above two equations leads to an expression for the orbital velocity of a satellite around the Earth:

$$v = \sqrt{\frac{Gm_E}{r}}$$

- Low Earth orbit refers to any orbits below an altitude of approximately 1000 km and will typically involve an orbital period of approximately one-and-a-half hours. A geostationary orbit is defined to be an orbit with an orbital period equal to that of the Earth, which places it permanently over a fixed point on the Earth's surface. This requires an altitude of approximately 35 800 km.
- Satellites in low Earth orbit are continually subjected to some small degree of atmospheric friction, which is influenced by many factors. This friction will eventually slow the satellite and decay its orbit.
- In order to re-enter the Earth's atmosphere, a spacecraft must deliberately lose velocity in such a way that it strikes the atmosphere at an optimum angle. If the angle is too shallow, the spacecraft may skip off the atmosphere and fail to re-enter. If the angle is too steep, the heat of re-entry and the g forces produced would be too great to ensure the survival of the spacecraft or its occupants.
- The heat of re-entry at its maximum produces a layer of ionised particles around a spacecraft that prevent radio communication with the spacecraft.

QUESTIONS

- Construct a diagram of a satellite orbiting the Earth, showing all forces acting on the satellite.
- A 400 g rock is tied to the end of a 2 m long string and whirled until it has a speed of 12.5 m s^{-1} . Calculate the centripetal force and acceleration experienced by the rock.
- A 900 kg motorcycle, travelling at 70 km h^{-1} , rounds a bend in the road with a radius of 17.5 m. Calculate the centripetal force required from the friction between the tyres and the road.
- (a) Define the orbital velocity of a satellite.
 (b) Describe its relationship to:
 - the gravitational constant G
 - the mass of the planet it is orbiting
 - the mass of the satellite
 - the radius of the orbit
 - the altitude of the satellite.
 (c) State this relationship in algebraic form.
- If a person were to hurl a rock of mass 0.25 kg horizontally from the top of Mt Everest, at an altitude of 8.8 km above sea level, calculate:
 - the velocity required by the rock so that it would orbit the Earth and return to the thrower, still waiting on Mt Everest. Note: The mass of the Earth is $5.97 \times 10^{24} \text{ kg}$, and its mean radius is 6380 km.
 - how long the thrower would have to wait for the rock's return.
 - the magnitude and direction of the centripetal force acting on the rock.
- Apollo rockets to the Moon would always begin their mission by launching into a low orbit with an approximate altitude of just 180 km. An orbit of this height can decay quite rapidly. However, they were never there for very long before boosting out of orbit towards the Moon. For their orbit calculate:
 - the required velocity in km h^{-1}
 - the period of the orbit
 - the acceleration of the spacecraft
 - the centripetal force acting on the spacecraft.
 Use 110 000 kg as the mass of the spacecraft in orbit. Additional data can be found in the previous question.
- Use Kepler's Law of Periods to calculate the time required for the following planets to complete an orbit of the Sun. Note that the radius of the Earth's orbit is $150 \times 10^6 \text{ km}$.

Planet	Radius of orbit (km)	Period of orbit (days)
Mars	228×10^6	687
Jupiter	5.2×10^8	12
Saturn	10.0×10^8	29
Uranus	19.2×10^8	84
Neptune	30.1×10^8	165

PLANET	RADIUS OF ORBIT ($\times 10^6$ km)	TIME TO ORBIT SUN (IN EARTH YEARS)
Mercury	58.5	
Venus	109	
Mars	229	
Jupiter	780	
Saturn	1430	

8. Compare, in words only, low Earth orbits and geostationary orbits.
9. Distinguish between a geostationary orbit and a geosynchronous orbit.
10. Calculate the altitude, period and velocity data to complete the following table.

TYPE OF ORBIT	ALTITUDE (km)	PERIOD (h)	VELOCITY ($m\ s^{-1}$)
Low Earth	360		
Geostationary		24	

11. Explain the limited lifetimes of low Earth orbits.
12. Describe the magnitude and direction of the g forces of re-entry into the Earth's atmosphere.
13. Construct a list of some of the dangers of atmospheric re-entry with a short discussion of each. Your list should include at least four items.



3.1 INVESTIGATING CIRCULAR MOTION

Aim

To examine some of the factors affecting the motion of an object undergoing uniform circular motion, and then to determine the quantitative relationship between the variables of force, velocity and radius.

Apparatus

rubber stopper
glass tube
50 g mass carrier
stopwatch
string
sticky tape
metre rule
50 g slot masses

Theory

Recall that the expression for the centripetal force that causes circular motion is as follows:

$$F_C = \frac{mv^2}{r}$$

where

F_C = centripetal force (N)

m = mass of object in motion (kg)

v = instantaneous velocity of mass (m s^{-1})

r = radius of circular motion (m).

In this experiment, the centripetal force is provided by the weight of some masses hanging on a mass carrier, so that here:

$$F_C = m_C g$$

where

m_C = mass of mass carrier + masses (kg)

g = acceleration due to gravity = 9.8 m s^{-2} .

You should also recall the relationship between the period of an object in circular motion and its orbital velocity:

$$T = \frac{2\pi r}{v}$$

where

T = period (s)

r = radius of motion (m)

v = orbital velocity (m s^{-1}).

Method

- Record the mass of the rubber stopper being used as a bob.
- Attach the rubber stopper to a length of string approximately 1.5 m long, then thread the loose end of the string through the glass tube.
- Attach the mass carrier to the loose end of the string as shown in figure 3.12.

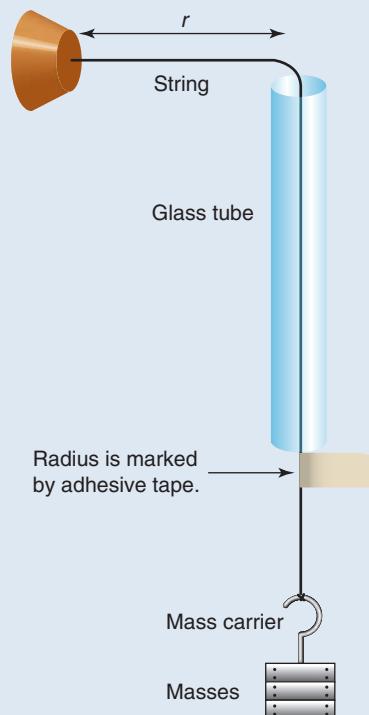


Figure 3.12

- Place a piece of sticky tape on the string at the point shown in figure 3.12 so that the distance, r , is 40 cm.
- Hold the glass tube and move it in a small circle so as to get the rubber bob moving in a circular path. The mass carrier will provide the centripetal force to keep the bob moving in its circular path. Adjust your frequency of rotation so that the sticky tape just touches the bottom of the glass tube. This will keep the radius of the bob's orbit steady.
- Record the time for the bob to complete 10 revolutions at a constant speed then calculate and record the period. Do this three times and then use the average of these as the correct period. Use the radius and period to calculate the orbital velocity of the bob.
- Repeat steps 4 to 6 for radii of 0.60 m, 0.80 m, and 1.0 m.
- Repeat steps 3 to 7 using masses of 100 g, then 150 g, and finally 200 g.

Results

Tabulate your results as shown in the table below.

FORCE	RADIUS	PERIOD (10 REVOLUTIONS)		MEAN PERIOD	ORBITAL VELOCITY v
$50\text{ g} \times$ gravity	1.0 m				
	0.8 m				
	0.6 m				
	0.4 m				
$100\text{ g} \times$ gravity	1.0 m				
	0.8 m				
	0.6 m				
	0.4 m				

Analysis

- From the results above, calculate the orbital velocity of the bob and complete the table.
- For each of the radii used with 50 g, construct a graph of v^2 versus r .
- Repeat this for the 100 g, 150 g and 200 g results.

Questions

- What is the relationship that these graphs indicate?
- What does the slope of your v^2 versus r graph represent?
- What role does gravity play in the results in this experiment?

CHAPTER 4 GRAVITY IN THE SOLAR SYSTEM



Figure 4.1 The orbit of the Moon around the Earth is determined by the gravitational forces that the two bodies exert on each other.

Remember

Before beginning this chapter, you should be able to:

- recall Kepler's Law of Periods and be able to use it to solve problems and analyse information:
$$\frac{r^3}{T^2} = \frac{GM}{4\pi^2}$$
- calculate the centripetal force acting on a satellite orbiting the Earth in uniform circular motion:

$$F = \frac{mv^2}{r}$$

- state the definition of momentum: $p = mv$
- state the definition of kinetic energy: $E_k = \frac{1}{2}(mv)^2$
- use the conservation of momentum and kinetic energy to analyse one-dimensional elastic collisions.

Key content

At the end of this chapter you should be able to:

- define Newton's Law of Universal Gravitation
- describe the gravitational field in the region surrounding a massive object such as a planet
- discuss the importance of Newton's Law of Universal Gravitation to an understanding of the orbital motion of satellites
- identify the slingshot effect as used by space probes.

Our solar system consists of eight planets and currently three known (but potentially many) dwarf planets performing orbital motion around the Sun. Most of those nine planets have smaller satellites performing orbital motion around them. In chapter 3, we looked at circular motion as the simplest model of orbital motion. We learned that this type of motion requires a centripetal force, that is, a force that acts continually on the orbiting body and is always directed towards the central body. We also noted that for planets and satellites, the force of gravity provides the centripetal force.

In this chapter we further examine the force of gravity, and also the nature of gravitational fields, within the solar system.

4.1

THE LAW OF UNIVERSAL GRAVITATION



Figure 4.2 Isaac Newton

You will recall from your work in the Preliminary Course that it was Isaac Newton who was first able to provide a formula to describe the manner in which gravity acts. This formula, now known as the *Law of Universal Gravitation*, is as follows:

$$F = G \frac{m_1 m_2}{r^2}$$

where

F = force of gravity between two masses (N)

G = universal gravitation constant

$= 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$

m_1, m_2 = the two masses involved (kg)

r = the distance between their centres of mass (m).

Note that this is always an attractive force and is exerted equally on both masses. It depends only upon the value of the two masses and their separation distance. Note also that the force is inversely proportional to the square of the distance. Hence, in any given situation, if the distance were to double, the value of the force would drop to one-quarter of its previous value.

SAMPLE PROBLEM

4.1

Determining gravitational forces in the Sun–Earth–Moon system

Given the following data, determine the magnitude of the gravitational attraction between:

- (a) the Earth and the Moon
- (b) the Earth and the Sun.

Mass of the Earth = 5.97×10^{24} kg

Mass of the Moon = 7.35×10^{22} kg

Mass of the Sun = 1.99×10^{30} kg

Earth–Moon distance = 3.84×10^8 m on average

Earth–Sun distance = 1.50×10^{11} m on average (one astronomical unit, AU)

SOLUTION

(a)

$$F = G \frac{m_E m_M}{r^2}$$

$$= \frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})(7.35 \times 10^{22})}{(3.84 \times 10^8)^2}$$

$$= 1.98 \times 10^{20} \text{ N}$$

That is, the magnitude of the gravitational force of attraction between the Earth and the Moon is approximately 1.98×10^{20} N.

$$\begin{aligned}
 \text{(b)} \quad F &= G \frac{m_E m_S}{r^2} \\
 &= \frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})(1.99 \times 10^{30})}{(1.50 \times 10^{11})^2} \\
 &= 3.52 \times 10^{22} \text{ N}
 \end{aligned}$$

That is, the magnitude of the gravitational force of attraction between the Earth and the Sun is approximately 3.52×10^{22} N, or about 180 times greater than the Earth–Moon attraction.

SAMPLE PROBLEM

4.2

Determining smaller gravitational forces

Determine the gravitational force of attraction between two 250 g apples that are lying on a desk 1.0 m apart.

SOLUTION

$$\begin{aligned}
 F &= G \frac{m_{\text{apple1}} m_{\text{apple2}}}{r^2} \\
 &= \frac{(6.67 \times 10^{-11})(2.5 \times 10^{-3})(2.5 \times 10^{-3})}{(1.0)^2} \\
 &= 4.2 \times 10^{-12} \text{ N}
 \end{aligned}$$

Comparing this tiny force with those forces calculated in sample problem 4.1, we can see that gravitation requires huge masses to produce significant forces.

Universal gravitation and the motion of satellites

In chapter 3, we learned that the force of gravity serves as the centripetal force that maintains the orbital motion of a satellite or planet, and we assumed this motion to be uniform circular motion. (Most orbits are ellipses with slight eccentricities that make them almost circular, so this assumption is a fair one that simplifies the situation.) This force holds the solar system together, tying the planets to the Sun and the satellites to their planets, keeping them on an orbital path and determining the speed of their rotations.

In chapter 3, we also derived an equation for the orbital velocity of a satellite from Kepler's Law of Periods. In order to emphasise the importance of gravity to the motion of satellites and planets, that same equation can be derived from Newton's expression for universal gravitation by equating it to the expression for centripetal force. The following focuses on the specific case of a satellite orbiting the Earth, but the equations are equally applicable to satellites orbiting other planets or planets orbiting the Sun.

First, the gravitational attraction between a satellite and the Earth would be given by the following expression:

$$F_G = G \frac{m_E m_S}{r^2}$$

where

$$\begin{aligned}
 m_E &= \text{mass of the Earth} \\
 &= 5.97 \times 10^{24} \text{ kg}
 \end{aligned}$$

$$m_S = \text{mass of the satellite (kg)}$$

$$r = \text{radius of the orbit (m)}$$

$$G = \text{universal gravitation constant.}$$

This gravitational force of attraction also serves as the centripetal force for the circular orbital motion, hence:

$$F_G = F_C.$$

Therefore, we can equate the formula for F_G with that for F_C :

$$G \frac{m_E m_S}{r^2} = \frac{m_S v^2}{r}$$

$$\therefore v = \sqrt{\frac{G m_E}{r}}$$

where

$$v = \text{orbital velocity (m s}^{-1}\text{)}.$$

Note: The radius of the orbit, r , is the sum of the radius of the Earth and the altitude of the orbit.

$$r = r_E + \text{altitude (m)}$$

where

$$r = \text{radius of orbit (m)}$$

$$\begin{aligned} r_E &= \text{radius of the Earth (m)} \\ &= 6.38 \times 10^6 \text{ m} \end{aligned}$$

$$\text{altitude} = \text{height of orbit above the ground (m)}.$$

Again, we may note that the orbital velocity required for a particular orbit depends only on the mass and radius of the Earth (or other central body) and the altitude of the orbit. Further, the greater the radius of the orbit, the slower the orbital velocity that is required to maintain the orbit.

SAMPLE PROBLEM

4.3

Determining orbital velocity

Calculate the orbital velocity of the Earth along its orbit around the Sun, and compare this to the value determined by considering the period of its motion.

SOLUTION

First, the orbital velocity of the Earth can be calculated directly:

$$\begin{aligned} v &= \sqrt{\frac{G m_S}{r}} \\ &= \sqrt{\frac{(6.67 \times 10^{-11})(1.99 \times 10^{30})}{(1.50 \times 10^{11})}} \\ &= 29\,700 \text{ m s}^{-1} = 29.7 \text{ km s}^{-1} \end{aligned}$$

Orbital velocity can also be calculated from the known period and radius of the Earth's orbit:

$$\begin{aligned} v &= \frac{2\pi r}{T} \\ &= \frac{2\pi(1.50 \times 10^{11})}{(365.26 \times 24 \times 3600)} \\ &= 29\,900 \text{ m s}^{-1} = 29.9 \text{ km s}^{-1} \end{aligned}$$

These two answers compare very well, and represent the same value after allowing for rounding errors.

SAMPLE PROBLEM**4.4****Comparing centripetal and gravitational forces acting on the Moon**

The period of the Moon's orbit around the Earth is 27.26 days. Use this information, along with the data in sample problem 4.1 (see pages 61–2), to calculate the orbital velocity of the Moon. Use this answer to then calculate the centripetal force acting on the Moon, assuming its orbit to be circular.

Finally, compare the calculated value for centripetal force to the gravitational force between the Earth and the Moon calculated in sample problem 4.1.

SOLUTION

$$\begin{aligned}\text{Orbital velocity } v &= \frac{2\pi r}{T} \\ &= \frac{2\pi(3.84 \times 10^8)}{(27.26 \times 24 \times 3600)} \\ &= 1024 \text{ m s}^{-1} \\ \text{Centripetal force } F_c &= \frac{m_{\text{Moon}} v^2}{r} \\ &= \frac{(7.35 \times 10^{22})(1024)^2}{3.84 \times 10^8} \\ &= 2.01 \times 10^{20} \text{ N}\end{aligned}$$

Referring back to sample problem 4.1 on pages 61–2, the value for the gravitational force between the Earth and the Moon was calculated to be 1.98×10^{20} N, which compares very closely to this result allowing for rounding error.

SAMPLE PROBLEM**4.5****Comparing centripetal and gravitational forces acting on the Earth**

Repeat sample problem 4.4, but this time focusing on the Earth in its orbit of the Sun.

SOLUTION

$$\begin{aligned}\text{Orbital velocity } v &= \frac{2\pi r}{T} \\ &= \frac{2\pi(1.50 \times 10^{11})}{(365.25 \times 24 \times 3600)} \\ &= 29\,900 \text{ m s}^{-1} \\ \text{Centripetal force } F_c &= \frac{m_{\text{Earth}} v^2}{r} \\ &= \frac{(5.97 \times 10^{24})(29\,900)^2}{1.50 \times 10^{11}} \\ &= 3.56 \times 10^{22} \text{ N}\end{aligned}$$

Referring back to sample problem 4.1, the value for the gravitational force between the Sun and the Earth was calculated to be 3.52×10^{22} N. Once again, the closeness of these values demonstrates that gravity functions as the centripetal force for the orbital motion of satellites and planets.

System mass or central mass?
 The derivation shown here uses the case of a small mass performing circular motion around a much larger mass. However, this is not what actually happens. Even assuming that the orbits are circular, the two bodies involved each orbit the common centre of mass of the two-body system. The distance of each mass from the centre of its orbit is thus not equal to the separation distance of the masses. This means that the distance variables in the gravitation and centripetal force expressions are not the same. When this difference is allowed for, the variable M in Kepler's Law of Periods becomes the mass of the system, rather than just the central mass. However, in the case of a satellite, or even a planet, this difference is insignificant.

The full derivation of this expression can be found in 'Astrophysics', chapter 16, on page 307, where it is applied to the situation of two stars orbiting each other.

Deriving the constant in Kepler's Law of Periods

In chapter 3, we learned that Kepler had originally stated his Law of Periods in the form $\frac{r^3}{T^2} = k$, but he was not able to determine an

expression for the constant k . When Isaac Newton was devising his Law of Universal Gravitation, he found that he was able to derive such an expression. The derivation begins by equating the gravitation and centripetal forces to give an equation for orbital velocity, just as we have already done.

$$G \frac{mM}{r^2} = \frac{mv^2}{r}$$

$$\therefore v = \sqrt{\frac{GM}{r}} \text{ where } M \text{ is the large central mass.}$$

The expression relating period to orbital velocity $v = \frac{2\pi r}{T}$ is then substituted into the equation, which can then be rearranged to the form of Kepler's Law of Periods.

$$\frac{2\pi r}{T} = \sqrt{\frac{GM}{r}}$$

$$\therefore \frac{r^3}{T^2} = \frac{GM}{4\pi^2} = \text{the constant } k$$

4.2 GRAVITATIONAL FIELDS

A **vector** is any quantity that has both magnitude and direction. Force is one example.

A **gravitational field** is a field within which any mass will experience a gravitational force. The field has both strength and direction.

A vector field can be said to exist in any space within which a force vector can act. You will recall that a **vector** is any quantity that has both magnitude and direction. Force is one example. Corresponding to the force vector that acts within it, a vector field also has strength and direction. Some examples of vector fields are magnetic fields, electric fields and gravitational fields.

A **gravitational field** is a field within which any mass will experience a gravitational force. Since the force of gravity will act on any mass near the Earth, a gravitational field exists around the Earth, and can be drawn as shown in figure 4.3. Note that on a small scale, such as the interior of a room, the field lines — or lines of force — appear parallel and point down since this is the direction of the force that would be experienced by a mass placed within the field.

The gravitational field of a planet or star extends some distance from it. Figure 4.3 shows the shape of the gravitational field around the Earth. We can see now that the field has a radial pattern with the field lines pointing towards the Earth's centre, again because this is the direction of the force that would be experienced by a mass within the field. Note that closer to the Earth, the field lines are closer together. This indicates that the field, and its force, are stronger in this region.

In chapter 1, we learned that the field vector g also describes the strength and direction of a gravitational field at some point within that field. Of course, g is more commonly known as 'acceleration due to gravity' and has a value of 9.8 m s^{-2} at the Earth's surface.

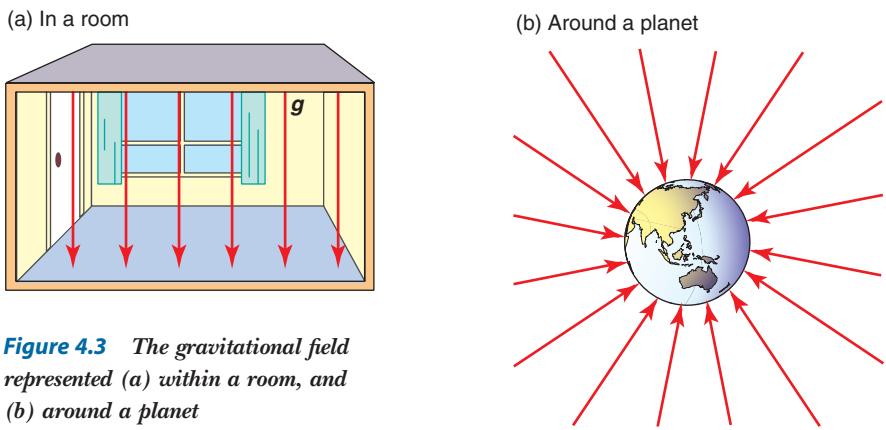


Figure 4.3 The gravitational field represented (a) within a room, and (b) around a planet

Of course, any large object near the Earth, such as the Moon, will have a gravitational field of its own, and the two fields will combine to form a more complex field, such as that shown in figure 4.4. Note that there is a point between the two, but somewhat closer to the Moon, at which the strength of the field is zero. In other words, the gravitational attraction of the Earth and that of the Moon are precisely equal but opposite in direction. Such points exist between any two masses, but become noticeable when considering planets and stars that are close enough to be gravitationally bound together.

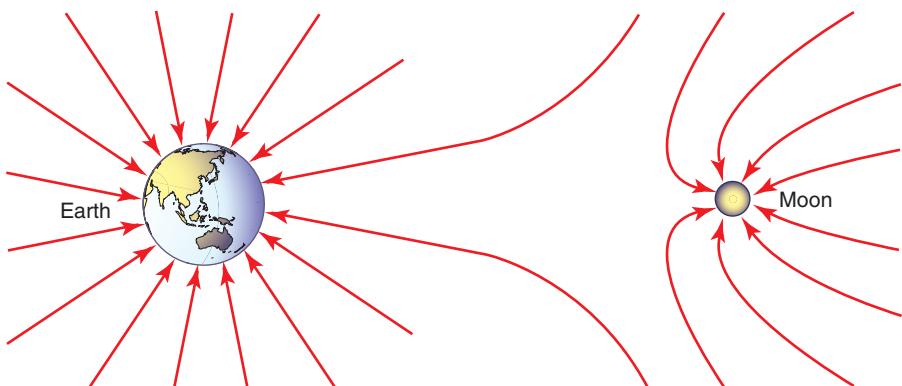


Figure 4.4 The gravitational field around the Earth and Moon. The overall shape depends upon the relative strengths of the two fields involved.

4.3 THE SLINGSHOT EFFECT

When the *Mariner 10* spacecraft reached Venus on 5 February 1974, it conducted a very successful survey of the planet, but that wasn't the end of its mission. Its flight path was shifted, aiming it closer in towards the planet, flinging it around and accelerating it towards Mercury. This was the first mission to use the **slingshot effect**, also known as a planetary swing-by or gravity-assist manoeuvre, to pick up speed and proceed on to another target. The *Mariner 10* arrived at Mercury a little under two months later, flying just 705 km above the surface of the planet.

During a slingshot, a spacecraft deliberately passes close to a large mass, such as a planet, so that the mass's gravity pulls the spacecraft in toward it. This causes the spacecraft to accelerate, and it heads around the planet and departs in a different direction. The departure speed of the spacecraft relative to the planet is the same as the approach speed relative to the planet, but the change in direction can result in a real change in velocity relative to the Sun. In general, a spacecraft will approach a planet at an angle to the planet's orbital path. By swinging

The **slingshot effect**, or planetary swing-by or gravity-assist manoeuvre, is a strategy used with space probes to achieve a change in velocity with little expenditure of fuel.

behind the planet an increase in speed can be achieved, and by swinging in front of the planet's path a decrease in speed is achieved. These trajectories are shown in figure 4.5. The major benefit of the slingshot effect is that the change in velocity is achieved with very little expenditure of fuel.

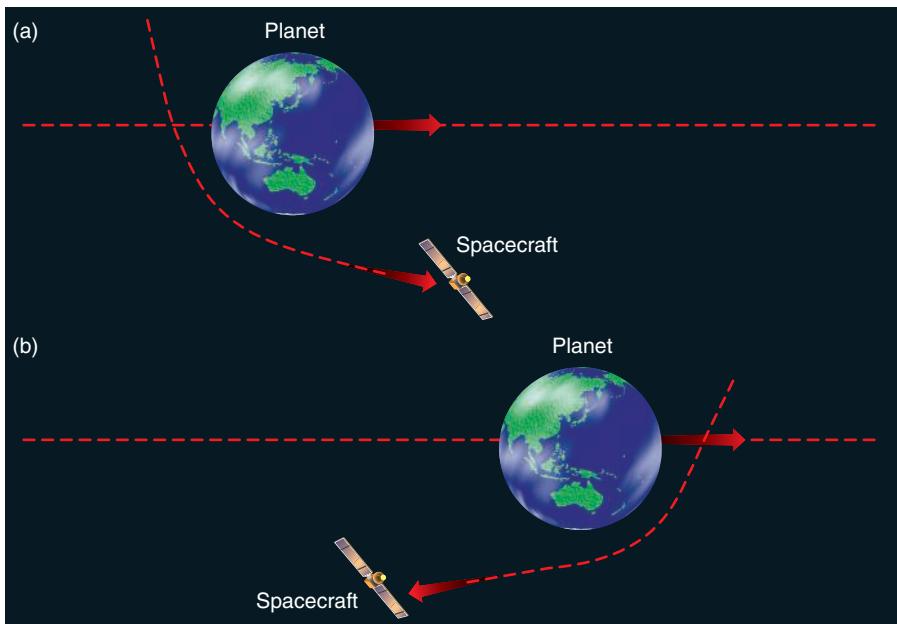


Figure 4.5 (a) Swinging behind a planet's path results in an increase in speed relative to the Sun.
(b) Swinging in front of a planet's path results in a decrease in speed.

In order to understand the slingshot effect, we need to analyse it as if it were a perfectly elastic one-dimensional collision. Even though there is no contact, the interaction behaves as a collision. However, because the bodies do not touch in any way, there are no energy losses and, hence, the 'collision' is elastic. Figure 4.6 represents the situation in this way. Note the variables we are using: v_i is the initial velocity of the spacecraft, V_i is the initial velocity of the planet, m is the mass of the spacecraft and $K \times m$ is the mass of the planet, where K has a very large value. Note: After the 'collision' there are two unknowns — the final velocity of the planet, V_f , and that of the spacecraft, v_f .

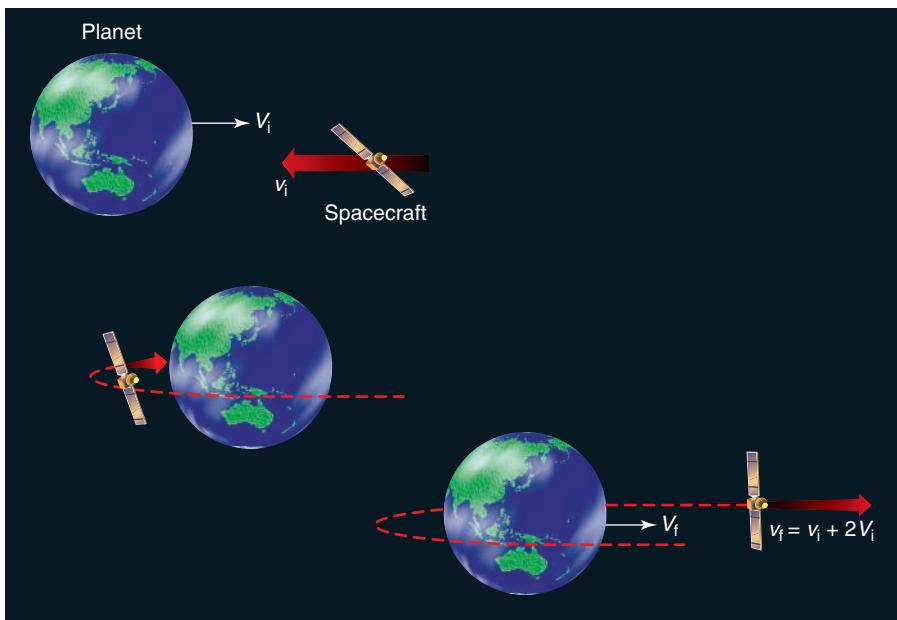


Figure 4.6 The velocity gained by the slingshot effect

Applying the conservation of momentum to this collision:

$$\begin{aligned}
 & \text{Initial momentum} = \text{final momentum} \\
 p_i \text{ of planet} + p_i \text{ of spacecraft} &= p_f \text{ of planet} + p_f \text{ of spacecraft} \\
 KmV_i + m(-v_i) &= KmV_f + mv_f \\
 KV_i - v_i &= KV_f + v_f
 \end{aligned} \tag{1}$$

Applying now the conservation of kinetic energy:

$$\begin{aligned}
 & \text{Initial kinetic energy} = \text{final kinetic energy} \\
 E_{ki} \text{ of planet} + E_{ki} \text{ of spacecraft} &= E_{kf} \text{ of planet} + E_{kf} \text{ of spacecraft} \\
 \frac{1}{2} KmV_i^2 + \frac{1}{2} m(-v_i)^2 &= \frac{1}{2} KmV_f^2 + \frac{1}{2} mv_f^2 \\
 KV_i^2 + v_i^2 &= KV_f^2 + v_f^2
 \end{aligned} \tag{2}$$

Solving equations [1] and [2] simultaneously for v_f leads to the following expression:

$$v_f = v_i + 2V_i$$

where

$$\begin{aligned}
 v_f &= \text{maximum exit velocity of spacecraft (m s}^{-1}\text{)} \\
 v_i &= \text{entry velocity of spacecraft (m s}^{-1}\text{)} \\
 V_i &= \text{velocity of planet (m s}^{-1}\text{).}
 \end{aligned}$$

This expression represents the maximum velocity achievable from the slingshot effect. This is achieved by a head-on rendezvous as described; at other angles, a spacecraft achieves lower velocities. The final velocity of the planet is marginally less than its initial velocity. Note that the kinetic energy of the system has been conserved, with the spacecraft gaining some kinetic energy and the planet losing an equivalent amount.

As an example, let us now consider a spacecraft approaching Jupiter almost head-on for a slingshot manoeuvre. Jupiter has an orbital velocity of about $13\,000 \text{ m s}^{-1}$ relative to the Sun, and let us assume that the spacecraft has a velocity of $15\,000 \text{ m s}^{-1}$ relative to the Sun. Using the equation above, the exit velocity of the spacecraft will be $(15\,000 + 2 \times 13\,000)$ or $41\,000 \text{ m s}^{-1}$.

In order to visualise how this effect occurs, it is helpful to consider the relative velocities. If we consider the velocity of the spacecraft relative to Jupiter then the situation is as if Jupiter were standing still. When a small object collides elastically with a very large, massive object, the small object will rebound without loss of speed — consider a ball bouncing against a wall. Similarly, our spacecraft approaches Jupiter at $(13\,000 + 15\,000)$ or $28\,000 \text{ m s}^{-1}$ relative to it, swings around behind it and then slingshots back out in front at $28\,000 \text{ m s}^{-1}$ relative to Jupiter.

Look now at what has happened to the velocity of the spacecraft relative to the Sun. If Jupiter's velocity is $13\,000 \text{ m s}^{-1}$ and the spacecraft is moving ahead $28\,000 \text{ m s}^{-1}$ faster than it, the spacecraft is now travelling at $41\,000 \text{ m s}^{-1}$ relative to the Sun.

A planetary swing-by of Jupiter very similar to that just described was performed by the space probe *Ulysses* in February 1992, as shown in figure 4.7. *Ulysses* used the swing-by to accelerate it out of the plane of the solar system but still in an orbit of the Sun. This allowed it to be the first spacecraft ever to explore this region of interplanetary space.

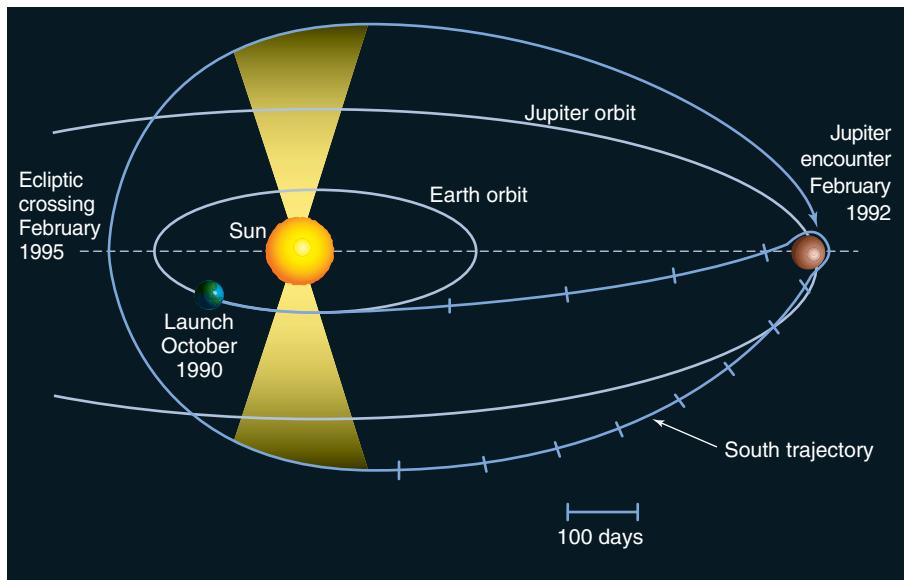


Figure 4.7 Ulysses used a slingshot around Jupiter to accelerate it out of the plane of the solar system.

Some slingshot manoeuvres have become quite complex and sophisticated. A case in point is the trajectory of the *International Sun–Earth Explorer 3*, or ISEE-3 space probe. It was initially placed into an orbit about the Langrangian L1 point between the Earth and the Sun. Here, it performed investigations on the solar wind and its connection with the Earth's magnetic field. In June 1982, it was removed from there and sent through a series of slingshots around the Moon and the Earth as shown in figure 4.8. The end result of this complicated series of manoeuvres was the probe heading into an orbit around the Sun and aiming for Comet Giacobini-Zinner. At this point the probe was renamed the *International Cometary Explorer* (ICE). It met up with Comet Giacobini-Zinner on 11 September 1985, making many important measurements of its plasma tail and was then able to travel on and meet up with Comet Halley in March 1986, making it the first spacecraft to investigate two comets in this way.

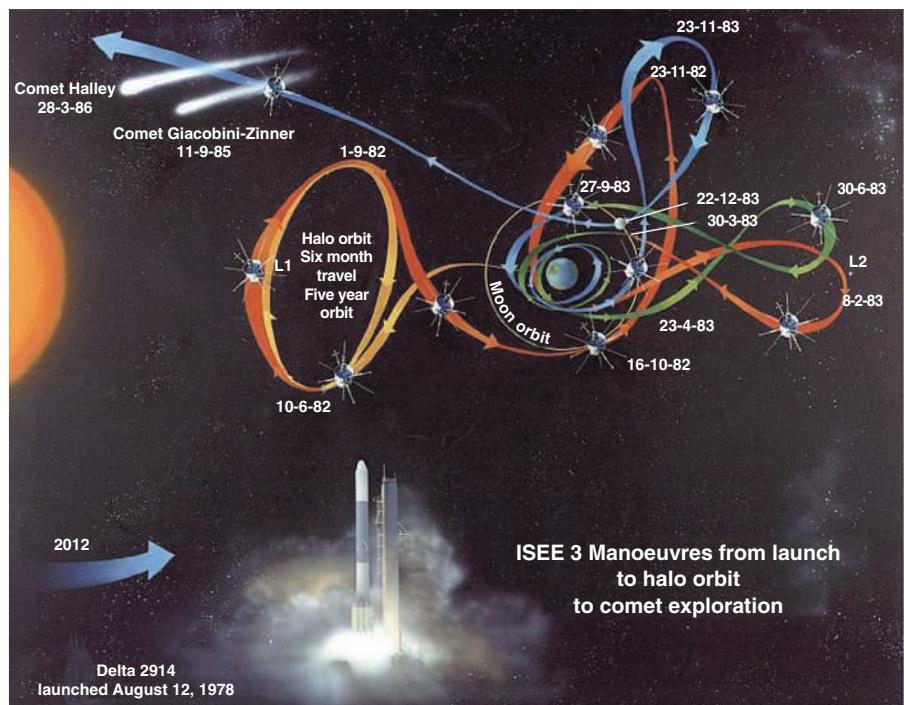


Figure 4.8 The trajectory of the ISEE-3/ICE space probe

SUMMARY

- The Law of Universal Gravitation was proposed by Isaac Newton to describe the force of attraction between any two masses. It is described by this equation:

$$F = G \frac{m_1 m_2}{r^2}$$

- Gravity acts as the centripetal force for the orbital motion of satellites and planets.
- Equating the formulas for universal gravitation and centripetal force allows the derivation of the orbital velocity formula.
- A gravitational field is a vector field surrounding any mass within which another mass will experience a gravitational force. Gravity is such a weak force that it takes a massive object (such as a planet) to create a significant gravitational field.
- The slingshot effect is a manoeuvre in which a spacecraft uses the gravity of a planet to make a change in velocity with very little expenditure of fuel.

QUESTIONS

- Define Newton's Law of Universal Gravitation.
- (a) List the variables upon which the force of gravity between two masses depends.
(b) Note that time is not one of the variables, implying that this force acts instantaneously throughout space, faster even than light. Does this seem reasonable to you? Explain your answer.
- State what would happen to the strength of the force of attraction between two masses if the distance between them was to halve and the masses themselves were each to double.
- Given the following data, calculate the magnitude of the gravitational attraction between:
(a) Jupiter and its Moon Callisto
(b) Jupiter and the Sun.
Mass of Jupiter = 1.9×10^{27} kg
Mass of Callisto = 1.1×10^{23} kg
Mass of the Sun = 1.99×10^{30} kg
Jupiter–Callisto distance = 1.88×10^9 m on average
Jupiter–Sun distance = 7.78×10^{11} m on average

- Calculate the magnitude of the force of gravitation between a book of mass 1 kg and a pen of mass 50 g lying just 15 cm from the centre of mass of the book.

- Calculate the force of gravity between a 72.5 kg astronaut and the Earth (mass = 5.97×10^{24} kg and radius 6.378×10^6 m), using the Law of Universal Gravitation:
(a) standing on the ground prior to launch
(b) at an altitude of 285 km after launch.

- For each of the orbiting bodies shown in the following table, calculate the orbital velocity from the period then use it to calculate the centripetal force. Also calculate the value of the gravitational force acting on the body and indicate how well the two forces compare.

ORBITING BODY	ORBITAL PERIOD	CENTRAL BODY	ORBITAL RADIUS
Satellite in low Earth orbit $m = 1360$ kg	90.6 minutes	Earth $m = 5.97 \times 10^{24}$ kg	Altitude = 300 km $\therefore r = 6.68 \times 10^6$ m
Venus $m = 4.9 \times 10^{24}$ kg	225 Earth days	The Sun $m = 1.99 \times 10^{30}$ kg	1.09×10^{11} m
Callisto $m = 1.1 \times 10^{23}$ kg	16.7 Earth days	Jupiter $m = 1.90 \times 10^{27}$ kg	1.88×10^9 m

- (a) Discuss the reasons why Newton's Law of Universal Gravitation is important to an understanding of the motions of satellites.
(b) Describe how the use of this law allows calculation of the motion of satellites.
- (a) State the nature of the slingshot effect or planetary swing-by.
(b) This is also known as a 'gravity-assist manoeuvre'. Briefly describe the role of gravity in this manoeuvre.
(c) State the laws of physics underlying this effect.
(d) Identify the benefits achieved by this manoeuvre.

CHAPTER

5

SPACE AND TIME



Figure 5.1 Albert Einstein in 1905

Remember

Before beginning this chapter, you should be able to:

- describe Newton's First Law of Motion
- apply the definition of velocity: $v = \frac{\Delta r}{t}$.

Key content

At the end of this chapter you should be able to:

- outline the features of the aether model for light transmission
- describe and evaluate the Michelson–Morley experiment
- discuss the role of experiments in the scientific method, with reference to the Michelson–Morley experiments and the use of thought experiments in relativity
- outline the nature of inertial frames of reference
- discuss the principle of relativity
- recognise the constancy of the speed of light and discuss the implications for theories of space and time
- discuss the definition of the metre in terms of time and the speed of light
- discuss qualitatively and quantitatively the following consequences of relativity: the relativity of simultaneity, time dilation, length contraction, mass dilation and the equivalence of mass and energy
- discuss the implications of time dilation, length contraction and mass dilation for space travel.

Much of the following physics of relativity sprang out of considerations of light — of what form it takes, how it moves from one place to another and how fast. For many centuries physicists argued over whether light takes the form of a shower of tiny particles, like buckshot from a shotgun, or whether it is in the form of a wave motion like sound waves. Then, in 1801, Thomas Young performed an experiment that showed that light rays could interfere with each other to produce a pattern, which is a property unique to waves. The early nineteenth century saw a series of further demonstrations of the wave nature of light.

In 1864, James Clerk Maxwell seemed to put the issue beyond doubt when he produced a brilliant set of equations to explain the behaviour of electric and magnetic fields, and then used the equations to show that these fields could move together as waves through space at the speed of light. Additionally, he proposed that light was also a form of these electromagnetic waves. The German scientists Helmholtz and Hertz were able, in 1887, to produce experimental evidence for the existence of these waves.

5.1 THE AETHER MODEL

Having concluded that light moves as a waveform, nineteenth-century physicists turned to other wave motions in order to better understand light. There were many others known, including sound waves, water waves, and earthquake waves. All of these waveforms need a medium through which to travel, and so it was believed that light waves would also require a medium. Nobody could find such a medium but belief in its existence was so strong that it was given a name, the ‘luminiferous **aether**’, and its properties were identified. The aether:

- filled all of space, had low density and was perfectly transparent
- permeated all matter and yet was completely permeable to material objects
- had great elasticity to support and propagate the light waves.

This list of properties may seem odd to us now and the whole concept of the aether may seem strange in hindsight, but bear in mind that nineteenth-century physicists were trying to understand a phenomenon completely unknown to them. It is not unlike the situation facing modern cosmologists in trying to understand why the universe seems to have much more matter than can be observed, and why the expansion of the universe seems to be accelerating. Some explanations of these modern-day puzzles attribute some similarly unusual properties to otherwise ‘ordinary’ space.

The search for the aether was to occupy physicists for several decades before it was finally accepted that (a) the aether does not actually exist, and (b) electromagnetic waves (including light) are unique in that they do not require a **medium** of any sort in order to move.

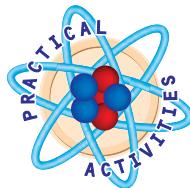
The Michelson–Morley experiment

If you were in a boat, how could you tell whether the boat was moving? You might simply look over the side to see if water is flowing past. If you wanted to be certain, you might dip your hand in to feel the flowing water.

Similarly, if the aether did exist, our Earth, moving through space at about 30 km s^{-1} as it orbits the Sun, should be moving through the aether. From our point of view, we should experience a flow of aether past us or, as it became known, an ‘aether wind’. However, the aether was thought to

The **aether** was the proposed medium for light and other electromagnetic waves, before it was realised that these waveforms do not need a medium in order to travel.

A **medium** is the material through which a wave travels.



5.1

Modelling the Michelson–Morley experiment

be extremely tenuous, so any aether wind would be hard to detect. There were many experiments designed and performed to detect it, but they all failed. The assumption was that the detection mechanisms were simply not sensitive enough.

The definitive experiment to detect the aether wind was performed by A. A. Michelson and E. W. Morley in 1887, for which they received the Nobel Prize in 1907. It was exceedingly sensitive.

In order to understand how the experiment worked, consider the analogy shown in figure 5.2. Two identical speedboats are going to have a race on a river, over two different courses. Boat A will head upstream for 2 km before turning around and heading back. Boat B will head directly across the 2 km-wide river before returning. Each boat is capable of a boat speed of 5 km h^{-1} and each completes a 4 km circuit. However, the current in the river affects the velocity of each boat differently, as boat A heads directly along the current while boat B heads across it. As figure 5.3 shows, each boat has a different effective velocity (relative to the bank) and boat A takes 15 minutes longer to complete the course.

Figure 5.2 Two identical speedboats race over different 4 km courses, one against and then back along the current, and the other across the current.

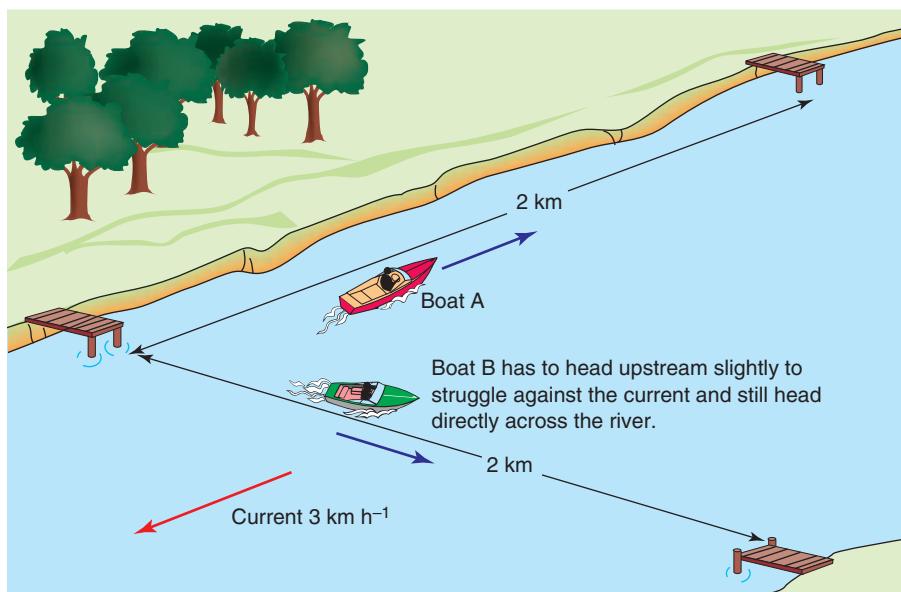


Figure 5.3 The current affects each boat differently, causing the boat that heads across the current to win every time.

(a) The situation for boat A heading along the river

Journey out (against current)

Boat speed = 5 km h^{-1} through the water
= 2 km h^{-1} as seen from the river bank

$$\therefore \text{time taken} = \frac{\text{distance}}{\text{speed}} = \frac{2 \text{ km}}{2 \text{ km h}^{-1}} = 1 \text{ h}$$

Current 3 km h^{-1}

Return journey (with current)

Boat speed = 5 km h^{-1} through the water
= 8 km h^{-1} as seen from the river bank

$$\therefore \text{time taken} = \frac{\text{distance}}{\text{speed}} = \frac{2 \text{ km}}{8 \text{ km h}^{-1}} = 0.25 \text{ h} = 15 \text{ min}$$

Hence, total time taken = 1 h 15 min

(b) The situation for boat B heading across the river

From Pythagoras' theorem,
boat speed = $\sqrt{5^2 - 3^2}$
= 4 km h^{-1}
as seen from the river bank

Boat speed = 5 km h^{-1} through the water, but this boat must head slightly upstream so that it can travel directly across.

Current 3 km h^{-1}

$\therefore \text{time taken} = \frac{\text{distance}}{\text{speed}} = \frac{2 \text{ km}}{4 \text{ km h}^{-1}} = 0.5 \text{ h each way}$

Hence, total time taken = 1 h and this boat wins!

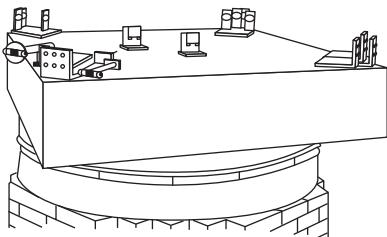


Figure 5.4 The apparatus of the Michelson–Morley experiment set up on a large stone block, to keep it rigid, and floating on mercury so that it can be easily rotated 90 degrees

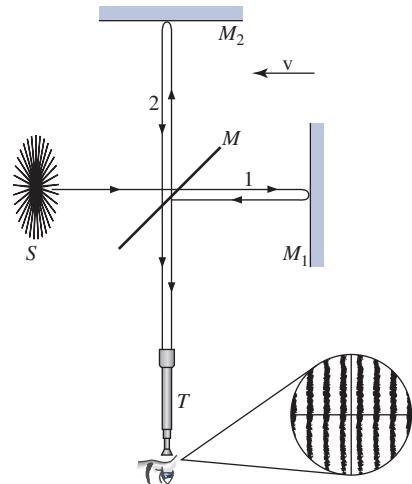
With hindsight, the result of the Michelson–Morley experiment has been able to help scientists of the twentieth century to reject the aether model and accept Einstein’s relativity. In this sense, it has been an important experiment in helping others to decide between the competing theories, along with the comparative success of relativity experiments.

It is important to note, however, that it did not sway scientific belief at the time. Aether supporters saw the null result only as an indication that their model needed improvement. Einstein, although apparently aware of the Michelson–Morley result, was not influenced by it and was unconcerned with proposed aether model modifications. He was approaching the problem from an entirely different direction.

If the object of the race was to determine the speed of the current, then it could be calculated from the difference in arrival times of the two speedboats. Also, by repeating the race with the boats interposed — A heading across the river and B heading along the river — any difference between the boats could be eliminated as a cause for the time difference as boat A should now win by the same margin.

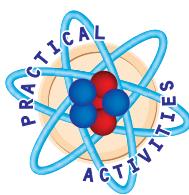
This is essentially what Michelson and Morley did — they raced two light rays over two courses, one into the supposed aether wind and one across it, then swung the apparatus through 90 degrees to interpose the rays. They were looking for a difference between the rays as they finished their race, from which they could calculate the value of the aether wind. Figure 5.4 shows their apparatus, while figure 5.5 uses a simplified diagram to show how it worked.

Figure 5.5 A simplified diagram showing the light rays’ paths in the Michelson–Morley experiment. A light ray from the source S is split into two by the half-silvered mirror. Ray A heads into the aether wind then reflects against mirror M_1 and returns. Ray B heads across the aether wind before reflecting back. Both rays finish their journey at the telescope, where they are compared.



The method of comparing the light rays involves a very sensitive effect called ‘interference’, and hence this apparatus is referred to as an ‘interferometer’. Essentially, when looking into the telescope a pattern of light and dark bands will be seen, as shown in figure 5.5. If the aether wind exists, so that one light ray is indeed faster than the other, then when the apparatus is rotated, so that the rays are interposed, the interference pattern should be seen to shift. However, no such shift was observed.

The experiment was repeated many times by Michelson and Morley, at different times of the day and year, but no evidence of an aether wind was ever found. The scientific community was not quick to abandon the aether model, however, and adapted the theory to keep it alive. One suggestion was that a large object such as a planet could drag the aether along with it. Another was that objects contract in the direction of the aether wind. However, none of these modifications survived close scrutiny. Further, the Michelson–Morley experiment has been repeated many times since 1887 by different groups with more and more sensitive equipment, and no evidence of an aether has ever been found. Yet belief in the necessity of the aether was so strong that physicists found it difficult to let go of the idea until, in 1905, Albert Einstein showed that the aether was not necessary at all.



5.2

SPECIAL RELATIVITY

Inertial frames of reference and the principle of relativity

Three hundred years before Einstein, Galileo posed a simple idea, now called the ‘principle of relativity’, which states that all steady motion is relative and cannot be detected without reference to an outside point.

An **inertial frame of reference** is a non-accelerated environment. Only steady motion or no motion is allowed. A non-inertial frame of reference experiences acceleration.

The idea can be found built into Newton's First Law of Motion as well. Put another way, if you are travelling inside a vehicle you cannot tell if you are moving at a steady velocity or standing still without looking out the window. You may have experienced this personally when sitting in a train and an adjacent train begins to roll — at first you may think that your own train is moving until you look out of a window on the other side of the carriage.

There are two points that must be reinforced:

- The principle of relativity applies only for non-accelerated steady motion; that is, standing at rest or moving with a uniform velocity. This is referred to as an **inertial frame of reference**. Situations that involve acceleration are called non-inertial frames of reference.
- This principle states that within an inertial frame of reference you cannot perform any mechanical experiment or observation that would reveal to you whether you were moving with uniform velocity or standing still.

As an example, if you were seated in a very smooth train and you held up a string with a small object tied to the end, the object would hang so that the string was vertical. However, as the train pulled away from the station you would notice the object swing backward so that the string was no longer vertical. This would continue until the train reached its cruising speed and stopped accelerating, at which point the object would move forward so that the string was once again vertical. When rounding bends in the track you would notice the string leaning one way or the other, but once the track straightened out the string would once again be vertical, just as it was when standing at the station. This plumb bob is operating as a simple accelerometer but it is unable to distinguish between being motionless and steady motion.

In the late nineteenth century, belief in the aether posed a difficult problem for the principle of relativity, because the aether was supposed to be stationary in space and light was supposed to have a fixed velocity relative to the aether. This meant that if a scientist set up equipment to measure the speed of light from the back of a train carriage to the front, and it turned out that the light was slower than it should be, the train must be moving into the aether. Put another way, this optical experiment provides a way to violate the principle of relativity where no mechanical experiment could.

A constant speed of light

Around the turn of the twentieth century, Albert Einstein puzzled over the apparent violation of the principle of relativity posed by the aether model. He had an ability to reduce a problem down to its simplest form and present it as a thought experiment. In this case the question he posed was: if I were travelling in a train at the speed of light and I held up a mirror, would I be able to see my own reflection? If the aether model was right, light could go no faster than the train. It could never catch up with the mirror to return as a reflection. The principle of relativity is thus violated because seeing one's reflection disappear would be a way to detect motion. On the other hand, if the principle of relativity were not to be violated, the reflection must be seen normally, which means that it is moving away from the mirror holder at $3 \times 10^8 \text{ m s}^{-1}$. However, this would mean that an observer on the embankment next to the train would see that light travelling at twice its normal speed.

This was a considerable dilemma but Einstein had a strong belief in the unity of physics and decided that the principle of relativity must not be violated and the reflection in the mirror must always be seen. This, in turn, meant that the aether did not exist. As a way out of the dilemma described, he also decided that the train rider and the person on the embankment must both observe the light travelling at its normal speed of $3 \times 10^8 \text{ m s}^{-1}$.

But how can this be? Einstein realised that if both observers were to see the same speed of light, and since speed = $\frac{\text{distance}}{\text{time}}$, then the distance and time witnessed by both observers must be different.

A published article

These ideas were explicitly stated in Einstein's 1905 paper titled 'On the Electrodynamics of Moving Bodies', which presented:

- a first postulate: the laws of physics are the same in all frames of reference; that is, the principle of relativity always holds
- a second postulate: the speed of light in empty space always has the same value, c , which is independent of the motion of the observer; that is, everyone always observes the same speed of light regardless of their motion
- a statement: the luminiferous aether is superfluous; that is, it is no longer needed to explain the behaviour of light. Einstein now had the confidence to set this concept aside.

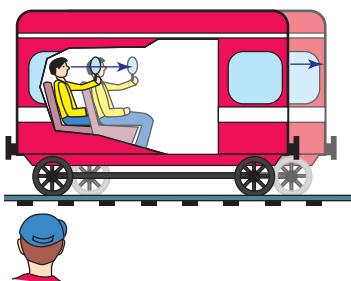


Figure 5.6 The train traveller in the light-speed train looks at his reflection in a mirror. An observer on the embankment outside the train sees the light travelling twice as far.

Space–time

In Newtonian physics, distance and velocity can be relative terms, but time is an absolute and fundamental quantity. Figure 5.6 uses the example of the train rider with the mirror, and shows how the velocity of the light is characterised by two events — the light leaving his face and the light arriving at the mirror (if we consider just the forward part of the light's journey). Remembering that the train is travelling at the speed of light, Newtonian physics says that the observer on the embankment outside the train records precisely double the distance of the journey of the light compared with that recorded by the train rider; however, they both record the same time.

Since velocity, $v = \frac{\text{distance covered}}{\text{time taken}}$, this means that the observer on

the embankment would measure a velocity of light twice that measured by the train rider.

However, Einstein's theory says that this is not what will occur. Rather, both the observer on the embankment and the train rider will measure precisely the same value for the velocity of light, called ' c '. He realised that this could only be true if the observer and the rider observed different times as well as different distances in such a way that distance divided by time always equals the same value, c .

Einstein radically altered the assumptions of Newtonian physics so that now the speed of light is absolute, and space and time are both relative quantities that depend upon the motion of the observer. In other words, the measured length of an object and the time taken by an event depend entirely upon the velocity of the observer. Further to this, since neither space nor time are absolute, the theory of relativity has replaced them with the concept of a space–time continuum. Any event then has four dimensions (three space coordinates plus a time coordinate) that fully define its position within its frame of reference.

PHYSICS FACT

Definition of the metre

The metre as a unit of length was first defined in 1793 when the French government decreed it to be 1×10^{-7} times the length of the Earth's quadrant passing through Paris. This arc was surveyed and then three platinum standards and several iron copies were made. When it was discovered that the quadrant survey was incorrect, the metre was redefined as the distance between two marks on a bar. In 1875 the Système Internationale (SI) of units was set up so that the definition became more formal: a metre was the distance between two lines scribed on a single bar of platinum–iridium alloy. Copies, or ‘artefacts’, were made for dissemination of this standard. There is always a need for the accuracy of a unit of measure to keep pace with improvements in

technology and science, so the metre has since been redefined twice.

The current definition of the metre uses the constancy of the speed of light in a vacuum ($299\,792\,458 \text{ m s}^{-1}$) and the accuracy of the definition of one second ($9\,129\,631\,770$ oscillations of the ^{133}Cs atom), to achieve a definition that is both highly accurate and consistent with the idea of space–time. One metre is now defined as the length of the path travelled by light in a vacuum during the time interval of $\frac{1}{299\,792\,458}$ of a second.

The term ‘light-year’ is a similar distance unit, being the length of the path travelled by light in a time interval of one year. One light-year is approximately equal to $9.467\,28 \times 10^{12}$ km.

5.3 CONSEQUENCES OF SPECIAL RELATIVITY

Part of the intriguing nature of relativity is that it begins with ideas that are so logical as to be inescapable. But these seemingly simple ideas lead to conclusions that can be amazing and, at first, difficult to accept. In this section, we will examine some of the more well-established consequences of the theory of relativity.

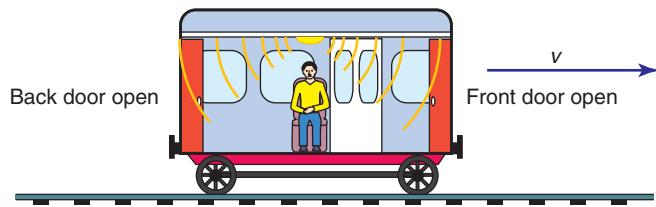
The relativity of simultaneity

As a way of better understanding how time is affected by relativity, Einstein analysed our perception of simultaneous events. He pointed out that when we state the time of an event, we are, in fact, making a judgement about simultaneous events. For example, if we say ‘school begins at 9 am’, then we are really saying that the ringing of a certain bell and the appearance of ‘9 am’ on a certain clock are simultaneous events.

Einstein contended that if an observer sees two events to be simultaneous then any other observer, in relative motion to the first, generally will not judge them to be simultaneous. In other words, simultaneous events in one frame of reference are not necessarily observed to be simultaneous in a different frame of reference. This is known as the relativity of simultaneity. In order to grasp his contention, Einstein offered the thought experiment shown in figure 5.7.

An operator of a lamp rides in the middle of a rather special train carriage. The doors at either end of the carriage are light-operated. At an instant in time when the operator happens to be alongside an observer on the embankment (outside the moving train), the operator switches on the lamp which, in turn, opens the doors.

(a) As seen by train traveller



(b) As seen by stationary observer

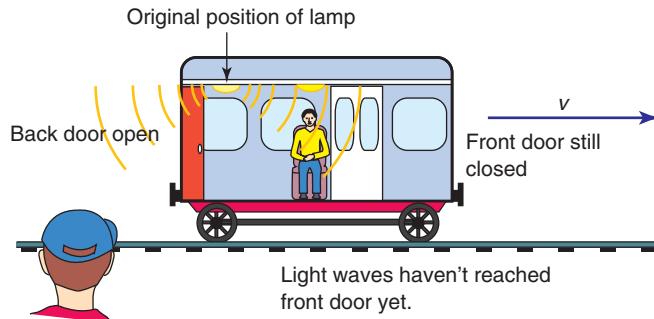


Figure 5.7 A thought experiment to illustrate the relativity of simultaneity

The operator of the lamp will see the two doors opening simultaneously. The distance of each door from the lamp is the same and light will travel at the same speed (c) both forward and backward so that each door receives the light at the same time and they open simultaneously.

The observer on the embankment, however, sees the situation differently. After the lamp is turned on, but before the light has reached the doors, the train has moved so that the front door is now further away and the back door is closer. He sees the light travelling both forward and backward at the same speed (c), but the forward journey is now longer than the backward journey, so that the back door is seen to open before the front door. They are most definitely not judged to be simultaneous events.

It is tempting to ask who is correct — the operator in the train or the observer on the embankment. The answer is that they both are. Both observers judged the situation correctly from their different frames of reference and this is a direct consequence of the constancy of the speed of light.

The relativity of time

We have already seen that time is perceived differently by observers in relative motion to each other. We are now going to determine an equation that shows this mathematically. Much of the work that follows uses the relationship that $\text{distance} = \text{velocity} \times \text{time}$ or, when we are talking about the passage of light, $\text{distance} = ct$. This relationship comes from the definition of velocity.

The following thought experiment uses a ‘light clock’. As shown in figure 5.8, a light pulse is released by a lamp, travels the length of the clock and is then reflected back to a sensor next to the lamp. When the sensor receives the pulse of light, it goes ‘click’.

For this thought experiment we shall return to the train scenario favoured by Einstein. Imagine a traveller, seated in a speeding train. The light clock is arranged vertically, with the lamp at the ceiling and the

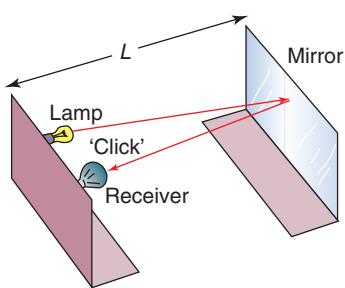


Figure 5.8 A light clock

A **rest frame** is the frame of reference within which a measured event occurs or a measured object lies at rest.

mirror on the floor. An observer is watching from the embankment outside the train. Our question now is this: when a light pulse is released, how long does it take to travel down to the mirror and return back to the ceiling, as seen by both the train traveller and the observer?

Let us first examine the situation as seen by the train traveller in the **rest frame**; that is, the frame within which the event occurs. If L is the height of the train carriage, for the total journey we can say that:

$$\text{distance} = 2L = ct_0$$

where

t_0 = time taken as seen by traveller

L = height of the carriage

so that

$$t_0 = \frac{2L}{c}$$

Examine now the situation as seen by the observer on the embankment. Figure 5.9 (page 80) compares the way the situation is viewed by each person. From outside the train the observer sees the light travelling along a much longer journey, and its length can be determined using Pythagoras' theorem:

$$\begin{aligned} \text{Total journey} &= ct_v = 2 \sqrt{L^2 + \left(\frac{vt_v}{2}\right)^2} \\ &= \sqrt{4L^2 + v^2 t_v^2}. \end{aligned}$$

Squaring this expression gives:

$$c^2 t_v^2 = 4L^2 + v^2 t_v^2.$$

Rearranging this leads to:

$$t_v^2 = \frac{4L^2}{(c^2 - v^2)}.$$

Taking the square root of both sides:

$$t_v = \frac{2L}{c \sqrt{1 - \frac{v^2}{c^2}}}$$

but, from above,

$$t_0 = \frac{2L}{c}$$

$$\text{Substituting this into the expression gives: } t_v = \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

(the time dilation equation)

where

t_0 = time taken in the rest frame of reference

= proper time

t_v = time taken as seen from the frame of reference in relative motion to the rest frame

v = velocity of the train

c = speed of light.

Note that t_0 is the time taken for the clock to go 'click' as observed by the train traveller, while t_v is the time taken as observed by the person on the embankment. Looking at the last expression above, we can see that

eBookplus

Weblink:

[Time dilation applet](#)

eModelling:

Time dilation calculator

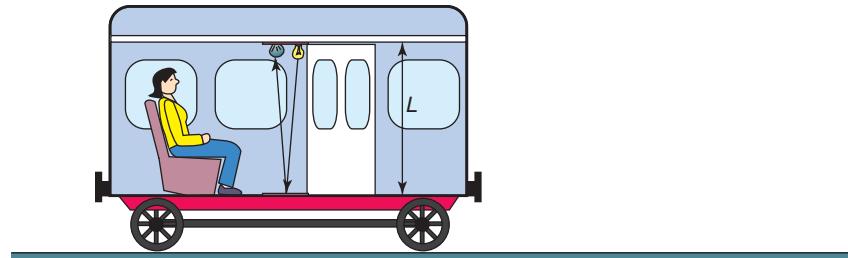
Use a spreadsheet to explore how speed affects time measurement.

doc-0037

the term $\sqrt{1 - \frac{v^2}{c^2}}$ is always less than one so that t_v is always greater than t_0 .

This means that the clock takes longer to go ‘click’ as observed by the person on the embankment or, put another way, the outside observer hears the light clock clicking slower than does the train traveller. Time is passing more slowly on the train as observed by the person outside the train!

(a) As seen by the train traveller



(b) As seen by an observer on an embankment outside the train

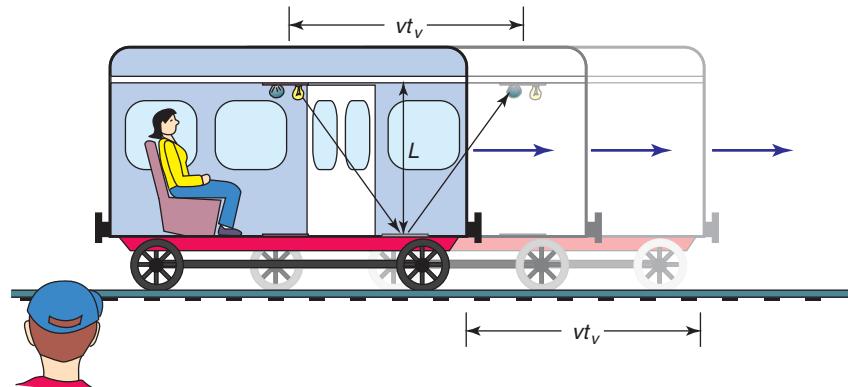


Figure 5.9 The length of the path of the light is perceived differently by the train traveller and the observer on the embankment. The observer sees the light travel further but with the same speed, hence time slowed down on the train.

Time dilation is the slowing down of events as observed from a reference frame in relative motion.

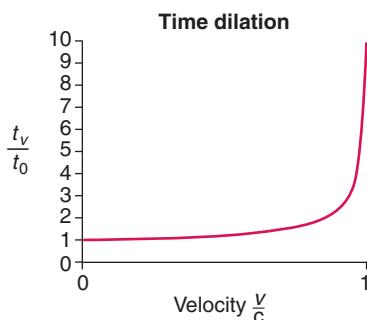


Figure 5.10 The degree of time dilation varies with velocity.

This effect is called **time dilation** and can be generally stated as follows: the time taken for an event to occur within its own rest frame is called the proper time t_0 . Measurements of this time, t_v , made from any other inertial reference frame in relative motion to the first, are always greater. The degree of time dilation varies with velocity as shown in figure 5.10.

It can be most simply stated as: moving clocks run slow.

This rather startling conclusion has been experimentally verified by comparing atomic clocks that have been flown over long journeys with clocks that have remained stationary for the same period. These experiments are possible only because of the extreme accuracy of atomic clocks built over the last few decades, even though Einstein predicted this effect about 100 years ago.

Further supporting evidence has been found in the abundance of mesons striking the ground after having been created in the upper atmosphere by incoming cosmic rays. What is surprising is that the mesons have a velocity of about $0.996c$ and, at that speed, should take approximately $16\ \mu s$ to travel through the atmosphere. However, when

measured in a laboratory, mesons have an average lifetime of approximately $2.2 \mu\text{s}$. This anomaly can be explained by the fact that $2.2 \mu\text{s}$ represents their proper lifetime, as measured in their rest frame, whereas $16 \mu\text{s}$ is a dilated lifetime due to their relativistic speed.

SAMPLE PROBLEM

5.1

A time-dilated sneeze

A train traveller sneezes just as his train passes through a station. The sneeze takes precisely 1.000 s as measured by another person seated next to the sneezer. If the train is travelling at half the speed of light, how long does the sneeze take as seen by a person standing on the platform of the station?

SOLUTION

The rest time, t_0 , is the time as observed within the sneezer's rest frame, and therefore is 1.000 s . The time dilation equation is needed to determine the time, t_v , as observed from the frame in relative motion; that is, the platform:

$$\begin{aligned} t_v &= \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}} \\ &= \frac{1.000}{\sqrt{1 - \left(\frac{0.5c}{c}\right)^2}} \\ &= 1.155 \text{ s.} \end{aligned}$$

SAMPLE PROBLEM

5.2

A time-dilated yawn

Continuing the last problem, if the person standing on the platform yawned just as the train was passing through, and this yawn lasted 2.000 s as measured by the yawner, what would be the duration of the yawn as measured by the train travellers?

SOLUTION

In this case the station platform is the rest frame of the yawn so that $t_0 = 2.000 \text{ s}$ and the time as measured from the train is t_v .

$$\begin{aligned} \therefore t_v &= \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}} \\ &= \frac{2.000}{\sqrt{1 - \left(\frac{0.5c}{c}\right)^2}} \\ &= 2.309 \text{ s.} \end{aligned}$$

Notice that each observer sees time dilated in the other frame of reference. This is central to relativity. There is no absolute frame of reference. No inertial frame is to be preferred over another and relativistic effects are reversible if viewed from a different frame.

eBook plus

eModelling: Length contraction calculator

Use a spreadsheet to explore how speed affects length measurement.
doc-0038

The relativity of length

As a consequence of perceiving time differently, observers in differing frames of reference also perceive length differently; that is, lengths that are parallel to the direction of motion. In order to understand how this occurs we will construct another thought experiment.

This time our train traveller has arranged the light clock so that it runs the length of the train, with the lamp and sensor located on the back wall

and the mirror on the front wall. As the train passes the observer on the embankment, the light clock emits a light pulse which travels to the front wall and then returns to the back wall where it is picked up by the sensor, which then goes ‘click’. This journey is observed by both people, but what is the length of the journey that each perceives?

Let us start with the situation as seen by the train traveller. Figure 5.11(a) shows that this is a simple situation:

$$\text{length of light journey} = ct_0 = 2L_0$$

where

$$L_0 = \text{length of train as perceived by train traveller}$$

$$t_0 = \text{time taken as perceived by train traveller.}$$

The situation seen by the observer at the side of the track is somewhat different because the train is moving at the same time, lengthening the forward leg of the light pulse’s journey and shortening the return leg, as shown in figure 5.11(b).

If t_1 is the time taken for the forward part of the journey, then:

$$\text{length of forward journey} = ct_1 = L_v + vt_1$$

$$\text{and hence, } t_1 = \frac{L_v}{c - v}$$

where

$$L_v = \text{length of the train as measured by the observer on the embankment.}$$

Similarly, t_2 is the time taken for the return, so that:

$$\text{length of the return journey} = ct_2 = L_v + vt_2$$

$$\text{and hence, } t_2 = \frac{L_v}{c + v}$$

The time for the whole journey as seen by the observer on the embankment is:

$$\text{time for journey, } t_v = t_1 + t_2 = \frac{L_v}{c - v} + \frac{L_v}{c + v}$$

$$\text{which can be rearranged to give } t_v = \frac{2L_v}{c\left(1 - \frac{v^2}{c^2}\right)}.$$

It is now crucial to appreciate that each observer perceives time differently. To take that into account, we need to equate the time dilation equation to the one just derived:

$$\frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{2L_v}{c\left(1 - \frac{v^2}{c^2}\right)} \text{ but we know that } t_0 = \frac{2L_0}{c}$$

$$\text{so } \frac{2L_0}{c\sqrt{1 - \frac{v^2}{c^2}}} = \frac{2L_v}{c\left(1 - \frac{v^2}{c^2}\right)} \text{ which reduces down to give}$$

$$L_v = L_0 \sqrt{1 - \frac{v^2}{c^2}} \text{ (the length contraction equation)}$$

where

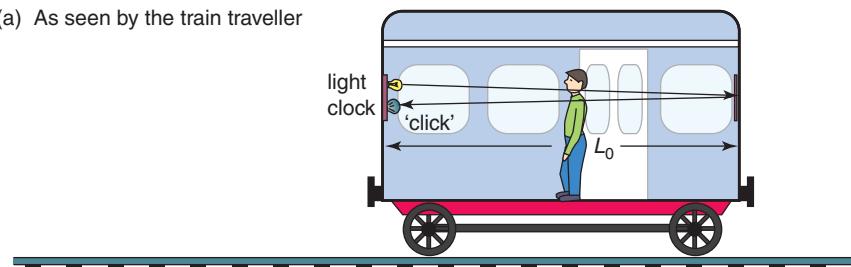
L_0 = the length of an object measured from its rest frame

L_v = the length of an object measured from a different frame of reference

v = relative speed of the two frames of reference

c = speed of light.

(a) As seen by the train traveller



(b) As seen by an observer outside the train

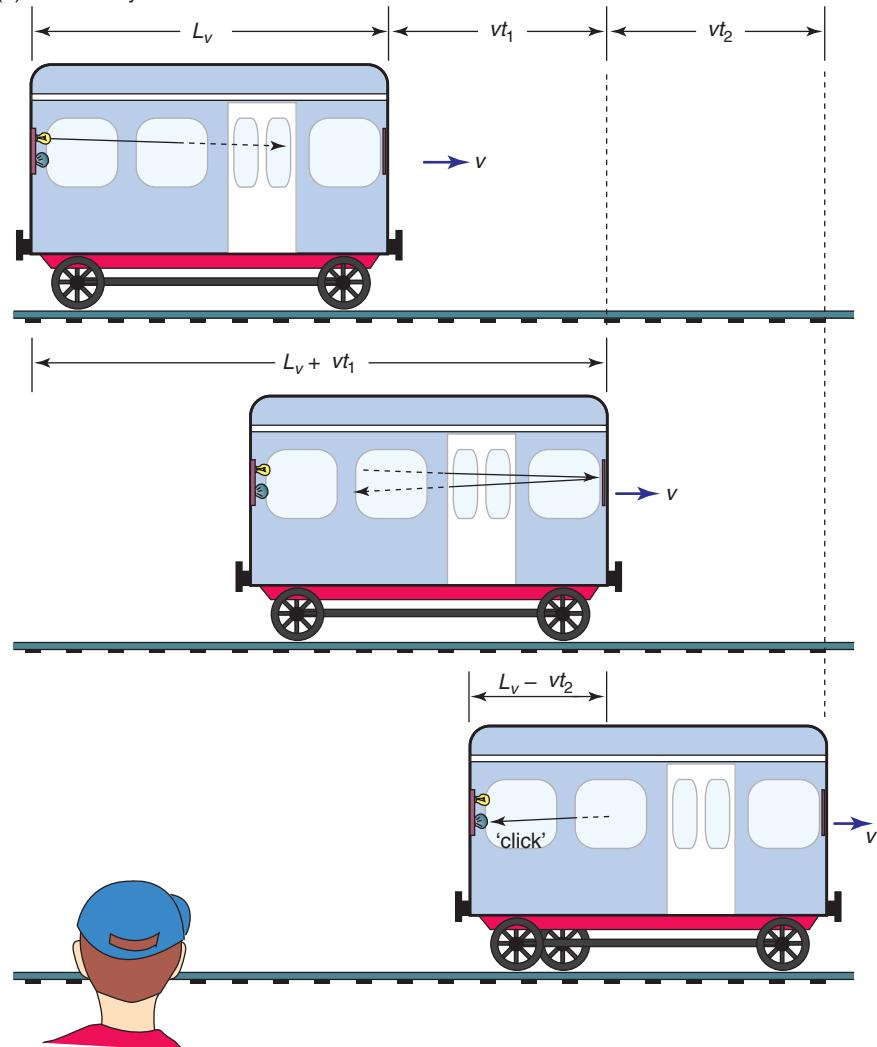


Figure 5.11 A thought experiment to derive the relativity of length

Returning to the thought experiment, this equation means that, since the term $\sqrt{1 - \frac{v^2}{c^2}}$ is always less than one, the length of the train as observed by the person on the embankment is less than that observed by the person inside the train. The person outside the train has seen the train shorten, and the faster it goes the shorter it gets!

Length contraction is the shortening of an object in the direction of its motion as observed from a reference frame in relative motion.

This effect is called **length contraction** and can be generally stated as follows: the length of an object measured within its rest frame is called its proper length, L_0 , or rest length. Measurements of this length, L_v , made from any other inertial reference frame in relative motion parallel to that length, are always less.

It can be most simply stated as: moving objects shorten in the direction of their motion.

This is another surprising result of relativity and it is interesting to see how the degree of length contraction varies with velocity, as shown in figure 5.12. Notice that as velocity approaches the speed of light, the observed length approaches zero. If this were a spaceship blasting past a planet at near light speed, the inhabitants of the planet would see a very short spaceship of nearly zero length, but the space travellers would notice no change at all to the length of their ship. They would, instead, briefly observe a wafer-thin planet in their windows, since from their inertial frame of reference it is the planet in rapid motion, not themselves.

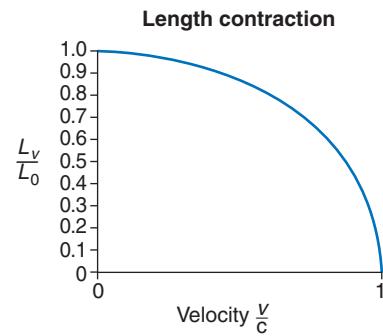


Figure 5.12 The degree of length contraction varies with velocity.

SAMPLE PROBLEM

5.3

A length-contracted train

When stationary, the carriages on the state's new VVFT (very, very fast train) are each 20 m long. How long would each carriage appear to a person standing on a station platform as this express train speeds through at half the speed of light?

SOLUTION

The proper length, L_0 , of a carriage is 20 m, while the length as seen from the platform is L_v .

$$\begin{aligned} L_v &= L_0 \sqrt{1 - \frac{v^2}{c^2}} \\ &= 20 \sqrt{1 - (0.5)^2} \\ &= 17.32 \text{ m} \end{aligned}$$

SAMPLE PROBLEM

5.4

A length-contracted person

An occupant of the VVFT looks out of a window and catches a quick glimpse of the person standing on the platform. If the thickness (from chest to back) of that person measured on the platform is 30 cm, what is the thickness observed from the train?

SOLUTION

The proper length (thickness in this case) L_0 is 30 cm and is measured from the platform since that is the rest frame of that person. The thickness observed from the train is L_v .

$$\begin{aligned} L_v &= L_0 \sqrt{1 - \frac{v^2}{c^2}} \\ &= 30 \sqrt{1 - (0.5)^2} \\ &= 25.98 \text{ m} \end{aligned}$$

Notice that observers in each frame of reference perceive lengths in the other frame to be contracted.

PHYSICS IN FOCUS

Faster than light?

Can anything travel faster than the speed of light? (so called ‘superluminal’ velocities)? According to the theory of relativity and the principle of causality, the answer to this question is ‘no’. The principle of causality says that a cause must happen before its effect. Yet, in July 2000, a team of physicists led by L. J. Wang made headlines when they claimed to have made a light pulse travel much faster than the speed of light, so fast that it went backwards in time. This result would appear to violate both relativity and causality, however, all is not as it seems.

To achieve their result the physicists passed a light wave through a specially prepared medium (caesium gas) to produce ‘anomalous dispersion’. In normal dispersion, such as occurs in glass, the blue light component in a light ray is slowed more than the red. In anomalous dispersion, the

red is slowed more than the blue. The effect of slowing the red is to change the way that the components of the light add together, to make the overall wave pulse appear to shift backward in time. It is thus a wave interference effect rather than a genuine superluminal velocity.

Wang et al. point out that this is the case, and that since it is a wave effect, no object with mass could travel this way. Additionally, because of the nature of the effect, no information could travel faster than light speed this way either. They note that their effect does not violate relativity or causality. Perhaps surprisingly, scientists have been performing this type of experiment for almost twenty years, but the light pulses have been so distorted that the results are inconclusive. The success of Wang et al. has been to design an experiment that avoids the distortions.

The relativity of mass

The first postulate of relativity states that the laws of physics are the same in all inertial frames of reference. Einstein felt strongly that this should also apply to the law of conservation of momentum, but it did not seem to. To demonstrate the problem consider the following example, illustrated in figure 5.13.

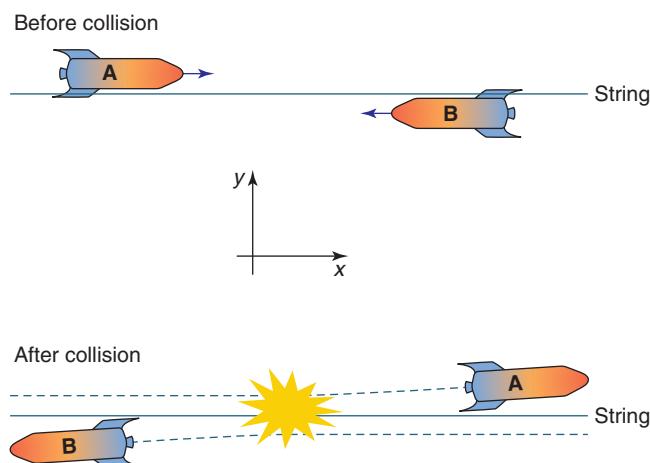


Figure 5.13 A collision between spacecraft

A long string is stretched through space. Two identical spacecraft travel toward each other on either side of the string, each with a velocity of $0.3c$ relative to the string. As they meet they collide with a glancing blow that marginally reduces their velocity along the string but also gives each a small velocity away from the string. The momentum prior to this collision is zero as the spacecraft have equal masses and equal but opposite velocities. Following the collision, the spacecraft are moving apart,

In relativity, velocities do not simply add together. For example, if two cars travel toward each other, each with a velocity of 90 km h^{-1} , then the velocity of one car as seen from the other will be 180 km h^{-1} . However, two spacecraft approaching each other with velocities of $0.9c$ will have a closing velocity of $0.99c$, rather than $1.8c$. The formula that applies here is:

$$\text{combined velocity} = \frac{v_1 + v_2}{1 + \frac{v_1 v_2}{c^2}}.$$

(The proof of this formula is not within the scope of this course.) Using this expression we can see that the two spacecraft referred to in the text, each with velocity $0.3c$, actually approach each other at $0.55c$, not $0.6c$ as mentioned.

however, their velocities in the x direction are still equal but opposite and almost unchanged. Due to the symmetry of the collision, the velocities of the spacecraft in the y direction, although very small, are also equal but opposite. As a result the total momentum in both the x and y directions is zero, hence momentum has been conserved.

The situation is different as seen from the point of view of one of the spacecraft, however, and the difference is due to time dilation and its effect on the y velocities.

As seen by a passenger of spacecraft A, prior to the collision spacecraft B speeds toward it with a velocity of $0.6c$ (actually $0.55c$ — see the note in the margin), strikes it a sideways blow and then departs at a slight angle to its original direction. Due to the presence of the string, the passengers of spacecraft A can identify that they are now moving slowly away from the string — covering, say, 10 metres in one second, giving a velocity of 10 m s^{-1} . Looking across, the passenger sees that the clocks in spacecraft B are running slow, so that it covers 10 metres in

$$\frac{1}{\sqrt{1 - 0.6^2}} = 1.25 \text{ s.}$$

The velocity of spacecraft B is thus $\frac{10 \text{ m}}{1.25 \text{ s}} = 8 \text{ m s}^{-1}$. Therefore, the y velocities of the spacecraft are not identical and momentum is not conserved in the y direction.

Algebraically:

$$\begin{aligned} p_y \text{ before collision} &= 0 \\ p_y \text{ after collision} &= m_A v_A + m_B v_B \\ &= m(10) + m(-8) \\ &= 2m \text{ where } m = m_A = m_B \end{aligned}$$

Hence, momentum is not conserved.

Einstein believed very strongly that momentum must be conserved in all inertial frames of reference. In order to solve this dilemma he suggested that the mass of an object must increase, or dilate, at relativistic speeds by a factor that compensates for the effect of time dilation on speed measurement. We can use this idea to derive a formula for mass dilation.

Referring back to the spacecraft problem, assume that total momentum is conserved in the y direction, as seen by the passenger of spacecraft A.

$$\text{momentum before collision} = \text{momentum after collision}$$

$$\begin{aligned} 0 &= m_A v_A + m_B v_B \text{ (as seen by A)} \\ &= m_A \left(\frac{r}{t_0} \right) - m_B \left(\frac{r}{t_V} \right) \end{aligned}$$

$$\text{hence } \frac{m_A}{t_0} = \frac{m_B}{t_V} \quad \text{and} \quad t_V = \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$\text{so that, as seen by A, } m_B = \frac{m_A}{\sqrt{1 - \frac{v^2}{c^2}}}$$

This expression shows how the mass of spacecraft B increases with speed, as seen by the passenger of spacecraft A. By substituting this new expression back into the problem we can see how the momentum works out.

$$p_{By} \text{ after collision} = m_B v_B$$

$$\begin{aligned} &= \left(\frac{m_A}{\sqrt{1 - \frac{v^2}{c^2}}} \right) \left(\frac{r}{t_0 / \sqrt{1 - \frac{v^2}{c^2}}} \right) \\ &= \left(\frac{m_A}{\sqrt{1 - \frac{v^2}{c^2}}} \right) \left(\frac{r \sqrt{1 - \frac{v^2}{c^2}}}{t_0} \right) \\ &= \frac{m_A r}{t_0} \end{aligned}$$

Hence, we can now say:

$$\begin{aligned} p_y \text{ before collision} &= 0 \\ p_y \text{ after collision} &= p_{Ay} + p_{By} \\ &= m_A v_A + m_B v_B \\ &= m_A \left(\frac{r}{t_0} \right) - \left(\frac{m_A r}{t_0} \right) \\ &= 0 \end{aligned}$$

Hence, momentum is now conserved.

The masses of the two spacecraft were originally identical, so the expression relating the masses can be generalised to the form

$$m_v = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

where

m_0 = mass measured in the rest frame of reference

= rest mass

m_v = mass as seen from the frame of reference in relative motion to the rest frame

v = velocity

c = speed of light

Mass dilation is the increase in the mass of an object as observed from a reference frame in relative motion.

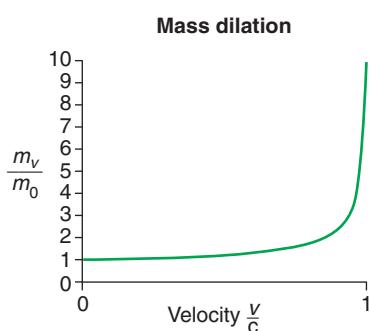


Figure 5.14 Mass as a function of velocity

This effect is called **mass dilation** and can be generally stated as follows: The mass of an object within its own rest frame is called its rest mass m_0 . Measurements of this mass m_v , made from any other inertial reference frame in relative motion to the first, are always greater. The degree of mass dilation varies with velocity as shown in figure 5.14. The effect can be most simply stated as: moving objects gain mass.

Experimental evidence for mass dilation came quickly. In 1909 it was noticed that beta particles (electrons) emitted by different radioactive substances possessed different charge to mass ratios. The various particles were travelling at significant fractions of the speed of light. Furthermore, the greater the speed of the beta particle, the smaller was its charge:mass ratio. When the effect of mass dilation was accounted for, the beta particles were all found to have the same charge:mass ratio. Modern particle accelerators, however, demonstrate mass dilation every time they are used. As they accelerate particles, such as electrons or protons, to relativistic speeds, ever greater forces are required as the particles' masses progressively increase.

SAMPLE PROBLEM**5.5*****The rest mass of an electron***

The rest mass of an electron is 9.109×10^{-31} kg. Calculate its mass if it is travelling at 80 per cent of the speed of light.

SOLUTION

$$\begin{aligned}m_v &= \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \\&= \frac{9.109 \times 10^{-31}}{\sqrt{1 - \frac{(0.8c)^2}{c^2}}} \\&= \frac{9.109 \times 10^{-31}}{0.6} \\&= 1.518 \times 10^{-30} \text{ kg}\end{aligned}$$

That is, the electron's mass is approximately 1.7 times its rest mass.

SAMPLE PROBLEM**5.6*****The rest mass of an electron near the speed of light***

Calculate the mass of an electron if travelling at 99.9 per cent of the speed of light.

SOLUTION

$$\begin{aligned}m_v &= \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \\&= \frac{9.109 \times 10^{-31}}{\sqrt{1 - \frac{(0.999c)^2}{c^2}}} \\&= \frac{9.109 \times 10^{-31}}{0.0447} \\&= 2.037 \times 10^{-29} \text{ kg}\end{aligned}$$

That is, the electron's mass this time is approximately 22 times its rest mass.

The equivalence of mass and energy

Note in figure 5.14 that as the speed of an object approaches the speed of light c , its mass approaches an infinite value. It is this enormous increase in mass that prevents any object from exceeding the speed of light. This is because an applied force is required to create acceleration. Acceleration leads to higher velocities, which eventually leads to increased mass. This means that further accelerations will require ever-greater force. As mass becomes infinite, an infinite force would be required to achieve any acceleration at all. Sufficient force can never be supplied to accelerate beyond the speed of light.

But herein lies a problem — if a force is applied to an object, then work is done on it. Another way to say this is that energy is given to the object. In the sort of situation we are considering, this energy would take the form of increased kinetic energy as the object speeds up. But at near light speed the object does not speed up as we would normally expect, so where is the energy going? The applied force is giving energy to the object and the object does not acquire the kinetic energy we would expect. Instead, it acquires extra mass, as shown in figure 5.14. Einstein made an inference here and stated that the mass (or inertia) of the object contained the extra energy.

Relativity results in a new definition of energy as follows:

$$E = E_k + mc^2$$

where

E = total energy

E_k = kinetic energy

m = mass

c = speed of light.

Notice that when an object is stationary, so that it has no kinetic energy, it still has some energy due to its mass. This is called its mass energy or **rest energy** and is given by:

$$E = mc^2$$

where

E = rest energy (J)

m = mass (kg)

c = speed of light (3×10^8 m s⁻¹).

This famous equation states clearly that there is an equivalence between mass and energy — that mass has an energy equivalent and vice versa. The speed of light squared is a very large number, however, and this means that if you were able to convert just a small amount of matter it would yield an enormous amount of energy. Just one kilogram of mass, for example, is equivalent to 9 million billion joules of energy.

This equivalence may seem strange when first seen; however, you must bear in mind that it has been proven experimentally many times, and demonstrated most dramatically as the energy released by a nuclear bomb.

SAMPLE PROBLEM

5.7

SOLUTION

The rest energy of an electron

What is the energy equivalent of an electron of mass 9.109×10^{-31} kg?

$$\begin{aligned}E &= mc^2 \\&= (9.109 \times 10^{-31})(3 \times 10^8)^2 \\&= 8.1981 \times 10^{-14}\text{ J}\end{aligned}$$

SAMPLE PROBLEM

5.8

SOLUTION

The rest energy of a uranium atom

What is the energy equivalent of an atom of uranium, which has a mass of 238 amu or 3.953×10^{-25} kg?

$$\begin{aligned}E &= mc^2 \\&= (3.953 \times 10^{-25})(3 \times 10^8)^2 \\&= 3.558 \times 10^{-8}\text{ J}\end{aligned}$$

Compare this to the energy yield of 3.2×10^{-11} J from a typical fission reaction in which only part of the mass of a uranium atom is converted to energy as the nucleus splits into two smaller parts.

Relativistic space flight

Designers of a new kind of spacecraft called a light sail make the remarkable claim that these craft could journey to Proxima Centauri, our closest neighbouring star and shortest interstellar journey, at a speed of 0.1c or 10% of the speed of light. This is far in excess of current achievable velocities. Assuming it to be true, how long would such a journey take?

The distance to Proxima Centauri is approximately four light-years, or 3.7869×10^{13} km. A speed of $0.1c$ is equal to 1.08×10^8 km h⁻¹. The time taken is easily calculated:

$$\begin{aligned}\text{speed} &= \frac{\text{distance}}{\text{time taken}} \\ \text{so } \text{time taken} &= \frac{\text{distance}}{\text{speed}} \\ &= \frac{37\,869\,000\,000\,000 \text{ km}}{108\,000\,000 \text{ km h}^{-1}} \\ &= 350\,640 \text{ hours} = 40 \text{ years.}\end{aligned}$$

When the distances and speeds are this large, a simpler calculation results if distance is expressed in light-years and speed is expressed in terms of c , as follows:

$$\begin{aligned}\text{time taken} &= \frac{\text{distance}}{\text{speed}} \\ &= \frac{4c \text{ years}}{0.1c} \\ &= 40 \text{ years.}\end{aligned}$$

However, this is the time taken as observed from Earth. The space travellers within the spacecraft will, according to relativity, record a slightly different travel time. There are two ways to calculate it:

- **Method 1:** If the time recorded on Earth, t_v , is 40 years, the rest time, t_0 , lapsed on the spacecraft, can be calculated using the time dilation equation:

$$\begin{aligned}t_v &= \frac{t_0}{\sqrt{1 - \left(\frac{v}{c}\right)^2}} \\ \therefore t_0 &= t_v \sqrt{1 - \left(\frac{v}{c}\right)^2} \\ &= 40 \sqrt{1 - (0.1)^2} \\ &= 39.799 \text{ years} \\ &= 39 \text{ years } 292 \text{ days.}\end{aligned}$$

In other words, the spacecraft reaches its destination 73.25 days, or almost two-and-a-half months, short of 40 years.

- **Method 2:** The occupants of the spacecraft see the distance they have to cover contracted according to the length contraction equation:

$$\begin{aligned}L_v &= L_0 \sqrt{1 - \left(\frac{v}{c}\right)^2} \\ &= 4 \sqrt{1 - (0.1)^2} \\ &= 3.9799 \text{ light years.}\end{aligned}$$

$$\begin{aligned}\text{Now, time taken} &= \frac{\text{distance}}{\text{speed}} \\ &= \frac{3.9799 c \text{ years}}{0.1c} \\ &= 39.799 \text{ years} \\ &= 39 \text{ years } 292 \text{ days.}\end{aligned}$$

This method produces the same result as the previous method — the spacecraft actually arrives 73.25 days short of 40 years.

This example illustrates the influence that relativity can have upon space travel when speeds become ‘relativistic’, which usually means 10% of the speed of light or faster. When speeds are less than this, the effects are almost negligible. When speeds become greater than this, the effects become significant. As figures 5.10 and 5.12 show, the effects intensify sharply with speeds faster than 0.9c.

Table 5.1 compares space travel at a variety of speeds, showing the time passed on board a spacecraft during one Earth day, as well as the length of external objects and distances as a percentage of the original length.

Table 5.1 A comparison of relativistic effects

SPACECRAFT	SPEED (km h ⁻¹)	RATIO $\frac{v}{c}$	TIME PASSED ON SPACECRAFT IN ONE EARTH DAY			CONTRACTED LENGTHS AS % OF ORIGINAL
			HOURS	MINUTES	SECONDS	
Space shuttle	28 000	0.000 026	23	59	59.999 972	99.999 999 97
Fast space probe	100 000	0.000 093	23	59	59.999 630	99.999 999 6
Light sail	108 000 000	0.1	23	52	46.92	99.499
<i>Starship Intastella</i>	972 000 000	0.9	10	27	40.89	43.59
<i>Starship Galactica</i>	1 079 892 000	0.999 9	0	20	21.85	1.4

Astronauts in orbit around the Earth will not observe any noticeable effect at all, since the length of their day is just 30 microseconds shorter than on Earth. Even in a current-day speedster at 100 000 km h⁻¹ the days are just 400 microseconds shorter and lengths are still 99.999 9996% of their former selves.

The situation is very different at 0.9c, however. While back on Earth 24 hours pass, at this speed less than 10.5 hours elapse and external objects such as planets have squashed up to 44% of their former lengths.

Consider now travelling in the *Galactica*. This flyer manages to zoom along at 99.99% of the speed of light and, in the course of one Earth day, just over 20 minutes have passed on board. Lengths have contracted to just 1.4% of their original lengths and the four light-year trip to Proxima Centauri would be completed in just over 20 days according to the ship’s clock. It would be natural at this point to wonder how this could be possible — if light takes four years to cover the distance, how could this starship, travelling at very nearly the speed of light, manage the journey in 20 days? The answer is that as observed from the Earth the starship does take four years; however, the clocks on the starship, both electronic and biological, are running slow so that by their reckoning only 20 days pass.

It should be pointed out that the energy costs of achieving these types of speeds would be prohibitive, even assuming that such speeds were technically possible. Acceleration is always the most energy costly phase of a space mission. As we have seen earlier, the effect of mass accumulation and time dilation is to require accelerations beyond 0.9c to involve ever greater forces and energy input for only marginal increases.

The twins paradox

Einstein himself suggested one of the strangest effects of relativity. His idea was that a living organism could be placed in a box and taken on a relativistic flight and then returned to its starting place almost without ageing, while similar organisms that had remained behind had long since died of old age.

eBookplus

Weblink:
Twins paradox
applet

Let us reconsider the problem using the flight of the *Starship Galactica* on the previous page. Consider twins, Martha and Arthur. Martha steps aboard the starship and speeds off to Proxima Centauri at maximum speed as described previously. This journey has taken Martha just 20.5 days but for Arthur, watching closely through a telescope, four years have passed. Martha immediately turns the starship around and returns to her brother. Upon arrival she is just 41 days older than when she left; however, Arthur has aged eight years waiting for her. It is not hard to appreciate that had the journey been to a star further away, Martha could have arrived back after her brother had grown old and died.

This problem is often considered a paradox, because the principle of relativity demands that no inertial frame of reference be preferred over others. In other words, relativity's effects should be reversible simply by looking at them from a different viewpoint. Length contraction, for instance, depends upon the viewer's frame of reference. An Earthbound viewer looking at the *Starship Galactica* will see it compressed to just 1.4% of its original length, while travellers within the starship would see the Earth squashed up to 1.4% of its original thickness.

If this problem is viewed differently, then it is Martha who sees Arthur disappearing away with the Earth at $0.9999c$ and then returning, so Arthur should be the younger. The apparent paradox is this: if both points of view are valid then each sibling will see the other as older than themselves. How can this be possible?

The answer is that this particular problem is not reversible. Martha's frame (the starship) has not remained inertial — it has accelerated and decelerated, turned around and then repeated its accelerations. It is very easy to tell the difference between the two frames simply by having each sibling carry an accelerometer. Hence, the two frames of reference are not equivalent and there is no paradox because Martha will definitely be younger and Arthur older.

SAMPLE PROBLEM**5.9****A relativistic interstellar journey**

If Martha were to throttle back the *Starship Galactica* so that it cruised to Proxima Centauri at just $0.8c$, how much older would Arthur be upon her immediate return?

SOLUTION

The trip time as observed by Arthur back on Earth is calculated as follows:

$$\begin{aligned}\text{time taken} &= \frac{\text{distance}}{\text{speed}} \\ &= \frac{2 \times 4c \text{ years}}{0.8c} \\ &= 10 \text{ years.}\end{aligned}$$

The trip time for Martha can be found using the time dilation equation:

$$\begin{aligned}t_v &= \frac{t_0}{\sqrt{1 - \left(\frac{v}{c}\right)^2}} \\ \therefore t_0 &= t_v \sqrt{1 - \left(\frac{v}{c}\right)^2} \\ &= 10 \sqrt{1 - (0.8)^2} \\ &= 6 \text{ years.}\end{aligned}$$

Hence, Arthur will be four years older than Martha upon her return.

SUMMARY

- The aether was the hypothesised medium for light and other electromagnetic waves. It was transparent and could not be detected, yet belief in its existence was strong since all other known waveforms require a medium through which to travel.
- The Michelson–Morley experiment used light and an effect called interference was devised to detect the aether. It was extremely sensitive yet failed to detect any indication of the existence of the aether.
- An inertial frame of reference is a non-accelerated environment. It allows for uniform velocity motion or a state of rest only.
- The principle of relativity states that it is not possible to detect uniform velocity motion while within a frame of reference, without referring to another frame. Classical physics established this principle for mechanics but not optics. Einstein included optics by extending the principle to include all the laws of physics.
- Einstein also postulated that the speed of light has the same value, c , in all reference frames; that is, to all observers.
- Time becomes a relative term once it is accepted that the speed of light is an absolute term. Distance, or space, is also a relative term.
- The SI unit of length, the metre, is defined in terms of the speed of light and time.
- Two events in different places that are judged by an observer to be simultaneous will not be simultaneous as judged by another observer in relative motion to the first.
- The time taken for an event to occur within its rest frame is called the proper time, t_0 . The time taken, t_v , as judged by observers in relative motion, will always be longer.

$$t_v = \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

- The length of an object measured within its rest frame is called its proper length, L_0 . Measured by observers in relative motion the length, L_v , will always be shorter.

$$L_v = L_0 \sqrt{1 - \frac{v^2}{c^2}}$$

- The mass of an object within its own rest frame is called its rest mass m_0 . Measurements of this mass m_v made by observers in relative motion will always be greater.

$$m_v = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

- The rest mass of an object is equivalent to a certain quantity of energy. Conversion between the two can occur under extraordinary circumstances:

$$E = mc^2$$

- Time dilation and length contraction could theoretically allow exceptionally long space journeys within reasonable periods of time, as judged by the travellers. However, relativity also indicates that the cost of energy to do this would be prohibitive.

QUESTIONS

- Outline the features of the aether model and the reasons that scientists believed that it needed to exist.
- List the supposed features of the aether.
- (a) Identify the objective of the Michelson–Morley experiment.
(b) Construct a diagram showing the paths of the light rays in the Michelson–Morley experiment.
(c) Write a one-paragraph description of how the apparatus worked.
(d) The experiment had a very definite result. Outline this result.
- Evaluate the success of the Michelson–Morley experiment in proving or disproving the aether model. Discuss the role it played in the development of ideas.
- Outline the essential aspects of an inertial frame of reference and identify a method to distinguish an inertial from a non-inertial frame of reference.
- You are in a spaceship heading, you think, in free motion towards Pluto; however, you are far from any reference point to check your progress. Suddenly a comet approaches from behind and overtakes you, heading in the same direction. Identify which of the following interpretations of events is correct and any method that could distinguish them.

- (a) The comet is also travelling towards Pluto but at a higher speed.
 (b) Your spaceship is stationary but the comet is heading towards Pluto.
 (c) The comet is stationary and you are travelling away from Pluto.
 (d) You are both travelling away from Pluto, but you have a higher speed.
7. Identify any method you could use to tell if your spacecraft, far from any planet or star, is standing still or travelling with a uniform velocity. What is the name given to this idea?
8. In classical physics, time was assumed to be absolute and the maximum velocity of interaction assumed infinite. Explain how these same concepts are viewed in the theory of relativity.
9. A traveller in a very fast train holds a mirror at arm's length and looks at his reflection. Both he and an observer outside the train see the same velocity of light for the traveller's image.
- Since velocity = $\frac{\text{distance}}{\text{time}}$, describe what has happened to the length of the traveller's arm, and the time taken for the reflection to return, as seen by the outside observer.
10. (a) Compare the current definition of a metre to the original metre standard.
 (b) Evaluate the light-time definition, such as the metre and the light-year.
11. Identify the technologies required to support the current definition of a metre.
12. Complete the table of distances shown below. You will need the following conversion factors:
- $$1 \text{ light-year} = 0.3066 \text{ parsecs} = 63\,240 \text{ AU}$$
- $$= 9.4605 \times 10^{12} \text{ km}$$
- $$1 \text{ parsec} = 3.2616 \text{ light-years} = 206\,265 \text{ AU}$$
- $$= 30.857 \times 10^{12} \text{ km}$$
- $$1 \text{ astronomical unit (AU)} = 1.5813 \times 10^{-5}$$
- $$\text{light-years} = 1.496 \times 10^8 \text{ km.}$$
- | STAR | DISTANCE
(light-years) | DISTANCE
(parsecs) | DISTANCE (km) |
|----------|---------------------------|-----------------------|-------------------------|
| Canopus | | 23 | |
| Rigel | 900 | | |
| Arcturus | | 10 | |
| Hadar | | | 3.0857×10^{15} |
13. In one paragraph, describe what is meant by the 'relativity of simultaneity'.
14. If two events are observed to occur at the same place and time by one person, will they be seen in the same way by all observers? Explain.
15. Pete the flying ace needs to hide his new experimental high-velocity plane in the hangar, away from the view of foreign spy satellites. 'You'll ne'er do it,' says Jock, the Scottish flight engineer, 'The hangar's barely 80 m long but yer-r-r plane there is over 120.'
- 'She'll be right, mate,' replies Pete, 'It's just a matter of going fast enough!' Jock stands at the hangar door while he watches Pete approach in his plane.
- (a) Jock sees the plane contract as it speeds up. Calculate how fast it must be travelling for him to be able to quickly close the door with the plane, at least momentarily, contained inside.
 (b) At the speed you determined in part (a), calculate the length of the hangar as seen by Pete in the approaching plane. Will the plane fit in the hangar as judged by Pete?
 (c) Discuss how it is possible for Pete and Jock to perceive the situation so differently.
16. As the Moon orbits the Earth it has an orbital speed of approximately 3660 km h^{-1} . If its proper diameter is 3480 km, calculate what we observe as its diameter.
17. A muon is a subatomic particle with an average lifetime of 2.2 microseconds when stationary. When in a burst of cosmic rays in the upper atmosphere, muons are observed to have a lifetime of 16 microseconds. Calculate their speed.
18. If our galaxy, the Milky Way, is 20 kiloparsecs or 65 000 light-years in radius, calculate how fast a spacecraft would need to travel so that its occupants could travel right across it in 45 years.
19. Calculate the travel time to the following destinations travelling at $0.999c$. Use the conversion factors given in problem 12.
- (a) Pluto when at a distance of 39.1 AU from the Earth
 (b) Proxima Centauri at a distance of 1.295 parsecs
 (c) Sirius at a distance of 8.177×10^{13} km
 (d) Alpha Crucis at a distance of 522 light-years
 (e) Andromeda galaxy at a distance of 690 kiloparsecs

20. The length of a spacecraft is observed, by someone on a nearby planet, to shrink to half its proper length. Calculate:
- the speed of the spacecraft relative to the planet
 - the observed mass of the spacecraft if its rest mass is 5×10^4 kg
 - the amount of time passed on the planet when one second has passed on the spacecraft, as observed from the planet.
21. A super rocket racer has a proper length of 30 m, a rest mass of 300 000 kg and can fly at 0.3c. Calculate:
- the length of the aircraft at speed when observed from the Earth
 - the time difference between the clocks of the pilot and his airbase if they were perfectly synchronised prior to lift-off and the racer was aloft and at speed for 10 h according to the pilot's clock
 - the mass of the speeding aircraft when observed from the Earth.
22. A Martian spaceship travelling at near-light speed passes a stationary Venusian spaceship. Each is made of glass and in each the occupants are holding a dance party.
- Describe the dancing Venusians as seen by the Martians.
 - Describe the dancing Martians as seen by the Venusians.
 - Are your answers to (a) and (b) contradictory? Explain.
23. Define rest energy.
24. Identify different situations in which rest energy is extracted.
25. Calculate the rest energy of the following objects:
- a proton of mass 1.673×10^{-27} kg
 - an alpha particle of mass 4.0015 amu or 6.6465×10^{-27} kg
 - a carbon atom of mass 12.0000 amu or 1.9932×10^{-26} kg
 - 5 mg of aspirin
 - 1 kg of sugar.
26. In terms of the energy required, accelerating a spacecraft to light speed is an impossibility. Explain why this is so.



5.1 MODELLING THE MICHELSON–MORLEY EXPERIMENT

Aim

To model the Michelson–Morley experiment.

Theory

The ‘luminiferous aether’ was the assumed medium for the propagation of light waves, proposed by nineteenth-century physicists. In 1881, Michelson and Morley performed a highly sensitive experiment to detect the ‘aether wind’. This is the apparent velocity of aether moving past us caused by the Earth moving through the stationary aether. The apparatus is shown in figure 5.4 (page 74). The apparatus produced two light rays and directed them along equal-length but separate paths as shown in figure 5.5 — one into the aether wind and one across it — and then compared the rays at the end of their journey to see if one was ahead of the other.

The model

This is a book exercise using an analogy as shown in figure 5.15. A large open-top shark cage, of the sort used by long-distance swimmers, is attached to the side of a boat. Its dimensions are 10 m by 10 m. The boat and the cage are both moving through the water at 25 cm s^{-1} . In one of the leading corners are two swimmers who are about to have a race. Swimmer A will head across the leading edge of the cage and back, while swimmer B will head along the side of the cage and back. Each is capable of swimming at 1 m s^{-1} , relative to the water. Who will win?

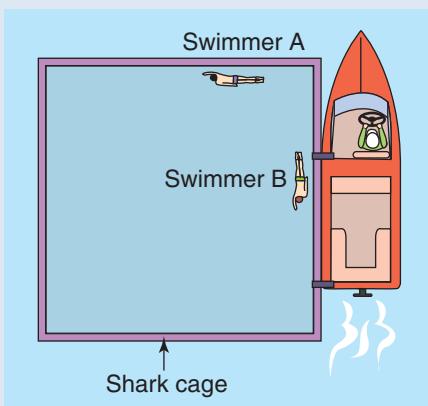


Figure 5.15
An analogy for the Michelson–Morley experiment. Two swimmers race, across and along, a large moving shark cage.

In this analogy of the Michelson–Morley experiment, the water represents the aether, the cage represents the Earth, the apparent current represents the aether wind and the swimmers represent the two light rays.

Analysis

Swimmer A

This swimmer will need to head into the current slightly in order to swim directly across. Add the swimmer’s velocity relative to the water to the apparent current to determine the swimmer’s velocity relative to the cage, as shown in figure 5.16. Use this information to determine the time taken to swim directly across the cage and back again.

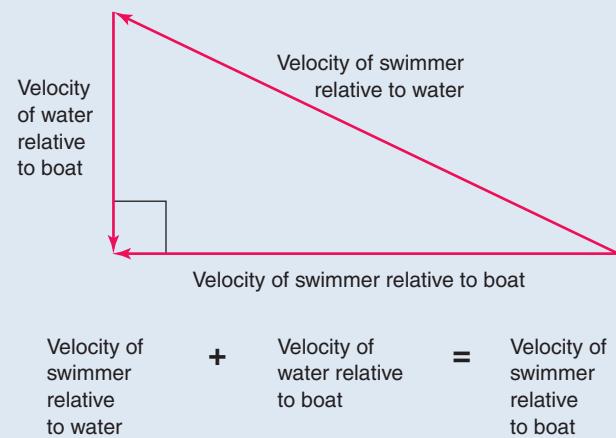


Figure 5.16 Adding the velocities of the swimmer who heads across the shark cage

Swimmer B

This swimmer’s forward journey is different from the return journey so each part needs to be treated separately. During the forward journey, the swimmer is swimming with the current but when returning, the swimmer is heading against the current.

For each part, determine the swimmer’s velocity relative to the cage and then use this information to determine the time taken to swim along the cage and back again. Add these two times to find the time for the total journey travelled by swimmer B.

Comparison

Now compare the times you have calculated for each swimmer. Which one wins the race?

Questions

- Will the winner of this race always win the race, regardless of boat speed?
- If the speed of the current is doubled, how would this affect the winning time?
- What does it mean if the swimmers’ race is a tie; that is, there is no winner?

4. In the Michelson–Morley experiment there was a null result (that is, no winner in our model). Given that their apparatus was supposed to be sensitive enough to detect the aether, what conclusions can be drawn from this?



5.2 NON-INERTIAL FRAMES OF REFERENCE

Aim

To use an accelerometer to distinguish between inertial and non-inertial frames of reference.

Apparatus

accelerometer (either a stand-alone device or data logger attachment)
dynamics trolley
string
50 g mass carrier with extra masses

Theory

A frame of reference is an environment within which an object resides. An inertial reference frame is one that is at rest or in uniform velocity. A non-inertial reference frame is one that is undergoing acceleration.

The principle of relativity states that when residing in an inertial reference frame it is not possible to tell whether the frame is at rest or in

uniform velocity without referring to another frame; it is only possible to distinguish between inertial and non-inertial frames. Since acceleration involves force, any force-detecting device can identify a non-inertial frame of reference.

An accelerometer is a device that identifies the direction and magnitude of an acceleration.

Method

- Familiarise yourself with the device by holding it horizontally and moving from side to side noting the change in its indicated acceleration.
- (a) Place the accelerometer upon a dynamics trolley arranged as shown in figure 5.17. Ensure that the string is sufficiently long so that the hanging mass strikes the floor well before the trolley reaches the pulley. When a mass of 100 g has been placed on the hanging end of the string, release the trolley and observe the scale. What was the reading of the acceleration? Once the mass reaches the floor the trolley will stop accelerating. What do you observe on the accelerometer?
(b) Repeat this procedure twice more, with the hanging mass set to 200 g and 400 g. Record the observed rate of acceleration before and after the mass strikes the floor.
- This part will require the cooperation of someone with a car.
(a) While seated in the car hold the accelerometer parallel with the sides of the car. Ask the driver to accelerate, coast for a while and then brake to a stop. Describe your observations.

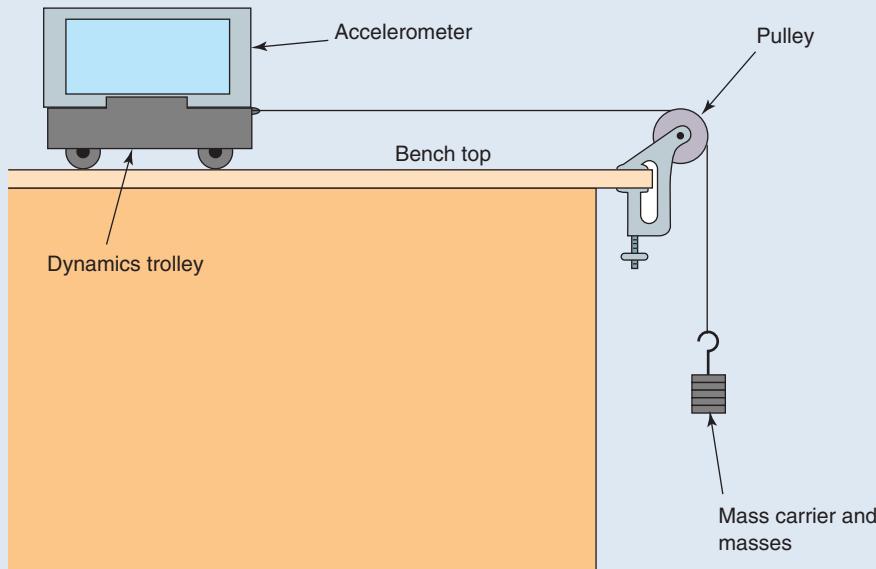


Figure 5.17

- (b) Now hold the accelerometer parallel with the front and back of the car. Ask the driver to drive around a few corners, preferably right-angle bends. Describe your observations.

Results

Part 2(a) Hanging mass 100 g

Indicated acceleration before mass strikes floor

$$= \underline{\hspace{2cm}} \text{ m s}^{-1}$$

What type of reference frame is this?

Indicated acceleration after mass strikes floor

$$= \underline{\hspace{2cm}} \text{ m s}^{-1}$$

What type of reference frame is this?

Part 2(b) Hanging mass 200 g

Indicated acceleration before mass strikes floor

$$= \underline{\hspace{2cm}} \text{ m s}^{-1}$$

Indicated acceleration after mass strikes floor

$$= \underline{\hspace{2cm}} \text{ m s}^{-1}$$

Hanging mass 400 g

Indicated acceleration before mass strikes floor

$$= \underline{\hspace{2cm}} \text{ m s}^{-1}$$

Indicated acceleration after mass strikes floor

$$= \underline{\hspace{2cm}} \text{ m s}^{-1}$$

Part 3

Observations when accelerating, coasting and braking: _____

Observations when cornering: _____

Questions

1. What is the effect of increasing the hanging mass?
2. How does the motion of an object change when undergoing acceleration?
3. What name is given to frames of reference that experience acceleration?
4. Is the accelerometer display different when travelling at a steady speed compared with standing at rest?
5. What name is given to the reference frames referred to in question 4?



Chapter 6

The motor effect and DC electric motors

Chapter 7

Generating electricity

Chapter 8

Generators and power distribution

Chapter 9

AC electric motors

MOTORS AND GENERATORS

CHAPTER

6

THE MOTOR EFFECT AND DC ELECTRIC MOTORS

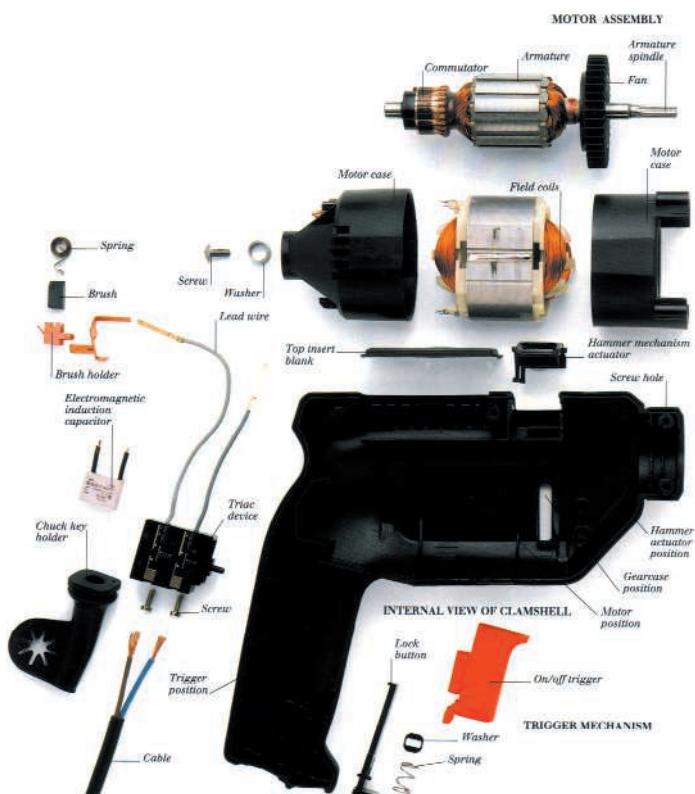


Figure 6.1 A disassembled motor from an electric drill. How does it work and what are the functions of all its parts?

Remember

Before beginning this chapter, you should be able to:

- describe the behaviour of the magnetic poles when they are brought close together
- define the direction of a magnetic field at a point as the direction that the N pole of a compass needle would point when placed at that point
- describe the magnetic field around single magnetic poles and pairs of magnetic poles
- describe the nature of a magnetic field produced by an electric current in a straight current-carrying conductor
- explain how the right-hand grip rule can determine the direction of current in, or the magnetic field lines around, a current-carrying conductor
- compare the nature and generation of magnetic fields by solenoids and a bar magnet.

Key content

At the end of this chapter you should be able to:

- identify the factors that affect the magnitude and direction of the force acting on current-carrying conductors in magnetic fields
- use the right-hand push rule to determine the direction of the force acting on current-carrying conductors in magnetic fields
- describe the force between long, parallel current-carrying conductors
- define torque
- describe the motor effect
- describe the main features of a DC electric motor and the role of each feature
- identify two methods for providing the magnetic field for a DC motor.

What would your life be like without electricity? Modern industrialised nations are dependent on electricity. Electricity is easy to produce and distribute, and is easily transformed into other forms of energy. Electric motors are used to transform electricity into useful mechanical energy. They are used in homes, for example in refrigerators, vacuum cleaners and many kitchen appliances, and in industry and transport.

In this module we will explore how electricity is used to drive electric motors, how it is produced and how it is distributed from the power stations to the consumers.

Use the box below to revise your work on magnetic fields from the Preliminary Course ('Electrical energy in the home'). This material is fundamental to the understanding of how DC electric motors operate.

PHYSICS IN FOCUS

Review of magnetic fields

- The law of magnetic poles states that opposite poles of magnets attract each other and like poles of magnets repel each other.
- Magnetic fields are represented in diagrams using lines. These show the direction and strength of the field. The density of the field lines represents the strength of the magnetic field. The closer the lines are together, the stronger the field.
- The direction of the magnetic field at a particular point is given by the direction of the force on the N pole of a magnet placed within the magnetic field. It is shown by arrows on the magnetic field lines.
- Magnetic field lines never cross. When a region is influenced by the magnetic fields of two or more magnets or devices, the magnetic field lines show the strength and direction of the resultant magnetic field acting in the region. They show the combined effect of the individual magnetic fields.
- The spacing of the magnetic field lines represents the strength of the magnetic field. It follows that field lines that are an equal distance apart represent a uniform magnetic field.
- Magnetic field lines leave the N pole of a magnet and enter the S pole.
- The following diagrams in figure 6.2 represent the magnetic fields around (a) a single bar magnet, (b) two N poles close to each other, and (c) a horseshoe magnet.
- In a diagram, as seen in figure 6.3, magnetic field lines going out of the page are represented using dot points (•). This is like an observer seeing the pointy end of an arrow as it approaches.

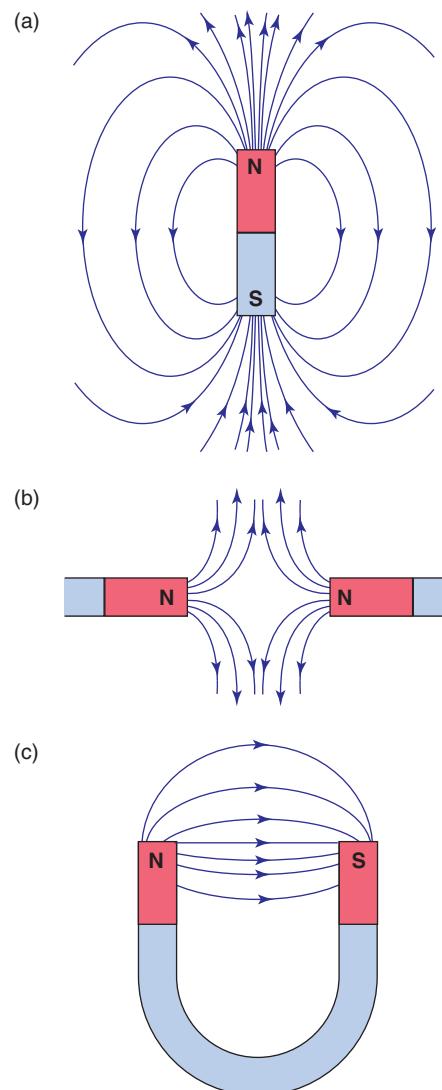


Figure 6.2 Magnetic field lines for (a) a single bar magnet, (b) two N poles close to each other and (c) a horseshoe magnet

(continued)

- Magnetic field lines going into the page, also seen in figure 6.3, are represented using crosses (\times). This is as an archer would see the rear end of an arrow as it leaves the bow.

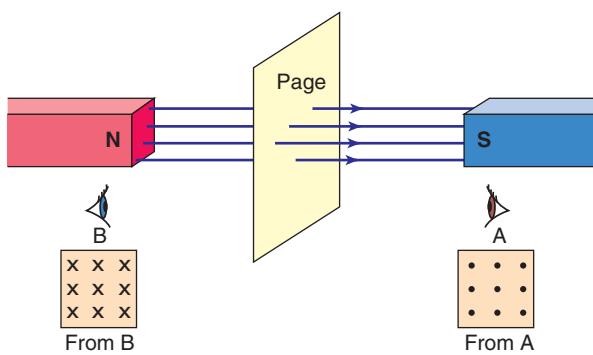


Figure 6.3 Magnetic field lines coming out of the page as observed by A, and going into the page, as observed by B

- The movement of charged particles, as occurs in an electric current, produces a magnetic field. The magnetic field is circular in nature around the current-carrying conductor, as shown in figure 6.4, and can be represented using concentric field lines. The field gets weaker as the distance from the current increases.

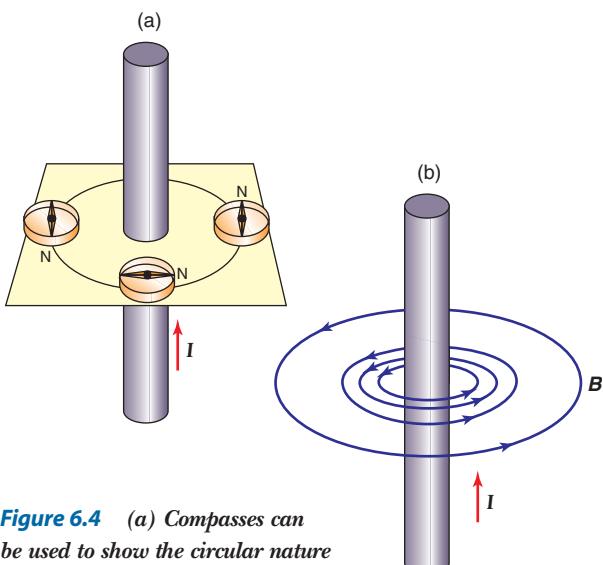


Figure 6.4 (a) Compasses can be used to show the circular nature of the magnetic field around a straight current-carrying conductor. (b) The magnetic field is circular and stronger closer to the wire.

- The direction of the magnetic field around a straight current-carrying conductor is found using the right-hand grip rule, as shown in figure 6.5. When the right hand grasps the conductor with the thumb pointing in the direction of conventional current, the curl of the fingers gives the direction of the magnetic field around the conductor.

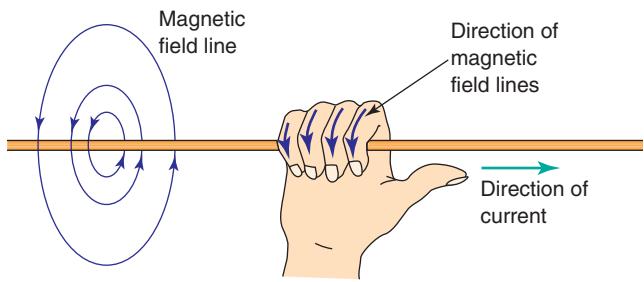


Figure 6.5 The right-hand grip rule

- When a current-carrying conductor is bent into a loop, the effect is to concentrate the magnetic field within the loop, as shown in figure 6.6.

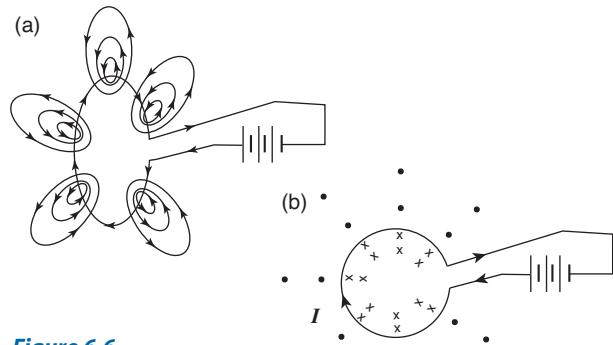


Figure 6.6
The magnetic field of a loop
(a) 3-D representation (b) 2-D representation

- A **solenoid** is a coil of insulated wire that can carry an electric current and is shown in figure 6.7. The number of times that the wire has been wrapped around a tube to make the solenoid is known as the number of ‘turns’ or ‘loops’ of the solenoid. The magnetic fields around each loop of wire add together to produce a magnetic field similar to that of a bar magnet. Note that the magnetic field goes through the centre of the solenoid as well as outside it.

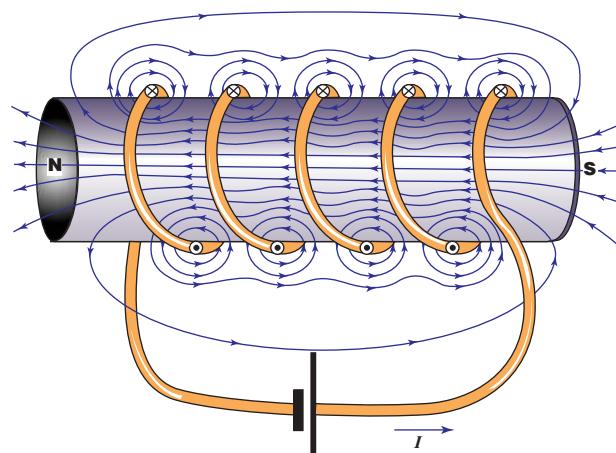


Figure 6.7 The magnetic field around a current-carrying solenoid

- The direction of the magnetic field produced by a solenoid can be determined using another right-hand grip rule; see figure 6.8. In this case, the right hand grips the solenoid with the fingers pointing in the same direction as the conventional current flowing in the loops of wire and the thumb points to the end of the solenoid that acts like the N pole of a bar magnet; that is, the end of the solenoid from which the magnetic field lines emerge.

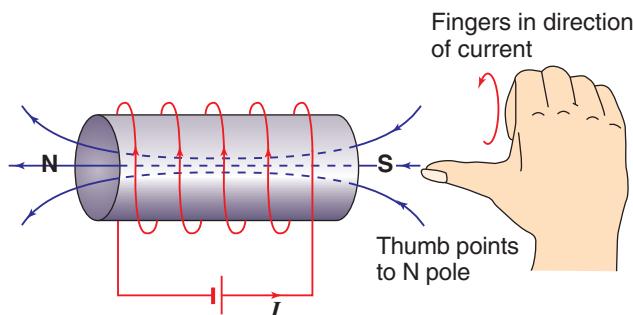


Figure 6.8 Determining the N pole of a solenoid

- Another method for determining the poles of a solenoid is to look at a diagram of the ends of the solenoid (see figure 6.9), and mark in the direction of the conventional current around the solenoid. Then mark on the diagram the letter N or S that has the ends of the letter pointing in the same

A **solenoid** consists of a coil of wire wound uniformly into a cylinder.

direction as the current. N is for an anticlockwise current, S is for a clockwise current.

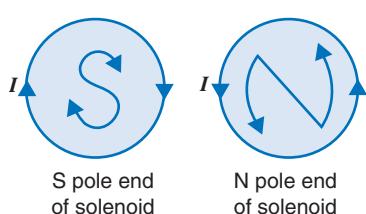
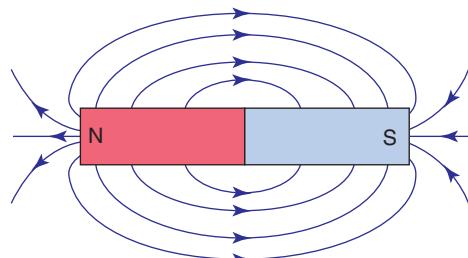
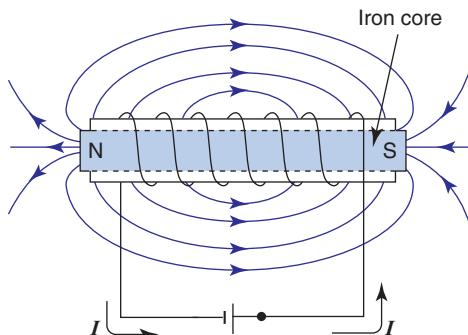


Figure 6.9 Another method for determining the poles of a solenoid

- An electromagnet is a solenoid that has a soft iron core. When a current flows through the solenoid, the iron core becomes a magnet. The polarity of the iron core is the same as the polarity of the solenoid. The core produces a much stronger magnetic field than is produced by the solenoid alone. In figure 6.10 the magnetic field of a permanent magnet is compared to that of an electromagnet.
- The strength of an electromagnet can be increased by:
 - increasing the current through the solenoid
 - adding more turns of wire per unit length for a long solenoid
 - increasing the amount of soft iron in the core.



(a) Permanent magnet



(b) Electromagnet

Figure 6.10 A permanent magnet and an electromagnet. Note the polarity of the iron core.

6.1 THE MOTOR EFFECT

A current-carrying conductor produces a magnetic field. When the current-carrying conductor passes through an external magnetic field, the magnetic field of the conductor interacts with the external magnetic field and the conductor experiences a force. This effect was discovered in 1821 by Michael Faraday (1791–1867) and is known as the **motor effect**. The direction of the force on the current-carrying

The **motor effect** is the action of a force experienced by a current-carrying conductor in an external magnetic field.

The **right-hand push rule** (also called the right-hand palm rule) is used to find the direction of the force acting on a current-carrying conductor in an external magnetic field.



6.1 The motor effect

conductor in an external magnetic field can be determined using the **right-hand push rule** and can be seen in figure 6.11 (discussed further in chapter 7).

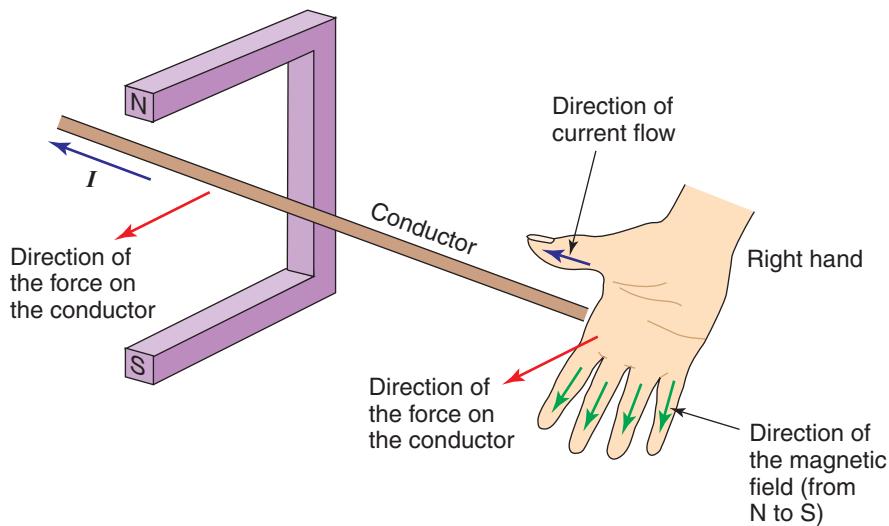


Figure 6.11 The right-hand push rule for a current-carrying conductor

Factors affecting the magnitude of the force

The magnitude of the force on a straight conductor in a magnetic field depends on the following factors:

- the strength of the external magnetic field. The force is proportional to the magnetic field strength, B
- the magnitude of the current in the conductor. The force is proportional to current, I
- the length of the conductor in the field. The force is proportional to the length, l
- the angle between the conductor and the external magnetic field. The force is at a maximum when the conductor is at right angles to the field, and it is zero when the conductor is parallel to the field. The magnitude of the force is proportional to the component of the field that is at right angles to the conductor. If θ is the angle between the field and the conductor, then the force is the maximum value multiplied by the sine of θ .

These factors are shown in figure 6.12 and can be expressed mathematically as:

$$F = BIl \sin \theta.$$

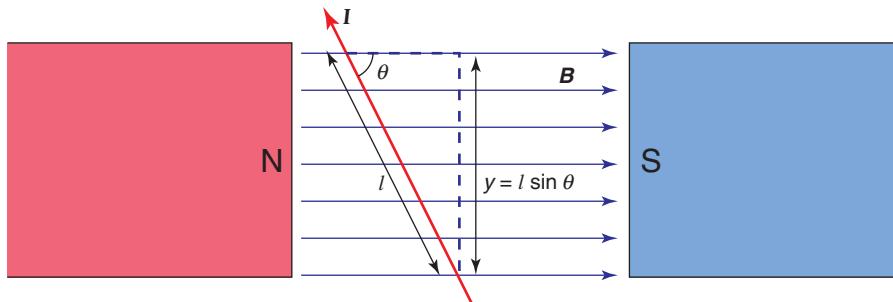


Figure 6.12 A conductor at an angle to a magnetic field

Force on a current-carrying conductor

If a conductor of length 8.0 cm carries a current of 30 mA, calculate the magnitude of the force acting on it when in a magnetic field of strength 0.25 T if:

- the conductor is at right angles to the field
- the conductor makes an angle of 30° with the field
- the conductor is parallel with the field.

SOLUTION

Use the equation:

$$F = BIl \sin \theta$$

where

$$l = 8.0 \times 10^{-2} \text{ m}$$

$$I = 3.0 \times 10^{-2} \text{ A}$$

$$B = 0.25 \text{ T.}$$

- $F = BIl \sin 90^\circ$
 $= 3.0 \times 10^{-2} \times 8.0 \times 10^{-2} \times 0.25 \times 1$
 $= 6.0 \times 10^{-4} \text{ N}$
- $F = BIl \sin 30^\circ$
 $= 3.0 \times 10^{-2} \times 8.0 \times 10^{-2} \times 0.25 \times 0.5$
 $= 3.0 \times 10^{-4} \text{ N}$
- $F = BIl \sin 0^\circ$
 $= 3.0 \times 10^{-2} \times 8.0 \times 10^{-2} \times 0.25 \times 0$
 $= 0$

6.2 FORCES BETWEEN TWO PARALLEL CONDUCTORS

If a finite distance separates two parallel current-carrying conductors, then each conductor will experience a force due to the interaction of the magnetic fields that exist around each.

Figure 6.13, shows the situation where two long parallel conductors carry currents I_1 and I_2 in the same direction.

Figure 6.13(a) shows the magnetic field of conductor 1 in the region of conductor 2. Conductor 2 is cutting through the magnetic field due to conductor 1. The right-hand push rule shows that conductor 2 experiences a force directed towards conductor 1.

Similarly, figure 6.13(b) shows the magnetic field of conductor 2 in the region of conductor 1. The right-hand push rule shows that conductor 1 experiences a force directed towards conductor 2. This means that the conductors are forced towards each other.



6.2

The force between two parallel current-carrying conductors

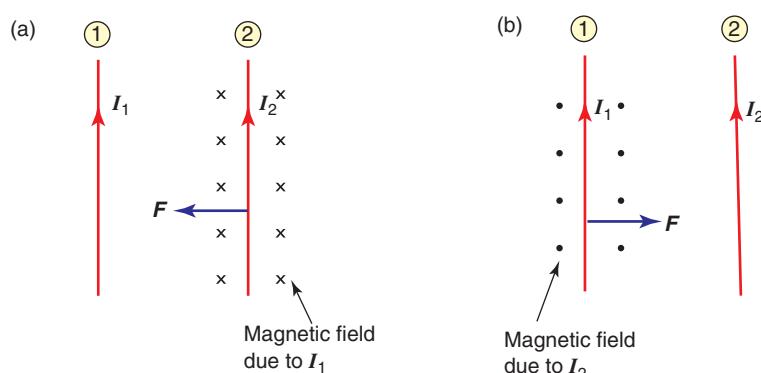


Figure 6.13 The forces acting on two long parallel conductors carrying currents in the same direction

Figure 6.14 shows the situation where two long parallel conductors carry currents I_1 and I_2 in opposite directions.

Figure 6.14(a) shows the magnetic field of conductor 1 in the region of conductor 2. The right-hand push rule shows that conductor 2 experiences a force directed away from conductor 1.

Similarly, figure 6.14(b) shows the magnetic field of conductor 2 in the region of conductor 1. The right-hand push rule shows that conductor 1 experiences a force directed away from conductor 2. This means that the conductors are forced apart.

Note that the magnitude of the forces acting on each pair of wires is equal, but the directions are opposite. This is true even if the conductors carry currents of different magnitudes.

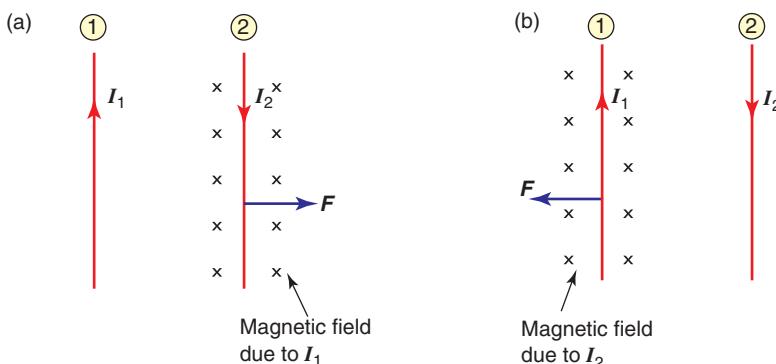


Figure 6.14 The forces acting on two long parallel conductors carrying currents in opposite directions

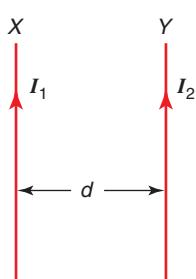


Figure 6.15 Two parallel current-carrying conductors

Determining the magnitude of the force between two parallel conductors

The magnetic field strength at a distance, d , from a long straight conductor carrying a current, I , can be found using the formula:

$$B = \frac{kI}{d}$$

where

$$k = 2.0 \times 10^{-7} \text{ N A}^{-2}.$$

Note that in this equation k is a constant derived from careful experimentation and that d is the perpendicular distance from the wire to the point at which B is to be calculated.

Figure 6.15 shows two parallel conductors, X and Y, that are carrying currents I_1 and I_2 respectively. X and Y are separated by a distance of d .

The magnetic field strength in the region of Y due to the current flowing through X is:

$$B_X = \frac{kI_1}{d}.$$

The magnitude of the force experienced by a length, l , of conductor Y due to the external magnetic field provided by conductor X is:

$$F = I_2 l B_X, \text{ or}$$

$$F = I_2 l \left(\frac{kI_1}{d} \right)$$

This can be rearranged to give the formula:

$$\frac{F}{l} = k \frac{I_1 I_2}{d}$$

A similar process can be used to show that the same formula will give the force experienced by a length, l , of conductor X due to the magnetic field created by the current flowing in conductor Y.

SAMPLE PROBLEM

6.2

SOLUTION

QUANTITY	VALUE
F	?
k	$2.0 \times 10^{-7} \text{ N A}^{-2}$
l	$5.0 \times 10^{-2} \text{ m}$
I_1	3.2 A
I_2	1.2 A
d	0.25 m

Use the equation:

$$\frac{F}{l} = \frac{k I_1 I_2}{d}$$

This transposes to give:

$$\begin{aligned} F &= \frac{k l I_1 I_2}{d} \\ &= \frac{2.0 \times 10^{-7} \times 5.0 \times 10^{-2} \times 3.2 \times 1.2}{0.25} \\ &= 1.5 \times 10^{-7} \text{ N.} \end{aligned}$$

To determine the direction of the force, first find the direction of the magnetic field at X, due to the current in Y, by using the right-hand grip rule. The field is out of the page. Next determine the direction of the force on X using the right-hand push rule. This shows that the force is to the right.

eBook plus

Interactivity:

Torque

int-0049

Lesson:

Torque

eles-0025

6.3

TORQUE

A **torque** can be thought of as the turning effect of a force acting on an object. Examples of this turning effect occur when you turn on a tap, turn the steering wheel of a car, turn the handlebars of a bicycle or loosen a nut using a spanner, as shown in figure 6.16 (on page 108). It is easier to rotate an object if the force, F , is applied at a greater distance, d , from the pivot axis. It is also easier to rotate the object if the force is at right angles to a line joining the pivot axis to its point of application.

The torque, τ , increases when the force, F , is applied at a greater distance, d , from the pivot axis. It is greatest when the force is applied at right angles to a line joining the point of application of the force and the pivot axis.

Torque is the turning effect of a force. It is the product of the tangential component of the force and the distance the force is applied from the axis of rotation.

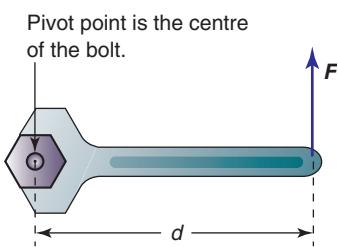


Figure 6.16 A force is applied to a spanner to produce torque on a nut and the spanner.

If the force is perpendicular to the line joining the point of application of the force and the pivot point, the following formula can be used:

$$\tau = Fd.$$

The SI unit for torque is the newton metre (N m).

If the force is not perpendicular to the line joining the point of application of the force and the pivot point, the component of the force that is perpendicular to the line (see figure 6.17) can be used. The magnitude of the torque can then be calculated using the following formula:

$$\tau = Fd \sin \theta$$

where θ is the angle between the force and the line joining the point of application of the force and the pivot axis.

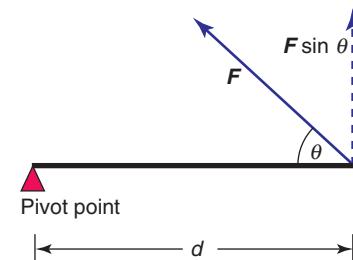


Figure 6.17 Calculating torque when F and d are not perpendicular

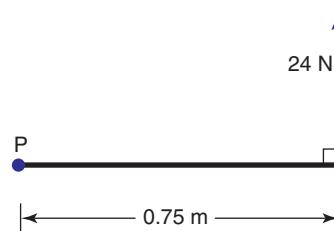
SAMPLE PROBLEM

6.3

Calculating torque

A lever is free to rotate about a point, P. Calculate the magnitude of the torque acting on the lever if a force of 24 N acts at right angles to the lever at a distance of 0.75 m from P. The situation is shown in figure 6.18.

Figure 6.18



SOLUTION

QUANTITY	VALUE
F	24 N
d	0.75 m
τ	?

$$\begin{aligned}\tau &= Fd \\ &= 24 \times 0.75 \\ &= 18 \text{ N m}\end{aligned}$$

SAMPLE PROBLEM

6.4

Calculating torque

What would be the magnitude of the torque in sample problem 6.3 if the force was applied at an angle of 26° to the lever, as shown in figure 6.19?

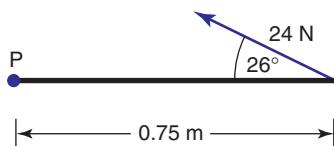


Figure 6.19

SOLUTION

QUANTITY	VALUE
F	24 N
d	0.75 m
θ	26°
τ	?

$$\begin{aligned}\tau &= Fd \sin \theta \\ &= 24 \times 0.75 \times \sin 26^\circ \\ &= 7.9 \text{ N m}\end{aligned}$$

6.4 DC ELECTRIC MOTORS

eBook plus

Weblink:
DC motor applet

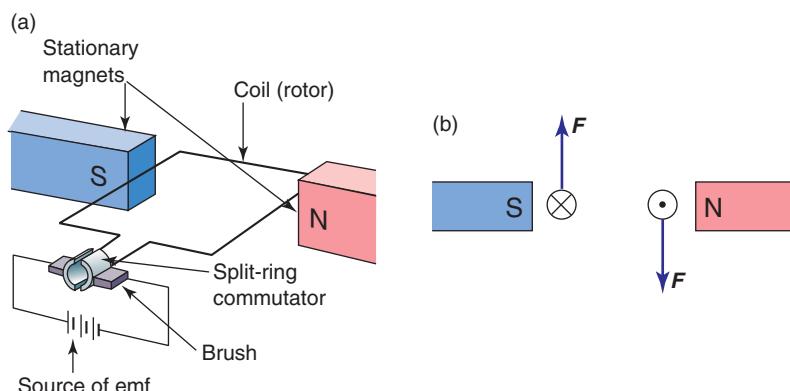
An electric motor is a device that transforms electrical potential energy into rotational kinetic energy. Electric motors produce rotational motion by passing a current through a coil in a magnetic field. Electric motors that operate using direct current (DC) are discussed in this section. The operation of electric motors that use alternating current (AC) is discussed in chapter 9.

Anatomy of a motor

A simplified diagram of a single-turn DC motor is shown in figure 6.20 (which shows only the parts of the DC motor that produce rotational motion).

The magnets provide an external magnetic field in which the coil rotates. As the magnets are fixed to the casing of the motor and are stationary, they are known as the **stator**. The stator sometimes consists of a pair of electromagnets.

The coil carries a direct current. In figure 6.20 the coil has only one loop of wire and this is shown with straight sides. This makes it easier to visualise how forces on the sides come about and to calculate the magnitudes of forces. The coil is wound onto a frame known as an **armature**. This is usually made of ferromagnetic material and it is free to rotate on an axle. The armature and coil together are known as the **rotor**. The armature axle protrudes from the casing, enabling the movement of the coil to be used to do work.



The **stator** is the non-rotating magnetic part of the motor.

The **armature** is a frame around which the coil of wire is wound, which rotates in the motor's magnetic field.

Figure 6.20 (a) The functional parts of a simplified electric motor (b) The direction of current flow in the coil and the direction of the forces acting on the sides

The force acting on the sides of the coil that are perpendicular to the magnetic field can be calculated using the previously discussed formula for calculating the force on a current-carrying conductor in a magnetic field:

$$F = BIl \sin \theta.$$

Real motor rotors have many loops or turns of wire on them. If the coil has n turns of wire on it, then these sides experience a force that is n times greater. In this case:

$$F = nBIl \sin \theta.$$

This extra force increases the torque acting on the sides of the coil.

The split-ring **commutator** and the brushes form a mechanical switch that change the direction of the current through the coil every half turn so that the coil continues rotating in the same direction. The operation of the commutator is discussed in a later section of this chapter.

A **commutator** is a device for reversing the direction of a current flowing through an electric circuit, for example, the coil of a motor.

The source of emf (electromotive force), for example a battery, drives the current through the coil.

How a DC motor operates

Figure 6.21 shows the simplified DC motor at five positions throughout a single rotation. The coil has been labelled with the letters K, L, M and N so that it is possible to observe the motion of the coil as it completes one rotation.

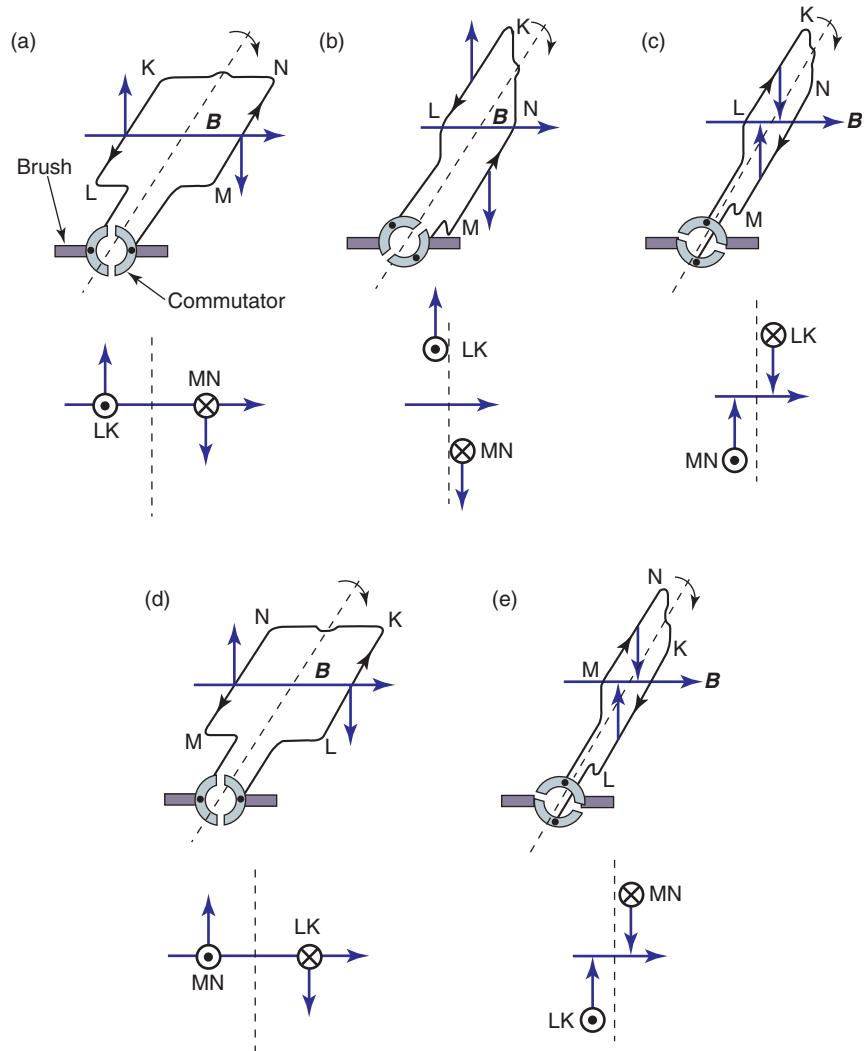


Figure 6.21 Forces acting on the sides of a current-carrying loop. The lower part of the diagram shows cross-sections of the coil.

In figure 6.21(a), the side LK has a force acting on it that is vertically upwards. Side MN has a force of equal magnitude acting on it that is vertically downwards. In this position the forces acting on the sides are perpendicular to the line joining the axle (the pivot line) to the place of application of the force. This means that the torque acting on the coil is at its maximum value. Note that the current is flowing in the direction of K to L.

In figure 6.21(b), the side LK still has a force acting on it that is vertically upwards. Similarly, side MN still has a force of equal magnitude acting on it that is vertically downwards. In this position the forces acting

on the sides are almost parallel to the line joining the axle (the pivot line) to the place of application of the force. This means that the torque acting on the coil is almost zero. It is just after this position that the commutator changes the direction of the current through the coil. The momentum of the coil keeps the coil rotating even though the torque is very small.

Figure 6.21(c) shows the situation when the coil has moved a little further than in the previous diagram, and the current direction through the coil has been reversed. The force acting on side LK is now downwards and the force acting on side MN is now upwards. This changing of direction of the forces and the momentum of the coil enable the coil to keep rotating in the same direction. If the current through the coil did not change its direction of flow through the coil, the coil would rock back and forth about this position. Note that the current is now flowing in the direction of L to K and the torque acting on the coil is still clockwise.

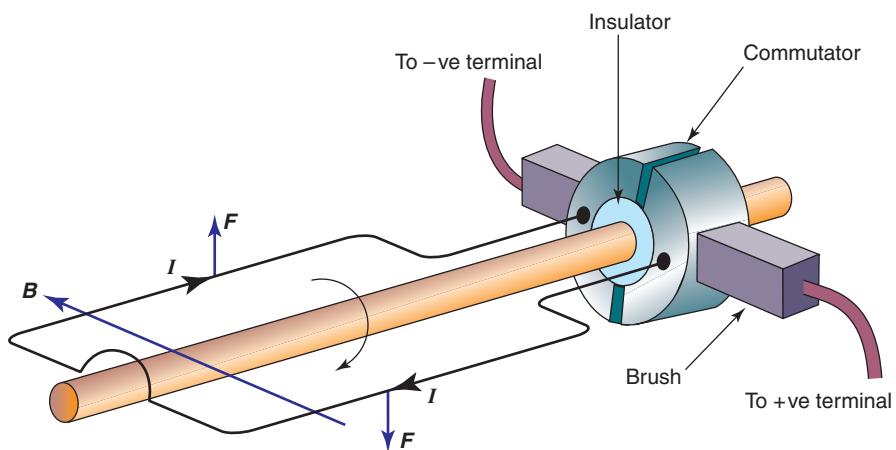
Figure 6.21(d) shows the position of the coil when the torque is again at a maximum value. In this case side MN has the upward force acting on it.

Figure 6.21(e) shows the position of the coil when the torque is again virtually zero and the current has again been reversed. Note that the current is again flowing in the direction of K to L and that there is still a clockwise torque acting on the coil.

The magnitude of the forces acting on sides LK and MN remained constant throughout the rotation just described. However, the torque acting on the coil changed in magnitude.

Commutators

The commutator is a mechanical switch that automatically changes the direction of the current flowing through the coil when the torque falls to zero. Figure 6.22 provides a close-up look at a commutator. It consists of a **split metal ring**, each part of which is connected to either end of the coil. As the coil rotates, first one ring and then the other make contact with a brush. This reverses the direction of the current through the coil. Conducting contacts called **brushes** connect the commutator to the DC source of emf. Graphite, which is used in the brushes, is a form of carbon which conducts electricity and is also used as a lubricant. They are called brushes because they brush against the commutator as it turns. The brushes are necessary to stop the connecting wires from becoming tangled.



A **split metal ring** is the two-piece conducting metal surface of a commutator. Each part is connected to the coil.

The **brushes** are conductors that make electrical contact with the moving split metal ring of the commutator.

Figure 6.22 A close-up look at a split-ring commutator

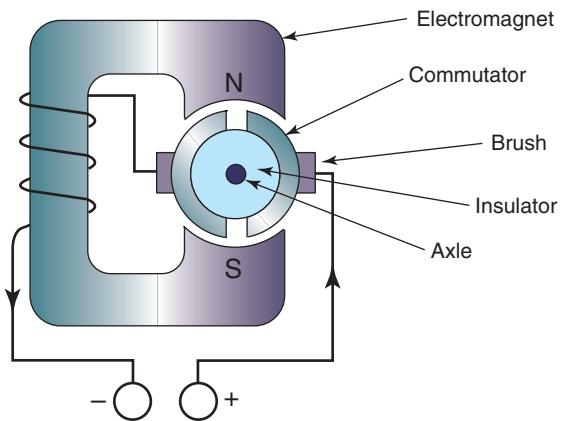
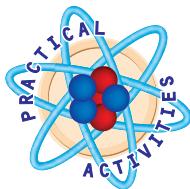


Figure 6.23 Using an electromagnet to provide the magnetic field. Note that the coil is not shown in this diagram!



6.3 A model DC motor

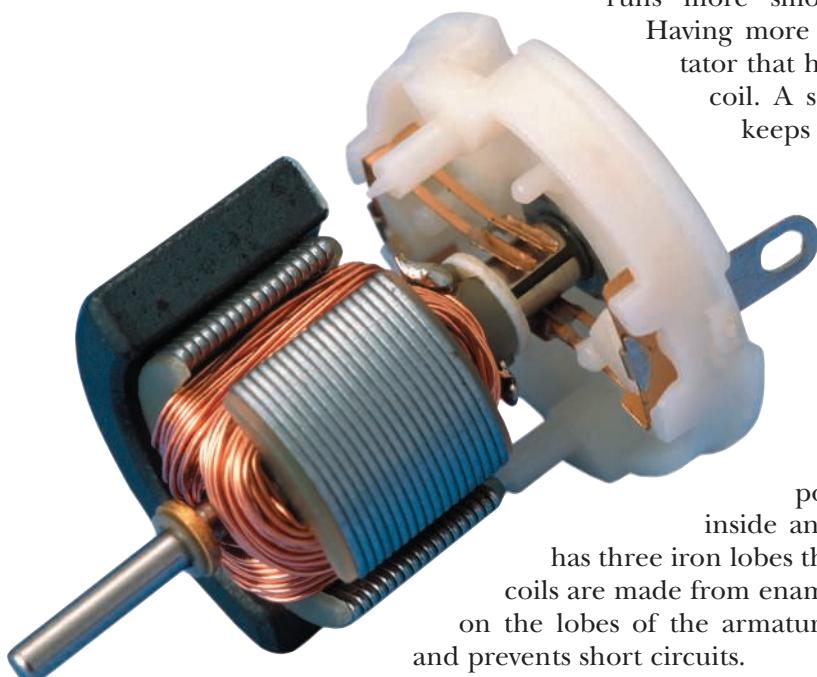


Figure 6.24 A cutaway look at a battery-operated DC motor

The magnetic field in a DC motor

The magnetic field of a DC motor can be provided either by permanent magnets (see figure 6.24) or by electromagnets. The permanent magnets are fixed to the body of the motor. Electromagnets can be created using a soft iron shape that has coils of wire around it. The current that flows through the armature coil can be used in the electromagnet coils. One arrangement for achieving this effect is shown in figure 6.23.

Changing the speed of a DC motor

Increasing the maximum torque acting on the sides can increase the speed of a DC motor. This can be achieved by:

- increasing the force acting on the sides
- increasing the width of the coil
- using more than one coil mounted on the armature.

The force can be increased by

- increasing the current in the coil (this is achieved by increasing the emf across the ends of the coil)
- increasing the number of loops of wire in the coil
- producing a stronger magnetic field with the stator
- using a soft iron core in the centre of the loop. (The core then acts like an electromagnet that changes the direction of its poles when the current changes direction through the coil). The soft iron core is a part of the armature.

Another method used to increase the average torque acting on the coil and armature is to have two or more coils that are wound onto the armature. This arrangement also means that the motor runs more smoothly than a single-coil motor.

Having more than one coil requires a commutator that has two opposite segments for each coil. A stator with curved magnetic poles keeps the force at right angles to the line joining the position of application and the axle for longer. This keeps the torque at its maximum value for a longer period of time.

Figure 6.24 shows many of these features in a small battery-operated DC motor. Note that only one of the stator magnets is shown and that it is curved. The poles of this magnet are on the inside and outside surfaces. The armature has three iron lobes that form the cores of the coils. The coils are made from enamelled copper wire wound in series on the lobes of the armature. The enamel insulates the wire and prevents short circuits.

PHYSICS FACT

- Michael Faraday came up with the idea of an electric motor in 1821.
- The first electric motor was created by accident when two generators were connected together by a worker at the Vienna Exhibition in 1873.
- The French engineer and inventor Zénobe-Théophile Gramme produced the first commercial motors in 1873.
- Direct current (DC) motors were installed in trains in Germany and Ireland in the 1880s.
- Nikola Tesla patented the first significant alternating current (AC) motor in 1888.

eBook plus

Weblink:
Electric motors

Calculating the torque of a coil in a DC motor

Consider a single coil of length, l , and width, w , lying in a magnetic field, \mathbf{B} . The plane of the coil makes an angle, θ , with the magnetic field. The coil carries a current, I , and is free to rotate about a central axis. This situation is shown in figure 6.25.

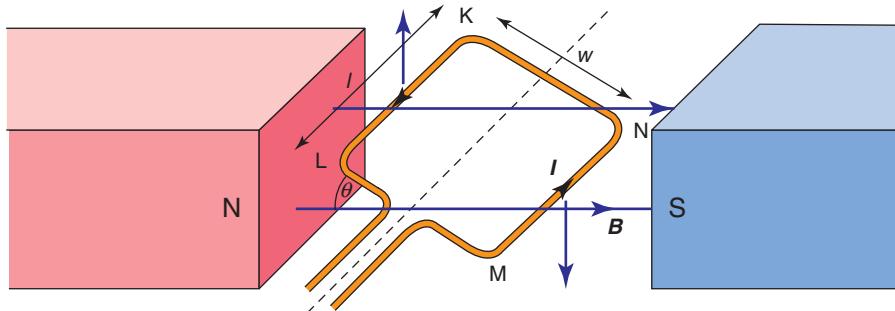


Figure 6.25 The plane of the coil at an angle, θ , to the magnetic field

Side KL experiences a vertically upward force of IlB . Side MN experiences a vertically downward force of IlB .

Both forces exert a clockwise torque on the coil. The magnitude of the torque on each side of the coil is given by:

$$\tau = Fd \sin \phi$$

where ϕ is the angle between side KL or NM and the magnetic field. Note that as ϕ decreases from 90° to 0° , θ , which is now the angle between the plane of the coil and the magnetic field, increases from 0° to 90° .

Also,

$$F = IlB, d = \frac{w}{2} \text{ and } \phi = (90 - \theta)^\circ.$$

Therefore the total torque acting on the coil is given by:

$$\tau = 2 \times IlB \times \frac{w}{2} \times \sin (90 - \theta)^\circ.$$

Since $l \times w = A$, the area of the coil and $\sin (90 - \theta)^\circ = \cos \theta$, the total torque acting on a coil can be expressed as:

$$\tau = BIA \cos \theta.$$

If the coil has n loops of wire on it, the above formula becomes:

$$\tau = nBIA \cos \theta.$$

(Remember that θ is the angle between the *plane of the coil* and the magnetic field.)

Calculating torque on a coil

A coil contains 15 loops and its plane is sitting at an angle of 30° to the direction of a magnetic field of 7.6 mT. The coil has dimensions as shown in figure 6.26 and a 15 mA current passes through the coil. Determine the magnitude of the torque acting on the coil and the direction (clockwise or anticlockwise) of the coil's rotation.

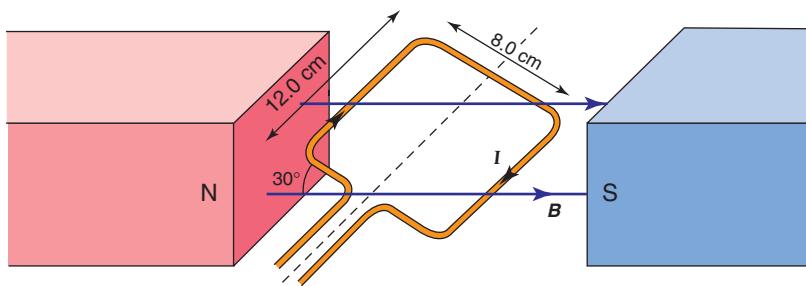


Figure 6.26

Use the relationship $\tau = nBIA \cos \theta$.

SOLUTION

QUANTITY	VALUE
n	15 loops
A	$9.6 \times 10^{-3} \text{ m}^2$
I	$1.5 \times 10^{-2} \text{ A}$
θ	30°
τ	?

$$\begin{aligned}\tau &= 15 \times 7.6 \times 10^{-3} \times 1.5 \times 10^{-2} \times 9.6 \times 10^{-3} \times \cos 30^\circ \\ &= 1.4 \times 10^{-5} \text{ N m}\end{aligned}$$

To determine the direction of rotation of the coil, apply the right-hand push rule to the left-hand side of the coil. This shows that the direction in this case is anticlockwise.

PHYSICS IN FOCUS**The galvanometer**

A galvanometer is a device used to measure the magnitude and direction of small direct current (DC) currents. A schematic diagram of a galvanometer is shown in figure 6.27.

The coil consists of many loops of wire and it is connected in series with the rest of the circuit so that the current in the circuit flows through the coil. When the current flows, the coil experiences a force due to the presence of the external magnetic field (the motor effect). The iron core of the coil increases the magnitude of this force. The needle is rotated until the magnetic force acting on the coil is equalled by a counter-balancing spring. Note that the magnets around the core are curved. This results in a radial magnetic field; the plane of the coil will always be parallel to the magnetic field and the torque will be constant no matter how far the coil is deflected. This also means that the scale of the galvanometer is linear, with the amount of deflection being proportional to the current flowing through the coil.

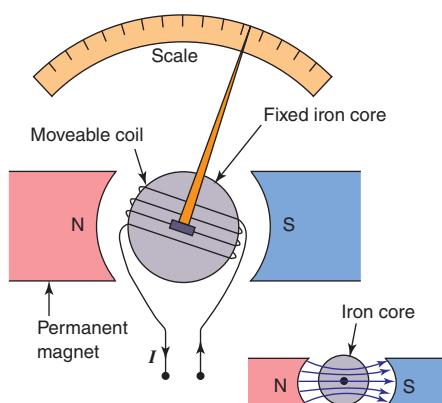


Figure 6.27 The galvanometer

PHYSICS IN FOCUS

Loudspeakers

Loudspeakers are devices that transform electrical energy into sound energy. A loudspeaker consists of a circular magnet that has one pole on the outside and the other on the inside. This is shown in figure 6.28.

A coil of wire (known as the voice coil) sits in the space between the poles. The voice coil is

connected to the output of an amplifier. The amplifier provides a current that changes direction at the same frequency as the sound that is to be produced. The current also changes magnitude in proportion to the amplitude of the sound. The voice coil is caused to vibrate or move in and out of the magnet by the motor effect.

The direction of movement of the voice coil can be determined using the right-hand push rule. This can be shown by examining figure 6.28(b). When the current in the coil is anticlockwise the force on the coil is out of the page. When the current is clockwise, the force on the coil is into the page. The voice coil is connected to a paper speaker cone that creates sound waves in the air as it vibrates. When the magnitude of the current increases, so too does the force on the coil. When the force on the coil increases, it moves more and the produced sound is louder.

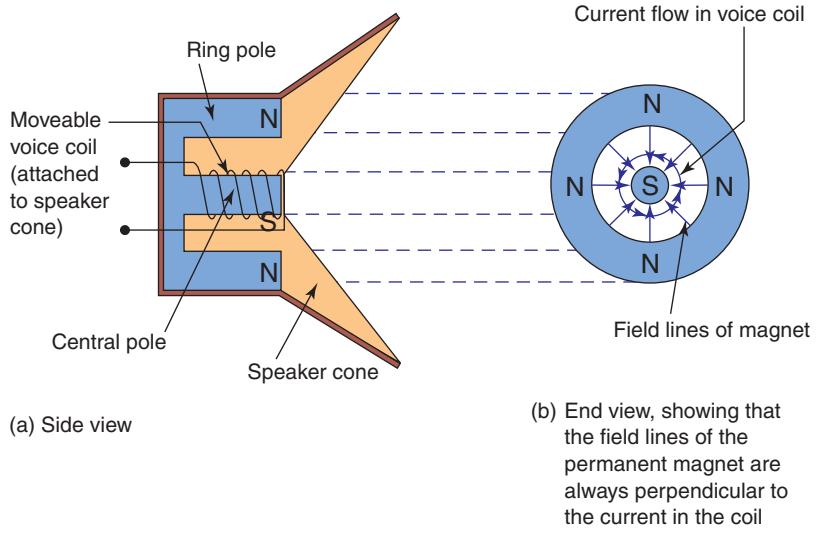


Figure 6.28 A schematic diagram of a loudspeaker

SUMMARY

- According to the motor effect, a current-carrying conductor in a magnetic field will experience a force that is perpendicular to the direction of the magnetic field. The direction of the force is determined using the right-hand push rule.
- The right-hand push rule is applied by:
 - extending the fingers in the direction of the magnetic field
 - pointing the thumb in the direction of the current in the conductor.
 The palm of the hand indicates the direction of the force.
- The magnitude of the force, F , on a current-carrying conductor is proportional to the strength of the magnetic field, B , the magnitude of the current, I , the length, l , of the conductor in the external field and the sine of the angle between the conductor and the field:

$$F = Bl \sin \theta$$

- If the conductor is parallel to the magnetic field, there is no force.
- Two long parallel current-carrying conductors will exert a force on each other. The magnitude of this force is determined using the following formula:

$$\frac{F}{l} = k \frac{I_1 I_2}{d}$$

- If the currents are in the same direction, the conductors attract each other. If the currents are in opposite directions, the conductors repel each other.
- Torque is the turning effect (moment) of a force. The magnitude of the torque is determined using the following formula:

$$\tau = Fd \sin \theta$$

where θ is the angle between the force and the line joining the point of application of the force and the pivot axis.

- The torque acting on the coil of an electric motor is given by the formula:

$$\tau = nBIA \cos \theta$$

where θ is the angle between the plane of the coil and the magnetic field.

- A DC electric motor is one application of the motor effect.
- A DC electric motor has a current-carrying coil that rotates about an axis in an external magnetic field.
- Galvanometers and loudspeakers are other applications of the motor effect.

QUESTIONS

- State the law of magnetic poles.
- Draw a bar magnet and the magnetic field around it. Label the diagram to show that you understand the characteristics of magnetic field lines.
- Are the north magnetic pole of the Earth and the north pole of a bar magnet of the same polarity? Explain your reasoning.
- Figure 6.29 shows three bar magnets and some of the field lines of the resulting magnetic field.
 - Copy and complete the diagram to show the remaining field lines.
 - Label the polarities of the magnets.

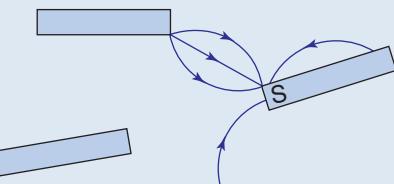


Figure 6.29

- Draw a diagram to show the direction of the magnetic field lines around a conductor when the current is (a) travelling towards you and (b) away from you.
- Each diagram in figure 6.30 represents two parallel current-carrying conductors. In each case, determine whether the conductors attract or repel each other. Explain your reasoning.



Figure 6.30

- Each empty circle in figure 6.31 represents a plotting compass near a coiled conductor. Copy the diagram and label the N and S poles of each coil, and indicate the direction of the needle of each compass.

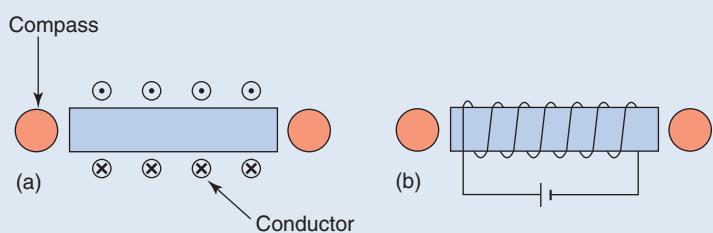


Figure 6.31

- The diagrams in figure 6.32 show electro-magnets. Identify which poles are N and which are S.

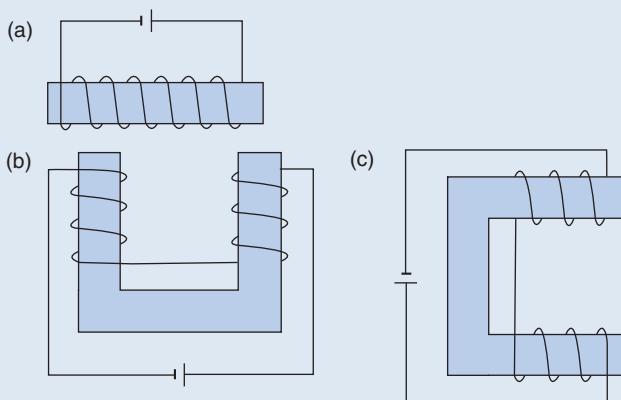


Figure 6.32

9. In figure 6.33 a current-carrying conductor is in the field of a U-shaped magnet. Identify the direction in which the conductor is forced.

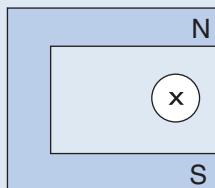


Figure 6.33

10. Identify the direction of the force acting on each of the current-carrying conductors shown in figure 6.34. Use the terms 'up the page', 'down the page', 'into the page', 'out of the page', 'left' and 'right'.

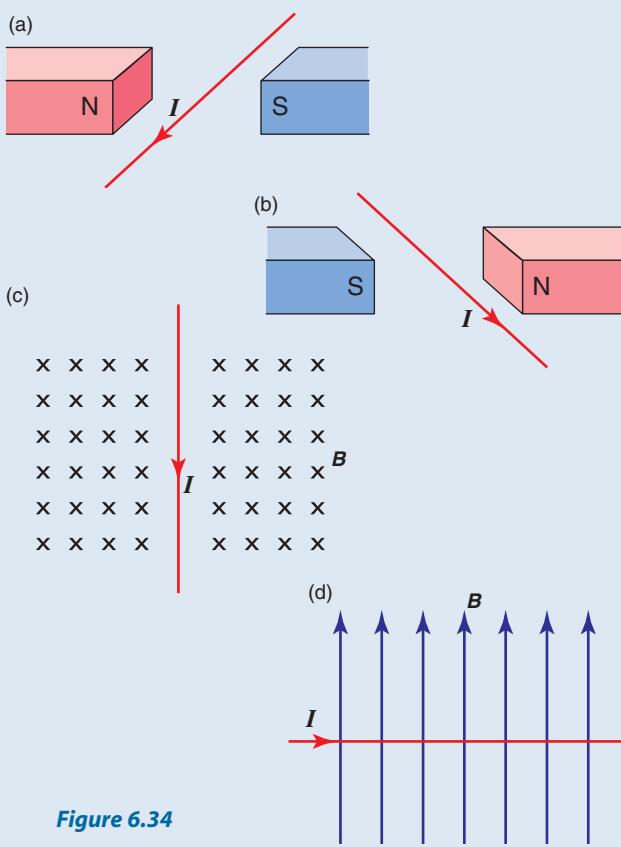


Figure 6.34

11. Deduce both the magnitude and direction of the forces acting on the lengths of conductors shown in figure 6.35.

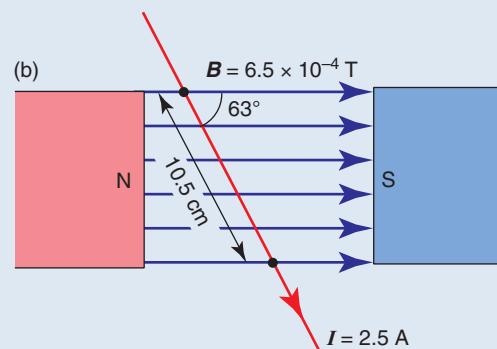
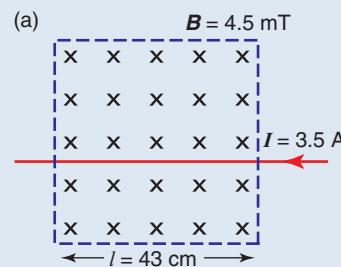


Figure 6.35

12. A student wishes to demonstrate the strength of a magnetic field in the region between the poles of a horseshoe magnet. He sets up the apparatus shown in figure 6.36.

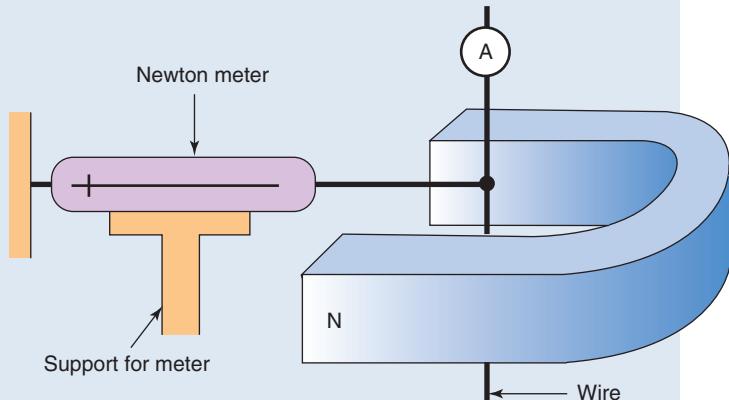


Figure 6.36

The length of wire in the magnetic field is 2.0 cm. When the ammeter reads 1.0 A, the force measured on the newton meter is 0.25 N.

- (a) What is the strength of the magnetic field?
 (b) In this experiment the wire moves to the right. In what direction is the current flowing, up or down the page?

13. A wire with the shape shown in figure 6.37 carries a current of 2.0 A. It lies in a uniform magnetic field of strength 0.60 T.

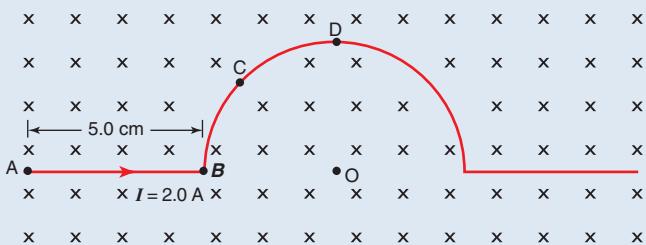


Figure 6.37

- Calculate the magnitude of the force acting on the section of wire, AB.
 - Which of the following gives the direction of the force acting on the wire at the point, C?
 - Into the page
 - Out of the page
 - In the direction OC
 - In the direction CO
 - In the direction OD
 - In the direction DO
 - Which of the following gives the direction of the net force acting on the semicircular section of wire?
 - Into the page
 - Out of the page
 - In the direction OC
 - In the direction CO
 - In the direction OD
 - In the direction DO
14. A wire of length 25 cm lies at right angles to a magnetic field of strength 4.0×10^{-2} T. A current of 1.8 A flows in the wire. Calculate the magnitude of the force that acts on the wire.
15. Two long straight parallel current-carrying wires are separated by 6.3 cm. One wire carries a current of 3.4 A upward and the other carries a current of 2.5 A downward.
- Evaluate the magnitude of the force acting on a 45 cm length of one of the wires.
 - Is the force between the wires attraction or repulsion?
16. Evaluate the magnitude of the force acting on a 40 cm length of one of the two long wires shown in figures 6.38 (a), (b) and (c).

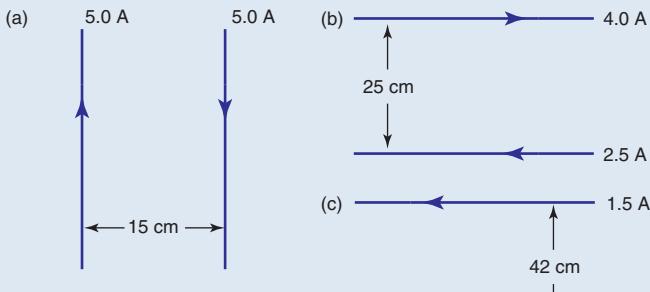


Figure 6.38

17. The diagram in figure 6.39 represents a side view of a single loop in a DC electric motor. Identify the direction of the forces acting on sides A and B of the loop.



Figure 6.39

18. Figure 6.40 shows the functional parts of a type of DC electric motor.

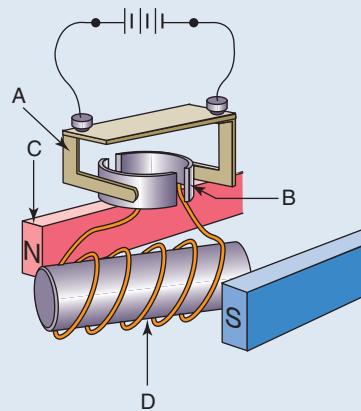


Figure 6.40

- Name the parts labelled A to D in the diagram.
- Describe the functions of the parts labelled A to D.

19. A coil is made up of 50 loops of wire and its plane is at an angle of 45° to the direction of a magnetic field of strength 0.025 T. The coil has the dimensions shown in figure 6.41 and a current of 1.5 A flows through it in the direction shown on the diagram.
- Identify the direction of the force acting on side AB.
 - Calculate the magnitude of the force acting on side AB.
 - Calculate the area of the coil.
 - Calculate the magnitude of the torque acting on the coil when it is in the position shown in the diagram.

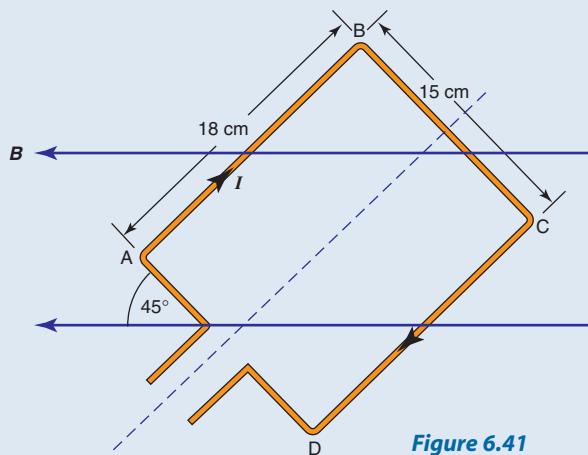


Figure 6.41

20. A student makes a model motor. She makes a rectangular coil with 25 turns of wire with a length of 0.050 m and width 0.030 m. The coil is free to rotate about an axis that is represented by a dotted line in figure 6.42.

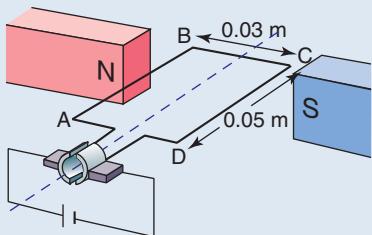


Figure 6.42

At the instant shown the plane of the coil is parallel to the direction of the magnetic field. The magnetic field strength is 0.45 T. When the current to the coil is activated it has a magnitude of 1.75 A in the direction ADCB.

- Calculate the magnitude and direction of the force acting on side CD when the current is flowing and the coil is in the position shown in the diagram.
- When the current begins to flow, the net force acting on the coil is zero yet the coil begins to rotate. Why does this occur?
- Describe what happens to the magnitude and direction of the force acting on side CD as the coil swings through an angle of 60°.
- Describe three things the student could do to get the coil to rotate at a faster rate.



6.1 THE MOTOR EFFECT

Aim

To observe the direction of the force on a current-carrying conductor in an external magnetic field.

Apparatus

variable DC power supply
variable resistor ($15\ \Omega$ rheostat)
connecting wires
retort stand
clamp
two pieces of thick card or balsa wood $10\text{ cm} \times 10\text{ cm}$
strip of aluminium foil $1\text{ cm} \times 30\text{ cm}$ (approximately)
two drawing pins
switch
horseshoe magnet

Method

1. Pin the foil strip between the pieces of card. Rest one card on the bench-top and support the other with the clamp and retort stand.
2. Connect the wires to the power pack's DC terminals, switch, variable resistor and strips as shown in figure 6.43. This will produce a current in the strip.

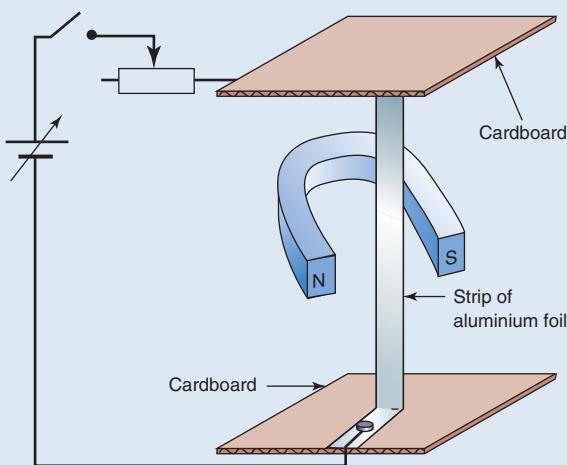


Figure 6.43 The set-up for the motor effect activity

3. Position the horseshoe magnet so that the strip is between the poles. Note the position of the poles of the magnet and the direction of the current through the strip when the switch is closed.
4. Set the power pack to its lowest value and turn it on.

5. Briefly close the switch and record the movement of the foil strip.
6. Turn the magnet over so that the magnetic field is in the opposite direction across the strip.
7. Briefly close the switch and record the movement of the foil strip.

Analysis

1. Did the strip experience a force when a current flowed?
2. Verify that the movement of the aluminium strip is in accordance with the right-hand push rule.



6.2 THE FORCE BETWEEN TWO PARALLEL CURRENT-CARRYING CONDUCTORS

Aim

To observe the direction of the forces between two parallel current-carrying conductors.

Apparatus

variable DC power supply
variable resistor ($15\ \Omega$ rheostat)
connecting wires
retort stand
clamp
two pieces of thick card or balsa wood $10\text{ cm} \times 10\text{ cm}$
two strips of aluminium foil $1\text{ cm} \times 30\text{ cm}$
(approximately)
four drawing pins
push switch

Method

1. Pin each foil strip between the pieces of card so that they are parallel when the top card is supported by the clamp and retort stand.
2. Connect the wires to the power pack's DC terminals, switch, variable resistor and strips as shown in figure 6.44(a). This will produce currents in the strips that are flowing in opposite directions.
3. Set the power pack to its lowest value and turn it on.
4. Briefly close the switch and record the movement of the foil strips.

5. Connect the wires to the power pack's DC terminals, switch, variable resistor and strips as shown in figure 6.44(b). This will produce currents in the strips that are flowing in the same direction.
6. Briefly close the switch and record the movement of the foil strips.

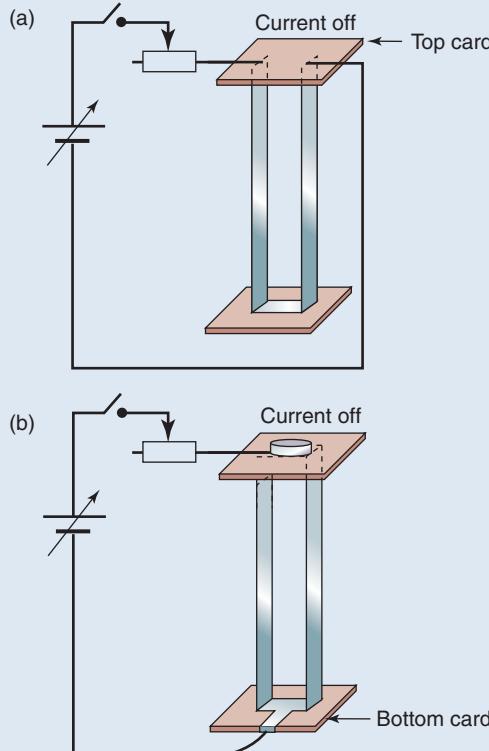


Figure 6.44 (a) The set-up for currents flowing in opposite directions (b) The set-up for currents flowing in the same direction

Analysis

Account for your observations.

6.3 A MODEL DC MOTOR

Aim

To build a model DC motor.

Apparatus

6 V, DC power supply
two bar magnets
thin insulated copper wire
two cylindrical corks, one thin and one thick
aluminium foil
glue

adhesive tape
five bamboo skewers
thick rectangular piece of polystyrene

Method

1. Push one of the skewers through the centres of the corks as shown in figure 6.45.
2. Glue two pieces of foil onto the small cork with two thin gaps between them to make a split ring commutator.
3. Wrap the thin copper wire around the thick cork 50 times, as shown. Hold in place with adhesive tape.
4. Strip the ends of the wire and connect to each of the foil strips of the commutator.
5. Make sure that the centres of the commutator strips line up with the centre of the coil windings.
6. Push the other skewers into the foil to support the coil and commutator, as shown.
7. Use the foil to make a set of brushes and use drawing pins to position them so that they touch opposite sides of the commutator, as shown.
8. Position the two magnets on opposite sides of the coil so that a N pole faces a S pole.
9. Connect the DC supply to the brushes and observe the motion of the armature.
10. Vary the spacing of the magnets.
11. Find two ways to reverse the direction of rotation of the armature.

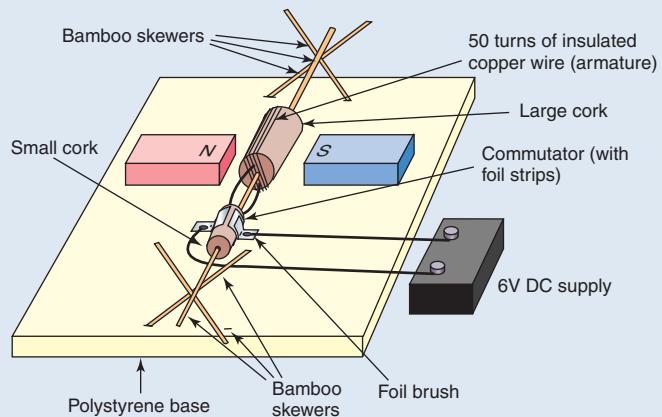


Figure 6.45 The set-up for a model motor

Analysis

1. Describe the effect of varying the spacing of the magnets on the speed of rotation of the armature. Account for this effect.
2. Describe two methods for reversing the direction of rotation of the armature.
3. Identify the role of the following:
 - magnets
 - coil.

CHAPTER

7

GENERATING ELECTRICITY



Figure 7.1 Faraday's magnetic laboratory, 1852
(watercolour on paper,
by Harriet Jane Moore (1801–1884))

Remember

Before beginning this chapter, you should be able to:

- define current, I , as the rate of flow of charge, Q , in a circuit
- use the formula $Q = It$
- recall that the potential difference, V , is the amount of energy transformed in a circuit element, per coulomb of charge passing through the circuit element
- recall that the emf (E or \mathcal{E}) is the amount of energy supplied by a source, per coulomb of charge passing through the source
- state Ohm's Law for metal conductors at a constant temperature, $V = IR$.

Key content

At the end of this chapter you should be able to:

- outline Michael Faraday's discovery of the generation of an electric current by a moving magnet
- define the magnetic field strength, B , as magnetic flux density
- describe the concept of magnetic flux in terms of magnetic flux density and surface area
- describe generated potential difference as the rate of change of magnetic flux through a circuit
- account for Lenz's Law in terms of conservation of energy, and relate it to the production of back emf in motors
- explain that, in electric motors, back emf opposes the supply emf
- explain the production of eddy currents in terms of Lenz's Law.

7.1 THE DISCOVERIES OF MICHAEL FARADAY

Michael Faraday (1791–1867) was the son of an English blacksmith. He started his working life at the age of twelve as an errand boy at a bookseller's store and later became a bookbinder's assistant. At the age of nineteen he attended a series of lectures at the Royal Institution in London that were given by Sir Humphrey Davey. This led to Faraday studying chemistry by himself. In 1813 he applied to Davey for a job at the Royal Institution and was hired as a research assistant. He soon showed his abilities as an experimenter and made important contributions to the understanding of chemistry, electricity and magnetism. He later became the superintendent of the Royal Institution.

In September 1821, following the 1820 discovery by Hans Christian Oersted (1777–1851) that an electric current produces a magnetic field, Michael Faraday discovered that a current-carrying conductor in a magnetic field experiences a force. This became known as the motor effect (see chapter 6).

Almost 10 years later, in August 1831, Faraday discovered **electromagnetic induction**. This is the generation of an emf and/or electric current through the use of a magnetic field. Faraday's discovery was not accidental. He and other scientists spent many years searching for ways to produce an electric current using a magnetic field. Faraday's breakthrough eventually led to the development of the means of generating electrical energy in the vast quantities that we use in our society today.

Electromagnetic induction is the generation of an emf and/or electric current through the use of a magnetic field.

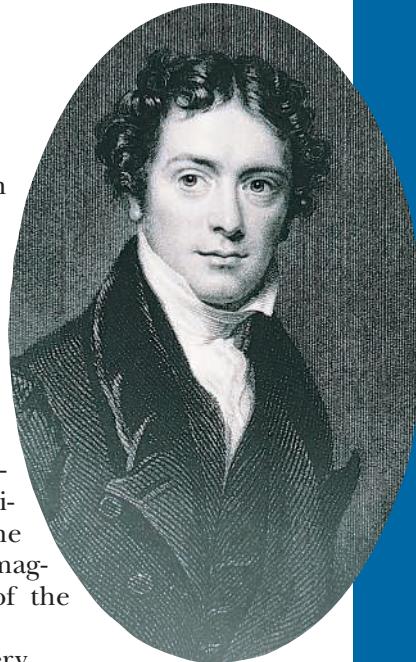


Figure 7.2
1830 portrait of Michael Faraday

PHYSICS FACT

Joseph Henry

American Joseph Henry (1797–1878) seems to have observed an induced current before Faraday, but Faraday published his results first and investigated the subject in more detail.

First experiments

In his first successful experiment, Faraday set out to produce and detect a current in a coil of wire by the presence of a magnetic field set up by another coil. He appears to have coiled about 70 m of copper wire around a block of wood. A second length of copper wire was then coiled around the block in the spaces between the first coil. The coils were separated with twine. One coil was connected to a **galvanometer** and the other to a battery. (A galvanometer is an instrument for detecting small electric currents. Faraday's early efforts to detect an induced current failed because of the lack of sensitivity of his galvanometers.) A simplified diagram of this experiment is shown in figure 7.3, on the following page.

When the battery circuit (or primary circuit) was closed, Faraday observed 'there was a sudden and very slight effect [deflection] at the

A **galvanometer** is an instrument for detecting small electrical currents.

galvanometer.' This means that Faraday had observed a small brief current that was created in the galvanometer circuit (or secondary circuit). A similar effect was also produced when the current in the battery circuit was stopped, but the momentary deflection of the galvanometer needle was in the opposite direction.

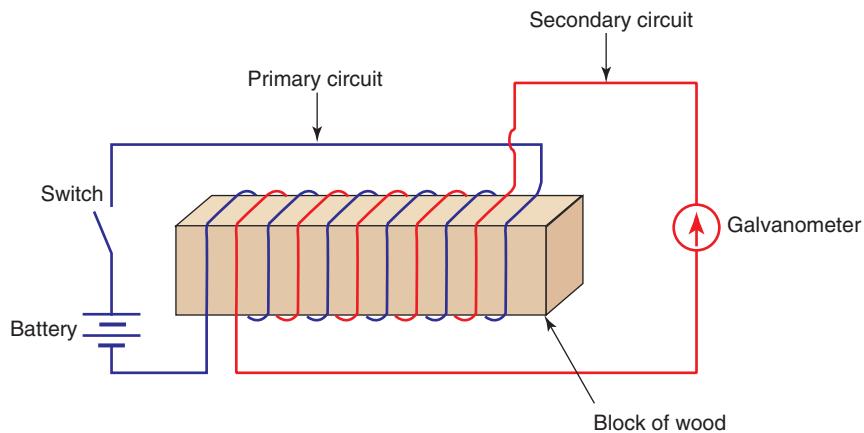


Figure 7.3 A simplified diagram of Faraday's first experiment

Faraday was careful to emphasise that the current in the galvanometer circuit was a temporary one and that no current existed when the current in the battery circuit was at a constant value.

Faraday modified this experiment by winding the secondary coil around a glass tube. He placed a steel needle in the tube and closed the primary circuit. He then removed the needle and found that it had been magnetised. This also showed that a current had been produced (induced) in the secondary circuit. It was the magnetic field of the induced current in the secondary circuit that had magnetised the needle.

The next experiment was to place a steel needle in the secondary coil when a current was flowing in the primary coil. The primary current was stopped and the needle was again found to be magnetised, but with the poles reversed to the direction of the first experiment.

Iron ring experiment

In a further experiment, Faraday used a ring made of soft iron (see figure 7.4). He wound a primary coil on one side and connected it to a battery and switch. He wound a secondary coil on the other side and connected it to a galvanometer. A simplified diagram of this experiment is shown in figure 7.5 on the opposite page.



Figure 7.4 Photograph of the apparatus (two coils of insulated copper wire wound around an iron ring) that Faraday used to induce an electric current on 29 August 1831

Faraday's iron ring apparatus, an iron ring with a primary and secondary coil wrapped around it, is the basis of modern electrical transformers.

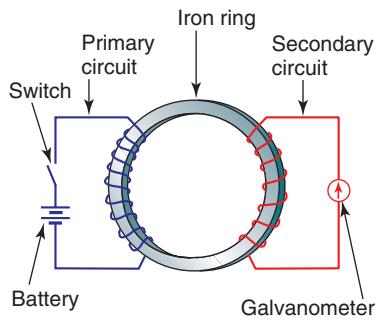


Figure 7.5 A simplified diagram of Faraday's iron ring experiment

When the current was set up in the primary coil, the galvanometer needle immediately responded, as Faraday stated, 'to a degree far beyond what has been described when the helices [coils] without an iron core were used, but although the current in the primary was continued, the effect was not permanent, for the needle soon came to rest in its natural position, as if quite indifferent to the attached electromagnet'. When the current in the primary coil was stopped, the galvanometer needle moved in the opposite direction. He concluded that when the magnetic field of the primary coil was changing, a current was induced in the secondary coil.

Using a moving magnet

Faraday was also able to show that moving a magnet near a coil could generate an electric current in the coil. The diagrams in figure 7.6 show the effect when the N pole of a magnet is brought near a coil, held stationary, and then taken away from the coil.

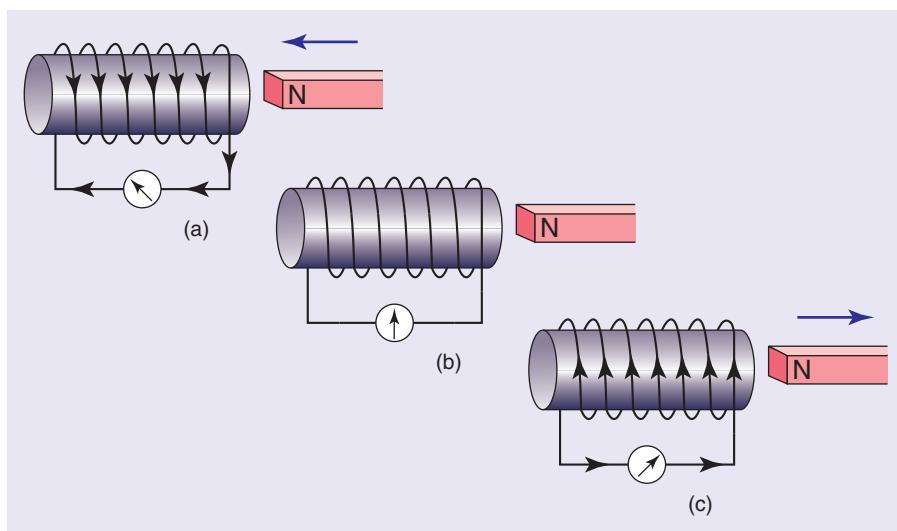


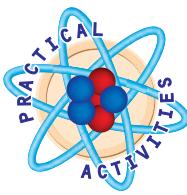
Figure 7.6 (a) When the N pole of a bar magnet is brought near one end of the coil, the galvanometer needle momentarily deflects in one direction, indicating that a current has been induced in the coil circuit. (b) When the magnet is held without moving near the end of the coil, the needle stays at the central point of the scale (no deflection), indicating no induced current. (c) When the N pole of the magnet is taken away from the coil, the galvanometer needle momentarily deflects in the opposite direction to the first situation, indicating that an induced current exists and that it is flowing in the opposite direction.

Similar results occur when the S pole is moved near the same end of the coil, except that the deflection of the galvanometer needle is in the opposite direction to when the N pole moves in the same direction.



7.1

Inducing current in a coiled conductor



7.2

Linking coils

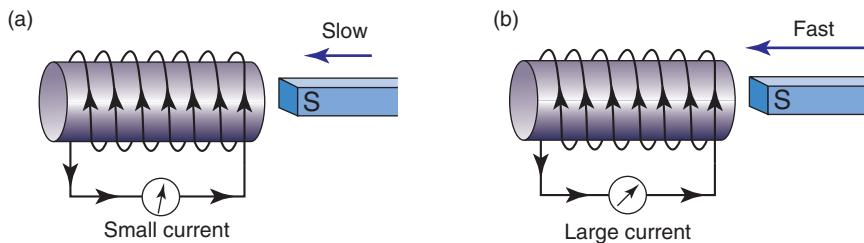


Figure 7.7 (a) A slow-moving magnet induces a small current.
(b) A fast-moving magnet induces a larger current.

You can repeat some of Faraday's experiments by doing practical activities 7.1 and 7.2.

7.2

ELECTROMAGNETIC INDUCTION

Induction can be defined as a process where one object with magnetic or electrical properties can produce the same properties in another object without making physical contact.

The word ‘flux’ comes from the Latin word *fluo* meaning ‘flow’. Flux is a state of flowing or movement. In physics, flux is the rate of flow of a fluid, radiation or particles.

Magnetic flux, Φ_B , is the amount of magnetic field passing through a given area. In the SI system, Φ_B is measured in weber (Wb).

The strength of a magnetic field, B , is also known as the **magnetic flux density**. In the SI system, B is measured in tesla (T) or weber per square metre (Wb m^{-2}).

Electromagnetic **induction** is the creation of an emf in a conductor when it is in relative motion to a magnetic field, or it is situated in a changing magnetic field. Such an emf is known as an induced emf. In a closed conducting circuit, the emf gives rise to a current known as an induced current.

Faraday demonstrated that it was possible to produce (or induce) a current in a coil by using a changing magnetic field. For there to be a current in the coil, there must have been an emf induced in the coil. We will now examine how this is achieved.

Magnetic flux

The magnetic field in a region can be represented diagrammatically using field or **flux** lines. You can imagine the magnetic field ‘flowing’ out from the N pole of a magnet, spreading out around the magnet and then ‘flowing’ back into the magnet through the S pole. The field lines on a diagram show the direction of magnetic force experienced by the N pole of a test compass if it were placed at that point. The closeness (or density) of the lines represents the strength of the magnetic field. The closer together the lines, the stronger the field.

Magnetic flux is the name given to the amount of magnetic field passing through a given area. It is given the symbol Φ_B . In the SI system, Φ_B is measured in weber (Wb). If the particular area, A , is perpendicular to a uniform magnetic field of strength B (as shown in figure 7.8 on the opposite page) then the magnetic flux Φ_B is the product of B and A .

$$\Phi_B = BA$$

The strength of a magnetic field, B , is also known as the **magnetic flux density**. It is the amount of magnetic flux passing through a unit area. In the SI system, B is measured in tesla (T) or weber per square metre (Wb m^{-2}).

The magnetic flux, Φ_B , passing through an area is reduced if the magnetic field is not perpendicular to the area, and Φ_B is zero if the magnetic field is parallel to the area. The above relationship between magnetic flux, magnetic flux density and area is often written as:

$$\Phi_B = B_{\perp}A$$

where B_{\perp} is the component of the magnetic flux density that is perpendicular to the area, A .

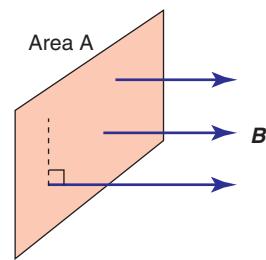


Figure 7.8 The magnetic field passing through an area at right angles

7.3 GENERATING A POTENTIAL DIFFERENCE

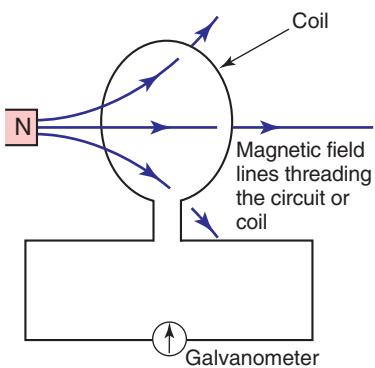


Figure 7.9 A galvanometer circuit showing magnetic flux threading the coil

For a current to flow through the galvanometer in Faraday's experiments there must be an electromotive force (emf, symbol E or \mathcal{E}). The magnitude of the current through the galvanometer depends on the resistance of the circuit and the magnitude of the emf generated in the circuit.

Faraday noted that there had to be change occurring in the apparatus for an emf to be created. The quantity that was changing in each case was the amount of magnetic flux threading (or passing through) the coil in the galvanometer circuit (see figure 7.9). The rate at which the magnetic flux changes determines the magnitude of the generated emf.

This gives Faraday's Law of Induction, which can be stated as follows:

The induced emf in a circuit is equal in magnitude to the rate at which the magnetic flux through the circuit is changing with time.

Faraday's law can be written in equation form as:

$$\mathcal{E} = -\frac{\Delta \Phi_B}{\Delta t}.$$

The negative sign in the above equation indicates the direction of the induced emf. This is explained in section 7.4, Lenz's Law (see page 128).

PHYSICS FACT

The symbol for the Greek letter delta is Δ . It is used in mathematics and physics to represent a change in a quantity.

The change in a quantity is calculated by subtracting the initial value from the final value. For example, the change in your bank balance over a month is the final balance minus your initial balance.

When calculating quantities using Faraday's Law of Induction,

$$\Delta \Phi_B = \Phi_{B\text{final}} - \Phi_{B\text{initial}}$$

Since $\Phi_B = B_{\perp}A$, a change in Φ_B can be caused by a change in the magnetic field strength, B , or in the area of the coil that is perpendicular to the magnetic field, or both.

If a coil has n turns of wire on it, the emf induced by a change in the magnetic flux threading the coil would be n times greater than that produced if the coil had only one turn of wire.

Rotating coils in uniform magnetic fields

When a coil rotates in a magnetic field, as occurs in generators and motors, the flux threading the coil is a maximum when the plane of the coil is perpendicular to the direction of the magnetic field. If the plane of the coil is parallel to the direction of the magnetic field, the flux threading the coil is zero, so rotating the coil changes the magnetic flux.

7.4 LENZ'S LAW

H. F. Lenz (1804–1864) was a German scientist who, without knowledge of the work of Faraday and Henry, duplicated many of their experiments. Lenz discovered a way to predict the direction of an induced current. This method is given the name Lenz's Law. It can be stated in the following way:

An induced emf always gives rise to a current that creates a magnetic field that opposes the original change in flux through the circuit.

This is a consequence of the Principle of Conservation of Energy. The minus sign in Faraday's Law of Induction is placed there to remind us of the direction of the induced emf.

Using Lenz's Law

When determining the direction of the induced emf, it is useful to use the field line method for representing magnetic fields. Figure 7.10 shows the effect of a magnet moving closer to a coil connected to a galvanometer. The coil is wound on a cardboard tube. As the magnet approaches the coil, the magnetic flux density within the coil increases. The induced current sets up a magnetic field (shown in dotted lines) that opposes this change. The approaching magnet increases the number of field lines pointing to the left that pass through the coil. The induced current in the coil produces field lines that point to the right to counter this increase.

The direction of the induced current in the coil can be deduced using the right-hand rule for coils. The thumb points in the direction of the induced magnetic field within the coil, the curl of the fingers holding the coil show the direction of the induced current in the coil. Note that magnetic field lines do not cross. Dotted lines have been used to show the general direction of the induced field lines, not the resultant field.



7.3

The direction of induced currents

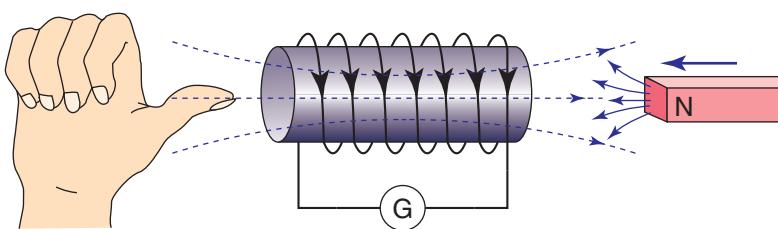


Figure 7.10 The N pole of a magnet approaches a coil. Note that the induced magnetic field of the coil repels the approaching N pole.

Induced current in a coil

A metal ring initially lies in a uniform magnetic field, as shown in figure 7.11. The ring is then removed from the magnetic field. In which direction does the induced current flow in the coil?

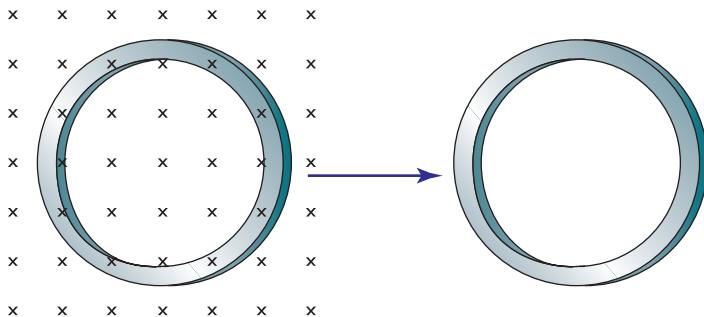


Figure 7.11

SOLUTION

Initially the magnetic field lines of the external field are passing into the page through the coil. As the coil is removed from the field, these field lines reduce in number. The induced current flows in such a way as to create a magnetic field to replace the missing lines. Therefore, the current in the ring must flow in a clockwise direction in the ring, as indicated by using the right-hand rule for coils. The current stops flowing when the entire ring has been removed from the external magnetic field.

eBook plus

Interactivity:
Magnetic flux and
Lenz's Law
int-0050

eLesson:
Magnetic flux and
Lenz's Law
eles-0026

Lenz's Law and the Principle of Conservation of Energy

What would happen if the opposite of Lenz's Law were true? That is, if a changing flux in a coil would produce a magnetic flux in the same direction as the original change of flux. This would lead to a greater change in flux threading the coil, which in turn would lead to an even greater change in flux. The induced current would continue to increase in magnitude, fed by its own changing flux. In fact we would be creating energy without doing any work. This clearly cannot occur.

The Principle of Conservation of Energy states:

Energy cannot be created nor destroyed, but it can be transformed from one form to another.

To create electrical energy in a coil, work must be done. Energy is required to move a magnet towards or away from a coil. Some of this energy is transformed into electrical energy in the coil.

Lenz's Law and the production of back emf in motors

Electric motors use an input voltage to produce a current in a coil to make the coil rotate in an external magnetic field. It has been shown that an emf is induced in a coil that is rotating in an external magnetic field. The emf is produced because the amount of the magnetic flux that is threading the coil is constantly changing as the coil rotates. The emf induced in the motor's coil, as it rotates in the external magnetic field, is in the opposite direction to the input voltage or supply emf. If this was not the case, the current would increase and the motor coil would go faster and faster forever. The induced emf produced by the rotation of a motor coil is known as the **back emf** because it is in the opposite direction to the supply emf.

The net voltage across the coil equals the input voltage (or supply emf) minus the back emf. If there is nothing attached to an electric motor to

Back emf is an electromagnetic force that opposes the main current flow in a circuit. When the coil of a motor rotates, a back emf is induced in the coil due to its motion in the external magnetic field.

SAMPLE PROBLEM**7.2****Currents in electric motors**

The armature winding of an electric motor has a resistance of 10Ω . The motor is connected to a 240 V supply. When the motor is operating with a normal load, the back emf is equal to 232 V.

- What is the current that passes through the motor when it is first started?
- What is the current that passes through the motor when it is operating normally?

SOLUTION

- (a) When the motor is first started, there is no back emf. The voltage drop across the motor is 240 V.

$$V = IR$$

$$I = \frac{V}{R}$$

$$= \frac{240}{10}$$

$$= 24 \text{ A}$$

QUANTITY	VALUE
V	240 V
R	10Ω
I	?

- (b) When the motor is operating normally, the voltage drop across the motor equals the input voltage minus the back emf.

$$\text{So } V = 240 \text{ V} - 232 \text{ V} = 8 \text{ V.}$$

QUANTITY	VALUE
V	8 V
R	10Ω
I	?

$$V = IR$$

$$I = \frac{V}{R}$$

$$= \frac{8}{10}$$

$$= 0.8 \text{ A}$$

This example shows that electric motors require large currents when starting compared with when they are operating normally.

7.5 EDDY CURRENTS

Charged particles moving in magnetic fields

The **right-hand grip rule** is used to find the direction of a magnetic field around a straight current-carrying conductor.

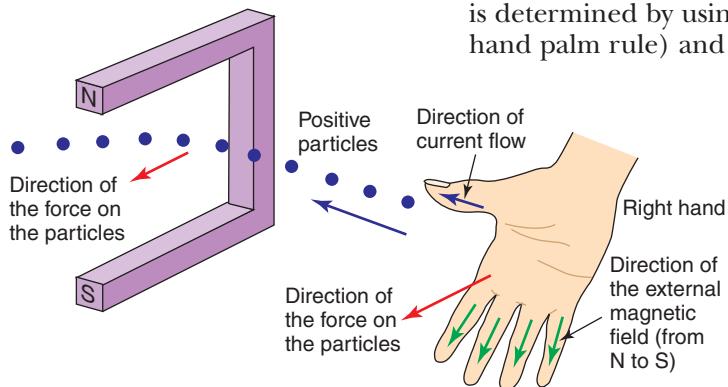


Figure 7.12 The right-hand push rule for moving charged particles

An **eddy current** is a circular or whirling current induced in a conductor that is stationary in a changing magnetic field, or that is moving through a magnetic field. They resemble the eddies or swirls left in the water after a boat has gone by.

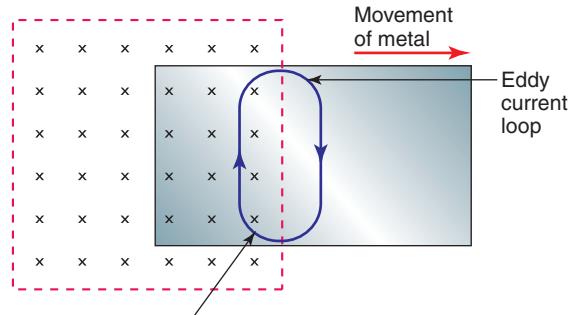


Figure 7.13 The production of eddy currents in a sheet of metal

Moving charged particles, for example electrons or alpha particles, produce magnetic fields. The direction of the magnetic field is found using the **right-hand grip rule** (see the boxed section *Review of magnetic fields* on page 101).

When moving charged particles enter an external magnetic field, the magnetic field created by the moving charged particles interacts with the external magnetic field. (An external magnetic field is one that already exists or that is caused by another source.)

When the moving charged particles enter the magnetic field at right angles to the field, they experience a force that is at right angles to the velocity and to the direction of the external field. The direction of the force is determined by using the right-hand push rule (also known as the right-hand palm rule) and is demonstrated in figure 7.12.

To use the right-hand push rule, position your right hand so that:

- the fingers point in the direction of the external field
- the thumb points in the direction of conventional current flow (this means in the direction of the velocity of positive charges or in the opposite direction to the velocity of negative charges)
- the direction of the force on the particles is directly away from the palm of the hand.

Magnetic fields and eddy currents

Induced currents do not occur in only coils and wires. They can also occur:

- when there is a magnetic field acting on part of a metal object and there is relative movement between the magnetic field and the object
- when a conductor is moving in an external magnetic field
- when a metal object is subjected to a changing magnetic field.

Such currents are known as **eddy currents**. Eddy currents are an application of Lenz's Law. The magnetic fields set up by the eddy currents oppose the changes in the magnetic field acting in the regions of the metal objects.

Figure 7.13 shows one method of production of an eddy current. A rectangular sheet of metal is being removed from an external magnetic field that is directed into the page. On the left of the edge of the magnetic field charged particles in the metal sheet experience a force because they are moving relative to the magnetic field.

By applying the right-hand push rule, it can be seen that positive charges experience a force up the page in this region. To the right of the edge of the magnetic field charged particles experience no force. Therefore the charged particles that are free to move at the edge of the field contribute to an upward current that is able to flow downward in the metal that is outside the field. This forms a current loop that is known as an eddy current.

The side of the eddy current loop that is inside the magnetic field experiences a force due to the magnetic field. The direction of the force on the eddy current can be determined using the right-hand push rule and it is always opposite to the direction of motion of the sheet. (This means, referring to figure 7.13, it is harder to

move the metal to the right when the magnetic field is present than when the field is not present.)

Eddy currents in switching devices

Induction switches are electronic devices that detect the presence of metals and switch on another part of a circuit. Walk-through metal detectors at airports use induction switching devices.

Induction switching devices consist of a high-frequency oscillator, an analysing circuit and a relay. The oscillator produces an alternating current in a coil. This produces an electromagnetic field with a frequency of up to 22 MHz. When a metal object comes near the coil, eddy currents are created in the object. The eddy currents place a load on the coil and the frequency of the oscillator is reduced. The analysing circuit monitors the frequency of the oscillator and, when it falls below a certain threshold value, switches on an alarm circuit using the relay. The threshold frequency can be adjusted so that small loads such as a few coins or metal buttons and zippers will not trigger the alarm, but larger loads such as guns and knives will.

PHYSICS IN FOCUS

Electromagnetic braking

Consider a metal disk that has a part of it influenced by an external magnetic field, as illustrated in figure 7.14(a). As the disk is made of metal, the movement of the metal through the region of magnetic field causes eddy currents to flow. Using the right-hand push rule, it can be shown that the eddy current within the magnetic field in figure 7.14 will be upwards. The current follows a downward return path through the metal outside the region of magnetic influence. This is shown in figure 7.14(b).

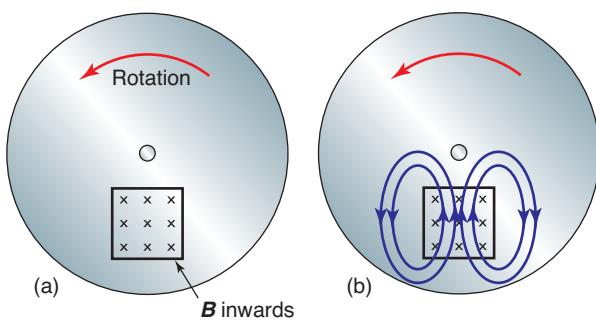


Figure 7.14 (a) A rotating metal disk acted upon by a magnetic field (b) The current that flows in the disk

The magnetic field exerts a force on the induced eddy current. This can be shown to oppose the motion of the disk in the example on the previous page by applying the right-hand push rule. In this way eddy currents can be utilised in smooth braking devices in trams and

trains. An electromagnet is switched on so that an external magnetic field affects part of a metal wheel or the steel rail below the vehicle. Eddy currents are established in the part of the metal that is influenced by the magnetic field. These currents inside the magnetic field experience a force that acts in the opposite direction to the relative motion of the train or tram, as explained below. In the case of the wheel, the wheel is slowed down. In the case of the rail, the force acts in a forward direction on the rail and there is an equal and opposite force that acts on the train or tram. *Note:* The right-hand push rule is used twice. The first time we use it, we show that an eddy current is produced. The thumb points in the direction of movement of the metal disk through the field because we imagine that the metal contains many positive charges moving through the field. We push in the direction of the force on these charges. This push gives us the direction of the eddy current.

The second time we use the right-hand push rule, we show that there is a force opposing the motion of the metal. Our thumb is put in the direction of the current in the field (the eddy current), then we push in the direction of the force on the moving charges (which are part of the metal disk). We then see that the force is always in the opposite direction to the movement of the metal.

PHYSICS IN FOCUS

Induction heating

Another effect of eddy currents is that they cause an increase in the temperature of the metal. This is due to the collisions between moving charges and the atoms of the metal, as well as the direct agitation of atoms by a magnetic field changing direction at a high frequency.

Induction heating is the heating of an electrically conducting material by the production of eddy currents within the material. This is caused by a changing magnetic field that passes through the material. Induction heating is undesirable in electrical equipment such as motors, generators and transformers, but it has been put to good use with induction cookers and induction furnaces.

Applying the principle of induction to cook tops in electric ranges

A gas stove top cooks food by burning gas to produce hot gases. The gases then flow across the bottom of a saucepan and transfer heat into it by conduction. However, a large amount of the thermal energy in the gases is carried away into the environment of the kitchen. The heat transferred to the saucepan is used to cook the food.

Some electric cook tops contain induction cookers instead of heating coils. An induction cooker sets up a rapidly changing magnetic field that induces eddy currents in the metal of the saucepan placed on the cook top. The eddy currents cause the metal to heat up directly without the loss of thermal energy that occurs with gas cooking. The heat produced in the metal saucepan is used to cook the food. The induction coils of the cooker are separated from the saucepan by a ceramic top plate. Induction cookers have an efficiency of about 80% while gas cookers have an efficiency rating of about 43%. A diagram of an induction cooker is shown in figure 7.15.

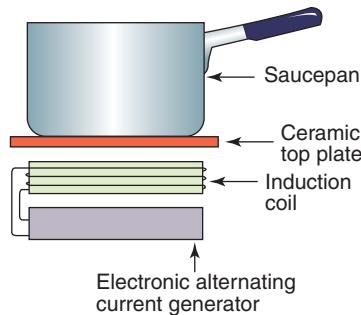


Figure 7.15
An induction cooker

As the current is alternating in the coil, there is a changing magnetic field that cuts through the metallic saucepan, causing eddy currents in the saucepan.

Induction furnaces

An induction furnace makes use of the heating effect of eddy currents to melt metals. This type of furnace consists of a container made from a non-metal material that has a high melting point and that is surrounded by a coil. The metal is placed in the container. The coil is supplied with an alternating current that can have a range of frequencies and this produces a changing magnetic field through the metal. Eddy currents in the metal raise its temperature until it melts. The eddy currents also produce a stirring effect in the molten metal, making the production of alloys easier. Induction furnaces take less time to melt the metal than flame furnaces. They are also cleaner and more efficient. A diagram of an induction furnace is shown in figure 7.16.

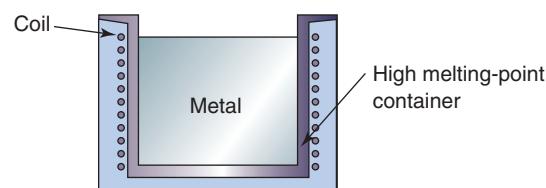


Figure 7.16 An induction furnace

SUMMARY

- Magnetic flux, Φ_B , is the amount of magnetic field passing through a given area. It depends on the strength of the field, B , as well as the area, A .
- The magnetic flux through a coil is the product of the area, A , of the coil and the component of the magnetic field strength, B , that is perpendicular to the area: $\Phi_B = B_{\perp}A$.
- Magnetic field strength is also known as magnetic flux density.
- Faraday's Law of Induction states that a changing magnetic flux through a circuit induces an emf in the circuit.
- The magnitude of an induced emf depends on the rate of change of magnetic flux through a circuit.
- Lenz's Law states that an induced emf always gives rise to a current that creates a magnetic field that, in turn, opposes the original change in flux through the circuit.
- Eddy currents are created when there is relative movement between a magnetic field and a metal object. The area of the magnetic field, however, does not cover the whole of the metal object. Eddy currents are also created when a conducting material is in the presence of a changing magnetic field.
- Eddy currents increase the temperature of metal objects.
- When a metal object is moving relative to a region affected by a magnetic field, the region of magnetic field exerts a force on the eddy currents that opposes the relative motion of the object to the field.

QUESTIONS

- Explain how Michael Faraday was able to produce an electrical current using a magnet.
- Define the concept of magnetic flux in terms of magnetic flux density and surface area.
- Calculate the magnetic flux threading (or passing through) the areas in the following cases.
 - An area of 1.5 m^2 is perpendicular to a magnetic field of flux density 2.0 T .
 - An area of 0.75 m^2 is perpendicular to a magnetic field of strength 0.03 T .

- A rectangle with a length of 4.0 cm and width 3.0 cm is perpendicular to a magnetic field of flux density $5.0 \times 10^{-3} \text{ T}$.

- A circle of radius 7.0 cm that is parallel with a magnetic field of flux density $5.0 \times 10^{-3} \text{ T}$.

- Evaluate the direction of the induced current through the galvanometer in each of the galvanometer circuit coils shown in figure 7.17. The arrows represent the motion of the coil or the magnet.

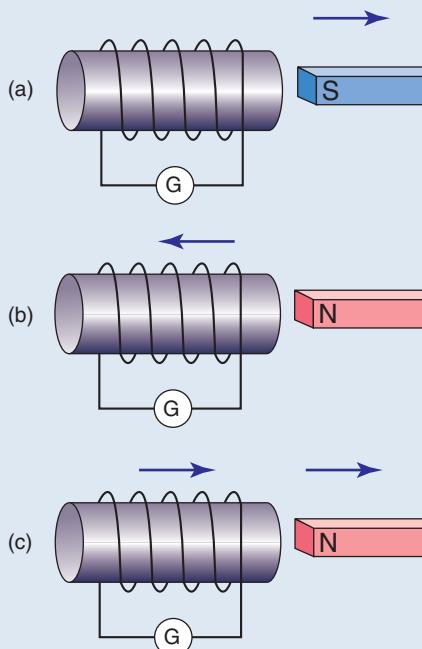


Figure 7.17

- Describe the effect that the speed of movement of a magnet has on the magnitude of a current induced in a coil.
- A magnet moving near a conducting loop induces a current in the circuit as shown in figure 7.18. The magnet is on the far side of the loop and is moving in the direction indicated by the dotted line. Describe two ways in which the magnet can be moving to induce the current as shown.

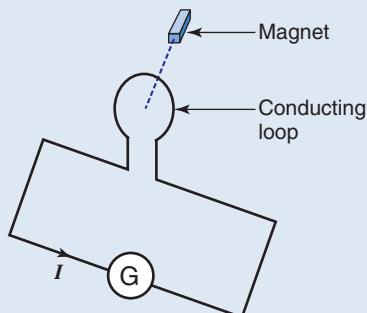


Figure 7.18

7. Deduce the direction of the induced current through the galvanometer in figure 7.19 when:
- the switch is closed
 - the switch remains closed and a steady current flows in the battery circuit
 - the switch is opened.

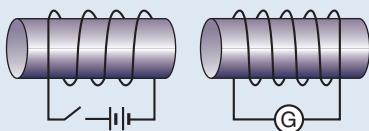


Figure 7.19

8. A flexible metal loop is perpendicular to a magnetic field as shown in figure 7.20(a). It is distorted to the shape shown in figure 7.20(b). Is the direction of the induced current in the loop clockwise or anticlockwise? Explain your answer.

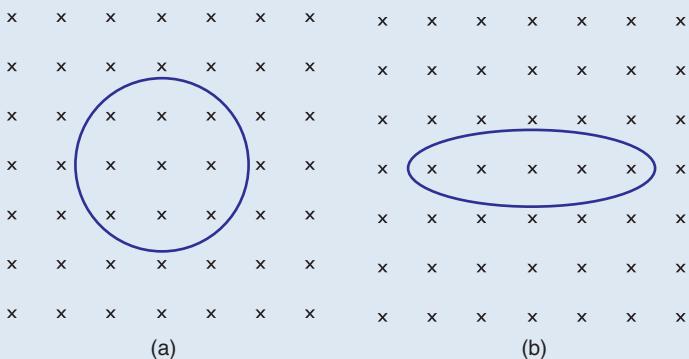


Figure 7.20

9. Figure 7.21 shows a loop of wire connected in series to a source of emf and a variable resistor. Describe the direction of the induced current in the central loop when the resistance of the outer loop circuit is increasing. Explain your reasoning.

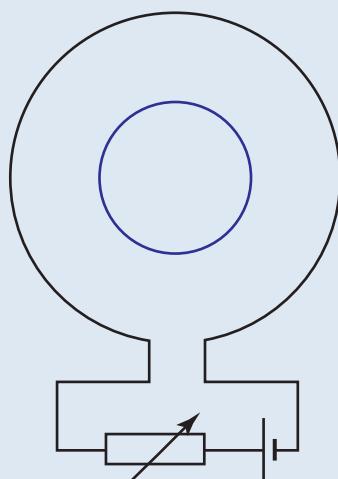


Figure 7.21

10. In what direction, clockwise or anticlockwise, is the induced current in the loop of wire in each situation shown in figure 7.22?

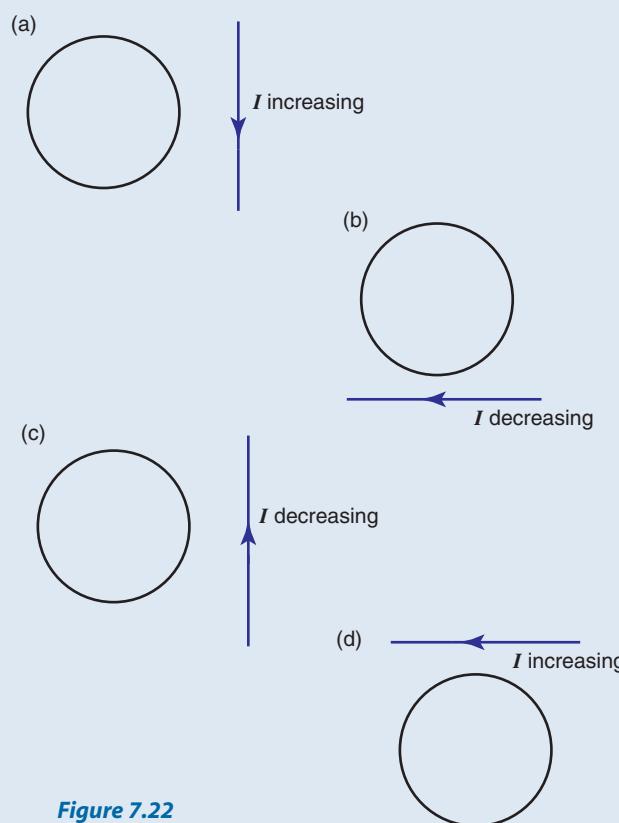


Figure 7.22

11. A metal rectangle has a length of 7.0 cm and a width of 4.0 cm. It is initially at rest in a uniform magnetic field of strength 0.50 T as shown in figure 7.23.

The rectangle is completely removed from the magnetic field in 0.28 s.

- What is the initial magnetic flux through the rectangle?
- In what direction, clockwise or anticlockwise, is the induced current in the rectangle when it is being removed from the magnetic field?

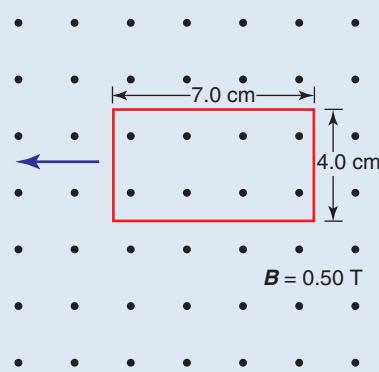


Figure 7.23

12. A square loop of wire has sides of length 6.5 cm. The loop is sitting in a magnetic field of strength 1.5×10^{-3} T as shown in figure 7.24. The magnetic field is reduced to 0 T in a period of 5.0 ms.
- What is the flux through the loop initially?
 - What will be the effect on the induced emf if a 25-loop coil is used, rather than a single loop?
 - In what direction will the current flow in the loop?

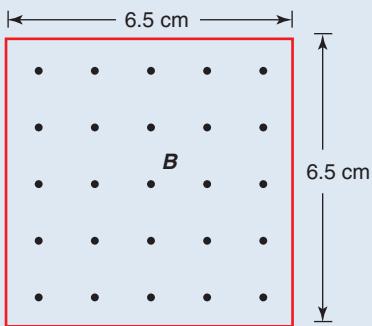


Figure 7.24

13. Use diagrams to show three ways to change the flux passing through a conductor loop.
14. (a) Explain the production of a back emf in an electric motor.
 (b) Describe how the back emf of an electric motor is produced.
 (c) Explain why the back emf in an electric motor opposes the supply emf.
 (d) How does the back emf determine the maximum rotating speed of an electric motor?
 (e) How can overloading an electric motor cause it to burn out?
15. The armature winding of an electric motor has a resistance of $5.0\ \Omega$. The motor is connected to a 240 V supply. When the motor is operating with a normal load, the back emf is equal to 237 V.
- Calculate the current that passes through the motor when it is first started.
 - Calculate the current that passes through the motor when it is operating normally.

16. The armature winding of an electric drill has a resistance of $10\ \Omega$. The drill is connected to a 240 V supply. When the drill is operating normally, the current drawn is 2.0 A.

- Calculate the current that passes through the drill when it is first started.
- Calculate the back emf of the drill when it is operating normally.

17. (a) What is an eddy current?
 (b) Discuss how eddy currents are produced.
 (c) Describe how eddy currents raise the temperature of metals.

18. A rectangular sheet of aluminium is pulled from a magnetic field as shown in figure 7.25.
- Copy the diagram and indicate the position and direction of an eddy current loop.
 - Which way does the force due to the external magnetic field and the eddy current act on the aluminium sheet?

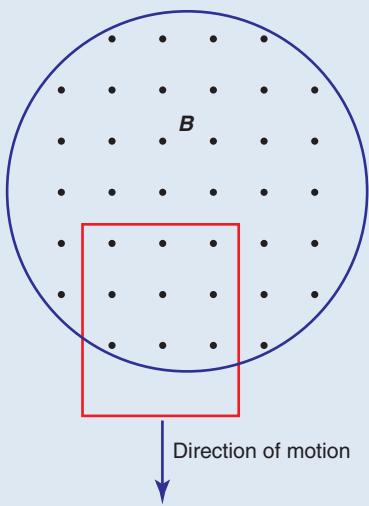


Figure 7.25

19. Discuss how eddy currents are utilised in the following situations:
- induction electric cook tops
 - electromagnetic braking of trains or trams
 - induction furnaces.



7.1 INDUCING CURRENT IN A COILED CONDUCTOR

Aim

- To study ways of inducing a current in a coiled conductor
- To study factors affecting the size of the induced current.

Apparatus

galvanometer

two coils having different numbers of turns of wire
two bar magnets, of different strengths, if possible
an iron core that fits into one of the coils. This could be made by taping large iron nails together.

connecting wires

Theory

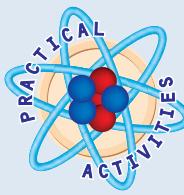
You will be reproducing some of Faraday's experiments. Modern galvanometers are much more sensitive than those available to Faraday.

Method

- Connect the coil with the fewest number of turns to the galvanometer. Push the N pole of a bar magnet into the coil. Describe what happens.
- Hold the magnet stationary near the coil. Describe what happens.
- Withdraw the N pole from the coil. Describe what happens.
- Repeat steps 1 and 3 at different speeds. Describe what happens.
- Hold the magnet stationary and move the coil in different directions. Rotate the coil so that first one end and then the other approaches the magnet. Describe what happens.
- Place the iron core in the coil and touch it with the N pole. Remove the magnet. Describe what happens.
- Design an experiment to examine the factors that affect the size of the induced current.

Analysis

- Relate your results to Faraday's law of electromagnetic induction.
- Describe how you would make a generator to create a relatively large current.



7.2 LINKING COILS

Aim

To see if the magnetic field of a current-carrying coil can induce a current in another coil.

Apparatus

galvanometer

two coils having different numbers of turns of wire, preferably one of which fits into the other
an iron core that fits into the smaller of the coils. This could be made by taping large iron nails together.
a $10\ \Omega$ resistor
a power supply
connecting wires

Theory

A current flowing in a coil will create a magnetic field that threads the second coil. When there is a changing magnetic field threading the second coil, an emf will be induced in the coil.

Method

- Set up the apparatus as shown in figure 7.26. Set the power supply to 2.0 V.

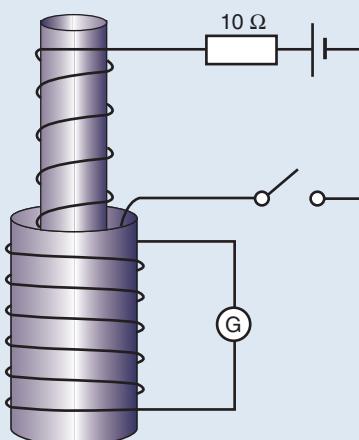


Figure 7.26

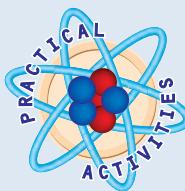
- Close the switch and observe the effects on the galvanometer.
- Open the switch and observe the effects on the galvanometer.
- Put an iron core in the smaller coil and repeat steps 2 and 3.

Analysis

Relate your results to Faraday's law of electromagnetic induction.

Questions

1. When was a current induced in the secondary (galvanometer) coil?
2. What happened when a steady current was flowing in the primary (power supply) coil?
3. What effect did the iron core have on the induced current?



7.3 THE DIRECTION OF INDUCED CURRENTS

Aim

To learn how to predict the direction of an induced current.

Apparatus

galvanometer
coil
bar magnet
connecting wires
battery

Theory

Lenz's Law states that the direction of an induced current in a coil is such that the magnetic field that it establishes opposes the change of the original flux threading the coil.

Method

1. Carefully examine the coil to see which way the wire is coiled around the cylinder.
2. Use a battery to establish which way the galvanometer deflects when currents flow through it in different directions.
3. Connect the coil to the galvanometer.
4. Push the N pole of the magnet towards the coil.
Note the deflection of the galvanometer.
Answer the following questions.
 - In which direction did the current flow in the coil?
 - Was the end of the coil nearest the magnet a N or S pole?
 - Does the magnetic field of the coil assist or oppose the motion of the magnet?
 - Does the magnetic field of the coil assist or oppose the change in flux threading the coil?
5. Pull the N pole away from the coil. Answer the questions of step 4.
6. Push the S pole of the magnet towards the coil.
Answer the questions of step 4.
7. Pull the N pole away from the coil. Answer the questions of step 4.

Analysis

Do your results verify Lenz's Law? Explain.

CHAPTER

8

GENERATORS AND POWER DISTRIBUTION



Figure 8.1 High-voltage transmission lines are used to distribute power from the generators to the consumers.

Remember

Before beginning this chapter, you should be able to:

- state Ohm's Law for metal conductors at a constant temperature:
 $V = IR$
- recall that the work done in a circuit component is equal to the amount of energy transformed in the component
- apply the formulas: $W = VQ$, $Q = It$, and $W = Vit$
- recall that power is the rate of doing work: $P = \frac{W}{t}$
- recall that the power dissipated in a metal conductor is given by the following formulas: $P = VI$, $P = \frac{V^2}{R}$ and $P = I^2R$
- recall that magnetic flux is the amount of magnetic field passing through an area
- recall that, through a circuit, a changing magnetic field induces an emf across the ends of the wire that makes the circuit
- apply Faraday's law: *The magnitude of the induced emf depends on the rate of change of magnetic flux through the coil.*
- apply Lenz's Law: *The induced emf in a coil is such that if a current were to flow, it would produce a magnetic field that opposes the change in flux threading the coil.*

Key content

At the end of this chapter you should be able to:

- describe the main components of a generator
- compare the structure and function of a generator to an electric motor
- describe the differences between AC and DC generators
- discuss the energy losses that occur as energy is fed through transmission lines from the generator to the consumer
- assess the effects of the development of AC generators on society and the environment
- outline the competition between Westinghouse and Edison to supply electricity to cities
- describe the purpose of transformers in electric circuits
- compare step-up and step-down transformers
- identify the relationship between the ratio of the number of turns in the primary and secondary coils of a transformer and the ratio of primary to secondary voltage
- explain why current transformations are related to the Principle of Conservation of Energy
- solve problems and analyse information about transformers using:
$$\frac{V_p}{V_s} = \frac{n_p}{n_s}$$
- explain the role of transformers in electricity substations
- discuss why some domestic electrical appliances use a transformer
- discuss the impact of the development of transformers on society.

If you have ever experienced a power blackout you will realise the dependence that society has developed for electricity. We use it for lighting, warmth, cooling systems and refrigeration of our food. It runs our computers, radios, televisions, DVD and CD players, and other appliances. Much of industry is also dependent on the supply of electrical energy for it to function. It provides safe, well-lit and comfortable office environments and powers machinery in factories, hospital equipment and communications technology.

Imagine what your life would be like if the principles of electromagnetic induction had not yet been discovered. There would be no cars as we know them because the ignition system relies on devices such as generators (alternators) and transformers (coils). What sort of music would you be listening to if there were no electric guitars and keyboard instruments?

Modern Western society is dependent on the production and transmission of electrical energy. In this chapter we will look at how electricity is produced by generators and how it is transferred from the power stations to homes and other consumers.

8.1 GENERATORS

eBookplus

Weblink:
Generator applet

A generator is a device that transforms mechanical kinetic energy into electrical energy. In its simplest form, a generator consists of a coil of wire that is forced to rotate about an axis in a magnetic field.

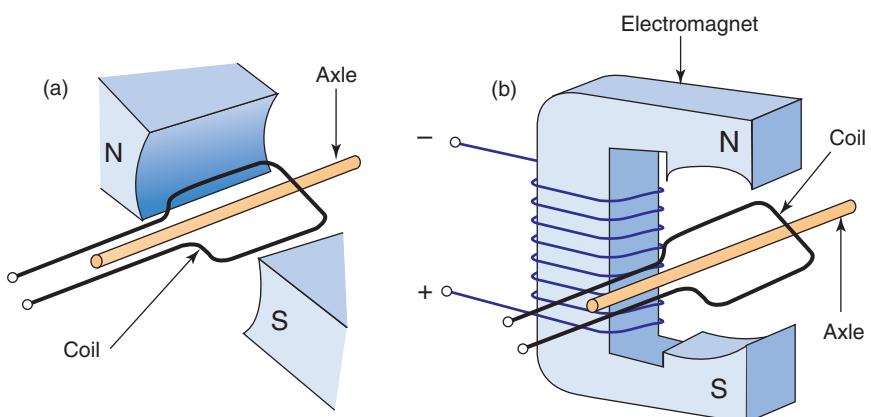
As the coil rotates, the magnitude of the magnetic flux threading (or passing through) the area of the coil changes. This changing magnetic flux produces a changing emf across the ends of the wire that makes up the coil. This is in accordance with Faraday's Law of Induction (see chapter 7), which can be stated as:

The induced emf in a coil is equal in magnitude to the rate at which the magnetic flux through the coil is changing with time.

The magnetic field of a generator can be provided either by using permanent magnets, as shown in figure 8.2(a) or by using an electromagnet, as shown in figure 8.2(b).

Figure 8.2 (a) Permanent magnets provide the magnetic field.

(b) An electromagnet provides the magnetic field.



The **stator** is the stationary part of an electrical rotating machine.

The **rotor** is the rotating part of an electrical rotating machine.

The stationary functioning parts of a generator are called the **stator**, and the rotating parts are called the **rotor**. In figure 8.2(a) and (b), the stators consist of the sections that produce the magnetic fields (permanent magnets or electromagnets). The rotors are the coils.

If the coil of a generator is forced to rotate at a constant rate, the flux threading the coil and the emf produced across the ends of the wire of the coil vary with time as shown in figure 8.3 below.

In figure 8.3 the magnetic field is directed to the right. The corners of the coil have been labelled L, K, M and N so that you can see how the coil is rotating.

Beneath the diagrams of the coil is an end view of the sides LK and MN showing the direction of the induced current that would flow through the sides at that instant if the generator coil was connected to a load. The arrows on this part of the diagram show the direction of movement of the sides of the coil.

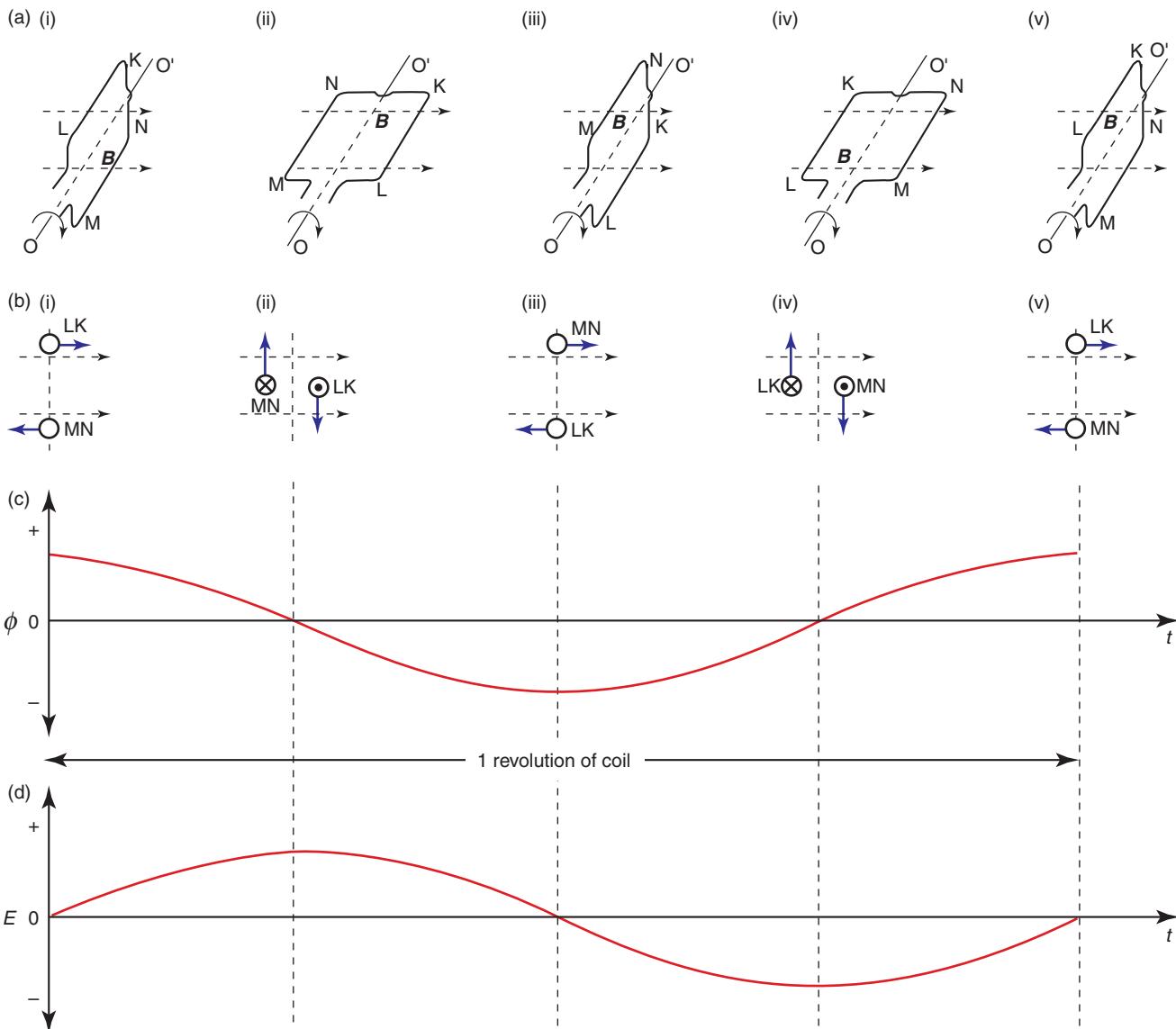


Figure 8.3 The variation of flux and emf of a generator coil as it completes a single revolution

The next section down in the diagram is a graph showing the variation of magnetic flux through the coil as a function of time.

The last section of the diagram is a graph showing the variation of emf that would be induced in the coil (if there was a gap in the coil between the points L and M) as a function of time. The emf is given by the negative of the gradient of a graph of magnetic flux threading the coil versus time.

In figure 8.3(a)(i) the flux threading the coil is at a maximum value. The emf is zero, as the gradient of the flux versus time graph is zero, which means that there is no change in flux through the coil at this instant.

In figure 8.3(a)(ii) the flux threading the coil is zero. The emf is at a maximum positive value, as the flux versus time graph has a maximum negative gradient. At this instant the change in flux is happening at a maximum rate.

In figure 8.3(a)(iii) the coil is again perpendicular to the magnetic field, but now the coil is reversed to its original orientation. The flux threading the coil is at a maximum negative value. The emf is zero, as the gradient of the flux versus time graph is again zero, meaning that at this instant there is again no change in flux.

In figure 8.3(a)(iv) the flux threading the coil is again zero. The emf now has its maximum positive value, as the gradient of the flux versus time graph has its maximum negative value. At this instant the change in flux is again happening at a maximum rate.

In figure 8.3(a)(v) the flux threading the coil is again at a maximum value. The emf is zero, as the gradient of the flux versus time graph is zero and there is no change in flux at this instant. And so the cycle continues.

The frequency and amplitude of the voltage produced by a generator depend on the rate at which the rotor turns. If the rotor is turning at twice the original rate, then the period of the voltage signal halves, the frequency doubles and the amplitude doubles. This is shown in figure 8.4.

The effectiveness of generators is increased by winding the coil onto an iron core armature. The iron core makes the coil behave like an electromagnet. This intensifies the changes in flux threading the coil as it is forced to rotate and increases the magnitude of the emf that is induced. This effect also occurs when the number of turns of wire on the armature is increased. The coil then behaves like a number of individual coils connected in series. If there are n turns of wire on the armature, the maximum emf will be n times that of a single coil rotating at the same rate.

AC generators

Figure 8.3, on page 141, shows how a coil forced to rotate smoothly in a magnetic field has a varying emf induced across the ends of the coil. The value of the emf varies sinusoidally with time. (This means that the graph of emf versus time has the same shape as a graph of $\sin x$ versus x .) If such an emf signal were placed across a resistor, the current flowing through the resistor would periodically alternate its direction. In other words, the emf across the ends of a coil rotating at a constant rate in a magnetic field produces an alternating current (AC). Alternating current electrical systems are used across the world for electrical power distribution.

This type of AC generator connects the coil to the external circuit or distribution system by the use of slip rings. Slip rings rotate with the coil. A slip ring system is shown in figure 8.5 on the following page.

In figure 8.5, side LK of the coil is connected to slip ring B while side MN is connected to slip ring A. Brushes make contact with the slip rings and transfer the emf (or current) to the **terminals** of the generator. In this case, the terminals are the external points of the generator where it connects to the load.

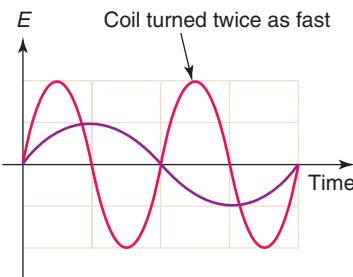


Figure 8.4 Doubling the frequency of rotation doubles the maximum induced emf.

A **terminal** is the free end of a cell or battery to which a connection is made to the rest of a circuit.

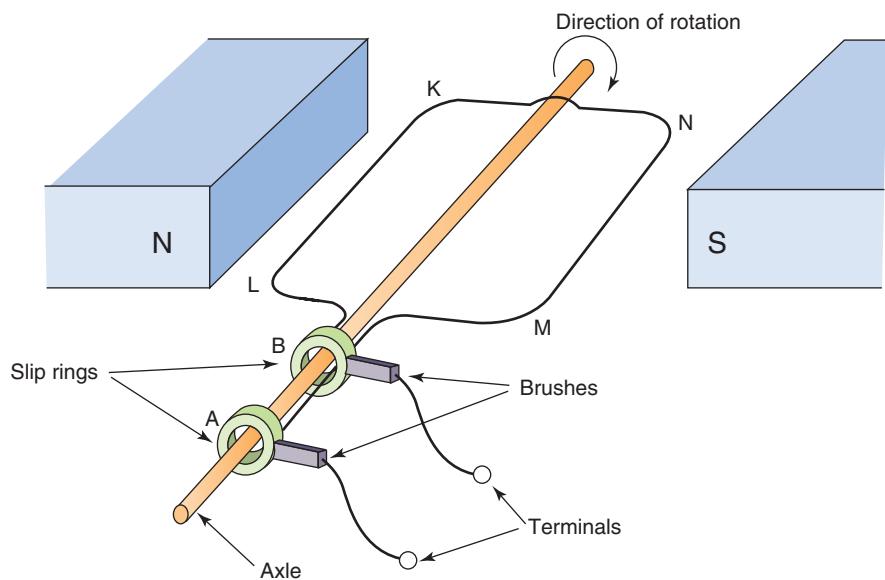


Figure 8.5 The functional parts of an AC generator

Which way will the current flow?

When asked to determine the direction of the current in the generator or some other part of a circuit connected to the generator, there are two methods that can be used.

The first method is to consider the magnetic force on a positive test charge in one side of the coil. The direction of the velocity of the charge in the magnetic field depends on the direction of the rotation of the coil. The direction of the magnetic force is determined using the right-hand push rule (see figure 8.6 below). The direction of the force acting on the test charge is also the direction of the current on that side of the generator. It is then a matter of following that direction around the coil to the terminals. Note that the terminal from which the current emerges at a particular instant is acting as the positive terminal.

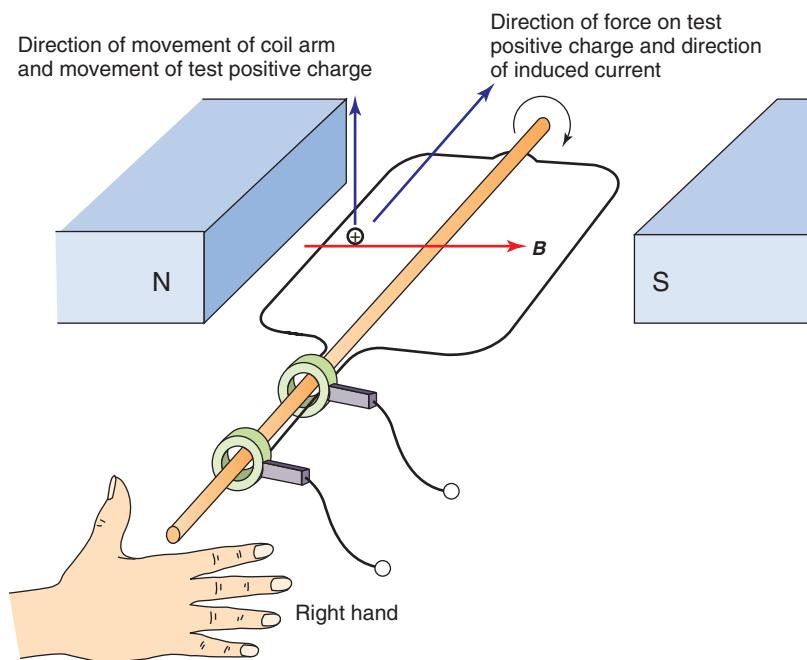


Figure 8.6 Using the right-hand push rule to determine the direction of current flow in a generator coil

SAMPLE PROBLEM**8.1**

The other method is to apply Lenz's Law to the coil. First determine the way in which the flux threading the coil is changing at the instant in question. The current induced in the coil will produce a magnetic field that opposes the change in flux through the coil. Once you have established the direction of the flux produced by the induced current, apply the right-hand grip rule for coils to determine the direction of the current around the coil.

Both methods are illustrated in sample problem 8.1.

Determining the polarity of a generator's terminals

Figure 8.7 shows an AC generator at a particular instant. At this instant, which of the terminals, A or B, is positive?

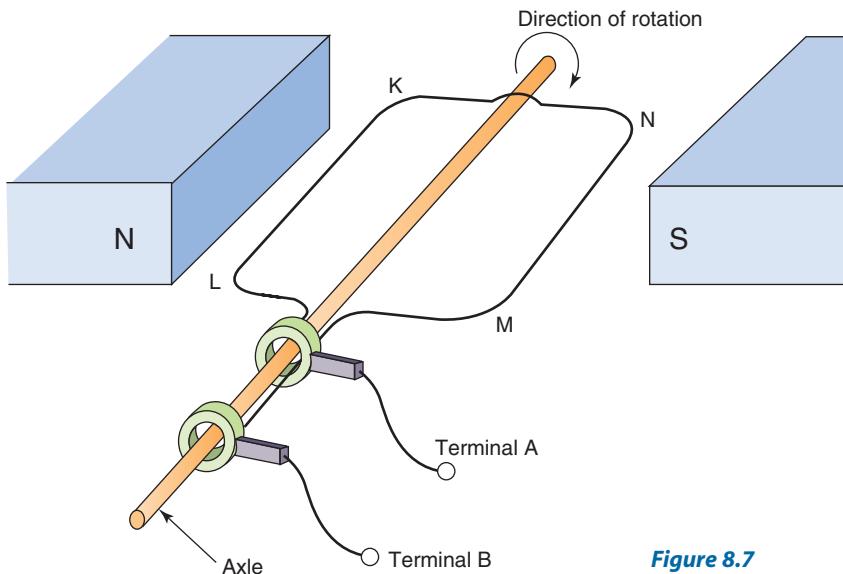


Figure 8.7

SOLUTION**Test charge method**

Consider a positive test charge in the side labelled LK. At the instant shown, this positive charge is moving upwards in a magnetic field directed to the right. Applying the right-hand rule, the positive charge is forced in the direction from L towards K. This situation is shown in figure 8.8.

Side LK is connected to the slip ring leading to terminal A. Side MN is connected to the slip ring leading to terminal B. If a current were to flow, it would emerge from terminal B. Therefore, terminal B is positive at the instant shown.

Using Lenz's Law

At the instant shown in figure 8.7, the flux is increasing to the right through the coil as it is forced to rotate in the indicated direction. The induced current in the coil will therefore produce a magnetic field that passes through the coil to the left to oppose the external change in magnetic flux through the coil. The right-hand grip rule for coils (thumb in the direction of the induced magnetic field through the coil, fingers grip the coil pointing in the direction of the current in the coil) shows that the induced current is clockwise around the coil as we view it. The current then emerges from the generator through terminal B. Therefore, terminal B is positive at the instant shown.

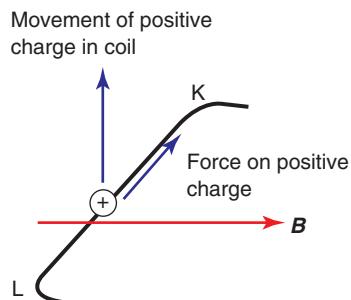
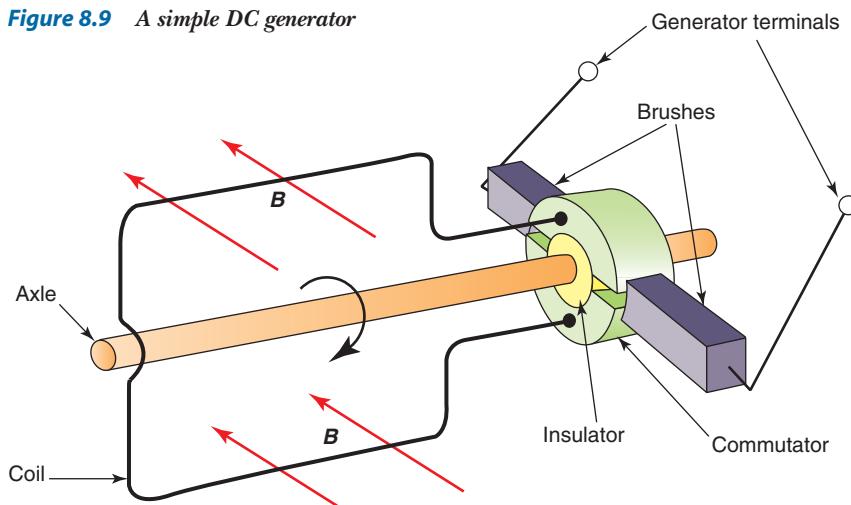


Figure 8.8 Positive charges are pushed in the direction from L towards K.

DC generators

A direct current (DC) is a current where the flow of charge is in one direction only. Direct currents provided by a battery or dry cell usually

Figure 8.9 A simple DC generator



have a steady value. Direct currents may also vary with time, but keep flowing in the same direction. DC generators provide such currents.

A simple DC generator consists of a coil that rotates in a magnetic field. (This also occurs in an AC generator.) The difference between an AC and a DC generator is in the way that the current is provided to the external circuit. An AC generator uses slip rings. A DC generator uses a split ring commutator to connect the rotating coil to the terminals. (Remember that a commutator is a switching device for reversing the

direction of an electric current.) The functional parts of a simple DC generator are shown in figure 8.9. The magnets have been omitted for clarity.

This diagram of a simple DC generator should remind you of a DC motor (see chapter 6), as it has the same parts. In the generator the coil is forced to rotate in the magnetic field. This induces an emf in the coil. The emf is transferred to the external circuit via the brushes that make contact with the commutator. When the emf of the coil changes direction, the brushes swap over the side of the coil they are connected to, thus causing the emf supplied to the external circuit to be in one direction only. The result of this process is shown in figure 8.10.

The output from a DC generator can be made smoother by including more coils set at regular angles on the armature. Each coil is connected to two segments of a multi-part commutator and the brushes make contact only with the segments connected to the coil producing the greatest emf at a particular time. A two-coil DC generator is shown in figure 8.11(a) and its output is shown in figure 8.11(b). Note that in this case the commutator has four segments.

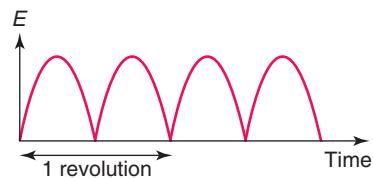


Figure 8.10 The output from a simple DC generator

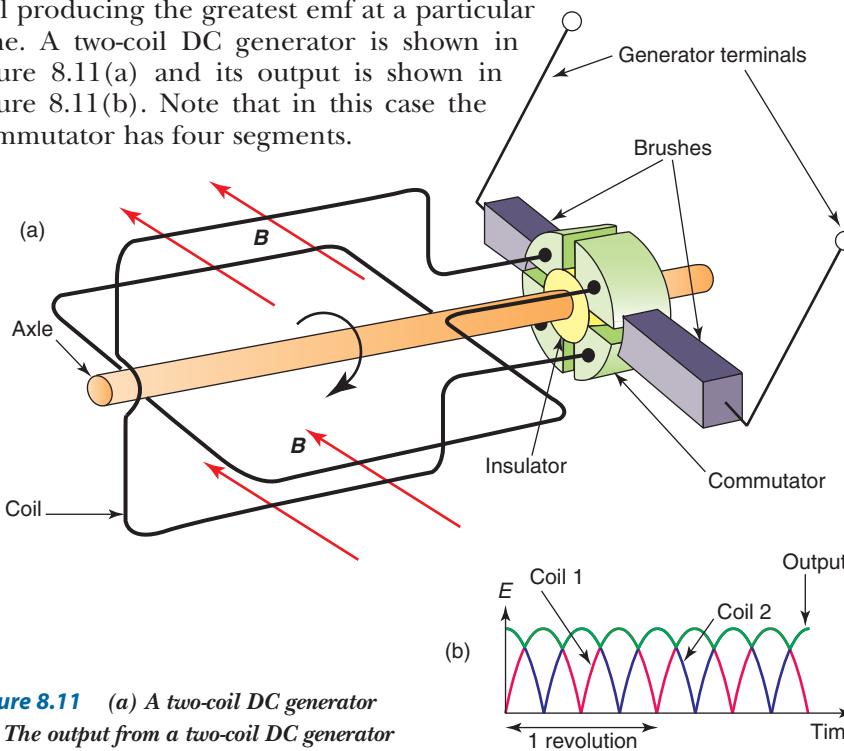


Figure 8.11 (a) A two-coil DC generator
(b) The output from a two-coil DC generator

You can investigate the operation and structure of an AC generator and a DC motor used as a DC generator by doing practical activity 8.1.



8.1

Observing the output of a hand-operated generator

8.2 ELECTRIC POWER GENERATING STATIONS

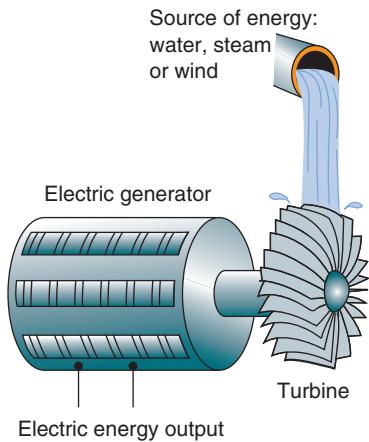


Figure 8.12 A turbine drives a generator.

Electric power generating stations provide electrical power to domestic and industrial consumers. In a power station, mechanical or heat energy is transformed into electrical energy by means of a turbine connected to a generator. A turbine is a machine whose shaft is rotated by jets of steam or water directed onto blades attached to a wheel. Figure 8.12 shows a simple turbine and generator combination.

The generators used in power stations have a different structure to those studied so far. A typical generator has an output of 22 kV. This requires the use of massive coils which would place huge forces on bearings if they were required to rotate. To eliminate this problem, a power station generator has stationary coils mounted on an iron core (making up the stator). The coils are linked in pairs on opposite sides of the rotor. The rotor is a DC supplied electromagnet that spins with a frequency of 50 Hz. A simplified diagram of a power station generator is shown in figure 8.13. In this diagram only one set of linked coils is shown.

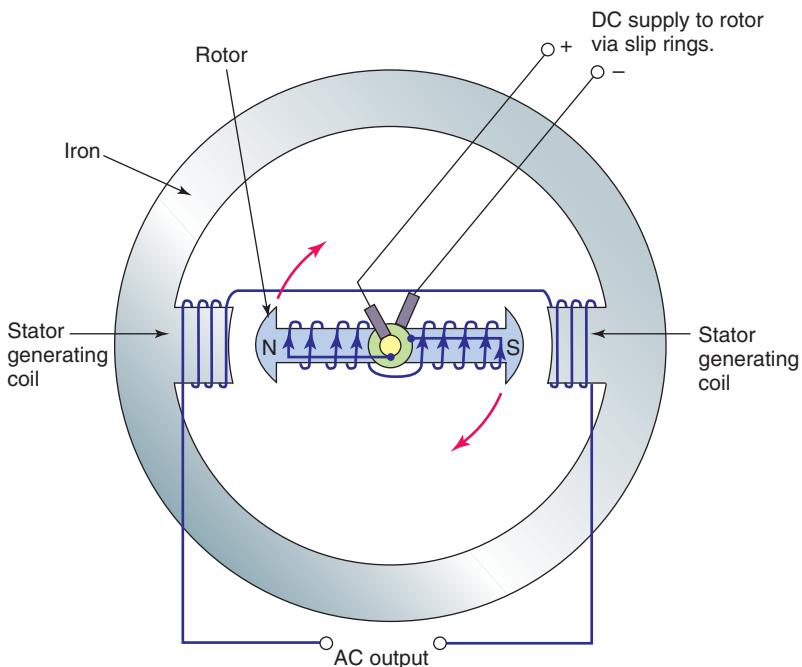


Figure 8.13 A single-coil generator

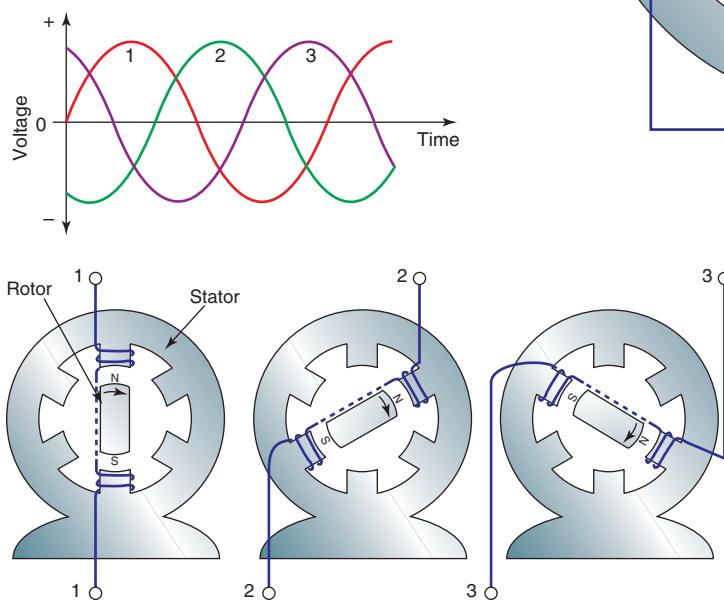


Figure 8.14 Three-phase power generation

Power station generators have three sets of coils mounted at angles of 120° to each other on the stator. This means that each generator produces three sets of voltage signals that are out of phase with each other by 120° . This is known as three-phase power generation. Each generator is connected to four lines, one line for each phase and a return ground line. Figure 8.14 shows the arrangement of the coils on the stator and the voltage outputs of each set of coils.

There are two main types of power station used in Australia: fossil fuel steam stations and hydroelectric stations.

PHYSICS FACT

AC versus DC: Westinghouse and Edison

Thomas Edison (1847–1931) is credited with many inventions, including the electric light bulb and the phonograph. He was the first person to establish a business supplying electricity to cities. Electricity was initially supplied to cities for lighting streets and houses. Edison General Electric Company opened the first electric power station in New York City in 1882 and began by installing street lighting systems.

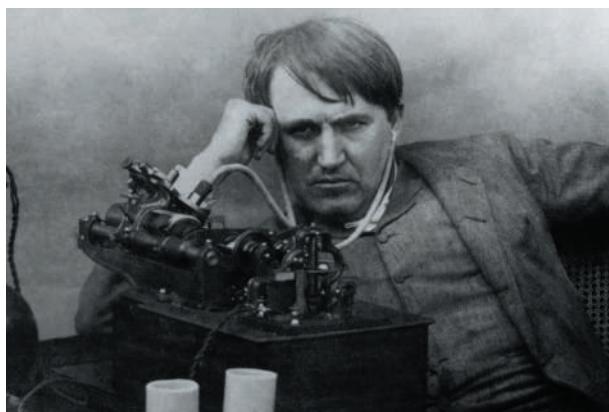


Figure 8.15 Thomas Edison

Edison used a direct current (DC) system. His generators — also known as dynamos — used a commutator to give a DC output. The commutator proved to be a problem with the high-speed, steam-driven generators of the 1890s. Edison's power stations could only supply areas a few kilometres away. Power losses in power lines vary directly with the square of the current in the lines and with the resistance of the lines. He therefore relied on thick copper cables to carry the electric current, as this reduced the resistance of the lines.

The first alternating current (AC) system was demonstrated in Paris in 1883. Nikola Tesla (1856–1943) was a Serbian-born scientist who worked for Edison Electric Company in France repairing one of the electrical plants. The company refused to pay him when he completed the project. He then moved to America and was hired by Thomas Edison. Edison promised to pay Tesla \$50 000 if he redesigned the dynamos used to produce DC electricity. When Tesla completed the project, Edison did not pay him — he said he had been joking about the money.

Tesla left the Edison Electric Company and worked for two years as a labourer while developing

an AC generating and transmission system. He also invented electrical generators and motors to use with AC.

George Westinghouse (1846–1914) saw the advantages of using AC for supplying cities. He purchased the patent rights for Tesla's generators and motors for \$1 000 000.

Edison saw the threat to his business posed by the Westinghouse AC system and tried to discredit it. He published an 83-page booklet entitled 'A Warning! From The Edison Electric Light Co.' in which he described the horrible deaths of people who had supposedly come into contact with Westinghouse's AC cables. In 1887, Edison held a public demonstration in New Jersey to show the dangers of AC electricity. He set up a 1000-volt Westinghouse AC generator attached to a metal plate and used it to kill a dozen animals.

The electric chair

The Edison research facility hired an inventor, Harold P. Brown, and his assistant Dr Fred Peterson to develop an electric chair using AC electricity. They acquired a 1000-volt Westinghouse generator and used it to kill dogs, cows and horses. They often invited the media to witness the experiments.

On 4 June 1888, electrocution became the method of execution in the state of New York. While still on the payroll of Edison, Dr Peterson headed a committee that advised the government on the best method of electrocution. The committee recommended the use of AC electricity. Westinghouse refused to sell generators to the New York state prisons for use in the electric chair, but Edison and Brown found a way to provide them.

(continued)

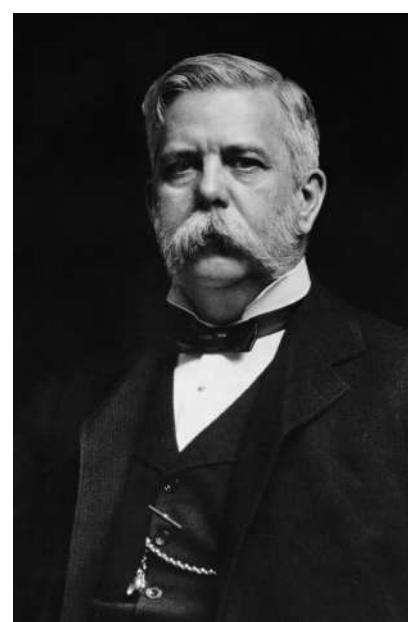


Figure 8.16
George Westinghouse

The first execution using the electric chair took place in New York in August 1890 and the victim was William Kemmler. Edison tried to describe the people who died in the electric chair as having been 'Westinghoused'. Westinghouse supposedly hired the best lawyer of the day to defend Kemmler and to attack electrocution as being a 'cruel and unusual form of punishment'. (Cruel and unusual forms of punishment are banned in the American Bill of Rights.)

Niagara Falls Power Plant

In 1887, a group of businessmen in Buffalo, USA, pledged a prize of \$100 000 to the inventors of the world to design a system that would use the power of Niagara Falls to provide electricity to the city, some 30 kilometres away. Westinghouse and Edison were competitors for the prize.

In 1891, the International Electrical Exhibition was held in Frankfurt, Germany. An AC line was set up to carry sizeable quantities of electrical power from Frankfurt to Lauffen, a

distance of about 180 kilometres. Tests of the system showed that only 23 per cent of the power was lost. This demonstration enabled Westinghouse to win the contract in 1893. The first AC generators were installed at Niagara Falls in 1896.



Figure 8.17 The first electric chair; used in the 1890s

8.3 TRANSFORMERS

A **transformer** is a magnetic circuit with two multi-turn coils wound onto a common core.

eBookplus

Weblink:
Transformer applet

Transformers are devices that increase or decrease AC voltages. They are used in television sets and computer monitors that have cathode ray tubes (see chapter 10) to provide the very high voltages needed to drive the cathode ray tubes. They are used in electronic appliances such as radios to provide lower voltages for amplifier circuits. They are also found in answering machines, cordless phones, digital cameras, battery chargers, digital clocks, computers, phones, printers, electronic keyboards, the electric power distribution system and many other devices.

Transformers consist of two coils of insulated wire called the *primary* and *secondary* coils. These coils can be wound together onto the same soft iron core, or linked by a soft iron core. The structure of the most common type of transformer and its circuit symbol are shown in figure 8.18.

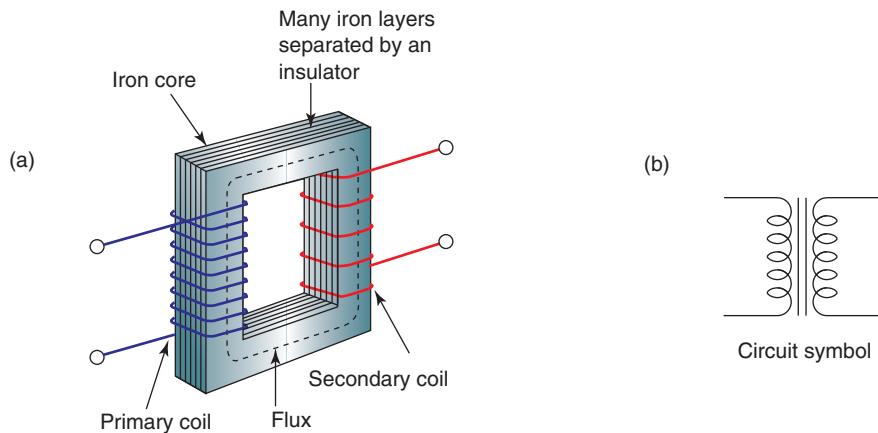


Figure 8.18 (a) A transformer with coils linked by a soft iron core
(b) Circuit symbol



8.2

Making a simple transformer

If a steady DC current were to flow through the primary coil, it would produce a constant flux threading the secondary coil. No voltage would be induced in the secondary coil, as a changing flux is needed.

A **step-up transformer** provides an output voltage that is greater than the input voltage.

A **step-down transformer** provides an output voltage that is less than the input voltage.



8.3

Transformer ins and outs

Transformers are designed so that almost all the magnetic flux produced in the primary coil threads the secondary coil. When an alternating current flows through the primary coil, a constantly changing magnetic flux threads (or passes through) the secondary coil. This constantly changing flux passing through the secondary coil produces an AC voltage at the terminals of the secondary coil with the same frequency as the AC voltage supplied to the terminals of the primary coil.

The difference between the primary voltage, V_p , and the secondary voltage, V_s , is in their magnitudes. The secondary voltage can be greater than or less than the primary voltage, depending on the design of the transformer. The magnitude of the secondary voltage depends on the number of turns of wire on the primary coil, n_p , and secondary coil, n_s .

If the transformer is ideal, it is 100% efficient and the energy input at the primary coil is equal to the energy output of the secondary coil. The rate of change of flux $\left(\frac{\Delta\Phi}{\Delta t}\right)$ through both coils is the same. Faraday's Law can be used to show that the secondary voltage is found using the formula:

$$V_s = n_s \frac{\Delta\Phi}{\Delta t}$$

Similarly, the input primary voltage, V_p , is related to the change in flux by the equation:

$$V_p = n_p \frac{\Delta\Phi}{\Delta t}$$

Dividing these equations produces the transformer equation:

$$\frac{V_p}{V_s} = \frac{n_p}{n_s}$$

If n_s is greater than n_p , the output voltage, V_s , will be greater than the input voltage, V_p . Such a transformer is known as a **step-up transformer**. If n_s is less than n_p , the output voltage, V_s , will be less than the input voltage, V_p . Such a transformer is known as a **step-down transformer**.

Transformers and the Principle of Conservation of Energy

The Principle of Conservation of Energy states that energy cannot be created nor destroyed but that it can be transformed from one form to another. This means that if a step-up transformer gives a greater voltage at the output, there must be some kind of a trade-off. The rate of supply of energy to the primary coil must be greater than or equal to the rate of supply of energy from the secondary coil. For example, if 100 J of energy is supplied each second to the primary coil, then the maximum amount of energy that can be obtained each second from the secondary coil is 100 J.

You cannot get more energy out of a transformer than you put into it. (Some energy is usually transformed into thermal energy in the transformer due to the occurrence of eddy currents in the iron core. In other words, eddy currents in the iron core cause the transformer to heat up.) There is a decrease in useable energy whenever energy is transformed from one form to another. The 'lost' energy is said to be dissipated, usually as thermal energy.

The rate of supply of energy is known as *power* and is found using the equation:

$$P = VI$$

In ideal transformers there is assumed to be no power loss and the primary power is equal to the secondary power. In this case:

$$P_p = P_s.$$

Substituting the power formula stated earlier, this equation becomes:

$$V_p I_p = V_s I_s.$$

Combining this equation with the transformer equation we get another very important relationship for transformers:

$$\frac{I_s}{I_p} = \frac{n_p}{n_s}.$$

SAMPLE PROBLEM

8.2

Transformer calculations

The transformer in an electric piano reduces a 240 V AC voltage to a 12.0 V AC voltage. If the secondary coil has 30 turns and the piano draws a current of 500 mA, calculate the following quantities:

- (a) the number of turns in the primary coil
- (b) the current in the primary coil
- (c) the power output of the transformer.

SOLUTION

(a)

QUANTITY	VALUE
V_p	240 V
V_s	12.0 V
n_s	30
n_p	?

$$\begin{aligned} \frac{V_p}{V_s} &= \frac{n_p}{n_s} \\ \Rightarrow n_p &= \frac{n_s V_p}{V_s} \\ &= \frac{30 \times 240}{12.0} \\ &= 600 \end{aligned}$$

Therefore the primary coil has 600 turns.

(b)

QUANTITY	VALUE
I_p	?
I_s	500 mA
n_s	30
n_p	600

$$\begin{aligned} \frac{I_s}{I_p} &= \frac{n_p}{n_s} \\ \Rightarrow I_p &= \frac{n_s I_s}{n_p} \\ &= \frac{30 \times 500}{600} \\ &= 25 \text{ mA} \end{aligned}$$

(c)

QUANTITY	VALUE
V_s	12.0 V
I_s	500 mA
P_s	?

$$\begin{aligned}P_s &= V_s I_s \\&= 12.0 \times 500 \\&= 6000 \text{ mW} \\&= 6.0 \text{ W}\end{aligned}$$

Reducing heat losses due to eddy currents

As we saw in chapter 7, eddy currents within a metal are circular movements of electrons due to a changing flux passing through the metal. These circular movements are at right angles to the direction of the changing flux.

By constructing the iron core from many layers of iron that are coated with an insulator, the size of the eddy currents is reduced and the losses due to heating effects are reduced. Such a core is called a laminated iron core. The cross-sections of the thin layers are perpendicular to the direction of the magnetic flux, so the size of the eddy currents is greatly reduced, as illustrated in figure 8.19.

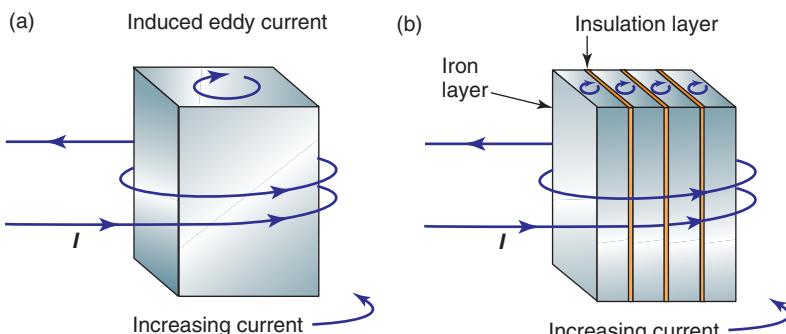


Figure 8.19 Eddy currents in (a) an ordinary iron core, (b) a laminated iron core

Another method for reducing eddy current losses in transformers is to use materials called *ferrites*, which are complex oxides of iron and other metals. These materials are good transmitters of magnetic flux, but are poor conductors of electricity, so the magnitudes of eddy currents are significantly reduced.

8.4 POWER DISTRIBUTION

Power stations are usually situated large distances from cities where most of the consumers are located. This presents problems with power losses in the transmission lines. Transmission lines are essentially long metallic conductors which have significant resistance. This means that they have a significant voltage drop across them when they carry a large current. This could result in greatly decreased voltages available to the consumer, as illustrated in sample problem 8.3 on page 152.

The resistance of a metallic conductor is proportional to its resistivity, ρ , its length, l , and is inversely proportional to its cross sectional area, A :

$$R = \frac{\rho l}{A}$$

The voltage drop, V , across a conductor equals the current, I , multiplied by the resistance, R . That is:

$$V = IR.$$

The rate of energy transfer in a conductor is called power, P , where:

$$P = VI.$$

If you know the current through a conductor and its resistance, the previous equation becomes:

$$P = I^2R.$$

Therefore, the power lost in a transmission line is given by the formula:

$$P_{\text{loss}} = I^2R$$

where

I = current flowing through the transmission line

R = the resistance of the transmission line.

SAMPLE PROBLEM

8.3

Transmission line calculations

A power station generates electric power at 120 kW. It sends this power to a town 10 km away through transmission lines that have a total resistance of 0.40 Ω . If the power is transmitted at 240 V, calculate:

- the current in the transmission lines
- the voltage drop across the transmission lines
- the voltage available in the town
- the power loss in the transmission lines.

SOLUTION

- For this calculation, use the station's power and the voltage across the transmission lines.

$$P = VI$$

$$\Rightarrow I = \frac{P}{V}$$

$$= \frac{120\,000}{240}$$

$$= 500 \text{ A}$$

$$(b) \quad V = IR$$

$$= 500 \times 0.40$$

$$= 200 \text{ V}$$

$$(c) \quad V_{\text{town}} = V_{\text{station}} - V_{\text{lines}}$$

$$= 240 - 200$$

$$= 40 \text{ V}$$

$$(d) \quad P_{\text{loss}} = I^2R$$

$$= (500)^2 \times 0.40$$

$$= 100\,000 \text{ W}$$

$$= 100 \text{ kW}$$

Using transformers to reduce power loss

Sample problem 8.3 demonstrates the difficulties involved in transmitting electrical energy at low voltages over large distances. The solution is to use transformers to step up the voltage before transmission. If the voltage is increased, the current is reduced. Recall that the power lost in transmission lines is given by the formula:

$$P_{\text{loss}} = I^2R$$

If the transmission voltage is doubled, the current is halved and the power loss is reduced by a factor of four. If the current is reduced by a factor of 10, the power loss is reduced by a factor of 100, and so on.



8.4

Transmission line power losses

eBookplus

eModelling: Modelling power transmission

Use a spreadsheet to model the energy loss incurred over a distance.

doc-0040

Using transformers enables electricity to be supplied over large distances without wasting too much electrical energy. This has had a significant effect on society. If transformers were not used in the power distribution system, either power stations would have to be built in the cities and towns or the users of electricity would have to be located near the power stations. The latter would mean that industries and population centres would have to be located near the energy sources such as hydro-electric dams and coal mines. The former would mean that fossil fuel stations would dump their pollution on the near-by population centres.

NSW electrical distribution system

The New South Wales electrical distribution system is shown in figure 8.20. The Bayswater power station in the Hunter Valley has four 660-megawatt generators that each have an output voltage of 23 kV. (Each set of coils in the generators has an output of 220 MW.) The three-phase power then enters a transmission substation where transformers step up the voltage to 330 kV.

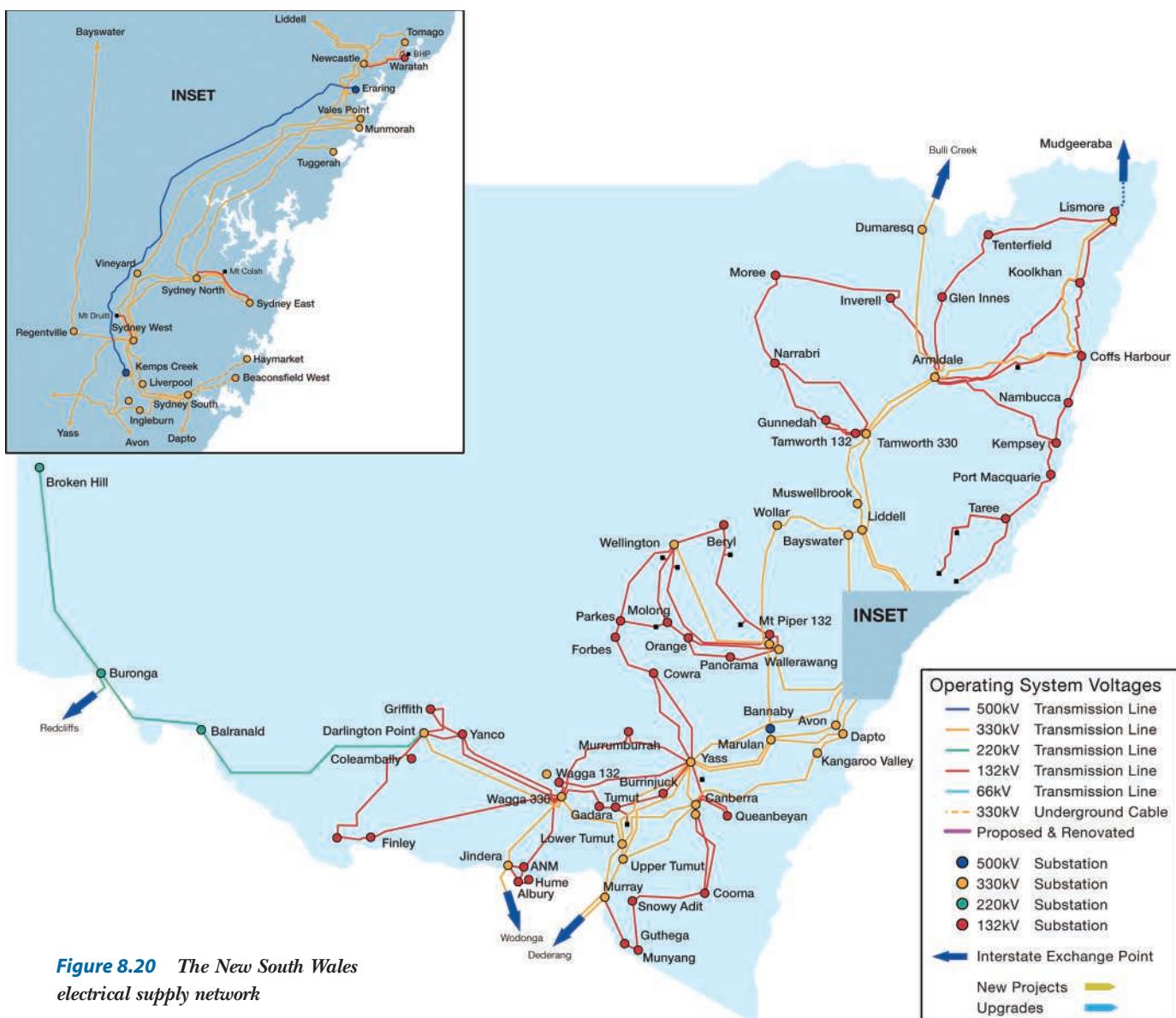


Figure 8.20 The New South Wales electrical supply network



The transmission lines end at a terminal station where the voltage is stepped down to 66 kV for transmission to zone power substations like the one shown in figure 8.21. There is usually a zone substation in each regional centre and in each municipality of a city. Here the voltage is stepped down to values of 11 kV and 22 kV. Power substations can perform three tasks:

1. step down the voltage using transformers
2. split the distribution voltage to go in different directions
3. enable, using circuit breakers and switches, the disconnection of the substation from the transmission grid or sections of the distribution grid to be switched on and off.

Figure 8.21 Transformers at a substation

Finally, pole transformers, as shown in figure 8.22, step the voltage down to 415 V for industry and 240 V for domestic consumption.



Figure 8.22 A pole transformer

PHYSICS IN FOCUS

Protecting power transmission lines from lightning

When lightning strikes, it will usually pass between the bottom of a thundercloud and the highest point on the Earth below. This means that it will strike tall trees, the tops of buildings such as church spires and the metal power towers used to support high voltage power transmission lines. Many such power towers have a cable running between them known as the continuous earth line. This cable

normally carries no current, but it may carry a current if a fault develops in the system. A second function of this cable is that it acts as a continuous lightning conductor. If this cable or a tower is struck by lightning, the electricity of the lightning will be conducted to the Earth by the metal towers and the transmission lines will not suffer from a sudden surge of voltage that could damage substations.

PHYSICS IN FOCUS

Insulating transmission lines

In dry air sparks can jump a distance of 1 cm for every 10 000 V of potential difference. Therefore, a 330 kV line will spark to a metal tower if it comes within a distance of 33 cm. In high humidity conditions the distance is larger. To prevent sparks jumping from transmission lines to the metal support towers, large insulators separate them from each other. It is important that these insulators are strong and have high insulating properties. Suspension insulators, illustrated in figure 8.23, are used for all high voltage power lines operating at voltages above 33 kV, where the towers or poles are in a straight line. Note that the individual sections of the insulators are disk shaped. This is because dust and grime collect on the insulators and can become a conductor when wet. Many wooden poles catch fire after the first rain following a prolonged dry period because a current flows across wet dirty insulators. The disk shape of the insulator sections increases the distance that a current has to pass over the surface of the insulator and so decreases the risk. There is also less chance that dirt and grime will collect on the undersides of the sections, and these are also less likely to get wet.

Both the continuous earth wire and insulators are clearly visible in figure 8.1 at the start of this chapter.

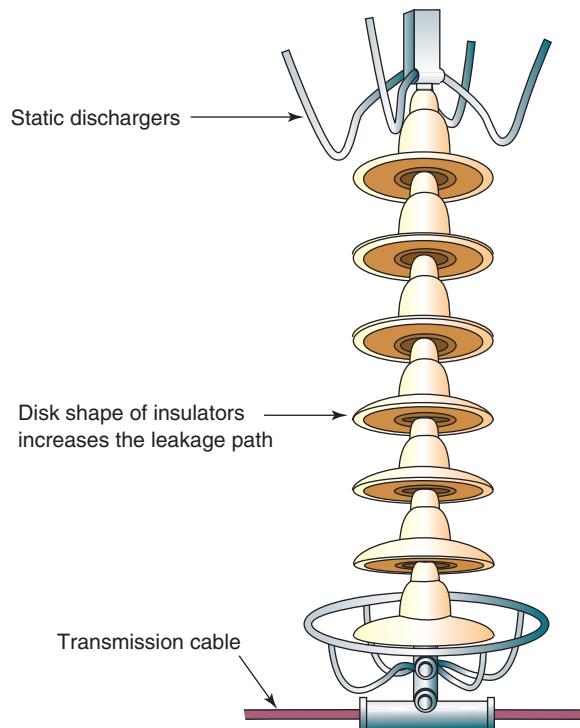


Figure 8.23 Suspension insulator used for high voltage transmission lines

PHYSICS IN FOCUS

Household use of transformers

Australian houses are provided with AC electricity that has a value of 240 V_{RMS}. Most electronic circuits are designed to operate at low DC voltages of between 3 V and 12 V. Therefore, household appliances that have electronic circuits in them will either have a 'power-cube' transformer that plugs directly into the power outlet socket, or have transformers built into them.

Power-cube transformers can be found in rechargeable appliances such as 'dust buster' vacuum cleaners, electric keyboards, answering machines, cordless telephones and laptop computers. You can probably find more in your own home. These transformers also have a rectifier circuit built into them that converts AC to DC.



The RMS value of an AC voltage is a way of describing a voltage that is continuously changing. The voltage actually swings between -339 V and +339 V at a frequency of 50 Hz. This voltage has the same heating effect on a metal conductor as a DC voltage of 240 V; hence, we usually describe it as 240 V.

Figure 8.24 Most electrical appliances operate at low voltages and have transformers built into them or come with power-cube transformers.

8.5 ELECTRICITY AND SOCIETY

The development of electrical generators and motors has affected many phases of modern life, but not always in the ways predicted.

It was first predicted that electric machines would do all physical labour. Workers would have more leisure time. Backbreaking housework would be eliminated by electrical gadgets so that people would have much more leisure time. What has happened instead is a reduction in unskilled jobs and an increase in unemployment.

Another prediction was that people could go back to living in the countryside (which was considered to be the ideal place to live) and that society would become more decentralised. This prediction arose during the industrial revolution of the nineteenth century when people were drawn to big cities where the power supplies were located. During this time the energy supplies were smog-producing steam driven machines. Electric power stations, however, are built at the source of their energy, the coal mines or where dams are built to provide hydro-electricity. The development of transformers and the power distribution system meant that cities, factories and other industries could be located at large distances from the power stations.

Rather than going back to living in the countryside, however, the middle classes moved out into the outer suburbs, living in bigger houses on bigger blocks of land. The poorer citizens were concentrated into poorer inner suburbs. However, the benefits of modern society that follow from the use of electrical energy are available to most rural communities.

The use of electrical generators and motors has also had a dramatic effect on the environment. Fossil fuel power stations have environmental effects such as thermal pollution, acid rain and air pollution due to the release of particles and oxides of nitrogen and sulfur. The huge amounts of carbon dioxide released by power stations contribute to the enhanced greenhouse effect that is thought to be raising the Earth's temperature.

SUMMARY

- A changing magnetic flux through a coil can induce a voltage across the terminals of the coil.
- A generator is a device that transforms mechanical energy into electrical energy.
- One type of generator has a rotating coil in an externally produced magnetic field.
- An AC generator uses slip rings to connect its terminals to the coil.
- A DC generator uses a split ring commutator to connect its terminals to the coil.
- Transformers are devices that can convert an AC input voltage signal to a higher or lower AC output voltage.
- A transformer consists of a primary and secondary coil, usually linked by a soft iron core.
- The following equations apply to ideal transformers:

$$\frac{V_p}{V_s} = \frac{n_p}{n_s} \text{ and } \frac{I_s}{I_p} = \frac{n_p}{n_s}$$

- Losses in transmission lines can be calculated using the formula $P_{\text{loss}} = I^2 R$.
- Power losses in transmission lines can be reduced by using step-up transformers to increase the transmission voltage, thereby reducing the transmission current.

QUESTIONS

- Identify the types of energy transformation that occur in electrical generators.
- Figure 8.25 shows a generator.
 - Name the parts of the generator labelled A, B, C, D and E.
 - Describe the function of each of these parts.
 - What type of generator (AC or DC) is this?

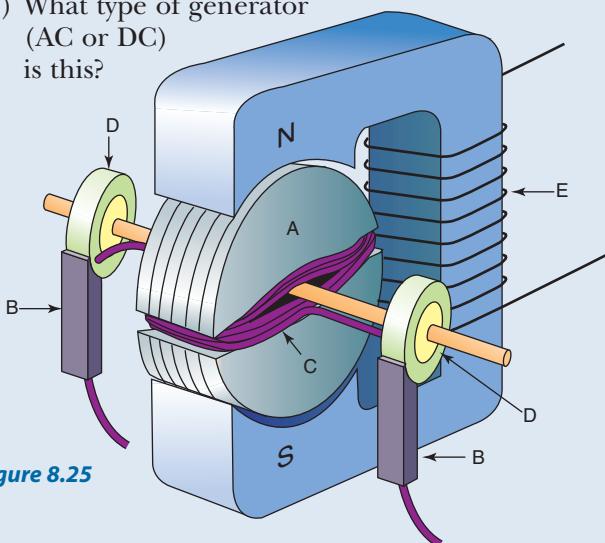


Figure 8.25

- A hand-turned generator is connected to a light globe in series with a switch. Is it easier to turn the coil when the switch is open or when it is closed? Explain your reasoning.
- Describe the effect of the following changes on the size of the current produced by a generator:
 - the number of loops of the coil is increased
 - the rate of rotation of the coil is decreased
 - the strength of the magnetic field is increased
 - an iron core is used in the coil rather than having an air coil.
- A student builds a model electric generator, similar to that shown in figure 8.26. The coil consists of 50 turns of wire. The student rotates the coil in the direction indicated on the diagram. The coil is rotated a quarter turn (90°).
 - In which direction does the current flow through the external load?
 - Sketch a flux versus time graph and an emf versus time graph as the coil completes one rotation at a steady rate starting from the instant shown in the diagram.

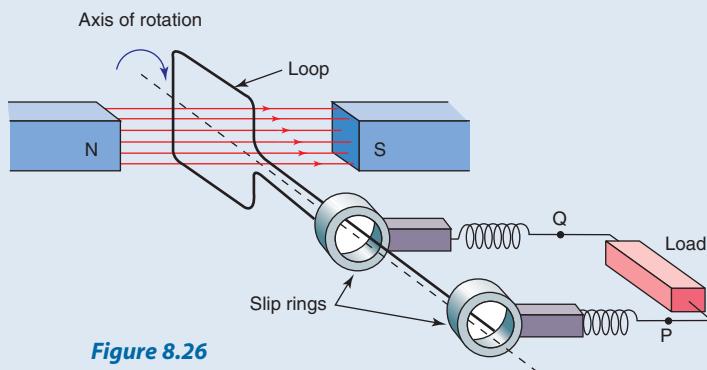


Figure 8.26

- A rectangular coil of wire is placed in a uniform magnetic field, B , that is directed out of the page. This is shown in figure 8.27(a). At the instant shown the coil is parallel to the page.

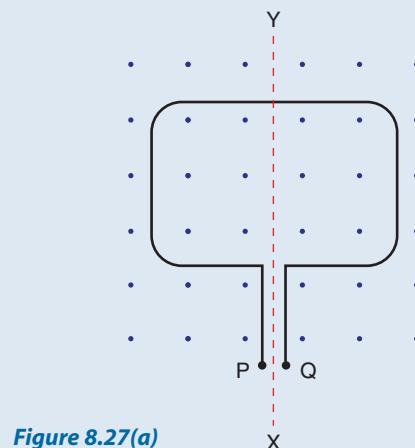


Figure 8.27(a)

The coil rotates about the axis XY at a steady rate. The time variation of the voltage drop induced between points P and Q for one complete rotation is shown in figure 8.27(b).

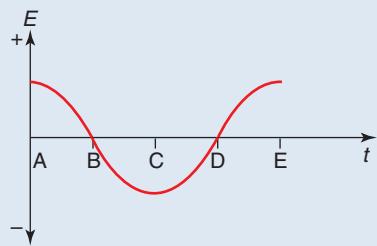


Figure 8.27(b)

- At which time(s) could the coil be in the position shown in figure 8.27(a)? Justify your answer.
- Which of the graphs in figure 8.27(c) shows the variation of voltage versus time if the coil is rotated at twice the original speed?

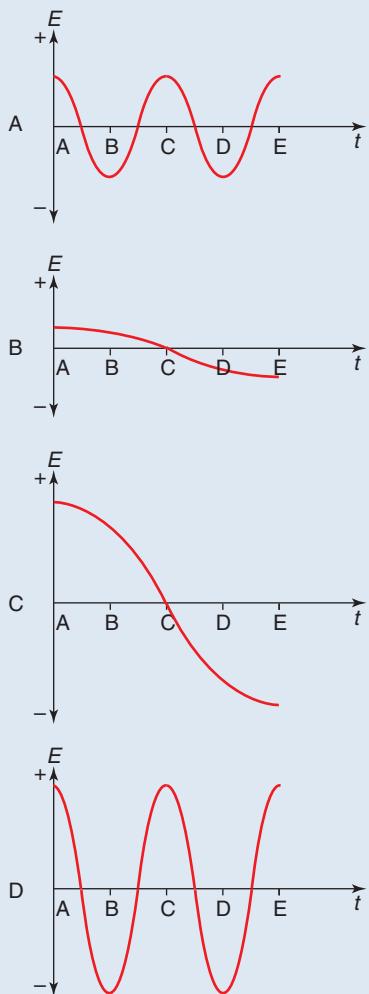


Figure 8.27(c)

- In a power station, an electromagnet is rotated close to a set of coils. The electromagnet is supplied with a direct current. An idealised diagram of this arrangement is shown in figure 8.28.

- In this arrangement, which part(s) make up the rotor and which make up the stator?
- Explain why it is necessary to rotate the electromagnet to produce an emf.
- Explain why energy must be provided to the electromagnet to keep it rotating at a constant speed.

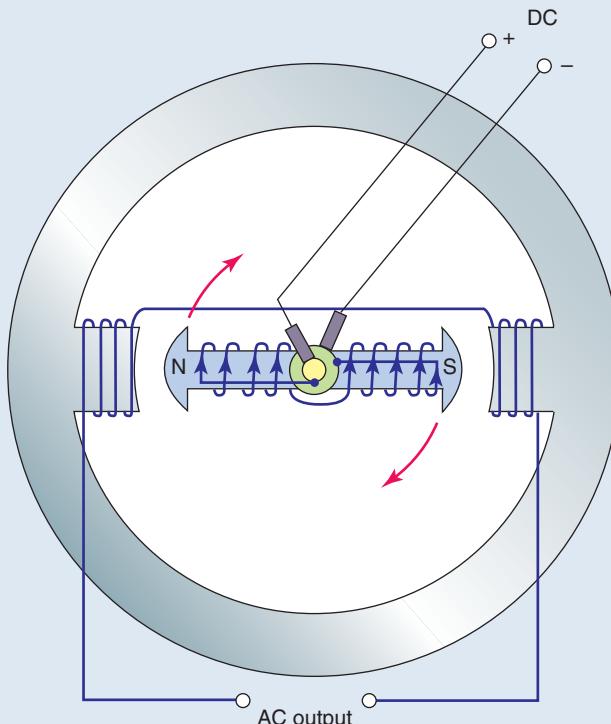


Figure 8.28

- Describe the main difference between AC and DC generators.
- Draw a labelled cross-section diagram of a simple transformer. Referring to your diagram, explain how it operates in terms of the principles of electromagnetic induction.
- Explain why a steady DC current input will not operate a transformer.
- A transformer changes 240 V to 15 000 V. There are 4000 turns on the secondary coil.
 - Identify what type of transformer this is.
 - Calculate how many turns there are on the primary coil.
- A doorbell is connected to a transformer that has 720 turns in the primary coil and 48 turns in the secondary coil. If the input voltage is 240 V AC, calculate the voltage that is delivered to the doorbell.
- A school power pack that operates from a 240 V mains supply consists of a transformer with 480 turns on the primary coil. It has two outputs, 2 V AC and 6 V AC.

- (a) Calculate the number of turns on the 2 V and 6 V secondary coils.
- (b) The power rating of the transformer is 15 W. Calculate the maximum current that can be drawn from the 6 V secondary coil.
14. An ideal transformer has 100 turns on the primary coil and 2000 turns on the secondary coil. The primary voltage is 20 V. The current in the secondary coil is 0.50 A. Calculate:
- the secondary voltage
 - the output power
 - the input power
 - the current flowing through the primary coil.
15. A transformer has 110 turns on the primary coil and 330 turns on the secondary coil.
- Identify this type of transformer.
 - By what factor does it change the voltage?
16. An ideal transformer is designed to provide 9.0 V output from a 240 V input. The primary coil is fitted with a 1.0 A fuse.
- Calculate the ratio

$$\frac{\text{number of turns on the primary coil}}{\text{number of turns on the secondary coil}}$$
 - Calculate the maximum current that can be delivered from the output terminals.
17. A neon sign requires 12 kV to operate. Calculate the ratio

$$\frac{\text{number of turns on the primary coil}}{\text{number of turns on the secondary coil}}$$
- of the sign's transformer if it is connected to a 240 V supply.
18. A 20.0 W transformer gives an output voltage of 25 V. The input current is 15 A.
- Calculate the input voltage.
 - Is this a step-up or step-down transformer?
 - Calculate the output current.
19. Describe the advantages gained by transmitting AC electrical power at high voltages over large distances.
20. A power station generates electric power at 120 kW. It sends this power to a town 10 kilometres away through transmission lines that have a total resistance of $0.40\ \Omega$. If the power is transmitted at 5.00×10^5 V, calculate:
- the current in the transmission lines
 - the voltage drop across the transmission lines
 - the voltage available at the town
 - the power loss in the transmission lines.
21. A generator coil at the Bayswater power station produces 220 MW of power in a single phase at 23 kV. This voltage is stepped up to 330 kV by a transformer in the transmission substation. Calculate:
- the ratio

$$\frac{\text{number of turns on the primary coil}}{\text{number of turns on the secondary coil}}$$

for the transformer
 - the output power of the transformer
 - the current in the transmission line.
22. A generator has an output of 20 kW at 4.0 kV. It supplies a factory via two long cables with a total resistance of $16\ \Omega$.
- Calculate the current in the cables.
 - Calculate the power loss in the cables.
 - Calculate the voltage between the ends of the cables at the factory.
 - Describe how the power supplied to the workshop could be increased.
23. Write a brief essay discussing one of the following topics.
- Describe the impact that the development of AC and DC generators has had on society.
 - Describe what life would be like if transformers had not been invented.



8.1 OBSERVING THE OUTPUT OF A HAND-OPERATED GENERATOR

Aim

- To observe the output voltage of an AC generator using a cathode ray oscilloscope.
- To use a DC motor as a generator and to observe its output voltage using a cathode ray oscilloscope.
- To observe what happens when the coil of a generator is rotated at different speeds.
- To compare the force required to rotate a coil of a generator when a load is connected to the generator.

Apparatus

hand-cranked model generator
DC motor
cathode ray oscilloscope (CRO)
2 V light globe
connecting wires

Theory

Rotating a coil in a magnetic field induces a voltage across the terminals of the coil.

An AC generator connects to the coil with slip rings.

A DC generator uses a split-ring commutator.

When a current flows through the coil of a generator there is a magnetic force that opposes the motion of the coil.

Method

- Connect the AC generator to the CRO as shown in figure 8.29.
- Turn the coil slowly and describe the trace on the CRO, noting the period, the peak voltage obtained and the shape of the trace. Sketch the trace.
- Repeat step 2, rotating the coil quickly.
- Connect the light globe across the terminals of the generator. Comment on the ease of rotating the coil compared with when the globe was not in use.
- Connect the CRO across the terminals of a DC motor.
- Turn the shaft of the motor by hand and describe the trace on the CRO, noting the period, the peak voltage obtained and the shape of the trace. Sketch the trace.

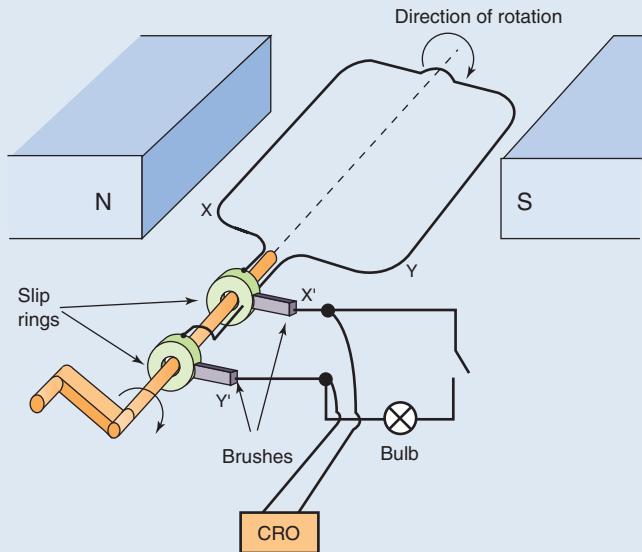


Figure 8.29

Analysis

Relate your observations to the theory presented in this chapter.



8.2 MAKING A SIMPLE TRANSFORMER

Aim

- To set up a simple transformer
- To observe that a steady DC current does not produce an output current from a transformer.

Apparatus

galvanometer
two coils having different numbers of turns, one coil fitting inside the other.
1.5 V cell
switch
variable resistor
connecting wires

Theory

A transformer consists of a primary and secondary coil. In this experiment the primary coil fits inside the secondary coil. The changing flux produced in the primary coil induces a current in the secondary coil.

Method

- Place the smaller (primary) coil in the larger (secondary) coil. Connect the secondary coil to the galvanometer. Connect the primary coil in series with the switch, variable resistor and cell as shown in figure 8.30.

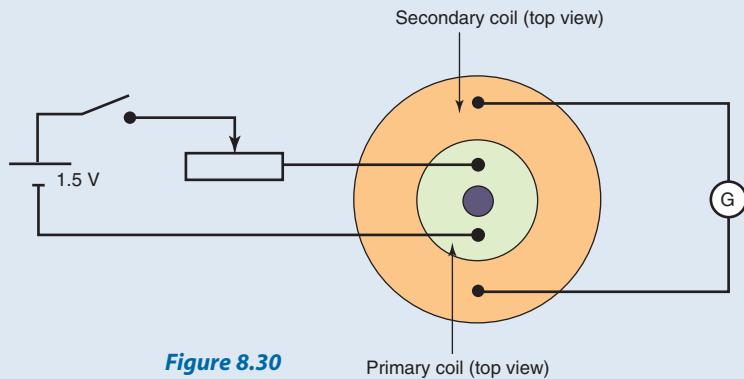


Figure 8.30

- Set the variable resistor to its lowest value.
- Observe the effects on the galvanometer as you close the switch, keep it closed for five seconds and then open the switch.
- Describe what happens.
- Close the switch and change the value of the variable resistor slowly and rapidly. Open the switch.
- Record your observations.

Analysis

- Relate your observations to Faraday's Law of Electromagnetic Induction.
- Describe the conditions necessary for the operation of a transformer.



8.3

TRANSFORMER INS AND OUTS

Aim

To use an AC input voltage to produce an AC output voltage and to compare their values.

Apparatus

two coils having a different number of turns of wire, preferably with the number of turns on each being known, one coil fitting inside the other

two 2 V light globes in holders.

AC power source
switch

dual trace cathode ray oscilloscope
connecting wires

long iron nails that fit inside the smaller coil to produce a soft iron core

Theory

An AC input voltage will induce an AC output voltage.

If the secondary coil has more turns than the primary, the device is a step-up transformer and the secondary voltage will be greater than the primary voltage.

The inclusion of an iron core increases the effectiveness of the transformer.

Method

- Place the smaller (primary) coil in the other (secondary) coil.
- Connect the secondary coil to a light globe.
- Connect the primary coil in series with the switch, a light globe and the AC source.
- Adjust the AC source to its lowest value (less than 2 V).
- Connect one set of input leads of the CRO across the terminals of the primary coil, and the other across the terminals of the secondary coil. The set-up is illustrated in figure 8.31.
- Close the switch and observe the traces on the CRO.
- Compare the primary and secondary peak voltages and frequencies.
- Insert iron nails in the primary coil and repeat steps 6 and 7.

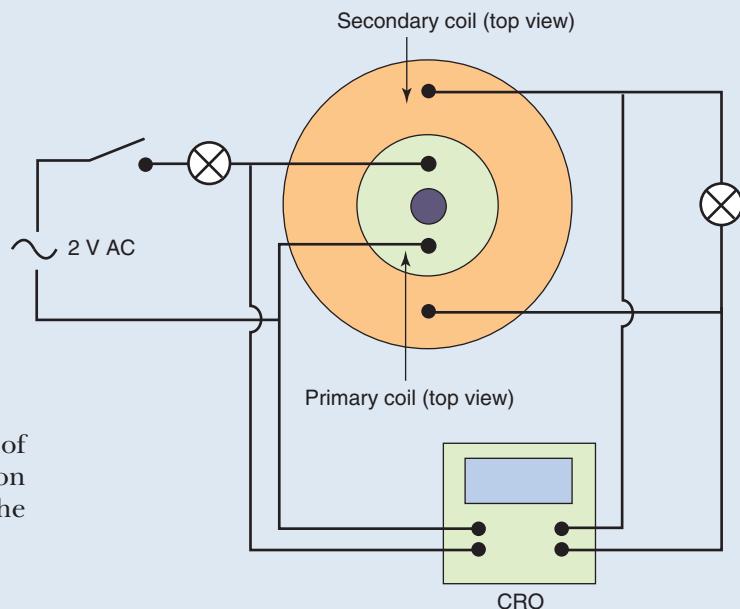
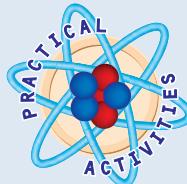


Figure 8.31

Analysis

- Explain your observations in terms of the theory you have studied.
- What effect did the insertion of the iron nails have on the effectiveness of the transformer?



8.4 TRANSMISSION LINE POWER LOSSES

Aim

- To investigate the effects of resistance in transmission lines
- To investigate the use of transformers in power distribution systems.

Apparatus

Transmission line experiment

(Available from Haines Educational P/L,
www.haines.com.au)

Theory

Power losses in transmission lines can be reduced by using transformers to step up the voltage for transmission and step it down again for use. This kit models transmission lines by using resistance wire.

Method

1. Set up the equipment as described in the instruction brochure.
2. Transmit power from an AC supply to the load globe using the transmission lines alone. Note the brightness of the globe and the current in the transmission lines.
3. Measure the voltage output of the supply and the voltage at the globe.
4. Repeat steps 1 and 2, this time using the transformers.

Analysis

Comment on your results.

CHAPTER 9 AC ELECTRIC MOTORS

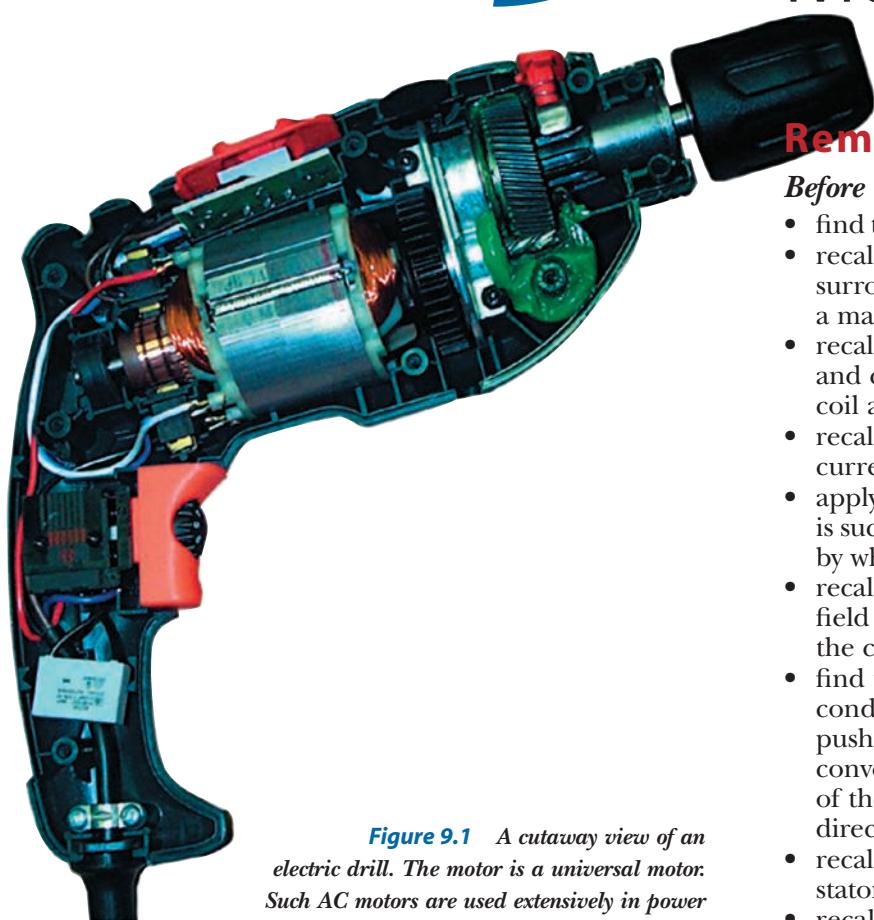


Figure 9.1 A cutaway view of an electric drill. The motor is a universal motor. Such AC motors are used extensively in power tools and household appliances.

Remember

Before beginning this chapter, you should be able to:

- find the polarity of a solenoid by using the right hand-rule
- recall that an electromagnet is a soft iron core surrounded by a coil of wire. An electromagnet acts like a magnet when a current flows through the coil.
- recall that voltages and currents can be induced in wires and coils by the relative movement between the wire or coil and a magnetic field
- recall that a changing magnetic field induces voltages and currents in wires and coils
- apply Lenz's Law: The direction of the induced current is such that its magnetic action tends to resist the change by which it is produced.
- recall that a current-carrying conductor in a magnetic field experiences a force. This force is at a maximum when the current is perpendicular to the magnetic field.
- find the direction of the force on a current-carrying conductor in a magnetic field by using the right-hand push rule: The thumb points in the direction of the conventional current. The fingers point in the direction of the external magnetic field. The palm indicates the direction of the force.
- recall that generators and motors have two main parts, the stator and the rotor
- recall that alternating current (AC) periodically changes direction. In Australia, AC electricity has a frequency of 50 Hz.
- recall that a power station has generators that have three sets of coils in their stators set 120° apart. They produce three AC currents that are 120° out of phase with each other. This is known as three-phase power generation.
- recall that a transformer consists of two coils, known as the primary and secondary, which are usually linked by a soft iron core. Transformers step up or step down AC voltages.

Key content

At the end of this chapter you should be able to:

- describe the main features of an AC motor
- identify some energy transfers and transformations involving conversion of electrical energy into more useful forms.

As we have seen in previous chapters, alternating current (AC) is widely used in today's world. It is easier to produce in power generating stations and easier to distribute over large distances with small energy losses due to the use of transformers. AC electricity is also produced at a very precise frequency. In Australia this frequency is 50 Hz.

AC motors are used when very precise speeds are required, for example in electric clocks. AC motors operate using an alternating current (AC) electrical supply. Electrical energy is usually transformed into rotational kinetic energy.

9.1 MAIN FEATURES OF AN AC MOTOR

As with the DC motors, AC generators and DC generators that have been studied earlier, AC motors have two main parts. These are called the stator and the rotor.

The stator is the stationary part of the motor and it is usually connected to the frame of the machine. The stator of an AC motor provides the external magnetic field in which the rotor rotates. The magnetic field produces a torque on the rotor.

Most AC motors have a cylindrical rotor that rotates about the axis of the motor's shaft. This type of motor usually rotates at high speed, with the rotor completing about one revolution for each cycle of the AC electricity supply. This means that Australian AC motors rotate at about 50 revolutions per second or 3000 revolutions per minute. If slower speeds are required, they are achieved using a speed-reducing gearbox. This type of motor is found in electric clocks, electric drills, fans, pumps, compressors, conveyors, and other machines in factories. The rotor is mounted on bearings that are attached to the frame of the motor. In most AC motors the rotor is mounted horizontally and the axle is connected to a gearbox and fan. The fan cools the motor.

Both the rotor and the stator have a core of ferromagnetic material, usually steel. The core strengthens the magnetic field. The parts of the core that experience alternating magnetic flux are made up of thin steel laminations separated by insulation to reduce the flow of eddy currents that would greatly reduce the efficiency of the motor.

The universal motor

There are two main classifications of AC motors. Single-phase motors operate on one of the three phases produced at power generation plants. Single-phase AC motors can operate on domestic electricity. Polyphase motors operate on two or three of the phases produced at power generation plants. One type of single-phase AC electric motor is the **universal motor**.

Universal motors are designed to operate on DC and AC electricity. They are constructed on similar lines to the DC motor studied in chapter 6. The rotor has several coils wound onto the rotor armature. The ends of these coils connect to opposite segments of a commutator. The external magnetic field is supplied by the stator electromagnets that are connected in series with the coils of the armature via brushes. The interaction between the current in a coil of the armature and the external magnetic field produces the torque that makes the rotor rotate. Even though the direction of the current is

A **universal motor** is a series-wound motor that may be operated on either AC or DC electricity.

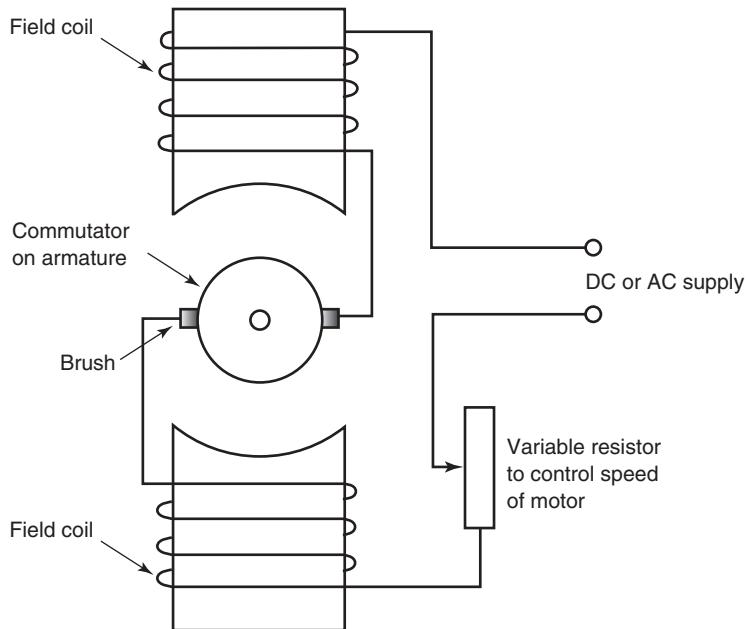


Figure 9.2 A universal motor

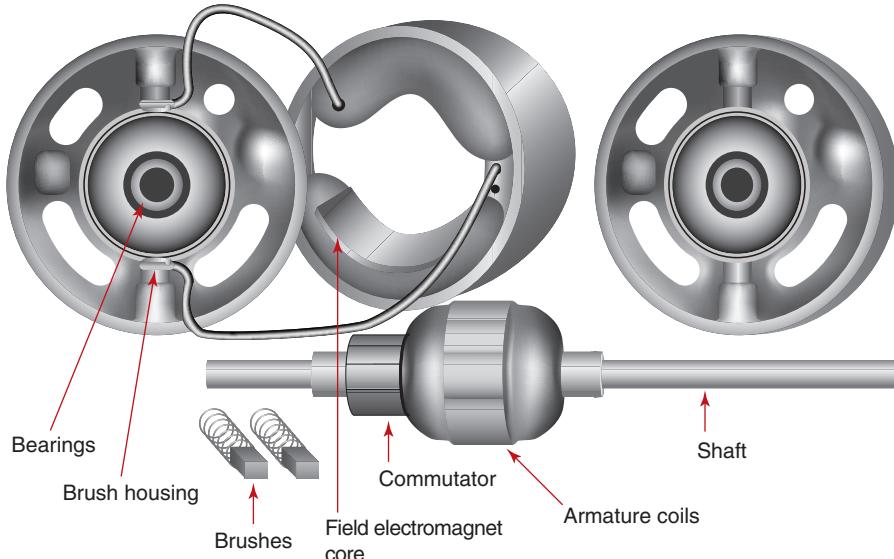


Figure 9.3 A dismantled universal motor

An **induction motor** is an AC machine in which torque is produced by the interaction of a rotating magnetic field produced by the stator and currents induced in the rotor.

changing 100 times per second when the motor is connected to the mains, the universal motor will continue to rotate in the same direction because the magnetic field flux of the stator is also changing direction 100 times every second.

A variable resistor controls the speed of a universal motor by varying the current through the coils of the armature and the field coils of the stator. The universal motor is commonly used for small machines such as portable drills and food mixers. Figure 9.2 shows a schematic diagram of a universal motor, and figure 9.3 shows a diagram of a universal motor that has been taken apart.

AC induction motors

Induction motors are so named because a changing magnetic field that is set up in the stator induces a current in the rotor. This is similar to what happens in a transformer, with the stator corresponding to the primary coil of the transformer and the rotor corresponding to the secondary. One difference is that in an induction motor the two parts are separated by a thin air gap. Another difference is that in induction motors the rotor (secondary coil) is free to move.

The simplest form of AC induction motor is known as the squirrel-cage motor. It is called a squirrel-cage motor because the rotor resembles the cage or wheel that people use to exercise their squirrels or pet mice. It is an induction motor because no current passes through the rotor directly from the mains supply. The current in the rotor is induced in the conductors that make up the cage of the rotor by a changing magnetic field, as explained later in this chapter.

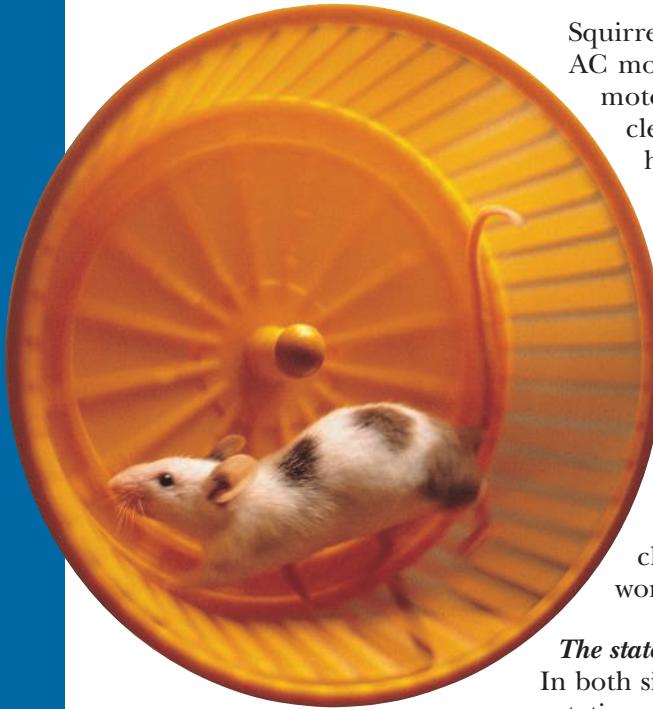


Figure 9.4 A mouse exercise wheel is similar to a squirrel cage.

Squirrel-cage induction motors are by far the most common types of AC motor used domestically and in industry. Squirrel-cage induction motors are found in some power drills, beater mixes, vacuum cleaners, electric saws, hair dryers, food processors and fan heaters, to name but a few.

The structure of AC induction motors

The easiest type of induction motor to understand is the three-phase induction motor. This operates by using each phase of AC electricity that is generated in power stations and supplied to factories.

Household electric motors are single-phase motors. This is because houses are usually supplied with only one phase of the three phases that are produced in power stations. It is not important to understand how the rotating magnetic field is achieved in single-phase AC induction motors; therefore, this chapter will concentrate on the three-phase motor, as its workings are easier to visualise.

The stator of three-phase induction motors

In both single- and three-phase AC induction motors, the stator sets up a rotating magnetic field that has a constant magnitude. The stator of a three-phase induction motor usually consists of three sets of coils that have iron cores. The stator is connected to the frame of the motor and surrounds a cylindrical space in which it sets up a rotating magnetic field. In three-phase induction motors, this is achieved by connecting each of the three pairs of field coils to a different phase of the mains electrical supply. The coils that make a pair are located on opposite sides of the stator and they are linked electrically. The magnetic field inside the stator rotates at the same frequency as the mains supply; that is, at 50 Hz. A cutaway diagram of a stator is shown in figure 9.5.

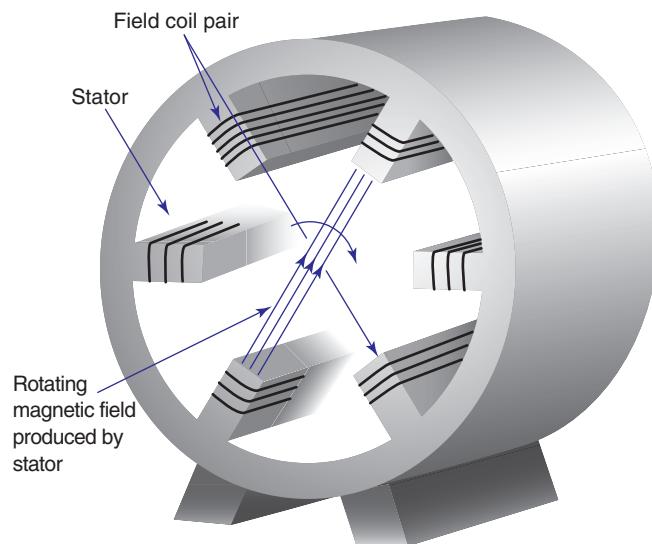


Figure 9.5 The rotating magnetic field set up by the stator. Note that in this stator there are three pairs of field coils and that each pair is connected.

The magnetic field rotates at exactly the same rate as the electromagnet in the power station generator that provides the AC electricity. Each pair of coils in the stator of the generator supplies a corresponding pair of coils in the stator of the motor. Therefore, the magnetic field in the motor rotates at exactly the same rate as the electromagnet in the generator. This is represented in figure 9.6.

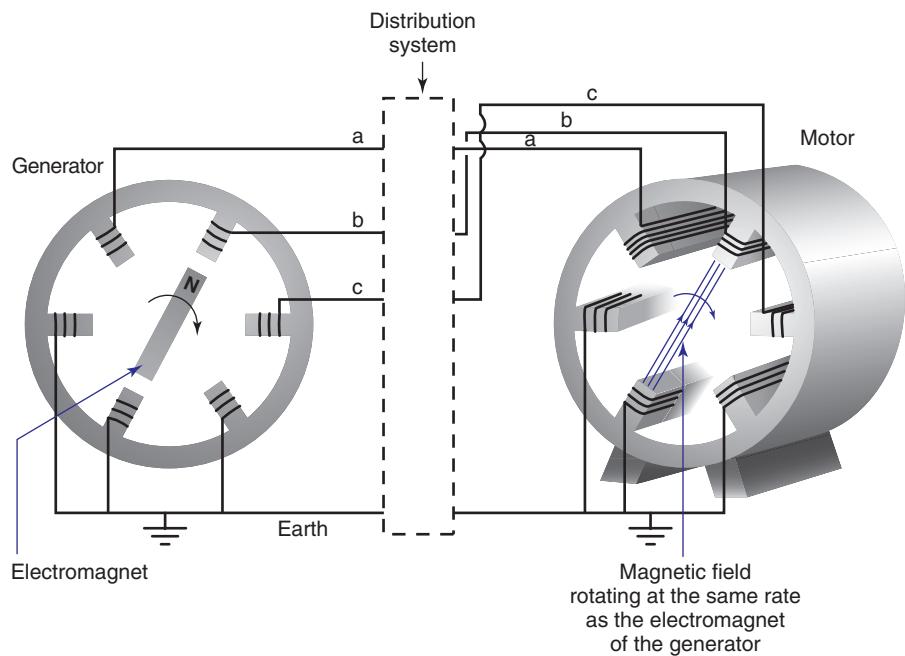


Figure 9.6 Supplying three-phase electrical power to the motor

A **squirrel-cage rotor** is an assembly of parallel conductors and short-circuiting end rings in the shape of a cylindrical squirrel cage.

As the electromagnet of the generator rotates, it influences three sets of linked coils. This produces three sets of AC voltage signals. These signals are ‘out of phase’ with each other. See figure 8.14 on page 146.

The squirrel-cage rotor

The rotor of the AC induction motor consists of a number of conducting bars made of either aluminium or copper. These are attached to two rings, known as end rings, at either end of the bars. This forms an object that is sometimes called a **squirrel-cage rotor** (see figure 9.7). The end rings ‘short-circuit’ the bars and allow a current to flow from one side to the other of the cage.

The bars and end rings are encased in a laminated iron armature as shown in figure 9.8. The iron intensifies the magnetic field passing through the conductors of the rotor cage and the laminations decrease the heating losses due to eddy currents. The armature is mounted on a shaft that passes out through the end of the motor. Bearings reduce friction and allow the armature to rotate freely.

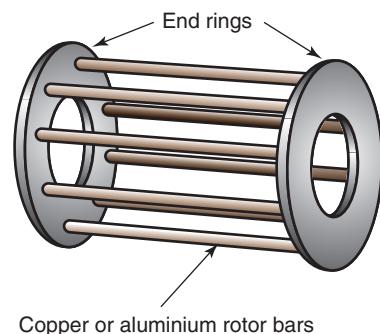


Figure 9.7 A squirrel-cage rotor

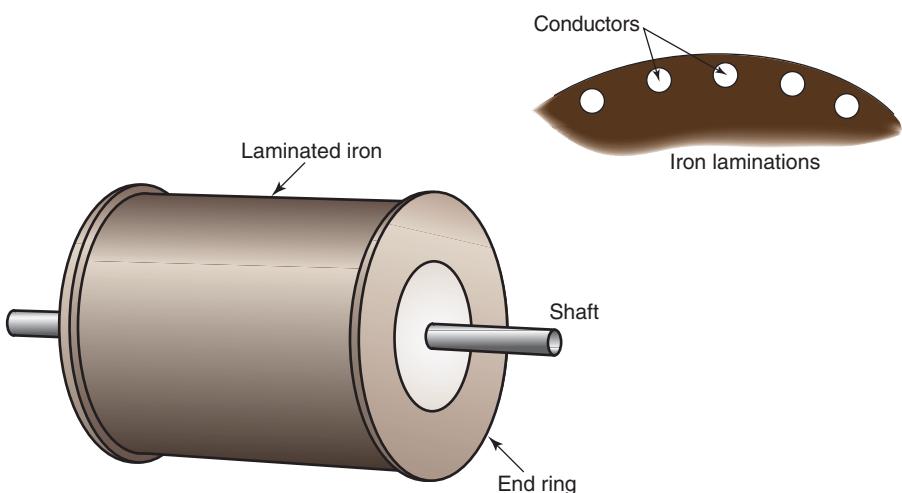


Figure 9.8 The rotor of an AC motor



9.1

Demonstrating the principle of an AC induction motor

Figure 9.9, below shows a cutaway model of a fully assembled induction motor. Note the field coils of the stator and the squirrel-cage rotor with a laminated iron core. Also note that the shaft in this case is connected to a gearbox so that a lower speed than 3000 revolutions per minute can be achieved, and that the cooling fan is mounted on the shaft.

The operation of AC induction motors

As the magnetic field rotates in the cylindrical space within the stator, it passes over the bars of the cage. This has the same effect as the bars moving in the opposite direction through a stationary magnetic field. The relative movement of the bars through the magnetic field creates a current in the bars. Bars carrying a current in a magnetic field experience a force. The discussion on the opposite page shows that the force in this case is always in the same direction as the movement of the magnetic field. The cage is then forced to 'chase' the magnetic field around inside the stator.

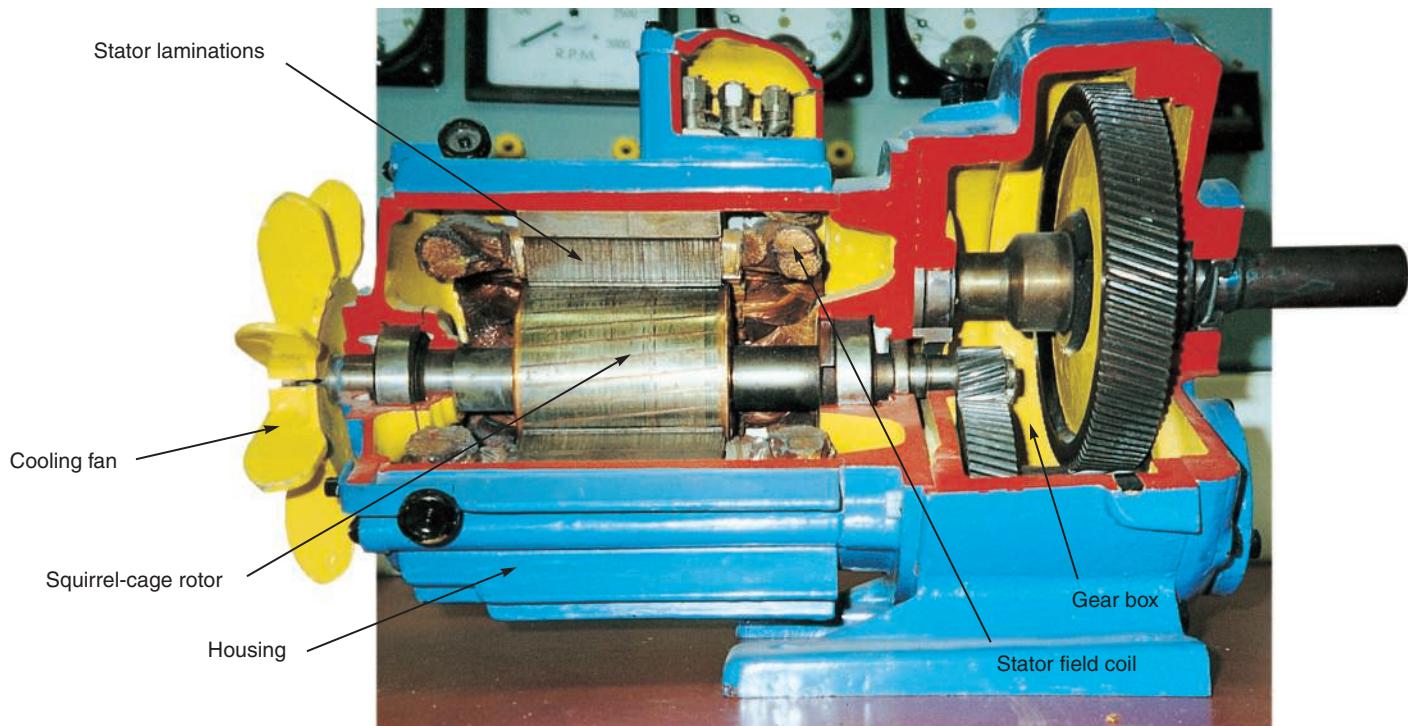


Figure 9.9 A cutaway view of an induction motor

Figure 9.10(a) on the following page shows an end view of the magnetic field as it moves across a conductor bar of the squirrel cage. The magnetic field moving to the right across the conductor bar has the same effect as the conductor bar moving to the left across the magnetic field. You can use the right-hand push rule to determine the direction of the induced current in the conductor. The thumb points to the left (the direction of movement for positive charges relative to the magnetic field), the fingers point up the page (the direction of the magnetic field) and the palm of the hand shows the direction of the force on positive charges and consequently the direction of the induced current. This will show that the current in the bar is flowing into the page.

There is now a current flowing in the conductor bar as shown in figure 9.10(b). The direction of the force acting on the induced current is determined using the right-hand push rule. Therefore, the force on the

conductor is to the right, which is in the same direction as the movement of the magnetic field. This is shown in figure 9.10(c).

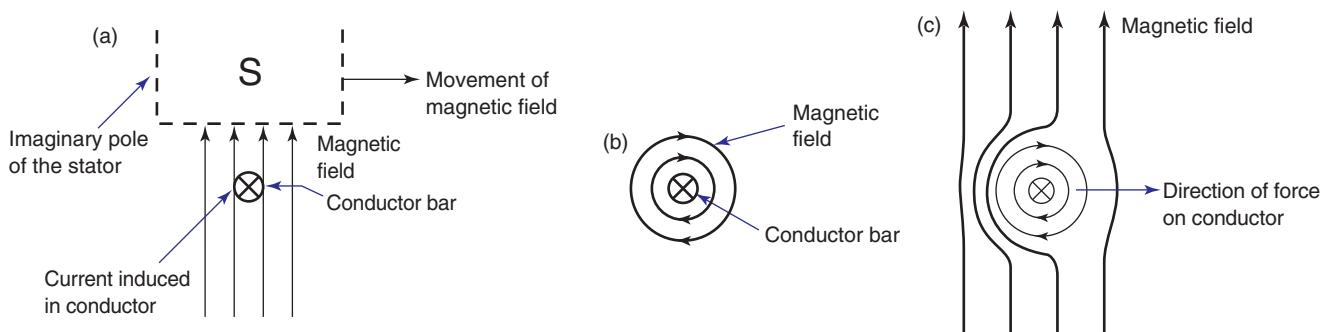


Figure 9.10 (a) The induced current in a conductor bar
 (b) The direction of the magnetic field of the induced current flowing in the bar
 (c) The force acting on the bar carrying the induced current

Slip

If the bars of the squirrel cage were to rotate at exactly the same rate as the magnetic field, there would be no relative movement between the bars and the magnetic field and there would be no induced current and no force. If the cage is to experience a force there must be relative movement, such as the cage constantly ‘slipping’ behind the magnetic field. When operating under a load, the retarding force slows the cage down so that it is moving slower than the field. The difference in rotational speed between the cage and the field is known as the **slip speed**. This means that the rotor is always travelling at a slower speed than the magnetic field of the stator when the motor is doing work.

When any induction motor does work, the rotor slows down. You can hear this happen when a beater mix is put into a thick mixture or when a power drill is pushed into a thick piece of wood. When this occurs, the amount of slip is increasing. This means that the relative movement between the magnetic field and the conductor bars is greater and that the induced current and magnetic force due to the current are increased.

Power of AC induction motors

Power is the rate of doing work. Work is done when energy is transformed from one type to another. Induction motors are considered to produce low power because the amount of mechanical work they achieve is low compared with the electrical energy consumed. The electrical power consumed by a motor is calculated using the formula $P = VI$, where V is the voltage at the terminals of the motor, and I is the current flowing through the coils of the stator. The ‘lost power’ of induction motors is consumed in magnetising the working parts of the motor and in creating induction currents in the rotor.

9.2 ENERGY TRANSFORMATIONS AND TRANSFERS

The Principle of Conservation of Energy states that energy cannot be created or destroyed, but it can be transformed from one type to another. ‘Transform’ means to change form. Energy transfers happen when energy moves from one place to another.

The energy transformations that occur when an electric appliance is operating depend on what the machine is doing. Consider the operation of a hair dryer as an example. The electric motor transforms electrical energy into mechanical energy (the rotor spins). Some electrical energy is also transformed into internal energy due to eddy currents in the laminated iron core. The mechanical energy of the rotor is transformed into sound and internal energy within the motor, and into the kinetic energy of air particles by a fan that is attached to the shaft of the rotor. Energy is transferred from the motor to the air by the rotor shaft and the fan. The air passes through a heating element where electrical energy is transformed into internal energy and light energy. Internal energy is then transferred out of the dryer by direct conduction to the air particles and by convection as the air particles carry the energy from the dryer.

SUMMARY

- AC electric motors are used in many machines in households and in industry.
- The basic operating principles of AC electric motors are the same as for DC motors. A current-carrying conductor in a magnetic field experiences a force, the direction of which is given by the right-hand push rule.
- The stator of an AC induction motor consists of field coils that set up a rotating magnetic field that has a constant magnitude. This field rotates 50 times every second.
- A squirrel-cage rotor has conductor bars that are short-circuited by two end rings. The bars are embedded in a laminated iron core. (The laminations reduce energy losses due to eddy currents.)
- A current is induced in the bars of the rotor as the magnetic field moves across them.
- The bars experience a force because they are carrying a current in a magnetic field. The direction of the force is the same as the direction of movement of the magnetic field. The rotor is therefore forced to chase the magnetic field.
- For the current to be induced in the rotor bars, the rotor must slip behind the magnetic field.

QUESTIONS

1. Explain why AC electric motors are used in situations that require accurate speed, such as clocks and tape recorders.
2. Describe how universal motors operate on both AC and DC electricity.
3. Describe the main features of an AC induction motor.
4. Figure 9.11 shows a cutaway diagram of an AC induction motor.
 - (a) Name the parts labelled A to E.
 - (b) Explain the functions of the parts labelled A to E.

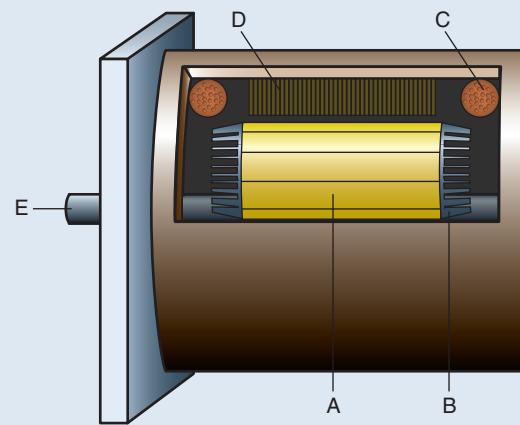
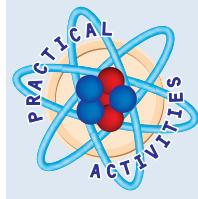


Figure 9.11

5. Briefly describe how the rotating magnetic field is produced in a three-phase AC electric motor.
6. Describe the construction of a squirrel-cage rotor.
7. Explain how a current is produced in the rotor bars of an AC induction motor.
8. Describe the purpose of the end rings of a squirrel cage.
9. (a) What is slip?
 (b) Explain why slip is necessary for the operation of a squirrel-cage induction motor.
10. Account for the ‘lost power’ of induction motors.
11. Describe the energy transformations and transfers that occur in the operation of a household electric mixer.



9.1 DEMONSTRATING THE PRINCIPLE OF AN AC INDUCTION MOTOR

Aim

- To investigate the direction of a current in a conductor that is moving relative to a magnetic field
- To investigate the factors affecting the magnitude of a current in a conductor that is moving relative to a magnetic field.

Apparatus

a copper or aluminium rod. If none are available, a length of copper wire will do.
two bar magnets or a horseshoe magnet
a galvanometer
connecting wires

Theory

A current is induced in a conductor when there is relative movement between it and a magnetic field.

The magnitude depends on the orientation of the conductor to the field, the speed of the relative motion between the conductor and the magnetic field and the strength of the magnetic field.

Method

Set up the apparatus as shown in figure 9.12. Place two bar magnets on two books with an N pole on the left and an S pole on the right.

Note the separation of the poles, as this affects the strength of the magnetic field.

1. Connect the galvanometer to the conductor.
2. Move the conductor downward slowly between the poles of the magnet. Note the magnitude and direction of the current through the conductor.

3. Move the conductor downward quickly between the poles of the magnet. Note the magnitude and direction of the current through the conductor.
4. Move the conductor up slowly between the poles of the magnet. Note the magnitude and direction of the current through the conductor.
5. Move the conductor upward quickly between the poles of the magnet. Note the magnitude and direction of the current through the conductor.
6. Place the conductor at an angle to the magnetic field and move it upward quickly between the poles of the magnet. Compare the magnitude of the current through the conductor with the result attained in step 5.
7. Double the separation of the magnetic poles. What effect will this have on the strength of the magnetic field?
8. Repeat step 5.
9. Return the magnetic poles to their original separation. Support the conductor and move the poles upward past the conductor. Do you get the same direction of current as when the conductor moved downward between the poles?

Analysis

Relate your observations to the theory presented in this chapter.

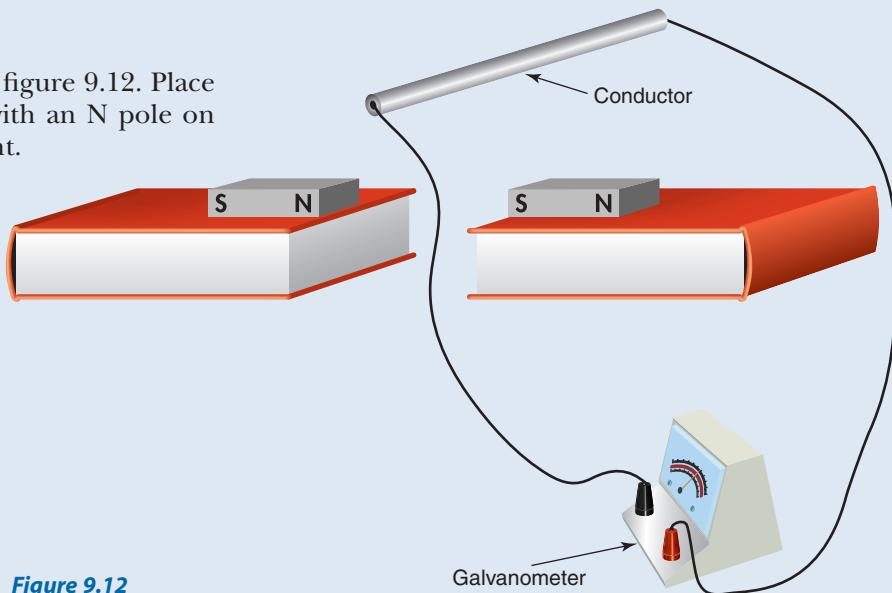
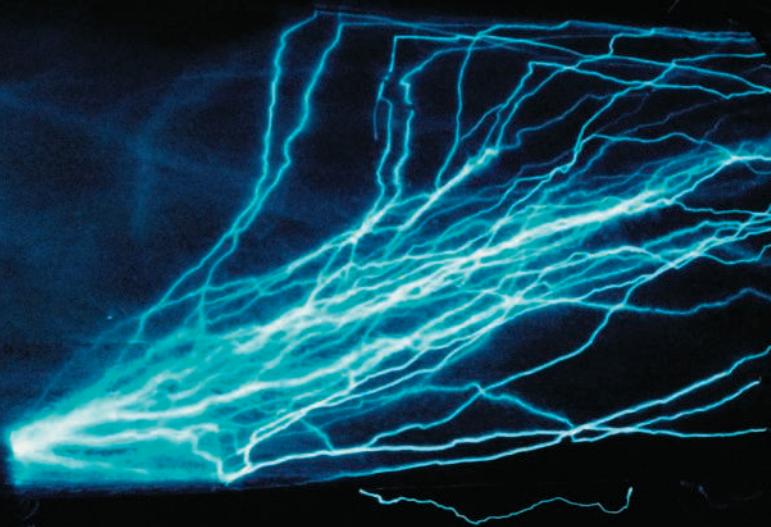


Figure 9.12



Chapter 10

Cathode rays and the development of television

Chapter 11

The photoelectric effect and black body radiation

Chapter 12

The development and application of transistors

Chapter 13

Superconductivity

FROM IDEAS TO IMPLEMENTATION

CHAPTER 10

CATHODE RAYS AND THE DEVELOPMENT OF TELEVISION

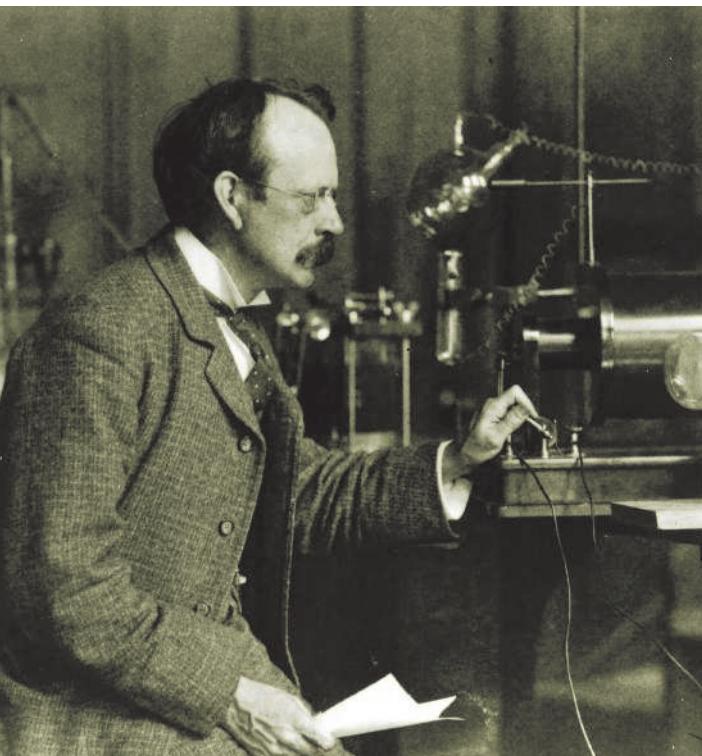


Figure 10.1 The famous physicist J. J. Thomson (1856–1940) was celebrated for his experiments with the electron.

Remember

Before beginning this chapter, you should be able to:

- describe the trajectory of a particle in a uniform gravitational field
- describe the properties of electrostatic charges and the fields associated with them
- describe electric potential difference, or the voltage between two points, as the work done in moving a unit charge between those points
- describe the effect that an electric field has on charged particles
- recall and perform calculations using $F = qE$, $P = VI$ and Energy = VI
- describe the fields produced by magnetic poles
- describe the magnetic fields produced around current-carrying conductors.

Key content

At the end of this chapter you should be able to:

- explain how cathode ray tubes allowed a stream of charged particles to be manipulated
- explain why the apparently inconsistent behaviour of cathode rays caused debate as to whether they were charged particles or electromagnetic waves
- list the properties of cathode rays and describe the key experimental observations from which these properties were deduced
- identify that moving charged particles in a magnetic field experience a force
- describe the effect that a magnetic field has on charged particles
- describe, qualitatively, the electric field between parallel charged plates
- outline Thomson's experiment in which he measured the charge-to-mass ratio of the electron
- sketch a cathode ray tube; label the electrodes, electron gun, the deflection plates or coils and the fluorescent screen, and describe their role in television displays and oscilloscopes
- perform calculations using $E = \frac{V}{d}$, $F = qE$ and $F = qvB \sin \theta$
- using the example of cathode rays, discuss the application of the scientific method to develop an understanding of this phenomenon

10.1 THE DISCOVERY OF CATHODE RAYS

In the early part of the nineteenth century, the discovery of electricity had a profound effect on the study of science. By the 1850s, much was known about which solids and liquids were electric conductors or insulators, and it was thought that gases were electric insulators.

The development of a vacuum, using pumps to remove the air from glass tubes, was also being actively researched at this time. As improved vacuum pumps were developed, scientists were able to experiment with gases at very low pressures. In 1855, a German physicist, Heinrich Geissler (1814–1879), refined a vacuum pump so that it could be made to evacuate a glass tube to within 0.01 per cent of normal air pressure. Geissler's friend, Julius Plucker (1801–1868), took these tubes and sealed a metal plate, called an electrode, to each end of the tube. The electrodes made electrical connections through the glass and were sealed to maintain the partial vacuum in the tube. These were then connected to a high-voltage source, as illustrated in figure 10.2. To their surprise, the evacuated tube actually conducted an electric current. What puzzled them more was the fact that the glass at the positive end, or **anode**, of the vacuum tube glowed with a pale green light. What type of invisible 'ray' caused this glow or **fluorescence**?

Whatever it was must have originated at the negative electrode, or **cathode**, of the vacuum tube. Another physicist, Eugene Goldstein (1850–1930), who was studying these same effects, named the rays that caused the glow 'cathode rays', and the tubes became known as **cathode ray tubes** or **discharge tubes** (see figure 10.3). Early experimenters used these tubes to investigate all of the properties of cathode rays and X-rays. Some modified the cathode ray tube to include a rectangular metal plate covered in zinc sulphide inside the tube. This plate had a horizontal slit cut into the end nearest the cathode and the plate was slightly bent so that the cathode rays formed a horizontal beam. When the cathode rays struck this material it appeared fluorescent and showed the path of the rays through the tube.

Cathode ray tubes have been refined and developed and are now used in television sets, computers and many other applications.

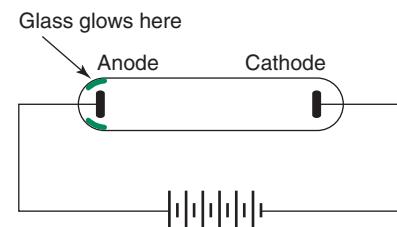


Figure 10.2 Production of cathode rays in a discharge tube, as used by Plucker

Partial vacuums are often described as 'rarefied air'.

Fluorescence is the emission of light from a material when it is exposed to streams of particles or external radiation.

Cathode rays are now known to be streams of electrons emitted within an evacuated tube from a **cathode** (negative electrode) to an **anode** (positive electrode). They were first observed in discharge tubes.

A **cathode ray tube** or a **discharge tube** is a sealed glass tube from which most of the air is removed by vacuum pump. A beam of electrons travels from the cathode to the anode and can be deflected by electrical and/or magnetic fields.

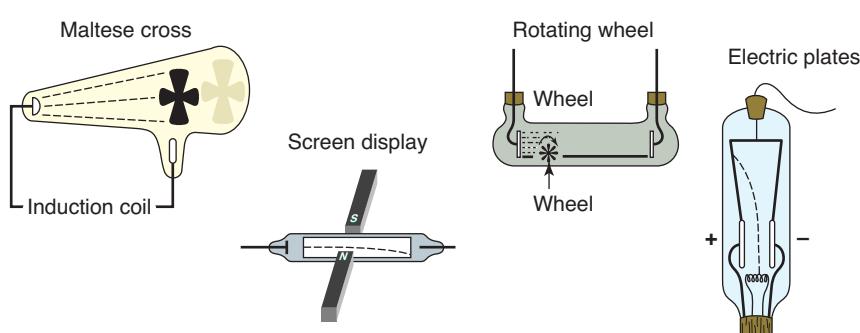


Figure 10.3 A variety of early discharge tubes used in experiments



10.1

Discharge tubes

Discharge tubes

Discharge tubes evacuated to different air pressures were found to produce different effects.

For example, in practical activity 10.1 (page 191), an induction coil acts as a step-up transformer, delivering a high voltage across the set of discharge tubes. At low pressures, electrons can accelerate to faster speeds before colliding with gas particles. Initially, a current will flow even though nothing can be seen. The first effect that can be observed is a steady luminous discharge known as a 'glow discharge'. As the pressure is lowered further, a number of colourful effects can be seen.

At first, most of the tube is occupied by a bright luminous region called a 'positive column' which appears to start from the anode and is broken up into a series of bands or striations (see figure 10.4). Near the anode, a weaker glow can be seen. The striations are separated by 'dark spaces'. These discharges and spaces are named after some of the scientists who examined them, for example, 'Ashton's dark space', 'Crookes' dark space' and 'Faraday's dark space'. The colours of the discharge depend on the gas used. In low pressure air, the positive column is a brilliant pink and the negative glow is deep blue.

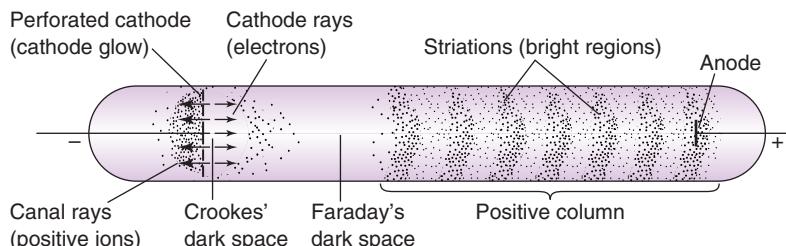


Figure 10.4 Some of the effects observed in discharge tubes

PHYSICS IN FOCUS

Everyday uses of discharge tubes

Neon signs colour the night in every city street. They are long tubes with most of the air removed. A small amount of gas is introduced which, when excited by a high potential, glows with a characteristic colour. For example, when the added gas is neon, the kinetic energy of the electrons is sufficient to ionise the gas around the cathode causing the emission of a reddish light.

Fluorescent tubes in the home contain mercury vapour at low pressure. The light produced is in the ultraviolet region of the electromagnetic spectrum. To produce visible light, a thin coating of a powder is spread on the inside surface of the tube. The ultraviolet radiation causes this coating to fluoresce with the familiar bright white light.



Figure 10.5 Discharge tubes are used in the neon lights that are often a feature of city skylines at night.

10.2 EFFECT OF ELECTRIC FIELDS ON CATHODE RAYS

Review your study of electric fields from the Preliminary Course topic, ‘Electrical energy in the home’.

You are familiar with three types of fields: gravitational, electric and magnetic. An electric field exists in any region where an electric charge experiences a force. There are two types of charge — positive and negative. We define the direction of the electric field as the direction in which a positive charge will experience a force when placed in an electric field.

This definition of an electric field allows us to describe the fields around a charge (see figure 10.6). Using Faraday’s ‘lines of force’, we see that these lines radiate from a point at the centre of the charge. For a positive charge, lines of force leave the centre of the charge and radiate in all directions from it. For a negative charge, the lines are directed radially into the centre of the charge.

If a positive charge is placed near another positive charge, it will experience a force of repulsion; that is, a force which acts in the direction of the arrow.

A number of rules apply to the interpretation of these lines of force diagrams (see figure 10.6).

- Field lines begin on positive charges and end on negative charges.
- Field lines never cross.
- Field lines that are close together represent strong fields.
- Field lines that are well separated represent weak fields.
- A positive charge placed in the field will experience a force in the direction of the arrow.
- A negative charge placed in the field will experience a force in the direction opposite to the arrow.

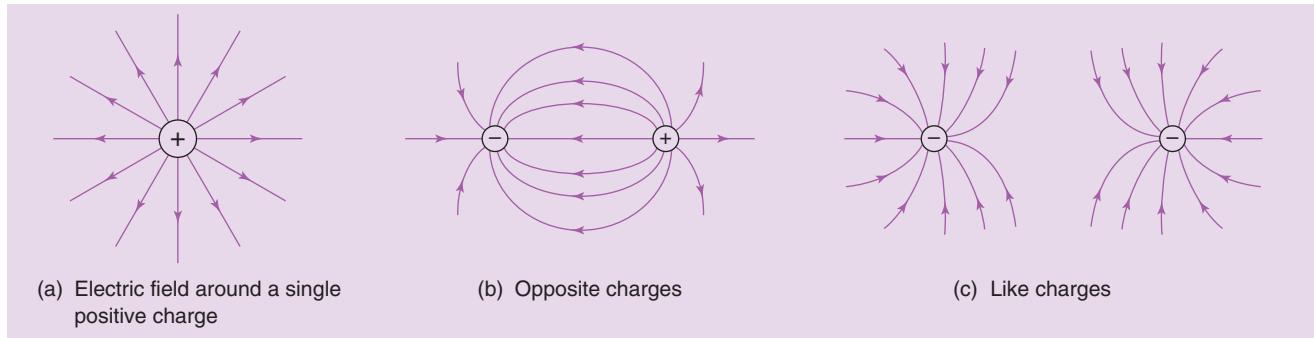


Figure 10.6 Electric fields around charges

Uniform electric fields

A uniform electric field can be made by placing charges on two parallel plates which are separated by a small distance compared with their length. These electric fields are very useful in physics and were used by prominent scientists such as Robert Millikan and J.J. Thomson (see page 180 onwards) when investigating the properties of small charged particles.

Consider the electric field between two plates that are separated by d metres, as shown in figure 10.7(a).

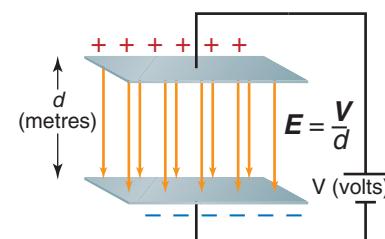


Figure 10.7(a) Electric field (E) between two parallel plates

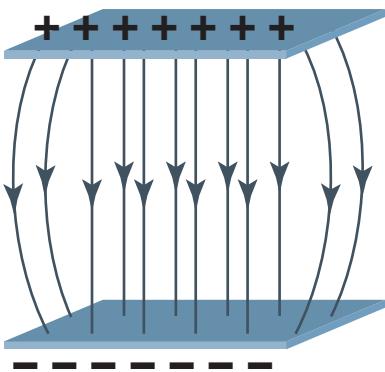


Figure 10.7(b) The electric field is uniform except at the edges of the plates where it bulges slightly.

The magnitude, or intensity, of an electric field is determined by finding the force acting on a unit charge placed at that point. The symbol for electric field is E .

$$E = \frac{F}{q}$$

where

E = electric field intensity (in newtons per coulomb)

F = electric force (in newtons, N)

q = electric charge (in coulombs, C).

When the potential difference, or voltage, is applied to the plates, a uniform electric field is produced. The strength of this field is the same at all points between the plates, except near the edges where it 'bulges' slightly (see figure 10.7(b)).

The magnitude of the electric field, E , is given by:

$$E = \frac{V}{d}$$

where

V = potential difference, in volts.

This can be derived by recalling that potential difference is the change in potential energy per unit charge moving from one point to the other. The amount of energy or work is given by:

$$W = qV$$

Also, the work done by a force is the product of the force and the distance moved, d . In this case, $F = qE$. Hence, the amount of work is given by:

$$W = Fd = qEd$$

It follows that: $V = Ed$ or $E = \frac{V}{d}$.

Remember that *work done is equal to the gain in energy*.

A small positive charge released next to the positive plate will experience a force that will accelerate the charge. The charge will increase its kinetic energy.

$$W = qV = \frac{1}{2}mv^2$$

This shows that the amount of work done depends only on the potential difference and the charge, and is the same for both uniform and non-uniform electric fields.

Electric field strength

What is the electric field strength between two parallel plates separated by 5.0 mm, if a potential difference of 48 volts is applied across them?

SOLUTION

$$V = 48 \text{ volts}$$

$$\begin{aligned} d &= 5.0 \text{ mm} \\ &= 5.0 \times 10^{-3} \text{ m} \end{aligned}$$

$$\begin{aligned} E &= \frac{V}{d} \\ &= \frac{48}{5.0 \times 10^{-3}} \\ &= 9600 \text{ V m}^{-1} \text{ or } \text{N C}^{-1} \end{aligned}$$

SAMPLE PROBLEM

10.1

SOLUTION

Moving charge through a potential difference

How much work is done moving a charge of $3.6 \mu\text{C}$ through a potential difference of 15 volts?

SOLUTION

$$q = 3.6 \times 10^{-6} \text{ C}$$

$$V = 15 \text{ volts}$$

$$\begin{aligned} W &= qV \\ &= 3.6 \times 10^{-6} \times 15 \\ &= 5.4 \times 10^{-5} \text{ J} \end{aligned}$$

SAMPLE PROBLEM

10.2

Velocity of a charge between plates

Two parallel plates are separated by a distance of 5.0 mm. A potential difference of 200 volts is connected across them. A small object with a mass of 1.8×10^{-12} kg is given a positive charge of $12 \mu\text{C}$. It is released from rest near the positive plate. Calculate the velocity gained as it moves from the positive plate to the negative plate.

SOLUTION

$$d = 5.0 \times 10^{-3} \text{ m}$$

$$V = 200 \text{ V}$$

$$q = 1.2 \times 10^{-5} \text{ C}$$

$$m = 1.8 \times 10^{-12} \text{ kg}$$

$$qV = \frac{1}{2} mv^2$$

$$v^2 = \frac{2(1.2 \times 10^{-5} \times 200 \times 10^3)}{1.8 \times 10^{-12}}$$

$$v = 5.2 \times 10^4 \text{ m s}^{-1}$$

PHYSICS IN FOCUS**Protection against lightning: pointed conductors**

Lightning strikes are an example of a massive electrical discharge over a short interval of time. Large cumulonimbus clouds generate a distribution of charge between the top of the cloud and the bottom. As the cloud moves over the ground the negative charge in the cloud repels electrons in the ground, producing a potential difference between the cloud and the ground. When this potential difference is large enough to overcome the resistance of the air, there is a discharge that we see as lightning. Uncontrolled discharge can be very destructive.

Benjamin Franklin, who experimented with electricity in the 1750s, came up with the first 'lightning rod' to act as a conductor and protect buildings from damage. The device is based on the fact that the electric field around a conducting object depends both on its charge and its shape. The field is strongest near sharp points on the object. The field can become sufficiently strong so that the air molecules lose electrons, becoming ions. Eventually, sufficient air molecules are ionised and the air surrounding the charged body becomes a conductor. The charge can then leak away into the air (see figure 10.8).

A Van der Graaf generator is round to allow a large electric charge to build up. When the charge is great enough, an electric spark can jump across a small gap. If we attach a fine wire with a sharp

point to the dome of the generator, no spark can be obtained. This is because the charge leaks away into the surrounding air before it builds to a high enough level for a spark to form.

For buildings, a lightning protection system involves attaching a pointed metal object at the highest part of the roof and running a system of metal straps from it to carry the charge safely to ground, where the strap is buried a metre into the earth. The charge from a lightning strike can drain quickly through the conductor and prevent a fire.

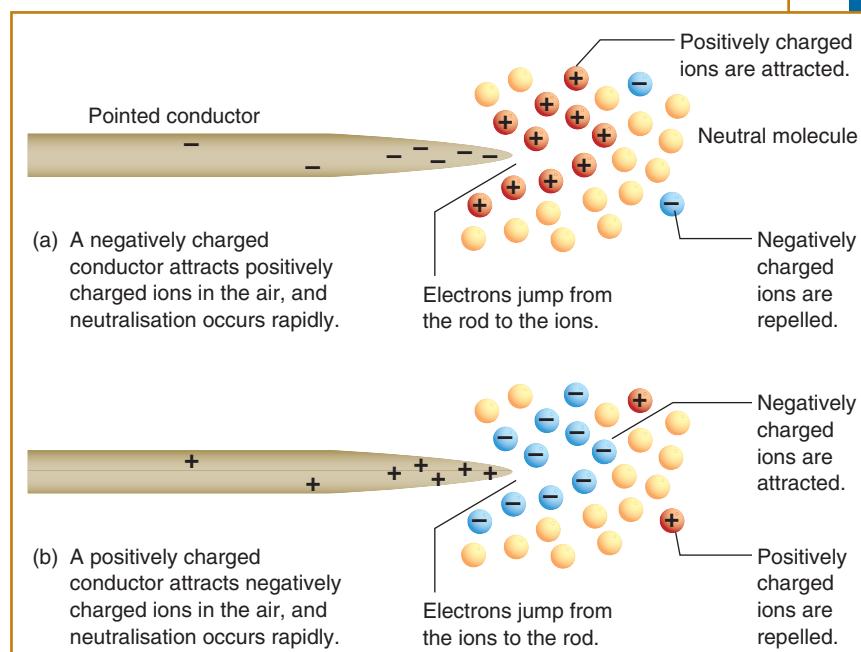


Figure 10.8 Pointed conductors discharge rapidly in air.

PHYSICS IN FOCUS

The photocopier machine: charged plates

Photocopiers which scan an image and produce a dry image (that is, one that does not use photographic solutions and special photosensitive paper) were first developed by Chester Carlson in a process called xerography, on 22 October 1938. The first commercial unit utilising a completely dry process was the Xerox-914 released in 1960.

Some semiconductors — selenium, arsenic and tellurium — act as ‘photoconductors’. That is, they act as insulators in the dark and electrical conductors in the light. A thin layer of semiconductor material is deposited onto the surface of a drum. At the start of a copy cycle, this drum is given a uniform electrostatic charge. The page to be copied is illuminated with a strong light and an image of the page is formed by a lens on the charged photoconductor. White areas light up the photoconductor and it becomes a conductor so that its charge leaks away to the metal backing. The black areas remain. This latent image is then developed.

A fine powder of small glass beads, covered with a black toner, is gently brushed over the drum. (The toner is actually a black coloured thermoplastic substance which melts and impregnates the paper.) The toner sticks only where there is remaining charge, that is, the black areas of the original image.

Next, a blank sheet of paper is rolled over the drum and the toner is transferred to the paper giving an exact image of the original. The image is then ‘fixed’ by heating this page. Finally, the drum is cleared and made ready for the next copy cycle.

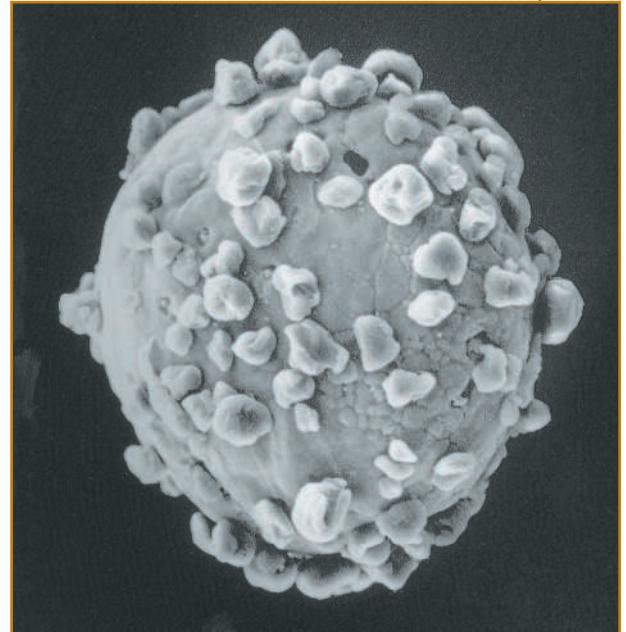
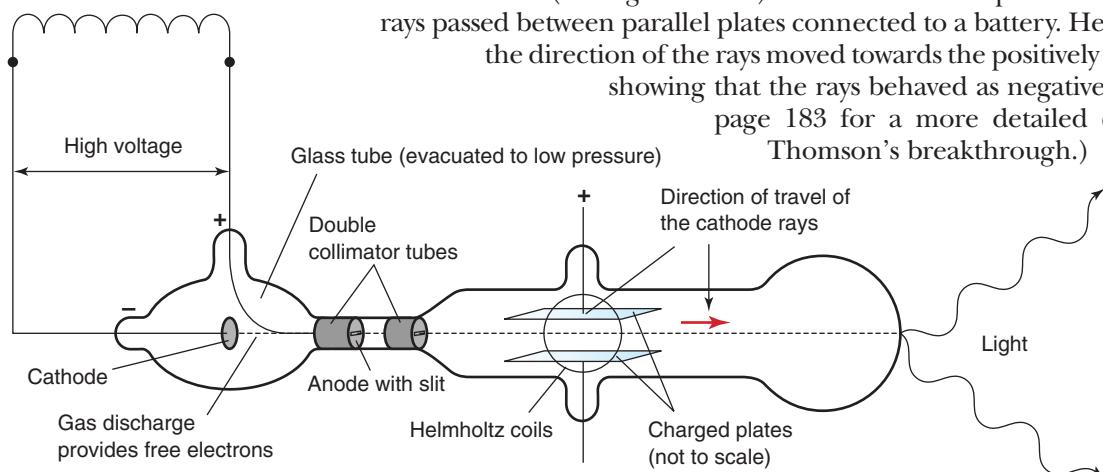


Figure 10.9 A magnified image from Xerox showing toner particles clinging to a tiny carrier bead by means of electrostatic forces. From the bead, the negatively charged toner particles are attracted to the charged drum and then to the paper.

J.J. Thomson

Figure 10.10 The apparatus for J. J. Thomson's experiments with cathode rays

Secondary coil of an induction coil



The work of English physicist Joseph John Thomson (1856–1940) centred around cathode rays (see figure 10.1, page 174). By incorporating charged plates inside the cathode ray tube, Thomson was able to verify an earlier hypothesis by Crookes that cathode rays would be deflected by electric fields (see figure 10.10). In Thomson's experiment, the cathode rays passed between parallel plates connected to a battery. He observed that the direction of the rays moved towards the positively charged plate, showing that the rays behaved as negative charges. (See page 183 for a more detailed description of Thomson's breakthrough.)

PHYSICS FACT

Millikan's oil drop experiment

In 1909, American physicist, Robert A. Millikan, was able to use the uniform electric field created between two parallel plates to investigate the properties of a charge. His set-up (see figure 10.11) involved an atomiser which sprayed a fine mist of oil drops into his apparatus (region A).

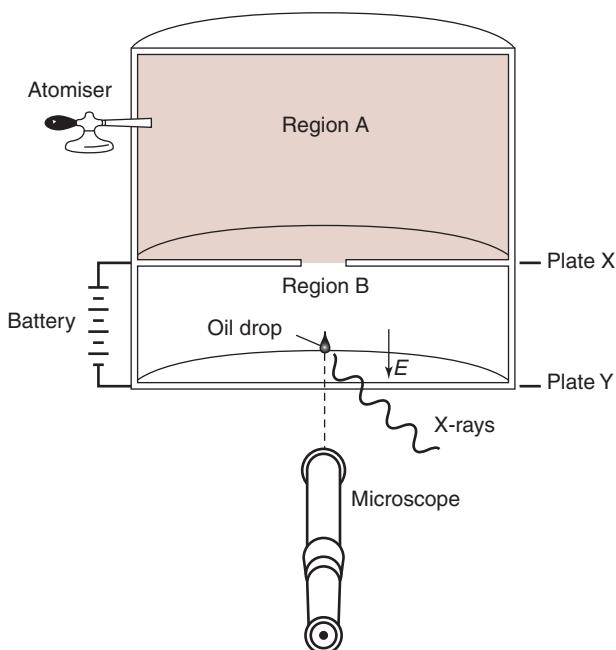


Figure 10.11 Apparatus for Millikan's oil drop experiment

Some drops drifted into region B and came under the influence of the electric field, E . As the oil drops entered this region they were momentarily exposed to a beam of X-rays, resulting in some of the oil drops becoming charged. The gravitational field of the Earth exerts a force

directed vertically down (weight) which can be counteracted by an electric field produced between parallel plates by a source of variable voltage. The spaces between the plates could be viewed through a microscope. By careful adjustments of the voltage it was possible for one drop to be held stationary, or made to travel with uniform velocity. That is, the forces acting on the drop were balanced.

$$\text{Weight force (down)} = \mathbf{F}_g = mg$$

$$\text{Electric force (up)} = \mathbf{F}_E = Eq$$

For this drop to be suspended between plates, $F_g = F_E$.

Having suspended an oil drop, Millikan could then determine the charge on that particular oil drop by solving for q ; that is, $q = \frac{mg}{E}$.

Millikan needed to determine the mass of the oil drop. His approach was to measure the terminal velocity of the oil drop when the electric field was turned off and it fell under the force of gravity alone. By using equations from fluid mechanics, he could calculate the radius of the oil drop. By using an oil with a known density, he was able to determine the mass of the oil drop.

Millikan's remarkable findings, for which he won a Nobel Prize in 1923, showed that the charge on an oil drop was not of just any arbitrary value. Instead, the charge always occurred in 'packets' or multiples of some smallest value. This value was calculated to be $1.6 \times 10^{-19} \text{ C}$ and was called the 'elementary charge', the charge found on an electron.

SAMPLE PROBLEM

10.4

Field strength and the charge on an oil drop

An oil drop of mass $6.8 \times 10^{-6} \text{ g}$ is suspended between two parallel plates which are separated by a distance of 3.5 mm, as shown in figure 10.12.

- What is the electric field strength between the plates?
- What is the charge that must exist on the oil drop?
- How many excess electrons must be present on the oil drop?

SOLUTION

- (a) Using the equation $E = \frac{V}{d}$:

$$E = \frac{110}{3.5 \times 10^{-3}}$$

$$= 3.1 \times 10^4 \text{ V m}^{-1} \text{ down.}$$

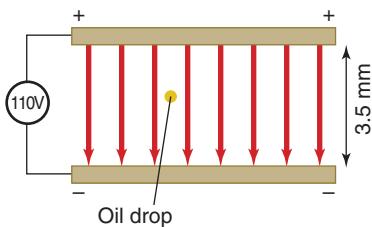


Figure 10.12

(b) Since the oil drop is suspended, $qE = mg$, $\therefore q = \frac{mg}{E}$.

$$\text{Now } 6.8 \times 10^{-6} \text{ g} = 6.8 \times 10^{-9} \text{ kg}$$

$$\therefore q = \frac{6.8 \times 10^{-9} \times 9.8}{3.1 \times 10^4} \\ = 2.1 \times 10^{-12} \text{ C.}$$

$$(c) \text{ Number} = \frac{\text{charge on drop}}{\text{charge on electron}} \frac{2.1 \times 10^{-12}}{1.6 \times 10^{-19}} = 1.3 \times 10^7 \text{ electrons.}$$

10.3 EFFECT OF MAGNETIC FIELDS ON CATHODE RAYS

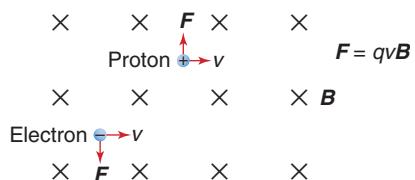


Figure 10.13 Forces on an electron and a proton moving perpendicularly to a magnetic field

Magnetic fields exert forces on electric currents; that is, on moving charged particles. If a particle with charge q is moving with velocity v , perpendicularly to a magnetic field of strength B , the particle will experience a magnetic force F , given by $F = qvB$.

The direction of the force is given by the right-hand rule. (If the particle has a positive charge, the direction of the conventional current is that of the velocity; if the particle has a negative charge, the direction of the conventional current is opposite to that of the velocity.) This is illustrated in figure 10.13 for an electron (negative charge) and a proton (positive charge).

If the velocity is at an angle θ to the magnetic field, the force is given by

$$F = qvB \sin \theta.$$

To find the direction of the force, use the component of the velocity perpendicular to the magnetic field and the right-hand rule. This is illustrated in figure 10.14 where the direction of the force is into the page.

If the charged particle is moving parallel to the magnetic field, $\theta = 0$, and therefore $F = 0$.

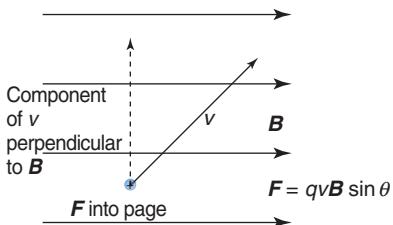


Figure 10.14 Force on a charged particle moving at an angle, θ , to a magnetic field

Review your study of magnetic fields from the Preliminary Course topic 'Electrical energy in the home' and the HSC Course topic 'Motors and generators'.

SAMPLE PROBLEM

10.5

Effects of magnetic fields on electrons

An electron of charge $-1.6 \times 10^{-19} \text{ C}$ is projected into a region where a magnetic field exists, as shown in the diagram. If the velocity of the electron is $2.5 \times 10^4 \text{ m s}^{-1}$, determine:

- the force on the electron at the instant it enters the magnetic field
- the shape of the path which the electron follows.

SOLUTION

$$(a) F = qvB \sin \theta$$

$$= 2.0 \times 10^{-2} \times 1.6 \times 10^{-19} \times 2.5 \times 10^4 \\ = 8.0 \times 10^{-17} \text{ N downwards.}$$

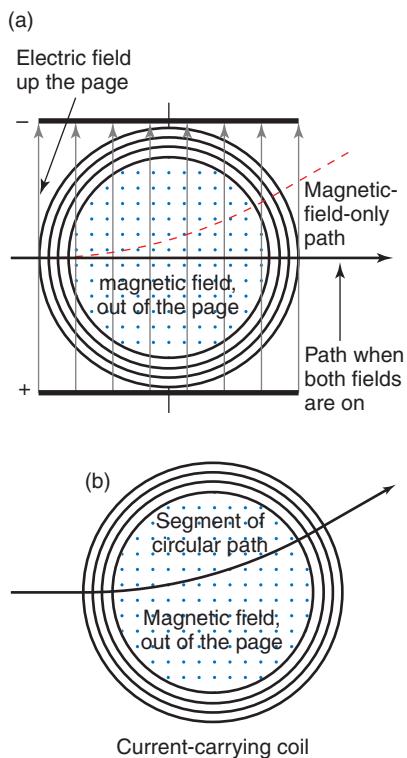
$$B = 2.0 \times 10^{-2} \text{ T}$$



Figure 10.15

10.4 DETERMINING THE CHARGE-TO-MASS RATIO OF CATHODE RAYS

The name ‘electrons’ was first suggested for the natural unit of electricity by G. J. Stoney in 1891.



J. J. Thomson was an intuitive and brilliant experimentalist. Following on from his experiment showing that cathode rays were deflected by electric fields, he succeeded in measuring the charge-to-mass ratio of the cathode ray particles, called electrons. Thomson built a cathode ray tube with charged parallel plates (called capacitor plates) to provide a uniform electric field and a source of uniform magnetic field. Using this apparatus, he investigated the effect of cathode rays passing through both fields (see figure 10.16). The fields were oriented at right angles to each other and this had the effect of producing forces on the cathode rays that directly opposed each other (see the ‘Physics fact’ below).

Thomson’s experiment involved two stages:

1. varying the magnetic field and electric fields until their opposing forces cancelled, leaving the cathode rays undeflected. This effect is shown in figure 10.16(a). By equating the magnetic and electric force equations, Thomson was able to determine the velocity of the cathode-ray particles.
2. applying the same strength magnetic field (alone) and determining the radius of the circle path travelled by the charged particles in the magnetic field (see figure 10.16(b)).

Thomson combined the results and obtained the magnitude of the charge-to-mass ratio for the charged particles that constituted cathode rays.

Figure 10.16 (a) A beam of negatively charged particles left undeflected by the combination of a magnetic field out of the page, and an electric field up the page (b) A negatively charged particle deflected by a magnetic field out of the page. The mechanics of circular motion describes the path, with the centripetal force provided by the magnetic force acting on the particle.

PHYSICS FACT

When charged particles enter an electric field they follow a trajectory under the influence of an electric force.

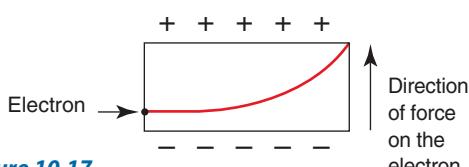


Figure 10.17

Similarly, when a charged particle enters a magnetic field, it experiences a magnetic force. The direction of this force is given by the right-hand palm rule.

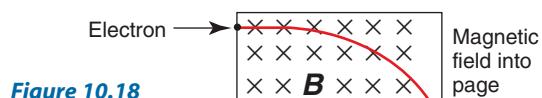


Figure 10.18

We can combine these two effects by arranging the electric field, magnetic field and the velocity of the particle at right angles to each other.

For example, by adjusting the strengths of the electric and magnetic fields, their effects on the motion

of a charged particle can cancel each other out. The particles can then travel along a straight path.

In figure 10.19 (see page 180), there are two sets of electric fields. The first accelerates the electrons through a set of collimators to produce a narrow beam. This beam then passes through a combination of electric and magnetic fields that can be adjusted.

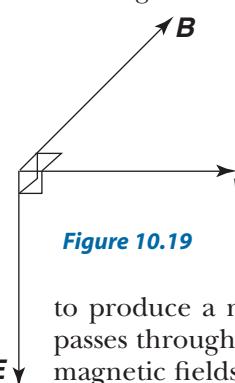


Figure 10.19

10.5 CATHODE RAYS — WAVES OR PARTICLES?

In 1875, twenty years after their discovery, William Crookes (1832–1919) designed a number of tubes to study cathode rays (some of these are shown in figure 10.3, page 175). He found that the cathode rays did not penetrate metals and travelled in straight lines. It was initially thought that the rays may be an electromagnetic wave because of the similarity in their behaviour to light. This was discounted when Crookes discovered that the cathode rays were deflected by magnetic fields, an effect which did not occur with light.

In a paper read to the Paris Academy of Science in 1885, Jean Perrin (1870–1942) described the two main hypotheses concerning the nature of cathode rays:

Some physicists think, with Goldstein, Hertz and Lenard, that this phenomenon is like light of very short wavelength. Others think, with Crookes and J. J. Thomson, that these rays are formed by matter which is negatively charged and moving with great velocity, and on this hypothesis their mechanical properties, as well as the manner in which they curve in a magnetic field, are readily explicable.

The way that physicists set out to understand the nature of cathode rays shows how the scientific method is used to solve problems. That is, observations from experiments are interpreted and a hypothesis developed to explain what is thought to be happening. Opposing models may arise, with supporters of each side arguing strongly for their belief. The argument may eventually be resolved either by improved experiments or with greater understanding of the phenomenon.

In this case, the debate about whether cathode rays were electromagnetic waves or streams of charged particles remained unsolved until 1897, when J. J. Thomson showed beyond doubt that the rays were streams of negatively charged particles, which we now call electrons. Why was the debate so prolonged? The problem was the apparently inconsistent behaviour of rays. For example, the following observations from cathode ray experiments fitted the wave model:

- they travelled in straight lines
 - if an opaque object was placed in their path, a shadow of that object appeared
 - they could pass through thin metal foils without damaging them.
- The following observations fitted the particle model:
- the rays left the cathode at right angles to the surface
 - they were obviously deflected by magnetic fields
 - they did not appear to be deflected by electric fields
 - small paddlewheels turned when placed in the path of the rays
 - they travelled considerably more slowly than light.

The main restriction for the charged particle theory was the absence of deflection in electric fields. However, Thomson showed that this was due to the rays themselves. He stated:

‘...on repeating the experiment I first got the same result, but subsequent experiments showed that the absence of deflection is due to the conductivity conferred on the rarefied gas by the cathode rays. On measuring this conductivity... it was found to



10.2

Properties of cathode rays

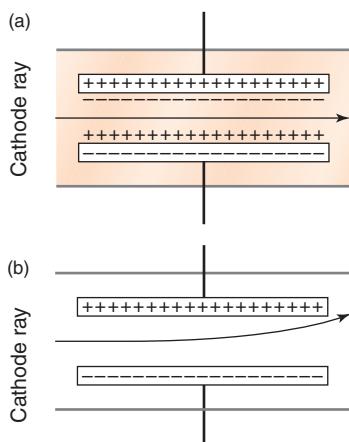


Figure 10.20 The path of cathode rays (a) at high gas pressure and (b) at low gas pressure

decrease very rapidly with the exhaustion of the gas...at very high exhaustions there might be a chance of detecting the deflection of cathode rays by an electrostatic force.'

Within the tube, the cathode rays ionised the gas. The ions were attracted to the plate with the opposite charge and the line-up of ions effectively neutralised the charge on the plate, allowing the cathode rays to pass by unaffected.

After evacuating the chamber, Thomson observed deflection and that the particles were always deflected towards the positive plate, which confirmed that they were negatively charged particles. The deflection of cathode rays in tubes of different gas pressure is shown in figure 10.20.

The ability of cathode rays to penetrate thin metal foils was still unexplained. The answer lay, not simply with the properties of cathode rays, but with the model of the atom. If the atom was not a solid object, but much more open, it might be possible for very small particles to pass through thin foil. Although not considered at this time, Ernest Rutherford (1871–1937) would use a similar approach to change the model of the atom (discussed in chapter 22, pages 419–423).

PHYSICS FACT

X-rays: discovery and application

In 1895, a type of radiation was discovered by Wilhelm Röntgen (1845–1923) while he was experimenting with cathode rays. He found that, in a dark room, a screen covered with a sensitive fluorescent material (barium platino-cyanide) glowed when it was placed near the end of a cathode ray tube (see figure 10.21).

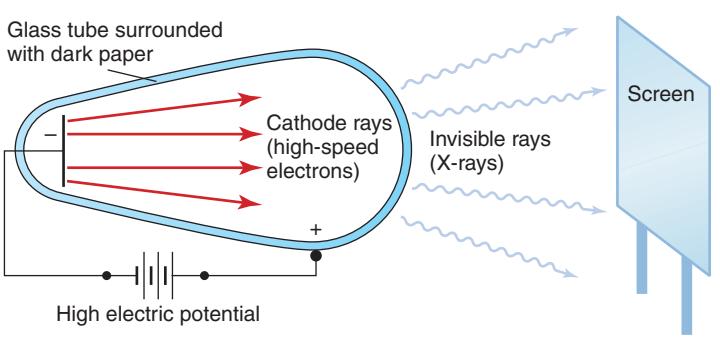


Figure 10.21 Röntgen's apparatus

Since cathode rays could not pass through the glass at the end of the tube, he deduced that this fluorescence must be due to a new form of radiation. He called this radiation 'X-rays' as their properties were not known. Later research showed that X-rays were produced when high-speed electrons interacted with matter, such as the glass in the cathode ray tube.

X-rays were later found to be electromagnetic waves, similar to light but with a much smaller

wavelength (see chapter 13, pages 235–239, for further discussion of X-rays).

Among the many characteristics that make X-rays so useful are the fact that they can:

- penetrate many substances
- expose photographic film
- cause certain substances to fluoresce
- be reflected and refracted.

The most common use of X-rays is in the field of medicine, for diagnosing illness or injury as well as treating illnesses such as cancer. X-ray machines are used widely — to check luggage at airports, analyse the welding of metal parts in an aircraft wing, and look at things that we otherwise could not see.

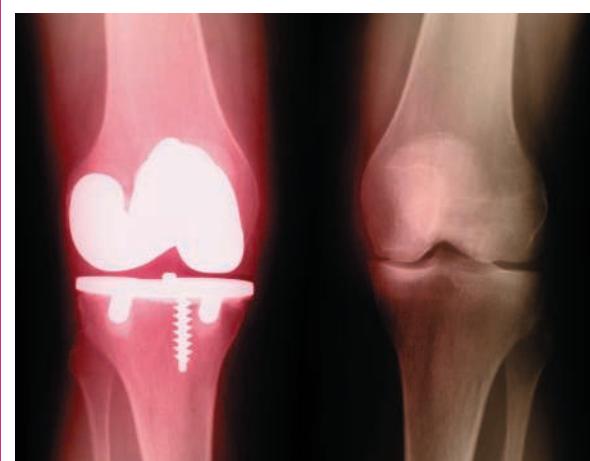


Figure 10.22 X-ray of a knee replacement

10.6 APPLICATIONS OF CATHODE RAYS

Cathode ray tubes have a multitude of applications, ranging from small screens at an automatic teller machine through to radar screens, television or computer monitors, and cathode ray oscilloscopes (CRO). They are also found in many different medical instruments. They are all based on the simple cathode ray tube.

Parts of a cathode ray tube

The main parts of a cathode ray tube are the electron gun, the deflecting plates and the fluorescent screen (see figure 10.23).

- In the *electron gun*, the heating filament heats the cathode, releasing electrons by thermionic emission. A number of electrodes are used to control the ‘brightness’ of the beam, to focus the beam and accelerate the electrons along the tube. Electrons are negatively charged particles and the positively charged anode develops a strong electric field that exerts a force on the electrons, accelerating them along the tube.
- Two sets of parallel *deflecting plates* are charged to produce an electric field that can deflect the beam of electrons separately, up or down and left or right. These fields are used to move the beam so that the electrons can be directed to all points on the fluorescent screen. Electric current passing through coils around the cathode ray tube produces magnetic fields that control the movement of the electron beam (as outlined on page 183). This ‘trace’ on the screen produces the visible output, such as the picture on a television.
- The glass *screen* is coated with layers of a fluorescent material. It emits light when high energy electrons strike it.

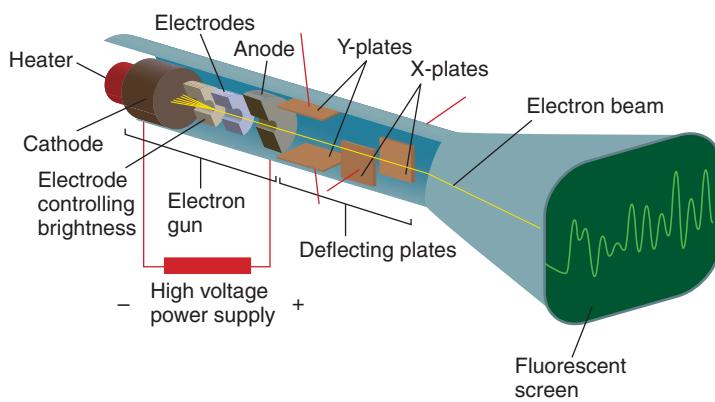


Figure 10.23 The parts of a simple cathode ray tube

Television

Television sets use a cathode ray tube as their output device. A colour television camera records images through three coloured filters — red, blue and green (see figure 10.24). The information is transmitted to the receiver which then directs the appropriate signal to one of three electron guns, each corresponding to one of the primary colours. The picture is then reconstituted on the screen by an additive process involving three coloured phosphors. Each electron gun stimulates its appropriate phosphor.

Each television image is made up of 625 horizontal lines of dots. The current in the coils energises the deflection coils and is varied to scan the screen twice for each image. The electrons sweep across the screen, building up the picture. Each picture is formed from two passes of the electron beam. The odd-numbered lines are drawn first, then the beam ‘flies’ back to the start and ‘draws in’ the even-numbered lines. Each scan takes one-fiftieth of a second. This is shorter than the time that the retina/brain system retains each image so that the screen does not seem to flicker.

Figure 10.24 (a) Simplified side view of a colour television (b) The main parts of a colour television

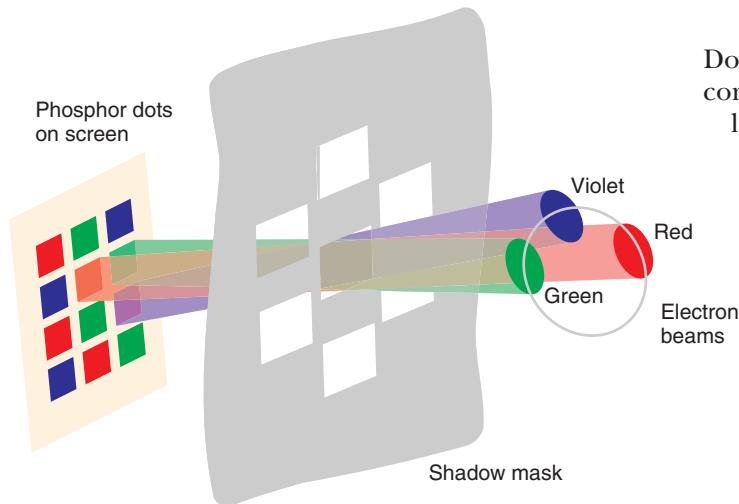
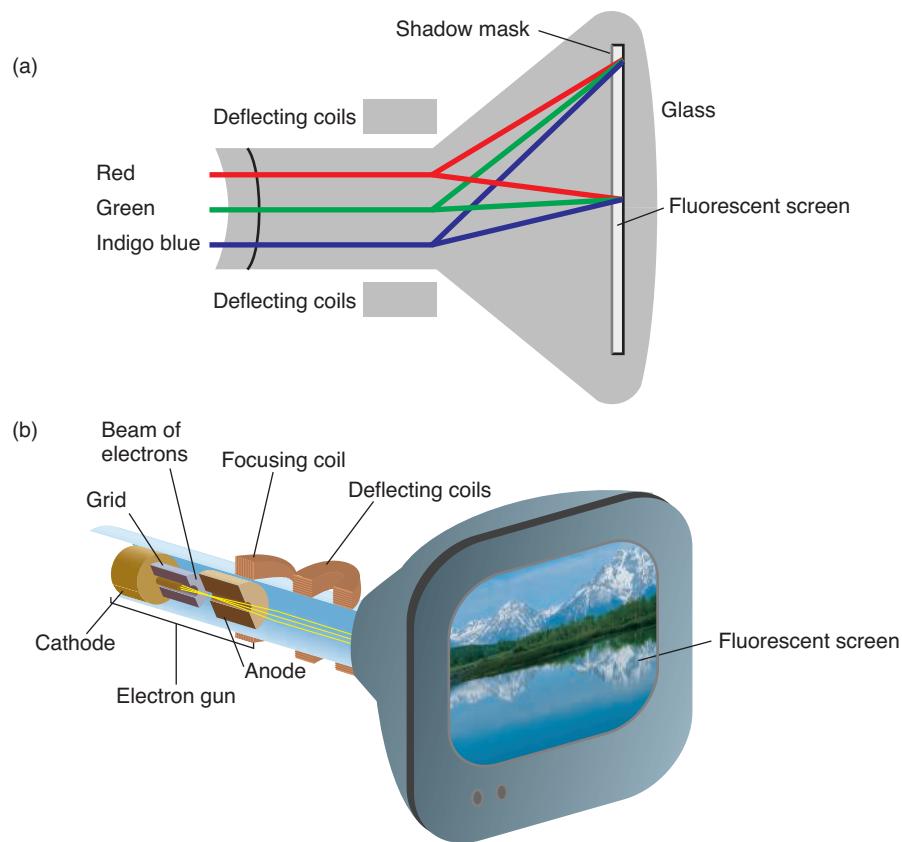


Figure 10.25 Enlarged section of the shadow mask in the cathode ray tube of a TV set. The beams of electrons are directed through the mask and hit a pattern of phosphor dots on the inside of the screen.

Dots of phosphorescent paint on the screen convert the energy of the electron beam into light. After being excited, the phosphorescence continues to emit light for a longer time, which helps to minimise screen flicker. In a colour television, the phosphorescent dots come in the three colours (red, green and blue) and the many colours that you see on the screen are formed from combinations of these three.

Three electron beams scan the screen. They come from slightly different directions through holes in a shadow mask to control the brightness of the three sets of phosphors (see figure 10.25).

Cathode ray oscilloscopes

The introduction of electronic control systems into all forms of science and industry has seen the cathode ray oscilloscope (CRO) become one of the most widely utilised test instruments. Because of its ability to make 'voltages' visible, the cathode ray oscilloscope is a powerful diagnostic and development tool.

A CRO uses a cathode ray tube to display a variety of electrical signals. The horizontal deflection is usually provided by a time base (or sweep generator), which allows the voltage (on the vertical axis) to be plotted as a function of time (on the horizontal axis). This enables complex waveforms or very short pulses to be displayed and measured. Figure 10.26 shows the basic controls and the front panel of a typical CRO.

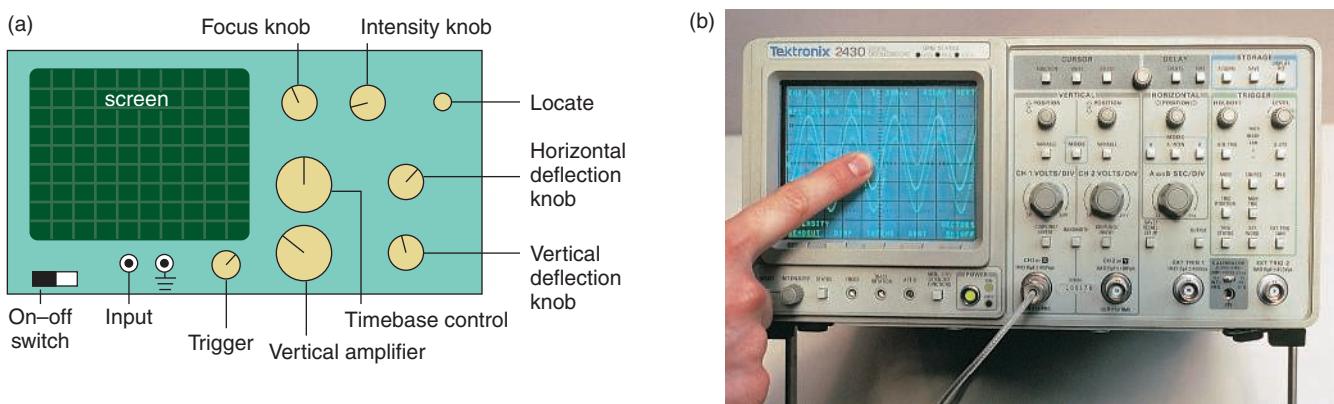


Figure 10.26 (a) The basic controls of a single trace CRO (b) The front panel of a typical CRO

The timebase control allows the technician to select a variety of sweep rates. This sets the ‘time per division’ for the figure drawn on the screen. The bottom rotary switches control the amplitude of the displayed waveform. Each centimetre of the grid can then be used to measure the voltage of the input waveform.

Figure 10.27 shows the major parts of the oscilloscope. The input waveform enters from the left on this diagram. As can be seen, one part of the signal is amplified and the voltage produced goes to the vertical deflecting plates. The other part passes to the trigger and time base and is then passed to the horizontal deflection plates.

The combination of both signals produce the waveform displayed on the front screen.

There are many types of cathode ray oscilloscope: single, double or multiple trace; analogue and digital; and storage CROs that allow technicians to record and store waveforms for comparison.

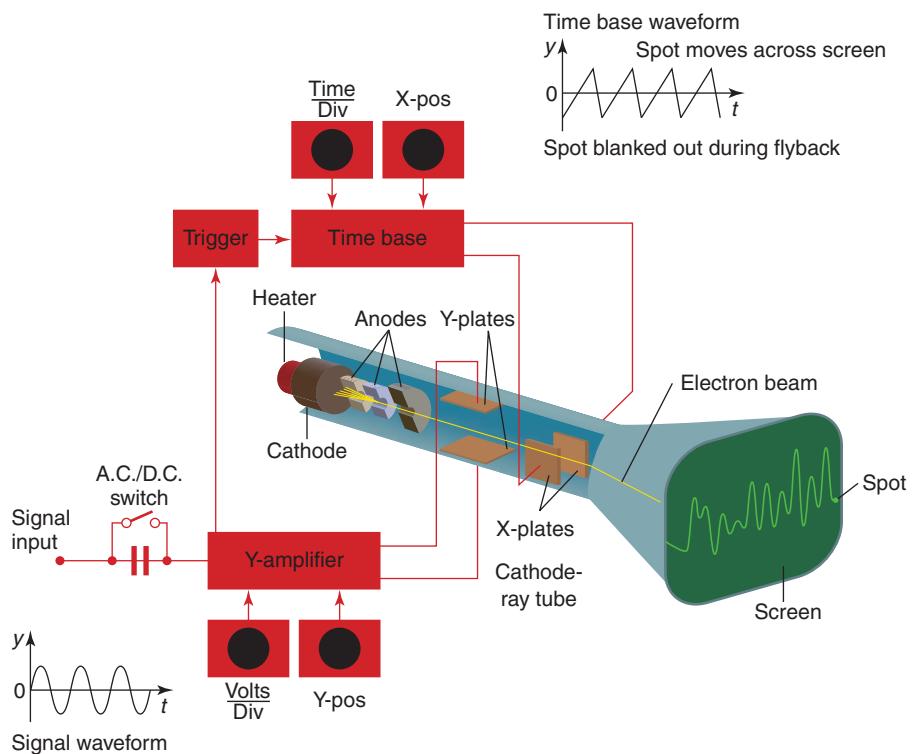


Figure 10.27 The main parts of a CRO

SUMMARY

- Cathode ray tubes were used to investigate the properties of cathode rays.
- Cathode rays were found to be negatively charged particles.
- Charged parallel plates produce a uniform electric field.
- A charged particle moving with a velocity, v , at an angle, θ , through a magnetic field of strength, B , experiences a force, F . The magnitude, in Newtons, is given by:

$$F = qvB\sin \theta.$$

F , v and B are all vector quantities, and each has a direction associated with it.

- The strength of a uniform electric field, E , in volts per metre, produced by parallel plates, separated by a distance, d , and charged by an applied voltage, V , is given by:

$$E = \frac{V}{d}.$$

- Thomson's experiment, using perpendicular electric and magnetic fields, allowed him to determine the charge-to-mass ratio of an electron. The value he determined was $1.759 \times 10^{11} \text{ C kg}^{-1}$.
- Cathode ray tubes are made up of a number of components including an electron gun, the electric field and a fluorescent screen.
- Cathode ray tubes have many applications including cathode ray oscilloscopes and televisions.
- Cathode ray oscilloscopes enable engineers to develop new electronic circuits by making the behaviour of pulses or waves visible.

QUESTIONS

Note: The charge on a single electron is taken as $-1.6 \times 10^{-19} \text{ C}$.

- A beam of electrons moves at right angles to a magnetic field of flux density $6.0 \times 10^{-2} \text{ T}$. The electrons have a velocity of $2.5 \times 10^7 \text{ m s}^{-1}$. What is the magnitude of the force acting on each electron?
- A stream of doubly ionised particles (missing two electrons and therefore carrying a positive charge of twice the electronic charge) move at a velocity of $3.0 \times 10^4 \text{ m s}^{-1}$ perpendicular to a magnetic field of $9.0 \times 10^{-2} \text{ T}$. What is the magnitude of the force acting on each ion?

- An electron is travelling at right angles to a magnetic field of flux density 0.60 T with a velocity of $1.8 \times 10^6 \text{ m s}^{-1}$. What is the force experienced by the particle?
- Given that the mass of the electron in question 3 is $9.1 \times 10^{-31} \text{ kg}$, what is the acceleration of the particle in the direction of the force acting on it?
- A pair of parallel plates is arranged as shown in figure 10.28. The plates are 5.0 cm apart and a potential difference of 200 V is applied across them.

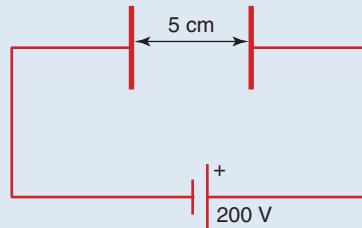


Figure 10.28

Data:

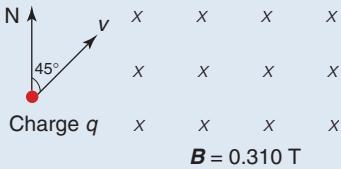
$$\text{Charge on electron} = -1.6 \times 10^{-19} \text{ C}$$

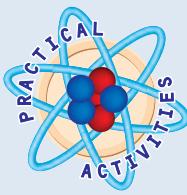
$$\text{Mass of electron} = 9.1 \times 10^{-31} \text{ kg}$$

$$\text{Mass of proton} = 1.67 \times 10^{-27} \text{ kg}$$

- Calculate the magnitude and direction of the electric field between the plates.
- Calculate the force acting on an electron placed between the plates.
- Calculate the force acting on a proton placed between the plates.
- Explain why these two forces are different.
- Calculate the work done in moving both the electron and the proton from one plate to the other.

- Look up the meaning of 'conservation of charge'. In a discharge tube, cathode rays were formed and moved from the cathode (the negative electrode). If these rays carried an electric charge, where was the corresponding amount of positive charge?
- Draw the electric field lines between a positive and negative charge of equal magnitude. In which area is the electric field strongest?
- Calculate the electric force on a charge of $1.0 \times 10^{-6} \text{ C}$ placed in a uniform electric field of 20 N C^{-1} .
- Draw the electric field lines between two parallel plates placed 5.0 cm apart.

10. The electric field between parallel plates can be considered ‘uniform’ only in the region between the plates that is well away from the edges of the plates. What is meant by this statement?
11. Negatively charged latex spheres are introduced between two charged plates and are held stationary by the electric field. Each sphere has a mass of 2.4×10^{-12} kg and the strength of the field required to counter their weight is 4.9×10^7 N C $^{-1}$. Sketch this arrangement, identifying the positive and the negative plate, and determine the charge on the spheres.
12. Two parallel plates are separated by a distance of 10.0 cm. The potential difference between the plates is 20.0 V.
- Calculate the electric field between the plates, assuming the field to be uniform.
 - A charge of $+2.0 \times 10^{-3}$ C is placed in the field. Calculate the force acting on this charge.
13. A charge of 5.25 mC, moving with a velocity of 300 m s^{-1} due north east, enters a uniform magnetic field of 0.310 T directed vertically downwards, into the page. Calculate the magnetic force on the charge.
- 
- Figure 10.29**
14. Why can cathode rays be observed and manipulated within a vacuum tube and not in air?
15. Describe the forces acting on charged particles entering a uniform magnetic field at right angles.
16. If charged particles enter a magnetic field at angles other than at right angles, describe their path.
17. List the properties of cathode rays which can be described:
- as wave motion
 - by a particle model.
18. Explain how the properties of cathode rays were demonstrated using the evacuated tubes in which a metal cross was mounted in the path of the rays, and in which a small paddle wheel was able to roll along glass rails.
19. Describe the path of an electron when it enters the region between parallel plates across which a potential difference of 1500 V is applied. Sketch the arrangement for an electron entering with a horizontal velocity of $2.4 \times 10^4 \text{ m s}^{-1}$ at right angles to the electric field.
20. Describe the conditions needed for an electron entering a magnetic field to undergo uniform circular motion.
21. In a tube similar to that used in the Thomson’s electromagnetic experiment, a magnetic field of 1.00×10^{-2} T is sufficient to allow the electrons to pass through the electric deflection plates. The plates are 10 mm apart and have a potential difference of 300 V across them.
- What is the strength of the electric field between the plates?
 - What was the speed of the electrons as they entered the region between the plates?
 - What was the strength of the magnetic force acting on the electrons?
22. Using the library or the internet, research how J. J. Thomson used his understanding of the nature of cathode rays to develop a new model of the atom.
23. Compare the methods used to control the movement of the cathode-ray beam in a CRO and in a television set.



10.1 DISCHARGE TUBES

Aim

To observe the effect that different gas pressures have on an electric discharge passed through a discharge tube.

Apparatus

power pack
two plug–plug leads
one set of discharge tubes
(with varying pressures)
induction coil
two plug–clip leads

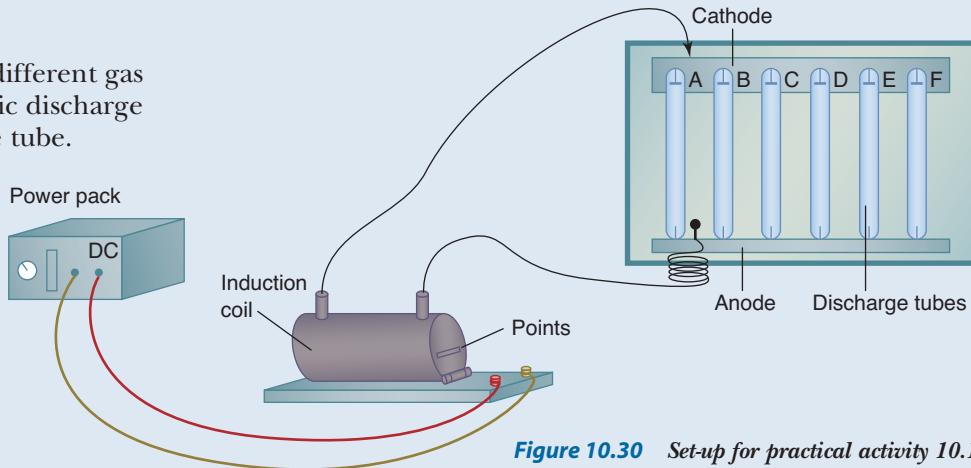


Figure 10.30 Set-up for practical activity 10.1

Theory

The high voltage produced by the induction coil is applied across the terminals inside the discharge tube. One plate (the cathode) becomes highly negative and releases a ray (cathode ray or electron). The electron passes through the gas in the tube and excites electrons in the atoms of the gas contained in the tube. The pressure of the gas determines the density of the atoms and therefore the nature of the collisions which take place between the electrons and atoms. Therefore, different discharge effects under different pressures can be observed (refer back to figure 10.4, page 176).

Method

Safety note

When the induction coil is connected to the discharge tube, X-rays are produced. However, it is the cathode rays hitting the glass or metal within the discharge tube that creates the X-rays, not the induction coil. If the experiment uses a minimum operating voltage these X-rays will be of a low energy and are significantly reduced after passing through the glass.

We need to deal with induction coils with extreme care because of the high voltages associ-

ated with them. Your teacher will set up the equipment related to the induction coil.

1. Attach the induction coil to the power pack using the two plug–plug leads. Adjust the points on the induction coil to obtain a continuous spark from the coil. Switch off the power pack.

2. Set the power pack at the correct setting for the induction coil (usually 6 volts) and turn it on.
3. Attach the negative terminal of the induction coil to the cathode of the discharge tube marked with the highest pressure (40 mmHg) and attach the positive terminal to the other end as shown in figure 10.30. Switch on the power pack.
4. Sketch a diagram of the pattern observed in this tube and describe it carefully.
5. Repeat the above procedure using each of the discharge tubes and see if you can observe streamers, Faraday's dark space, cathode glow, Crookes' dark space, striations and the positive column. Carefully describe each pattern, identifying each of the effects mentioned. (Tubes to be used should be 40 mmHg, 10 mmHg, 6 mmHg, 3 mmHg, 0.14 mmHg and 0.03 mmHg. These are represented as A, B, C, D, E and F in figure 10.30).

Questions

1. What effects were common throughout all tubes?
2. If the striations are produced by electrons (cathode rays) striking atoms and causing light to be released, give an explanation for the occurrence of variation in the patterns for different pressures.



10.2 PROPERTIES OF CATHODE RAYS

Aim

To determine some of the properties of the rays which come from the cathode of a discharge tube.

Apparatus

two power packs
two plug–plug leads
one pair of magnets
induction coil
four plug–clip leads
discharge tubes (maltese cross, electric plates, rotating wheel, screen display)

Theory

This experiment will most likely be performed as a class demonstration by your teacher. The discharge tubes used are illustrated in figure 10.3, page 175 and are similar to those Sir William Crookes would have used.

Method

Before starting, it would be advisable to read the ‘Analysis’ section of this experiment so as to plan what you should record during the experiment.

1. Connect the power pack to the induction coil and set it at 6 volts. Adjust the points on the

induction coil so that a strong steady spark is being produced, as in practical activity 10.1.

2. Connect the terminals of the induction coil to the discharge tube containing the maltese cross (Crookes’ tube). Observe the end of the tube containing the cross when the cross is down and when it is up.
3. Replace the Crookes’ tube with the tube containing the electric plates and connect the terminals of the plate to its high DC voltage supply. Observe the effects of the electric field on the cathode rays.
4. Connect the tube with the fluorescent screen display to the induction coil and record the effect of placing a set of bar magnets around the cathode rays as shown in figure 10.3 (page 175).
5. Finally, attach the tube containing the glass wheel on tracks to the induction coil and observe the effects that the cathode rays have on the wheel when the tube is horizontal.

Analysis

1. For each of the tubes placed in the circuit, sketch a diagram of the tube and the effect caused by the cathode rays.
2. Using the laws of electromagnetism, determine the charge that is evident on the cathode rays.

Questions

1. What are five properties of cathode rays which can be deduced from this experiment?
2. From these results, can we conclusively say that the cathode rays are electrons? Why or why not?

CHAPTER 11

THE PHOTOELECTRIC EFFECT AND BLACK BODY RADIATION

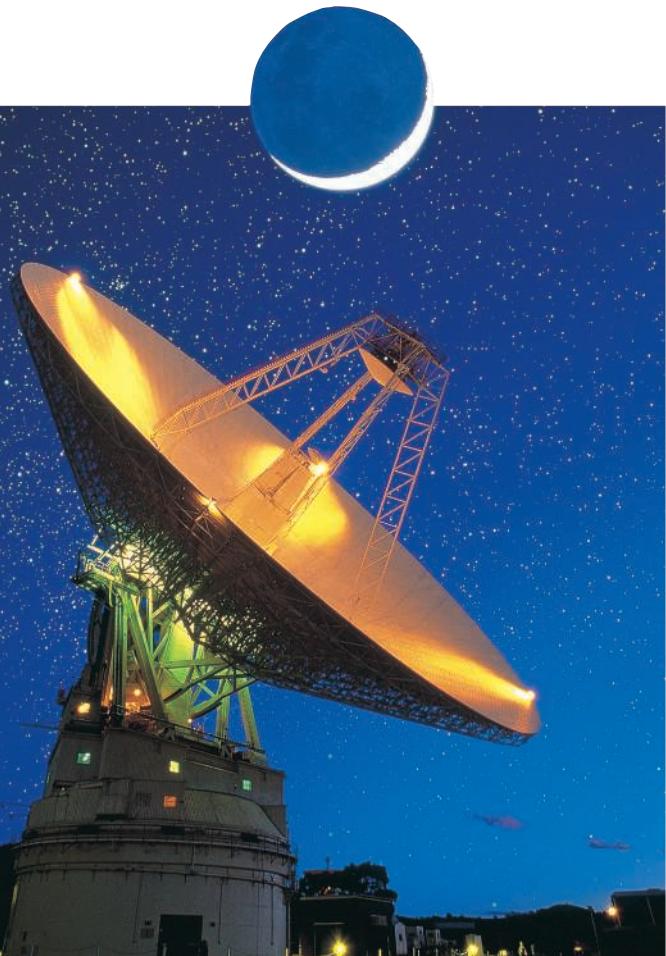


Figure 11.1 Radio telescopes are pointed into space.

They collect radio waves and other electromagnetic radiation from galaxies. The SETI Project (Search for Extra-Terrestrial Intelligence) utilises radio telescopes to 'listen' for intelligent signals from other intelligent beings. The electromagnetic spectrum we know today extends from the wavelength of gamma rays, as small as 10^{-14} m, through to radio waves with wavelengths of 10^5 m. That knowledge has come from the work of two of the giants of science, James Clerk Maxwell and Heinrich Hertz.

Remember

Before beginning this chapter, you should be able to:

- define and apply the terms *medium, displacement, amplitude, period, crest, trough, transverse wave, frequency, wavelength* and *velocity* to the wave model
- recall the terms *velocity, frequency* and *wavelength*, and their appropriate units, and solve numerical problems using $v = f\lambda$
- recall that different types of radiation make up the electromagnetic spectrum and that they are propagated through space at a constant speed, c
- explain that the relationship between the intensity of electromagnetic radiation and the distance from a source is determined by the inverse square law $I \propto \frac{1}{d^2}$
- recall the arguments in the previous chapter about the nature of cathode rays
- recall that the wavelength of the radiation emitted from an object depends on its temperature (black body radiation).

Key content

At the end of this chapter you should be able to:

- outline Hertz's experiments with the speed of radio waves, their properties compared to light, and the photoelectric effect
- describe the model of the black body and its role in understanding the particle nature of light
- identify the contributions that Planck and Einstein made to the concept of quantised electromagnetic energy and describe how this relates to the particle model of light
- identify the relationships between photon energy, frequency, speed of light and wavelength by using the formulas $E = hf$ and $c = f\lambda$
- outline the use of the photoelectric effect in breathalysers, solar cells and photocells.

In this chapter we will look at the changing ideas regarding the nature of light in the latter part of the nineteenth century and the early twentieth century. We will also look at how changes in theory and experimentation led to an understanding of black body radiation and the photoelectric effect and also introduced the ‘quantum theory’.

11.1 MAXWELL'S THEORY OF ELECTROMAGNETIC WAVES

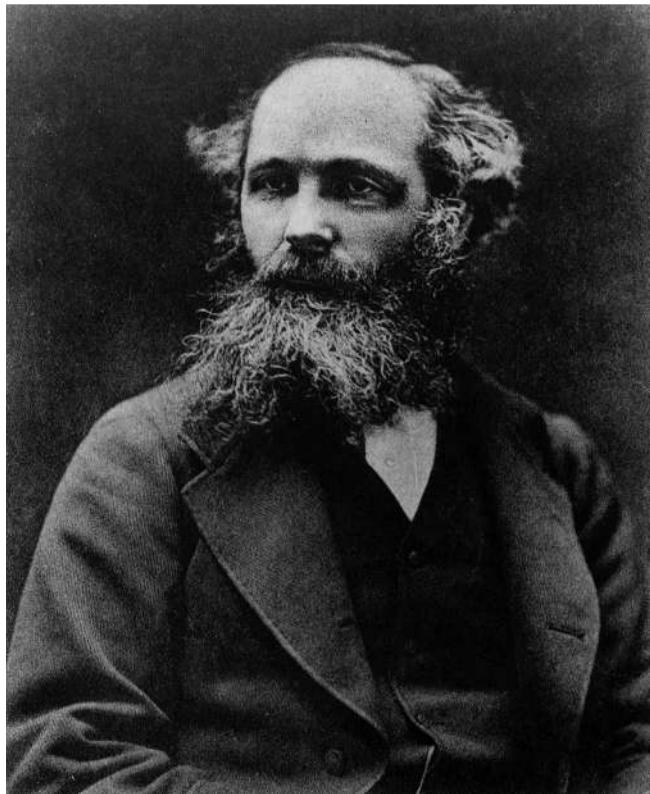


Figure 11.2 James Clerk Maxwell (1831–1879)

The passage of light across the vast universe was apparent to all who looked toward the heavens and saw the stars. Explaining how that could be was another matter. According to previous wave theory, waves were propagated through a medium. What was the medium in which light travelled?

One of the problems that nineteenth century scientists had in understanding how electromagnetic waves carry energy through the vacuum of space was that mechanical waves vibrate in a medium. Was there a medium filling space? One name given to this unproven medium was the luminiferous *aether*. The presence or absence of the aether was hotly debated. Proof of its presence or absence became one of the great goals of science. In the end, Albert Einstein simply said that its existence or absence was irrelevant. It could not be detected and made no difference to the passage of light.

Based on observations that a changing magnetic field induces an electric field in the region around a magnet, and that a magnetic field is induced in the region around a conductor carrying an electric current, James Clerk Maxwell concluded that the mutual induction of time- and space-changing electric and magnetic fields should allow the following unending sequence of events.

- A time-varying electric field in one region produces a time- and space-varying magnetic field at all points around it.
- This varying magnetic field then similarly produces a varying electric field in its neighbourhood.
- Thus, if an electromagnetic disturbance is started at one location (for example, by vibrating charges in a hot gas or in a radio antenna) the disturbance can travel out to distant points through the mutual generation of electric and magnetic fields.
- The electric and magnetic fields propagate through space in the form of an ‘electromagnetic wave’ (illustrated in figure 11.3).

The upshot of this sequence of events was clear. Light, and indeed any electromagnetic wave, does not need a medium to propagate. Electromagnetic waves are self-propagating. Once started, they have the capacity to continue forever without continuous energy input. You are probably familiar with the idea that the light we see coming from distant stars took millions, if not billions, of years to reach the Earth. It is possible that the origin of that light no longer exists, yet you can still see the light that emanated from the object.

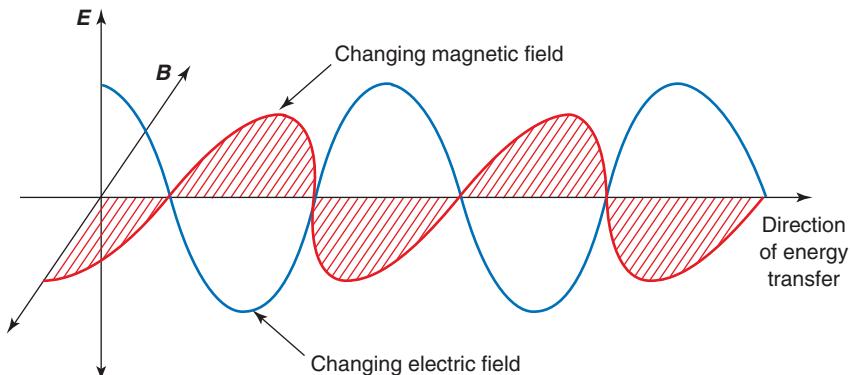


Figure 11.3 A diagrammatic representation of an electromagnetic wave

Maxwell's theory gave a definite connection between light and electricity. In a paper titled 'A dynamical theory of the electromagnetic field', which he presented before the Royal Society in 1864, Maxwell expressed four fundamental mathematical equations that have become known as 'Maxwell's equations'.

Maxwell's equations predicted that light and electromagnetic waves must be transverse waves and that the waves must all travel at the speed of light. They also implied that a full range of frequencies of electromagnetic waves should exist. In other words, the equations suggested the existence of an electromagnetic spectrum.

At the time of these predictions, only light and infra-red radiation was known and confirmed to exist. One look at the spectrum shown in figure 11.4 allows you to see how little was known of the complete electromagnetic spectrum known to exist today. Maxwell's equations also suggested that the speed of all waves of the full electromagnetic spectrum, if they did exist, was a definite quantity that he estimated as $3.11 \times 10^8 \text{ m s}^{-1}$. Maxwell's theoretical calculations were supported by the experimental data of French physicist Armand Hippolyte Louis Fizeau (1819–1896) who had determined a figure very close to this for the speed of light. In 1849, Fizeau's experiments to measure the speed of light had obtained a value of $3.15 \times 10^8 \text{ m s}^{-1}$.

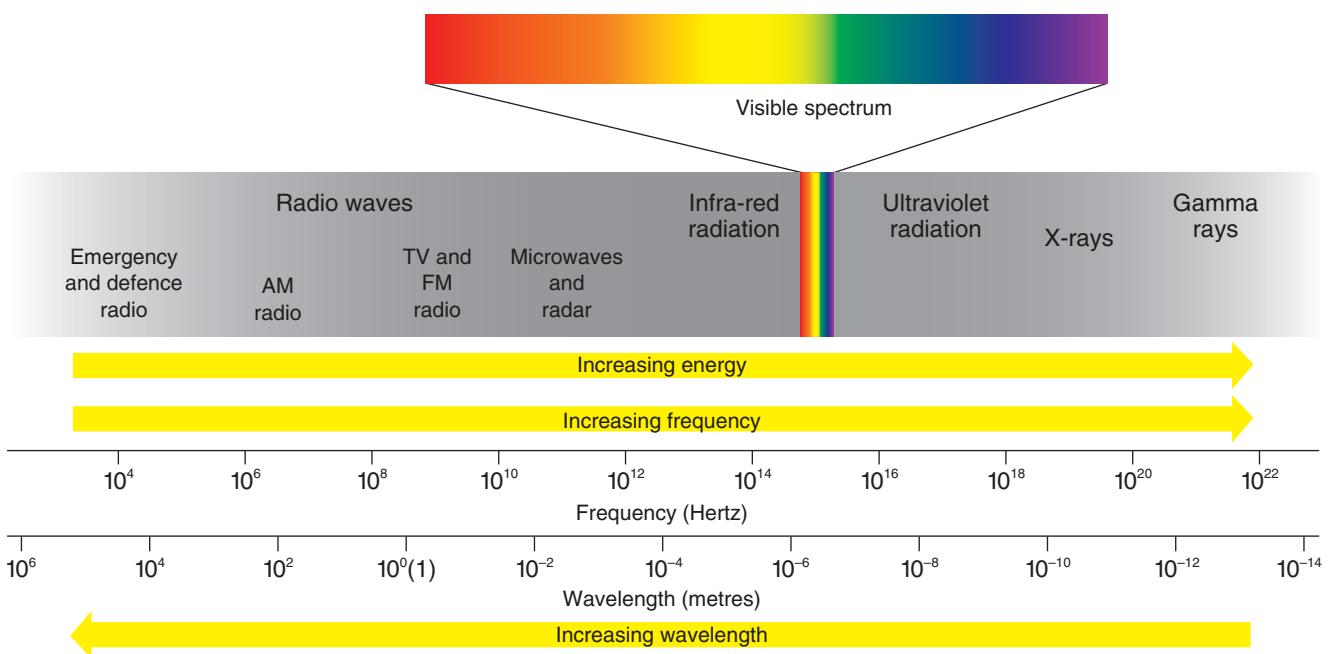


Figure 11.4 The electromagnetic spectrum

11.2 HEINRICH HERTZ AND EXPERIMENTS WITH RADIO WAVES



Figure 11.5
Heinrich Hertz (1857–1894)

Maxwell's two most important predictions were that:

- electromagnetic waves could exist with many different frequencies
- all such waves would propagate through space at the speed of light.

In 1886, Heinrich Hertz conducted a series of experiments that verified these predictions. Unfortunately, Maxwell had died in 1879 and did not see this experimental confirmation of the theoretical predictions of his equations.

Hertz reasoned that he might be able to produce some of the electromagnetic waves with frequencies other than that of the visible light predicted by Maxwell's equations. He thought he could produce some of these electromagnetic waves by creating a rapidly oscillating electric field with an induction coil that caused a rapid sparking across a gap in a conducting circuit.

In his experiments that confirmed Maxwell's predictions, Hertz used an induction coil to produce sparks between the spherical electrodes of the transmitter. He observed that when a small length of wire was bent into a loop so that there was a small gap and held near the sparking induction coil, a spark would jump across the gap in the loop. He observed that this occurred when a spark jumped across the terminals of the induction coil (see figure 11.6). This sparking occurred even though the loop was not connected to a source of electrical current.

Hertz concluded this loop was a detector of the electromagnetic waves generated by the transmitter. This provided the first experimental evidence of the existence of electromagnetic waves.

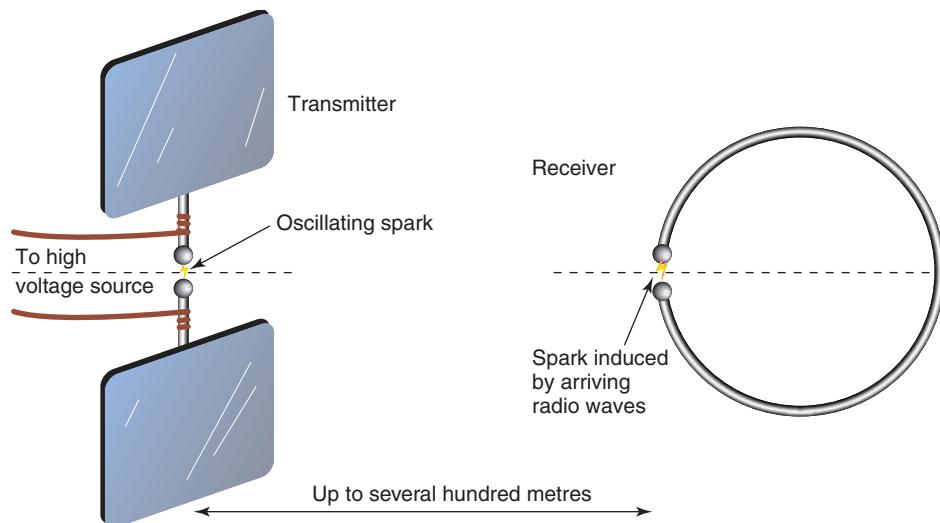


Figure 11.6 Hertz, using an induction coil and a spark gap, succeeded in generating and detecting electromagnetic waves. He measured the speed of these waves, observed their interference, reflection, refraction and polarisation. In this way, he demonstrated that they all have the properties characteristic of light.

Hertz then showed that these new electromagnetic waves could be reflected from a metal mirror, and refracted as they passed through a prism made from pitch. This demonstrated that the waves behaved similarly to light waves in that they could be reflected and refracted.

Additionally, Hertz was able to show that, like light, the new electromagnetic waves could be polarised. Hertz showed that the waves originating from the electrodes connected to the induction coil behaved as if they were polarised by rotating the receiver loop. When the detector loop was perpendicular to the transmitter gap, the radio waves from the gap produced no spark (see figure 11.7). The spark in the receiver was caused by the electric current set up in the conducting wire. When the detector loop was parallel to the spherical electrodes attached to the induction coil (see figure 11.8), the spark in the receiver was at maximum. At intermediate angles it was proportionally less. This was a behaviour similar to that shown by polarised light waves after the light has passed through an analyser, such as a sheet of Polaroid. It demonstrated that the newly generated electromagnetic waves were polarised.

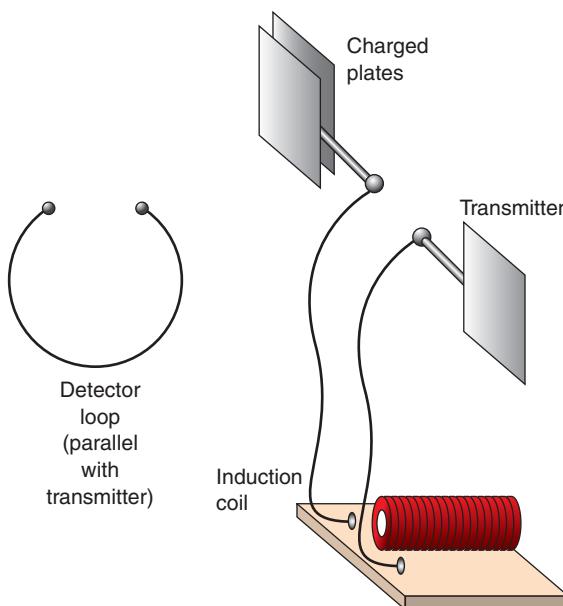


Figure 11.7 No spark was detected when the detector loop was rotated.

Hertz was the first to observe what we refer to as the ‘photoelectric effect’. This is discussed further on pages 202–208.

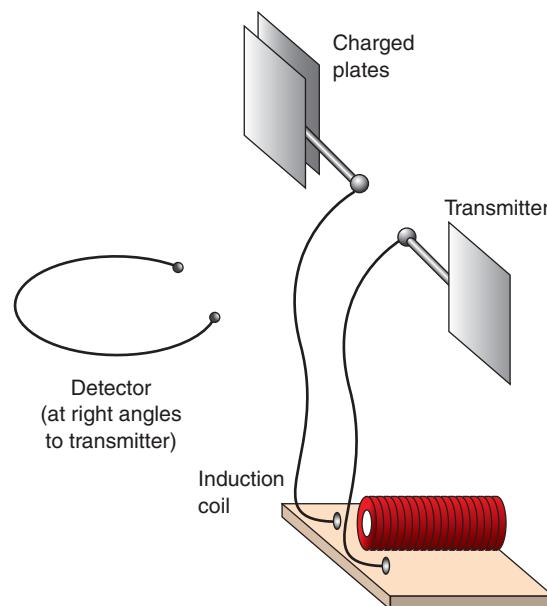


Figure 11.8 Hertz detected the waves when the detector loop was placed like this.

If Hertz’s interpretation was correct, and electromagnetic waves did travel through space from the coil to the loop, he reasoned there must be a small delay between the appearance of the first and second spark. The spark in the detector cannot occur at exactly the same time as the spark in the induction coil because even travelling at the speed of light it takes a finite time for the wave to move from one point to another. Hertz measured this speed in 1888 by using a determined frequency from an oscillating circuit and a measured wavelength as determined by interference effects for the waves produced and found that it corresponded to the speed predicted by Maxwell’s equations; it was the same as the speed of light.

To measure this wavelength, Hertz connected both the transmitter and the detector loop with a length of wire. He had already shown, by rotating the second loop, that the waves produced by the sparking

behaved as if they were polarised. The spark in the receiver was caused by the electric current set up in the conducting wire. At intermediate angles, interference of the currents provided a measure of the wavelength of the radio waves through the air.

The speed of transmission of the sparks was measured using a technique taken from light. Lloyd's mirror uses interference of two separate beams of light. One beam travels directly from the source to a detector. The other reflects a beam from the source from a mirror set at a small angle. Both beams interfere both constructively and destructively when they arrive at the detector. It is possible to use the pattern produced to determine the wavelength of the waves.

Hertz carried out a modification of this experiment, reflecting the sparks from a metal plate. He suggested that the waves produced had a wavelength larger than light, which should make measurements easier. Knowing the frequency of the sparks and their wavelength, he obtained a value for the speed of transmission. His value was similar to the speed of light measured by Fizeau (see page 195).

The invention of this set of experiments and procedures was the first time that electromagnetic waves of a known frequency could be generated. Today these waves originally produced by Hertz in his experiments are known as radio waves. Hertz never transmitted his radio waves over longer distances than a few hundred metres. The unit for frequency was changed from 'cycles per second' to the 'hertz' honouring the contribution of Heinrich Hertz.

Using the microwave apparatus available in schools, large wax prisms and metal plates, students can reproduce many of these results — demonstrating that electromagnetic waves have all the properties of light.

Radio waves and their frequencies

The discovery of radio waves was made by Hertz; however, the development of a practical radio transmitter was left to the Italian, Guglielmo Marconi (1874–1937). Marconi's experiments showed that for radio waves:

- long wavelengths penetrate further than short wavelength waves
- tall aerials were more effective for producing highly penetrating radio waves than short aerials.

The earliest radio messages were sent in 1895 by Marconi across his family estate, a distance of approximately three kilometres. By 1901 he was sending radio messages across the Atlantic Ocean from Cornwall, England to Newfoundland, Canada.

Different frequency radio waves can be generated easily and precisely by oscillating electric currents in aerials of different length. This is because the frequency of the waves generated faithfully matches the frequency of the AC current generating them. This ease of generation allows radio waves to be utilised extensively for many purposes. Applications include communication technologies, such as radio, television and mobile phones, and other technologies such as microwave cooking and radar. Essentially, the only difference between any of these waves used for these different purposes is the frequency of the waves generated by the transmitting aerials. For communications, sections of the available electromagnetic spectrum or bands of spectrum are used (see figure 11.4 on page 195). These chunks of spectrum use many single frequencies to transmit without interference.



11.1

Producing and transmitting radio waves

PHYSICS IN FOCUS

How do radio aerials work?

One form of aerial is the dipole antenna. Using an alternating electric current in the antenna, the electrons are continually accelerated back and forth. Electric charges (electrons) move back and forth along a length of conductor. The electric and magnetic fields of these moving charges produce an electromagnetic radiation that has the same frequency as the alternating current operating in the antenna. The electric current travels along the entire length of the aerial. Therefore, the greater the length of the aerial the longer the time for each cycle and hence, the lower the frequency of the electromagnetic wave generated. Long antennae therefore generate long-wavelength radio waves whereas

short-length antennae produce short-wavelength radio waves, which are called microwaves.

For communication bands such as radio waves, the length of the antenna can be as small as one half of the wavelength of the carrier wave (see figure 11.9(a) and (b)). When you consider the wavelength of radio waves, you can understand why radio antennae must be so tall. To avoid building very large antennae they are often mounted on buildings. One end is on the roof and the other end is grounded. This is to reduce the overall length of the actual antenna structure while still producing an antenna that is the height of the building and the antenna structure.

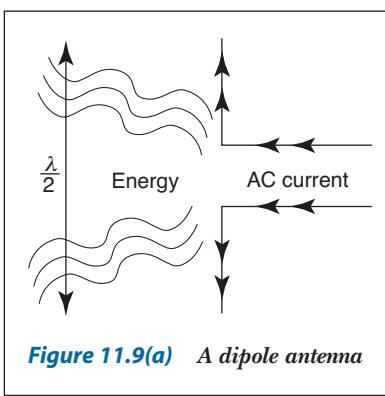


Figure 11.9(a) A dipole antenna

ducing an antenna that is the height of the building and the antenna structure.

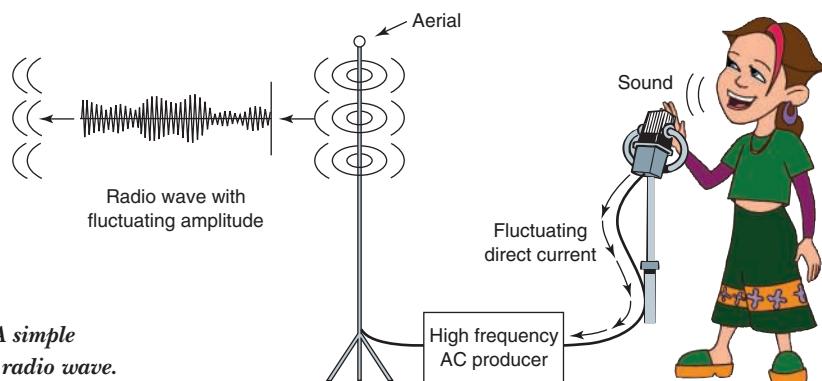


Figure 11.9(b) A simple aerial transmits a radio wave.

11.3 THE BLACK BODY PROBLEM AND THE ULTRAVIOLET CATASTROPHE

When an object such as a filament in a light globe is heated (but not burned) it glows with different colours: black, red, yellow and blue-white as it gets hotter. To understand how radiation is emitted for all objects, and how the wavelength of the radiation varies with temperature, creative experiments involving the behaviour of standard objects called 'black bodies' were required. A black body is one which absorbs all incoming radiation. The use of black bodies was necessary because all objects behave slightly differently in terms of the radiation they emit at different temperatures. Scientists could use the standard black body in experiments to study the nature of radiation emitted at different temperatures, and then extrapolate their findings for other objects.

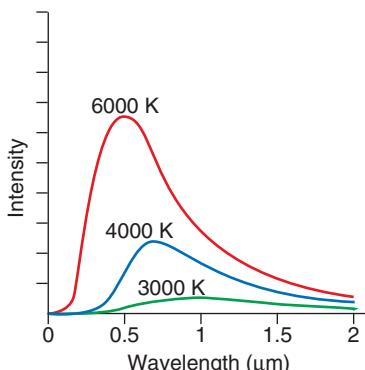


Figure 11.10 The peak in intensity moves to lower wavelength and higher frequency radiation with increasing temperature.

As an example of an object used to model a black body, imagine you drilled a very small hole through the wall of an induction furnace (an efficient oven in which the temperature can be set to known values). At a temperature of 1000°C, the walls of such a model black body will emit all types of radiation, including visible light and infra-red and ultra-violet radiation, but they will not be able to escape the furnace except through the small hole. They will be forced to bounce around in the furnace cavity until the walls of the furnace absorb them. As the walls absorb the radiation they will increase in energy. This causes the walls to release radiation of a different wavelength, eventually establishing an equilibrium situation. All radiation entering through the small hole is absorbed by the walls, so the radiation leaving the hole in the side of the furnace is characteristic of the equilibrium temperature that exists in the furnace cavity. This emitted radiation is given the name black body radiation.

As figure 11.10 shows, the radiation emitted from a black body extends over all wavelengths of the electromagnetic spectrum. However, the relative intensity varies considerably and is characteristic of a specific temperature.

Black bodies absorb all radiation that falls on them. That energy is spread throughout the object. The cavity walls within the black body also get hotter. As the walls of the cavity get hotter, the emission of more intense, shorter wavelength radiation from the cavity occurs. Physicists used a spectrometer to measure how much light of each colour, or wavelength, was emitted from the hole in the side of the black body models they constructed. The shape of the radiation versus intensity curves on the graphs that they created presented a problem for the physicists attempting to explain the intensity and wavelength variations that occurred quantitatively.

The problem was how to explain the results theoretically. The traditional mathematics based on thermodynamics predicted that the pattern of radiation should be different to that which the physicists found occurred.

The ‘classical’ wave-theory of light predicted that, as the wavelength of radiation emitted becomes shorter, the radiation intensity would increase. In fact, it would increase without limit. This would mean that, as the energy (that was emitted from the walls of the black body and then re-absorbed) decreased in wavelength from the visible into the ultraviolet portion of the spectrum, the intensity of the radiation emitted from the hole in the black body would approach infinity. This increase in energy level would violate the principle of conservation of energy and could not be explained by existing theories. This effect was called the ‘ultraviolet catastrophe’.

The experimental data from black body experiments (see figure 11.11) showed that the radiation intensity curve corresponding to a given temperature has a definite peak, passing through a maximum and then declining. This could not be explained.

The German scientist, Max Planck, arrived at a revolutionary explanation for the nature of the radiation emitted in experiments. Planck proposed that energy would be exchanged between the particles of the black body and the equilibrium radiation field. Using an analogy with the transmission of radio waves (with an aerial of specific length indicating the frequency of the radio wave radiation produced), the relatively high frequency of light emitted by a black body required an ‘aerial’ of a size similar to that of the atom for its production. The question was, how did this come to be?

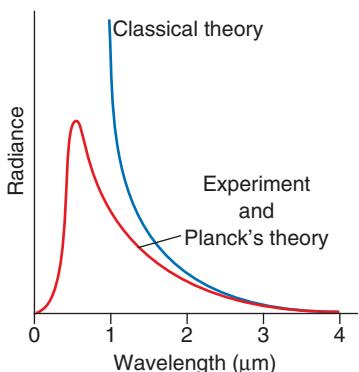


Figure 11.11 The predicted curve for the classical model and the actual curve obtained from experiment

Planck came up with a revolutionary idea to explain the results observed in experiments. He assumed that the radiant energy, although exchanged between the particles of the black body and the radiant energy field in continuous amounts, may be treated statistically as if it was exchanged in multiples of a small ‘lump’. Each lump is characteristic of each frequency of radiation emitted. He described this small, average packet as a ‘quantum’ of energy, that could be described by hf , where f was the frequency, and h a small constant, now called ‘Planck’s constant’ ($h = 6.63 \times 10^{-34} \text{ J s}$).

Therefore:

$$E = hf$$

where

E = energy, measured in joules

h = Planck’s constant = $6.63 \times 10^{-34} \text{ J s}$

f = frequency in hertz.

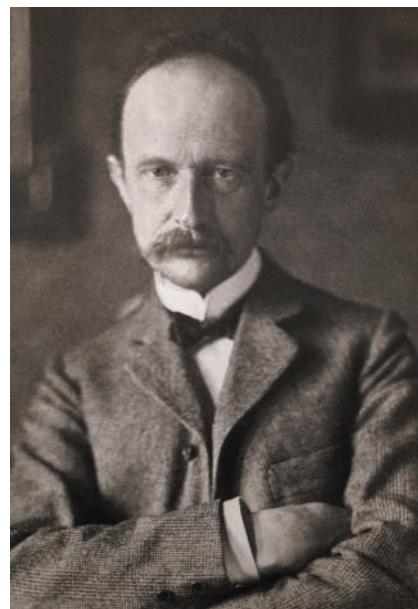


Figure 11.12 Max Planck (1858–1947) has come to be recognised as one of the key people involved in the development of modern physics.

This equation models the quantum relationship. This modification was seen by Planck as a small correction to classical thermodynamics. It turned out to be a most significant step towards the development of a totally new branch of physics: the quantum theory. Although he is considered to be the first to introduce this theory, Planck was never comfortable with the strict application of the quantum theory. He had invented the quantum theory but believed that all he had really done was to invent a mathematical trick to explain the results of black body radiation experiments. He failed to accept the quantisation of radiation until later in his career when the quantum theory was backed up with more examples and supporting evidence.

Armed with the Planck relationship and the knowledge that the speed of electromagnetic radiation ($c = 3 \times 10^8 \text{ m s}^{-1}$) was the product of the frequency of the radiation and the wavelength of the radiation ($c = f \times \lambda$), the energy in one quantum, or **photon**, of light of any known wavelength was then able to be determined.

The term **photon** describes a unit (or ‘packet’) of energy relating to the quantum description of matter. It is also a particle of electromagnetic energy with a zero rest mass.

SAMPLE PROBLEM

11.1

SOLUTION

Calculating photon energy

What is the energy of an ultraviolet-light photon, wavelength = $3.00 \times 10^{-7} \text{ m}$?

$$c = f\lambda \text{ so } f = \frac{c}{\lambda}$$

$$E = hf$$

$$= h \frac{c}{\lambda}$$

$$= \frac{6.63 \times 10^{-34} \times 3.00 \times 10^8}{3.00 \times 10^{-7}}$$

$$= 6.63 \times 10^{-19} \text{ J}$$

In this way the energy of a light photon of any known wavelength of light can be determined.

11.4 WHAT DO WE MEAN BY 'CLASSICAL PHYSICS' AND 'QUANTUM THEORY'?

A useful analogy is that of a slippery dip. Classical physics says that the difference in height from the top to the bottom is a continuous 'slide'.

Quantum mechanics says that there are a set number of small steps (the ladder) between the top of the slide and the ground.

Classical physics can be described, in broad terms, as physics up to the end of the nineteenth century. It relied on Newton's mechanics and included Maxwell's theories of electromagnetism. Classical physics still applies to large-scale phenomena and to the motion of bodies at speeds very much less than the speed of light. The quantum theory applies to the very small scale, particularly at the atomic level. Energy is believed to occur in discrete 'packets' or 'quanta'. Energy packets can be absorbed by an atom, and then re-radiated. Classical physics predicts that the emission of electromagnetic radiation is continuous; that is, can occur in any amount.

The difference between the classical description of 'continuous' energy and the new discrete, or quantised description of energy may be understood with a simple 'thought experiment'. Consider an old-fashioned balance, pivoted in the centre, with large pans suspended from each side. On one side we place a bucket with water and on the other we add house bricks to balance the bucket and water. We can increase the weight of bricks by adding or subtracting one brick at a time. If each brick has a mass of 2 kg, then our smallest packet, or quantum, of mass is 2 kg. On the other side, we can increase the weight by adding water. Without extending this example too far, we have a discrete variable (the bricks) and a continuous variable (the water) which can be added in smaller and smaller drops.

The continuous variable (water) is similar to the classical model of energy, and the bricks correspond to the quantum model in which energy can be exchanged in multiples of a small number. This simple example uses only one size of brick, or quantum. In fact the size of the quantum value varies with the energy effect we are studying.

Why don't we see these 'jumps' in light? The constant, h , now called 'Planck's constant', is equal to 6.63×10^{-34} J s. The quantum of energy is very small and, for all practical purposes, it is too small to be observed.

11.5 THE PHOTOELECTRIC EFFECT

The **photoelectric effect** is the name given to the release of electrons from a metal surface exposed to electromagnetic radiation. For example, when a clean surface of sodium metal is exposed to ultraviolet light, electrons are liberated from the surface.

The **photoelectric effect** is one of several processes for removing electrons from a metal surface. The effect was first observed by Hertz in 1887 when investigating the production and detection of electromagnetic waves using a spark gap in an electric circuit (see pages 196–198). Hertz used an induction coil to produce an oscillating spark. Hertz called the transmitting loop, spark A, and the detecting loop, spark B. In Hertz's own words, describing his detection of the photoelectric effect:

'I occasionally enclosed spark B in a dark case so as to more easily make the observations; and in so doing I observed that the maximum spark length became decidedly smaller inside the case than it was before. On removing, in succession, the various parts of the case, it was seen that the only portion of it which exercised this prejudicial effect was that which screened the spark B from spark A. The glass partition exhibited this effect not only in the immediate neighbourhood of spark B, but also when it was interposed at greater distance from B between A and B.'

eBook plus

Weblinks:

[Explaining the photoelectric effect](#)

[The photoelectric effect](#)

eModelling:

[Modelling the photoelectric effect](#)

A simple spreadsheet can easily be set up to enable playing with the idea of the 'work function' of a metal. Setting up the formulae for the spreadsheet will also help reinforce the relationship between the various physical quantities.

[doc-0042](#)

Hertz had discovered the photoelectric effect. He had found that illuminating the spark gap in the receiving loop with ultraviolet light from the transmitting gap gave stronger sparks in the receiving loop. Glass used as a shield between the transmitting and receiving loops blocked the UV. This reduced the intensity of sparking in the receiving loop. When quartz was used as a shield there was no drop in the intensity of sparking in the receiving loop. Quartz allowed the UV from the transmitted spark to fall on the detector.

Wilhelm Hallwachs (1859–1894) read the extract written by Hertz describing the photoelectric effect in a journal and designed a simpler method to measure this effect. He placed a clean plate of zinc on an insulating stand and attached it by a wire to a gold leaf electroscope. He charged the electroscope negatively and observed that the charge leaked away quite slowly. When the zinc was exposed to ultraviolet light from an arc lamp, or from burning magnesium, charge leaked away quickly. If the electroscope was positively charged, there was no fast discharge (see figure 11.13).

Figure 11.13 A positively charged electroscope is not affected by illumination with UV light, while the charge on a negatively charged electroscope discharges.

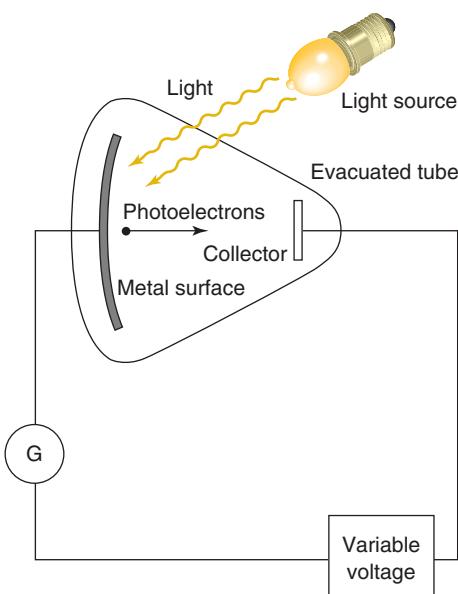
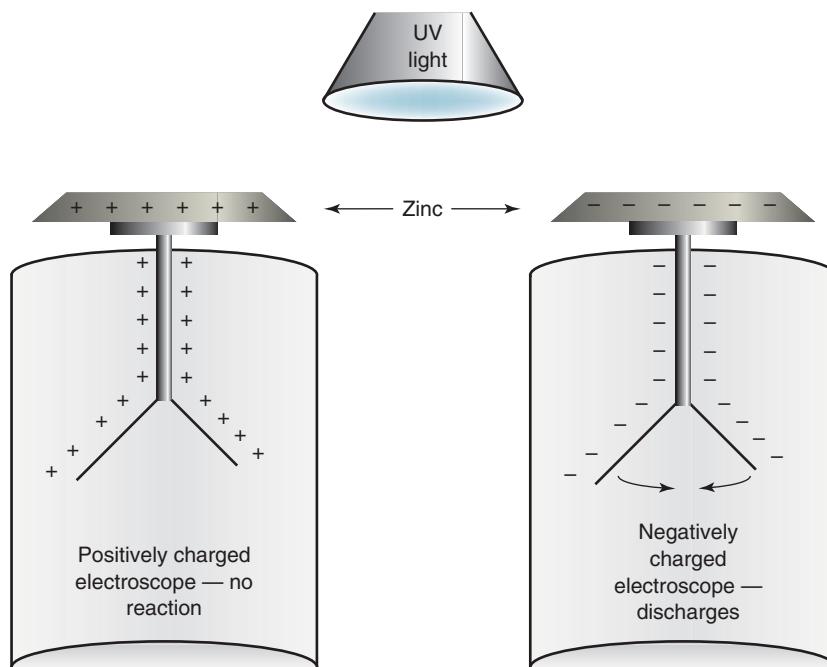


Figure 11.14 Apparatus used to demonstrate the photoelectric effect

In 1899, J. J. Thomson established that the ultraviolet light caused electrons to be emitted from a sheet of zinc metal and showed that these electrons were the same particles found in cathode rays. He did this by enclosing the metal surface to be exposed to ultraviolet light in a vacuum tube (see figure 11.14).

The new feature of this experiment was that the electrons were ejected from the metal by radiation rather than by the strong electric field used in the cathode ray tube. At the time, recent investigations of the atom had revealed that electrons were contained in atoms and it was proposed that perhaps they could be excited by the oscillating electric field.

In 1902, Hungarian-born German physicist Philipp von Lenard (1862–1947) studied how the energy of emitted photoelectrons varied with the intensity of the light used. He used a carbon arc lamp with which he was able to adjust the light intensity. He found in his investigations using a vacuum tube that photoelectrons emitted by the metal cathode struck another plate, the collector.

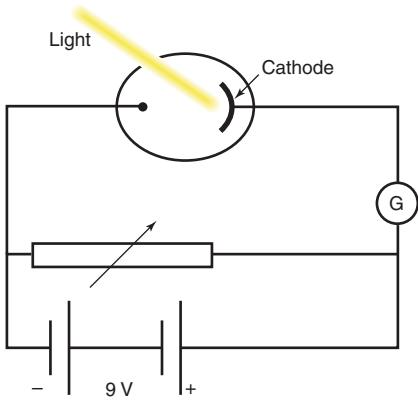


Figure 11.15 The voltage applied across the variable resistor opposes the motion of the photoelectrons. The electrons that reach the opposite electrode create a small current, measured by the galvanometer. The value of the voltage at which the current drops to zero is known as the stopping voltage.

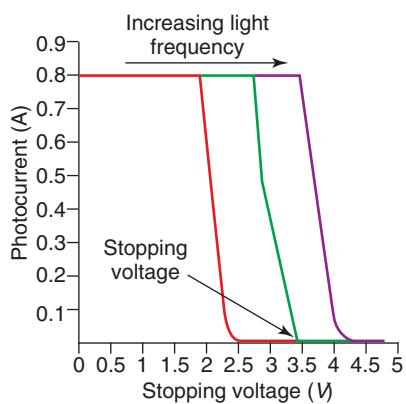


Figure 11.17 For a given light intensity, increasing the frequency of the light increases the maximum kinetic energy with which the photoelectrons are emitted.

When each electron struck the collector, a small electric current was produced that could be measured. To measure the energy of the electrons emitted, Lenard charged the collector negatively to repel the electrons. By doing so, Lenard ensured that only electrons ejected with enough energy would be able to overcome this potential hill (see figure 11.15). Surprisingly, he found that there was a well-defined minimum voltage, V_{stop} (see figure 11.16).

Lenard was also able to filter the arc light to investigate the effect that different frequencies of incident electromagnetic radiation had on photoelectron emission.

Lenard observed that:

- doubling the light intensity would double the number of electrons emitted
- there was no change in the kinetic energy of the photoelectrons as the light intensity increased
- the maximum kinetic energy of the electrons depends on the frequency of the light illuminating the metal, as figure 11.17 shows.

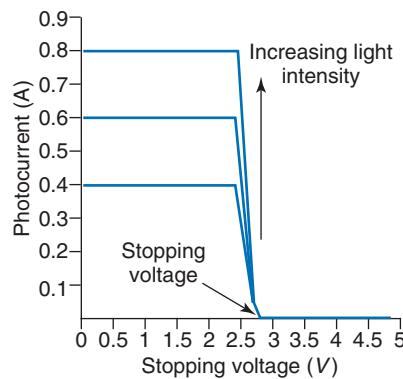


Figure 11.16 For a given frequency, photoelectrons are emitted with the same maximum kinetic energy because the electrons are all stopped by the same voltage. Increasing the intensity of the light increases the number of electrons released from the surface, causing an increase in the photocurrent.

Explanation of the photoelectric effect

Classical physics was unable to explain the photoelectric effect. Maxwell had predicted electromagnetic radiation, and Hertz had confirmed that light was electromagnetic radiation. Light was a wave phenomenon, and according to the classical theory of Maxwell, somehow or other the light waves falling on the surface of a metal would cause the emission of electrons.

The key observations a theory had to explain are:

1. If the radiation falling on the metal surface is going to cause emission of photoelectrons, they may be emitted almost immediately after the light falls on the metal. There may be no significant time delay, even if the intensity of the light, and hence the rate at which energy is being transferred to the metal surface, is very low.
2. There is a cut-off (or threshold) frequency. Radiation of lower frequency than a particular value will not cause the emission of photoelectrons, regardless of how bright the light source is. (Very intense light, which carries a large amount of energy, cannot cause emission of photoelectrons if the frequency of the light is less than the threshold frequency.)
3. If the light does cause the emission of photoelectrons, increasing the intensity of the light will increase the number of photoelectrons emitted per second.
4. The energy of the photoelectrons does not depend on the intensity of the light falling on the surface of the metal, but it does depend on the frequency of the light. Light of higher frequency causes the emission of photoelectrons with higher energy.

Classical physics had difficulties with three of these four observations.

- According to a wave theory of light, the light waves would distribute their energy across the whole of the metal surface. It might be expected that all electrons in the atoms of the metal (or at least all of the outer electrons) would gain energy from the light waves. If the light was very faint, it could take a considerable time for the electrons to gain sufficient energy and for one electron to be able to escape from the metal surface, but this is not what is observed.
- There was no way to explain the cut-off frequency.
- There was no way to explain the relationship between the frequency of the light and the kinetic energy of the most energetic electrons.
- The only observation that could be explained by a wave theory was the fact that increasing the intensity of light increases the rate of emission of photoelectrons.

The problems associated with explaining the photoelectric effect were solved by Einstein in 1905. Although his paper is often referred to as his photoelectric paper, it was in fact a paper on the quantum nature of light, and the photoelectric effect was just one of several examples used by Einstein to illustrate the quantum nature of light.

In this paper, Einstein states: ‘The simplest conception is that a light quantum transfers its entire energy to a single electron.’ In other words, the light quantum is acting as a particle in a collision with an electron.

This ‘light quantum’ model of light is able to explain all four observations of the photoelectric effect. (The term photon was not introduced until 1926, but we will use it at this stage to refer to a ‘light quantum’.)

If a photon strikes a metal surface, it will collide with a single electron in an atom in the metal. All of the energy of the photon will be passed on to the electron. This electron may gain sufficient energy to escape from the metal surface, so there is no problem with a time delay. Even with very faint light, the first photon to strike the metal surface could possibly cause the emission of an electron.

The energy of a photon is related to its frequency ($E = hf$). A certain minimum energy is required to cause the emission of an electron from the metal surface (this is the **work function** for that metal). If the energy of the photon is less than this value, the photon cannot cause an electron to be emitted. This explains the threshold frequency.

Increasing the intensity of the light increases the rate at which photons fall on the metal surface, hence the rate of emission of photoelectrons increases.

As the frequency of the light is increased beyond the cut-off frequency, more energy is provided to the electrons, hence the kinetic energy of the most energetic electrons increases.

This last idea played an important part in the development of the particle nature of light in the photoelectric effect. In 1905, observations of sufficient accuracy were not available to Einstein to test his equation; it was not until 1916 that Millikan, reluctantly, provided that evidence. Little attention had been paid to Einstein’s ideas of photoelectric effect in the intervening ten years. Millikan announced his results in 1916 but said, ‘The Einstein equation accurately represents the energy of the electron emission under irradiation with light [but] the physical theory upon which the equation is based [is] totally unreasonable.’ However, Millikan also stated that his results, combined with Einstein’s equation, provided ‘the most direct and most striking evidence so far obtained for the reality of Planck’s h ’.

The **work function** is the minimum energy required to release the electron from the surface of a particular material.

Einstein's photoelectric equation

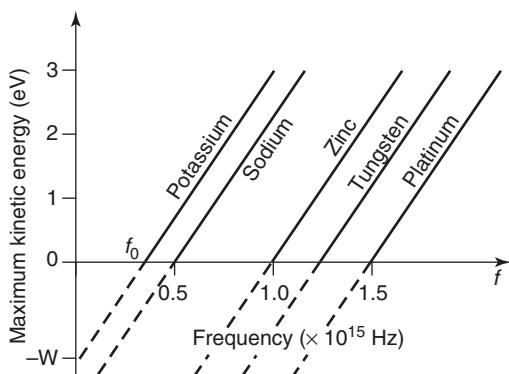


Figure 11.18 This graph shows the maximum kinetic energy with which the photoelectrons are emitted versus the frequency of light, for five different metals. Note that the gradient of all the lines is equal to Planck's constant.

As previously discussed, when different metal surfaces are illuminated with monochromatic light, electrons may be ejected from the metal surface. These electrons are called photoelectrons. Different metals hold electrons with different forces. Providing the photons of light illuminating the metal have sufficient energy (are of a high enough frequency) to overcome the energy holding the electrons in the metal, the electrons may be emitted. Only a small proportion of such electrons will in fact escape from the metal surface, and the emitted electrons will have a spread of energies, as some electrons may have required energy to move them to the metal surface. We will deal with the most energetic electrons emitted.

If a graph is plotted of the maximum kinetic energy of the emitted electrons versus the frequency of the light, the gradient of the lines representing different metals is the same (see figure 11.18). The point at which the lines intercept the frequency axis is a measure of the threshold frequency for that metal. If the frequency of the monochromatic light is below this threshold frequency, no photoelectrons will be emitted from the metal surface.

The lines for all of the metals are parallel and have a gradient equal to Planck's constant.

If we apply the general gradient equation $y = mx + b$ to any of the lines on this graph, we find that:

$$E_{k \max} = hf - W$$

This equation is an energy equation:

$E_{k \max}$ = the maximum kinetic energy of an emitted electron

W = the minimum energy required to remove the electron from the metal surface (the work function of the metal)

hf = the energy of the incident photon.

The energy of Einstein's 'light quantum' is hf , so this equation represents an interaction between an individual quantum of light (a photon) and an individual electron.

Of course, we now have light behaving as a particle in the photoelectric effect but as a wave in other phenomena (such as interference and diffraction). The photon has a dual wave and particle nature.

PHYSICS FACT

Both Planck and Einstein lived in Germany during the early part of the twentieth century. Working in the same area of physics, they were firm friends. However, during World War I, this friendship became strained. Einstein was a pacifist, while Planck strongly supported the German cause, even though he lost his son in battle in 1915.

With the rise of Hitler and the anti-Semitism movement, Einstein, who was Jewish, emigrated to the United States during the 1930s. Planck was able to continue his academic career in Berlin, even in the face of the hostility of anti-Semitism

groups towards the 'decadent Jewish science' of relativity and the quantum theory.

It is difficult to understand the pressures experienced by Planck, who tried to protect his Jewish friends and students throughout the period. Unlike Einstein, he did not see the moral imperative of opposing Hitler but tried to compromise and work within the system. Einstein remained a pacifist, yet some would say that he compromised some of these ideals with his support of the development of the atom bomb. He later wrote of the pain he experienced when the bombs were finally used. See chapter 25 for more detail on the nuclear bomb.

Applications of the photoelectric effect

Solar cells — an important use of the photovoltaic effect



Figure 11.19 A panel of commercial solar cells

A **photocell** is a device that uses the photoelectric effect. These devices include photovoltaic, or solar cells, which convert electromagnetic energy, such as sunlight, into electrical energy. Other examples are photoconductive cells and phototubes.

A **photocell** is a device that converts energy from sunlight into electrical energy. The first true solar cell was made in 1889 by Charles Fritts, using a thin selenium wafer covered with a thin layer of gold. By 1927, light-sensitive devices, such as light meters for photography, became generally available.

Modern solar cells use silicon and gallium arsenide. They use focusing devices, such as lenses, to achieve efficiencies of greater than 37% in converting light energy into electricity. An efficiency of 37% means that 37% of the light energy falling onto the cell is converted into electricity.

The work function of most materials requires the electromagnetic energy to have a frequency near that of ultraviolet light to allow electrons to be emitted. This limits the application of photocells using the photoelectric effect. Devices that use p-n junctions are more commonly found in the generation of power and in the detection and measurement of light.

In a solar cell (see figure 11.20), light energy is applied to the junction region of a semiconductor diode where p-type silicon is in contact with n-type silicon. Electrons are released from the silicon crystal lattice because of the photoelectric effect. This has the effect of raising the junction voltage. For the solar cell to work, the n-type layer is exposed to light and the p-type layer is not. On the light-exposed side of the solar cell a fine grid of metal provides electrical contacts. These contacts are able to collect the photoelectrons emitted from the light-exposed n-type silicon surface.

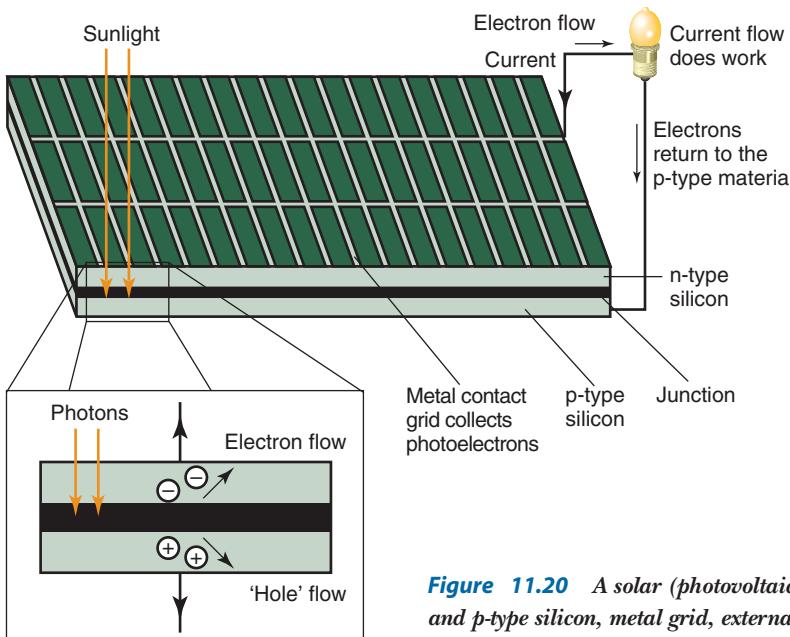


Figure 11.20 A solar (photovoltaic) cell showing the junction between the n-type silicon and p-type silicon, metal grid, external circuit and electron path. Note that the direction of hole and electron migration from the junction is opposite.



Figure 11.21 An example of the use of solar cells to provide energy to power a telephone in a remote location

A **photoconductive cell**, or photo-resistor, uses the fact that electrical resistance is affected by light falling onto it.



Figure 11.22 Example of the use of a photoconductive cell

Photoconductive cells

Most modern **photoconductive cells** are made of semiconductor material. This material is described more fully in chapter 12. Briefly, it is a mixture of elements that has some, but not all, electrons free to conduct electricity. The rest are confined in a lattice structure between atoms. To produce the photoconductive effect, light energy causes the electrons to be released from their valence bonds in a material. The number of free or mobile electrons in a semiconductor is limited and the addition of the light-released electrons raises the conductivity, or reduces the resistance, of the semiconductor. The resistance change (in ohms) may be in the order of many thousands of times.

Many different substances are photoconductive, that is, they conduct when exposed to light but do not when in the dark. Examples of photoconductive materials include lead

sulfide (PbS), lead selenide (PbSe) and lead telluride (PbTe). These are extremely sensitive and become conductive when exposed to electromagnetic radiation in the infra-red range. Cadmium sulfide (CdS) is also sensitive and becomes photoconductive when exposed to light in the visual range.

The photoconductive cell or photo-resistor was the first photoelectric device created. It was developed in the late nineteenth century. Photoconductive cells are used as switches to turn street lights on and off according to the amount of light energy naturally available at any particular time. They can also be used as light gates. You have used light gates to measure velocity by detecting the time that a light beam is blocked by a glider moving along a linear air-track. Such gates are used in industry as counting devices on production lines, and in packaging, and can be employed in alarm systems.

Because photoconductive cells can register the amount of light, they can be used to ascertain when concentrations of particles exceed certain levels in liquids and gases. When light scatters from particles in the water, for example, less light passes through into the detector. This can be calibrated to measure the amount of pollution. They can also be used to test the purity of commercially produced drinks and water supplies.

Photoconductive cells are part of the sensor that is used to scan bar-codes on grocery items at the supermarket checkout. They are also used in light meters to measure the intensity of illumination in photography.

Phototubes

Phototubes, also known as photocells, are commonly used as the ‘electric eyes’ to open automatic doors in shopping centres. In public toilets, the cells are used to turn taps on and off when people wash their hands. Some holiday resorts use the ‘eyes’ to trigger a presentation of descriptive audio or visual information when a tourist enters a room. They can also be used in astronomy to measure electromagnetic radiation from celestial objects such as stars and galaxies. Photomultiplier cells are extremely sensitive detectors. They effectively multiply the number of electrons released by a factor of a million.

SUMMARY

- In 1887, Hertz showed that both radio and light waves are electromagnetic waves — involving varying or changing electric fields, coupled with a changing magnetic field that was perpendicular to the electric field.
- The relationship $v = f\lambda$, relates the frequency (Hz) and the wavelength (m) to the velocity (m s^{-1}). In a vacuum, all electromagnetic waves travel at the speed of light (usually referred to as c) and thus $c = f\lambda$.
- The photoelectric effect is the release of electrons from materials, usually metals, by the action of light or some other electromagnetic radiation such as X-rays or gamma radiation.
- An oscillating or vibrating atom can emit electromagnetic energy only in discrete packets, or quanta. For light, in particular, the basic quantum of energy is a photon.
- The relationship between the frequency, f , and the amount of energy emitted, E , is given by the formula $E = hf$ where h is Planck's constant.
- Quantum mechanics states that the amount of energy emitted from a black body (or any other atomic-level energy transformation) is quantised so that it can increase only in certain small steps.
- Albert Einstein explained the photoelectric effect, based on Planck's work on black body radiation, and included the development of the idea of a quantum of energy or photon being the basic unit of energy.
- As scientists, Einstein and Plank both experienced pressure from social and political forces and coped in different ways.
- When a photon hits electrons in a metal it will release either all or none of its energy. An individual electron cannot accept energy from more than one photon.
- The amount of energy required to overcome the attractive forces of an electron in the electron 'sea' is called the work function of the electron.
- The maximum kinetic energy of the electron after it has been ejected from the metal's surface can be determined by measuring the stopping voltage, v_{stop} .
- A photocell is a device which uses the photoelectric effect to generate or control an electric current. Applications include solar cells and breathalyser.

QUESTIONS

- A beam of monochromatic light falls onto a cold, perfect black body and imparts 0.10 mW of power to it. If the wavelength of the light is 5.0×10^{-7} m, calculate:
 - the frequency of the light
 - the energy per photon for the light
 - the number of photons per second striking the black body.
- A beam of UV light of frequency 7.0×10^{15} Hz is incident on the apparatus shown in figure 11.23. If the maximum kinetic energy of an emitted electron is 9.0×10^{-19} J, calculate:
 - the potential required to stop electrons reaching the collector
 - the work function of the material on which the light is shining
 - the threshold frequency of the material.

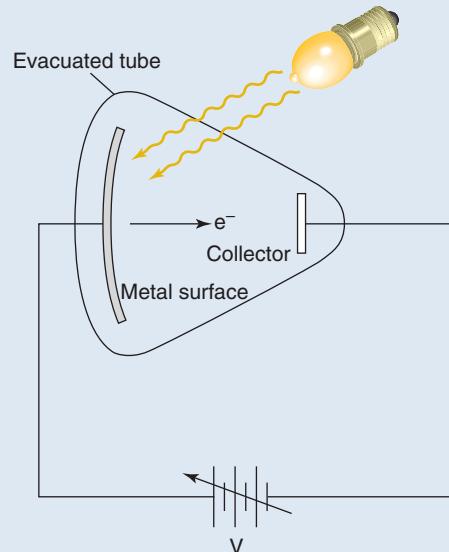


Figure 11.23

- Examine figures 11.7 and 11.8 (page 197). Explain why Hertz came to the conclusion that the waves produced by the sparks were polarised.
- Demonstrate the dependence of colour on the temperature of a black body. In figure 11.10 (page 200), what is the significance of the peak of the curve?
- Maxwell, Michelson and Hertz carried out experiments on electromagnetic waves of different frequencies. Compare their observations and discuss how an understanding of the electromagnetic spectrum was developed.

6. A scientist is investigating the effect of different types of radiation on the surface of a piece of sodium metal. A method is devised to cut a new surface across the sodium plate while in vacuo, since sodium is highly reactive and oxidises quickly. The apparatus is finally set up as shown in figure 11.24. The two variables under investigation will be the frequency, f , of the radiation and the kinetic energy, E_k , of the photo-electrons.
- (a) Should the sodium plate be positively or negatively charged in order to make the proper investigations?
- (b) Results for the experiment are as follows:
- | FREQUENCY OF INCIDENT LIGHT ($\times 10^{14}$ Hz) | STOPPING POTENTIAL (volts) |
|--|----------------------------|
| 5.4 | 0.45 |
| 6.8 | 1.00 |
| 7.3 | 1.15 |
| 8.1 | 1.59 |
| 9.4 | 2.15 |
| 11.9 | 2.91 |
- Record these results on a $E_{k\text{ max}}$ versus frequency graph.
- (c) Determine the threshold frequency for the sodium metal.
- (d) Determine a value for Planck's constant, h , from your graph.
- (e) What is the work function for sodium?
7. Above a particular (and specific to each material) threshold frequency of electromagnetic radiation, electrons are ejected immediately. Below this threshold frequency, electrons are never ejected. Explain how the photon model for light, rather than the wave model, explains this behaviour.
8. A photon collides with an electron and is scattered backwards so that it travels back along its original path. Describe and explain the expected wavelength of the scattered light.
9. The light from a red light-emitting diode (LED) has a frequency of 4.63×10^{14} Hz. What is the energy change of electrons that produce this light?
10. We can detect light when our eye receives as little as 2.00×10^{-17} J. How many photons of light, with a wavelength of 5.50×10^{-7} m, is this?
11. A red laser emitting 600 nm ($\lambda = 6.0 \times 10^{-7}$ m) wavelength light and a blue laser emitting 450 nm light emit the same power. Compare their rate of emitting photons.
12. One electron ejected from a clean zinc plate by ultraviolet light has kinetic energy of 4.0×10^{-19} J.
- What would be the kinetic energy of this electron when it reached the anode, if a retarding voltage of 0.90 V was applied between anode and cathode?
 - What is the minimum retarding voltage that would prevent this electron reaching the anode?
 - All electrons ejected from the zinc plate are prevented from reaching the anode by a retarding voltage of 4.3 V. What is the maximum kinetic energy of the electrons ejected from the zinc?
 - Sketch a graph of photocurrent versus voltage for this metal's surface. Use an arbitrary photocurrent scale.
13. The waves of the electromagnetic spectrum share some similarities and have some differences. What are their similarities? What are their differences?
14. (a) What is the wavelength of the radio waves broadcasting station 2MMM in Sydney if the frequency of the broadcast is 104.9 MHz?
- (b) What is the energy of a photon of that wave?
15. Arrange the following electromagnetic waves in order of increasing energy levels:
long-wave radio waves, gamma ray, blue light, red light, infra-red light, microwave radio waves, X-rays.

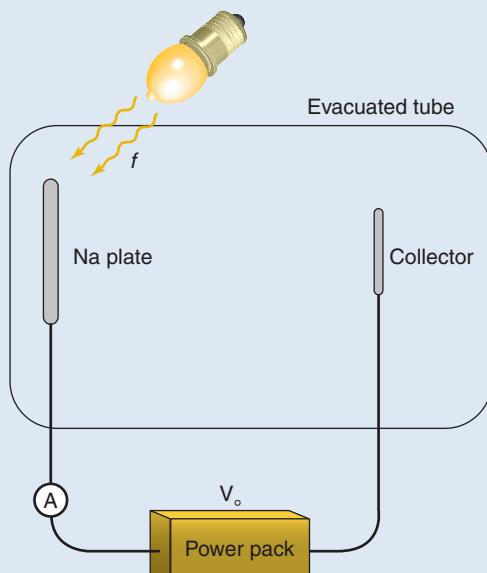


Figure 11.24

16. What were the features of radio waves that were demonstrated by Hertz's experiments?
17. What is the energy of an X-ray of wavelength 2.5×10^{-11} m?
18. If a black body was to gain sufficient energy to raise its temperature from 2000 K to 4000 K, describe how a plot of its radiant intensity versus radiant wavelength of the electromagnetic radiation would change.



11.1

PRODUCING AND TRANSMITTING RADIO WAVES

Aim

To demonstrate the production and transmission of radio waves.

Apparatus

induction coil
transformer rectifier with leads
small transistor radio

Method

1. Adjust the gap on the induction coil to about 5 mm and adjust the transformer to 6 V DC.

Warning: The spark across the gap of an induction coil generates long-wavelength X-rays and short-wavelength ultraviolet radiation. These are potentially dangerous, and students should not stand closer than one metre.

2. Adjust the tuner of the radio, so that it does not receive a station.
3. Move around the room and try to estimate where the radio can receive the static noise from the spark.
4. Adjust the gap to 10 mm, and repeat the exercise.
5. Change the tuner of the radio and scan across the range of wavelengths.

Questions

1. What is the maximum distance from the induction coil at which the radio receives static noise from the 5 mm spark?
2. Is the distance different when the spark is produced by a 10 cm gap?
3. Can you detect any pattern in the static received at different wavelengths?
4. An induction coil is an example of a transformer. What can you infer about the voltage across the gap and the resulting charge movement observed as the spark?
5. In what form is the energy transferred from the spark to the radio? In what manner must charges move to produce this energy?

CHAPTER 12

THE DEVELOPMENT AND APPLICATION OF TRANSISTORS

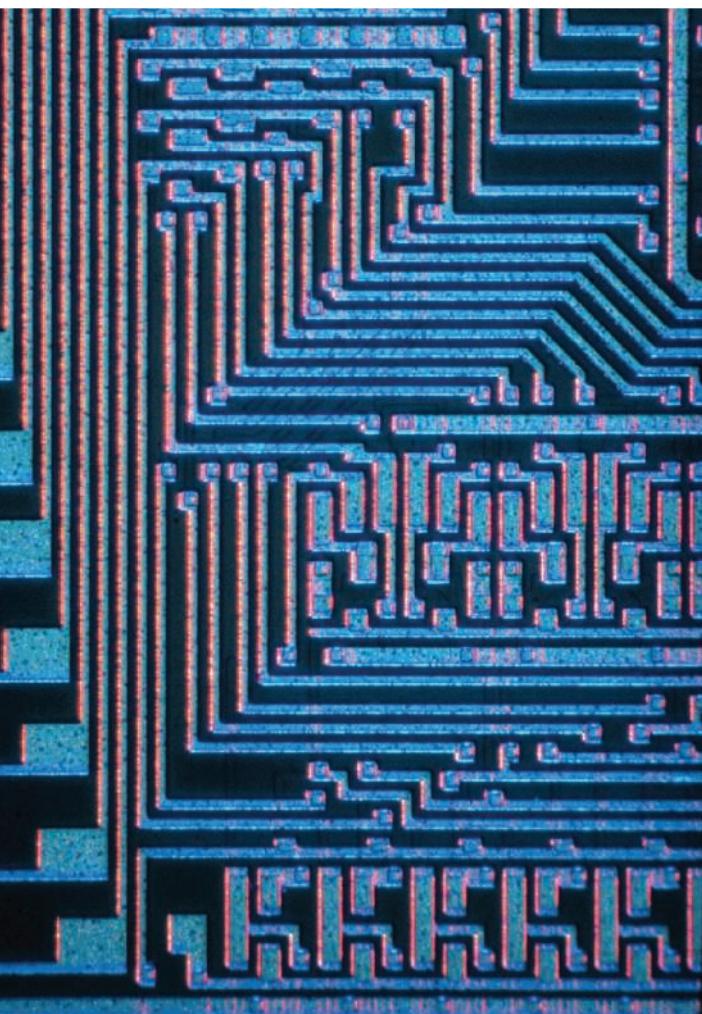


Figure 12.1 An example of an integrated circuit

Remember

Before beginning this chapter, you should be able to:

- discuss how the length, cross-sectional area, temperature and type of material affect the movement of electricity through a conductor
- describe the relationship between potential difference, current and power dissipated
- recall the arguments relating to the ‘wave–particle’ debate of matter and energy.

Key content

At the end of this chapter you should be able to:

- identify that some electrons in solids are shared between atoms and can move freely
- describe the difference between conductors, insulators and semiconductors in terms of band structures and relative electrical resistance
- identify absences of electrons in a nearly full band as holes, and recognise that both electrons and holes help to carry current
- compare the relative number of free electrons that can drift from atom to atom in conductors, semiconductors and insulators
- discuss the use of germanium and silicon as raw materials in transistors
- describe how ‘doping’ a semiconductor can change its electrical properties
- identify the differences in p-type and n-type semiconductors in terms of the relative number of negative charge carriers and positive holes
- discuss the differences between solid state and thermionic devices and discuss why solid state devices have largely replaced thermionic devices
- assess the impact on society of the invention of transistors, relating particularly to microchips and microprocessors.

In this chapter you will extend your knowledge of the electrical nature of matter and how this electrical nature led to the development of transistors and integrated circuits. Thermionic (radio) valves, although often unreliable, were used in all electronic appliances during the first part of the twentieth century. The invention of the transistor — that did the same job more reliably in most applications — enabled the miniaturisation of electronics. This miniaturisation has enabled the invention of the integrated circuit used in portable radios, CD players, computers and digital phones.

PHYSICS IN FOCUS

Band structure in solids

You may previously have encountered the shell structure of electrons in atoms, and you may have related this to the spectra emitted when atoms are excited by electrical discharge or heating. (Sprinkle some sodium chloride into a Bunsen flame to observe the bright orange spectrum of sodium.)

These spectral lines are produced by electrons being excited into a higher energy shell and jumping back to a lower energy shell, emitting light of a particular frequency as they do so. All atoms of an element have the same electron shells or energy levels when they exist as individual atoms, but this situation changes when they are present in solids.

In 1925, Wolfgang Pauli proposed what became known as the Pauli Exclusion Principle. It can be stated simply that no two electrons can simultaneously occupy the same energy state. (This is important in the electron structure of individual atoms. All the electrons cannot collapse into the lowest energy shell in an atom.)

In a gas, there is no problem with two well-separated atoms having electrons in precisely the same energy state, but this does become a problem as the atoms are brought closer together and the electrons from different atoms begin to interact with each other. This interaction results in a slight change in energy of the levels so that no two electrons have identical energy. As more and more atoms are pushed closer together, this results in what were precise energy levels in the individual atoms being spread into energy bands in the solid.

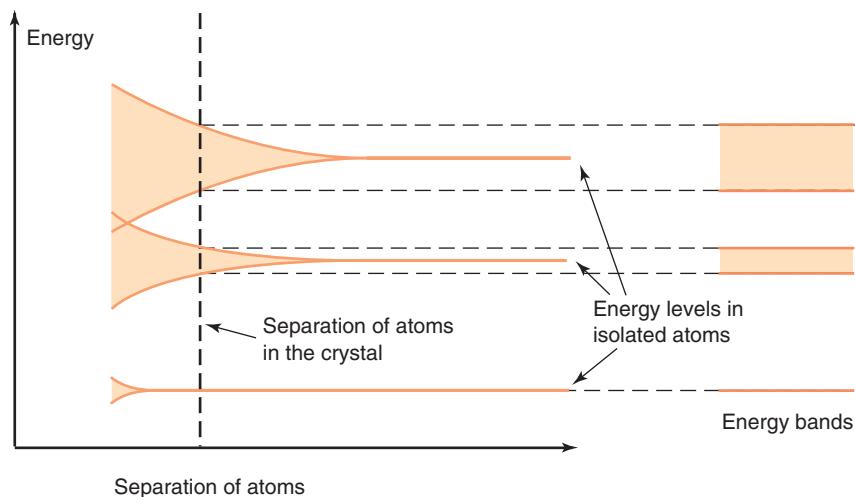


Figure 12.2 The band structure of different types of solids, semiconductors in particular, is important in the development of the solid state devices studied in this chapter.

12.1 CONDUCTORS, INSULATORS AND SEMICONDUCTORS

Different materials vary greatly in their ability to conduct electricity. Their conduction strength depends on the ease with which electrons are able to move through the crystal lattice. In materials that are good insulators, the atoms in the lattice are held by strong covalent bonds in which electron pairs are shared between atoms. This sharing means that

electrons are held tightly and are not available to conduct electricity through the lattice.

Metal lattices, on the other hand, consist of an orderly array of positive metal ions (see figure 12.3). To maintain stability, valence electrons are ‘delocalised’, or free to move like a cloud of negative charges throughout the lattice. These electrons can conduct electricity through the lattice.

Delocalised electrons in the metal lattice move randomly between atoms. Under the influence of an electric field, the random motion of the electrons decreases and begins to have a net motion in a direction opposite to the electric field. This net motion produces the electric current.

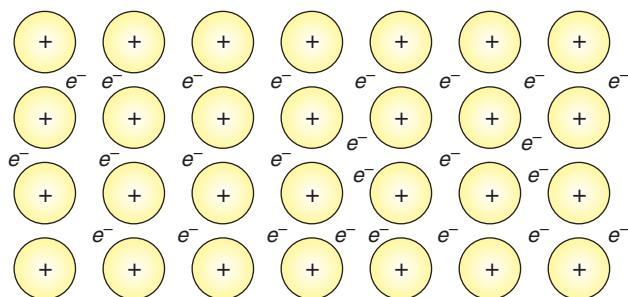


Figure 12.3 In a metal, the positive ions of the lattice are surrounded by delocalised electrons.

The de Broglie model of electron orbitals around the nucleus of an atom (see pages 215–216) explained why electrons in an atom can have only certain, well-defined energies. For any particular element, the highest energy level available for electrons to occupy may, or may not, be completely filled. Elements that do not have their outermost energy level or shell filled will try to get it filled by bonding with other elements. They may do this by:

- gaining electrons from other atoms of an element, forming ionic bonds
- giving electrons to other atoms, forming ionic bonds
- sharing electrons with other atoms, such as in covalent bonds.

The aim of this giving of electrons in ionic bonds and the sharing of electrons in covalent bonds, is to fill electron shells.

When atoms of any type of substance, including insulators, **semiconductors** and conductors, are very close together (such as in a solid) their highest electron energy levels overlap in a continuous fashion. Regions within these highest energy levels are called energy bands. The highest energy band occupied by electrons at absolute zero is called the **valence band**. In a conductor, the valence band and the conduction band overlap (see figure 12.4(a) below). Electrons from the valence band are able to move freely because the band is only partially filled.

In an insulator, electrons completely fill the valence band, and the gap between it and the conduction band is large (see figure 12.4(b) below). The electrons cannot move under the influence of an electric field unless they are given sufficient energy to cross the large energy gap to the conduction band.

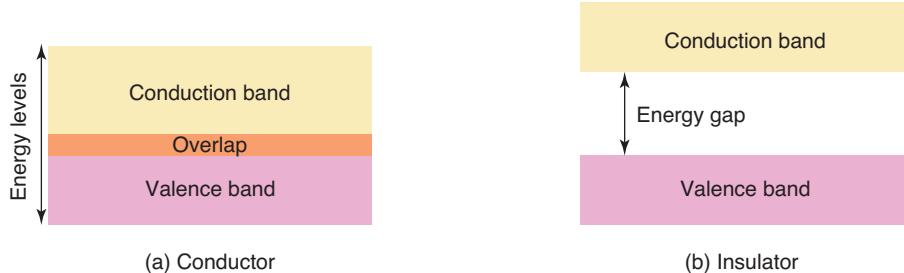
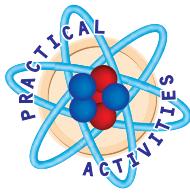


Figure 12.4 Energy bands for (a) a conductor and (b) an insulator



12.1

Band structures

In an ionic crystal such as sodium chloride, the bond between the sodium and the chlorine atom has been accomplished by the sodium atom giving away an electron in its outermost, unfilled electron orbital level to the chlorine atom. This enables the chlorine to fill its outermost electron orbital level without overlapping of bands. Both the positively charged sodium (that has lost an electron) and the negatively charged chlorine (that has gained the electron from the sodium atom) are ions. They have filled outer electron orbitals corresponding to completely filled bands.

Solid sodium chloride therefore behaves as an insulator. The sodium and chlorine ions have no surplus electrons in their outer electron orbital energy levels to transfer from atom to atom. The application of an electric field cannot cause the net movement of a current of electrons. However, it is not impossible for an electric current to flow in an insulator. If the applied electric field is sufficiently large, even an insulator can conduct.

PHYSICS FACT

de Broglie's wave model

The concept of the wave–particle duality of light states that light can act both as a wave and as a particle at different times. If, for example, we carry out an experiment to study the interference of light, we are observing light behaving as a wave.

Similarly, by experimenting with the photoelectric effect, we are observing light behaving as a particle.

French physicist Louis de Broglie (1892–1987) proposed that electrons should also demonstrate a wave-like nature.

Working with Einstein's theory and Planck's quantum theory, he derived an expression relating momentum and wavelength.

He combined Einstein's equation, $E = mc^2$, and Planck's equation, $E = hf$, to show that

$$mc^2 = hf.$$

Using momentum, $p = mv$, and replacing v with c (since c = the velocity of light), he obtained

$$pv = hf.$$

After substituting $\lambda = \frac{v}{f}$, $\lambda = \frac{h}{p} = \frac{h}{mv}$

is obtained

where

m = mass (kg)

v = velocity (m s^{-1})

c = speed of light in a vacuum

p = momentum = mv (N s)

h = Planck's constant = $6.626 \times 10^{-34} \text{ J}$

λ = wavelength (m)

f = frequency (Hz).

In about 1923, de Broglie put forward the idea that, just as light could be thought of as having particle characteristics, electrons could act as a wave. Describing his approach many years later,

he said that as he observed the patterns formed by standing waves in a string he wondered what would happen if the string was bent into a circle. Stable patterns would form when multiples of the wavelength corresponded to the circumference (or length) of the string. Extending this to the circumference of the Bohr orbit in an atom, and using the wavelength of an electron, only certain orbits would be stable — exactly those for which the circumference was equal to a multiple of the wavelength. Here was an opportunity to explain the assumptions made by Bohr. The assumptions were:

- electrons could occupy stable, non-radiating orbits. Electrons moving in circular motion undergo acceleration, and will radiate electromagnetic energy. Losing energy would drive the electron into increasingly smaller orbits, finally collapsing into the nucleus. Since we know that atoms are stable, the energy levels that the electrons occupy must also be stable and cannot radiate energy. These are the 'non-radiating' states referred to by Bohr.
- radiation emission and absorption by atoms can only occur in quantised amounts. Electrons can only move between these discrete levels and must absorb or emit only the amount of energy needed to move between energy levels.

The cornerstone of de Broglie's idea was that the electron orbiting the atom must have a standing-wave pattern of vibration so that its orbit does not destructively interfere with itself. Since the orbital level represents an energy level, only electron energy levels where the electron orbits the nucleus with a standing wave

(continued)

pattern consisting of a ‘whole’ number of wavelengths is possible. Intermediate electron energy levels can’t be stable as they would produce an interfering wave character in the orbiting electron. The electron can absorb energy and move to a higher standing-wave energy level but this energy absorption must be of a specific amount; that is, a quantum. The same electron can move to a lower energy orbital (that produces a standing-wave pattern of orbit) by emitting radiation energy as a photon. This photon must

have an energy precisely equal to the energy difference between the orbital energy levels. Again this energy release must be of a specific size — a quantum. De Broglie’s hypothesis is discussed further in the option topic ‘From Quanta to Quarks’ (see pages 444–447).

De Broglie’s wave model of electrons allowed electrons to orbit the nucleus only when the circumference of the circular orbit was a whole number of wavelengths (see figure 12.5).

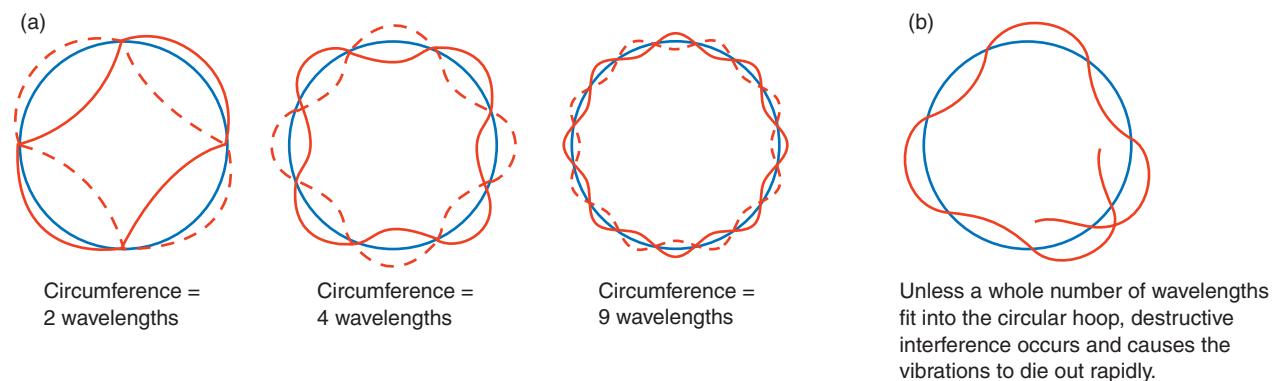


Figure 12.5 A model of the atom showing an electron as a standing wave

12.2 BAND STRUCTURES IN SEMICONDUCTORS

In a semiconductor, the gap between the valence and the conduction bands is smaller than that in an insulator (see figure 12.5).

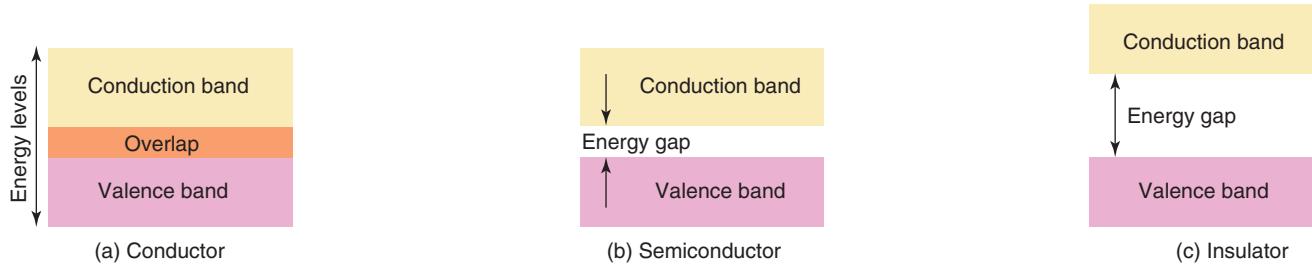


Figure 12.6 Energy bands in (a) a conductor (b) semiconductor and (c) an insulator

Table 12.1 Differences between the structure of conductors and non-conductors

	TYPE OF MATERIAL		
	INSULATOR (NON-CONDUCTOR)	SEMI-CONDUCTOR	CONDUCTOR
Valence band	Completely filled	Almost filled	Partly filled
Conduction band	Well separated	Just separated	Overlapping
Energy gap	Large	Small	Very small/ non-existent

In many materials, including metals, resistivity increases with temperature. In some materials, such as the semiconductors germanium and silicon, resistivity decreases markedly with increasing temperature (see figure 12.7). The increased thermal energy causes some electrons to move to levels in the conduction band from the valence band. Once there, they are free to move under the influence of an electric field.

At absolute zero, all of the electrons in a semiconductor occupy the valence band and the material acts as an insulator. As the temperature of the semiconductor material increases, thermal energy allows some electrons to cross the gap into the conduction band. This leaves the valence band unfilled. This means that holes have been created in the valence band where the electrons have left. These holes actually act as a positive flow of current moving in the opposite direction to the electron current flow. (Note that electron current refers to the movement of the electrons and is in the opposite direction to conventional current.) Thus, conduction is possible in both the conduction band as a flow of electrons and in the valence band as a flow of positive holes. The hole current flows towards the negative potential while the electron current flows towards the positive potential. The speed of the electron current is, however, much greater. This is because the hole current must move from single atom to single atom whereas the electron current can simply flow through the overlapping conduction bands of adjacent atoms.

A **dopant** is a tiny amount of an impurity that is placed in an otherwise pure crystal lattice to alter its electrical properties.

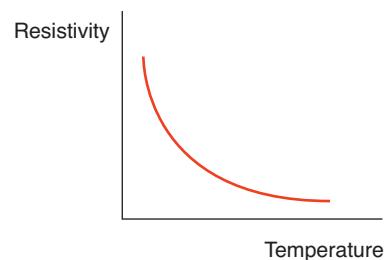


Figure 12.7 Resistivity as a function of temperature for a pure silicon semiconductor

Sometimes an impurity atom (**dopant**), of a different type to the atom making up the main crystal lattice of a semiconductor material is present in a semiconductor crystal lattice. If that dopant atom has a different number of valence electrons from the atom of the semiconductor it replaces, extra energy levels can be formed within the energy gap between the valence and conduction bands. This means it is easier for these materials to conduct because the energy difference between the valence and conduction bands for such dopant atoms is less. The number of dopant atoms needed to create a difference in the ability of a semiconductor to conduct is very small. It will occur with the replacement of a semiconductor atom by only one in every 200 000 atoms.

There are two main types of semiconductor materials: intrinsic and extrinsic.

- *Intrinsic* — the semiconducting properties of the material occur naturally. No doping of the crystal lattice is necessary to enable the material to act as a semiconductor. Examples include the elements silicon and germanium.
- *Extrinsic* — the semiconducting properties of the material are manufactured to behave in the required manner. Generally this means that the material is a naturally occurring semiconductor that has its semiconducting properties modified by the addition of dopant atoms. These are generally silicon or germanium with small impurity levels of dopants, such as phosphorous or boron.

Nearly all of the semiconductors used in modern electronics are extrinsic and based on silicon.

Table 12.2 Some comparisons of resistivity between conductors and semiconductors

MATERIAL	TYPE	RESISTIVITY (Ωm)
Aluminium	Metallic conductor	2.5×10^{-8}
Copper		1.6×10^{-8}
Iron		9.0×10^{-8}
Silver		1.5×10^{-8}
Constantan (copper, nickel) Nichrome (Ni, Cr, Fe)	Alloy metal conductor	4.9×10^{-7} 1.1×10^{-6}
Germanium Silicon	Semi-metal semiconductor	0.9 2000
Glass Mica Polystyrene Wax	Non-metallic insulator	From 10^{10} to 10^{14} From 10^{11} to 10^{15} From 10^{15} to 10^{19} From 10^{12} to 10^{17}

Making a semiconductor

The most widely used semiconductor materials are made from crystals of elements from Group 4 of the periodic table. These elements have four electrons in their valence band. They fill the valence band to eight electrons by sharing an electron with each of four adjacent atoms. Each of these four atoms also contributes a single electron — forming a pair of electrons that is a bond between the atoms. In turn, each of the four atoms bonded to the first atom share a single electron with four adjacent atoms. In this way, each of the atoms has its own four valence band electrons and shares four single electrons from four adjacent atoms. The atom then appears to have eight electrons in its valence band and the band is filled. This sharing of electrons between atoms to form a bond between the atoms is known as **covalent bonding**.

Two Group 4 elements, silicon and germanium, were predicted to be ideal for the production of electronic components because of their semiconducting properties.

Silicon

The conducting properties of silicon can be related to its crystal structure. Silicon crystal forms the so-called diamond lattice where each atom has four nearest neighbours at the vertices of a tetrahedron (see figure 12.8). The tetrahedron consists of a silicon atom at the centre with the four other silicon atoms bonded to it and forming a triangular prism about it.

This fourfold tetrahedral coordination uses the four outer (valence) electrons of each silicon atom. According to the quantum theory, the energy of each electron in the crystal must lie within well-defined bands. The next higher band above the valence band, where the outer four electrons exist, is the conduction band. The conduction band is separated from the valence band by an energy gap. Heating the semiconducting material enables some electrons to gain enough energy to jump that gap from the valence band to the conduction band. This means one bond of the tetrahedron is no longer complete. This incomplete bond is a hole.

A **covalent bond** is a strong chemical bond formed between atoms by the sharing of electrons in the valence band.

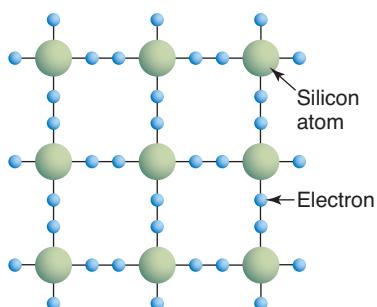


Figure 12.8 The lattice structure of silicon

Germanium and zone refining

Germanium was the first Group 4 element that could be sufficiently purified to behave as a semiconductor. Germanium as an element is

relatively rare, making up less than 1.5 parts per million of the Earth's crust. It is never found in an uncombined form in nature, existing only as a compound.

Early diodes and transistors were made from germanium because suitable industrial techniques were developed to purify the germanium to the ultrapure level required for semiconductors during World War II.

Germanium has one major problem when used in electronic components: it becomes a relatively good conductor when it gets too hot. The conductivity level means that hot germanium electronic components allow too much electric current to pass through them. This can damage the electronic equipment and cause it to fail to perform the task for which it was designed. The problem is that the resistance to electric current flow that makes the semiconductor useful in electronic components also generates heat.

Advantages of silicon

Silicon was the other element with semiconducting properties that was predicted to be ideal for the production of electronic components. Unlike germanium, silicon is very common in the Earth's crust. Like germanium, silicon never appears as a free element in nature. Silicon is always combined into chemical compounds so it has to be purified before it can be used in the production of semiconductors. Almost every grain of sand you see is made of silicon dioxide, so silicon as a raw material is far more plentiful than the rare germanium.

The problem with using silicon in electronic components is that it is more difficult to purify. However, silicon makes the most useful semiconductors for electronics. It is affected less by higher temperatures in terms of maintaining its performance level.

The first silicon transistors were made in 1957 by Gordon Teale working for Texas Instruments. After the production of those first silicon transistors, the germanium transistors were largely phased out of production, except for specialised applications. From the 1960s onwards, silicon became the material of choice for making solid state devices. It is much more abundant than germanium and retains its semiconducting properties at higher temperatures.

12.3 DOPING AND BAND STRUCTURE

A pure semiconductor (called an *intrinsic semiconductor*) has the right number of electrons to fill the valence band. Semiconductors can conduct electricity only if electrons are introduced into the conduction band, or are removed from the valence band to create holes. Electrons are the *negative charge carriers* in the conduction band, and holes in the valence band act as *positive charge carriers*.

As we have seen, the process of doping is one method of enhancing the conductivity of a semiconductor. A tiny amount of an impurity atom is introduced into the semiconductor crystal lattice to alloy with the material. This process produces designer semiconductors that are said to be extrinsic semiconductors.

Extrinsic semiconductors

There are two types of extrinsic semiconductors: n-type semiconductors and p-type semiconductors.

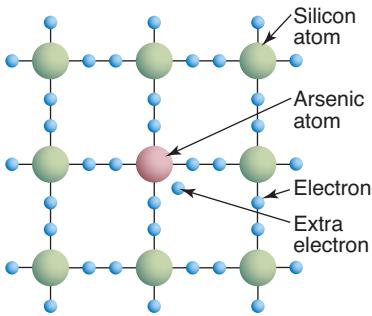


Figure 12.9 Lattice structure of silicon doped with arsenic

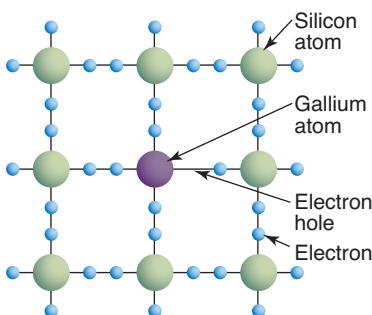


Figure 12.10 Lattice structure of silicon doped with gallium

N-type semiconductors are formed when a Group 5 impurity atom (such as phosphorous or arsenic) is substituted into a silicon crystal lattice, replacing an atom of silicon (see figure 12.9). Group 5 atoms have five electrons in the valence band whereas silicon and germanium, as Group 4 atoms, have only four electrons in the valence band. When this occurs, four of the five outermost electrons from the substituting doping atom fill the valence band just like electrons from a silicon atom would. The one extra electron is promoted to the conduction band. These impurity atoms are in this case called donor atoms. The extra electrons from the donor atoms in the conduction band are mobile. Since they are electrons, and hence carry a negative electrical charge, a semiconductor doped in such a way so as to produce an excess of negative charge carriers is called an n-type semiconductor.

P-type semiconductors are formed in a similar fashion to n-type semiconductors. A Group 3 atom (such as boron or gallium) is substituted into the crystal lattice in place of a Group 4 atom of silicon or germanium (see figure 12.10). The Group 3 atom has only three electrons in the valence band. This means that when such an atom replaces a Group 4 atom, there is one electron short in the tetrahedral structure. This means that a hole has effectively been incorporated into the crystal lattice without the need to elevate an electron from the valence band to the conduction band. The holes act as positive charge carriers that are mobile and carry current.

When the doping impurity has only three electrons, for example, when indium is added to germanium, there is one site unfilled by electrons. This hole actually acts as a mobile positive charge carrier when an electric field is applied. The movement of the electron is shown in figure 12.11. Under the effect of an electric field, electrons move towards the positive terminal of the source and move into any hole. Since there are no free electrons available, the deficient indium atoms near the positive terminal attract electrons from their neighbours, disrupting covalent bonds. This creates new holes in the adjacent atoms. As electrons continue to move towards the positive terminal the holes move in the opposite direction.

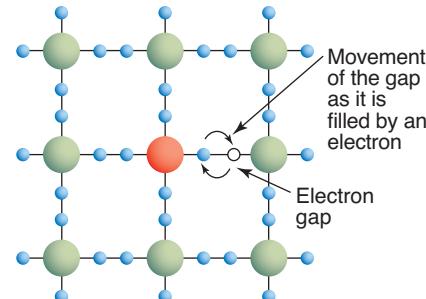


Figure 12.11 Lattice structure of silicon doped with indium showing the electron movement

12.4

THERMIONIC DEVICES

A **valve** is a thermionic device in which two or more electrodes are enclosed in a glass vacuum tube. The name comes from the rectifying property of the device; that is, the current flows in only one direction.

A **diode** contains only two electrodes.

In electrical appliances there is often a need to control the direction of current flow, convert AC into DC, switch current flow on or off or amplify a current. Devices within the appliance control this. Prior to the invention of solid state devices, such as the solid state **diode** or the transistor, thermionic or **valve** devices accomplished these tasks.

Thermionic devices utilise heated filaments and terminals set in glass vacuum tubes, such as the older radio valves. The filament in the vacuum valve is heated by an electric current, causing it to liberate electrons and act as a cathode. These electrons are then accelerated by a high potential difference towards an anode. The simplest example of such a valve is the diode (see figure 12.12).

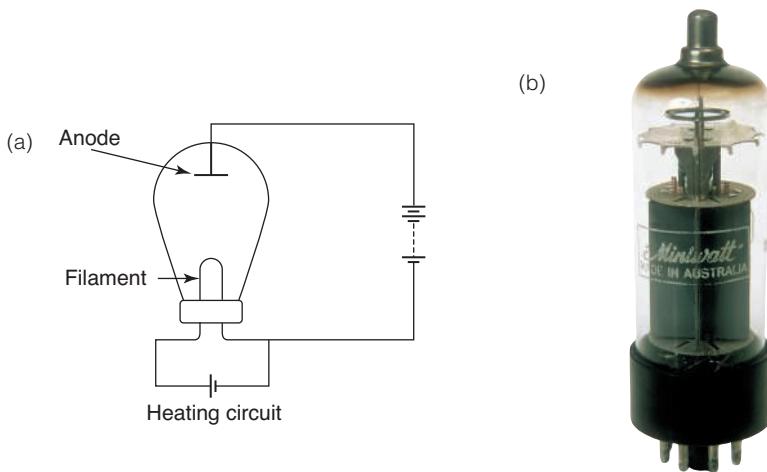


Figure 12.12 (a) Example of a thermionic diode. Electrons are emitted from a heated filament and accelerated to the anode. (b) A photograph a thermionic valve, of the type used in the mid-twentieth century

A diode has two elements inside the glass vacuum tube: a cathode and a plate (or anode). The cathode may be heated directly (with current flowing through the cathode) or indirectly (with a separate filament). With the negative terminal attached to the cathode, electrons will flow through the diode — creating an electric current. If the battery is connected in reverse, no current will flow. Such ‘unidirectional conduction’ makes a diode suitable as an ‘electronic switch’ and for converting AC current into DC current. This is called rectification.

Thomas Edison (1847–1931) (of electric light fame) observed that an electric current flowed between the cathode and the positively charged plate in the first diode. Lee de Forest (1873–1961) added a third electrode to the vacuum diode and demonstrated that the valve now acted as a current amplifier. This third electrode was called the ‘grid’ and the valve was called the triode.

The grid circuit can be adjusted separately from the anode (or plate) circuit. It is closer to the cathode than the anode. A voltage placed on the grid has a much larger effect on the electric field within the valve. The grid can therefore be used to control the anode current (see figure 12.13(a) on the following page).

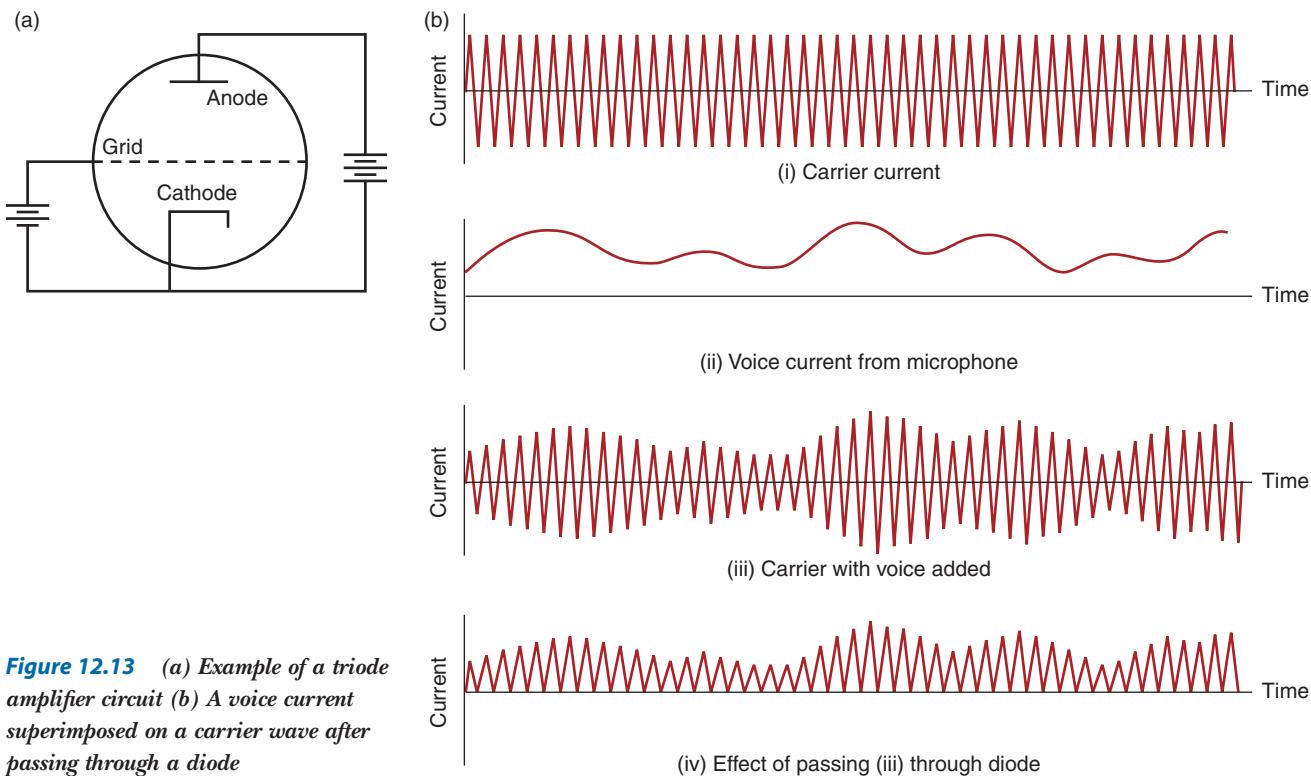
When an alternating voltage, or signal, is applied to the grid, the electron current is an amplified replica of the signal voltage. Because a high voltage is applied between the cathode and the anode, small variations in the grid current produce amplified signals in the anode circuit.

PHYSICS FACT

Lee de Forest’s invention of the triode valve in 1906 began the electronic age. He called the triode the ‘audion’. He did not realise the potential of his device until 1912 when he discovered that the triode could be used to amplify sound and electromagnetic waves.

The invention of this device made possible the whole range of electronics we accept as normal today: radio, television, radar, computers and long-distance telephony.

This was important in the early days of radio reception. A radio wave consists of a ‘carrier wave’ with the signal superimposed. The wave hitting an antenna generates an alternating electrical current. The carrier wave is removed, leaving a small AC current signal. This small AC signal current can be applied to the grid, with the amplified signal passing to a loudspeaker to produce sound.



12.5

SOLID STATE DEVICES

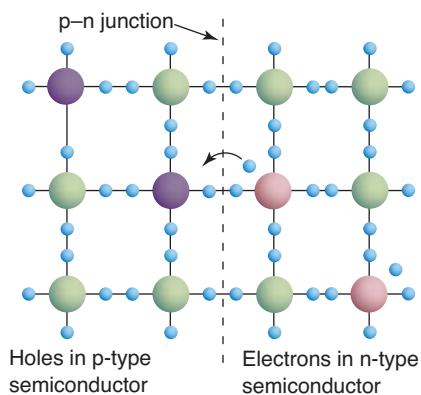


Figure 12.14 The holes and electrons in a p-n junction

Modern appliances utilise solid state devices, such as transistors and integrated circuits. Solid state devices are made from semiconductor materials. For example, a junction or interface between a p-type semiconductor and an n-type semiconductor acts as a diode (see figure 12.14). This combination allows current to flow in only one direction. Electrons move across the junction from the n-type semiconductor (see figure 12.15), neutralising available holes in the p-type semiconductor material (see figure 12.16). This zone adjacent to the interface is called the ‘depletion zone’.

The depletion zone exerts a ‘force’ that resists the movement of any more electrons into the region. This effectively means that the conventional electrical current flow is confined to one direction only; that is, from the positive terminal of the battery, into the p-type semiconductor material and then out of the n-type semiconductor material to the negative terminal of the battery. A diode connected in such a way is said to have a forward bias applied to it.

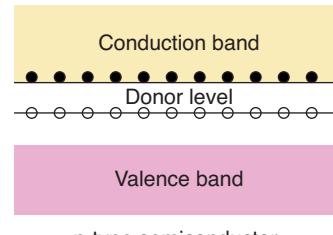


Figure 12.15 Energy band changes in a doped n-type semiconductor. The level is called ‘donor’ because electrons are donated to the conduction band.

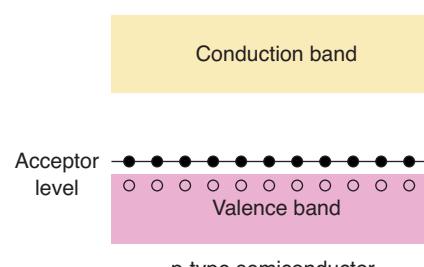


Figure 12.16 Energy band changes in a doped p-type semiconductor. The level is called ‘acceptor’ because electrons are accepted from the conduction band.

eBookplus

Weblink:
Semiconductors:
p- and n-type

If the battery is connected the other way — that is, with the p-type material connected to the negative terminal of the battery and the n-type material connected to the positive terminal — no current, or only very minute currents, can flow through the diode. The diode in such a case is acting as a large resistor and is said to be reverse biased (see figure 12.17).

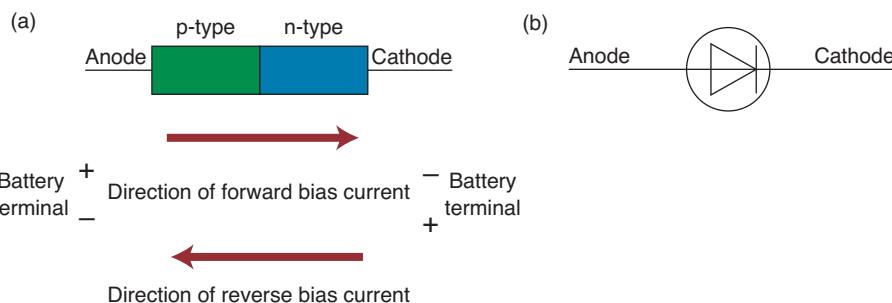


Figure 12.17 (a) A junction diode.

(b) Its circuit symbol. The arrow on a diode indicates the direction of conventional current flow possible through the diode. Note that it is unidirectional.

PHYSICS IN FOCUS

The p-n junction

A small region at the junction of a p-type semiconductor and an n-type semiconductor is the key to the operation of diodes, transistors and photovoltaic cells. This region is called the depletion region or depletion zone, and it is formed by the diffusion of charge carriers from one type of material to the other.

In n-type material, the dominant charge carriers are the electrons from the donor atoms. The n-type material will contain some positively charged donor ions and some electrons contributed by the donors. There will be a relatively small number of electrons and positive holes due to the material's intrinsic semiconductor properties.

In p-type material, the dominant charge carriers are the positive holes from the acceptor atoms. The p-type material will contain some negatively charged acceptor ions and the corresponding positive holes. Of course, there will also be a relatively small number of electrons and positive holes due to the material's intrinsic semiconductor properties.

The mobile charge carriers from one type of material will diffuse into the other type of material. The thickness of the depletion zone is much less than the distance that the charge carriers can diffuse through the material.

As electrons from the n-type material diffuse into the p-type material, they will introduce negative charge to the p-type material and leave the n-type material positively charged.

Similarly, as holes from the p-type material diffuse into the n-type material, they will introduce positive charge to the n-type material and leave the p-type material negatively charged.

The diffusion of electrons from n-type to p-type and the diffusion of holes from p-type to n-type both contribute to the build-up of positive charge on the n-type material and negative charge on the p-type material. Hence an electric field is established across the depletion zone. As the diffusion of charge carriers increases, this electric field also increases, until it opposes further diffusion of the charge carriers across the junction.

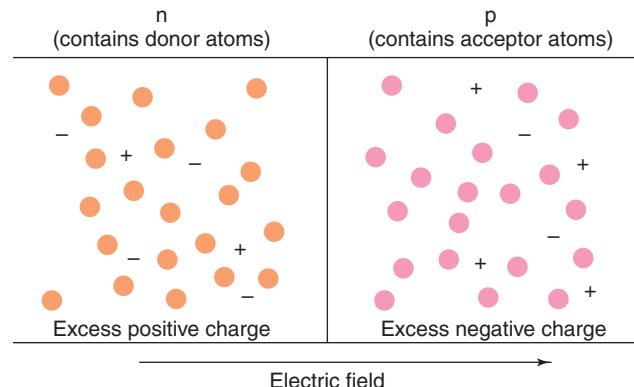


Figure 12.18 The electric field at a p-n junction. The orange dots represent donor ions. These are the dopant atoms that have lost their 'extra' electron and become positively charged ions. The electrons have diffused across the boundary into the p-type material. The red dots represent acceptor ions. These are the dopant atoms that have accepted an electron (which has diffused across from the n-type material). The electrons have effectively filled the positive holes in the acceptor atoms, and they have become negatively charged. The + signs represent a very small number of positive holes that still exist, and the - signs represent the very small number of conduction band electrons that are still present.

(continued)

As we can see in figure 12.18, the depletion zone at the junction contains donor and acceptor ions and very few electrons or positive holes, so it is almost devoid of mobile charge carriers.

(Even in equilibrium, there is a small flow of electrons from n-type to p-type through the depletion zone. These electrons undergo recombination with holes in the p-type material. There is an equal flow of electrons from p-type to n-type. These electrons are produced by thermal generation in the p-type material. The recombination current and thermal generation current must be equal and opposite unless either of the two is altered by the application of a potential difference across the junction.)

The presence of the electric field can be used to explain both the diode nature of the p-n junction and also how the p-n junction acts as a photovoltaic cell.

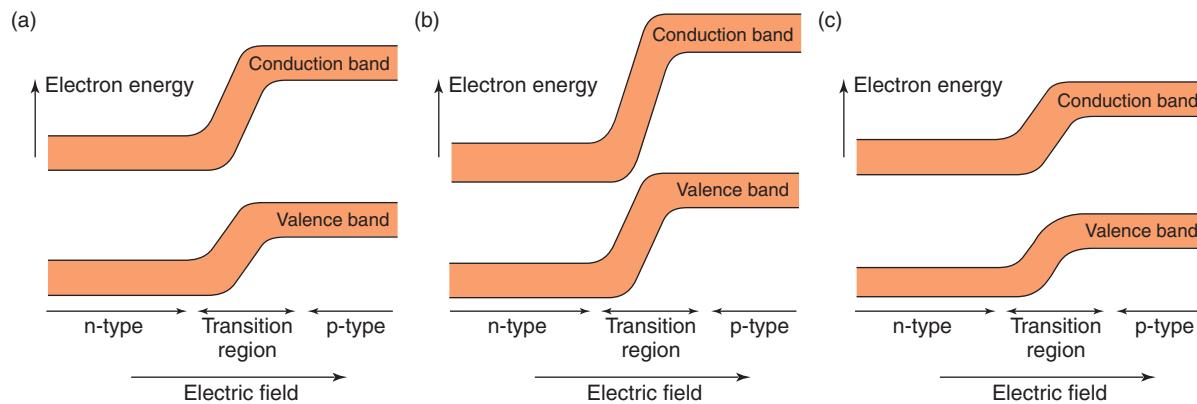


Figure 12.19 (a) The energies of the valence and conduction bands change across a p-n junction when the junction is in thermal equilibrium and there is no electric potential difference applied across the junction. (b) The change in the band energies when a reverse bias voltage is applied. Neither electrons nor holes will be able to flow across the junction. (c) The band energies when a forward bias voltage is applied. Large numbers of electrons will be able to flow from the n-type material to the p-type material with a correspondingly large flow of holes from p-type to n-type.

Energy bands in a p-n junction diode

The electric field is responsible for a change in energy of the valence and conduction bands across the junction.

When a positive potential is applied to the n-type material and a negative potential is applied to the p-type material, as shown in figure 12.19(b), the energy difference is increased and almost no electrons are able to pass from the n-type to the p-type. (It is a similar case for holes, which are unable to travel from p-type to n-type.) This is known as reverse bias.

When a negative potential is applied to the n-type material and a positive potential is applied to the p-type material, as shown in figure 12.19(c), the energy difference is decreased. Now large numbers of electrons will be able to flow from n-type to p-type, and there will be a correspondingly large flow of holes from p-type to n-type. This is known as forward bias.

12.6 THERMIONIC VERSUS SOLID STATE DEVICES

Thermionic devices, such as radio valves and amplifiers, cannot match the efficiency, cost or reliability of solid state devices. Appliances utilising valves had a number of disadvantages compared with devices utilising solid state devices.

- Thermionic devices and appliances were bulky. Even radios advertised as ‘portable’ could be lifted only with difficulty by a child. So much power was required that batteries had to be large or numerous. A twelve, D-sized battery radio that was considered portable was not unusual.
- A large amount of heat was developed by the valves. This required engineering solutions to protect surrounding electronics.

- Valves are fragile. Like a light globe, there is a seal between the evacuated glass tube and the Bakelite (an early plastic) base through which the leads pass from internal connections to the pins in the base. This meant radios or tape recorders could not be carried as easily, or treated as roughly as modern tape recorders or portable CD players.
- The cathode was coated with a metal that released large numbers of electrons. The heat that was produced slowly boiled off the metal coating and the coating reacted with the traces of gas present in the tube.
- Valves had a relatively short lifetime. Technicians would start testing a malfunctioning appliance by testing the valves, even replacing them to see if the fault disappeared. Solid state devices are now among the least likely faults. One of the original uses for the valve was in telephone exchanges. As telephone networks began to expand rapidly in the late 1940s and 1950s, the unreliability of the valve began to be intolerable.
- Individual sockets and valves were mounted on a metal chassis. Components were connected by insulated wires to other discrete components. There was often movement between the chassis and sockets, leading to broken solder joints. The glass envelopes of the valve were fragile and the seals frequently broke, allowing air into the valve and destroying it.
- High voltages were required to correctly bias the triodes to amplify signals. This is in contrast to a silicon transistor that requires around 0.6 V to do the same job.

12.7

INVENTION OF THE TRANSISTOR

A **transistor** is a tiny switch that changes the size or direction of electric current as a result of very small changes in the voltage across it. Transistors are used in sound amplifiers and in a wide range of electronic devices. Today, a single chip of silicon can hold many microscopic transistors and is called an integrated circuit.



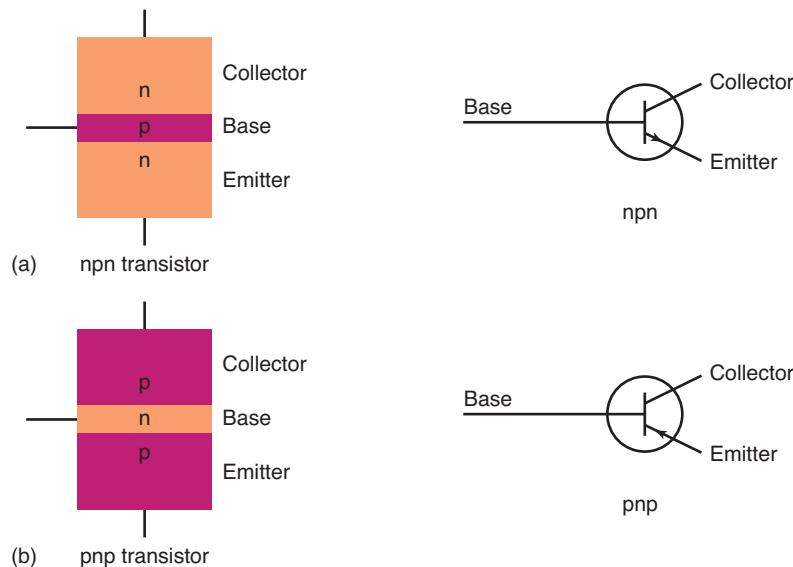
12.2

Invention of the transistor

Figure 12.20 npn and pnp transistors with their symbols

William Shockley is considered the ‘father’ of the **transistor**. However, he was not present on 17 November 1947 when scientists John Bardeen and Walter Brattain, at Bell Laboratories, observed that when electrical contacts were applied to a crystal of germanium, the output power was larger than the input. Shockley saw the potential and, over the next few months, greatly extended the understanding of the physics of semiconductors. (For more detail on the team’s work, see Practical activity 12.2, page 231.)

A transistor is a semiconductor device that can act as a switch or as part of an amplifier. There are two types: npn transistors and pnp transistors (see figure 12.20).



A transistor is simply a combination of two junctions. One consists of a thin layer of n-type material between two sections of p-type material. The other is the reverse, with a thin section of n-type between two sections of p-type material. The three connections are called the emitter, base and collector (see figure 12.21).

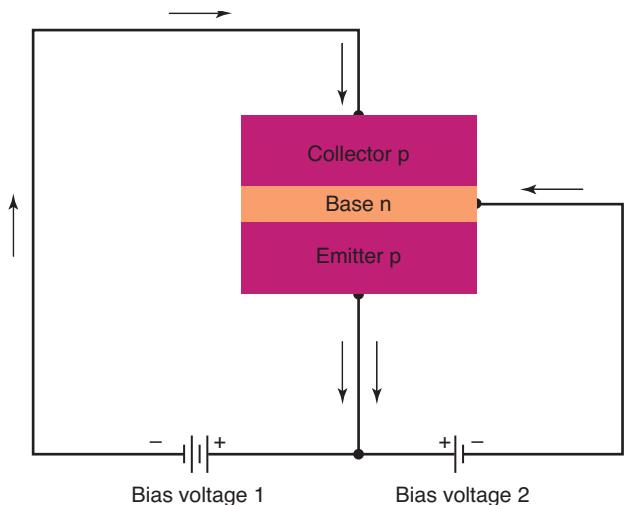


Figure 12.21 A block diagram of a pnp transistor. Arrows indicate electron movement.

Each of these junctions must be correctly biased in order for the transistor to work.

For a pnp transistor, mobile electrons in the n-region initially move away from the junctions towards the positive terminal. The holes in each of the p-regions also move away from the junctions towards the negative terminals. When the emitter is slightly positive, or forward biased, holes move across the junction into the n-region, or base. Note that the junction that is forward biased is known as the emitter. Most of the holes do not recombine with electrons in the base but flow across the second junction into the collector. The input impedance (electrical resistance) between the emitter and the base is low, whereas the output impedance is high. Small changes in the voltage of the base cause large changes in the voltage drop across the collector resistance, making this arrangement an amplifier.

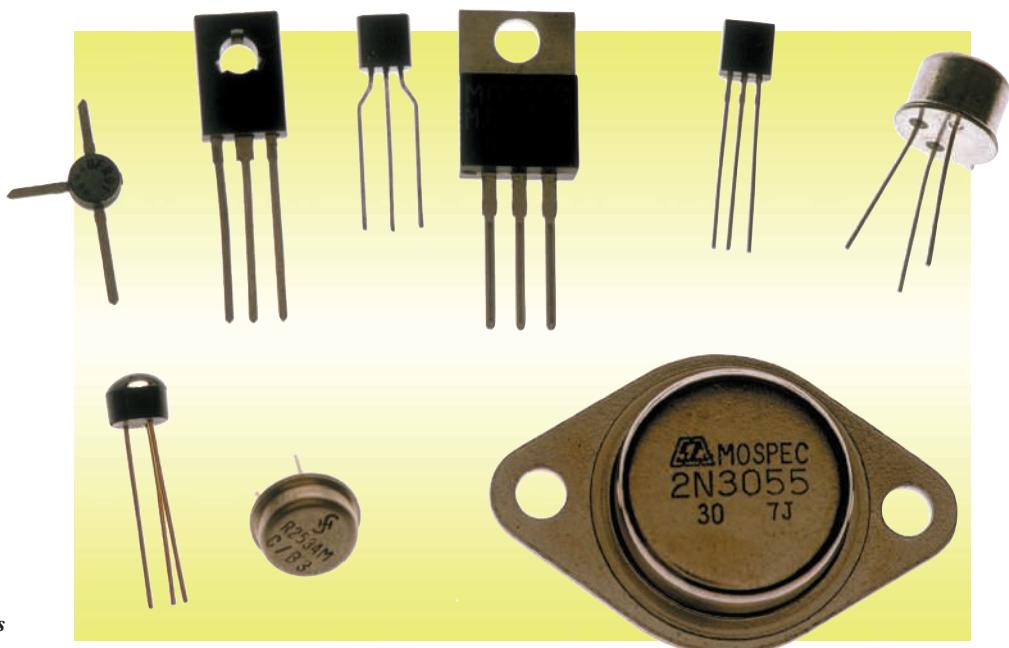


Figure 12.22 Some of the numerous transistor cases and types

12.8 INTEGRATED CIRCUITS

Integrated circuits (ICs) are tiny electronic circuits used to perform a specific electronic function. An IC is formed as a single unit by diffusing impurities into a single crystal of silicon. Each IC is really a complex circuit built up in a three-dimensional manner in layers (see, for example, figure 12.1, page 212). Each vertical section of the circuit can be etched into the silicon by means of light beams.

Large-scale integrated circuits (LSI) have as many as five million circuit elements, such as resistors and transistors, combined on one square of silicon often less than 1.3 cm on a side. Hundreds of these can be arrayed on a single wafer. All of the interconnections between the components are built into the chip. This reduces the problems of variable resistance in wiring and soldering components and improves the speed of signal transmission. These chips are assembled into packages containing all of the external connections to facilitate insertion into printed circuit boards. (A summary of the process is shown in figure 12.24.)

The solid state revolution

The earliest solid state devices were discrete components; that is, a single transistor or diode. That means that each component was a separate item. They were small, used much less power and were more reliable.

The invention of integrated circuits made miniaturisation of electronic circuits possible. Whole amplifiers could be built on a single chip and connected to a circuit board. This meant that there were fewer connections, less wiring and less heat produced than in comparable thermionic device appliances. An added advantage of solid state devices was that the time taken for signals to move around the circuit was much shorter. Computers run much faster and handle incredible amounts of data. Large-scale integration (LSI) has since led to hand-held calculators that have more computing power than the computer onboard the *Voyager* spacecraft.

By 1960, vacuum tubes were being supplanted by transistors, because of their reliability and their cheaper construction costs. Computers changed from a room full of cabinets requiring airconditioning and stable power supplies to table-top models and eventually to hand-held devices.

Discrete transistors soon gave way to integrated circuits (ICs), allowing complete appliances to be built onto a single chip. The first silicon chip was made in 1958 and, by 1964, ICs contained about ten individual components on a chip approximately 3 mm square. In 1970, the same size chip held one thousand components and, importantly, cost no more to produce.

In 1971, Intel produced its first microprocessor. As microcomputers became popular, the demand for mass-produced ICs increased, further reducing the cost. By 1983, over half a million components could fit on a single chip. Now they hold millions of components and microprocessors can be found in most household appliances.

In late 2007 Intel introduced a new line of power-efficient microprocessors codenamed Penryn. They are based on a 45-nanometre process that uses smaller and lower-power transistors. The laptop version was introduced in 2008 and consumes 25 watts of power, a significant reduction from the 35 watts of the older 65-nanometre chips. As well as reducing power usage, the new chips can operate at higher clock speeds, delivering 40% to 60% improvement in video and imaging performance.



Figure 12.23 Thousands of tiny components make up the integrated circuits of an everyday calculator.

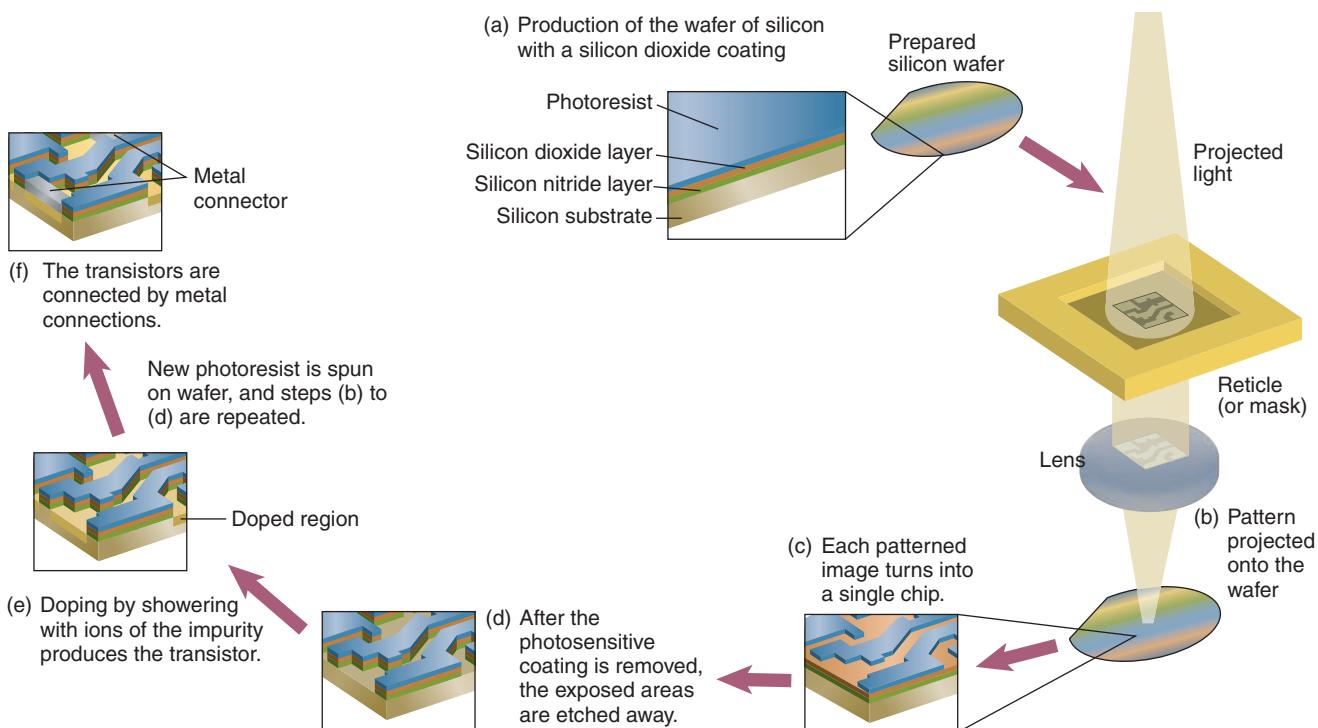
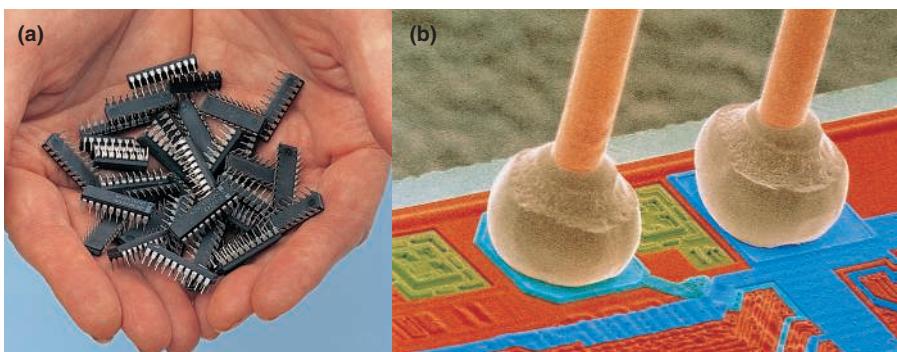


Figure 12.24 Chip fabrication is the result of many steps and many chips are produced on a single wafer of silicon.



Socially, the advent of transistors has made continuing changes to the way we live our lives. This book is being written using integrated circuits. We communicate with friends using mobile phones that use chips. The way we do business, travel, shop and are entertained — almost all our activities rely on silicon microchip technology.

PHYSICS FACT

Applications of semiconductors: photovoltaic cells

One of the most important uses for silicon transistors is in photovoltaic (PV) cells or solar cells. The operation of solar cells is based on the formation of a junction. The natural potential difference of the junction permits current to flow from the p side to the n side, remembering that the definition of conventional current is from the positive terminal through the circuit to the negative terminal.

To make solar cells, discs or wafers of crystalline silicon undergo a number of steps, such as:

- grinding and cleaning
- doping
- metallisation
- anti-reflection coating.

The resulting cell is shown in figure 12.26 with the p-type and n-type materials joined to produce a ‘sandwich’.

When light strikes the cell, some is absorbed within the semiconductor material and the energy of the absorbed light is transferred to the semiconductor. The light photons knock the electrons from the valence band to the conduction band, allowing them to move freely. By placing metal contacts on the top and bottom of the solar cell, the current of moving electrons can be used externally. Common uses of photovoltaic cells include the powering of calculators, telephones, lights and radio warning beacons.

Every PV cell has an electric field which forms where the n-type and p-type silicon are in contact. The free electrons in the n side move into the free holes on the p side. Before this movement of electrons, the p- and n-type semiconductor silicon wafers making up the PV cell are electrically neutral, but now the electrons move to the p side. The holes at the p-n junction fill first, making it difficult for later electrons to move across the junction.

Eventually, a level of equilibrium is reached in terms of the charge distribution about the junction. This equilibrium produces an electric field between the p- and n-type sides of the PV cell. This field acts as a diode, allowing electrons to flow from the p side to the n side of the PV cell, but not the reverse. It acts like an energy hill — electrons can easily go down the hill (to the n side), but cannot reverse up the hill (to the p side). This one-way current can be stimulated to continue if the electrons are collected by a metal plate on the n-type side of the PV cell and fed back to the p-type side of the PV cell by a circuit wire. The current of electrons in the external circuit can do work because of the potential difference that exists between the p- and n-type sides of the PV cell.

The mechanism of current production by a PV cell is as follows. When light, in the form of photons, hits a cell, its energy frees electron-hole pairs. Each photon with more than the minimum energy level will free exactly one electron. As a result, the freeing of the electron produces a

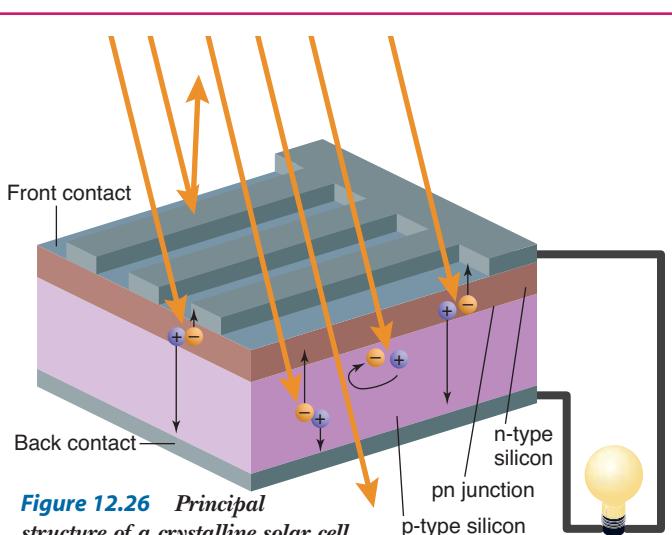


Figure 12.26 Principal structure of a crystalline solar cell

hole. The electric field will then send the electron to the n side, and the hole will tend to migrate to the p side of the PV cell.

If an external circuit is provided (as shown in figure 12.27), electrons will flow through the circuit path to the p side, releasing energy and doing work.

To minimise energy losses from the PV cell, it is covered by a metallic contact grid that shortens the distance that electrons have to travel. The metal conducting grid also covers only a small proportion of the surface of the wafer. In spite of this, some photons are blocked by the metal conducting grid. The grid cannot be too small or its own resistance will increase and be unacceptably high. There is a trade-off between grid size and current generation.

An anti-reflective coating (see figure 12.28) is applied to the top of the PV cells to reduce reflection losses, as every photon that is reflected is a photon that cannot be used for energy. The PV cell is covered with glass for protection. Most PV cells are made up into modules of many cells. Modules must be made by connecting several cells to increase both current (connecting in parallel) and voltage (connecting in series) and adding terminals on the back.

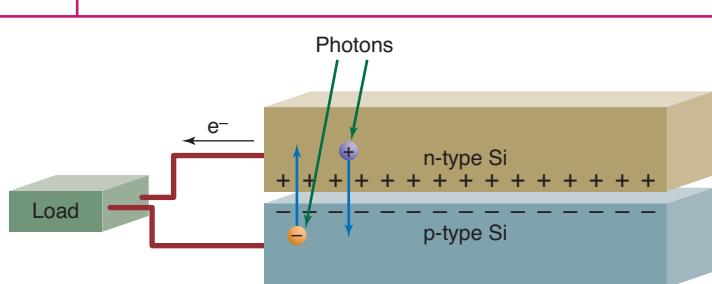


Figure 12.27 Photons hitting electrons, causing them to be released from their structure and move in the diode, thus releasing energy to the load (which may, for example, be a calculator or a light)

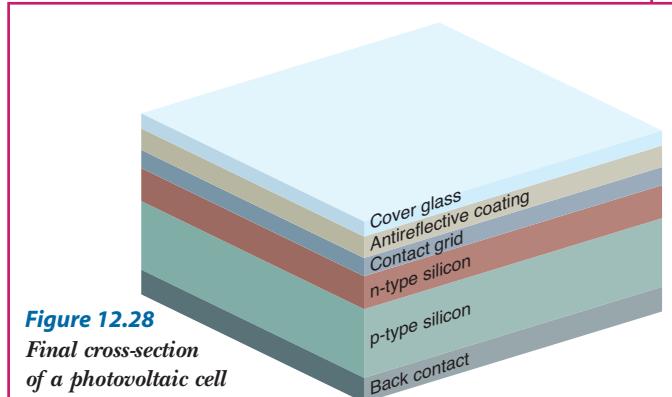


Figure 12.28
Final cross-section of a photovoltaic cell

SUMMARY

- The number of freely moving electrons dictates whether a material will be a conductor, semiconductor or insulator.
- The valence energy band is used for electrons when atoms bond. The conduction band is for electrons that can move from one atom to another and conduct electricity.
- Conducting materials contain a very small energy band gap, a partially filled valence band with gaps or holes and an overlap between the valence band and the conduction band.
- Insulators have strong bonding links between electrons and large energy band gaps between the filled valence band and conduction band of electrons.
- Semiconductors contain weaker bonding links between the electrons and a smaller energy gap between the electron-filled valence band and the conduction band.
- Substituting other elements into the silicon crystal lattice (doping) is used to vary the conductivity of the silicon.
- Intrinsic semiconductors mainly conduct current because of the excitation of a valence electron into the conduction band level.
- Extrinsic semiconductors mainly conduct using either excess electrons or excess holes in their structure.
- n-type extrinsic semiconduction is produced by the addition of an impurity with an extra valence electron above four valence electrons. p-type extrinsic semiconductors are produced by the addition of an impurity with one less valence electron below four valence electrons.
- Diodes and triodes are two examples of solid state electronic devices made from semiconducting material. When large numbers of transistors and other components are combined on a single chip it is called an integrated circuit.
- Diodes are used as rectifiers and transform alternating current into direct current.
- Semiconductors and transistors have a large number of advantages over the large, hot, evacuated tubes called thermionic devices.

They allow the passage of information with very low power, an increase in speed of transmission with a much lower energy use, more components to fit in a confined space, portability and flexibility and are strong enough to be used in many situations.

QUESTIONS

1. Calculate the wavelength of an electron travelling at a speed of $5.00 \times 10^5 \text{ m s}^{-1}$.
2. Explain, using band theory, why it is possible for an electric current to flow in an insulator.
3. Use the band theory to explain why metals are good conductors.
4. What is a hole current and how does a hole current differ from an electron current?
5. Describe how you could model the production and movement of electrons and holes in a semiconductor that has an applied electric field from left to right using a Chinese checker set and marbles.
6. Outline why germanium was used for early transistors instead of the silicon now typically used.
7. Explain why doping with an atom that has five valence electrons makes the doped silicon a much better conductor than undoped silicon.
8. Describe the difference between an n-type semiconductor and a p-type semiconductor.
9. Explain why it was so desirable to develop solid state devices to replace the thermionic devices such as the triode and the diode valve.
10. The transistor can act as an amplifier or as a switching device. The use of the transistor as a switching device led directly to the development of the microchip and microprocessor. What is a transistor?
11. Describe the difference between an n-type transistor and a p-type transistor.
12. Why does a hole current always move in the opposite direction to an electron current operating in the same transistor?



12.1 BAND STRUCTURES

Aim

To investigate a model of the difference between conductors, insulators and semiconductors in terms of band structures.

Apparatus

A computer with internet access

eBook plus

Weblink:
Conductors

Theory

Materials are labelled as conductors, semiconductors or insulators depending on their ability to conduct electrical charge. Conduction occurs when a substance contains electrons with sufficient energy to occupy the conduction band energy level.

Method

1. Log on to the web site given above. You should be viewing the page with the heading ‘Let’s Imagine’.
2. View each page, finishing with the heading ‘Let’s Summarise’.

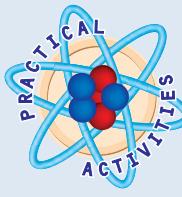
Analysis

Describe the model used for the behaviour of electrons in:

- (a) conductors
- (b) semiconductors
- (c) insulators.

Questions

What is the essential difference in terms of band structure between conductors, semiconductors and insulators?



12.2 INVENTION OF THE TRANSISTOR

Aim

To research the work of the team of Nobel Prize winners who developed the transistor and, eventually, integrated circuits.

Apparatus

A computer with internet access

eBook plus

Weblinks:
The development
of the transistor
William Shockley

Theory

The invention of the transistor is an interesting example of team dynamics. William Shockley considered himself the ‘ideas man’. He would think of a design concept and pass it on to John Bardeen, the theorist, who would try to find a way to make it work. Bardeen then passed the design to Walter Brattain, the experimentalist, who could build anything in a lab.

For two years, Bardeen and Brattain tried unsuccessfully to make Shockley’s idea of a field transistor work. Finally they worked on an idea of their own and, in 1947, invented the world’s first transistor — the point contact transistor. Spurred on by this, Shockley immediately designed the very successful bipolar transistor. The trio won the Nobel Prize for their work, but, sadly, jealousies split up the team.

Another scientist, John Atalla, later developed Shockley’s original idea. Ironically, most of today’s applications, including integrated circuits, use the field effect transistor.

Method

Access the web sites quoted above. Using the biographies and the timeline (under ‘Transistorized!’), work in groups to compile reports on the contributions of each of the scientists involved in the development of the transistor, and present the reports to the class.

CHAPTER 13

SUPER-CONDUCTIVITY



Figure 13.1 A maglev train. This Japanese train operates on mutually repelling magnetic fields and achieves speeds of over 500 km h^{-1} .

Remember

Before beginning this chapter, you should be able to:

- recall that electrical conduction in metals is related to the movement of electrons through the lattice
- recall the principle of superposition.

Key content

At the end of this chapter you should be able to:

- outline the methods used by the Braggs to determine the structure of crystals
- explain that metals possess a crystal lattice structure
- identify that the conducting properties of a metal are related to the large numbers of electrons able to move through the crystal lattice
- discuss how the lattice structure impedes the paths of electrons, resulting in the generation of heat
- identify that resistance in metals is increased by the presence of impurities and scattering of electrons by lattice vibrations
- describe the occurrence in superconductors below their critical temperature of a population of electron pairs unaffected by electrical resistance
- describe how superconductors and magnetic fields have been applied to the development of a maglev train
- discuss the BCS theory
- discuss the advantages and limitations of superconductors and possible applications in electricity transmission.

In 1911, research into the structure and electrical behaviour of materials led to the identification of some materials for which the electrical resistance almost disappeared when the temperature approached absolute zero. This property became known as superconductivity.

An explanation of superconductivity depended on an understanding of the crystalline structure of conductors and the way in which electrons interacted with it. W. Bragg used the diffraction of X-rays from a regular crystal to determine its structure.

13.1 INTERFERENCE

When you have previously studied the behaviour of waves, you probably observed that the amplitude displacements produced by the waves would combine when they passed through each other. This effect is called superposition. The amplitude of the resultant wave at every point was found by adding the displacements of each wave. You should recall that a wave disturbance due to two or more sources can usually be taken as the algebraic sum of the individual disturbances. In the special case where the sources vibrate with the same frequency and with a constant phase, some fascinating **interference** effects occur.

Figure 13.2 shows how waves in water from two slits in a barrier produce a characteristic pattern of interference. This is a property of all waves.

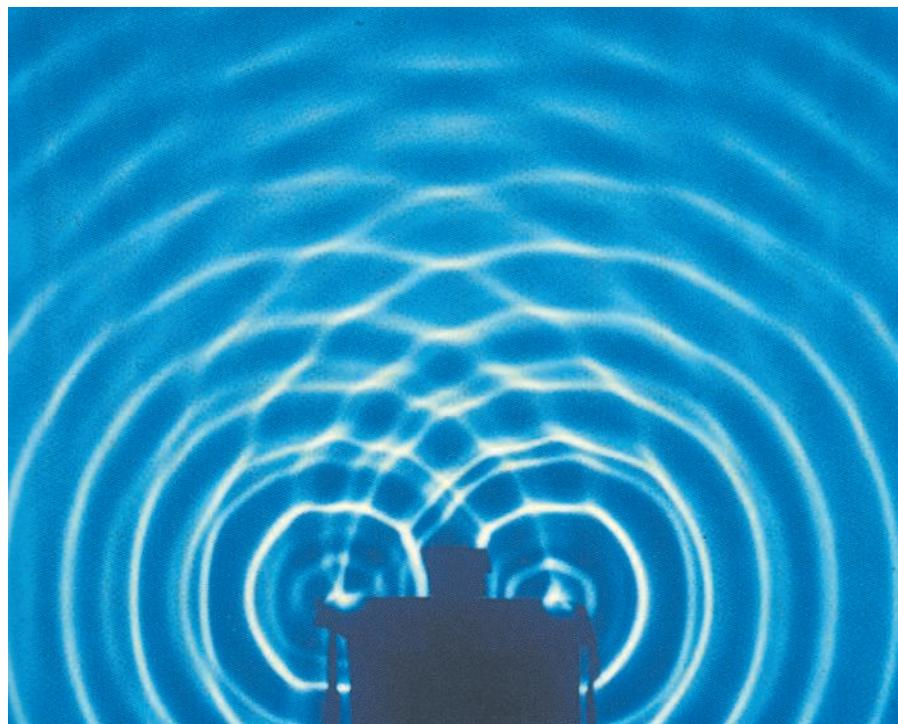


Figure 13.2 Constructive and destructive interference produced by two in-phase sets of circular water waves

Light waves are **coherent** when there is a constant phase difference between them; that is, the peaks line up and the troughs line up.

The two slits act as separate sources of **coherent** circular waves. As they spread out, they interfere with each other. In some directions the waves combine constructively, making waves of larger amplitude. In other directions they combine destructively, so that there is little or no resulting wave amplitude. The wedge-shaped areas of sharp contrast indicate the crests (light) and troughs (dark) of strongly reinforced waves. In some cases the waves arrive at the same point in time and space totally (180°) out of phase. The resulting wave amplitude is zero so there are no contrasting lines, and we see a region with no wave motion.

eBookplus

Interactivity:
Young's experiment
(interference effects
with white light)
int-0051
eLesson:
Young's experiment
(interference effects
with white light)
int-0027

The condition for complete constructive interference is that waves from two sources arrive at the same point with the same phase. Since the waves are continuous, each point one full wavelength apart will have the same phase. The condition for constructive interference may be described in terms of the difference in path length. Points will be in phase so long as the difference in the path travelled by each wave is a whole number of wavelengths. If we let n be any whole number of wavelengths, then we can describe this difference in path lengths as $\Delta D = n\lambda$ where λ = wavelength.

Light waves also demonstrate interference effects. Young's 'double slit' experiment provides a clear example of the geometry, and the relationships necessary for constructive interference. In figure 13.3 the light from a small aperture falls on two small slits.

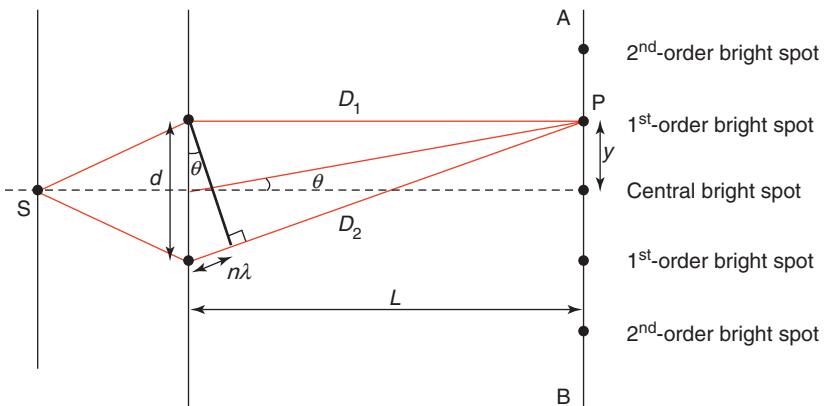


Figure 13.3 Geometrical construction for describing the interference pattern of two point sources (Young's experiment)

The two slits are at a distance, d , apart and a screen is placed a distance, L , from the slits. The light from each slit travels a distance shown as D_1 and D_2 respectively and meets at a point, P , a distance, y , above the midpoint of the screen. If the waves meet at P when they are in phase, they will interfere constructively. The points of the waves are in phase if the path difference that each wave has travelled with respect to the other is a whole number multiple of the wavelength λ .

The difference in path lengths $\Delta D = D_1 - D_2$.

The constructive interference, also known as a 'maximum', occurs when

$$d \sin \theta = n\lambda \text{ where } n = \text{integer and } n\lambda = y \frac{d}{L}.$$

This relationship permits the determination of the wavelength of light of a particular frequency using simple measurements.

When monochromatic light (light of a single frequency) is passed through a pair of closely spaced slits in a screen or grating of some kind, the light passing through each slit acts as a new point source of light. Because the light is from a single monochromatic source it is coherent when passing through the slits. The resulting interference of the light from these two sources causes a pattern of alternating light and dark lines, or fringes, on a screen placed some distance away from the slits that is solely the product of the path distance between the sources and the screen.

The bright areas that result on the screen are called **maxima**; the dark areas **minima**. The centre of each maxima, where it is the brightest, is where the light from the two sources is completely in phase at the screen. The darkest area in the centre of each minima is where the light from the two sources is completely out of phase at the screen. The values of y , d and L can easily be measured from a screen placed some distance from a monochromatic light source that is transmitted through a double slit in

The term **maxima** refers to points on an interference pattern where the peaks of each set of waves coincide. This produces a bright spot when light is used and is a point of constructive interference.

The term **minima** refers to points on an interference pattern where peaks of one wave coincide with troughs of the other. This produces a dark spot and is a point of destructive interference.

A **diffraction grating** is a device consisting of a large number of slits and is used to produce a spectrum.

13.2

Diffraction refers to the spreading out of light waves around the edge of an object or when light passes through a small aperture.

eBook plus

eModelling:
Modelling interference
and diffraction
Spreadsheets help explore
interference and diffraction and
their combined effects
doc-0041

a slide or **diffraction grating**. However, when doing this it is important to make sure that all measurements are made to the central bright region of any maximum that is formed on the screen.

DIFFRACTION

We are familiar with shadows cast on a wall by an object and know that the shadow has the same shape as the object. However, if we look carefully, we will see that the edges of the shadow are a little fuzzy, that is, they are not perfectly sharp. This lack of sharply defined edges on the shadow is due to the phenomenon of **diffraction**.

Young's double slit experiment showed that light does not travel past an object in straight lines, but spreads out around the object's edges as waves. These waves can interfere with each other as they spread out. This spreading out of light that occurs around an object or when light is passing through a small aperture is called diffraction. It is pronounced when the waves have to travel different paths to a point some distance from the source and in doing so travel paths that have differences in length that approach either multiples of half or full wavelengths. A diffraction pattern is shown in figure 13.4.

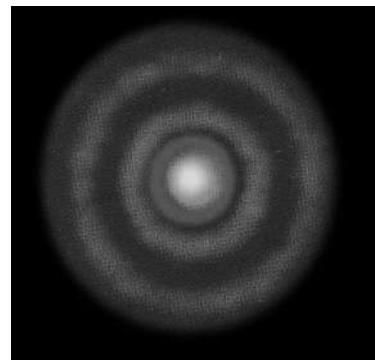


Figure 13.4 Diffraction pattern of a circular aperture (an opening through which light passes), showing maxima and minima

13.3

All modern X-ray tubes are basically the same as the type developed by Coolidge. This device became the standard method of producing X-rays. In the X-ray tube, electrons are accelerated by a high potential difference or voltage to produce the X-rays as they hit a target within the Coolidge tube. Figure 13.5 shows a Coolidge X-ray tube.

eBook plus

Weblink:
X-ray diffraction

X-RAY DIFFRACTION

Diffraction is a property of all waves, including electromagnetic radiation. Diffraction effects increase as the physical dimension of the aperture approaches the wavelength of the waves. Diffraction of waves results in interference that produces dark and bright rings or spots. The precise nature of these effects is dependent on the geometry of the object causing the diffraction.

X-rays were discovered by Röntgen towards the end of the nineteenth century (see the Physics fact in chapter 10, page 185). A study of their nature revealed they were electromagnetic waves. Although similar to light and radio waves, X-rays were determined by experiment to have a wavelength much shorter than that of visible light, in the order of 10^{-10} m. Within a short period, scientists studying these new electromagnetic waves were able to reliably produce X-rays of a specific frequency.

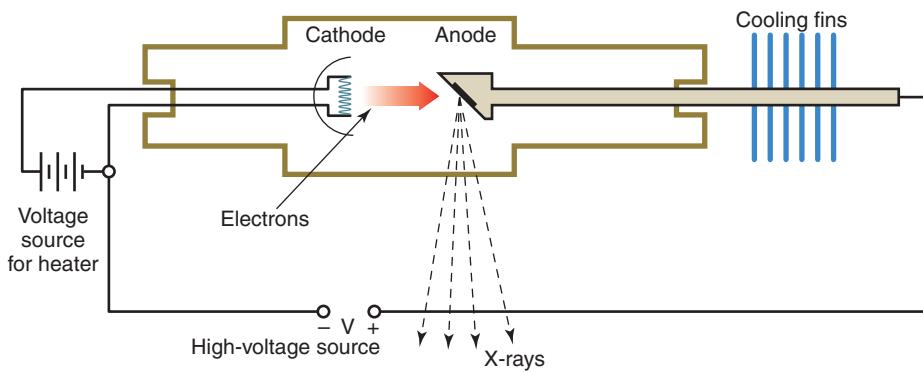


Figure 13.5 A Coolidge X-ray tube, invented in 1913 by American physicist William D. Coolidge

A diffraction grating for visible light is a device for producing interference effects such as spectra. A grating consists of a large number of equidistant parallel lines engraved on a glass or metal surface. The distance between the lines is of the same order as the wavelength of the light. Note that the larger the number of slits on a grating, the sharper the image obtained. This is why a diffraction grating produces a well-separated pattern of narrow peaks (see figure 13.6).

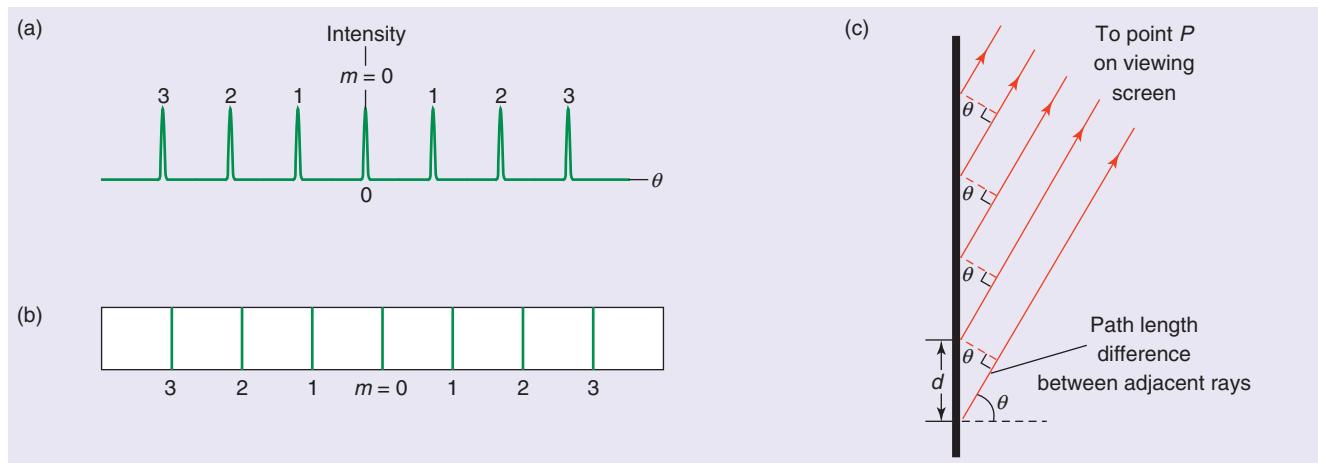


Figure 13.6 A diffraction grating is a set of accurately ruled lines on a glass that produces a sharp pattern of maxima and minima.

X-rays are electromagnetic radiation with a wavelength of the order of 1×10^{-10} m. Compare this with a wavelength of 5.5×10^{-7} m for green light in the middle of the visible spectrum. A standard optical diffraction grating cannot be used to discriminate between different wavelengths in the X-ray wavelength range as it can in the visible spectral range. An optical grating can split visible light up into a series of spectral lines or the rainbow of colours you are probably familiar with from earlier learning. The discrimination of different wavelengths in the X-ray range requires an instrument capable of measuring an angle of less than 0.0019° required for their diffraction.

In 1912, the German physicist Max Von Laue (1879–1960) proposed that the regular spacing of a **crystal**, such as sodium chloride, might form a natural three-dimensional ‘diffraction grating’ for X-rays.

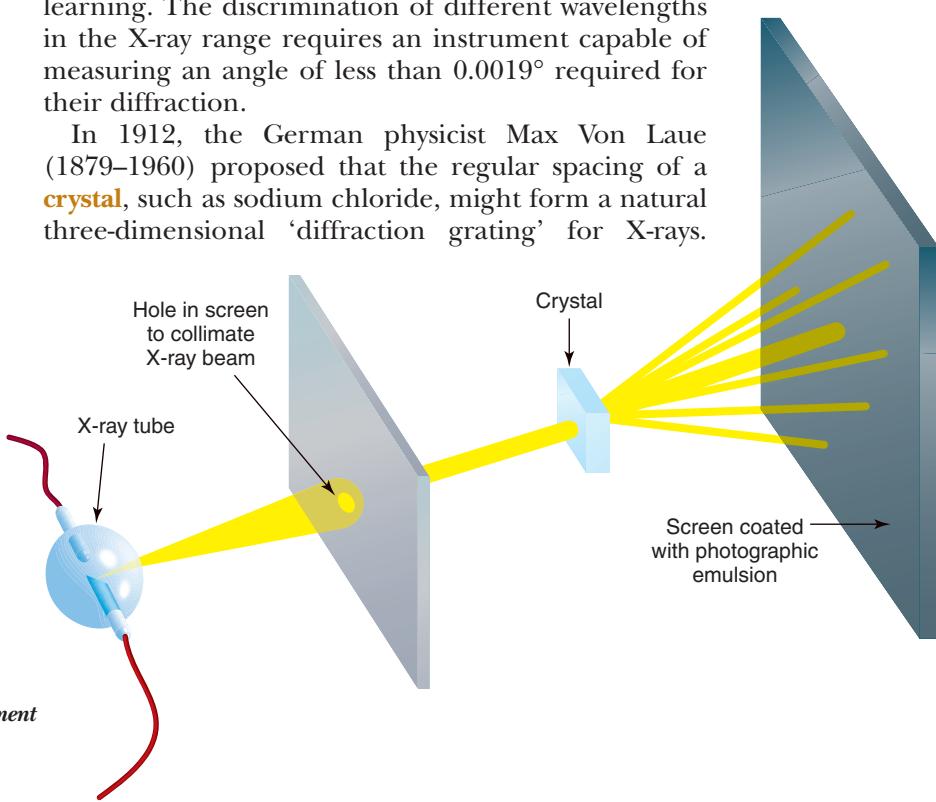


Figure 13.7 A representation of Von Laue’s original diffraction experiment

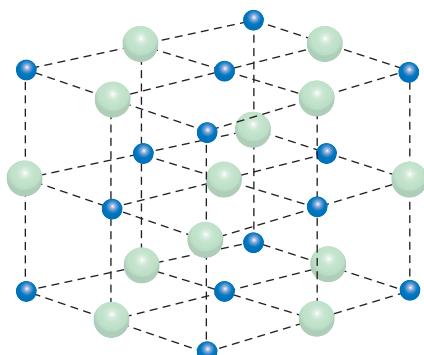


Figure 13.8 Crystal structure of sodium chloride

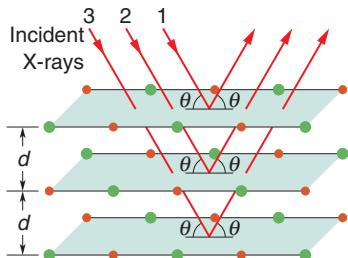


Figure 13.9 The interference in emitted X-rays is caused by some X-rays reflecting from lower levels or adjacent atomic layers.

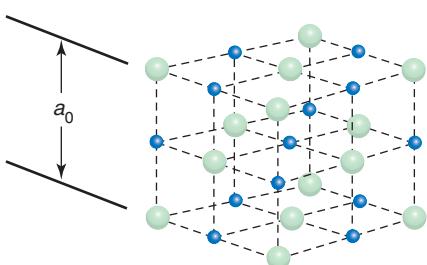


Figure 13.10 a_0 is the width of the 'unit cell' of a sodium chloride crystal.

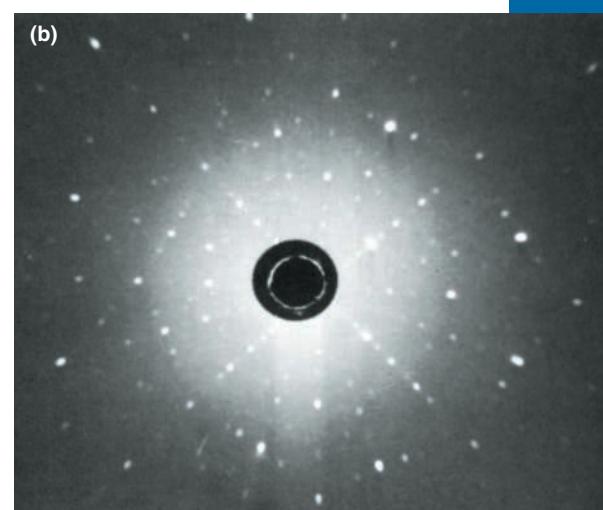
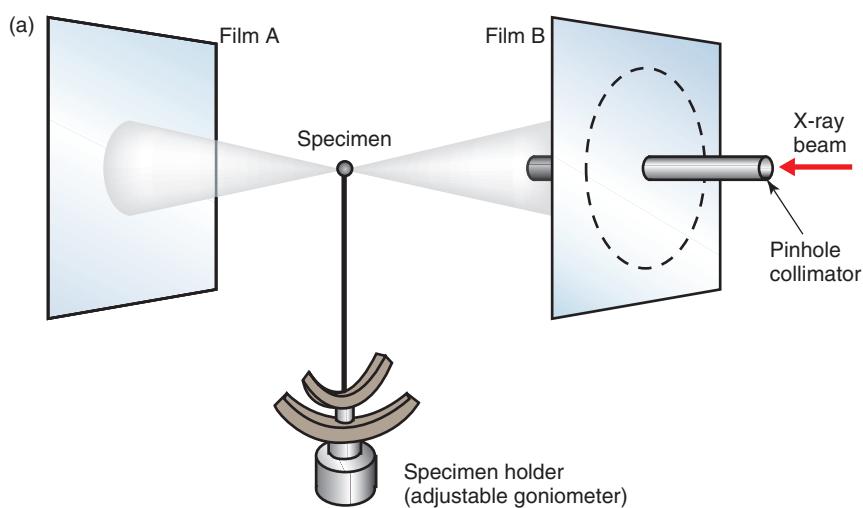


Figure 13.11 (a) A flat plate camera used for X-ray diffraction of a crystal (b) A Laue pattern of a silicon crystal

This experiment was carried out by his colleagues, W. Friedrich and P. Knipping, who bombarded a crystal of zinc sulphide. They obtained a diffraction pattern on photographic film (see figure 13.7).

British physicist Sir William Henry Bragg (1862–1942) and his Australian-born son Sir William Lawrence Bragg (1890–1971) developed an X-ray spectrometer to systematically study diffraction of X-rays from crystal surfaces. They proposed that X-rays, because of their short wavelength (in the order of the size of the atom, 10^{-10} m), could penetrate the surface of matter and 'reflect' from the atomic lattice planes within the crystals.

When X-rays enter a crystal such as sodium chloride (see figure 13.8), they are scattered (absorbed and re-emitted) in all directions.

In some directions the scattered waves undergo destructive interference, resulting in an intensity minimum; in other directions, the interference is constructive, resulting in an intensity maximum. This process of scattering and interference is a form of diffraction. The Braggs observed that the maxima occurred in specific directions. They concluded that the X-rays were reflected from the regularly spaced parallel planes of the crystal that were formed by the arrangement of atoms in the crystal lattice. This effect is shown in figure 13.9. For the first time, thanks to the work of the Braggs, it was possible to look at the arrangement of the atoms in a solid material.

The work of W. L. Bragg provided a mathematical analysis of their experiments, deriving the relationship between the spacing of the crystal planes, the wavelength of the radiation and the angle of reflection. X-ray diffraction provides a tool for studying both X-ray spectra and the arrangement of atoms in a crystal.

To study spectra, a crystal is chosen with a known interplanar spacing, d . A detector is mounted on a device called a goniometer (see figure 13.11) and can be rotated through a range of angles to measure the crystal rotation angles at which the maxima occur. A chart recorder produces a trace of X-ray intensity against rotation angle.

Alternatively, the crystal itself can be studied with a monochromatic X-ray beam of known wavelength. This enables the determination of the spacing of various crystal planes as well as the structure of the crystal unit cell (smallest possible crystal unit, shown in figure 13.10).

Bragg's results revealed a predictable relationship among several factors. They were:

- the distance between similar atomic planes of a crystal (the interatomic spacing), known as the 'd-spacing', measured in angstroms
- the angle of incidence (the angle θ), measured in degrees
- the wavelength of the radiation, λ , which in this case is approximately 1.54 angstroms (1.54×10^{-10} m).

Bragg's Law summarises this relationship between the maxima and the diffraction angles as:

$$n\lambda = 2d \sin \theta$$

where

n = an integer taking values of 1, 2, 3 ... etc.

θ = diffraction angle in degrees.

A crystal was found to be made of atoms arranged in a regular three-dimensional pattern. The diffraction pattern obtained by transmission existed as a pattern of spots. Analysis of these patterns made it possible to determine the position of atoms in the crystal.

13.4 BRAGG'S EXPERIMENT

eBookplus

Weblink:

Bragg diffraction experiment with microwaves

eBookplus

Weblink:

Bragg's law and diffraction applet

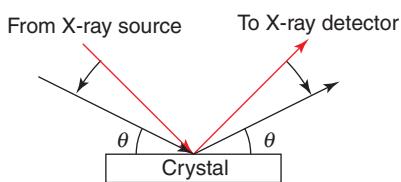


Figure 13.13 As a result of X-ray diffraction the intensity of the detected X-rays varies according to the angle θ .

An X-ray diffraction instrument has a source of X-rays. This is an X-ray tube operating at about 40 000 volts. The target of the high-energy electrons in the tube may be copper or chromium and is mounted on a base of metal through which cooling water flows. The X-rays produced as the accelerated electrons hit the target material are collimated using parallel plates of metal covered in molybdenum. This causes the X-ray beam to become parallel. The parallel X-ray beam strikes the sample under investigation and the scattered X-rays coming from the material are then detected very precisely. (The Braggs actually used an ionisation chamber to detect the X-rays. An ionisation chamber is a gas-filled detector that produces a current pulse when X-rays ionise the gas inside the detector. This enabled them to determine the position and intensity of the diffracted X-rays.)

The process of X-ray diffraction by a crystal is quite complex. The pattern of maxima and minima occurs as if the X-rays were reflected by a system of parallel planes. However, the X-rays are not actually reflected, but scattered. This model uses reflection to simplify the description and calculations.

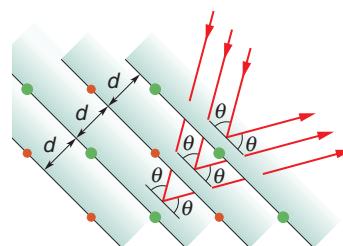
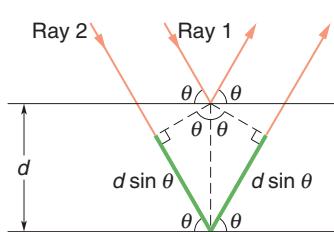


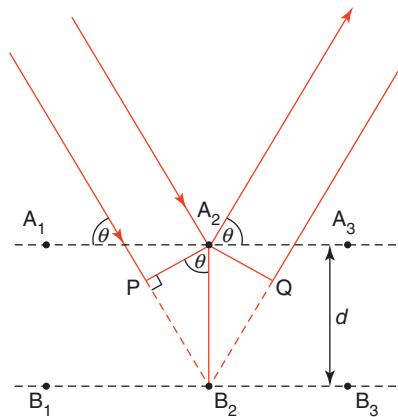
Figure 13.12 A set of reference planes and Bragg diffraction along the planes of a crystal

The Bragg experiment utilised X-rays reflected from adjacent atomic planes within the crystal, as shown in figures 13.12 and 13.13. The reflected X-rays interfered constructively and destructively, producing the familiar pattern. Measurement of the angles allows the spacing and arrangement of the crystal to be determined.

As shown in figure 13.14, the rays of the incident beam are always in phase and parallel until the top beam strikes the top layer of the crystal at an atom (labelled A₂). The second beam continues to the next layer where it is reflected by atom B₂. The second beam must travel the extra distance PB₂ + B₂Q. If the total distance PB₂ + B₂Q is a whole number of wavelengths then the two waves will be in phase.

$$n\lambda = PB_2 + B_2Q$$

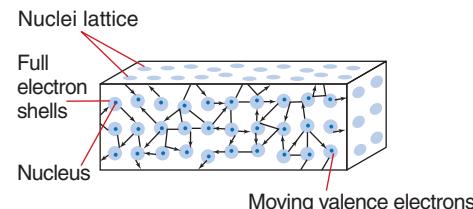
Figure 13.14 Describing Bragg's law requires the use of both geometry and trigonometry. The lower beam must travel the extra distance (PB₂ + B₂Q) to continue travelling parallel and adjacent to the top beam.



13.5 THE CRYSTAL LATTICE STRUCTURE OF METALS

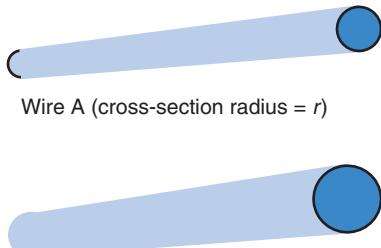
As you saw in chapter 12, the properties of solids depend on the type of bonding. The classical model of metals (see figure 13.15) describes valence electrons as being common property of all of the atoms in the metal, forming an 'electron cloud'. These electrons are said to be 'delocalised'. Because of the random direction of movement of these electrons, with equal numbers moving in each direction, a steady state is established. That is, there will be no net transport of electric charge.

Figure 13.15 Random motion of electrons in a metallic lattice



When an electric field is applied, it produces a small component of velocity in the direction opposite to the field (because electrons have a negative charge). The applied electric field creates a force that drives the electrons in a common direction. At any point in a conductor, the average velocity of the electrons is proportional to the strength of the electric field. If 'piling up' of electrons occurs it can create a reverse potential. This can oppose the electric field. If, for example, there is not a complete circuit, there cannot be a continual flow of electrons and charge will 'pile up'. This means that there will be a short interval after the application of a potential difference, in which a current will flow. However, the build-up of charge opposes the potential difference and the current stops. As such, potential difference can exist across a conductor in an incomplete circuit without any current flowing through it. This is exactly the situation in a dry cell battery. A potential difference exists between the terminals of a dry cell battery without a current necessarily flowing in the circuit.

When a current flows, charge moves under the effect of the applied field. As the charge moves, work is continually being done by the electric field. For each coulomb of charge moving through a potential difference of one volt, one joule of work is done. For a constant current, no kinetic energy is gained, so all of the work done must go into heat. This is generally called ‘joule heating’. It might be noted that joule heating is independent of the direction of current flow and is, in fact, irreversible.



Wire A (cross-section radius = r)

Wire B (cross-section radius = $2r$)

Figure 13.16 Cross-sectional area determines the amount of space in which electrons can travel.

Since direct current and low-frequency, alternating current travel through the body of the metal, the greater the cross-sectional area of the wire (see figure 13.16), the more electrons there are to move along the wire, and the greater the current which can flow for a given value of applied voltage. The atoms that form the lattice vibrate more as their temperature increases. As the electrons begin to move, they collide with impurities and tiny imperfections in the lattice.

The resistance increases as a direct result of these collisions with irregularities in the crystal lattice (see figure 13.17). Inside a superconductor the behaviour of electrons is very different. The impurities and crystal lattice are still there, but the movement of electrons is significantly different. The electrons pass almost unobstructed through the lattice. Because the electrons do not collide with anything, superconductors can transmit electric current with no appreciable loss of energy.

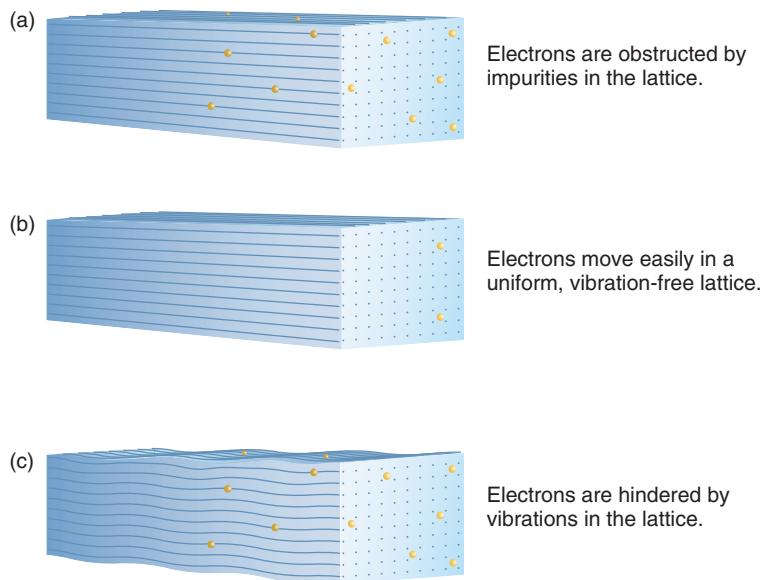


Figure 13.17 The presence of impurities and lattice structure determines the motion of electrons through a conductor.

13.6 SUPERCONDUCTIVITY

A photograph of a magnet floating above a curved disk is one of the most widely published images of superconductivity working (see figure 13.18).

We have seen that electric power is better transmitted as alternating current at very high voltage than as direct current. One advantage of AC transmission is that less energy is lost due to the heating effects caused by the resistance of wires to the flow of electric current. AC transmission also allows the use of very efficient transformers to step up and down the voltage to the required difference in potential.

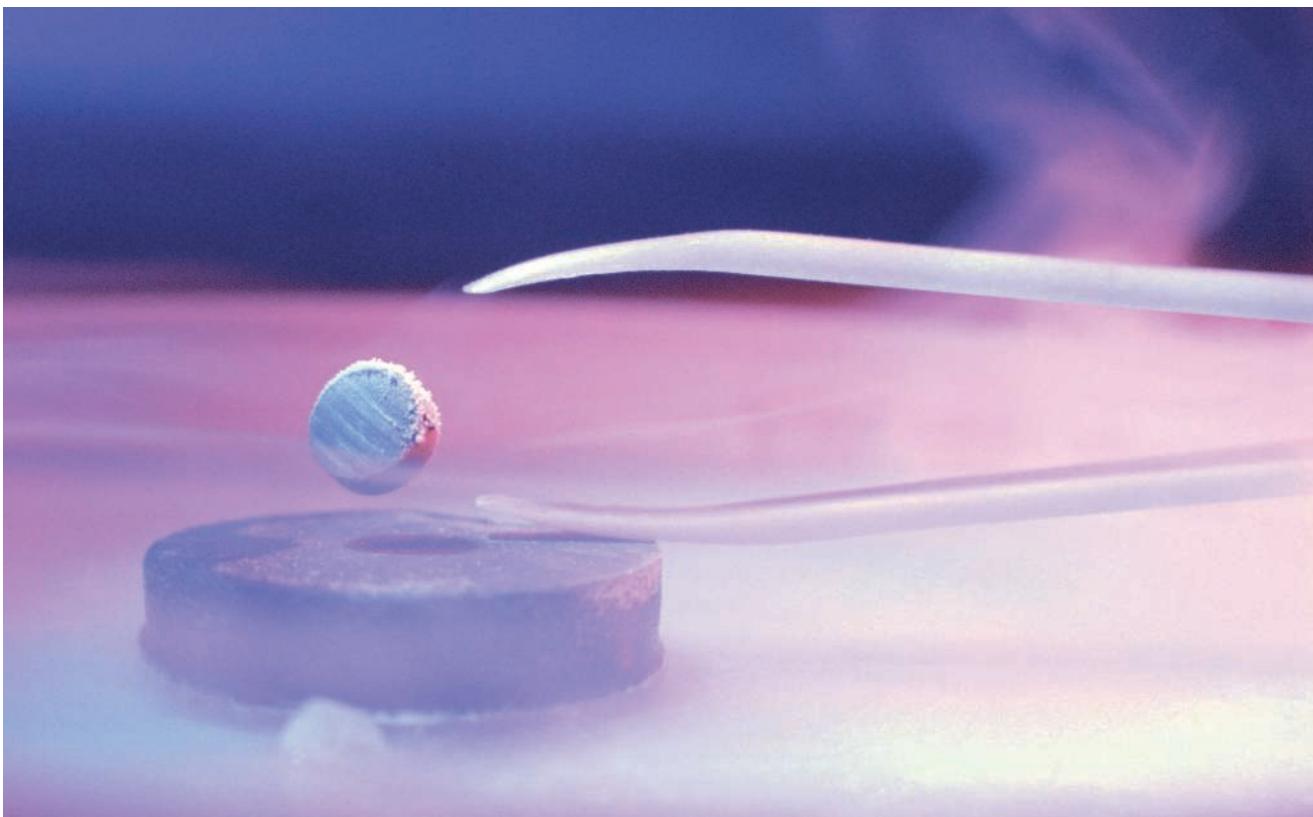


Figure 13.18 A small permanent magnet hovering above a superconducting ceramic disc

You should recall that the Kelvin scale of temperature defines absolute zero (0 K) as 273.16 degrees Celsius below zero.

The battle over whether transmission of power would be via AC or DC was apparently settled during the latter part of the nineteenth century, with AC transmission winning out. Today the advent of the superconductor may swing the advantage back to a system of DC transmission.

The problem is that even copper wire — which is an excellent conductor of electricity — offers some resistance to the passage of that electricity. A copper wire carrying a current of a mere 100 A, if the copper has a resistance of $8 \Omega \text{ km}^{-1}$, will dissipate a huge amount of power for each kilometre the electricity must be transported. The amount of power dissipated per kilometre is determined from the relationship $P = I^2 R$. For the case described, a 100 A current moving through a distance of only 1 km will dissipate energy at a rate of 80 000 W. When you consider that almost every domestic dwelling has the capacity to draw a current of around 100 A through the total of its circuits, and then consider the number of kilometres between the user and the power station supplier generating the electricity, it is clear that the generation and production of power for domestic distribution is a relatively inefficient and energy wasting process.

Imagine the advantages of removing resistance to the flow of electricity along the copper wires from the power station altogether. The study of superconductivity suggests this is possible. (See also pages 246–249.)

The earliest experiments demonstrating that electrical resistance in a wire disappeared at low temperatures occurred in 1911. Before that, in 1906, H. K. Onnes (1853–1926) discovered a superconductor. That superconductor was mercury. Onnes' results are shown in figure 13.19.

The processes for cooling matter to near absolute zero involve using a succession of liquefied gases down to about 4.2 K. Lower temperatures may be achieved by successive magnetisation and demagnetisation. In

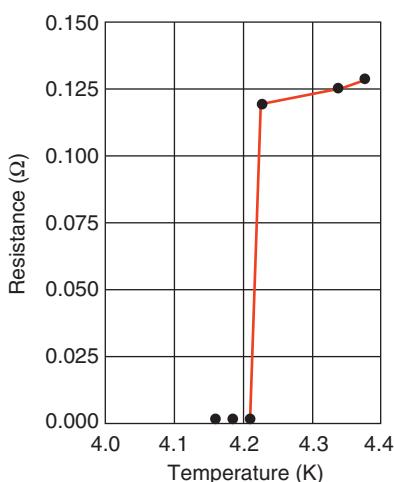


Figure 13.19 A graph of resistance versus temperature in mercury wire as measured by Onnes

1906, Onnes liquefied hydrogen (20.4 K) by continuous vacuum pumping to ensure maximum expansion of the cooling gas. In 1908, using liquid hydrogen, Onnes succeeded in liquefying helium at 4.2 K. In 1911, he used that liquid helium as a coolant and discovered that the electrical resistance of some metals dropped rapidly to almost zero below a temperature that was characteristic of that metal (see figure 13.20).

The characteristic temperature at which a metal becomes superconducting is called its critical temperature, T_c . Table 13.1 gives the critical temperature for some elements.

Since the first superconducting metals were discovered, a number of ceramics have been developed that have a much higher critical temperature and, therefore, are easier and cheaper to use as superconductors than metals or metal alloys. The resistance of the ceramic YBCO at varying temperatures is shown in figure 13.21.

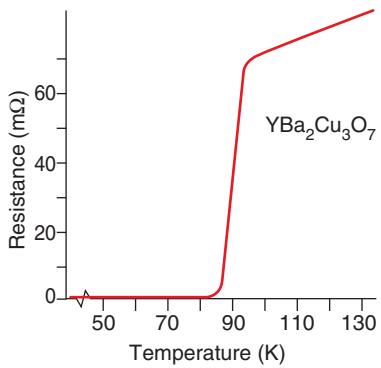


Figure 13.21 Resistance versus temperature for the ceramic YBCO

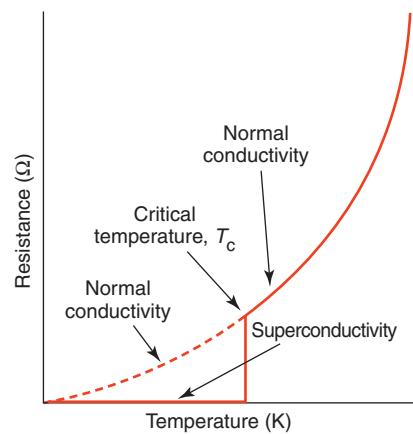


Figure 13.20 The rapid change in resistance with temperature

Table 13.1 Critical temperature (T_c) values for some elements. The ceramics have a much higher critical temperature and therefore are easier to use as superconductors.

ELEMENT/ALLOY	T_c (K)	T_c (°C)
Aluminium	1.20	-271.95
Hafnium	0.35	-272.8
Lead	7.22	-265.93
Mercury	4.12	-269.03
Niobium-aluminium-germanium alloy	21	-253.15
Technetium	11.2	-261.95
Tin	3.73	-269.42
Tin-niobium alloy	18	-255.15
Titanium	0.53	-272.62
Uranium	0.8	-272.35
Metal oxide ceramics		
$\text{YBa}_2\text{Cu}_3\text{O}_7$ (YBCO)	90	-183.15
$\text{HgBa}_2\text{Ca}_2\text{Cu}_3\text{O}_8$	133	-140.15

Very pure metals that become superconductors are known as Type I superconductors and high-temperature ceramics that become superconductors are known as Type II superconductors.

13.7 HOW IS SUPERCONDUCTIVITY EXPLAINED?

It was nearly 50 years after the discovery of superconductivity that a satisfactory explanation was provided. After Onnes discovered superconductivity, many great physicists attempted to find an explanation. Albert Einstein, Werner Heisenberg, Niels Bohr and Wolfgang Pauli are some of those who tried but failed. Some of them had worked on it before quantum mechanics was developed, and it is no wonder that they failed, as superconductivity can only be explained using quantum mechanics.

Success finally came to John Bardeen (1908–1991), Leon Cooper (1930–) and J. Robert Schrieffer (1931–) in 1957. In 1955, John Bardeen, who had already won a Nobel Prize for the invention of the transistor, was keen to recruit Leon Cooper, who had recently received his PhD, to work with him on the problem of superconductivity. Bardeen had left the Bell Laboratories, the scene of his transistor success, and Robert Schrieffer was then a graduate student. Bardeen had already made two unsuccessful attempts, but in 1957, with the help of Cooper and Schrieffer, it was a case of third time lucky. The three of them were awarded the Nobel Prize for Physics in 1972.

Their theory, now called the BCS theory (for the initials of Bardeen, Cooper and Schrieffer), has not received the general recognition it deserved, possibly because it is deeply founded in quantum mechanics, and attempts at classical explanations fail badly. In 2007, at a conference honouring the 50th anniversary of their paper, Dr Cooper recalled the difficulty of the calculations relating to the collective quantum behaviour of millions and millions of electrons. At the time of publication in 1957, many physicists who read their paper failed to comprehend it. At the 2007 conference, Dr Cooper jokingly recalled that when Bardeen was trying to recruit him to work on superconductivity, Bardeen failed to mention his previous two failures!

A clue to the nature of superconductivity had been provided by Walther Meissner and Robert Ochsenfeld in 1933 when they measured the magnetic field in a superconductor and discovered it was precisely zero. They also showed that if a magnetic field was present inside a superconductor above its critical temperature, it would become zero when the material was cooled below its critical temperature and became superconducting. This is now known as the Meissner effect.

In his presentation in 2007, Dr Cooper recalled ‘the simple facts of superconductivity (as of 1955)’. He mentioned the discovery of superconductivity by Kammerling Onnes and the discoveries of Meissner and Ochsenfeld. He also mentioned that there was evidence (from specific heats) to suggest that there was an energy gap involved. The so-called isotope effect, which showed that the critical temperature depended on the mass of the ions present in the lattice, provided evidence for the involvement of phonons, lattice vibrations, in the process.

Bardeen, Cooper and Schrieffer considered that superconductivity occurred because of an interaction between electrons. They initially tried to use all the methods then available for dealing with such a problem, but did so without success.

They knew that electrons repel each other with the Coulomb force between the negatively charged electrons. This contributes an average energy of 1 electron volt per atom. The average energy associated with the superconducting transition could be estimated as 10^{-8} electron volts per atom. This huge difference in energy made the problem very difficult.

Bardeen and David Pines (a student of Bardeen before the arrival of Cooper) had shown that there could be an interaction between electrons and phonons in the lattice, that under some conditions electrons could interact through the exchange of phonons, and that such an interaction could possibly be attractive.

(We have seen in section 13.5 that lattice vibrations increase the resistance of a metal because lattice vibrations, or phonons, scatter electrons. However, sometimes phonons can be exchanged between electrons and produce an attraction between the electrons.)

The idea of an energy gap is important. Somehow, by exchanging phonons, an electron is able to reach a lower energy state. Once it is in that state, there is an energy gap that the electron must overcome to get out of that state. In other words, the energy gap helps to keep electrons in the superconducting state. Bardeen believed that understanding this energy gap was the key to understanding superconductivity.

Dr Cooper found that solving equations with millions and millions of electrons was virtually impossible. However, he found that from his equations he could establish a pairing of electrons; hence the idea of the 'Cooper pair' was born. This was a step beyond the previous discovery of electrons interacting via phonons. Cooper found that when the electrons did interact, they grouped into pairs. Even though these pairs may be constantly breaking and reforming, the key to superconductivity is the electrons interacting and forming Cooper pairs. When Cooper came up with this idea, it was apparent that the two electrons in the pair must have a considerable separation, otherwise the Coulomb repulsion would completely dominate the attractive force between the electrons of the pair. Cooper noted that there would be an overlap of Cooper pairs and that something like 10^6 or 10^7 pairs could occupy the same volume.

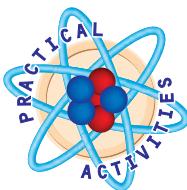
We encountered the Pauli exclusion principle when we first met band structure in solids. Electrons normally obey the Pauli exclusion principle because they are fermions (or spin $\frac{1}{2}$) particles. The electrons in a Cooper pair have opposite spins, so the Cooper pairs are bosons and are not restricted by the Pauli exclusion principle. In fact, all the Cooper pairs can occupy the same energy state, and even though huge numbers of Cooper pairs overlap, they do not interact with each other.

The mathematical problem of dealing with vast numbers of particles still remained, though Cooper had set them on the right track. Robert Schrieffer made the next breakthrough. He applied statistical methods to solve the difficulties associated with the Cooper pairs. He realised that the Cooper pairs seemed to merge into one large group that moved along together. He provided a simple analogy of a line of skaters who are linked arm in arm. If one of the skaters hits a bump, that skater is supported by all the others and continues to move with the group. Although this may sound simple, the mathematics associated with it is very complex.

Once the mathematical problems had been resolved, the BCS team had a theory that could explain all the phenomena related to type I superconductivity. Bardeen, Cooper and Schrieffer were awarded the Nobel Prize for Physics in 1972. (This was Bardeen's second Nobel prize in Physics.)

However, the problem of type II superconductors remains unsolved. BCS theory cannot explain it, and so far there are no satisfactory theories.

In summary, the key to understanding superconductivity in type I superconductors is the formation of Cooper pairs. A Cooper pair consists of two electrons that are a considerable distance apart. The attractive force between the two electrons is provided by the exchange of phonons (lattice vibrations). Formation of a Cooper pair lowers the energy of the



13.1

Temperature change in superconductors



13.2

Levitation and the Meissner effect

electrons, and there is an energy gap they have to cross to jump out of the superconducting state. (Of course, raising the temperature will achieve this and destroy the superconducting state.) The two electrons in a pair have opposite spin, so the Cooper pairs are not constrained by the Pauli exclusion principle. There will be perhaps 10^6 or 10^7 Cooper pairs overlapping each other, but they can all be in the same energy state. It can be considered that all the members of this large association of Cooper pairs assist all the other members to move through the lattice.

When a superconducting material in its normal state is placed in a magnetic field, the magnetic field strength inside the material is almost the same as the magnetic field strength outside the material. If the material is in its superconducting state, currents flow in the superconductor to produce a magnetic field that cancels the applied magnetic field inside the superconductor. That is, the magnetic field inside the superconductor is always zero.

This expulsion of the magnetic field from inside a superconductor is called the Meissner effect and is illustrated in figure 13.22.

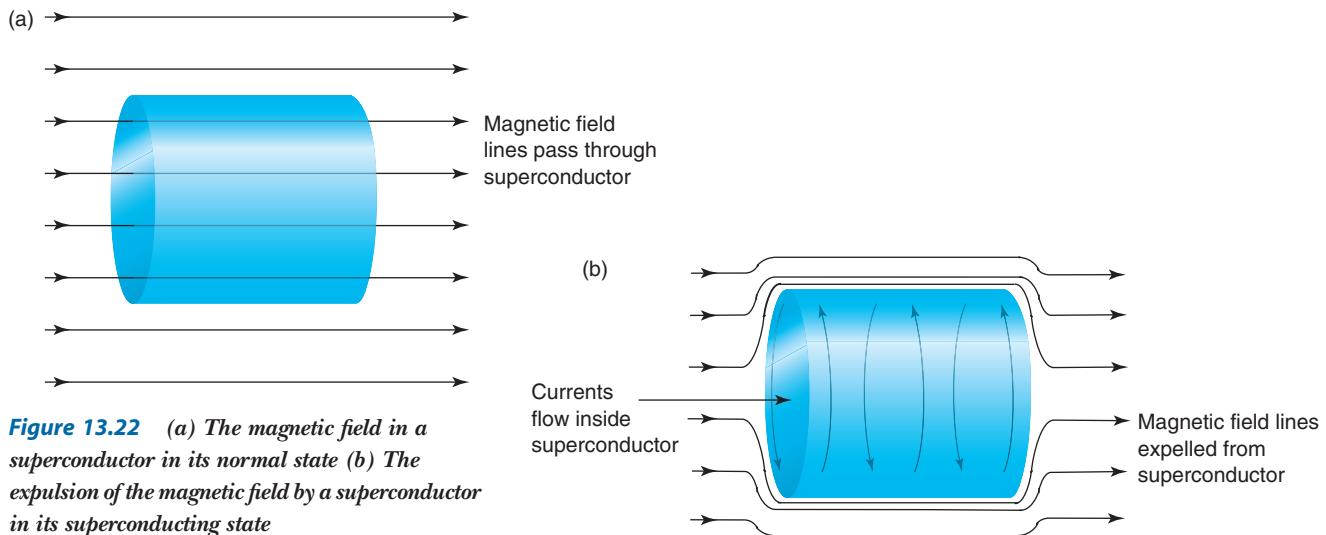


Figure 13.22 (a) The magnetic field in a superconductor in its normal state (b) The expulsion of the magnetic field by a superconductor in its superconducting state

If a magnet is brought near to a superconductor, the currents in the superconductor that expel the magnetic field create magnetic poles which cause repulsion between the magnet and the superconductor. If a small magnet is placed above a superconductor, the repulsive force on the magnet can balance the weight of the magnet — causing it to be suspended above the superconductor. This is shown in figure 13.23 and explains the levitation of the magnet shown in figure 13.18, page 241.

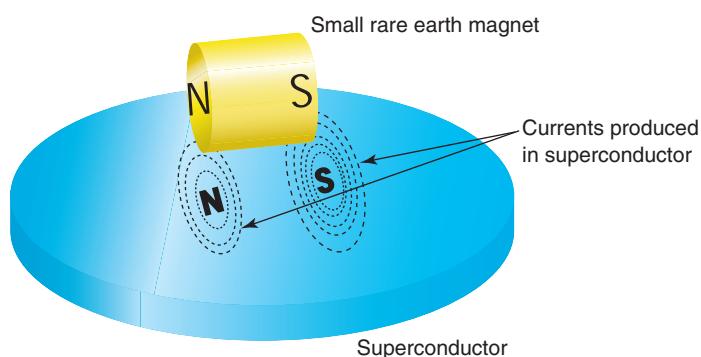


Figure 13.23 The currents produced in a superconductor lead to levitation of a magnet.



13.3

Resistance and superconductors

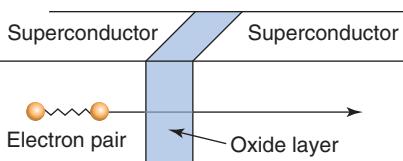


Figure 13.24 The Josephson effect allows superconducting material to act as a switch.

The effects that are described here — zero resistance and the Meissner effect — are the macroscopic properties of materials that become superconducting.

There are also very important microscopic or quantum effects shown by superconductors. One such property is the ‘tunnelling effect’. This is a property of the wave-electron, transporting the electron through spaces that are forbidden by classical physics because of high electric potential. In 1962, Brian Josephson (1940–) discovered that, if two superconducting metals were separated by an insulating barrier, it was possible for electron pairs to pass through the barrier without resistance.

There are two important consequences of this. Even in the absence of a potential difference between two identical superconductors, a tunnelling current will flow across a Josephson junction. Secondly, when a constant potential difference is applied, the current that flows will oscillate with a constant frequency. This second effect allows the most precise measurement of the fundamental quantum constant $\frac{e}{h}$ since $f = \frac{2eV}{h}$.

Such a Josephson junction acts as a superfast switch (see figure 13.24). This property is a major advantage in computers where the processing time depends on the speed at which signal pulses can be transmitted. It also makes the very precise measurement of magnetic flux possible.

Applications of superconductivity

One of the major problems facing the widespread use of superconductivity is the very low temperatures required to produce superconductivity effects. Researchers are working on ‘high-temperature superconductors’ which will retain their superconductive properties at temperatures that can be more easily managed.

The potential application for superconductors is almost unlimited. We will look at a few examples.

Power transmission

The ability to conduct electricity without losing power in heat stimulated research into engineering applications. Electrical transmission lines lose an appreciable amount of energy due to the resistance of the wires. If materials can be developed which overcome the physical problems that make high-temperature superconductors brittle, very large current densities could be conducted in relatively thin wires. This would reduce the cost of power and the need for the ever increasing demand for new power stations. Superconducting wires could carry three to five times as much current as conventional transmission lines. The current in such transmission lines would of course be DC rather than the conventional AC. This is because the constant direction-switching in AC causes energy losses and heating. That would defeat the purpose of the thinner wires and would counter the low superconducting temperatures. One experimental electricity transmission line uses an HTS (high temperature superconductor) material wound around a hollow core which carries a liquid helium coolant.

Power generation

At the point of generation, superconducting magnets that would not require the presence of an iron core would potentially be only a fraction the size and mass of present generators. Less fossil fuel would be required to produce electricity which would reduce the emissions of

greenhouse gases and other pollutants from power plants. Figure 13.25 shows an example of the use of superconductors in the power generation industry.

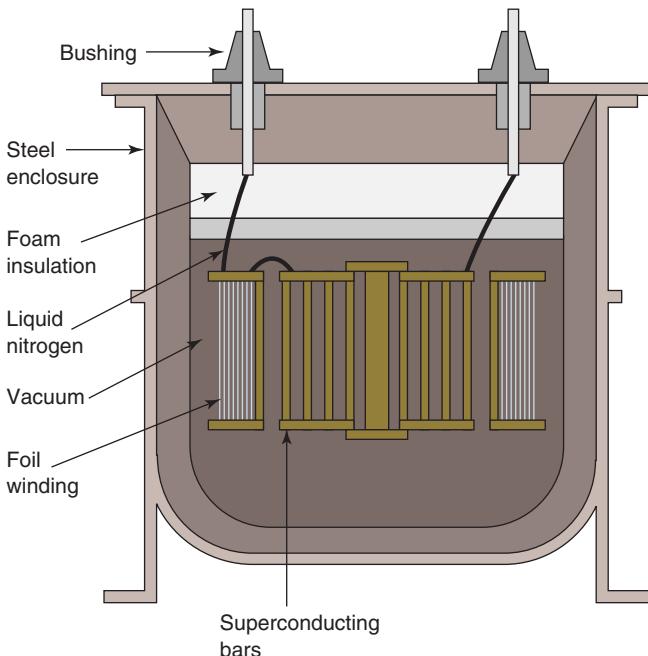


Figure 13.25 A section of the prototype for a high-temperature superconducting fault current limiter (FCL). FCLs, sometimes called ‘chokes’, are important devices in the energy industry for controlling faults in power supplies. The superconducting elements are housed in a stainless steel vacuum flask filled with nitrogen.

Power storage

One of the major problems faced by power stations is that electrical energy cannot be stored easily. Essentially, electricity must be used immediately. Superconducting magnetic energy storage (SMES) is one possible answer to this problem. These facilities use a large ring structure constructed using a HTS material and refrigerators. Electrical energy in the form of a DC current can be introduced into the device. The current is introduced as DC because the constant switching of direction in AC produces some energy loss. The DC electrical current would flow around the SMES device’s circular path indefinitely without energy loss until required, whereupon it can be retrieved and converted into AC current for delivery to domestic users and industry. Alternatively, it could be transported by a superconducting transmission system as DC. The big advantage of the SMES electricity storage system is that the power generation machinery can continuously operate at peak efficiency levels no matter whether demand is at a maximum or minimum. This minimises the need to build new power stations and potentially opens the way to the use of large-scale solar power stations with energy produced during the daylight being stored for 24-hour use.

Electronics

There is enormous scope for the use of superconductors in electronics. The speed and further miniaturisation of computer chips are limited by the generation of heat (due to resistance of the electric current flow required to make them run), and by the speed with which signals can be

conducted. Superconductive film, used as connecting conductors, may result in more densely packed semiconductor chips. These could transmit information several orders of magnitude faster. Superconducting digital electronic components have achieved switching times of 9 picoseconds (9×10^{-12} s). This is much faster than any conventional switching device. The development of Josephson junctions has led to very sensitive microwave detectors and magnetometers that are used in geological surveys.

Medical diagnostics

Superconducting magnets have a vital role in the development of new diagnostic tools. The intense magnetic fields used in magnetic resonance imaging (MRI) instruments, are ideal applications of superconductors. The magnetic solenoids need to be large enough to allow a person to enter. To produce a magnetic field of four tesla, an ordinary solenoid winding would have to be a metre thick. There would be several hundred kilowatts of power dissipated as heat for every metre length of conductor. With superconducting alloys of tin-niobium, magnetic fields of 4 T can be established easily. Furthermore, once the desired strength of magnetic field is energised, that is, the current level in the superconducting solenoid is achieved, it runs in a ‘persistent current mode’. In other words, the current is simply cycling around the solenoid without energy loss. Therefore, the device does not require the input of any more electric power to maintain the magnetic field. Normal solenoids would require constant power input. The MRI works by tunnelling radio frequencies to produce photons that have energies similar to the difference between the ‘spin up’ and ‘spin down’ states in a hydrogen atom in the human body. The signal produced is essentially a measure of the concentration of hydrogen atoms. From this information a measure of the soft tissue in a person’s body can be computer generated without the need for invasive surgery or without the inherent disadvantages of using seismic energy in ultrasounds. Smaller magnetic solenoids can be placed around a person’s body to follow the contours of the body. By changing the radio frequencies of the instrument and with the use of computer reconstruction, three-dimensional maps of the soft tissues of a patient can be made that are highly detailed and accurate. The ability to measure magnetic flux precisely is used in a SQUID (Superconducting Quantum Interference Device). This is an instrument to measure tiny magnetic fluxes that generate electrical impulses in the device. Magnetic fields as small as 10^{-13} T occur with small variations in current within the brain or the heart. These provide important diagnostic information for doctors.

The magnetically levitated train

Magnetic levitation (maglev) suspends an object so that it is free of contact with any surface. This has the effect of providing a frictionless contact with the ground, making it particularly appropriate for high-speed trains. A typical maglev train is streamlined to reduce air resistance and travels along a guideway (see figure 13.1, page 232, and figure 13.26). Once the train is levitated, by continually changing the polarity of alternate magnets along the track, a series of attractions and repulsions is generated that provides the force to overcome air resistance and accelerate the train along the guideway. Speeds of 517 km h^{-1} have been achieved in Japan. The main obstacle to higher speeds is the air resistance encountered by the train during motion. The enormous amount of electrical power needed by the train is an obstacle to its wider use.

There are two different maglev systems.

- The electromagnetic suspension system (EMS), currently used in Germany, uses conventional electromagnets mounted under the train on structures that wrap around the guideway to provide the lift and to create the frictionless running surface. This system is unstable because of the varying distances between the magnets and the guideway. This instability needs to be monitored closely and computers have provided the control to correct the instability. The lifting force is produced by arrays of electromagnets of like polarity in both the train and the guideway. The magnets repel each other to lift the train above the track.
- The electrodynamic suspension (EDS) system, developed in Japan, uses superconducting magnets on the vehicle and electrically conductive strips or coils in the guideway to levitate the train. This does not require the same degree of computer monitoring and adjustment while travelling, but the requirement for very low temperatures means that, for the moment, this is not a practical system. The system for accelerating the train along a guideway is similar to the EMS system.

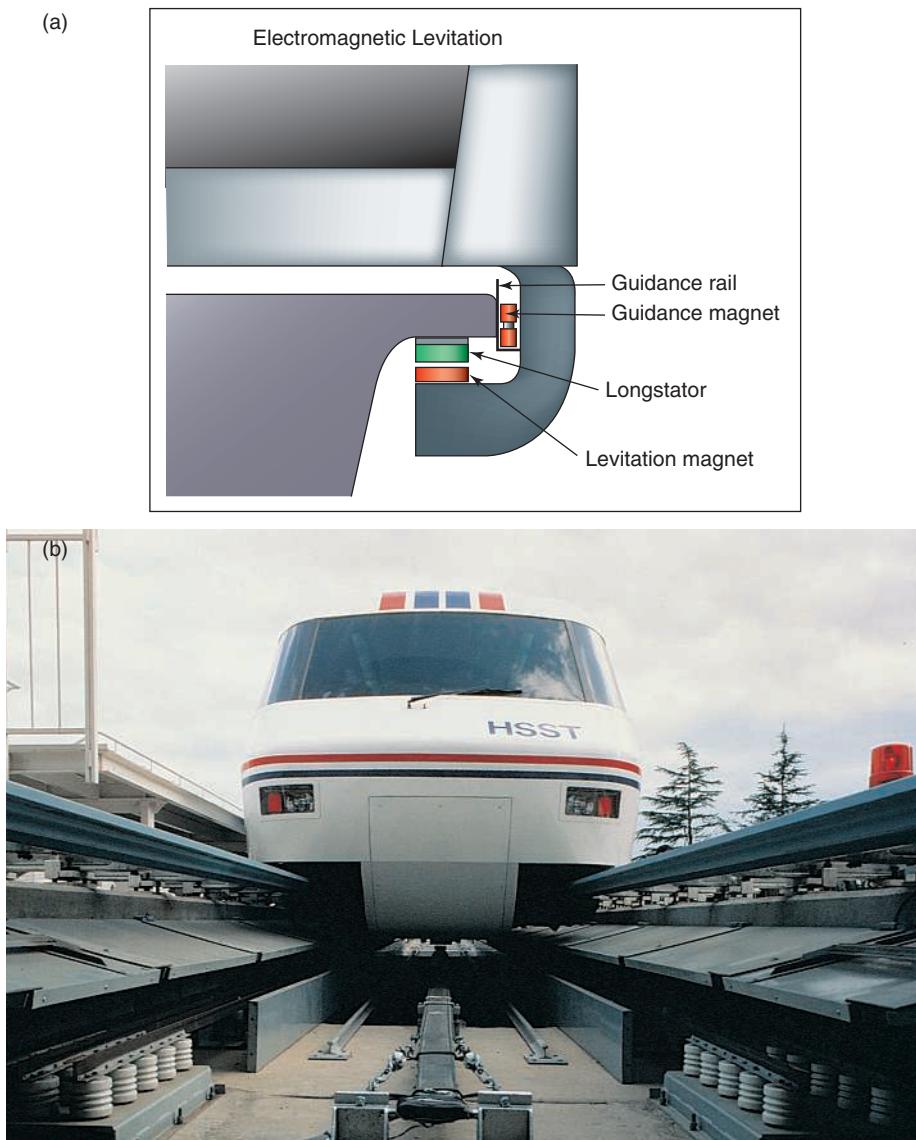


Figure 13.26 (a) Use of magnets in suspending the train (b) A maglev train, based on the EMS system

Particle accelerators

Most high energy particle accelerators now use superconducting magnets. The Tevatron at Fermilab uses 744 superconducting magnets in a ring of circumference 6.2 km.

The Large Hadron Collider (LHC) at CERN has a circumference of 27 km. It uses 1232 large dipole superconducting magnets, as well as many smaller ones, bringing the total to over 1700 magnets. Each of the large dipole magnets is 15 metres long and weighs 35 tonnes. A total length of over 7000 kilometres of niobium–titanium superconducting cable is used in these magnets. It has been estimated that if conventional electromagnets had been used, the accelerator would have had to be over 120 kilometres in circumference to produce the same energies, and its electricity demands would have been phenomenal. (For more information on the LHC, see ‘Recent developments in accelerator research’, page 512.)

Table 13.2 A time line of the development of superconductivity

1911	Dutch physicist Heike Kamerlingh Onnes discovers superconductivity in mercury at a temperature of 4 K. Onnes immediately predicts many uses for superconductors. One prediction involves the ability to produce electrical motors without the need for an iron core in the electromagnet because the current carried would be so much greater than possible with conventional wiring. Onnes’ vision is thwarted by the difficulties of finding superconductors that can be made into wires easily and carry the required current densities required to realise his vision. The search for HTS begins.
1912	Onnes is awarded the Nobel Prize in physics for his research into the properties of matter at low temperature.
1933	W. Meissner and R. Ochsenfield discover the Meissner effect.
1941	Scientists report superconductivity in niobium nitride at 16 K.
1953	Vanadium-3 silicon found to superconduct at 17.5 K.
1962	Westinghouse scientists develop the first commercial niobium-titanium superconducting wire.
1962	English physicist Brian Josephson predicts the ‘tunnelling’ phenomenon in which pairs of electrons can pass through a thin insulating strip between superconductors.
1968	IBM begins a research program to produce a Josephson junction computer.
1972	John Bardeen, Leon Cooper and John Schrieffer win the Nobel Prize in physics for the first successful theory of how superconductivity works.
1972	The Japanese test their first magnetically levitated (maglev) rail vehicle using a niobium-titanium superconductor.
1982	The first MRI machines are placed in hospitals for evaluation. These use superconducting wires that create a powerful magnetic field. These machines are considered the most significant advance in imaging devices since the X-ray machine.
1986	IBM researchers A. Muller and G. Bednorz make a ceramic compound of lanthanum, barium, copper and oxygen that superconducts at 35 K.
1987	University researchers at Houston use yttrium instead of lanthanum to produce a superconductor that operates at 92 K.
1988	Allen Hermann of the University of Arkansas makes a superconducting ceramic containing calcium and thallium that superconducts at 120 K, well above the boiling point of liquid nitrogen (78 K).
1993	A. Schilling, M. Cantoni and J. Guo produce a superconductor from mercury, barium and copper with a maximum transition temperature of 133 K.
1996	US researchers demonstrate a 200 horsepower motor and a 2.4 kilowatt current limiter based on HTS. A 50-metre HTS transmission line is built.

SUMMARY

- Max von Laue first showed that X-rays were diffracted into different patterns by rock crystals.
- Bragg's Law is the description of the relationship between the wavelength of X-rays hitting and bouncing off a crystal surface and the distance between the particle layers in the crystal's structure. It is described by the simple equation:

$$n\lambda = 2d \sin \theta.$$
- X-ray crystallography deals with the scattering of X-rays by arrays of atoms. When such arrays are very regular, such as in crystals, it is possible to interpret the results of the photographic pattern caused by this diffraction in terms of the atomic array in the structure of the crystal.
- Metal possesses a crystal lattice structure.
- The temperature at which a metal achieves superconducting ability is called the critical temperature, T_c . Ceramic alloys have a much higher critical temperature and therefore are easier to use as superconductors than metals or metal alloys.
- The Meissner effect is the expulsion of a magnetic field from inside a superconductor.
- The BCS describes pairs of electrons (Cooper pairs) which perform a coordinated motion inside the crystal structure of the conductor.
- There are two types of superconductor: Type I are made of metals, have a low critical temperatures and produce small magnetic fields; Type II are made from ceramic compounds which have a higher T_c and produce more useful magnetic fields.
- Superconductors are used in superconducting magnets (for example, the maglev train), transmission of current and computer switching.

QUESTIONS

- In an X-ray diffraction experiment, a maximum intensity of X-rays was found to be at 15° and the interatomic distance was 2.5×10^{-10} m.
 - Would you expect to observe maximum intensities at other angles?
 - Explain why this happens.

- Explain the significance of the Braggs' use of regular crystal planes instead of a diffraction grating with X-rays.
- Describe the role that Cooper pairs of electrons play in superconductivity.
- (a) Describe two main differences between Type I and Type II superconductors.
 (b) When you experiment with superconductors in the lab, you use Type II superconductors. Explain why you would not normally use Type I.
- (a) What is the resistance of a superconductor in the normal state if 300 mA of current are passing through the sample and 4.2 mV are measured across the voltage probes?
 (b) What potential difference is required to force 300 mA through the same superconductor, if the resistance is $1.0 \times 10^{-4} \Omega$?
- Consider wiring the superconductor from question 6(b) in series with a 10Ω resistor and connected to a 1.5 V battery. How much electrical current will flow through the superconductor?
- Explain how a flat superconductor is able to levitate a small magnet above its surface.
- The following data were obtained from a YBCO superconductor.
 - Calculate the resistance (and complete the table on page 252) for each trial given that a constant current of 100 mA was flowing through the sample.
 - Explain why some of the voltages obtained in this experiment show negative values.
 - Using the information in the table on the following page, graph resistance versus temperature.
 - Find the point on the graph with the largest slope.
 - Estimate the critical temperature, T_c , from your graph.
 - Is this a Type I or a Type II superconductor? Explain.
- List and describe two applications of superconductors.
- (a) List the problems that scientists must overcome before superconductors can be used effectively.
 (b) Suggest some ways of overcoming these problems.
- What are the advantages of 'high temperature' superconductors?
- Describe the differences between conductors and superconductors.

VOLTAGE (V)	TEMPERATURE (K)	RESISTANCE (Ω)
0.001 0370	118.2	
0.0010270	116.1	
0.001 0600	114.8	
0.001 0490	112.9	
0.001 0350	110.9	
0.001 0200	109.1	
0.001 0090	106.9	
0.001 0010	105.0	
0.000 9890	103.5	
0.000 9750	102.2	
0.000 9670	100.0	
0.000 9510	97.9	
0.000 9440	95.8	
0.000 9180	95.0	
0.000 9110	94.3	
0.000 8920	93.8	
0.008 440	93.5	
0.007 830	93.2	
0.006 390	93.0	
0.000 5050	92.6	
0.000 3790	92.3	
0.000 2430	92.1	
0.000 0930	91.7	
0.000 0100	91.4	
0.000 0030	91.0	
0.000 0002	90.8	
-0.000 0002	90.1	
-0.000 0001	89.9	
0.000 0003	89.5	
-0.000 0001	88.8	
0.000 0001	88.5	



13.1 TEMPERATURE CHANGE IN SUPERCONDUCTORS

Aim

To determine the T_c of a superconductor using the Meissner effect.

Theory

One way to measure the critical temperature of a superconductor is by using the Meissner effect. When the temperature of a superconductor is lowered to below the critical temperature, T_c , the superconductor will push the field ‘out of itself’. This will result in the magnet being forced to levitate and float above the superconductor.

By noting the temperature changes as this levitation occurs it is possible to obtain the critical temperature.

Apparatus

YBCO superconductor with attached thermocouple

small magnet

digital voltmeter

liquid nitrogen

If available, a temperature probe attached to a data logger is more useful. Data collected can be transferred to a computer and can be processed by all students using an appropriate software package.

Method

1. Attach the thermocouple lead from the superconductor to a digital voltmeter set to the millivolt range.
2. Completely immerse the superconducting disc in liquid nitrogen with the thermocouple on the bottom. Calibrate your thermocouple according to the manufacturer’s specifications.
3. Balance your magnet above the superconducting material and observe the levitation due to the Meissner effect. When the liquid nitrogen has almost completely boiled away the temperature will begin to increase.
4. Observe the magnet as the disc warms.

Note: Even though the top of the disc warms up before the bottom, the Meissner effect will continue (but diminish) until the entire superconductor is above the critical temperature.

Results

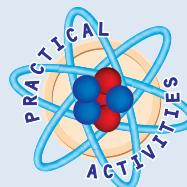
Observe the effects of each step of the method and record the temperature when the magnet comes to a complete rest on the superconducting disc.

Analysis

Predict what will happen as the liquid nitrogen boils away.

Questions

Why should the thermocouple be placed on the bottom during step 2 of the method?



13.2 LEVITATION AND THE MEISSNER EFFECT

Aim

To observe the Meissner effect — the levitation of a magnet above a superconductor.

Apparatus

superconducting pellet

neodymium-iron-boron (or other strong) magnet

liquid nitrogen

petri dish

dewar flask or styrofoam cup

non-magnetic (non-metallic) tweezers

insulated gloves

Theory

Levitation of a magnet above a superconductor is an example of the Meissner effect. If the temperature of a superconductor is lower than its critical temperature, a magnetic field cannot penetrate the superconductor.

Method

1. Carefully fill the styrofoam cup with liquid nitrogen. Place the petri dish on top of the styrofoam cup and carefully pour in enough liquid nitrogen until the liquid is about 1 cm deep. Wait until the boiling subsides.

- Using non-metallic (non-magnetic) tweezers, carefully place the superconductor in the liquid nitrogen in the petri dish. Again wait until the boiling subsides.
- Using the same non-metallic tweezers, carefully place a small magnet about two mm above the centre of the superconductor pellet. Release the magnet.
- If the magnet jumps away from the superconductor, try placing the magnet on the pellet and then placing both in the petri dish and wait until the boiling subsides.
- While the magnet is suspended above the superconducting pellet, gently rotate the magnet using the non-magnetic tweezers.

Results

Describe what you observe and note any changes that occur over time.

Questions

- What temperature would you predict that the magnet to move away from the superconductor?
- Predict what should happen using the alternative procedure outlined in step 4 of the method. Is your prediction supported by your observations? Explain.
- Explain what you observe when the magnet is gently rotated using non-magnetic tweezers.



13.3 RESISTANCE AND SUPER- CONDUCTORS

Aim

To produce a switch and hence observe the resistance within superconductors.

Apparatus

YBCO superconductive wire with attached leads
two dry cell batteries with holder/attachments
three volt light globe with holder
liquid nitrogen
dewar flask or styrofoam cup (as an alternative)

Method

- Connect the superconductor, light globe and batteries in series. When the superconductor is at room temperature it is in the normal state and will therefore have a high resistance. As a result of having a high resistance in series in the circuit, the globe will not light.
- Your teacher will place the superconductor into the liquid nitrogen.
- Your teacher will then remove the superconductor from the liquid nitrogen. The globe will begin to dim and eventually go out.

Analysis

At each stage describe what you observe and explain the observations.

ASTROPHYSICS

Chapter 14

Looking and seeing

Chapter 15

Astronomical measurement

Chapter 16

Binaries and variables

Chapter 17

Star lives



CHAPTER 14

LOOKING AND SEEING

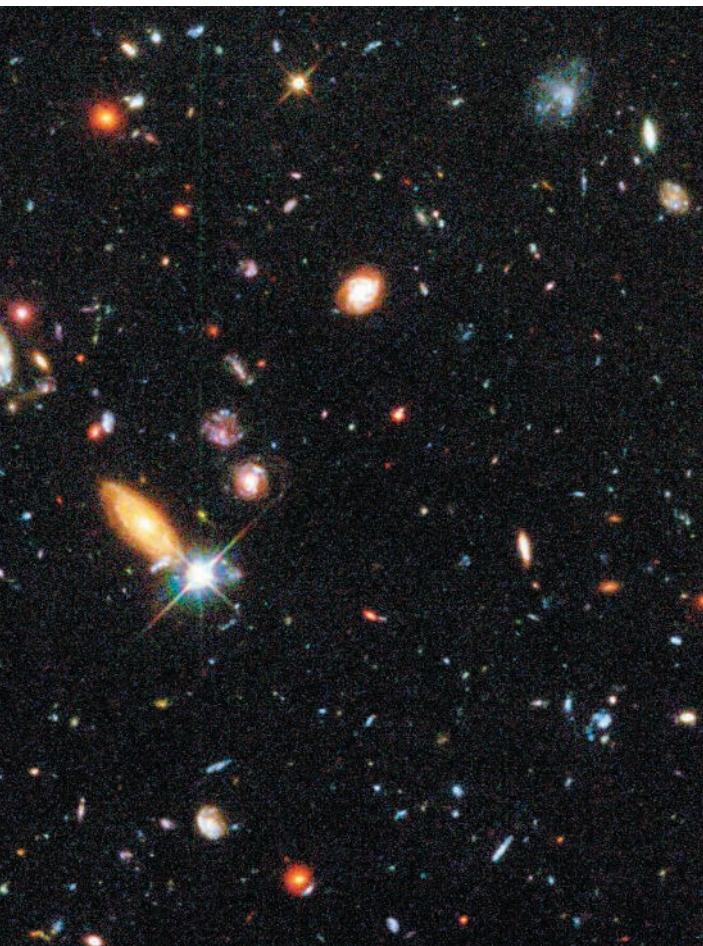


Figure 14.1 The Hubble Deep Field is a small portion of the sky selected for its absence of foreground stars. Although a very small portion of the northern sky — about the size of a grain of sand held at arm's length — the Hubble Deep Field shows over 3000 galaxies at various distances. This image is a compilation of over 300 separate exposures taken by the Hubble Space Telescope.

Remember

Before beginning this chapter, you should be able to:

- describe the electromagnetic spectrum and its components
- describe atmospheric filtering of the electromagnetic spectrum.

Key content

At the end of this chapter you should be able to:

- describe the selective absorption of the electromagnetic spectrum by the atmosphere and relate this to the need to observe those wavelengths from space
- discuss Galileo's use of his telescope to observe the features of the Moon
- define the terms resolution and sensitivity of telescopes, and be able to calculate the resolution of a variety of telescopes
- demonstrate why telescopes need a large diameter objective lens or mirror for sensitivity and resolution
- discuss the problems associated with ground-based astronomy in terms of resolution and selective absorption of electromagnetic radiation
- outline methods being employed to try to improve the resolution and/or sensitivity of ground-based systems, including active optics, adaptive optics and interferometry.

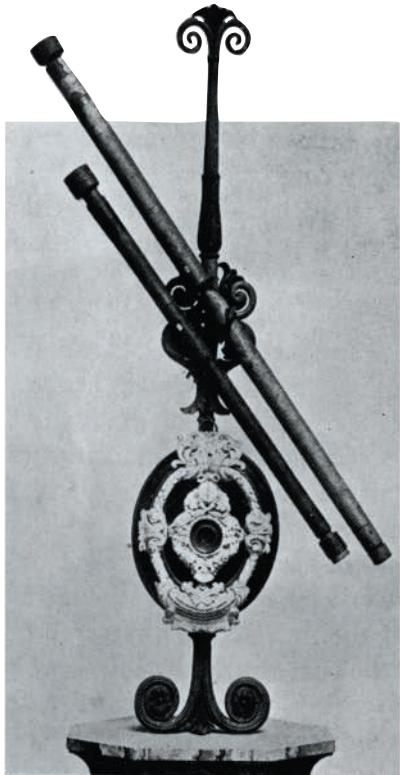


Figure 14.2 Two of Galileo's telescopes on display at the museum of science in Florence, Italy

14.1 GALILEO'S TELESCOPES

In 1609, an Italian scientist named Galileo Galilei heard the news that a Dutchman had pieced together a few glass lenses in such a way as to make distant objects seem much closer. Galileo set about constructing such a device for himself. First he had to work out the optics involved, then grind the lenses and then build it. His first attempt was quite poor but clearly demonstrated its potential. He built more and better (two are shown in figure 14.2) telescopes until he had one with a magnification of over thirty times.

Five years earlier, Galileo witnessed the appearance of a new star in the heavens — something that was not supposed to occur according to the prevailing Aristotelian view, endorsed by the Roman Catholic Church, that the heavens were perfect and unchanging. That event sparked within him an interest in astronomy so that, when finally he had a telescope of his own, it was natural for him to point it upward, at the night sky. He was the first person to do so.

The first heavenly object to catch his attention was the Moon. Of that he wrote:

...the surface of the moon is not smooth, uniform, and precisely spherical as a great number of philosophers believe it (and other heavenly bodies) to be, but is uneven, rough, and full of cavities and prominences, being not unlike the face of the Earth, relieved by chains of mountains and deep valleys.



Figure 14.3 A sample of the sketches Galileo made of the Moon

Galileo was even able to devise a means of calculating the height of a mountain on the Moon from a measurement of its shadow. These observations of the Moon were startling enough but, during the course of that year and the next, Galileo made a number of other startling discoveries, such as the moons of Jupiter, that would fundamentally challenge the way science would regard the heavens.

Such was the impact of the first application of a telescope to the field of astronomy. Telescopes literally provide us with a window to the heavens and, as technology has improved, so too has our ability to build bigger and better telescopes which, in turn, yield more discoveries and better science.

14.2 ATMOSPHERIC ABSORPTION OF THE ELECTROMAGNETIC SPECTRUM

The electromagnetic spectrum was discussed in Physics 1: Preliminary Course, chapter 3. It is also illustrated in this text in chapter 11 (see page 195).

You will recall from your Year 11 studies of Physics that visible light is but one part of a larger spectrum of electromagnetic radiation. The various components of the spectrum are summarised in table 14.1. Note that there is no clear boundary between adjacent parts of the spectrum, and so there is some overlap in wavelength.

Table 14.1 Components of the electromagnetic spectrum

EM SPECTRUM	WAVELENGTH (m)	COMMENT
Gamma rays	$< 10^{-10}$	Absorbed by the atmosphere.
X-rays	10^{-11} to 10^{-7}	Absorbed by the atmosphere.
Ultraviolet	10^{-8} to 4×10^{-7}	Mostly absorbed by the atmosphere.
Visible light	4×10^{-7} to 7×10^{-7}	Not absorbed by the atmosphere.
Infra-red	7×10^{-7} to 1×10^{-2}	Freely penetrates haze but is incompletely absorbed by the atmosphere.
Radio waves	1×10^{-3} to 1×10^6	A broad grouping of microwaves and radio bands — uhf, vhf, hf, mf and lf. Not absorbed by the atmosphere.

As indicated in table 14.1, not all of the electromagnetic spectrum can penetrate the atmosphere of the Earth. Despite the fact that all components of the electromagnetic spectrum strike the outer atmosphere from space, only visible light, radio waves and some UV and IR make it through to the ground (see figure 14.4).

This, in turn, means that ground-based telescopes can operate only in the visible spectrum (optical telescopes such as the Anglo-Australian Telescope, shown in figure 14.5) or in the radio bands (radio telescopes such as Parkes, shown in figure 14.6 on page 260). Observations of other frequencies must be carried out either from a plane or high-altitude balloon in the upper atmosphere or from a spacecraft above the atmosphere, such as the Hubble Space Telescope (see figure 14.7, page 260). For ground-based telescopes, some atmospheric effects still have to be taken into account as discussed in section 14.4 (page 265).

See the ‘Physics in focus’ boxes on pages 264, 266 and 268 for information on the latest in space and ground-based telescopes around the world.

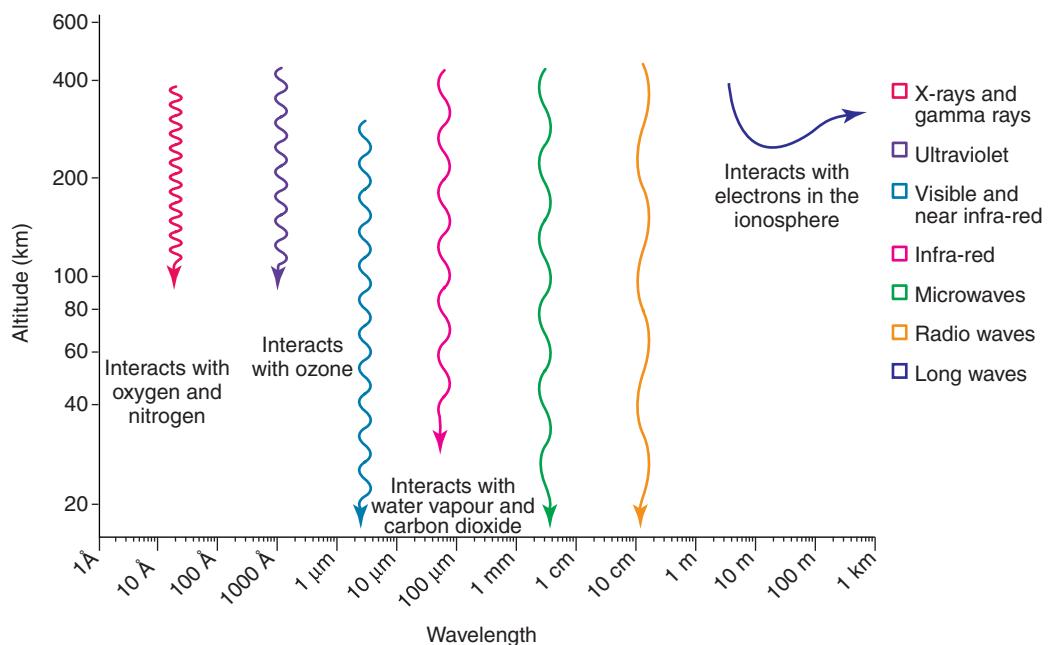


Figure 14.4 The Earth's atmosphere filters out most of the electromagnetic radiation coming from space. Two notable 'windows' exist in this barrier — visible light and radio waves.



Figure 14.5 The 3.9 metre Anglo-Australian Telescope (AAT) at Siding Springs

Figure 14.6 The Parkes 64-metre radio telescope



Figure 14.7 The Hubble Space Telescope (HST)

14.3 TELESCOPES

There are many different designs for telescopes, yet all of the popular designs are based upon just two basic arrangements — refracting telescopes and reflecting telescopes.

Refracting telescopes

A refracting telescope, such as that shown in figure 14.8, is the style of telescope that most people recognise. Lenses are used to gather and focus the starlight by refraction, or bending, of the rays. As figure 14.9 shows, the light enters at one end and is focused by two lenses to form an image in an observing eye located at the other end. This arrangement of lenses causes an image to be seen upside-down and back-to-front; however, this is not a problem when observing stars.

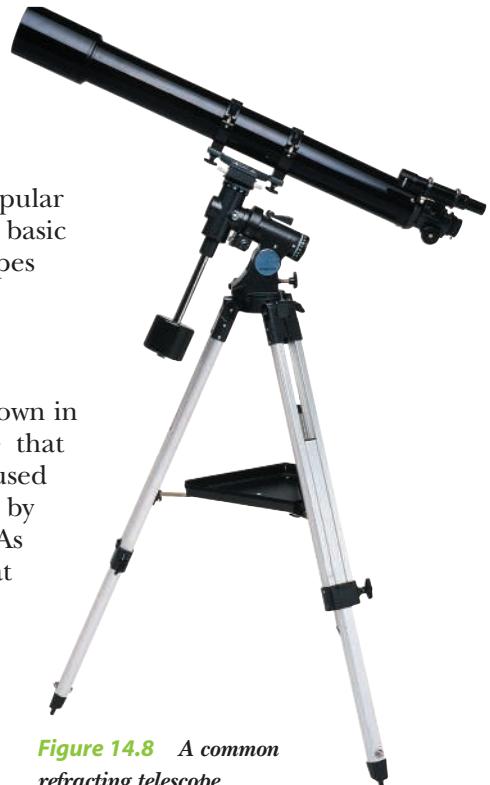


Figure 14.8 A common refracting telescope

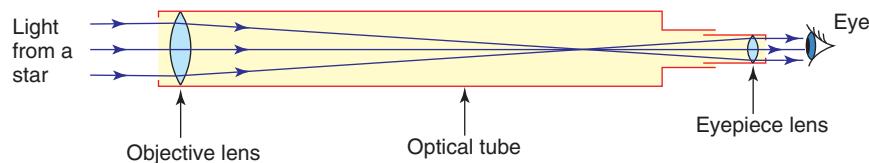


Figure 14.9 The arrangement of lenses inside an astronomical refracting telescope



Figure 14.10 A common Newtonian reflecting telescope

Refracting telescopes are preferred for planetary and lunar observations but not for observing stars because the lenses can introduce image errors (called aberrations), and because large lenses are expensive to manufacture accurately. In addition, the unobstructed light path of a refractor results in good image contrast, which is important when observing planets.

Reflecting telescopes

Figure 14.10 shows the type of reflecting telescope found in NSW high schools. This type of telescope uses a parabolic concave mirror to gather and focus the starlight by reflection of the rays. Figure 14.11, on the following page, shows a variety of common designs. The most basic design, shown in figure 14.11(a), is the prime focus. This is the design used by radio telescopes, with the signal coming from the detector in electronic form. For optical work, however, it is necessary to direct the light out of the telescope tube. School telescopes use the design shown in figure 14.11(b), known as a Newtonian reflector since Isaac Newton first suggested it. Larger research telescopes use the Cassegrain design shown in figure 14.11(c), which directs the light through a hole in the primary mirror. This design can be produced on a large scale far less expensively than similarly sized refracting telescopes.

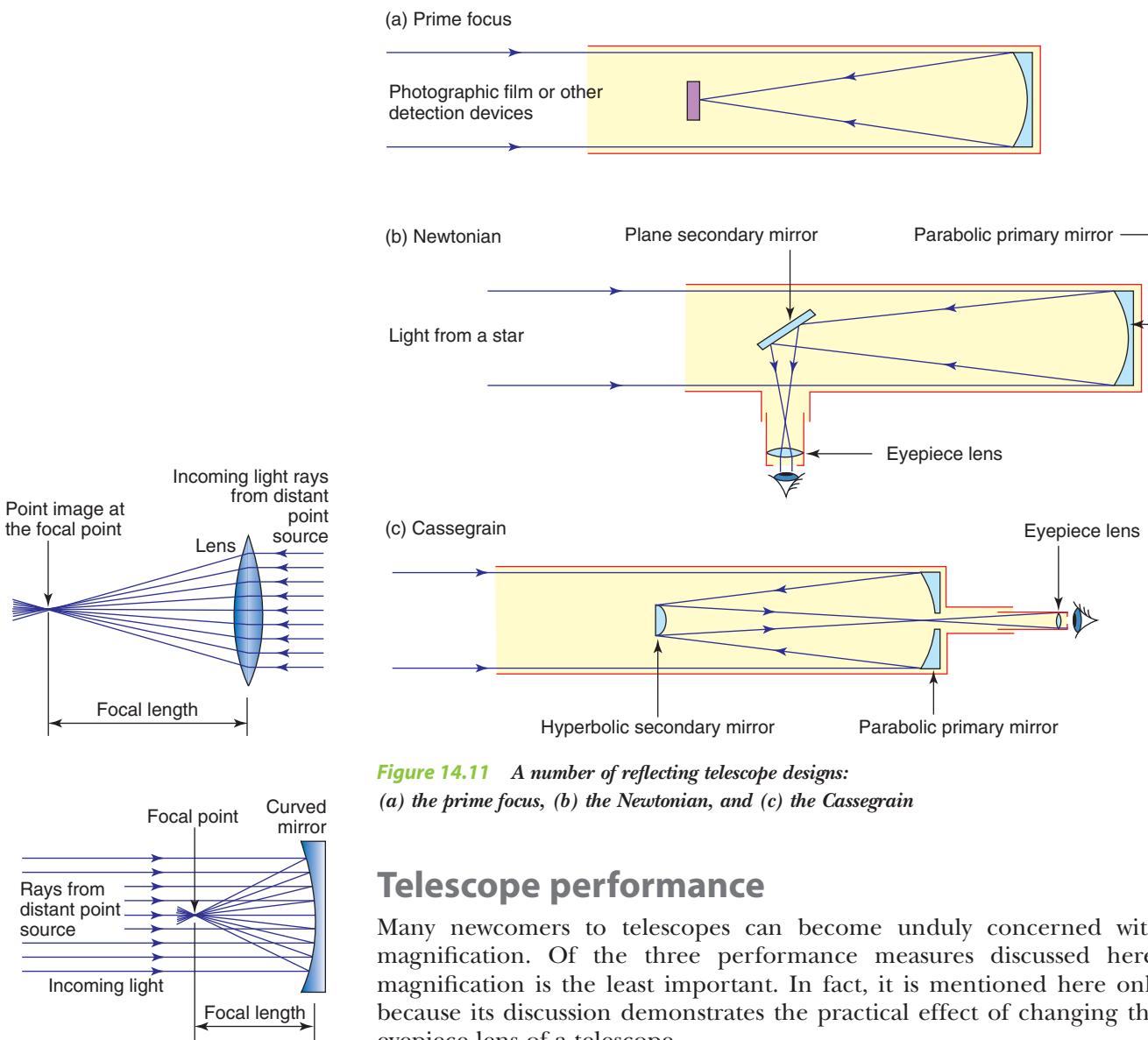


Figure 14.11 A number of reflecting telescope designs:
(a) the prime focus, (b) the Newtonian, and (c) the Cassegrain

Telescope performance

Many newcomers to telescopes can become unduly concerned with magnification. Of the three performance measures discussed here, magnification is the least important. In fact, it is mentioned here only because its discussion demonstrates the practical effect of changing the eyepiece lens of a telescope.

Any convex lens or concave mirror has a focal length, as shown in figure 14.12. A telescope has two focal lengths of concern — the focal length, f , of the telescope itself (that of the objective lens in a simple refractor, or that of the primary mirror in a simple reflector) and the focal length, f_e , of the telescope eyepiece. The magnification, m , of the telescope can be calculated using the expression:

$$m = \frac{f}{f_e}$$

Calculating magnification

A telescope has a focal length of 125 cm and it is fitted with an eyepiece with a focal length of 12.5 mm. Determine its magnification.

SOLUTION

$$\begin{aligned} m &= \frac{f}{f_e} \\ &= \frac{1.25 \text{ m}}{0.0125 \text{ m}} \\ &= 100 \times \text{magnification} \end{aligned}$$

SAMPLE PROBLEM**14.2****Changing the eyepiece**

The telescope in sample problem 14.1 is now fitted with a different eyepiece, this time with a focal length of 20 mm. What is the new magnification?

SOLUTION

$$\begin{aligned} m &= \frac{f}{f_e} \\ &= \frac{1.25 \text{ m}}{0.020 \text{ m}} \\ &= 62.5 \times \text{magnification} \end{aligned}$$

**14.1****Comparing the light-gathering ability of different sized lenses**

The **theoretical resolution** of a telescope is its ability to distinguish two close objects as separate images. It is measured as an angle.

The sensitivity of a telescope is its ability to pick up faint objects for observation, or its light-gathering power. This depends upon the collecting area of the lens or mirror, since a larger area means more light is being gathered and focused to form an image. The collecting area of the telescope depends upon its radius, or diameter (which is the dimension usually quoted). Therefore, a larger diameter telescope will usually mean a more sensitive one. A 100 mm school telescope, such as the one shown in figure 14.10, will be much less sensitive than the 3.9 m AAT shown in figure 14.5 simply because of the significant difference in diameter.

The **theoretical resolution** of a telescope is its ability to distinguish two close objects as separate images. It is measured as an angle and depends upon the wavelength of light, or other electromagnetic radiation being collected, as well as the diameter of the telescope. The following formula for resolution is sometimes called the ‘Dawes limit’:

$$R = \frac{2.1 \times 10^5 \lambda}{D}$$

where

R = resolution (arcsec, or seconds of arc)

λ = wavelength (m)

D = diameter (m).

Note that a smaller angle indicates a higher resolution.

SAMPLE PROBLEM**14.3****Theoretical resolution of the Parkes telescope**

What is the theoretical resolution of the 64 m Parkes radio telescope when observing radio waves of wavelength 3 cm?

SOLUTION

$$\begin{aligned} R &= \frac{2.1 \times 10^5 \lambda}{D} \\ &= \frac{2.1 \times 10^5 \times 0.03}{64} \\ &= 98 \text{ arcsec} \end{aligned}$$

SAMPLE PROBLEM**14.4****Theoretical resolution of a small telescope**

What is the theoretical resolution of a 100 mm Newtonian reflecting telescope when observing starlight with a wavelength of approximately 500 nm?

SOLUTION

$$\begin{aligned} R &= \frac{2.1 \times 10^5 \lambda}{D} \\ &= \frac{2.1 \times 10^5 \times 500 \times 10^{-9}}{0.100} \\ &= 1.05 \text{ arcsec} \end{aligned}$$

Theoretical resolution of the Anglo-Australian Telescope

Determine the theoretical resolution of the 3.9 m Anglo-Australian Telescope when observing starlight of wavelength 500 nm.

SOLUTION

$$\begin{aligned} R &= \frac{2.1 \times 10^5 \lambda}{D} \\ &= \frac{2.1 \times 10^5 \times 500 \times 10^{-9}}{3.9} \\ &= 0.027 \text{ arcsec} \end{aligned}$$

**14.2*****The Australian Telescope
Compact Array***

We can see from the sample problems above that a radio telescope, by the nature of the wavelengths it observes, is restricted to very poor resolutions. However, they usually have very large collecting areas and therefore can be very sensitive devices. Another factor contributing to their sensitivity is that radio signals can be amplified with very little increase in noise using electronic amplifiers. This cannot be done with light signals.

By comparison, when looking at the stars with just a 100 mm optical telescope you will be enjoying a far superior resolution of about 1 arcsec. However, this telescope is much less sensitive than a radio telescope. To look at the stars with increased sensitivity, we will need to move to a larger optical telescope such as the Anglo-Australian Telescope which, by virtue of its 3.9 m mirror, enjoys a much brighter field of view. Theoretically, it should also enjoy a much greater resolution than the small telescope. Ironically, however, it does not because of atmospheric blurring, or ‘seeing’.

PHYSICS IN FOCUS***The S-Cam***

The S-Cam (Superconducting Camera) is an example of technology pushing optical telescope sensitivity to its limits. It incorporates a cryogenic light sensor built using superconductors and cooled to just 1 K. It is able to register individual photons of light, very quickly recording their position

as well as directly measuring their colour. The information is accumulated in a database that allows the examination of very quick variations in light. Such variations are typical of some astronomical events, such as the optical explosions associated with gamma-ray bursts; these events could not be adequately studied with previous technology. The S-Cam is currently fitted to the 4.2 metre William Herschel Telescope, located at the Observatorio de Roque de los Muchachos on the island of La Palma in the Canary Islands.

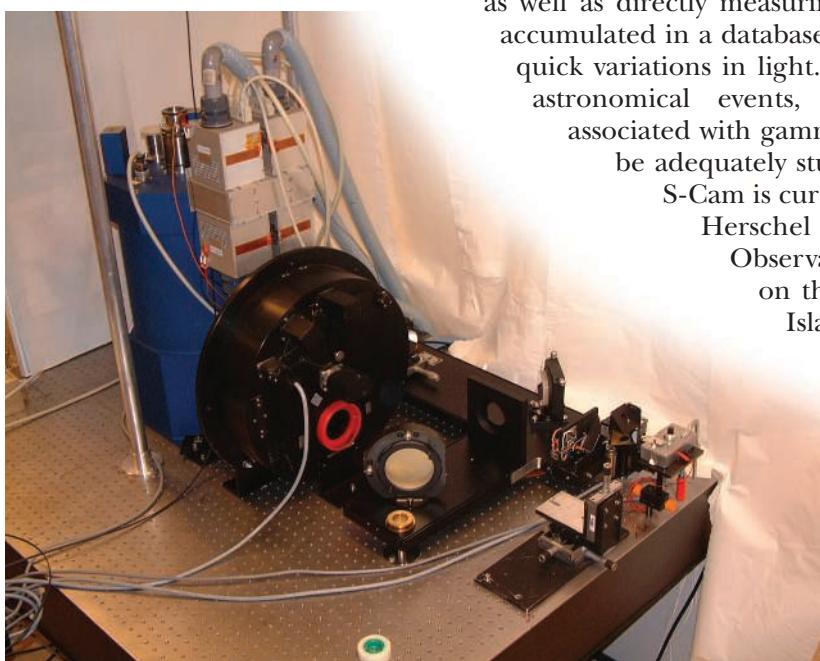


Figure 14.13 The S-Cam has been developed by the European Space Agency (ESA).

14.4 SEEING

‘**Seeing**’ refers to the twinkling and blurring of a star’s light due to atmospheric distortion.

If you look across a car park or along a road on a very hot day you will see ripples rise from the surface. You notice it because the moving hot air distorts the light passing through it. The same thing occurs to starlight entering the Earth’s atmosphere. Turbulent air distorts the path of the starlight through it, making the stars appear to twinkle, and blurs their image. This effect is known as ‘**seeing**’, and will normally blur the image of a star to about 1 arcsec. The best locations in the world for looking at stars, such as Mauna Kea, Hawaii, have a seeing of about 0.5 arcsec.

This imposes a practical limit on the achievable resolution from a large optical telescope. The Anglo-Australian Telescope, when opened in 1974, was restricted to a seeing of about 1 arcsec, despite its theoretical resolution of approximately 0.03 arcsec. Ironically, this is no better than a small 100 mm telescope with a theoretical resolution of 1 arcsec, although the AAT’s view is much brighter.

Radio telescopes are not affected as much by seeing, by virtue of the longer wavelengths they observe. There is some effect when observing wavelengths of a few millimetres — water vapour and oxygen in the atmosphere tend to absorb radio signals of this wavelength. In addition, rain can be a factor since raindrops are a few millimetres in size. However, wavelengths longer than this are not affected by atmospheric blurring.

There is one other obstacle to viewing that should be mentioned — the Sun. Obviously the Sun interferes with optical viewing, restricting optical astronomers to night viewing. Less obviously, the Sun is also a source of interference for radio astronomers since it is a strong radio source. This usually prevents radio telescope observations within 90° of the Sun, unless a particularly strong radio source is being viewed, such as certain quasars.

14.5 MODERN METHODS TO IMPROVE TELESCOPE PERFORMANCE

The capability limits of telescopes appeared to have been reached several decades ago. Radio telescopes are quite sensitive and are not bothered by seeing conditions but have quite poor resolution. Optical telescopes are expensive to manufacture in large diameters but they must be large to be sensitive. However, seeing limits their effective resolution to no better than a 200 mm telescope even at the very best locations in the world. Recently there have been many innovative approaches to overcoming these barriers to effective ground-based astronomy.

Interferometry

The resolution problem of radio telescopes can be overcome by using many radio dishes laid out in a large pattern, and then combining their signals together to make them behave as a single radio telescope with a much larger diameter.

This has been done in New Mexico, USA, to create the Very Large Array (VLA) shown in figure 14.14. The VLA is made up of 27 radio dishes set out in a large Y pattern up to 36 km across. Each dish is 25 m in diameter but, when combined electronically, they provide the resolution of a dish 36 km in diameter and the sensitivity of a dish 130 m in diameter.

Figure 14.14 The Very Large Array (VLA) in New Mexico. The dishes can be moved along tracks to give four different arrangements of 36 km across, 10 km across, 3.6 km across or just 1 km across.



We can use the resolution formula on the largest array size of 36 km when observing the shortest receivable wavelength of 7 mm, to calculate the maximum theoretical resolution:

$$\begin{aligned} R &= \frac{2.1 \times 10^5 \lambda}{D} \\ &= \frac{2.1 \times 10^5 \times 0.7 \text{ cm}}{3.6 \times 10^6 \text{ cm}} \\ &= 0.04 \text{ arcsec.} \end{aligned}$$

Interferometry is a technique used to combine the data from several elements of an antenna array in order to achieve a higher resolution.

eBookplus

Weblink:
The Square
Kilometre Array

This means that at 0.04 arcsec the VLA is challenging large optical telescopes for theoretical resolution and, in practical terms, is bettering them since it is not bothered as much by seeing.

The VLA is an interferometer, which means that it combines the data from each element of the array to form an interference pattern. Computers are used to mathematically analyse these patterns to reveal information about the structure of the radio source. This technique is known as **interferometry**.

Interferometry techniques have also been used to ‘unblur’ the images from large optical telescopes. ‘Speckle interferometry’ uses many images from a telescope, keeping each exposure short enough to freeze the atmospheric blur. A computer is then used to process the many exposures and extract more exact information about the star or other object.

PHYSICS IN FOCUS

The Square Kilometre Array (SKA)

The SKA is a new generation radio telescope array that may place Australia at the forefront of radio astronomy. An international project that may be located on Australian soil, the SKA will be an array of radio antennas linked to work as a radio telescope with an effective collecting area of one square kilometre. The antennas will be arranged into groups of about one hundred, each group forming an array

station. Approximately 80 of these stations will be arranged in a spiral pattern up to 400 km from a core array as shown in figure 14.15. There will also be several array stations located even further from the core in order to give the SKA a high resolution capability.

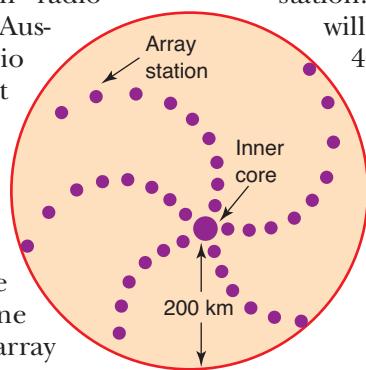


Figure 14.15 The planned layout of the SKA, a new generation radio telescope array

Active optics

The most recent developments for optical telescopes are active and adaptive optical systems. These systems seek to detect the errors in starlight caused by atmospheric blurring and then to optically correct them automatically. If done properly, the telescope operator should be aware only of improved seeing.

Active optics use a slow feedback system to correct sagging or other deformities in the primary mirror of large modern reflector telescopes.

Active optics use a slow feedback system to correct sagging or other deformities in the primary mirror of large modern reflector telescopes. In the past, large telescopes such as the AAT used primary mirrors with a thickness about one sixth of their diameter in order to ensure that they did not deform as the telescope was moved around the sky. However, there is a new generation of 8 to 10 m reflecting telescopes that use thin mirrors — just 20 cm thick approximately. These mirrors will certainly change shape as the telescope changes direction or heats up or cools down. However, the back of the mirror is fitted with many actuators that can push or pull the mirror back into the correct shape.

When the light leaves the primary mirror, but before it reaches the final lens (where the eyepiece is in a small telescope), it is slowly sampled by a ‘wavefront sensor’. This is a type of interferometer, which can detect how the incoming light has been altered. By sampling slowly, the effect of atmospheric turbulence is eliminated and any remaining effect is then due to deformities in the primary mirror. A computer calculates the required shape adjustments and then moves the actuators as required every few minutes.

Adaptive optics use a fast feedback system to attempt to correct for effects of atmospheric turbulence.

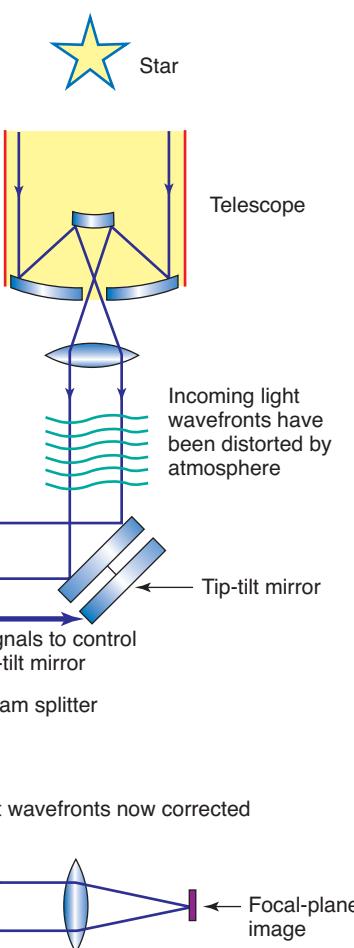


Figure 14.16
A typical adaptive optical system layout

The first telescope to use active optics was the 3.5 m New Technology Telescope in Chile, which uses 75 actuators under its primary mirror. The most notable use of active optics is in the new 10 m telescopes Keck I and Keck II in Hawaii (see page 269). The enormous mirrors in these telescopes are each made up of 36 separate pieces of appropriate shapes. The position and curvature of each piece is controlled by the active system and adjusted twice per second.

Adaptive optics

Adaptive optics use a more aggressive approach in an attempt to correct effects of atmospheric turbulence. A wavefront sensor is still employed between the primary mirror and the lens, as shown in figure 14.16. This time, however, rapid computer-calculated corrections are fed to one or two secondary mirrors that ‘straighten out’ the light. These corrections are made at up to 1000 times per second, and this speed is the major difference between adaptive and active systems. Figure 14.17, on the following page, shows how adaptive optics allows a binary star to be seen more clearly.

One of the possible secondary mirrors is called a ‘tip-tilt’ mirror, which is able to adjust for slight changes in the position of the light. (A tip-tilt system is used in the Anglo-Australia Telescope.) The other is a deliberately deformable mirror to adjust for deformities in the light. Making this type of image correction presents a considerable technological challenge and some development is still required before many large telescopes can successfully adopt adaptive optics.

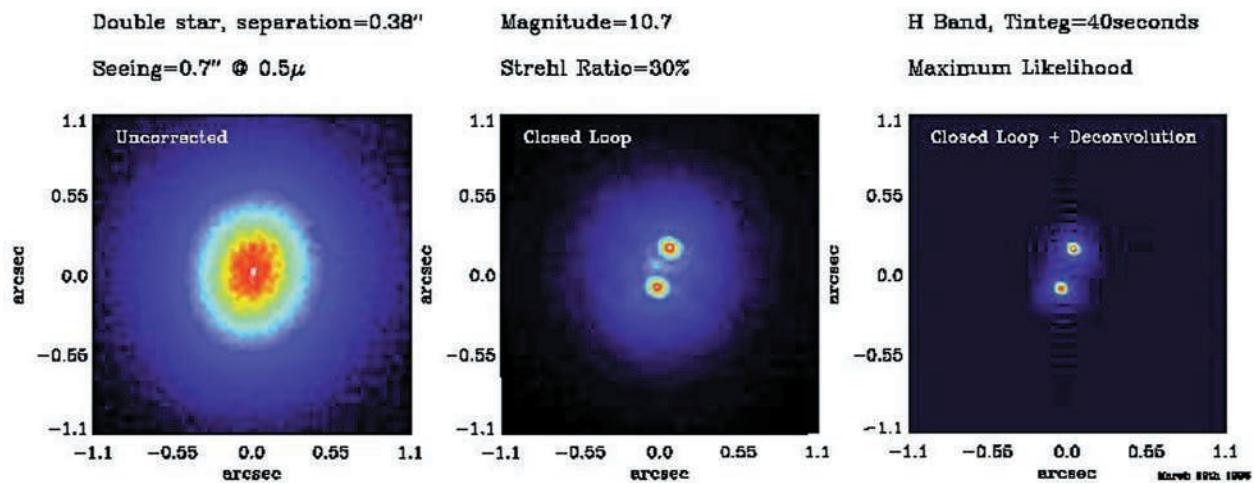


Figure 14.17 Adaptive optics used to improve the seeing on the Canada–France–Hawaii Telescope (CFHT) at Mauna Kea Observatory, Hawaii. On the left can be seen a raw image of an unresolved close binary (double) star system. The two stars are 0.38 arcsec apart but seeing is approximately 0.7 arcsec. In the centre, the adaptive optics have been turned on and the two stars can now be seen quite clearly. In the third image, computer enhancement has been added. © CFHT, 1996. Used with permission.

PHYSICS IN FOCUS

Advanced telescope technology

NASA’s ‘Great Observatories Program’ has worked to place four telescopes in space to cover the whole electromagnetic spectrum, including regions not observable from the ground. Being above the atmosphere also eliminates atmospheric blurring of images. The four telescopes are described below.

1. The Hubble Space Telescope (HST) was put into orbit in 1990. It needed repairs soon after but has since demonstrated the remarkable clarity possible with no seeing to blur its images. HST detects visible and ultraviolet light, so that its view of the universe is much as we see it.
2. The Compton Gamma Ray Observatory (GRO), which detected high-energy gamma rays. GRO was put into orbit in 1991 but was brought down on 4 June 2000 after failure of one of its gyroscopes six months earlier, and crashed into the Pacific Ocean. Gamma rays are produced by high-energy processes. This meant that GRO could look at events such as solar flares, supernovae and hot, spinning

disks of matter around black holes. These gamma rays are not visible to us so the view that GRO could see was a view of the world foreign to our eyes.

3. The Chandra X-Ray Observatory (formerly called AXAF) was named after Nobel Prize winner Subrahmanyan Chandrasekhar. It was launched into orbit in 1999 to observe X-rays that, although less energetic than gamma rays, are produced by similar events — the hot matter associated with objects such as supernovae, quasars or black holes.
4. The Spitzer Space Telescope (formerly called SIRTF) was launched in 2003 for what became a five-year mission. Spitzer observes the infrared light produced by cool objects such as nebulae discs in which stars are born, and discs of dust or planets around other stars. To observe these, Spitzer was placed into an Earth-trailing orbit around the Sun, where its cryogenically cooled instruments suffer less heating than if it were in an Earth orbit.

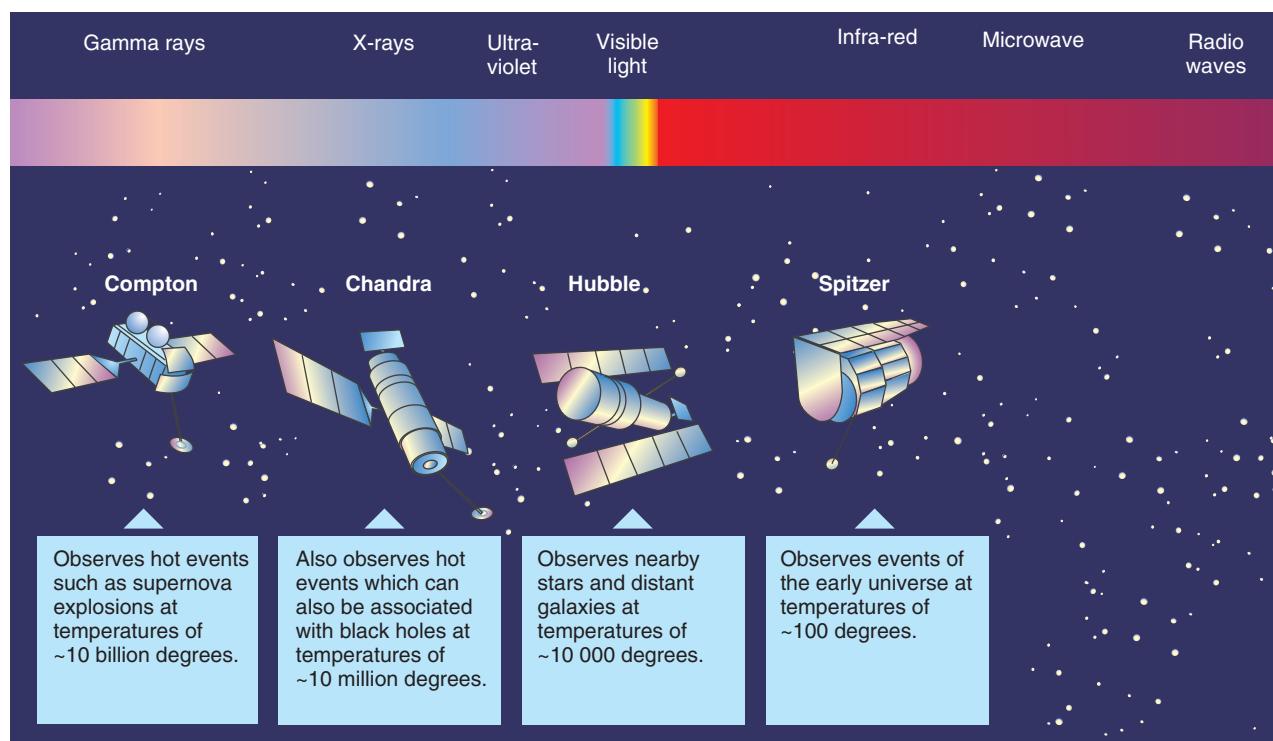


Figure 14.18 NASA's space observatories span the electromagnetic spectrum. Compton and Chandra observe hot events such as supernovae and black holes. Hubble observes nearby stars and distant galaxies. Spitzer observes the birth places of planets, stars and even whole galaxies.

Since the 1990s, many advanced ground-based telescope facilities have been initiated, such as the pair of telescopes known as Keck I and Keck II, located on Mauna Kea in Hawaii (shown in figure 14.19). Each has a 10 m primary mirror, easily the largest in the world. The mirrors themselves are extraordinary, being made up of 36 smaller hexagonal mirrors 1.8 m in diameter. Each has the appropriate shape ground into it, and each is held in place by a computer-controlled active optical system so that they act together as a single near-perfect mirror.

A 10 m mirror has 17 times the light-gathering area of the Hubble Space Telescope. Being free of the atmosphere, the HST can see finer detail, but the Keck telescopes are more sensitive so they can see fainter objects. A common approach of research teams is to first use the HST to find and pinpoint distant objects and then to use the Kecks to explore those objects.

Mauna Kea, a dormant volcano in Hawaii, offers some of the best seeing conditions in the world. Surrounded by thermally stable ocean, the surrounding air is unusually still. Despite this, the Kecks have an adaptive optical system, which eliminates even the seeing from this site. In addition to

this, the two telescopes were linked by interferometry in 2001 to deliver the resolving power of a telescope with a mirror 85 m in diameter!

As remarkable as the Keck telescopes are, there are plans to apply this same technology to telescopes with mirrors of more than 20 m diameter. These are known as Extremely Large Telescopes (ELTs). Two examples that are currently being planned are the European Extremely Large Telescope and the Thirty Metre Telescope.



Figure 14.19 Telescopes Keck I and Keck II on Mauna Kea, Hawaii

SUMMARY

- Galileo was the first person to point a telescope at the sky and the object of his first studies was the Moon. He was able to observe the texture of the surface, commenting on the craters and even measuring the height of a mountain.
- All of the various electromagnetic wavelengths impinge upon the upper atmosphere, yet only visible light, some infra-red, microwaves and radio waves successfully penetrate through to the ground; the other wavelengths are absorbed by the atmosphere.
- Ground-based telescopes are therefore restricted to detecting visible light and radio waves.
- Space-based telescopes are used to observe all electromagnetic wave bands without the restrictions of ground-based telescopes.
- Two basic designs of telescopes are refractors (which use lenses) and reflectors (which use mirrors).
- The sensitivity of a telescope is its ability to pick up faint objects for observation, or its light-gathering power. This depends upon the collecting area of the lens or mirror, and hence the square of its diameter.
- The theoretical resolution of a telescope is its ability to distinguish two close objects as separate images.

$$R = \frac{2.1 \times 10^5 \lambda}{D}$$

- Seeing refers to atmospheric blurring of a star's light. It is caused by turbulence within the Earth's atmosphere and severely restricts the practical resolution of large optical telescopes.
- Radio telescopes, which are normally quite sensitive, can improve their resolution by being connected into an array and using interferometry to give an effectively much larger dish diameter.
- Large ground-based optical telescopes are combating atmospheric blurring by introducing computer-controlled adaptive optics to smooth out the twinkling of stars and unblur their images.
- The latest space telescopes include the Hubble Space Telescope, the Compton Gamma Ray Observatory, the Chandra X-ray Observatory and the Spitzer Space Telescope. A new generation of 8 m ground-based telescopes includes the twin 10 m Keck telescopes in Hawaii.

QUESTIONS

- State the years in which Galileo built and began to use his telescopes.
- Identify the initial object of his observations and what he learnt from those observations.
- (a) List the components of the electromagnetic spectrum in order from least to most energetic. Which has the highest frequency?
 (b) Identify the components that are absorbed by the atmosphere as they attempt to penetrate it from space.
 (c) Discuss strategies that can be employed to systematically observe these radiations.
- State which components of the electromagnetic spectrum can be observed by ground-based telescopes.
- Identify the two basic designs used in telescopes. Draw a sketch of each.
- Define the sensitivity of a telescope.
- Define the resolution of a telescope.
- Define the magnification of a telescope.
- Calculate the theoretical resolution of the following optical telescopes when observing light of wavelength 500 nm:
 (a) a 50 mm refracting telescope
 (b) a 50 mm reflecting telescope
 (c) a 100 mm Newtonian reflector
 (d) a 200 mm Cassegrain reflector
 (e) a 3 m Schmidt telescope
 (f) an 8 m Schmidt telescope.
- Calculate the resolution of a 200 mm optical telescope when viewing light of wavelength:
 (a) 500 nm
 (b) 600 nm
 (c) 700 nm.
- Calculate the resolution of the following radio telescopes when observing 3 cm radio waves:
 (a) a 50 mm dish (why is this impractical?)
 (b) a 10 m dish
 (c) a 30 m dish
 (d) a 70 m dish
 (e) a 200 m array
 (f) a 1 km array
 (g) a 15 km array.
- Calculate the resolution of a 50 m radio telescope when observing the following wavelengths:
 (a) 1 mm (b) 1 m (c) 1 km.

13. Calculate the magnification and resolution of a 300 mm Cassegrain telescope used to observe light of wavelength 650 nm. The telescope has a focal length of 1000 mm and is fitted with:
 - (a) a 25 mm eyepiece
 - (b) a 10 mm eyepiece.
14. The Very Large Array in New Mexico, USA, is a set of 27 radio antennas linked by interferometry to give the resolution of a single dish of diameter 36 km and the sensitivity of a dish of diameter 130 m.
 - (a) With reference to the sensitivity, calculate the effective collecting area of the array.
 - (b) Calculate the theoretical resolution of the VLA when observing radio waves of wavelength 10 cm.
15. The Square Kilometre Array, planned to be constructed in Australia, will be an array of up to 1000 radio antennas with an effective collecting area of 1 km^2 .
 - (a) Calculate the diameter a single dish would need in order to have this collecting area.
 - (b) Referring to figure 14.14, if the outlying array stations were 400 km from the centre of the array, what would be the theoretical resolution of the SKA when observing radio waves of wavelength 10 cm?
16. Radio telescope facilities across Australia have formed a network called the Australian Long Baseline Array (LBA). In this system, selected telescopes are used to observe the same object at much the same time; however, these observations are made independently of each other. Data from each telescope are transported to a correlator in Sydney and the interferometry is then applied. Although results are delayed, very long baselines can be used to give extraordinary resolutions.
17. If the Parkes radio telescope and the 15 m radio telescope at Perth are used, then a baseline of approximately 3000 km is achieved. Calculate the theoretical resolution for such an observation.
18. Another cooperating telescope is at Haartebeesthoek in South Africa. If this is used with Narrabri then a baseline of 9853 km is achieved. Calculate the effective theoretical resolution.
19. The network provides a variety of baselines down to 113 km, between Narrabri and Mopra. Discuss reasons why a radio astronomer would deliberately choose to use a smaller baseline than the maximum available.
20. Describe the condition known as ‘seeing’ and how it is caused.
21. Discuss how ‘seeing’ severely restricts the capabilities of large ground-based telescopes.
22. Discuss strategies currently being employed in modern, large, ground-based telescopes to counter the effects of ‘seeing’.
23. Compare the resolution and sensitivity of a typical radio telescope to that of a large optical telescope.
24. Outline strategies being employed in modern radio telescope facilities to improve their resolution.
25. Compare the resolution and sensitivity of a radio telescope array such as the VLA or SKA to that of a large optical telescope.
26. Describe NASA’s ‘Great Observatories Program’. What does it hope to achieve?
27. Describe the twin 10 m telescopes Keck I and Keck II.
28. Compare the capabilities of the Hubble Space Telescope and the Keck I telescope.



14.1 COMPARING THE LIGHT-GATHERING ABILITY OF DIFFERENT SIZED LENSES

Aim

To demonstrate the extra light-gathering ability of larger lenses.

Apparatus

at least two biconvex lenses of different diameter
light meter
two polarising filters
one sunny day

Warning!

Do not use the lenses to look at the Sun. Blindness can result.

Method

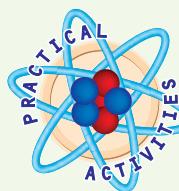
1. Set up the light meter in the sunlight. The needle on the meter may already indicate a maximum value. If this is the case, place both polarising filters, one upon the other, over the sensor panel and then rotate the upper filter only until the needle on the meter slides back to a central reading.
2. Starting with the smallest lens, measure its diameter and record this information. Use the lens to gather the Sun's light onto the panel of the light meter. Do not focus the light to a point, but rather create a circle of light that fills the panel of the light meter. Record the reading from the light meter.
3. Repeat this method with each lens of different diameter, in ascending order.

Results

LENS DIAMETER (cm)	LENS AREA (m^2)	LIGHT METER READING

Questions

1. Based on the information in your table, what can be said about the reading on the light meter as lens diameter is increased?
2. What does this observation tell us about lenses of larger diameter?
3. What are the implications of this finding for telescopes?



14.2 THE AUSTRALIAN TELESCOPE COMPACT ARRAY

Aim

To use the internet to find out more about radio astronomy at the Australian Telescope Compact Array.

Apparatus

Internet access is all that is required for this activity.

Method

The Australian Telescope Compact Array (ATCA) is a radio telescope array of six 22 m antennas located at Culgoora, NSW. It is operated by the Australian Telescope National Facility (ATNF), which is part of the CSIRO, and also operates several other installations such as the Parkes radio telescope.

eBookplus

Weblink:
The Australian Telescope
National Facility

You will be presented with access to each of the telescope facilities operated by the ATNF. Investigate each of the options, and when you are ready, select 'ATCA (Narrabri)'. When the page loads, select 'information for the public', then select the 'ATCA Live!' option. Alternatively, if time is short, you can go directly to this page using the following link:

eBookplus

Weblink:
ATCA Live!

Use the information on this web page to answer the following questions.

Questions

1. Write down the date and time at which you are completing this exercise.
2. What are the current weather conditions at Culgoora?
3. Which antenna(s), if any, are currently offline?
4. (a) What object(s) are the other antennas tracking?
(b) What is the right ascension and declination of this object?
(c) Consult a star map. Within which constellation does this object lie?
(d) What is the closest star, of magnitude 6 (approximate naked-eye limit) or brighter to this object?
5. (a) What frequencies are being observed?
(b) To what radio bands do these frequencies correspond? (Refer to figure 14.4.)
6. How have the antennas been configured? Draw their configuration in your practical book.
7. Notice that the telescopes are arranged along an east–west line. This gives good resolution in this direction but poor resolution in the north–south direction. How does the ATCA overcome this difficulty? An explanation can be found at the following link.

eBook plus

Weblink:
[Virtual radio interferometry](#)

CHAPTER 15

ASTRONOMICAL MEASUREMENT



Figure 15.1 Photographic astronomer David Malin at the prime focus of the Anglo-Australian Telescope

Remember

Before beginning this chapter, you should be able to:

- describe the dimensions of an ellipse
- describe the concept of a black body
- explain the Bohr model of the atom
- describe star groups such as giants, main sequence, and white dwarfs
- describe the Hertzsprung–Russell diagram and the star groupings found within it, in particular main sequence stars, red giants and white dwarfs.

Key content

At the end of this chapter you should be able to:

- define the terms parallax, parsec and light-year
- explain how trigonometric parallax can be used to determine the distance to stars
- calculate the distance to a star given its trigonometric parallax, using $d = \frac{1}{p}$
- discuss the relative limitations with trigonometric parallax measurements
- compare the relative limits of ground-based astrometric measurements to space-based measurements
- account for the production of emission and absorption spectra, and compare these with a continuous black body spectrum
- describe the technology needed to measure astronomical spectra
- identify the general types of spectra — continuous, emission and absorption, and identify astronomical objects that produce each
- describe the key features of stellar spectra and explain how these are used to classify stars
- describe how stellar spectra can provide a variety of information such as chemical composition, surface temperature, rotational and translational velocity, and density
- define apparent and absolute magnitude
- calculate brightness ratios using $\frac{I_A}{I_B} = 10^{\frac{(m_B - m_A)}{5}}$
- determine the distance to a star using the distance modulus $M = m - 5 \log\left(\frac{d}{10}\right)$
- outline the method of distance approximation called spectroscopic parallax
- explain how colour index is obtained and why it is useful
- describe the advantages of photoelectric technologies over photographic methods for photometry.

15.1 ASTROMETRY

Astrometry is the careful measurement of a celestial object's position, and changes of position, to a high order of accuracy.

Astrometry is positional astronomy; the branch of astronomy concerned with the careful measurement of position, and changes of position, of a star or other celestial object, to a high order of accuracy. These apparent position changes can be due to the real motion of the body, or the motion of the Earth around its orbit, representing a shifting point of observation. This latter case is of particular interest here because it allows a measurement of distance, and the technique involved is the focus of this section.

Parallax

Try this little experiment. Hold up your index finger so that it is vertical and about 10 cm in front of your face. Close your left eye and note the position of your finger against the background. Now open your left eye and close the right. You will notice that the position of your finger has apparently shifted against the background. This effect is known as parallax.

Parallax is the apparent change in position of a nearby object as seen against a distant background due to a change in position of the observer.

Parallax is the apparent shift in position of a close object against a distant background due to a change in position of the observer.

Trigonometric parallax is a method of using trigonometry to solve the triangle formed by parallax to determine distance.

Trigonometric parallax

Trigonometric parallax is a method of determining distances by using triangulation together with parallax. The method is used by surveyors to determine terrestrial distances, and is used by astronomers to determine distances to certain nearby stars.

As shown in figure 15.2, if the length of the baseline (formed by the motion of the observer) is known, and the angle of deviation, θ , measurable, then the distance to the object can be calculated using trigonometry.

$$\tan \theta = \frac{\text{baseline}}{\text{distance}}$$

$$\text{so distance} = \frac{\text{baseline}}{\tan \theta}$$

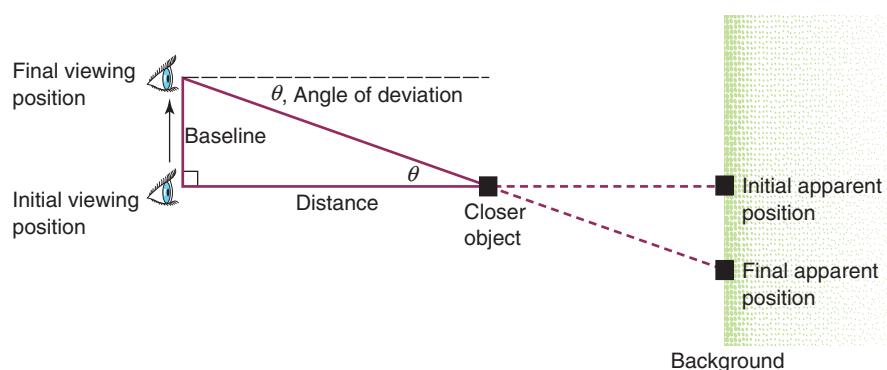


Figure 15.2 Parallax allows distance to be calculated because a triangle is formed and trigonometry can be applied.

For astronomical purposes, our viewing position of the heavens changes regularly in several ways. The rotation of the Earth results in a diurnal parallax, in which the viewing position changes over the course of one evening. In this case the length of the baseline, that is, the diameter of the Earth, is so small that this method can really be used to determine distances only within our own solar system.

The motion of the Earth in its orbit around the Sun results in an annual parallax, and this provides a larger and much more useful baseline.

Annual parallax

Observations of a nearby star during the course of a year will show apparent shifts in its position against the background of more distant stars. The largest shift will be between observations made six months apart.

Annual parallax, p , is half the angle through which a nearby star appears to shift against the backdrop of distant stars, over a particular six-month period.

The **annual parallax, p** , of a star is half the angle through which the star appears to shift as the Earth moves from one side of its orbit to the other. It can be seen from the triangle formed in figure 15.3 that:

$$\text{Distance of star from Earth} = \frac{\text{radius of Earth's orbit}}{\sin p}.$$

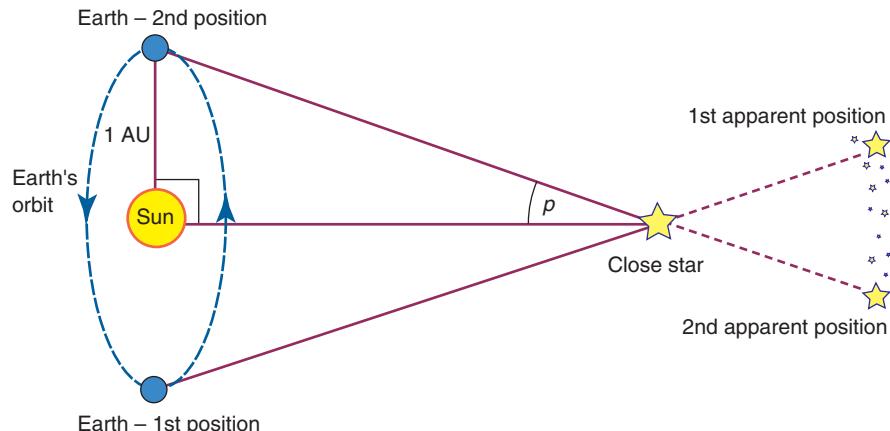


Figure 15.3 The formation of the annual parallax angle, p

The largest annual parallax observed belongs to Proxima Centauri and is 0.772 seconds of arc, as measured by the HIPPARCOS satellite (see page 278). With angles this small we can make the following approximation:

$$\sin \theta = \tan \theta = \theta.$$

In addition to this, if the radius of the Earth's orbit is expressed as 1 AU (Astronomical Unit), then the above formula for distance can be stated as:

$$d = \frac{1}{p}$$

where

d = distance from Earth (parsecs, pc)

p = parallax (arcsec).

Hence, distance is inversely proportional to parallax. The greater a star's annual parallax the closer it must be to us, and vice versa. Note that this expression has led to the definition of a new unit of astronomical distance known as the parsec.

The parsec (parallax-second)

One **parsec** (pc) is the distance from the Earth to a point that has an annual parallax of one arcsecond, as shown in figure 15.4.

In fact, no star is as close as 1 pc to us. As already mentioned, the closest star, Proxima Centauri, has an annual parallax of 0.772 arcsec, which places it at a distance of 1.29 pc.

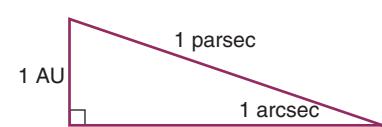


Figure 15.4 When $p = 1 \text{ arcsec}$, then $d = 1 \text{ parsec}$.

One **light-year** is the distance travelled through space in one year by light or other electromagnetic wave. It corresponds to a distance of 0.3066 parsecs or 9.4605×10^{12} km.

SAMPLE PROBLEM

15.1

SOLUTION

$$\begin{aligned} d &= \frac{1}{p} \\ &= \frac{1}{0.286} \\ &= 3.50 \text{ pc} \\ &= 3.50 \times 3.26 = 11.4 \text{ light-years} \end{aligned}$$

The parallactic ellipse

It was stated earlier that annual parallax was measured between two observations of a star made six months apart, and this was based on geometry such as that shown in figure 15.3. However, not just any six-month period will do. It must be the right six-month period.

As shown in figure 15.5, if the position of a nearby star is determined frequently throughout a year, then it will appear to trace out an ellipse, called the parallactic ellipse. The ellipse is due to the fact that most stars are at some angle to the plane of the Earth's orbit, measured from the Sun. (If they were located perpendicularly, as shown in figure 15.3, then the star would trace out a circle.)

The annual parallax is more rightly defined as the angle subtended by the semi-major axis of the parallactic ellipse. It occurs only twice in the course of a year, and that is when the Earth-Sun-star angle is 90 degrees. These two events are six months apart, and this is the specific six-month period that must be used to determine the angle.

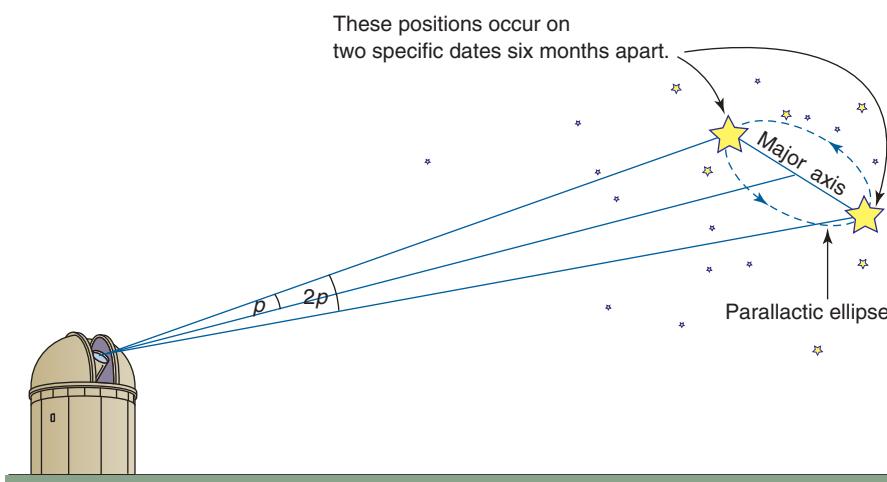


Figure 15.5 The annual parallax is half the angle subtended by the major axis of the parallactic ellipse. This ellipse is the apparent path traced out by a nearby star over the course of a year.

For the purpose of comparison between various length units, note that:

$$\begin{aligned} 1 \text{ parsec} &= 30.857 \times 10^{12} \text{ km} \\ &= 206\,265 \text{ AU} \\ &= 3.2616 \text{ light-years} \end{aligned}$$

where

$$\begin{aligned} 1 \text{ light-year} &= \text{the distance travelled through space in one year by light} \\ &= 9.4605 \times 10^{12} \text{ km} \\ &= 0.3066 \text{ parsecs.} \end{aligned}$$

Calculating distance using annual parallax

Determine the distance, in pc and light-years, of Procyon with an annual parallax of 0.286 arcsec.

Limitations

The very small angular measurements described on the previous page have traditionally been made photographically using large ground-based optical telescopes. However, such observations are affected by atmospheric blurring or ‘seeing’ (as described in chapter 14, page 265) so that the smallest parallax that can be measured from the ground is approximately 0.03 arcsec. This corresponds to a maximum distance, measurable to reasonable accuracy, of approximately 30 pc. In astronomical terms this is only a very short distance; however, this astrometric technique has provided a foundation of measurements upon which other techniques have built.

PHYSICS IN FOCUS

Astrometric satellites

The perfect way to overcome atmospheric blurring is to get above the atmosphere and make observations from space. Before NASA launched its Hubble Space Telescope, the European Space Agency (ESA) put into orbit a 290-mm astrometric telescope aboard a satellite called HIPPARCOS. Between 1989 and 1993, it was able to measure the parallax of approximately 120 000 stars to a precision of 0.001 arcsec. This is over 10 times more precise than ground-based measurements, and extends the maximum distance determined by astrometric means to 1000 pc. Its results are available in the ‘HIPPARCOS Catalogue’, which can be accessed on the internet.

ESA’s planned next-generation astrometric satellite has been dubbed ‘Gaia’. It is intended that Gaia will measure star positions and parallaxes to a precision of 10 microarcsec, which is

100 times more precise than that achieved by HIPPARCOS. This will allow it to determine star distances right across our galaxy to a good accuracy (10–20 per cent), provided that a star is bright enough to be measured. Approximately one billion such stars will be logged, which represents approximately one per cent of the stars in our galaxy, the Milky Way.



15.1

Accessing star data



15.2

Annual parallax precision

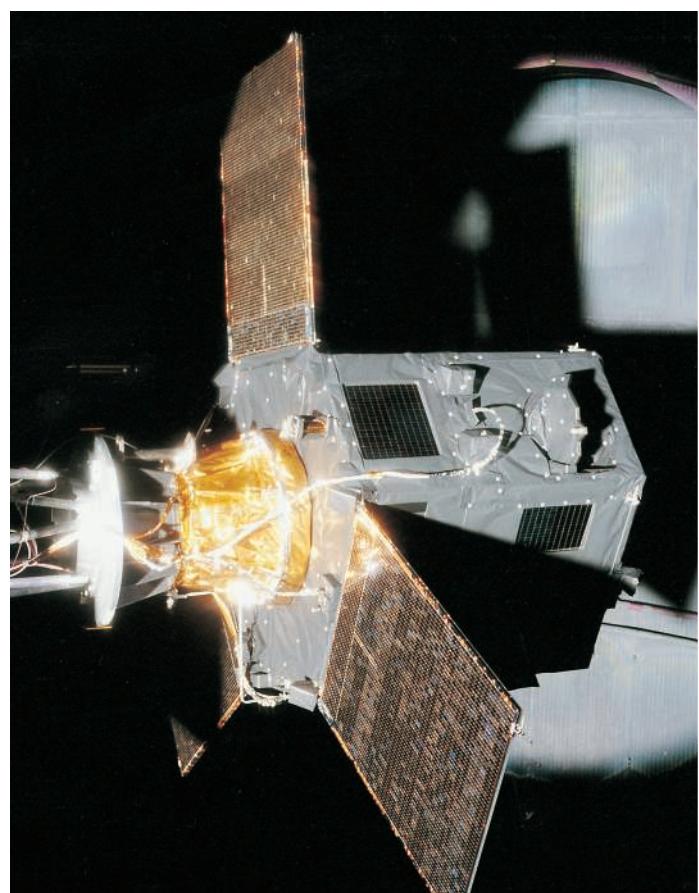


Figure 15.6 The HIPPARCOS satellite

15.2 SPECTROSCOPY

Consider the following three observations:

- looking at a torch light which has been covered over with yellow plastic
- looking at the yellow flame produced by a spray of sodium chloride solution into a Bunsen burner flame
- looking at a reflection of the yellow light produced by the Sun.

In each case you are observing yellow light, but the composition of that light is different. The difference is not apparent to your naked eye, however, and in order to discover this difference you will need to use a device known as a spectroscope. You will then be able to examine the components of the light and draw many inferences about the material that produced it. This is the field of spectroscopy and, by using it, astronomers have been able to learn a great deal about the observable objects in the universe.

Making spectra

A **spectroscope** is a device used to spread a light into its spectrum. It can be attached to the eyepiece of a telescope to examine the spectra of starlight.

We must first be aware that most light is a mixture of wavelengths or colours. If we were able to spread out or disperse a light ray, we would be able to observe the spectrum of colours within that light. This is what a **spectroscope** does and it can be attached to the eyepiece of a telescope to examine the spectra of starlight. It is made up of several elements working together, as shown in figure 15.7 below. There must first be a light source and this will be followed by several slits to form the light into a flat, vertical beam. The light then enters either a triangular prism or a diffraction grating, both of which have the ability to disperse light out into its spectrum. Because the light is in the form of a flat beam, the spectrum spreads out as a rectangular strip. The spectrum can then be recorded on a photographic plate or examined in more detail with a small telescope.

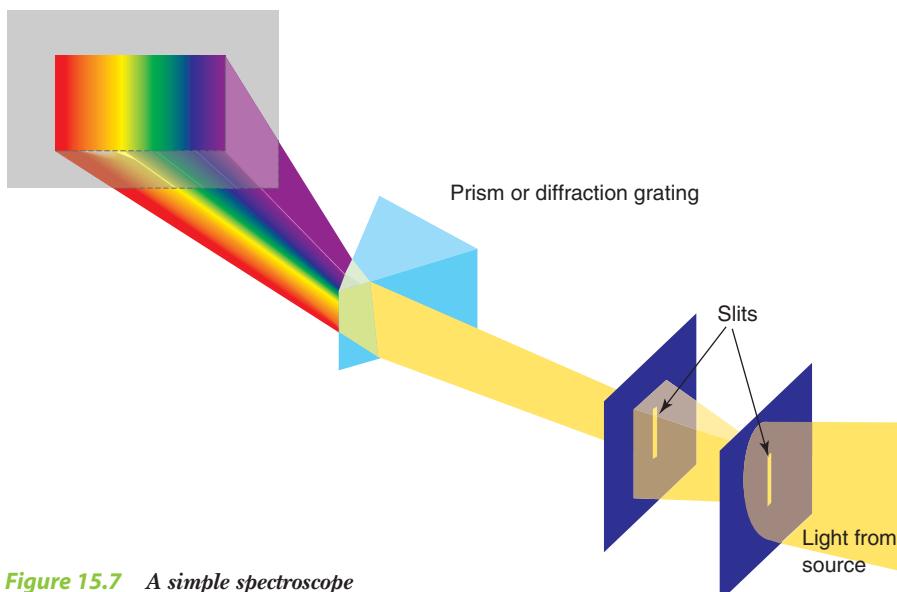


Figure 15.7 A simple spectroscope

If a photographic plate is used, then spectra such as those shown in figure 15.8(a) on page 280 will result. If a small telescope is used, then an electronic sensor, called a photometer, can be attached and used to sense the intensity of each wavelength. This can produce a spectrum in the form of a graph, such as that shown in figure 15.8(b) on page 280. The device is then known as a spectrophotometer.

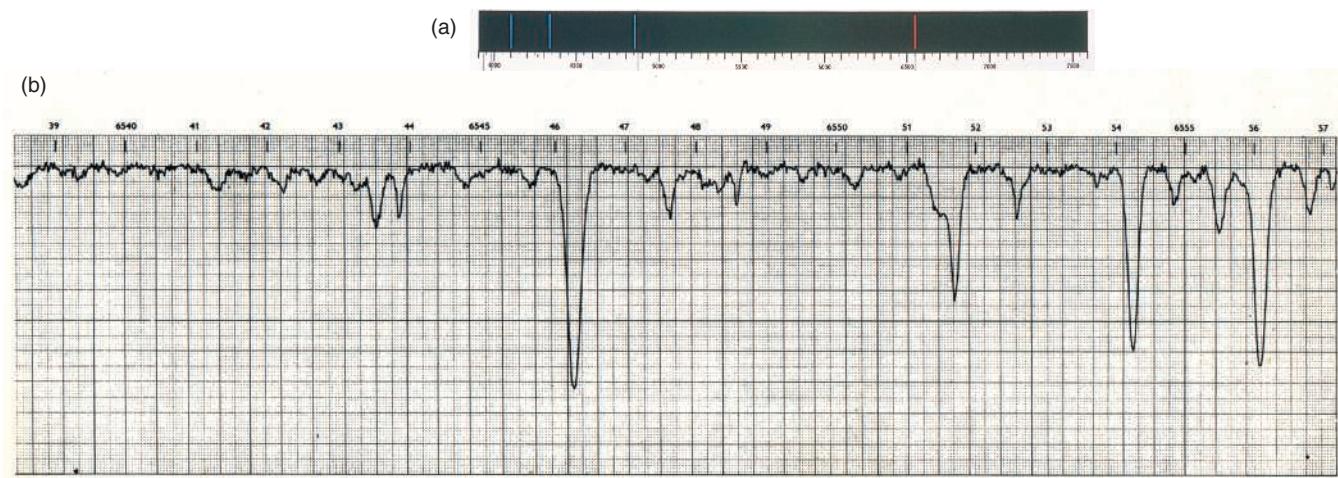


Figure 15.8 (a) An example of a visible line spectrum of hydrogen atoms produced using a photographic plate in a spectroscope (b) Using an electronic sensor can produce an intensity graph.

It is understandable that spectroscopy was first based on the visible spectrum of light, these wavelengths being most apparent to us and most easily observed. With the use of appropriate electronic sensors, however, spectroscopy is no longer restricted just to the electromagnetic radiation that we can see.

PHYSICS IN FOCUS

The S-Cam: Spectrophotometer in a chip

On page 264 we discussed the S-Cam, a new CCD (video camera) for electronic astronomical observations that greatly increases the sensitivity of an optical telescope. Although it is still in development, this cryogenic superconducting camera has the ability to record the position and colour of individual photons of light as they are received. All this information is quickly compiled into a database by a computer. This also gives the S-Cam the ability to function as an extremely accurate spectrophotometer, making observations quickly and simply without the need for intervening filters or prisms that would normally reduce sensitivity.

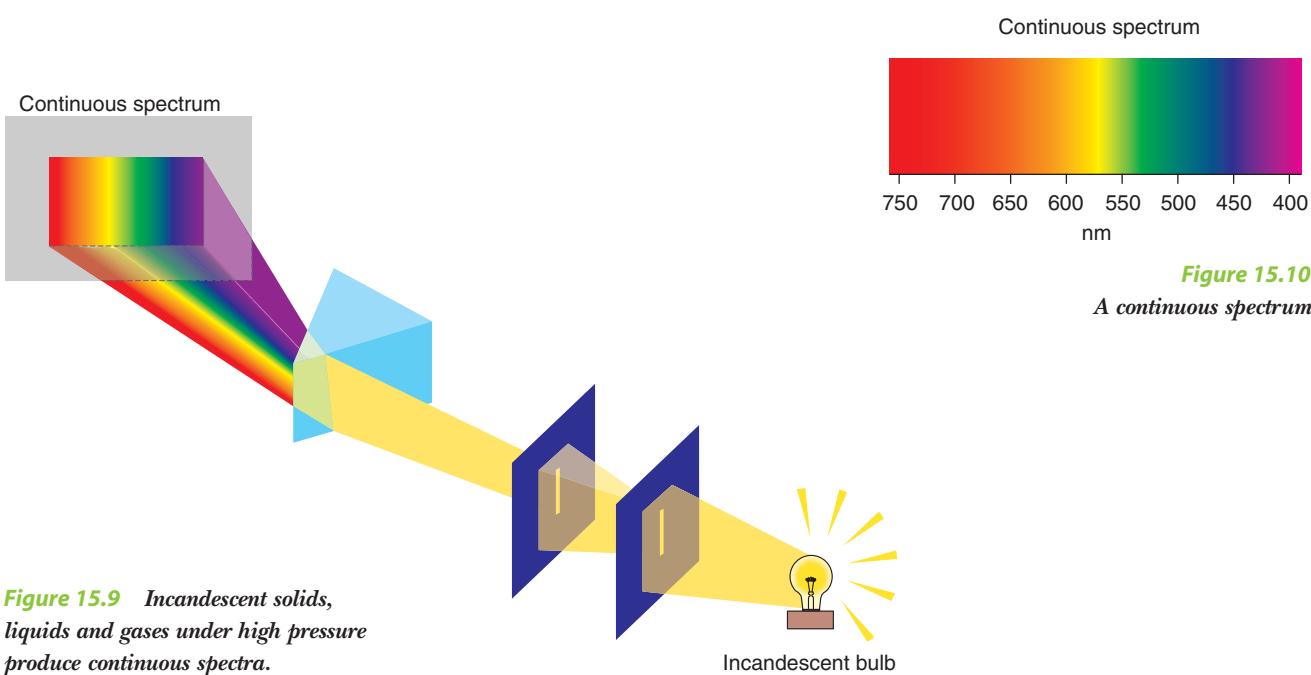
Types of spectra

At the beginning of this section there were three different examples of light sources mentioned: a light globe, a vapour and a star. Each produces a different type of spectrum — continuous, emission and absorption.

Continuous spectra

If the light source for a spectroscope is a hot, glowing solid, liquid or high-pressure gas, then a continuous rainbow-like spectrum will be produced such as that shown in figures 15.9 and 15.10. A common, everyday source is an ordinary **incandescent** light globe. ‘Incandescent’ means bright or glowing. Like black bodies, most substances become incandescent when they become hot enough.

Incandescent means bright or glowing. Like black bodies, most substances become incandescent when they become hot enough.



Continuous spectrum

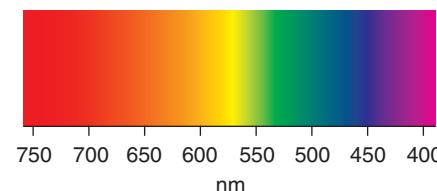


Figure 15.10

A continuous spectrum

PHYSICS IN FOCUS

Black body radiation

Black body radiation is the electromagnetic radiation that is emitted by a black body at a particular temperature. It is distributed continuously, but not evenly, across the various wavelengths, as shown in figure 15.11.

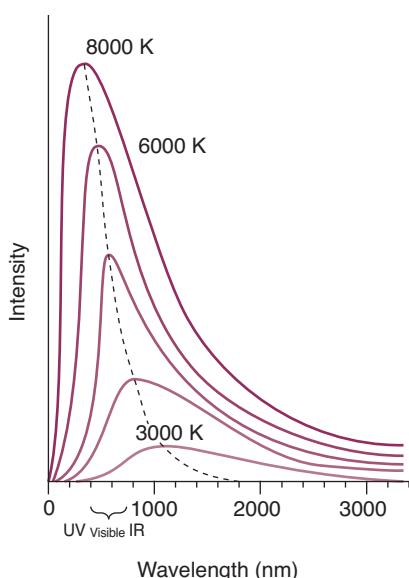


Figure 15.11 Black body radiation curves. Note that as the temperature increases, the curve becomes higher, indicating greater energy output, and the peak of the curve shifts to shorter wavelengths.

In figure 15.11, you will notice that each curve is specific to a particular temperature and has a peak intensity corresponding to a particular frequency. There are several noticeable trends which are explained here for interested students (although not essential content for this course):

- As the temperature increases, the peak moves toward the shorter wavelengths. At lower temperatures the radiation lies mostly in the infra-red region but, as temperature increases, the peak moves into the visible spectrum, and at higher temperatures the peak has moved well into the ultraviolet. This relationship can be written as:

$$\lambda_{\max} T = W$$

where

λ_{\max} = wavelength of maximum output (m)

T = temperature (K)

W = a constant

$$= 2.9 \times 10^{-3} \text{ m K}$$

This is known as Wien's Law, and it can be used to determine the approximate surface temperature of a star. If, when observing the light from a star, the wavelength of maximum output can be measured (using a spectrophotometer) then the surface temperature can be calculated.

(continued)

2. The colour of a black body changes with temperature. Referring to figure 15.11, as the temperature increases, the distribution of wavelengths within the visible spectrum changes. Applying this to stars we can say that low-temperature stars appear red. As the temperature increases, the wavelengths peak in the yellow, and the star will appear yellow. At slightly higher temperatures the spread is more even and a star will appear white. As the peak moves beyond the visible spectrum and into the ultraviolet with yet higher temperatures, the distribution within the visible spectrum is concentrated at the blue end, and so the star will appear blue.
3. As temperature increases, the black body radiation curve becomes higher and broader, indicating that more total energy is being emitted. This can be expressed by the following relationship, known as Stefan's Law:

$$L = 4\pi R^2 \sigma T^4$$

where

$$\begin{aligned} L &= \text{luminosity} \\ &= \text{total energy output per second} \\ &\quad (\text{joules/second, J s}^{-1}) \\ &= \text{total power output} \\ &\quad (\text{watts, W}) \\ T &= \text{temperature} \\ &\quad (\text{kelvin, K}) \\ R &= \text{radius of the star} \\ &\quad (\text{metres, m}) \\ \sigma &= \text{Stefan's constant} \\ &= 5.6705 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}. \end{aligned}$$

Note that, while a star's energy output is very sensitive to its temperature (proportional to the fourth power of T), it also depends upon the square of its radius. As an example, let us compare the two stars Castor and Deneb. Both are almost equally bright stars in our night sky and both have the same surface temperature. However, Deneb is much further away and much more luminous than Castor. Its greater energy output is due to its much larger radius — it is a bright supergiant, whereas Castor is a main sequence star, like the Sun.

In chapter 11 black bodies were discussed. Their consideration is important here also. A black body is a hypothetical object that is a perfect absorber and emitter of electromagnetic radiation. When at the temperature of its surroundings it emits as much radiation as it absorbs. The emitted radiation has a continuous distribution of wavelengths and depends only on the temperature of the surface of the body.

At optical to infra-red wavelengths, incandescent bodies, including stars, produce a continuous spectrum that, in the manner that they radiate energy, approximates a black body. Much is known of the nature of black body radiation. This allows information to be learned about a star from the close examination of the range of wavelengths and the intensity of the radiation it produces.

Emission spectra

An emission spectrum has the appearance of a long, dark rectangle upon which appears discrete bright coloured bands. They are produced

by hot glowing (incandescent) gases of low density like that shown in figure 15.12.

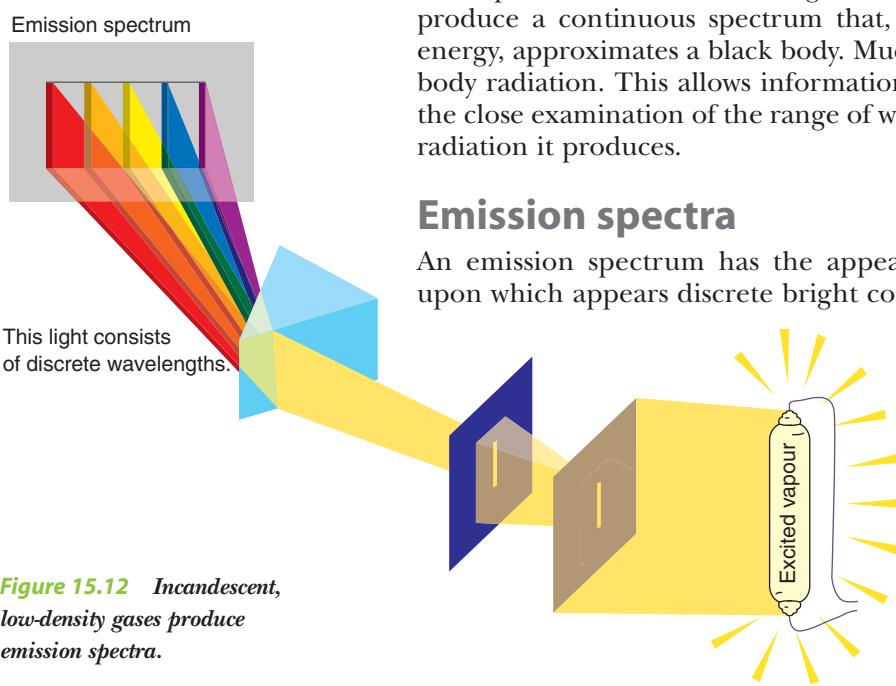


Figure 15.12 Incandescent, low-density gases produce emission spectra.

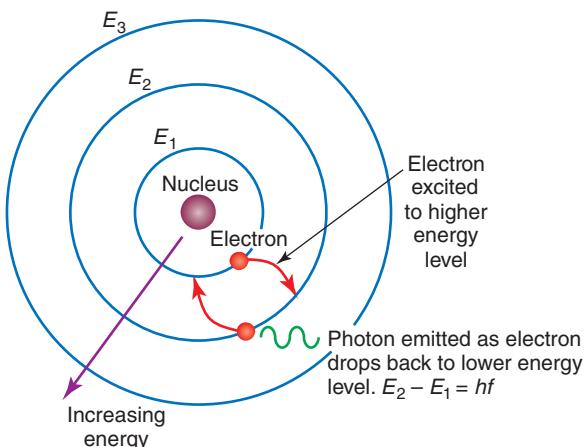


Figure 15.13 The different electron energy levels suggested by Niels Bohr

A **photon** is a quantum (or discrete packet) of electromagnetic radiation. It can be thought of as an elementary particle with zero rest mass and charge, travelling at the speed of light.

In order to understand how these discrete spectral lines are produced, we need to consider a simple model of the atom. In this model, shown in figure 15.13, the nucleus can be considered as a single positively charged body at the centre, and the electrons as tiny negatively charged particles orbiting the nucleus. In 1913, the Danish physicist Niels Bohr (1885–1962) suggested that the electrons were not free to orbit anywhere around the nucleus, but could orbit only at specific radii. Each specific radius represents a specific energy level.

Bohr restricted his original treatment to the hydrogen atom as this is the simplest, having just one electron. He suggested that the electron, which normally occupied the lowest energy level or ‘ground state’, E_1 , could be given some energy which would cause it to jump up to a higher energy level, or ‘excited state’, say E_2 . The energy could be given by a collision with other particles or with light.

Soon after occupying the excited state, the electron will drop back to the more stable ground state. The energy it loses in doing this ($E_2 - E_1$) is given off in the form of a photon, or packet of electromagnetic radiation.

When Bohr suggested this process, Max Planck (1858–1947) had already suggested that electromagnetic radiation occurred in packets of energy called **photons**. These can be thought of as elementary particles with zero rest mass and charge, travelling at the speed of light. A beam of light is thus a shower of photons, with the intensity of the beam dependent upon the number of photons in the shower.

Planck had further deduced that the energy of each photon depended only upon the wavelength of the radiation it contained, so that

$$E = hf$$

where

- E = energy (joules, J)
- f = frequency (hertz, Hz)
- h = Planck’s constant
 $= 6.6 \times 10^{-34}$ J s.

However, in the case of our atom, the energy of the photon must equal the difference in energy of the two states involved, so that:

$$E_2 - E_1 = hf$$

where

- E_1 = ground state
- E_2 = excited state.

If the electron had been excited to an even higher excited state, then it may return to the ground state in a single large jump, or alternatively in a set of smaller jumps. Each particular jump down between different energy levels represents a different amount of energy, and therefore a photon of radiation of different frequency or wavelength given off.

As a result, a hydrogen atom that has been excited tends to produce many photons, with a set of discrete frequencies unique to its own set of energy levels. When this light is directed through a spectroscope, then the spectrum produced will contain only discrete wavelengths, or lines, rather than a continuous spread of colours. The set of lines seen is so unique to hydrogen that it can be regarded as a fingerprint of that element. If the spectrum of an unknown mix of elements is examined and that particular set of lines appears, then it can be stated with confidence that the mixture contains hydrogen.

It is more difficult to fully explain the spectra produced by larger atoms or molecules, but the principle that each produces a characteristic set of spectral lines that can be regarded as a unique identifying fingerprint, still holds true. Figure 15.14 shows the characteristic emission spectra of several elements.

Emission spectra can be produced by certain hot interstellar gas clouds known as emission nebulae. These nebulae are heated by ultraviolet radiation from a nearby star until they become hot enough to shine by their own light (usually red light, which is characteristic of the hydrogen gas that makes up most of these nebulae). Quasars are another type of astronomical object that produces emission spectra. These are extremely distant and old objects that emit more energy than a hundred galaxies combined. Their spectra can be spread right across the electromagnetic spectrum, with the bulk lying within infra-red wavelengths.

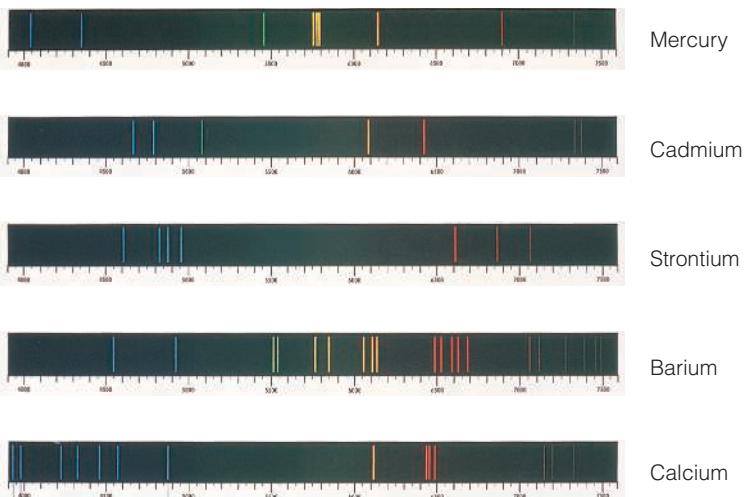


Figure 15.14 The emission spectra of several elements. Each has a unique and characteristic pattern.

Absorption spectrum

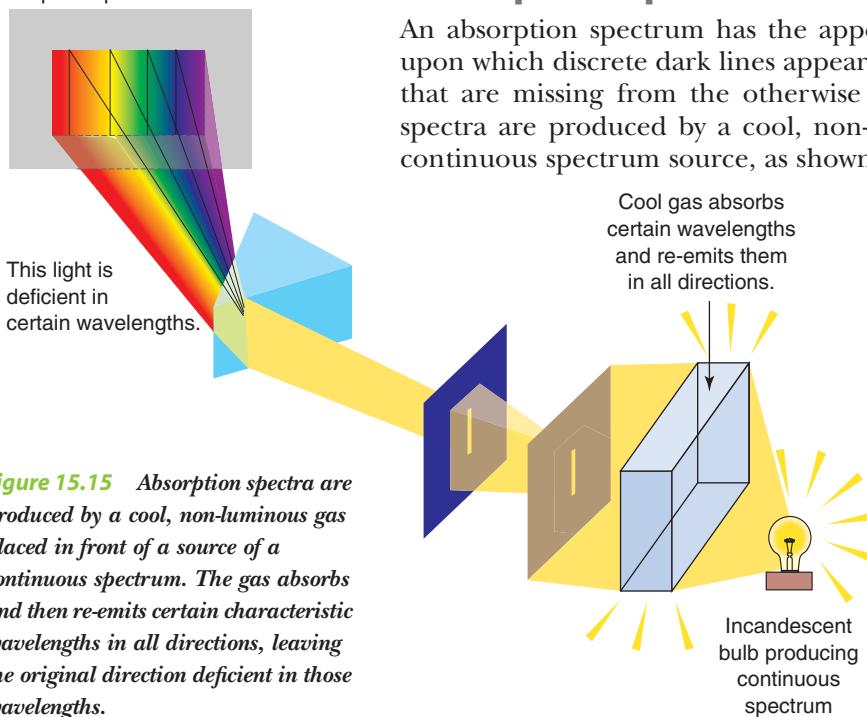


Figure 15.15 Absorption spectra are produced by a cool, non-luminous gas placed in front of a source of a continuous spectrum. The gas absorbs and then re-emits certain characteristic wavelengths in all directions, leaving the original direction deficient in those wavelengths.

Absorption spectra

An absorption spectrum has the appearance of a continuous spectrum upon which discrete dark lines appear. These lines represent wavelengths that are missing from the otherwise continuous spectrum. Absorption spectra are produced by a cool, non-luminous gas placed in front of a continuous spectrum source, as shown in figure 15.15.

Cool gas absorbs certain wavelengths and re-emits them in all directions.

As the continuous spectrum shines through the gas, the atoms of the gas will absorb those particular wavelengths that they would emit if the gas were hot. These are the frequencies that correspond to differences between energy levels within the atoms. Absorbing these wavelengths raises their electrons to excited states. The radiation is immediately re-emitted as the electrons drop back down to the ground state. However, this radiation is re-emitted in all directions, not just the original direction of the incident light.

This means that the original light is now deficient in those particular wavelengths. In general, the wavelengths missing from an absorption spectrum correspond to the bright lines in the emission spectrum of the same gas if it were hot and glowing. Therefore, the absorption spectrum of a cool gas contains the same identifying pattern of lines that are contained within emission spectra of hot gases.

This idea is the basis of stellar spectroscopy, since stars produce absorption spectra. The reason for this is that the main body of the star is hot, dense gas and therefore produces a continuous spectrum. Surrounding the star is a cooler and less dense atmosphere, which absorbs certain wavelengths and re-emits them away, resulting in an absorption spectrum. This is shown in figure 15.16.

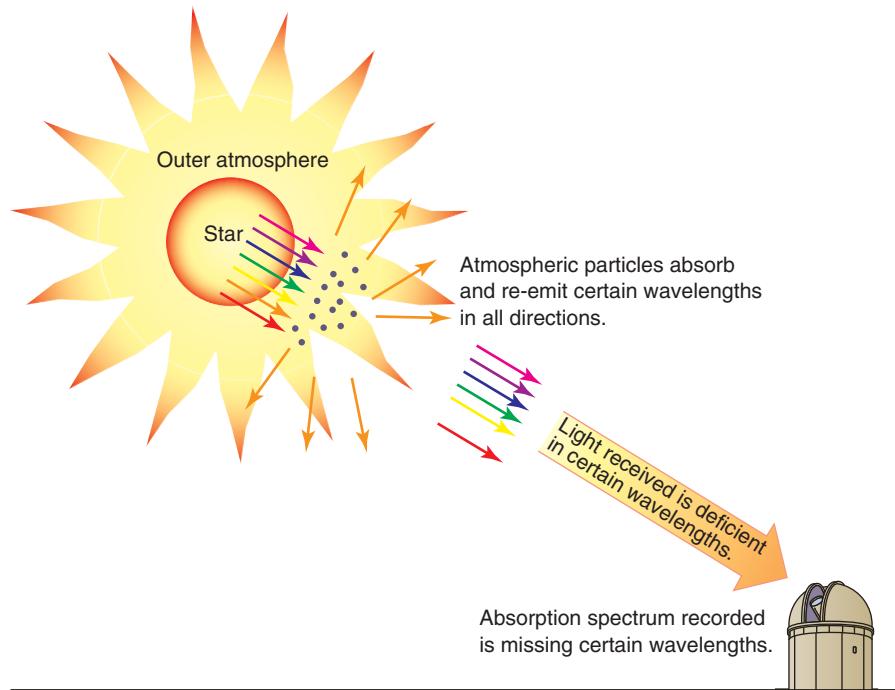


Figure 15.16 Stars produce absorption spectra. The hot, dense gas which forms the star is the continuous spectrum source. The surrounding atmosphere is the cool, non-luminous gas which absorbs certain wavelengths from the spectrum before re-emitting them away in all directions.



15.3 Spectra

Table 15.1 summarises the three types of spectra, how they are produced and what types of celestial objects produce them.

Table 15.1 Types of spectra and their production by celestial objects

TYPE OF SPECTRUM	GENERALLY PRODUCED BY:	CELESTIALLY PRODUCED BY:
Continuous	Hot solids, liquids, gases under pressure	Galaxies, inner layers of stars
Emission	Incandescent low-density gases	Emission nebulae, quasars
Absorption	Cool gases in front of continuous spectrum	Atmosphere of stars

Spectral analysis of starlight

In the late 1800s, the technique of spectroscopy began to be applied to the light of celestial bodies. Since then it has become one of the most valuable tools of the professional astronomer, as the spectral ‘fingerprints’ reveal the chemical composition of stars, nebulae and galaxies, as well as the atmosphere of the other planets of our own solar system.

Stellar spectroscopy is the examination of the spectra of stars in order to learn more about their composition, surface temperature, velocity, density, etc.

Stellar spectroscopy, concentrating on the spectra of stars, can deduce a great deal of other information, such as surface temperature, velocity and density, as the following sections describe.

Spectral classification

Most stars are made up of a very similar set of elements and compounds, yet their spectra can vary considerably. The reason is that different atoms and molecules produce spectral lines of very different strengths at different temperatures. At lower temperatures, molecules can exist near the surface of a star and they produce particular spectral lines. At hotter temperatures, these molecules can no longer exist and the spectral lines produced belong to neutral atoms. At higher temperatures the atoms become ionised, and the spectral lines produced are characteristic of these particles.

The stars have been classified into a set of spectral classes, each designated by a letter, according to the main spectral lines evident. When placed in order of decreasing temperature they are the seven spectral classes O, B, A, F, G, K and M. This can be remembered using the mnemonic ‘Oh, Be A Fine Girl (or Guy), Kiss Me’. Table 15.2 summarises the main spectral features of each class, as well as the temperature and colour (recalling that the colour of a star depends upon its temperature) that correspond to each.

PHYSICS FACT

In recent years, astronomers have discovered a new class of stars, which they have named class L. These are dwarf stars cooler than M class stars, with a surface temperature less than 2500 K and possibly as low as 1600 K.

Table 15.2 Spectral classifications and their corresponding features. Note that in astronomy, the term ‘metal’ refers to any element other than hydrogen or helium.

SPECTRAL CLASS	TEMPERATURE (K)	COLOUR	SPECTRAL FEATURES
O	28 000–50 000	Blue	Ionised helium lines Strong UV component
B	10 000–28 000	Blue	Neutral helium lines
A	7500–10 000	Blue-white	Strong hydrogen lines Ionised metal lines
F	6000–7500	White	Strong metal lines Weak hydrogen lines
G	5000–6000	Yellow	Ionised calcium lines Metal lines present
K	3500–5000	Orange	Neutral metals dominate Strong molecular lines
M	2500–3500	Red	Molecular lines dominate Strong neutral metals

Each spectral class has been further divided into subgroups by attaching a digit, from 0 to 9, following the letter. As an example, a small section of the classification system would be as follows:

-B8-B9-A0-A1-A2-A3-A4-A5-A6-A7-A8-A9-F0-F1-F2-

PHYSICS IN FOCUS

Luminosity classes

When considering black body radiation and temperature on page 281, we saw that stars with similar temperatures could still have very different luminosity (energy output) due to a difference in size. To account for this, the spectral classification system was extended in 1941, by the inclusion of eight luminosity classes as listed in table 15.3. Each is indicated by a Roman numeral that appears as a suffix behind the spectral class. For example, Castor is an A2V star (a blue-white main sequence star) while Deneb is an A2Ia star (a blue-white bright supergiant). Each luminosity class corresponds to specific regions of a Hertzsprung–Russell diagram (see figure 15.17).

Table 15.3 The eight luminosity classes

Ia	Bright supergiant
Ib	Supergiant
II	Bright giant
III	Giant
IV	Subgiant
V	Main sequence
VI	Subdwarf
VII	White dwarfs

The luminosity class of a star can be learned from its spectra, in particular the pressure broadening of its spectral lines. (See the text under ‘Density’ on page 289.)

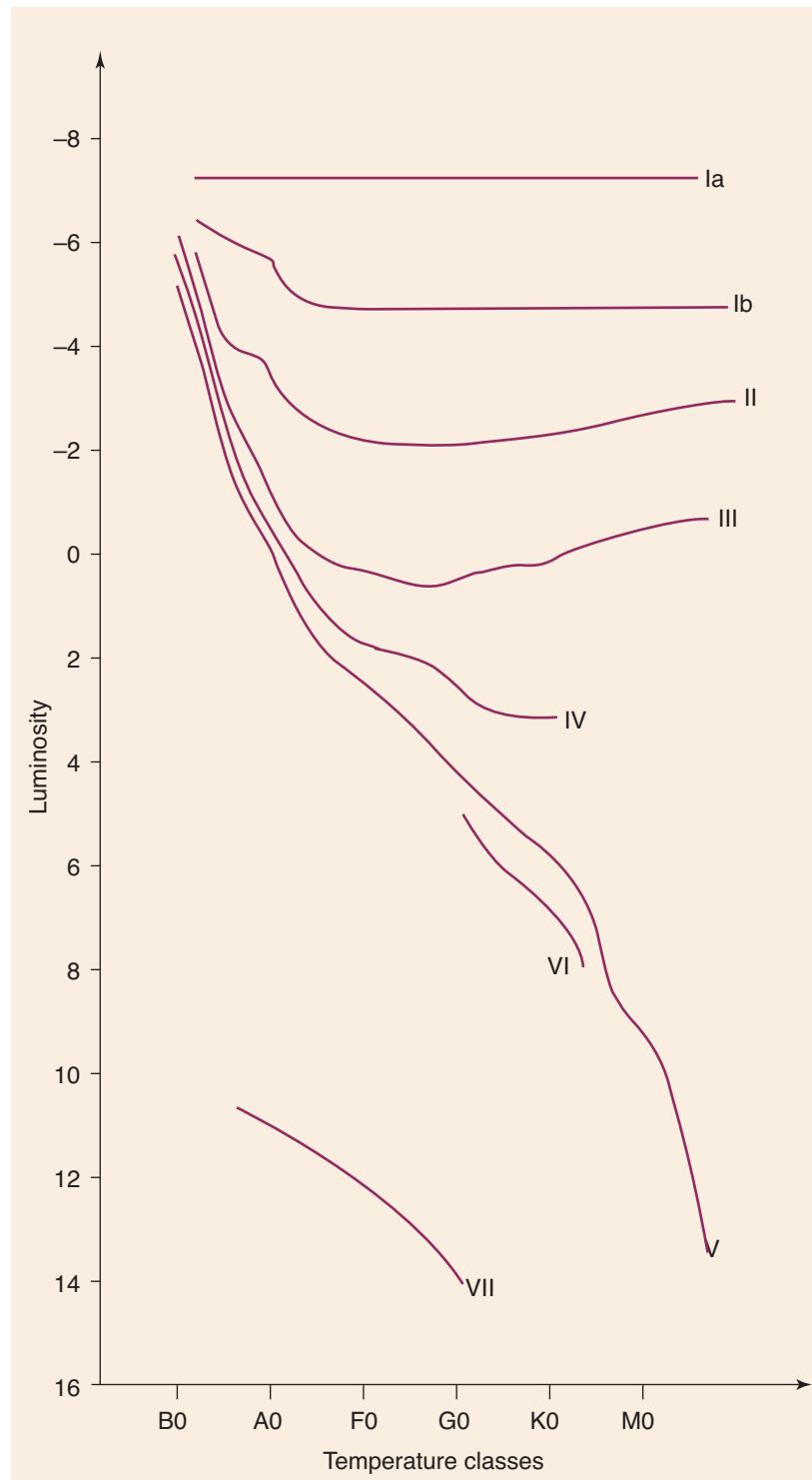


Figure 15.17 The eight luminosity classes plotted on a Hertzsprung–Russell diagram

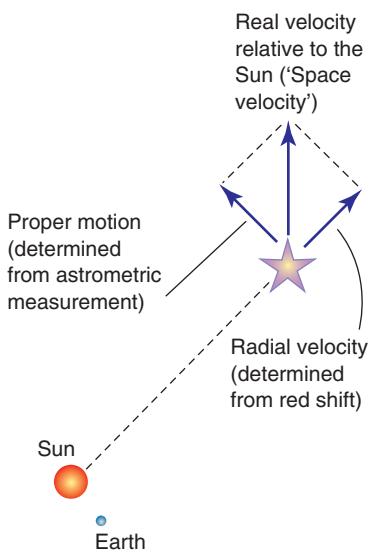


Figure 15.18 If a star's radial velocity is combined with its proper motion then its real velocity relative to the Sun can be determined.

Other inferred information

Temperature

As can be seen from table 15.2 on page 286, once a star's spectral class has been identified from its main spectral lines, then its temperature can be deduced. (There is another method. Using a spectrophotometer, each wavelength can be examined for intensity in order to discover the wavelength of maximum energy output, λ_{\max} . Once this has been determined, the temperature of the star can be calculated using Wien's Law, as described on page 281. This measurement would also suggest the spectral class and, hence, composition of the star.)

Translational velocity

If a star moves away from us then the patterns of recognisable lines within its spectrum, as we observe them, appear at slightly longer wavelengths. We say they have red-shifted because the lines now appear a few nanometres closer to the red end of the spectrum. Similarly, if the star were to move toward us then its spectral lines would be blue-shifted. This is due to the Doppler effect and, by measuring the extent of the wavelength shift, the radial velocity of the star (velocity toward or away from us) can be calculated. As figure 15.18 shows, if this is combined with the proper motion of the star (sideways velocity as seen by us) then the star's real velocity relative to the Sun can be computed.

PHYSICS IN FOCUS

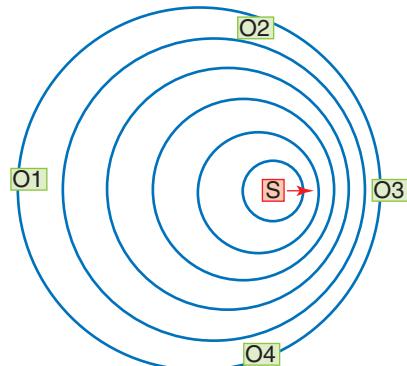
The Doppler effect

The Doppler effect is an apparent change in the frequency or wavelength of a wave as a result of the source and the observer moving relative to each other (see figure 15.19).

If a source of waves (these could be sound waves, light waves, microwaves, radio waves, etc.) is moving at a significant fraction of the speed of those waves, then the wavelength is effectively shortened ahead of its motion, and lengthened behind it. Observers at these locations will experience these changes in wavelength and frequency and, by measuring the change, the velocity of the source can be calculated.

The same effect can be produced if the source is stationary and the observer is moving, so it is really relative motion between the source and observer that is significant.

The Doppler effect is used by police radar units to measure the velocity of speeding cars. In astronomy it is used to measure the velocity of stars relative to us. Very fast and distant celestial objects are usually assumed to owe their velocity to the general expansion of the universe, and in this case the degree of Doppler shifting can also give an indication of distance.



Legend:

- [S] = Source
- [O1] = Observer 1
- [O2] = Observer 2
- [O3] = Observer 3
- [O4] = Observer 4

Figure 15.19 The Doppler effect. When the source is moving quickly compared to the speed of the waves it produces, then observer 1 'sees' a longer wavelength than emitted, observer 3 sees a shorter wavelength, while observers 2 and 4 see the same wavelength as that emitted by the source.

Rotational velocity

Smaller Doppler shifts (both red and blue) can be caused by a star's own rotational velocity, or by its participation in a rotating double star system. In the case of a single, rapidly rotating star, the atoms moving quickly

away from us on one side and atoms moving quickly toward us on the other side combine to produce a slight but simultaneous red and blue shift which broadens the spectral lines. The faster the star rotates, the greater this Doppler broadening effect is.

In the case of a rotating double star system seen from its edge, at certain times one star is blue-shifted while the other is red-shifted. At some later period this situation will reverse and, by keeping track of this, the rotational period and velocities can be calculated.

Density

High density and pressure within the atmosphere of the star can also broaden its spectral lines. The effect is progressive — the greater the atmospheric density and pressure, the greater the ‘pressure broadening’. The spectral lines of a supergiant star (with a particularly low density atmosphere) are much narrower than those of a more dense main sequence star (such as our Sun) of the same spectral class.

15.3 PHOTOMETRY

Photometry is the measurement of the brightness of a source of light or other radiation.

Photometry is the measurement of the brightness of a source of light or other radiation. In astronomy this is applied to the light from stars as well as other celestial objects. Astronomical photometry has a long history, beginning in Greece over two thousand years ago. Early measurements of star brightness were judged by eye. More recently, photographic techniques were applied with a corresponding increase in accuracy. Most recently, electronic devices have been employed to measure star brightness with further improvements in sensitivity and faster response times. These devices may be photomultiplier tubes that offer high sensitivity to very low light levels, or charge coupled devices (CCDs), such as those found within video cameras, that can produce digitised images for computer processing.

However it is measured, much can be learned from the knowledge of a star’s brightness and how it compares to other stars.

Measuring brightness and luminosity

When looking at the night sky it is obvious that the stars vary in brightness. A star’s brightness, in watts per square metre, is a measure of the intensity of the radiation reaching the Earth from the star. This depends upon the luminosity of the star as well as its distance from us.

We briefly considered luminosity on page 282 and saw that luminosity depends upon the radius of the star and, especially, upon its temperature. The luminosity of a star, in watts, is the total energy radiated by it per second. Since it is the rate of energy output, it is also the power output of the star and is sometimes called intrinsic or absolute brightness (see figure 15.20).

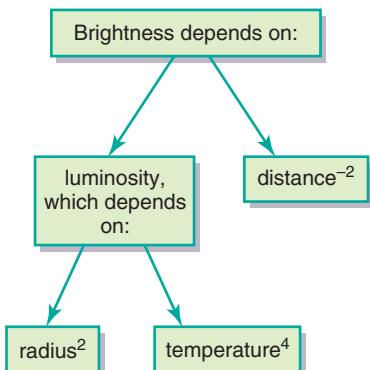


Figure 15.20 Brightness and luminosity are related.

PHYSICS FACT

Detectors aboard satellites have measured the amount of radiant energy per square metre per second reaching us from the Sun. From this, astronomers have been able to calculate the total power output of the sun as 3.83×10^{26} watts, and this is designated by the symbol L_O . The Sun’s power output, L_O , is used as a standard for comparison with other stars.

In this method, a measurement of the Sun’s brightness and a knowledge of its distance has allowed a calculation of its luminosity.

Magnitudes

During the second century BC, the Greek astronomer Hipparchus established a scale to record the brightness of stars, an extension of which is still in use today. He defined the brightest star that he could see as being of magnitude 1 and the faintest star as magnitude 6. Therefore, his scale was a reverse one, with lower numbers indicating brighter stars.

Since Hipparchus established his scale, stars have been found that are brighter than magnitude 1 (the brightest star in the night sky is Sirius with a magnitude of -1.4) and dimmer than magnitude 6 (large telescopes can observe stars fainter than magnitude 25). Of course, Hipparchus did not have a telescope to help him.

In the 1850s scientists began to realise that the human eye did not respond to light increases in a linear way. For example, a light that is doubled in intensity is not perceived to be twice as bright. William Herschel had observed that a magnitude 1 star was approximately 100 times brighter than a star of magnitude 6. In 1856 Norman Pogson fixed this observation as a mathematical definition:

$$\text{If } m_B - m_A = 5 \text{ then } \frac{I_A}{I_B} = 100$$

where

m_A = magnitude of star A (brighter star)

m_B = magnitude of star B (dimmer star)

$\frac{I_A}{I_B}$ = brightness ratio of the two stars.

This means that:

$$\begin{aligned} \text{If } m_B - m_A = 1 \text{ then } \frac{I_A}{I_B} &= \text{5th root of 100} \\ &= 2.512. \end{aligned}$$

In other words, a star of magnitude, say, 4 is 2.512 times brighter than a star of magnitude 5. This is known as Pogson's ratio, and the compounding nature of this ratio can be seen in table 15.4. The magnitude scale, now based upon the mathematical relationship described, is known as Pogson's scale.

Table 15.4 Brightness ratios using the Pogson scale

MAGNITUDE DIFFERENCE BETWEEN TWO STARS	BRIGHTNESS RATIO OF THE STARS
0	$2.512^0 = 1$
1	$2.512^1 = 2.512$
2	$2.512^2 = 6.310$
3	$2.512^3 = 15.85$
4	$2.512^4 = 39.81$
5	$2.512^5 = 100$
6	$2.512^6 = 251$
7	$2.512^7 = 631$
8	$2.512^8 = 1585$
9	$2.512^9 = 3981$
10	$2.512^{10} = 10\,000$

In general, the brightness ratio of any two stars can be calculated using the following formula:

$$\frac{I_A}{I_B} = 100^{\frac{(m_B - m_A)}{5}}$$

where

m_A = magnitude of star A (brighter star)

m_B = magnitude of star B (duller star)

$\frac{I_A}{I_B}$ = brightness ratio of the two stars.

SAMPLE PROBLEM

15.2

Calculating a brightness ratio

The Sun has a magnitude of -26.8 . The brightest star in the night sky is Sirius with a magnitude of -1.4 . How much brighter does the Sun appear compared to Sirius?

SOLUTION

$$\begin{aligned}\frac{I_A}{I_B} &= 100^{\frac{(m_B - m_A)}{5}} \\ \frac{I_{\text{Sun}}}{I_{\text{Sirius}}} &= 100^{\frac{(m_{\text{Sirius}} - m_{\text{Sun}})}{5}} \\ &= 100^{\frac{(-1.4 - -26.8)}{5}} \\ &= 100^{5.08} \\ &= 1.4 \times 10^{10}\end{aligned}$$

In other words, the Sun appears over ten billion times brighter than Sirius.

SAMPLE PROBLEM

15.3

Another brightness ratio

Proxima Centauri is the closest star to our solar system and yet it is quite faint, with a magnitude of 11 . Calculate the brightness ratio of Algol (magnitude 2.1) compared to Proxima Centauri.

SOLUTION

$$\begin{aligned}\frac{I_A}{I_B} &= 100^{\frac{(m_B - m_A)}{5}} \\ \frac{I_{\text{Algol}}}{I_{\text{Proxima}}} &= 100^{\frac{(m_{\text{Proxima}} - m_{\text{Algol}})}{5}} \\ &= 100^{\frac{(11 - 2.1)}{5}} \\ &= 100^{1.78} \\ &= 3600\end{aligned}$$

That is, Algol appears 3600 times brighter than Proxima Centauri, although a good telescope is required to see Proxima Centauri.

The concept of magnitudes has developed over the years for various purposes and we now need to be more specific about what type of magnitude we are dealing with at any time.

Apparent magnitude

Apparent magnitude, m , is the magnitude given to a star as viewed from Earth.

Apparent magnitude, given the symbol m , is the magnitude given to a star as viewed from Earth. This is the same magnitude that we have been discussing in the previous section of work. Apparent magnitude is a measure of the brightness of a star and is therefore influenced by the distance of the star as well as its intrinsic brightness (and any intervening matter such as interstellar dust, which can make a star look dimmer than it otherwise would be). Measurements of apparent magnitude can be performed photographically or photoelectrically.

Absolute magnitude

Absolute magnitude, M , is the magnitude that a star would have if it were viewed from a standard distance of 10 parsecs.

Absolute magnitude, given the symbol M , is defined to be the magnitude that a star would have if it were viewed from a standard distance of 10 parsecs. Because distance has been set to a standard, it is no longer an influence and so absolute magnitude is a measure of the intrinsic brightness or luminosity of a star. Although it is not a quantity that is directly measurable, there are other ways that absolute magnitudes can be deduced, and it proves to be a useful means to make direct comparisons between stars. For example, Achernar and Betelgeuse are both bright stars with apparent magnitudes of 0.45; however, Achernar has an absolute magnitude of -2.8 while Betelgeuse's is much brighter again at -5.14. The reason the two stars appear the same is that Betelgeuse is much further away from us than is Achernar.

The distance modulus

Consider the following comparison: Castor and Vega are both A class main sequence stars with absolute magnitudes of approximately 0.6. Vega lies at a distance of 7.76 pc and has an apparent magnitude of 0.03. However, Castor lies roughly twice as distant at 15.8 pc and therefore is not as bright, with an apparent magnitude of 1.58.

The close relationship between apparent magnitude, absolute magnitude and distance is expressed in the following formula:

$$M = m - 5 \log\left(\frac{d}{10}\right)$$

where

M = absolute magnitude

m = apparent magnitude

d = distance (pc).

Rearranging this equation produces a number of other accepted forms of the expression, as follows:

$$m - M = 5 \log\left(\frac{d}{10}\right)$$

$$m - M = 5 \log d - 5.$$

The term $(m - M)$ is known as the **distance modulus**. It is directly related to the distance of a star.

SAMPLE PROBLEM

15.4

SOLUTION

Calculating distance using the distance modulus

Achernar has an apparent magnitude of 0.45 and an absolute magnitude of -2.77. Calculate its distance.

$$M = m - 5 \log\left(\frac{d}{10}\right)$$

$$-2.77 = 0.45 - 5 \log\left(\frac{d}{10}\right)$$

$$-3.22 = -5 \log\left(\frac{d}{10}\right)$$

$$0.644 = \log\left(\frac{d}{10}\right)$$

$$\therefore \frac{d}{10} = 10^{0.644} = 4.4$$

$$\therefore d = 4.4 \times 10 = 44 \text{ pc}$$

Utilising annual parallax and the distance modulus

According to the HIPPARCOS Catalogue, Altair has a parallax of 194.44 milliarcsec and an apparent magnitude of 0.76. Calculate:

- its distance, and
- its absolute magnitude.

SOLUTION

$$\begin{aligned}
 (a) \quad d &= \frac{1}{p} = \frac{1}{0.194\ 44} \\
 &= 5.14 \text{ pc} \\
 (b) \quad M &= m - 5 \log \left(\frac{d}{10} \right) \\
 &= 0.76 - 5 \log \frac{5.14}{10} \\
 &= 0.76 - 5 \log 0.514 \\
 &= 0.76 - 5(-0.289) \\
 &= 2.2
 \end{aligned}$$

Spectroscopic parallax

Spectroscopic parallax is a method of using the H-R diagram and the distance modulus formula to determine the approximate distance of a star.

Spectroscopic parallax is the name given to a method of using the Hertzsprung–Russell (H–R) diagram and the distance modulus formula to determine the approximate distance of a star.

When you studied ‘The Cosmic Engine’ in Year 11, you were introduced to the H–R diagram. This is a graph of luminosity or absolute magnitude (vertical axis) versus temperature or spectral class (horizontal axis). When many stars are plotted onto an H–R diagram, certain star groupings become apparent, such as the main sequence, red giants and white dwarfs, as shown in figure 15.21.

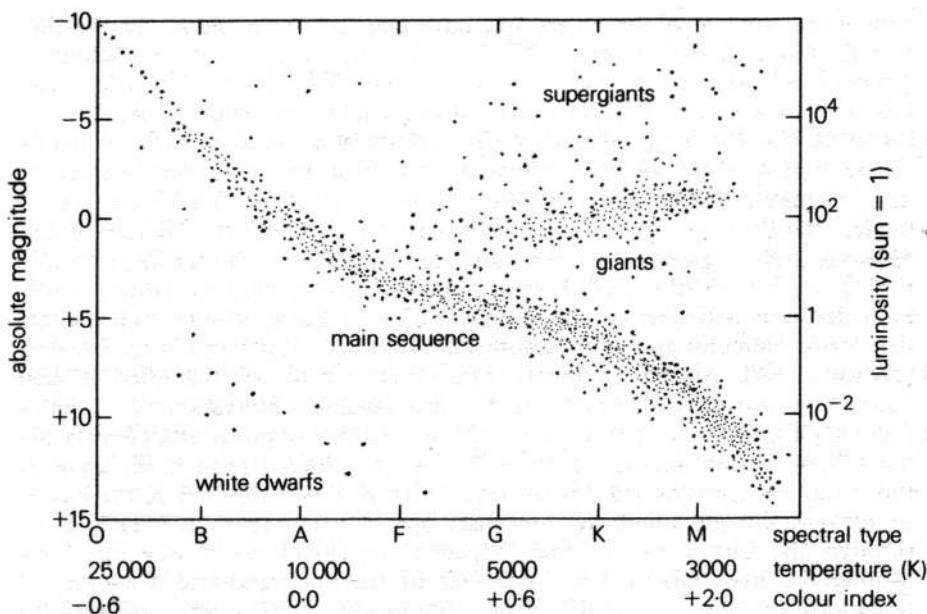


Figure 15.21 An H–R diagram showing the various star groupings

The method involves the following steps:

- Using photometry, measure the apparent magnitude, m , of the star in question.
- Using spectroscopy, determine the spectral class of the star. It is also important to note the luminosity class, in order to determine to which

group the star belongs. Earlier in this chapter we discussed how the width of the spectral lines is an indicator of the luminosity class of a star.

3. The H–R diagram is now consulted. Starting on the horizontal axis at the appropriate spectral class, a vertical line is drawn to the middle of the correct star group. This will place the star on the H–R diagram in approximately the correct position. From this position a horizontal line is drawn across to the vertical axis, so that the absolute magnitude, M , can be read off.
4. Now that m and M are both known, the distance modulus formula is applied in order to calculate the distance, d .

This technique can be applied to many stars; however, the value determined for the absolute magnitude can carry a large percentage error and, therefore, so too will the calculated distance. In other words, the distance determined by this method is approximate only.

SAMPLE PROBLEM**15.6*****Calculating the distance to Aldebaran using spectroscopic parallax***

Aldebaran is a bright red star with an apparent magnitude of 0.9 and belongs to spectral class K5 and luminosity class III, which means that it is a red giant. Use spectroscopic parallax to determine its approximate distance.

SOLUTION

With reference to the H–R diagram in figure 15.21, we locate spectral class K5 along the horizontal axis and move from here vertically up to the middle of the giants. From this point, slide horizontally across to read off an absolute magnitude, M , of approximately 0. The distance modulus formula is then applied:

$$\begin{aligned} M &= m - 5 \log\left(\frac{d}{10}\right) \\ 0 &= 0.9 - 5 \log\left(\frac{d}{10}\right) \\ -0.9 &= -5 \log\left(\frac{d}{10}\right) \\ 0.18 &= \log\left(\frac{d}{10}\right) \\ \therefore \frac{d}{10} &= 10^{0.18} = 1.5 \\ \therefore d &= 1.5 \times 10 = 15 \text{ pc.} \end{aligned}$$

This problem serves as a good example of the approximate nature of this technique. The annual parallax of Aldebaran has been accurately measured to be 50 milliarcsec. Therefore, its distance can be much more accurately calculated using:

$$\begin{aligned} d &= \frac{1}{p} = \frac{1}{0.05} \\ &= 20 \text{ pc.} \end{aligned}$$

The source of the error lies in the approximation of M , as Aldebaran's real absolute magnitude lies closer to –0.6.

SAMPLE PROBLEM**15.7*****Calculating the distance to Proxima Centauri using spectroscopic parallax***

Proxima Centauri, the closest star to our solar system, is an M5V star with an apparent magnitude of 11. Determine its distance using spectroscopic parallax.

SOLUTION

Spectral class M5 indicates that Proxima is a red star. Luminosity class V indicates that it is a main sequence star (these stars are known as red

dwarfs). From the H–R diagram in figure 15.21 the absolute magnitude can be approximated at 12.5. Applying the distance modulus equation:

$$\begin{aligned}M &= m - 5 \log\left(\frac{d}{10}\right) \\12.5 &= 11 - 5 \log\left(\frac{d}{10}\right) \\1.5 &= -5 \log\left(\frac{d}{10}\right) \\-0.3 &= \log\left(\frac{d}{10}\right) \\\therefore \frac{d}{10} &= 10^{-0.3} = 0.5 \\\therefore d &= 0.5 \times 10 = 5 \text{ pc.}\end{aligned}$$

It is well known that Proxima Centauri actually lies at a distance of 1.3 pc. Once again the source of the inaccuracy lies in the approximation of M , as Proxima Centauri's actual absolute magnitude is almost 15.5.

The sample problems above demonstrate the technique, as well as the approximate nature of the distances it produces. What then is its use? It is not used when there is a more accurate solution available. Rather, it can be used to give a 'ball-park figure' when no other technique can.

Measuring colour

It has already been noted that the stars vary in colour. This variation is not obvious when looking with the naked eye. Aldebaran, found in the constellation of Taurus, and Betelgeuse, found in Orion, are both red giants yet their colour is a subtle pink when studied with unaided eyes. Look with a good set of binoculars or a telescope, however, and the colour becomes obvious. Similarly, the blue colour of stars such as Rigel, also found in the constellation of Orion, is quite faint until looked at with a telescope. All of these stars are quite bright, but just how bright they appear depends very much upon the colour sensitivity of the device and method used.

Early measurements of star magnitudes were done by eye, which can be a surprisingly sensitive discriminator. However, the human eye is most sensitive to the yellow-green portion of the visible spectrum. As a result, the red and blue stars mentioned earlier are not judged by the eye to be as bright as they really are. Magnitude determined this way is referred to as **visual magnitude**.

Later measurements of star magnitudes were made photographically, and called 'photographic magnitudes'. However, a problem arose. Photographic magnitudes were inconsistent with visual magnitudes. The source of the inconsistency was that photographic film is most sensitive to the blue end of the visible spectrum, so that blue stars were measured to be brighter than by eye, while yellow and red stars were measured to be fainter.

Today, magnitudes are measured using photometers. Not only are these devices very sensitive and accurate, but they are also sensitive to a much wider range of wavelengths such as ultraviolet and infra-red, to which the human eye is quite insensitive. In order to maintain consistency with naked eye observations, stars are observed through a yellow-green filter, called a V (for visual) filter, when apparent magnitudes are being measured.

Visual magnitude refers to magnitude as judged by eye, or more accurately by a photometer fitted with a yellow-green filter.

Colour magnitudes

The influence of colour sensitivity upon magnitude measurements, as described on the following page, has been taken advantage of to devise a way to quantify star colours. By placing a standard set of coloured filters in front of a photometer, three different colour magnitudes for each star can be measured. The filters are described in table 15.5 on the following page. Each filter transmits a broad band of wavelengths, and the wavelength at the centre of this band is also indicated in the table.

Note that one of these filters is the yellow-green filter (V), used to simulate visual magnitudes. Another is a blue filter (B), which is used to simulate photographic magnitudes. The ultraviolet filter (U) utilises the extra sensitivity available from the photometer. This has become an internationally accepted system referred to as the UBV system.

Table 15.5 Colour filters used in the UBV system

NAME	COLOUR	CENTRE WAVELENGTH	BASED UPON
U	Ultraviolet	365 nm	—
B	Blue	440 nm	Photographic magnitude
V	Yellow-green	550 nm	Visual magnitude

The letters U, B and V are also used to denote a star's apparent magnitude as measured through each of the filters. Note that a red star, such as Betelgeuse, will appear brightest through the V filter, so its V magnitude will be lower than its B or U. Similarly a blue star, such as Rigel, will appear brightest through the B filter so that its B magnitude will be lower than its V or U.

Comparisons such as these can be useful. However, because these colour magnitudes are numbers, comparisons can be made numerically, and this is the purpose of the colour index.

PHYSICS FACT

The UBV system of stellar magnitudes has been extended into the red and infra-red wavelengths, better utilising the capabilities of photometers and creating a system able to produce two-colour values for a much greater range of stars. The filters that have been added are listed in table 15.6.

Table 15.6 Filters used to extend the UBV system

NAME	CENTRE WAVELENGTH
R	700 nm or 0.7 μm
I	900 nm or 0.9 μm
J	1250 nm or 1.25 μm
H	1.6 μm
K	2.2 μm
L	3.4 μm
M	5.0 μm
N	10.2 μm
Q	21 μm

Colour index

By subtracting one colour magnitude from another, a numerical two-colour value will result that expresses the colour of a star. The most standard of this type of numerical comparison is the colour index.

Colour index is the difference between the photographic magnitude, B, and the visual magnitude, V.

$$\text{Colour index} = B - V$$

The application of this formula results in a numerical scale that expresses colour. To see how this occurs, recall that a red star is brighter through a V filter than a B, so that its V magnitude is lower than its B magnitude. Therefore, the expression $B - V$ will result in a small positive number. A blue star is the reverse of this. A blue star is brightest through a B filter, and so its B magnitude will be less than its V magnitude. The expression $B - V$ will result in a small negative number. By definition, stars of spectral class A0 have a colour index of zero. These stars have a surface temperature of 10 000 K and a blue-white colour.

Table 15.7 shows the range of the colour index scale and how it correlates with colour, as well as temperature and spectral class.

Table 15.7 The correlation of colour index, colour, temperature and spectral class

COLOUR INDEX	COLOUR	SPECTRAL CLASS	TEMPERATURE (K)
-0.6	Blue	O	28 000–50 000
	Blue	B	10 000–28 000
0	Blue–white	A	7500–10 000
	White	F	6000–7500
+0.6	Yellow	G	5000–6000
	Orange	K	3500–5000
+2.0	Red	M	2500–3500

Note that the relationship between colour index and temperature is not linear — colour index -0.6 to zero covers a temperature range of 40 000 K, colour index zero to $+0.6$ covers a range of 4000 K, and colour index $+0.6$ to $+2.0$ covers about 4000 K. If a spectrophotometer is available, Wein's Law can be used to give a more accurate value for the star's temperature (see page 281).

SAMPLE PROBLEM

15.8



15.4

Colour filters

Using colour index information

Three stars are measured to have colour indexes of $+0.5$, 0 and -0.5 . What can be said of each star?

SOLUTION

Refer to table 15.7. Of the first star (colour index $+0.5$) we can say that its colour is white-yellow, its spectral class is about F5 and its surface temperature is approximately 6500 K. Of the second star (colour index 0) we can say that its colour is blue-white, its spectral class is A0 and its temperature is approximately 10 000 K. Of the third star (colour index -0.5) we can say that its colour is blue, its spectral class is about O5, and its temperature is approximately 30 000 K to 40 000 K.

PHYSICS IN FOCUS

Photoelectrics versus photographics for photometry

Photometry involves the measurement of the brightness or magnitude of a source of light such as a star.

This can be done photographically, using specially prepared emulsions, in the field known as photographic photometry. This method involves making a photograph of a portion of the sky. When the photograph is developed, a measurement is made of the size and density of the spot made by each star. Brighter stars expose a larger area of film and hence appear on a photograph as a larger, denser spot (as shown in figure 15.22). Each spot is compared to standard spot sizes and densities to determine the stars' magnitudes.

Photographic emulsions are restricted to the visible spectrum including the near-infra-red and near-ultraviolet; however, particular emulsions can restrict this range further. In addition, fine detail can be recorded photographically, often to a higher resolution than can be achieved electronically.

Photoelectric photometry is more common. These systems use a combination of a filter and an electronic sensor such as a Charge Coupled Device (CCD) which is a light-sensing array also found in a video camera — CCDs for astronomy have a higher resolution. A photomultiplier tube may also be used. Both devices convert the light input into an electronic signal that can be multiplied, digitised, analysed and stored electronically. All of this can be done quite quickly and remotely if necessary.

These devices have a much wider range of wavelengths to which they are sensitive than do photographic emulsions. Together with a suitable filter they can sense intensities over broad wavelength bands, as in the UBV system, or very narrow bands, useful when searching for the presence of a particular element in a celestial object. In addition, the electronic detectors in use today are more sensitive to faint light sources than is photographic emulsion, although they cannot, as yet, achieve the same resolution.

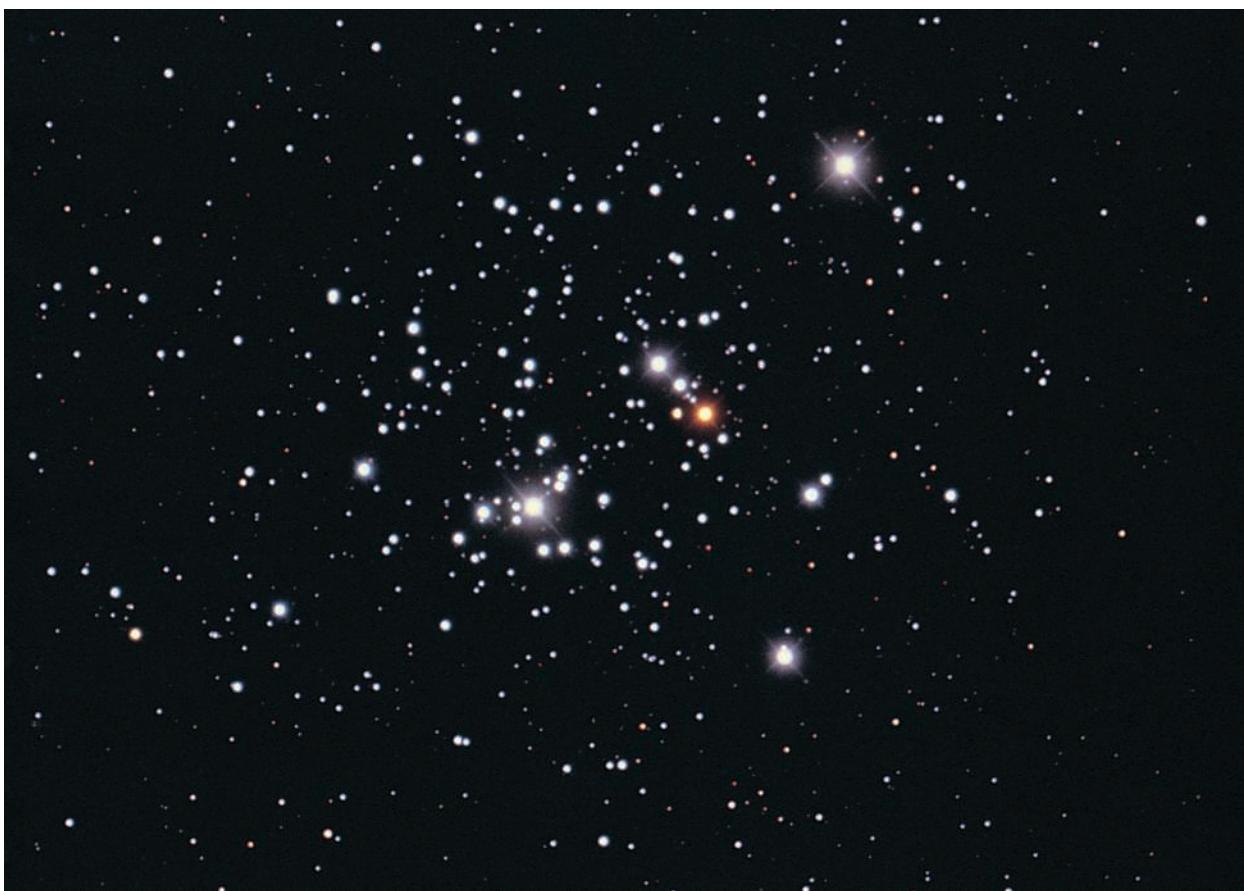


Figure 15.22 On a photograph, brighter stars appear as larger spots.

SUMMARY

- Annual parallax, ρ , is half the angle, in seconds of arc, through which a nearby star appears to shift as the Earth moves between two positions six months apart.
- The distance, in parsecs, to a nearby star is given by $d = \frac{1}{\rho}$.
- Atmospheric blurring reduces the accuracy of ground-based astrometric measurements. Space-based measurements avoid this problem, allowing the ability to measure more distant stars.
- A spectroscope is a device that uses a prism or a diffraction grating to separate a light into its spectrum.
- There are three different types of spectra — continuous (produced inside stars), emission (produced by quasars and certain nebulae) and absorption (produced by stars).
- Black body radiation is approximated by a star's energy output in optical and infra-red wavelengths. It is spread continuously but not evenly across the electromagnetic spectrum. It has a peak output that is dependent upon temperature.
- The position and appearance of the lines in a spectrum can reveal details of a star's composition, temperature, velocity and density.
- The spectral classes, from hottest to coolest, are O, B, A, F, G, K, M.
- Apparent magnitude, m , is the magnitude given to a star as viewed from Earth. It is a measure of a star's brightness.
- A star that is one magnitude lower than another is 2.512 times brighter. The brightness ratio of any two stars can be calculated using the following formula:

$$\frac{I_A}{I_B} = 100^{\frac{(m_B - m_A)}{5}}.$$

- Absolute magnitude, M , is the magnitude that a star would have if viewed from a standard distance of 10 parsecs. It is a measure of a star's luminosity.
- If m and M are known, then a star's distance can be calculated using the following formula:

$$M = m - 5 \log \frac{d}{10} \quad \text{or} \quad m - M = 5 \log d - 5.$$

- Spectroscopic parallax is a method of approximating a star's distance using an H-R diagram to estimate M , and a measurement of m , to then calculate d .
- Star magnitudes can be measured through colour filters to produce the colour magnitudes U, B and V.
- Colour index = $B - V$. This is a numerical measure of the colour of stars and produces a scale from -0.6 (blue) to 0.0 (blue-white) to +2.0 (red).
- Photoelectric photometry has a greater range, degree of sensitivity and ability to produce easily digitised images. Photographic photometry can achieve a higher resolution.

QUESTIONS

Astrometry

1. Complete the following table of conversion factors:

	km	AU	l-y	pc
1 km =	1			
1 AU =		1		
1 light-year =			1	
1 parsec =				1

2. Define:
- an astronomical unit (AU)
 - annual parallax
 - parsec
 - light-year.
3. (a) Construct a triangle to explain the method of astronomical distance determination using trigonometric parallax.
(b) Identify the baseline in this triangle.
(c) What mathematical assumption was made in the development of this technique?
4. Using the HIPPARCOS Catalogue web site the following annual parallaxes were found. Use them to calculate the distance to each star, in parsecs.
- Achernar, $\rho = 0.0227$ arcsec
 - Acrux, $\rho = 0.0102$ arcsec
 - Aldebaran, $\rho = 0.0501$ arcsec
 - Algol, $\rho = 0.0351$ arcsec
 - Altair, $\rho = 0.1944$ arcsec
 - Antares, $\rho = 5.4$ milliarcsec (mas)
 - Arcturus, $\rho = 88.9$ mas
 - Barnard's Star, $\rho = 549.0$ mas
 - Hadar, $\rho = 6.2$ mas
 - Mira, $\rho = 7.8$ mas

5. Use the following star distances to calculate the corresponding annual parallaxes (data based on information from the HIPPARCOS Catalogue).
 - (a) Castor, $d = 15.8$ pc
 - (b) Pollux, $d = 10.3$ pc
 - (c) Rigel, $d = 237.0$ pc
 - (d) Sirius, $d = 2.6$ pc
 - (e) Bellatrix, $d = 74.5$ pc
 - (f) Betelgeuse, $d = 131.1$ pc
 - (g) Canopus, $d = 95.9$ pc
 - (h) Capella, $d = 12.9$ pc
 - (i) Deneb, $d = 990.1$ pc
 - (j) Fomalhaut, $d = 7.7$ pc
6. For each of the stars listed in question 5, convert their distance from parsecs to light-years.
7. (a) State the main limiting factor with the trigonometric technique when used with ground-based telescopes.
 (b) Discuss means that can be used to overcome this limitation.
8. (a) Compare the precision of astrometric measurements made by traditional ground-based telescopes with that of space-based telescopes.
 (b) Extend this comparison to the limits of distance determinations using each type of telescope.
9. In chapter 14, on pages 268 and 269, we looked at some highly advanced, ground-based telescopes. Discuss the improvements to astrometric measurements that these telescopes offer.

Spectroscopy

10. Define a black body, and black body radiation.
11. Describe how the theory of black bodies can be applied to stars.
12. Referring to figure 15.11 (page 281), explain the changes that occur to a black body's radiation curve as its temperature increases.
13. Use the graph of intensity versus wavelength shown in figure 15.11 (page 281) to predict the approximate surface temperature of a star with its peak intensity:
 - (a) at 1000 nm
 - (b) at ultraviolet wavelengths
 - (c) in the visible spectrum.
14. The spectrum of a star has its peak intensity at a wavelength of 400 nm. The star has a radius of 6.90×10^8 m. Using the equations on pages 281 and 282 calculate:
 - (a) the surface temperature of the star
 - (b) the total energy output per second (power output) of the star.

15. Describe the parts of a spectroscope.
16. List the three types of spectra, along with the astronomical objects that may produce each type.
17. Describe the process that produces discrete bright lines in an emission spectrum.
18. Describe the process that produces discrete dark lines in an absorption spectrum.
19. Describe the general trend of dominant lines in stellar spectra, beginning with cool stars and moving to hot stars.
20. List the names of the spectral classes, beginning with hot stars. Next to each write down the dominant spectral feature and the temperature range for each class.
21. Describe the method by which the spectrum of a star can be used to deduce information about that star's:
 - (a) surface temperature
 - (b) chemical composition
 - (c) translational velocity
 - (d) rotational velocity
 - (e) density.

Photometry

22. Describe the difference between the brightness and luminosity of a star.
23. Describe in point form three basic features of the magnitude scale as devised by Hipparchus.
24. (a) What was the definition that Pogson used to describe Hipparchus' magnitude scale mathematically?
 (b) Identify the significance of Pogson's ratio.
25. (a) If two stars differ in brightness by one magnitude, state how much brighter one is compared to the other.
 (b) If the two stars differ in magnitude by five, calculate their brightness ratio.
26. Calculate the brightness ratios of each of the pairs of stars below. Be sure to state clearly which star is the brighter.
 - (a) Achernar ($m = 0.45$) and Algol ($m = 2.09$)
 - (b) Antares ($m = 1.06$) and Arcturus ($m = -0.05$)
 - (c) Barnard's star ($m = 9.54$) and Hadar ($m = 0.61$)
 - (d) Pollux ($m = 1.16$) and Procyon ($m = 0.4$)
 - (e) Mira ($m = 6.47$) and Proxima Centauri ($m = 11.01$)
 - (f) Mira ($m = 6.47$) and Algol ($m = 2.09$)

27. The apparent magnitude of the Sun is -26.7 while that of the full Moon is -12.5 . Calculate how much brighter the Sun is compared to the Moon.
28. (a) Define apparent magnitude and absolute magnitude.
 (b) State what characteristic of a star each measures.
29. (a) If a star's apparent magnitude, m , were numerically greater than its absolute magnitude, M , what does this tell you about its distance from us? Identify the sign of the distance modulus.
 (b) If $M > m$, what does this tell you about a star's distance? Identify the sign of the distance modulus now.
 (c) If $M = m$, what can be inferred about a star's distance? Identify the value of the distance modulus.
30. Acrux has an apparent magnitude of 0.77 and an absolute magnitude of -4.19 .
 (a) Calculate how much brighter this star would be if it were located at a distance of 10 pc rather than its true distance.
 (b) Calculate the true distance of Acrux using the distance modulus formula.
31. Use the distance modulus formula to calculate the missing data in the following table.

STAR	m	M	d
Rigel	0.18	-6.69	
Bellatrix	1.64		74.5
Capella		-0.48	12.9
Sirius	-1.44		2.64
Deneb	1.25	-8.73	
Altair		2.2	5.14
Achernar	0.45		44.1
Spica	0.98	-3.55	

32. The following is an extract of data from the HIP-PARCOS Catalogue. Use the parallax formula $(d = \frac{1}{p})$ as well as the distance modulus formula to complete the table. Limit your answers to three significant figures.

STAR	PARALLAX (mas)	DISTANCE (pc)	m	M
Fomalhaut	130.08		1.17	
Vega	128.93		0.03	
Canopus	10.43		-0.62	
Betelgeuse	7.63		0.45	
Rigel Kent	742.12		-0.01	

33. Describe the steps involved in spectroscopic parallax.
34. Distances determined by spectroscopic parallax can involve a high degree of error. Identify the source of this error.
35. Fomalhaut is an A3V star while Vega is an A0V. Determine their distances using the spectroscopic parallax techniques, and then compare these distances to those calculated in question 34. Comment on any differences.
36. Canopus is an F0I star. Determine its distance using spectroscopic parallax and compare this figure to that calculated in question 32.
37. Explain the difference that occurs between apparent magnitudes determined by eye and those determined photographically.
38. Identify the three filters used in the UBV system. Include wavelengths for each filter.
39. (a) Compare an orange star's apparent magnitude as measured through a B filter, to that measured through a V filter.
 (b) Make the same comparison for a blue star.
 (c) Repeat the comparison, this time for a white star.
40. (a) Define colour index.
 (b) Construct a number line to represent the colour index scale. Upon this line write down star colours that correspond to particular values.
41. Aldebaran has a colour index of 1.538 . State its approximate:
 (a) colour
 (b) spectral class
 (c) temperature.
42. Spica has a colour index of -0.235 . Discuss what can be inferred about this star based solely upon this single piece of information.



15.1 ACCESSING STAR DATA

Aim

To access the HIPPARCOS Catalogue search facility in order to access up-to-date star data.

Apparatus

Internet access

Background information

During the period 1989 to 1993 the HIPPARCOS satellite surveyed 118 218 stars, measuring their annual parallax to an accuracy of one milliarcsec. It took until 1997 to analyse and compile the data into the HIPPARCOS Catalogue. Auxiliary instruments aboard the satellite recorded other parameters of over a million other stars, and these data have been compiled to form the Tycho Catalogue.

Method

Portions of the HIPPARCOS and Tycho Catalogues can be accessed at their search facility web site. To locate the page either perform a web site using their name as the search string, or use the weblink provided.

eBookplus

Weblink:
The HIPPARCOS
Catalogue

To retrieve information on particular stars you will first need to know the HIP number for the star. There is a link for this function, known as ‘Simbad’, at the bottom of the Research Tools page on the HIPPARCOS Catalogue site. Find this link and determine the HIP number for the star known as Procyon. (The HIP number can also be found at any number of third party database websites.) Now go back to the search page and enter this number in the field labelled ‘HIPPARCOS Identifier’. Print out the results of your search. On the printout, highlight the annual parallax of this star, then include a calculation of the distance.

There is something unusual about Procyon. Can you spot what it is?

You should now be able to retrieve information on other stars of your choice.



15.2 ANNUAL PARALLAX PRECISION

Aim

To compare the limits of annual parallax as measured by ground-based and space-based methods.

Background information

Errors can be propagated through a calculation. In general, if

$$A = \frac{1}{B}$$

then

percentage error in A = percentage error in B . Hence, since

$$d = \frac{1}{p}$$

we can say that

percentage error in d = percentage error in p .

Method

By reference to this textbook, or otherwise, determine the accuracy limit, in seconds of arc, of measurements of annual parallax made from the ground-based telescopes, as well as from the HIPPARCOS satellite. For the purpose of comparison, we will also include the anticipated accuracy of the future GAIA satellite. Enter this information into a results table as shown.

Next, determine parallax angles for which each accuracy limit represents a 1% error and record this in the appropriate spaces. Repeat this process, determining angles that would have 10% and 50% errors.

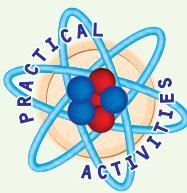
Finally, use these parallax angles to calculate the corresponding distances that would have a 1%, 10% and a 50% error in each case.

Results

ACCURACY LIMIT	GROUND-BASED TELESCOPE	SPACE-BASED HIPPARCOS	SPACE-BASED GAIA
Parallax with 1% error			
Parallax with 10% error			
Parallax with 50% error			
Distance with 1% error			
Distance with 10% error			
Distance with 50% error			

Questions

- What is the main source of error for ground-based measurement?
- How many times more accurate is HIPPARCOS than ground-based measurements?
- How many times more accurate will GAIA be than HIPPARCOS (as planned)?
- Bearing in mind that there is always some error in any measurement, what would you regard as an acceptable percentage error to an astronomical distance, if it is to be stated with confidence?



15.3 SPECTRA

Aim

To observe each of the three types of spectra — continuous, emission and absorption.

Apparatus

spectroscope
gas discharge tubes
incandescent lamp
coloured solutions

Background information

Each of the three types of spectra are produced by different types of sources as outlined in the table below.

SPECTRUM	PRODUCED BY	LAB SOURCE
Continuous spectra	Incandescent solids, liquids and high density gases	Incandescent globe
Emission spectra	Incandescent low density gases	Gas discharge tubes
Absorption spectra	Non-luminous fluid in front of a continuous spectrum	Coloured solution in front of a globe

Method

- Turn on the incandescent globe and examine its spectrum with the spectroscope. Describe and then draw its appearance using coloured pencils.

- Turn on the available gas discharge tubes in turn, allowing each time to warm up. Examine each with the spectroscope, then describe and draw its appearance.
- Turn on the incandescent globe and place a large beaker of coloured solution in front of it. Using the spectroscope, view the light through the beaker. Try several different colourings such as copper sulfate or potassium permanganate. Describe and draw the spectra you observe.

Questions

- Part (a) should have produced a continuous spectrum. What astronomical object produces this type of spectrum?
- Part (b) should have produced emission spectra. What astronomical objects produce this type of spectrum?
- Part (c) should have produced absorption spectra. What astronomical objects produce this type of spectrum?



15.4 COLOUR FILTERS

Aim

To demonstrate the use of colour filters for photometric measurements.

Apparatus

light ray kit with colour filters
data logger with light sensor or a light meter

Theory

The human eye is most sensitive to the yellow-green portion of the spectrum and this sensitivity is reflected in visual magnitudes. Early photographic emulsions were most sensitive to the blue end of the spectrum, and this is reflected in photographic magnitudes. Photoelectric photometry simulates these by the use of colour filters. The UBV system uses an ultraviolet filter (U), a blue filter (B) and a yellow-green filter (V).

Method

- Place a red filter before the light ray kit lamp, to simulate a red star. Darken the room and place your light sensor at a distance from the lamp that produces appropriate output levels. Record

the reading from the light sensor. Next, place a yellow filter (which we will use to simulate the V filter) in front of the sensor and record the output reading that results. Finally, replace the yellow filter with a blue filter (for the B simulation) and record the output reading.

- (b) Replace the red filter in front of the lamp with a blue filter to simulate a blue star. Repeat the measurements made in part (a).

Results

(a) Red star: No filter reading = _____

V filter reading = _____

B filter reading = _____

(b) Blue star: No filter reading = _____

V filter reading = _____

B filter reading = _____

Questions

1. (a) Compare the ‘no filter’ readings from parts (a) and (b) above. Were the readings the same, or did the device show a greater sensitivity to one of them?
(b) If they were different, can you suggest another explanation for the difference?
2. (a) Through which colour filter is the red star brightest?
(b) Explain how this would result in a positive colour index for a red star.
3. (a) Through which colour filter is the blue star brightest?
(b) Explain how this would result in a negative colour index for a blue star.

CHAPTER 16

BINARIES AND VARIABLES

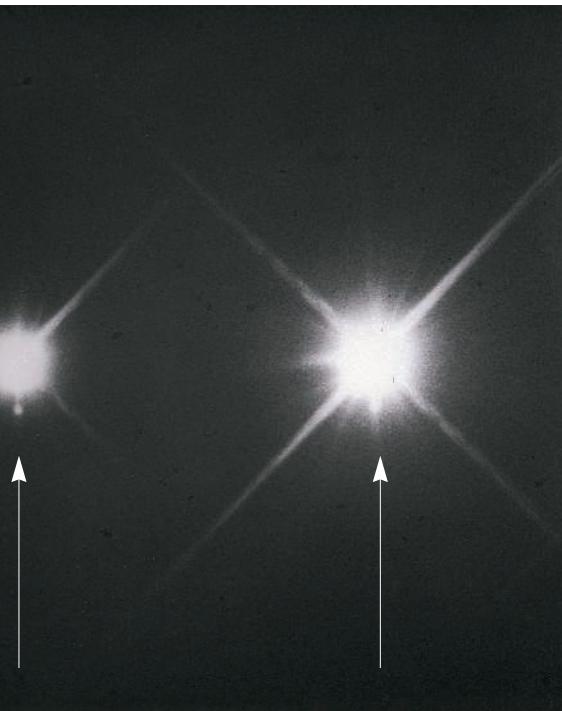


Figure 16.1 Two exposures of a binary or double star — the Dog Star, Sirius A (the brightest star in the sky, after the Sun) and the Pup, Sirius B (the first white dwarf to be discovered). Sirius B is indicated by the arrows.

Remember

Before beginning this chapter, you should be able to:

- describe the Law of Universal Gravitation
- describe the role of centripetal force in circular motion
- describe and draw a Hertzsprung–Russell (H–R) diagram and the placement of main sequence, red giants and white dwarfs upon it.

Key content

At the end of this chapter you should be able to:

- describe binary stars in terms of the means of their detection; that is, visual, eclipsing, spectroscopic and astrometric
- explain the importance of binary stars in determining stellar masses
- use Kepler's Third Law to calculate the mass of a binary star system and solve problems using
$$m_1 + m_2 = \frac{4\pi^2 r^3}{(G T)^2}$$
- classify variable stars as either extrinsic or intrinsic and non-periodic or periodic
- explain the importance of the period–luminosity relationship for determining the distance of Cepheid variables.

Binary stars are double star systems. Variable stars are stars that vary in brightness. These two apparently different stellar objects have at least two things in common. Firstly, an unresolved binary can appear to be a variable star. Secondly, the study of each type of object has made an invaluable contribution to the advancement of astronomy. The study of binaries has increased our knowledge of the mass of stars; while the study of variables has given us a reliable distance-measuring tool.

16.1 BINARIES

A binary star system consists of two stars in orbit about their common centre of mass. It may come as a surprise to learn that more than half of the main sequence stars known are not single stars, but are members of binary (double) or other multiple star systems. Binaries, in particular, are useful because their motion can be analysed to determine their masses.

To understand the significance of this we need to realise that there is no direct way to measure the mass of an isolated star. As we have seen already in studying this astrophysics option, by analysing the light from a single star we can infer a great deal about its size, temperature, luminosity, composition, density and velocity. However, none of this information will lead us directly to the knowledge of a star's mass.

In order to determine the mass of a star we need to observe its gravitational effect on another object. For instance, the mass of the Sun can be determined by analysing its effect on the motion of the Earth (or other planet). The vast majority of single stars show no such effect that can be analysed; however, binary stars do. In a sense, binaries have functioned as stellar scales because through them so much has been learned of the masses of stars that even the mass of single stars can be inferred.

Binary systems have been classified according to the way that they have been detected, placing them into the following four groups: visual, eclipsing, spectroscopic and astrometric. Bear in mind, however, that there is no physical difference between any of these binary systems.

Visual binaries

A visual binary can be resolved by a telescope; that is, a good telescope can clearly show both stars in the system. The brightest of the pair is called the primary and is designated with the letter A; the other star is the secondary and carries the letter B. Examples that are easily observed are Alpha Crucis A and B, Gamma Andromeda A and B, and Alpha Centauri A and B, although this last example is actually a triple system with the third star, Alpha Centauri C, not visible with a small telescope.

Suspected visual binaries may appear close only by line-of-sight, so it is sometimes necessary to observe the stars for many years to be sure that they are in motion around each other. Each star follows an elliptical orbit around the centre of mass of the system, with the star of larger mass tracing out a smaller ellipse. This is shown in figure 16.2.

By observing closely in order to measure the period of the motion and the separation of the stars, it is possible to calculate the total mass of the system. To see how this is possible, consider the simplified system shown in figure 16.3 in which the ellipses have been simplified to circles.

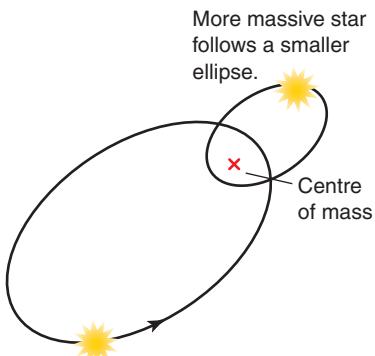


Figure 16.2 Each star follows an elliptical path around the centre of mass of the system with the more massive star closer.

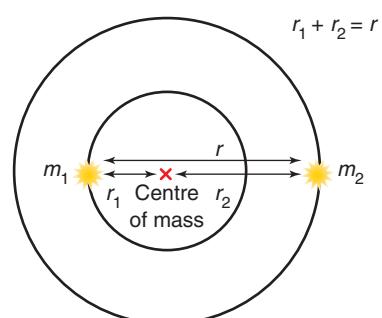


Figure 16.3 A simplified binary system

The centre of mass of the system is the point at which

$$m_1 r_1 = m_2 r_2 \quad \text{but } r_2 = r - r_1$$

so that

$$m_1 r_1 = m_2 (r - r_1)$$

$$\therefore r_1 = \frac{m_2 r}{m_1 + m_2}$$

or

$$r_1 = \frac{m_2 r}{M} \quad \text{where } M = m_1 + m_2.$$

You will recall from your work for the module ‘Space’ (page 41) that it is the gravitational attraction between each star that acts as the centripetal force that keeps each star in its orbit.

Therefore

$$F_{\text{gravitational}} = F_{\text{centripetal}}$$

$$\frac{G m_1 m_2}{r^2} = \frac{m_1 v^2}{r_1}.$$

Recall that the orbital period, T , is related to orbital speed by:

$$v = \frac{2\pi r_1}{T}.$$

Substituting this into the previous equation gives:

$$\frac{G m_2}{r^2} = \frac{4\pi^2 r_1}{T^2}.$$

The expression for r_1 above is substituted into the expression:

$$\frac{G m_2}{r^2} = \frac{4\pi^2 m_2 r}{T^2 M}$$

and therefore

$$M = \frac{4\pi^2 r^3}{G T^2}$$

where

M = total mass of the binary

system (kg) = $M_1 + M_2$

m_1 = mass of star 1 (kg)

m_2 = mass of star 2 (kg)

r = separation distance

of the stars (kg)

T = orbital period of the
binary system (s).

Note that by rearranging this formula it becomes Kepler’s Third Law:

$$\frac{r^3}{T^2} = \frac{GM}{4\pi^2}.$$

By using this equation it is possible to calculate the mass of the binary system, but not the individual star masses. In order to do that a measurement must be made of the distance from one of the stars to the centre of mass. This is not an easy measurement to make, since the inclination of the orbit relative to us must be known. However, if the measurement can be made then, by combining the following three relations (used above), the individual masses m_1 and m_2 can be calculated:

$$m_1 r_1 = m_2 r_2$$

so that

$$\frac{r_1}{r_2} = \frac{m_2}{m_1}.$$

Now, if

$$r_2 = r - r_1$$

and

$$M = m_1 + m_2$$

then

$$\frac{r_1}{r - r_1} = \frac{M - m_1}{m_1}$$

which simplifies to

where

r_1 = distance of star 1 from the centre of mass of the binary system (m).

$$m_1 = \frac{M(r - r_1)}{r}$$

This expression allows each star's mass to be calculated once its distance from the centre of mass of the system is known.

SAMPLE PROBLEM

16.1

Calculating the mass of a binary system

Sirius A and B, shown in figure 16.1, are known as the 'Dog Star' and 'The Pup'. Sirius A is a white, A-class main sequence star and Sirius B was the first white dwarf to be discovered. The system has a trigonometric parallax of 379.2 milliarcsec, which means that it lies at a distance of just 2.64 pc. The pair is observed to have a period of 18 295.4 days (just over 50 years). If their separation is 3.0×10^9 km, calculate the mass of the system.

SOLUTION

Note that $18\ 295.4$ days = 1.58×10^9 s

and 3.0×10^9 km = 3.0×10^{12} m

$$\begin{aligned} M &= \frac{4\pi^2 r^3}{GT^2} \\ &= \frac{4\pi^2 (3.0 \times 10^{12})^3}{(6.672 \times 10^{-11})(1.58 \times 10^9)^2} \\ &= 6.4 \times 10^{30} \text{ kg} \end{aligned}$$

SAMPLE PROBLEM

16.2

Calculating the mass of the stars within a binary system

The stars in a visual binary system are observed to have an orbital period of 1.8×10^8 s and are 5.0×10^8 km apart. Further measurement determines that the more massive star is 1.5×10^8 km from the centre of mass of the system. Determine:

- the total mass of the system
- the masses of each star.

SOLUTION

(a)

$$\begin{aligned} M &= \frac{4\pi^2 r^3}{GT^2} \\ &= \frac{4\pi^2 (5.0 \times 10^{11})^3}{(6.672 \times 10^{-11})(1.8 \times 10^8)^2} \\ &= 2.3 \times 10^{30} \text{ kg} \end{aligned}$$

(b)

$$\begin{aligned} m_1 &= \frac{M(r - r_1)}{r} \\ &= \frac{2.3 \times 10^{30} (5.0 \times 10^{11} - 1.5 \times 10^{11})}{5.0 \times 10^{11}} \\ &= 1.6 \times 10^{30} \text{ kg} \\ m_2 &= M - m_1 \\ &= 2.3 \times 10^{30} - 1.6 \times 10^{30} \\ &= 0.7 \times 10^{30} \text{ kg} \end{aligned}$$

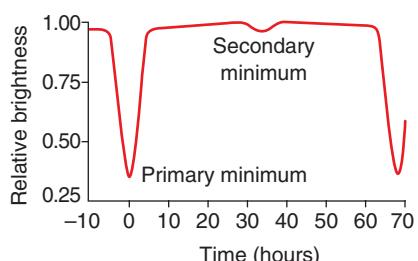
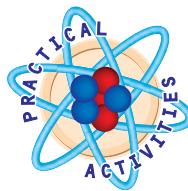


Figure 16.4 The light curve of Algol

Eclipsing binaries

An eclipsing binary is a binary system whose orbital plane is seen edge on by us. This means that at some stage through an orbit each star eclipses the other and blocks its light. These binaries are characterised by their light curve; a graph of brightness versus time. A typical example is the light curve of Algol, the first eclipsing binary discovered, shown in figure 16.4.



16.1

Eclipsing binaries

Algol A is a spectral class B8 main sequence star, while Algol B is a fainter spectral class K2 sub-giant. When the stars are side-by-side from our point of view, then the system produces maximum light. When Algol B is in front of Algol A (the primary eclipse) the brighter star is hidden and the light received by us drops significantly. When Algol A is in front of Algol B (the secondary eclipse) the dimmer star is hidden and the light received by us drops again but not as much as during the primary eclipse. As a result of these variations the light curve shows a regular pattern of asymmetrical dips and the period of the motion can easily be measured as the time between successive primary or secondary minima.

The duration of the eclipses can also reveal the diameter of each star, and in Algol's case the two stars are of very similar size (2.9 solar radii for A and 3.5 solar radii for B) although their masses are quite different (3.7 solar masses for A and 0.8 solar masses for B).

Figure 16.5 presents an example with two stars of quite different sizes. In this case a small hot star and a large cool star orbit each other. While their luminosities may be quite similar, the primary eclipse in this case is with the larger star in front of the smaller star (which produces more light than a similar sized area of the larger star).

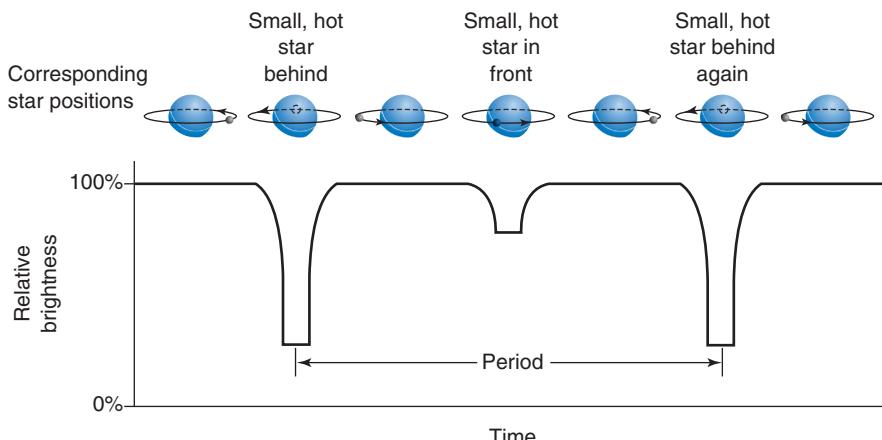
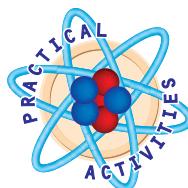


Figure 16.5 The light curve typical of a binary system with a small hot star and a large cool star



16.2

Spectroscopic binaries

A spectroscopic binary is an unresolved pair whose binary nature is revealed by alternating Doppler shifting of their spectral lines. Spectroscopic detection of a binary system is most likely if the period of the motion is short and the individual star velocities are high. Consequently, most spectroscopic binaries are close binary systems.

Figure 16.6, on the following page, shows four positions in the rotation of a binary system. The system is viewed from its edge; that is, along the plane of the orbits of the stars. When in positions 1 and 3, the stars are moving across our line of sight so that a single regular absorption spectrum is observed. However, when in position 2 star A is moving away from us, then its spectral lines are slightly red-shifted. At the same time star B is moving towards us so that its lines are slightly blue-shifted. This results in a doubling (or splitting) of the spectral lines, and a measurement of the degree of shifting can lead to the velocities of each star. In position 4 the situation is similar to position 2; however, the motions are interposed.

Regular observation of the spectrum of such binaries will reveal their period. This, in combination with the velocities of the stars, allows the circumference of the orbit, and hence the separation of the two stars, to be calculated. Kepler's Third Law can then be employed to calculate the total mass of the system.

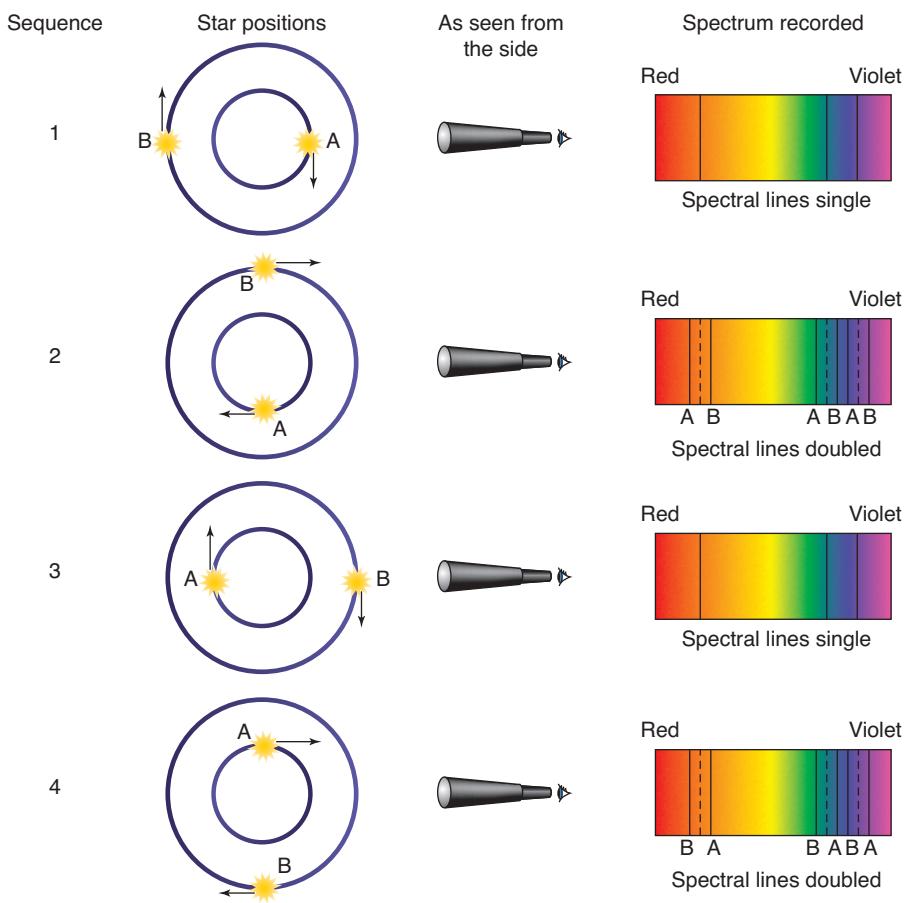


Figure 16.6 The doubling of spectral lines characteristic of spectroscopic binaries

Note that this analysis works properly only if the orbits are viewed from the edge. Therefore, it is quite possible that a spectroscopic binary can also be an eclipsing binary. Unfortunately, it is quite difficult to determine the angle of inclination of the plane of the orbit to us, and without this vital piece of information we cannot calculate the individual masses of the stars.

Astrometric binaries

In an astrometric binary one of the stars is too faint to be observed; however, the visible star can be seen to have an orbital motion. This shows itself as a detectable ‘wobble’ in the star’s proper motion, or motion relative to the rest of the sky. From this, astronomers infer the presence of the unseen partner. Astrometric measurement of the visible star’s wobble can reveal the period of the orbit as well as its size, leading to an estimation of the mass of the system and, possibly, the individual star masses.

PHYSICS FACT

The unseen partners in astrometric binaries have been found to have a very wide range of masses. Very high mass partners have been found, providing early evidence of the existence of black holes. Modern astrometric techniques (see chapter 15) have detected stars with very small wobbles indicating very small mass unseen partners, providing the first evidence of the existence of planets outside our solar system. Most of these planets have masses similar to Jupiter and many are positioned quite close to their star.

The mass-luminosity relationship

A major benefit of the study of binary systems has been the ability to determine star masses. If the luminosities of the main sequence stars are also measured and plotted against their masses, then a relationship between the two becomes apparent. Known as the mass-luminosity relationship, as shown in figure 16.7, it shows that for most main sequence stars the luminosity is proportional to the fourth power of its mass.

$$L \propto \text{mass}^4$$

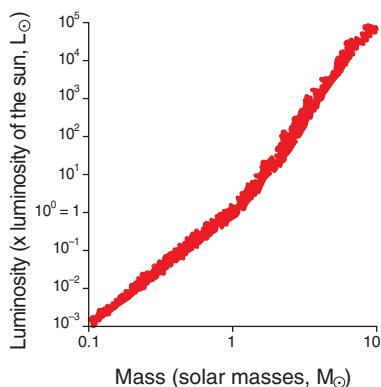


Figure 16.7 The mass–luminosity relationship for main sequence stars

This has implications for interpreting the main sequence on an H–R diagram. As luminosity increases up the sequence, so too does mass, although not by as much (because of the fourth power relationship). Also, the brighter, more massive stars have shorter lifetimes. This is because, although they are larger and have more fuel, they are also much more luminous and therefore are burning that fuel at a much faster rate. This new information about the main sequence, all flowing from the study of binaries, is summarised in figure 16.8.

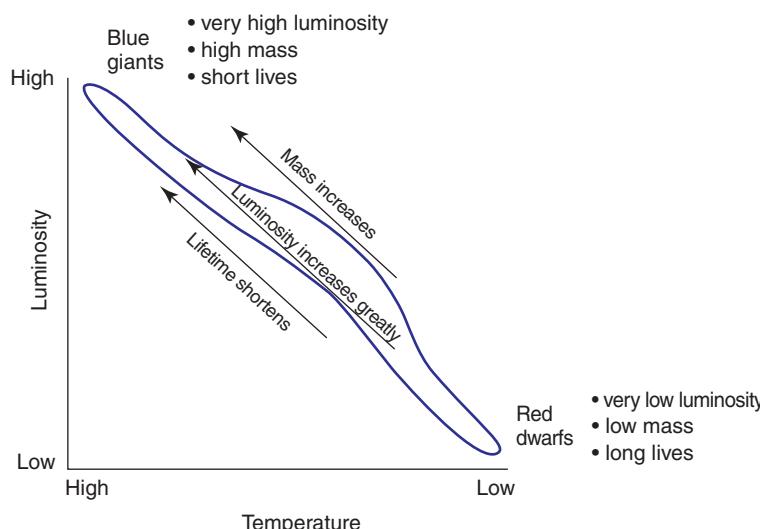


Figure 16.8 The main sequence moves from dim red dwarfs to bright blue giants. Moving up the main sequence luminosity increases, mass increases and length of lifetime decreases.

PHYSICS IN FOCUS

Naming stars

Johann Bayer (1572–1625), a German astronomer, introduced the system of naming stars in 1603. In this system a star's name is a letter (or letter combination) followed by the Latin version of the name of the constellation within which it lies. The letters are assigned in order of brightness, beginning with the letters of the Greek alphabet, followed by lower-case letters, followed by upper-case letters. Hence, Alpha Centauri is the brightest star in the constellation of

Centaurus, and Delta Cephei is the fourth brightest star in the constellation of Cepheus.

In 1862 the system was modified for variable stars, whereby the letters R to Z were reserved for this purpose, and when these were exhausted in a particular constellation then two letter combinations were used, beginning with RR, RS, RT, and so on, down to ZZ. Two well known examples are T Tauri and RR Lyrae.

16.2 VARIABLES

These are stars that vary in brightness with time. There is a variety of types of variable stars, along with a variety of apparent causes. Approximately 30 000 variable stars have been recorded to date. To put some order to this assortment of stars, they can be classified using the system shown in figure 16.9. Each type shown in this figure is discussed in the following sections.

The initial step in this system is to classify the stars as either extrinsic or intrinsic variables.

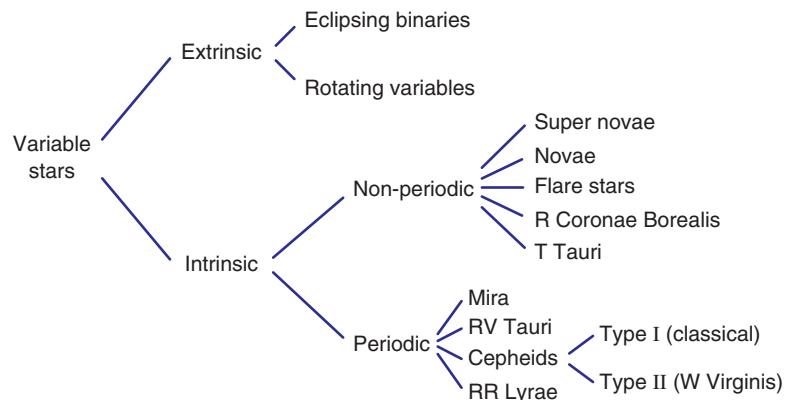


Figure 16.9 Classification of variable stars

Extrinsic variables

With extrinsic variables the variation in brightness is due to some process external to the star. This includes eclipsing binaries, already discussed in this chapter, as well as rotating variables. The latter group are stars with hotter or cooler areas on their surface that move in and out of view as the star rotates, thereby altering its brightness. Extrinsic variables are summarised in table 16.1.

Table 16.1 Extrinsic variables

TYPE	DESCRIPTION
Eclipsing binaries	Orbiting stars periodically eclipse each other.
Rotating variables	Large cool/hot spots change star's brightness as it rotates.

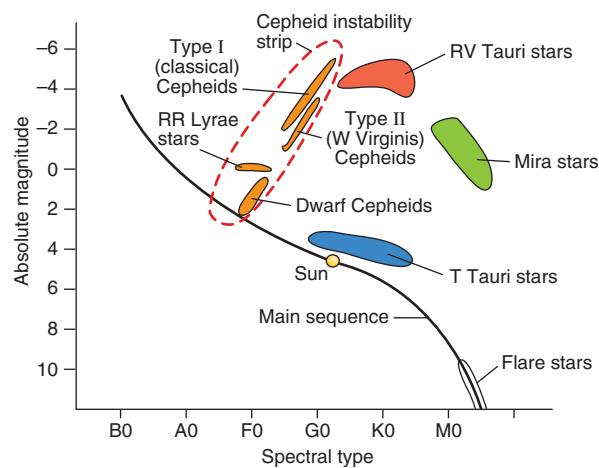


Figure 16.10 Locations of specific variables on an H-R diagram

Intrinsic variables

The brightness variation in this case is due to changes within the star itself. Many of the intrinsic variable stars occupy specific locations on an H–R diagram, and these are shown in figure 16.10. Intrinsic variables are further classified as non-periodic or periodic variables.

Non-periodic variables

These intrinsic variables show irregular variations in brightness. This group shows a variety of types, summarised in table 16.2. They include supernovae, novae, flare stars, R Coronae Borealis, and T Tauri stars.

Table 16.2 Types of non-periodic variables

TYPE	BRIGHTNESS VARIATION	DESCRIPTION
Supernovae	Temporary increase to $M < -15$ before fading away	A violent explosion destroying the star, leaving behind a compact, high density object (neutron star or black hole) and an expanding shell of gas.
Novae	Sudden increase of about 10 magnitudes before returning to normal	A close binary pair in which hydrogen-rich material is drawn from one star to the other, a white dwarf. Eventually enough material accumulates to react, creating the nova explosion. The star returns to normal though a shell of gas may have been ejected.
Flare stars (UV Ceti stars)	Sudden increase >2 magnitudes, returning to normal within an hour	Red dwarfs which experience intense outbursts of energy from small areas of their surface.
R Coronae Borealis	Sudden decrease of about 4 magnitudes, slowly fluctuating back to normal	Supergiant stars rich in carbon which periodically accumulates in the outer atmosphere, strongly absorbing light, before being blown away.
T Tauri	Irregular, unpredictable variations. Light is usually obscured by gas cloud, necessitating observations in the infra-red.	Young protostars still contracting from the gas cloud in which they lie. They are rotating rapidly and losing mass. Light variation is due to this activity in the outer layers.

PHYSICS FACT

In an extreme version of the nova process described in table 16.2, a white dwarf can accumulate too much mass from a companion star, producing a runaway reaction that leads to the star's destruction in a 'type I' supernova.

Periodic variables

Periodic variables display a regular pattern of brightness variation. The various types of periodic variables can be characterised by their light curve parameters as shown in figure 16.11.

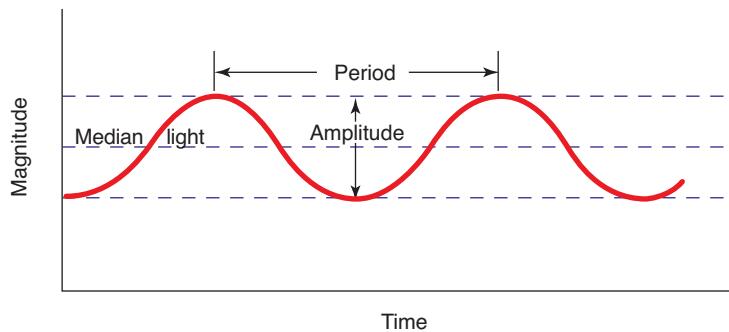


Figure 16.11 A generic light curve for a periodic variable, showing the parameters used for description

Also known as pulsating variables, the regular variation in brightness is, in general, due to a disequilibrium that exists between the two forces that act upon a star to determine its size. These two forces are the gravitational force and its radiation pressure, as represented in figure 16.12. This can occur whenever there is a change in radiation pressure.

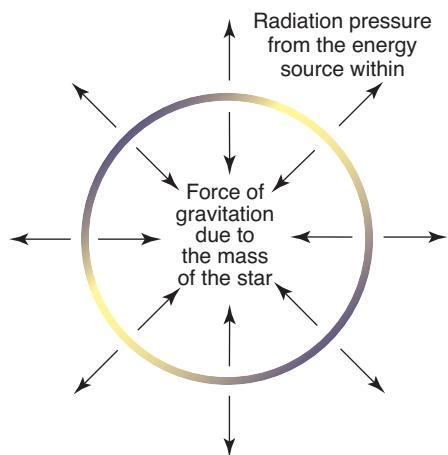


Figure 16.12 The two forces that determine a star's size are its radiation pressure pushing outwards, and its own gravitation pushing inward. When these two forces are in disequilibrium then the star will pulsate in size, temperature, luminosity and brightness.

Included in this category are Mira variables, RV Tauri variables, Cepheids and RR Lyrae variables. The properties of each are summarised in table 16.3.

Table 16.3 Periodic (pulsating) variables

TYPE	PERIOD (DAYS)	AMPLITUDE (MAGNITUDES)	MEDIAN LIGHT (MAGNITUDE)	COMMENT
Mira	80–1000	2.5–10	No typical value	Long period, pulsating red giants and supergiants.
RV Tauri	20–150	No typical value	No typical value	Yellow supergiants. Alternating deep and shallow minima on light curve.
Cepheid	1–50	0.1 to 2	−1.5 to 5	Very luminous yellow supergiants. Type I (young) and Type II (older).
RR Lyrae	< 1	< 2	0 to +1	Old giants. Always have $M \approx +0.6$.

Cepheid and RR Lyrae variables are of particular importance to astronomers as they offer another means of distance measurement.

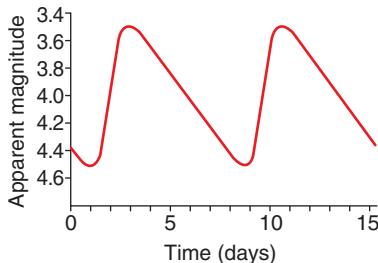


Figure 16.13 The typical light curve of a Cepheid variable

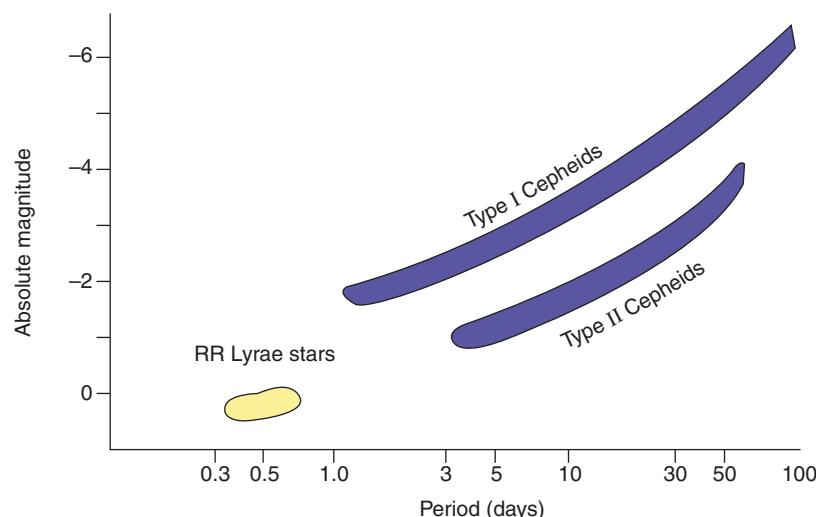
Period-luminosity relationship

From 1908 to 1912, Henrietta Leavitt studied Cepheids located in the Small Magellanic Cloud, which is one of two small irregular galaxies close to our own (the other is the Large Magellanic Cloud). The Cepheids she studied are therefore all at a similar distance from us. Cepheid variables can be recognised by their characteristic light curve, which displays a sharp increase in brightness followed by a slower decrease to complete the oscillation, as shown in figure 16.13. Henrietta Leavitt recognised that those Cepheids with longer periods of oscillation were also, on average, more luminous. This is now known as the period–luminosity relationship.

It was later discovered that there are two types of Cepheids — dubbed type I and type II. Type I (or classical) Cepheids are massive, young, second generation stars while type II (or W Virginis stars) are small, old, red, first generation stars.

The period–luminosity relationship for both types is shown in figure 16.14. It is this relationship that allows the distance to a Cepheid to be calculated. The steps in the method are:

1. Establish the type of Cepheid being observed (by spectral analysis).
2. From the light curve of the Cepheid, determine the period.
3. From the period–luminosity relationship, use the period to determine the star's average absolute magnitude, M .
4. From direct observation, measure the star's average apparent magnitude, m .
5. Use the distance modulus formula $m - M = 5 \log \left(\frac{d}{10} \right)$ to calculate the distance to the star.



This method of distance determination can be complicated by interstellar dust, which can make a star appear dimmer than it otherwise would. This will lead to a calculated distance greater than the correct value. Nevertheless, this technique has proved to be a particularly useful tool and has been used to find distances within our galaxy as well as distances to neighbouring galaxies.

Figure 16.14 The period–luminosity relationship

PHYSICS FACT

RR Lyrae variables are more populous than Cepheids and just as useful for distance measurement. Several thousand of these old giant stars are known, and their value lies in the fact that they always have a similar average absolute magnitude of +0.6. This is also shown in figure 16.14. Once a RR Lyrae variable is recognised, a measurement of its average apparent magnitude will immediately allow a distance calculation using the distance modulus formula.

SUMMARY

- A binary star system consists of two stars in orbit about their common centre of mass.
 - Binary systems have been classified into four groups — visual, eclipsing, spectroscopic and astrometric.
 - A visual binary can be resolved by a telescope.
 - If the period of the motion and the separation of the stars are known, then it is possible to calculate the total mass of a binary system by using a rearranged form of Kepler's Third Law:
- $$m_1 + m_2 = \frac{4\pi^2 r^3}{GT^2}.$$
- An eclipsing binary is a system whose orbital plane is seen edge on by us, meaning that at some stage through an orbit each star eclipses the other and blocks its light.
 - Eclipsing binaries demonstrate a light curve with a characteristic asymmetrical, double-minimum cycle.
 - A spectroscopic binary is an unresolved pair that shows an alternating Doppler shifting of its spectral lines. This creates a regular doubling of the spectral lines.
 - An astrometric binary has one star too faint to be observed; however, the visible star can be seen to have an orbital motion.
 - The study of binaries has revealed the mass-luminosity relationship, which allows masses of main sequence stars to be inferred.
 - Variables are stars that vary in brightness with time. They can be classified as either extrinsic or intrinsic variables.
 - Extrinsic variables owe their variation in brightness to some process external to the star. Intrinsic variables owe their variation to changes within the star itself. They can be further classified as non-periodic or periodic variables.
 - Non-periodic variables are intrinsic variables that show irregular variations in brightness.
 - Periodic (or pulsating) variables display a regular pattern of brightness variation.
 - Cepheids are a type of periodic variable that possesses a period–luminosity relationship. It allows a calculation of distance to the variable star.

QUESTIONS

Binaries

- Define a binary star.
- Draw and describe the orbits of a binary pair.
- Describe the nature of a visual binary.
- (a) State the measurements that need to be taken of a binary in order to calculate the mass of the system.
(b) What further piece of information is required in order to calculate the masses of the individual stars?
- Use the information in the following table to calculate the mass, in kilograms and solar masses, of the binary systems specified. Data: 1 solar mass = 1.989×10^{30} kg.

BINARY SYSTEM	PERIOD (HOURS)	SEPARATION (km)	TOTAL MASS OF SYSTEM (kg)	TOTAL MASS OF SYSTEM (SOLAR MASSES)
a	39.5	5.50×10^6		
b	52.6	8.88×10^6		
c	123	1.05×10^7		
d	426	5.89×10^7		
e	752	1.04×10^8		
f	1150	5.82×10^7		
g	3590	1.66×10^8		
h	10 500	3.94×10^8		
i	27 800	4.00×10^8		
j	43 800	5.55×10^8		

- Data required: solar mass = 1.989×10^{30} kg
solar radius = 696 000 km
 - A close binary pair is observed to have a period of just 55 hours and a separation of 27.5 solar radii. Calculate the total mass of the binary system in kilograms and solar masses.
 - The ratio of orbital radii (distance from star to centre of mass) is observed to be 3 : 5. Calculate the individual star masses.

7. Describe the nature of an eclipsing binary.
8. (a) Construct and describe the light curve of an eclipsing binary, in which one star is small and hot while the other is large and cool.
(b) Your sketch should show two unequal minima per cycle. Identify the primary minimum. Explain your choice and describe the positions of the stars at this point.
9. (a) Compare the primary eclipse to the secondary eclipse for an eclipsing binary.
(b) Explain the brightness difference between the primary and the secondary eclipses.
10. Describe the information that can be learned from the light curve of an eclipsing binary.
11. Describe the nature of a spectroscopic binary.
12. Construct a diagram to explain the spectral line doubling observed with spectroscopic binaries.
13. Describe the information that can be learned from the spectrum of a spectroscopic binary.
14. Identify the most serious limitation to the analysis of spectroscopic binaries.
15. Define the nature of an astrometric binary.
16. Describe the mass–luminosity relationship and how it was discovered.
17. Describe what the mass–luminosity relationship tells us about main sequence stars.
18. Stars at the top of the main sequence are bright and therefore massive. This means that they have more fuel to ‘burn’ compared to duller, smaller stars. Explain why they should be expected to have shorter lifetimes?

Variables

19. Define a variable star.
20. Outline the system of classification used for variable stars.
21. Define an extrinsic variable and identify some examples.
22. Define an intrinsic variable.
23. Describe the nature of a non-periodic variable. Identify some examples.
24. (a) Describe the nature of a periodic variable.
(b) Explain why these stars are also known as pulsating variables.
25. One of the most important types of periodic variable to astronomers is Cepheids.
 - (a) Describe the period–luminosity relationship of Cepheids.
 - (b) Explain why this relationship is important.
 - (c) Describe the process of distance determination using Cepheids.
 - (d) Identify when a Cepheid variable is brightest — when it is largest or smallest. Explain why. (You will need to do some extra research to discover the answer to this.)



16.1 ECLIPSING BINARIES

Aim

To model the light curve produced by an eclipsing binary using a computer simulation.

Apparatus

There are several computer programs available to perform this simulation; however, there are also several internet sites which run the simulation as a java applet. This activity utilises an applet developed at Cornell University and provided via the following weblink.

eBookplus

Weblink:
[Eclipsing binary applet](#)

Theory

If we have a side view of a binary system then the light received from it will vary as the stars eclipse each other. The primary eclipse is when the brighter of the two stars is eclipsed and it produces the greatest drop in light intensity received. This is clearly seen on the system's light curve as the

deepest of the two dips produced by the eclipses. The period of the motion is the time taken between successive primary eclipses, or twice the time between the primary and secondary eclipse.

Recall that Kepler's Third Law links the period of a binary system to its total mass and star separation.

$$\frac{r^3}{T^2} = \frac{GM}{4\pi^2}$$

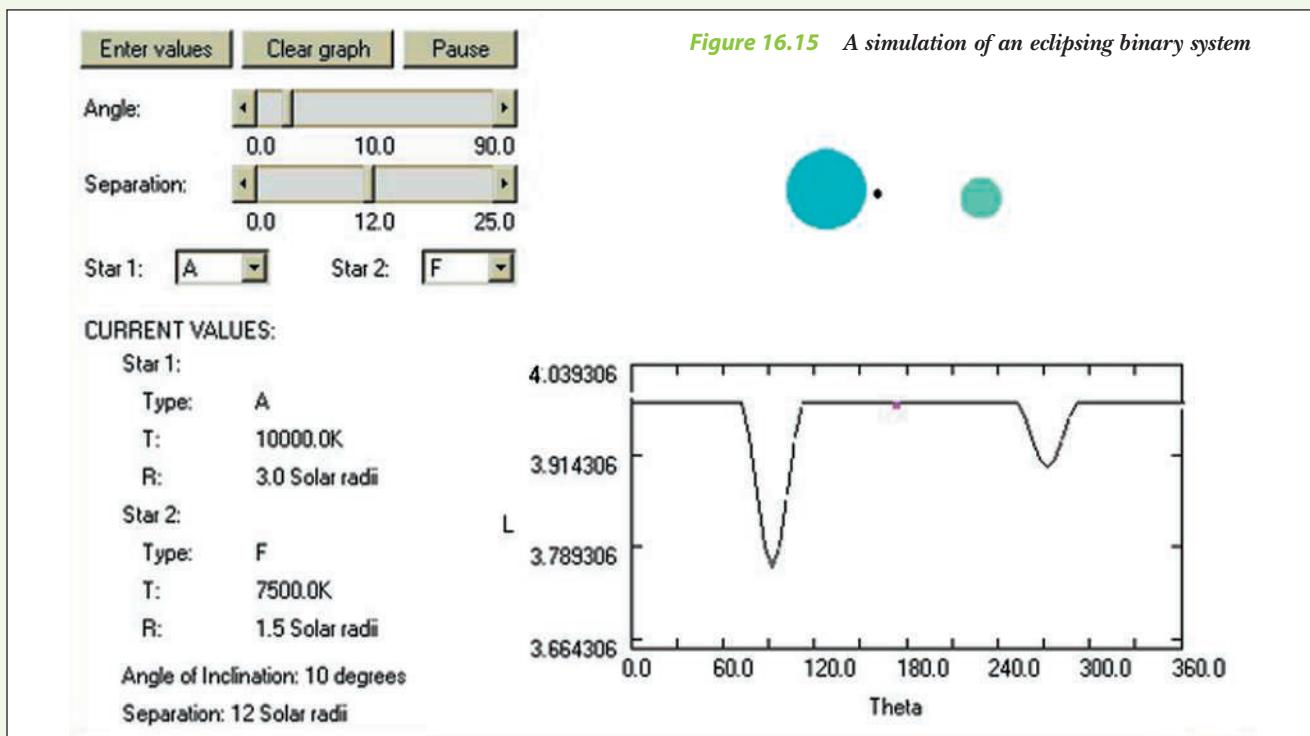
Method

Figure 16.15 shows the java applet at Cornell University. The screen shows a simulation of the orbiting stars and the light curve they produce, and a small point moves along the light curve to indicate the progress of the system. The applet provides control over the viewing angle, the star separation, and the spectral class of each star. Notice the effect of each of the following actions on the period of the motion, being sure to press the 'enter values' button after each alteration.

1. Increasing and decreasing the separation of the stars.
2. Altering the star types. The applet begins with stars that are similar but not identical. Try making the stars the same (both O, both F, both M), and quite different (one O and one M, one B and one K).

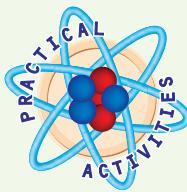
Finally, note that the light curve will alter if the system is not viewed exactly from the side. Try varying the viewing angle, and note the change in the light curve.

Figure 16.15 A simulation of an eclipsing binary system



Questions

1. How did the period change when the star separation was:
 - (a) increased
 - (b) decreased?
2. Describe and draw the light curve produced when the stars are:
 - (a) both large
 - (b) both average mass
 - (c) both small
 - (d) very different
 - (e) slightly different.



16.2 SPECTROSCOPIC BINARIES

Aim

To model the spectral line doubling observed from a spectroscopic binary.

Apparatus

Internet access. This activity utilises a java applet developed at Cornell University and provided via the following weblink.

eBook plus

Weblink:
[Spectroscopic binary applet](#)

Background information

Refer to figure 16.6. An unresolved close binary can produce regular doubling of its spectral lines as the two stars move towards and away from us. The cause of the line shifting is the Doppler effect. Regular observations can determine the period as well as the radial velocities of the two stars. This can lead to a knowledge of the mass of the binary system.

Method

Figure 16.16 shows the java applet at Cornell University. The screen shows a simulation of the orbiting stars, a graph of their velocities relative to us, and their spectral lines. As the

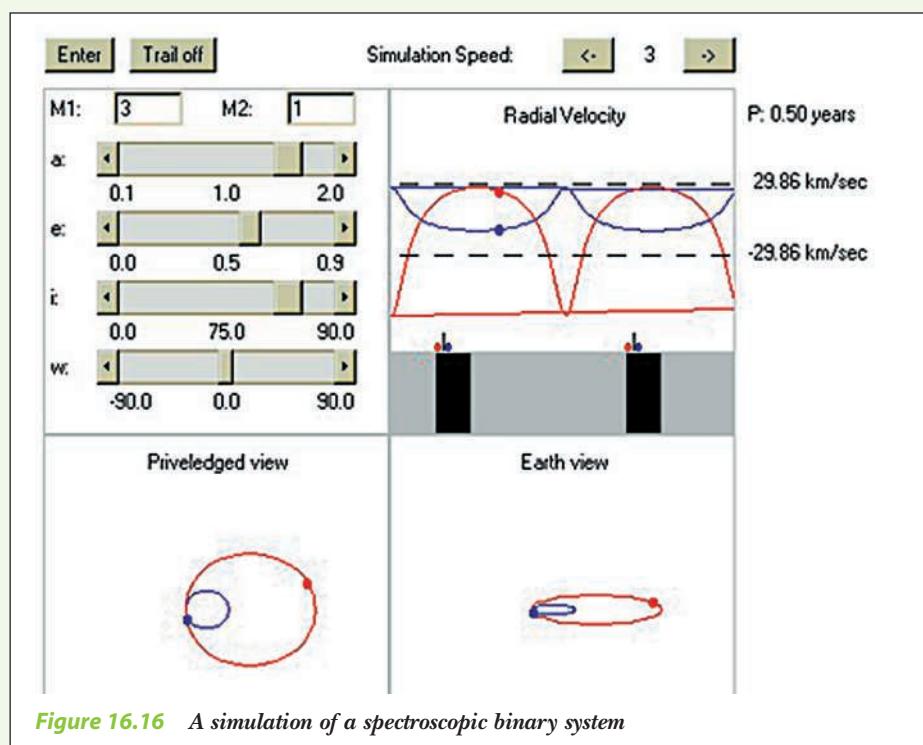
stars orbit, a small point moves along the graph to indicate the progress of the system, and the spectral lines periodically double. The applet provides control over the masses of each star (M_1, M_2), the star separation, a , the inclination, i , of the orbit to us, as well as other parameters.

Notice the effect of each of the following actions on the spectral line doubling, being sure to press the 'enter' button after each alteration.

1. Vary the masses M_1 and M_2 so that they are similar but high, similar but low, and quite different.
2. Vary the star separation a , increasing as well as decreasing it.
3. Vary the angle of inclination i between 90 and 0 degrees.

Questions

1. How did the doubling change when the star masses were:
 - (a) similar but high
 - (b) similar but low
 - (c) quite different?
2. What is radial velocity?
3. What effect does varying the star separation have on:
 - (a) the radial velocities of the stars
 - (b) the doubling of the spectral lines?
4. Describe the effect of a combination of high values for both M_1 and M_2 as well as a low star separation.
5. What happens to the line doubling effect when the angle of inclination is reduced?



CHAPTER 17

STAR LIVES



Figure 17.1 The Great Nebula in Orion is one of the very few nebulae that can be seen with a pair of binoculars. It is a region of hydrogen, ionised by the hot stars within it. The dark regions contain dust, which obscures background light. Behind the nebula lies a large molecular cloud, which is a site of star formation.

Remember

Before beginning this chapter, you should be able to:

- describe the nature of the force of gravity
- interpret a written nuclear equation, including the various symbols used
- interpret a Hertzsprung–Russell diagram.

Key content

At the end of this chapter you should be able to:

- describe the processes involved in stellar formation
- outline the key stages in a star's life
- describe the nuclear processes within a star that correspond to the stages outlined in the above point
- discuss the synthesis of elements inside stars, including the heavier elements on the periodic table
- explain how the age of a globular cluster can be determined from its turn-off point on an H–R diagram
- explain the concept of star death in relation to planetary nebulae, supernovae, white dwarfs, neutron stars/pulsars and black holes
- draw the evolutionary tracks of stars of 1, 5 and 10 solar masses, on an H–R diagram.

In this chapter we are going to discuss the life of a star. Of course, a star is not really alive — this is simply a useful analogy. Stars form from giant interstellar clouds, they experience a long period of activity and then, when their fuel is spent, they shut down. We will refer to these periods as the star's birth, lifetime and death, respectively. Each is spectacular in its own way.

17.1 STAR BIRTH

The interstellar medium

The **interstellar medium** consists of gas and dust.

The **interstellar gas** occurs as regions of neutral atoms, ions or molecules. It is mostly hydrogen.

The **interstellar dust** is made of grains of silicates and ices in a core and mantle structure, just one micrometre across.

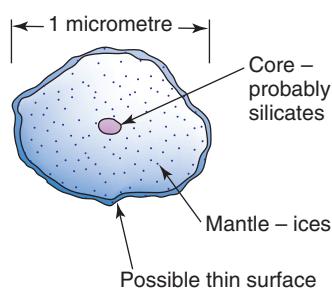


Figure 17.2 A model for the structure of an interstellar dust grain

Many people think that interstellar space, the space between the stars, is a vacuum but this is not so. This space is filled with a sparse and irregular gas as well as even more sparse grains of dust. Stars are born out of this **interstellar medium**, so we shall examine it in more detail.

The **interstellar gas** is mostly hydrogen with some helium and traces of other elements. It covers broad regions in the form of neutral atoms, charged ions or molecules, and is distributed in clouds, or 'nebulae', or intercloud gas. The hotter regions of ionised hydrogen are easily observed as nebulae around hot stars, such as the Great Nebula in Orion shown in figure 17.1. These regions absorb ultraviolet radiation given off by the stars and re-emit it at visible wavelengths. The neutral hydrogen, mostly concentrated in the plane of the galaxy, has been detected only more recently using radio telescopes, since it gives off radiation with a wavelength of 21 cm.

Most of the molecular gas is located in enormous cold clouds. Easily most common in these clouds is molecular hydrogen (about half the hydrogen in our galaxy appears to be in this form), but other molecules have been detected. These include water, ammonia, carbon monoxide, methane, ethanol and other carbon-based molecules. Occasionally these molecules collide with each other, become excited and radiate at ultra-violet, visible and infra-red wavelengths, but especially at millimetre (radio) wavelengths.

The molecular clouds are several tens of light years across, have densities of several billion molecules per cubic metre, and masses of about a thousand solar masses. Further, these clouds are arranged in huge cloud complexes, and play a vital role in the process of star birth.

The **interstellar dust** is much more tenuous — just one grain per cubic metre on average. It is thought to be formed in the outer atmosphere of cool supergiant stars before being blown away by the star's stellar wind. Each grain of dust is thought to be composed of a core and a mantle, with the core consisting of silicates (or iron or graphite) and the mantle made up of a mixture of ices (water, carbon dioxide, methane, ammonia). This structure is represented in figure 17.2.

Dust clouds can be detected because they affect any light trying to pass through them, by a reddening or extinction of the light. The reddening effect is caused by the scattering of the bluer wavelengths and is similar to the reddening of the Sun's light that we see at sunset. Extinction refers to the dimming or complete blocking of light, and is caused by scattering as well as absorption of the light. An interstellar cloud with sufficient dust to completely block the light of stars or a nebula behind it is called a dark nebula. A good example is the Horsehead Nebula shown in figure 17.3 on the following page.



Figure 17.3 The Horsehead Nebula

In the interstellar medium, dust and molecules are associated. It appears that the dust grains act as a site of molecule formation. Consequently, molecular clouds invariably contain plenty of dust. Therein lies a problem for astronomers, because molecular clouds are where stars are born but the dust blocks our view and prevents us seeing the process of star birth, at least in visible light. Infra-red and radio wave radiation penetrates the dust so that infra-red and radio telescopes can reveal what optical telescopes cannot. As we saw in chapter 14 (page 268), NASA's space infra-red telescope, the Spitzer Space Telescope, is designed for just this purpose.

Gravitational collapse

A molecular cloud that is sufficiently cool and massive will contract under its own gravity. It begins slowly but, as it draws itself in, the gravitational freefall speeds up. The density increases more quickly at its centre and, being denser, it experiences greater gravity and contracts even faster. The cloud now has two parts — a rapidly contracting core and the slower contracting surroundings.

A **protostar** is a new star before it begins to produce any nuclear energy in its core.

As the core contracts, the gravitational potential energy of its gas particles is being converted to kinetic energy, so that it heats up. This heat creates an outwards pressure that works against the gravitational collapse, but only slightly at first. As the core gets hotter and hotter, this pressure builds, slowing and eventually stopping the collapse and stabilising the size of the core, which is then called a **protostar** (shown in figure 17.4). This process takes approximately one million years.

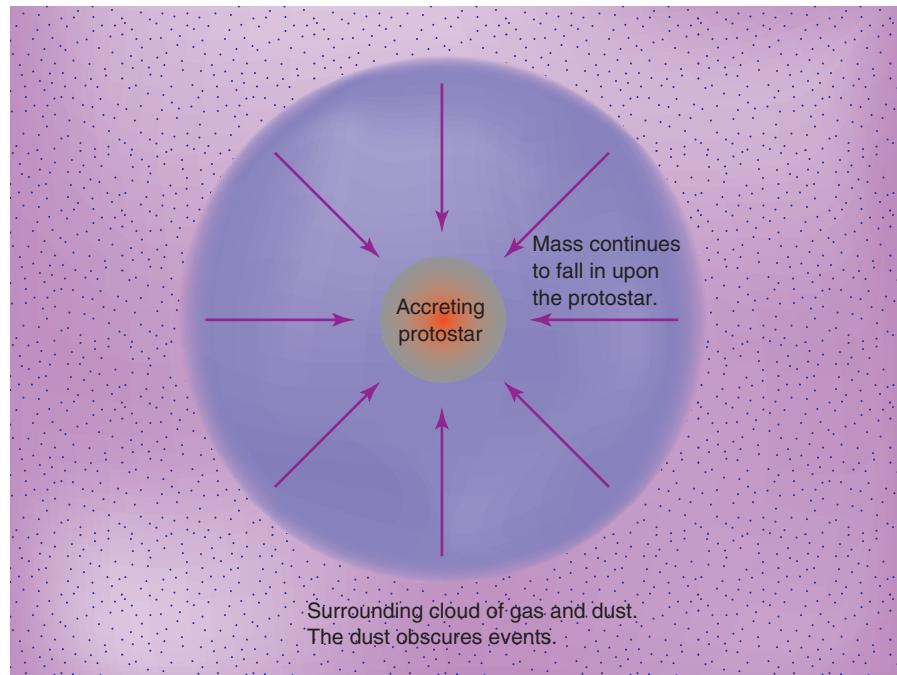


Figure 17.4 The formation of a protostar

The protostar is hidden from our view because the surrounding molecular cloud contains obscuring dust. However, this surrounding material is still contracting and continues to rain in upon the protostar. The protostar slowly increases its mass by this accretion. It then begins to behave as a T Tauri variable, developing strong stellar winds that blow away the remnants of the surrounding cloud. Finally, we can see the forming star with visible light.

With no source of energy, the protostar begins a slow shrinkage. This decrease in size causes it to become less luminous but also heats its core further. Eventually the core may reach a temperature high enough to trigger the nuclear fusion of the hydrogen within it (approximately eight million kelvin). This new long-lasting energy source stabilises the star. It is now a zero-age main sequence star — a smaller and less luminous but more stable object than the protostar it once was. Its mass is somewhere between 0.01 and 100 solar masses. Were it smaller than this, the protostar would not have heated sufficiently to begin nuclear fusion; if it were larger than this, the protostar would have overheated and blown itself apart.

Note that a plot of the main sequence using only zero-age stars is referred to as the **zero-age main sequence (ZAMS)**. It forms the complete diagonal, main sequence shape shown on most generic H–R diagrams.

So far this discussion has focused on the formation of a single star. However, the portion of cloud that has suffered gravitational collapse is usually of several solar masses and is spinning as well. As it contracts, conservation of the angular momentum of the spinning cloud makes it spin faster, and this causes the cloud to fragment into smaller spinning parts, so that a group of stars is formed. Each smaller part is also spinning, which eventually causes further fragmentation and leads to systems of planets around the stars.

The **zero-age main sequence (ZAMS)** is a plot of the main sequence using only zero-age stars.

PHYSICS FACT

Angular momentum is the rotational equivalent of linear momentum. When a spinning object becomes smaller (not less massive), it must spin faster in order to conserve its angular momentum. This idea is used by divers, gymnasts and figure skaters who spin faster when they tuck in their arms or legs.

The process of star birth described can be traced on an H–R diagram, but the pathway is slightly different for different mass stars, as shown in figure 17.5. For a star of approximately one solar mass, the process takes about 50 million years.

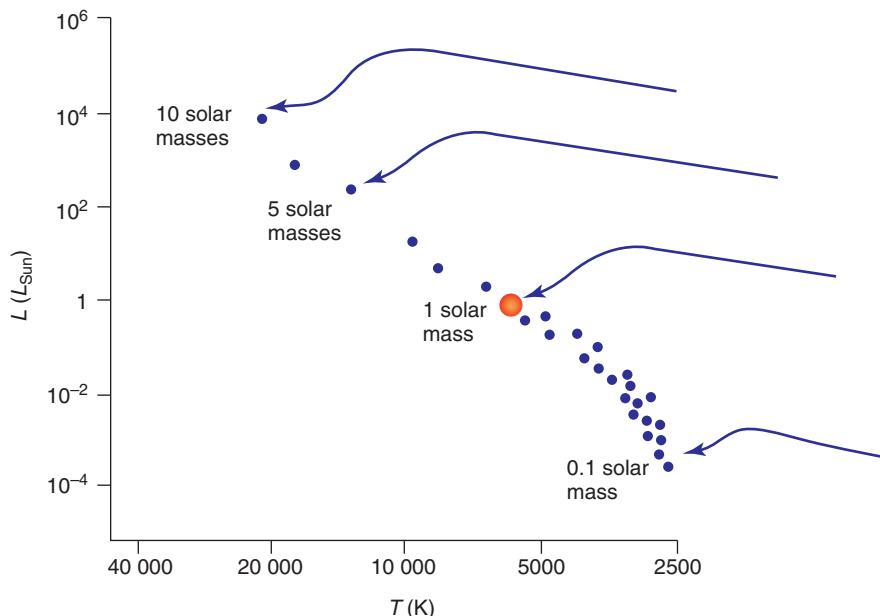


Figure 17.5 The pathways of stellar birth on an H–R diagram

17.2 MAIN SEQUENCE STAR LIFE

A **main sequence star** is characterised by the fusion of hydrogen to helium in its core, surrounded by unused, or non-reacting, hydrogen layers. The stability of the star comes from the equilibrium it has achieved, both hydrostatic and thermal. The hydrostatic equilibrium is the balance between the outward radiation pressure and the inward gravitational force, as represented in figure 16.12. The thermal equilibrium refers to the balance between the rate at which energy is produced in the core of the star and the rate at which energy is radiated away from the surface.

Hydrogen ‘burning’

The source of the energy in the core of the star is the fusion of hydrogen. This is commonly referred to as ‘hydrogen burning’, although it must be remembered that this process is not the chemical reaction of combustion. Rather, it is the joining, under high temperatures, of hydrogen nuclei to form helium nuclei. This nuclear reaction results in a slight decrease in mass, and the lost mass is transformed into the energy released in accordance with Einstein’s equation, $E = mc^2$. The net reaction can be written as follows:



where

e^+ = positron (positive electron)

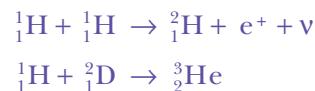
ν = neutrino (small, massless, chargeless particle).

While this is the net reaction in all main sequence stars, there are two different mechanisms to achieve it.

The proton–proton chain

The **proton–proton (p–p) chain** is the hydrogen fusion mechanism that is first to occur in main sequence stars.

The first is known as the **proton–proton (p–p) chain**, and it is the first to start in all stars when they reach the main sequence. This process begins with the following two reactions:



where

H_2^2 = deuterium (heavy hydrogen)

He_2^3 = light helium.

These two reactions must both occur twice before the final reaction can take place:



Note that six H_1 's go into the reactions but two are returned so that the net reaction has four hydrogens combining to produce a helium. In figure 17.6 we have tried to represent this chain in a flow diagram.

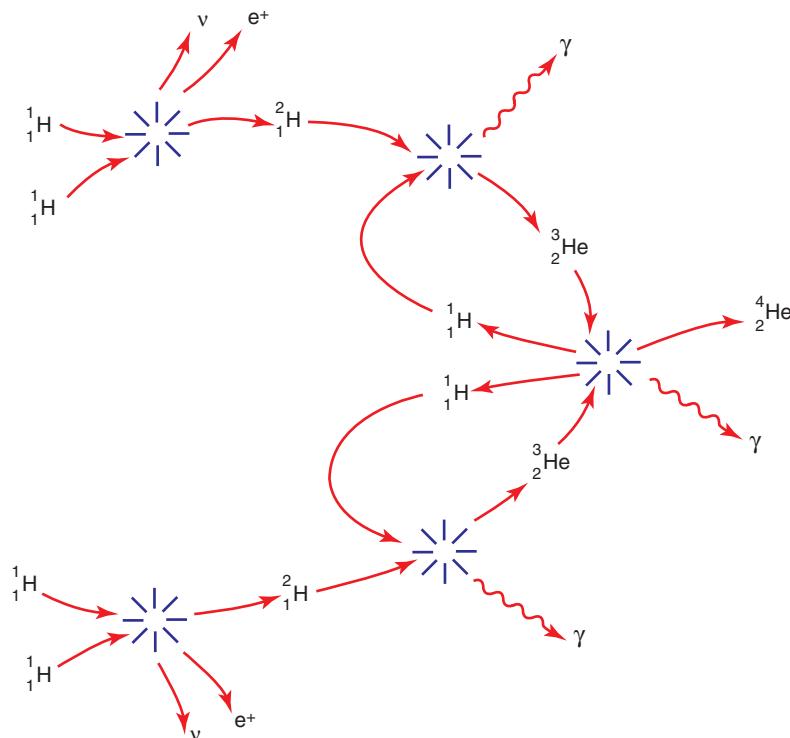


Figure 17.6 The proton–proton chain

The CNO cycle

The **CNO cycle** is the hydrogen fusion mechanism that dominates in hotter main sequence stars.

There is another mechanism present, known as the **carbon–nitrogen–oxygen (CNO) cycle**, although in smaller, cooler stars it does not produce much energy. However, in stars more massive than the Sun the core temperature exceeds 1.6×10^7 K, and at approximately this temperature the CNO cycle becomes the dominant process.

The CNO cycle is a six-stage process in which carbon acts as a catalyst. That is, a $^{12}_6\text{C}$ nucleus goes into the first reaction and is returned in the last. The reactions are listed below in the order they occur, and are drawn in figure 17.7 as a cycle. Note that the net reaction is still that four hydrogens combine to produce a helium.

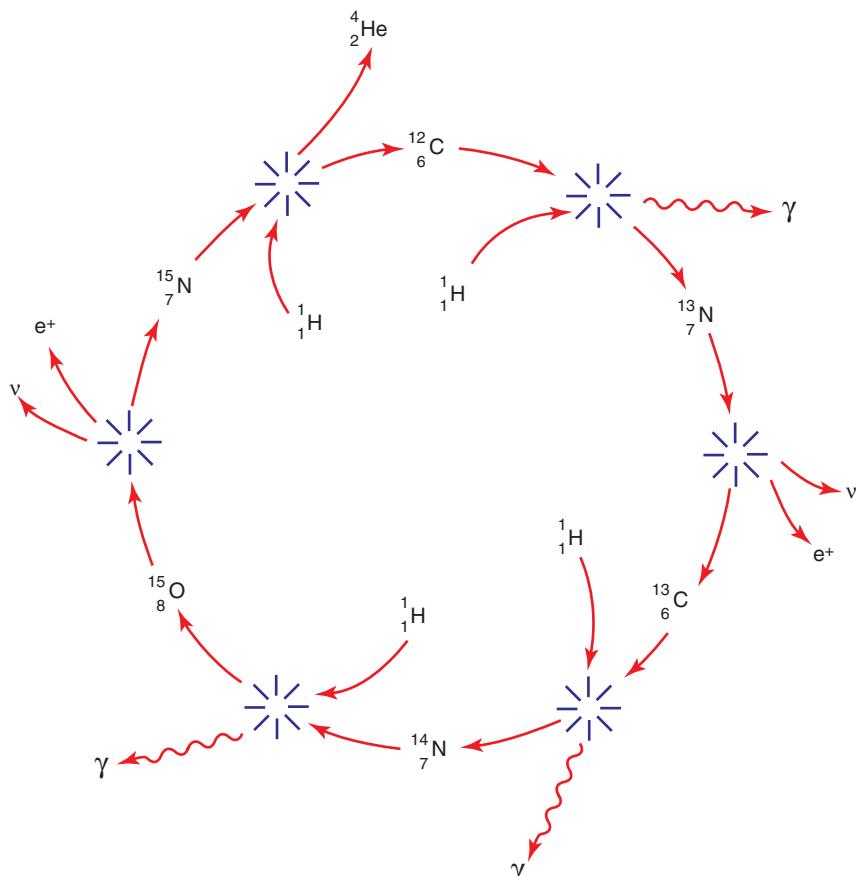
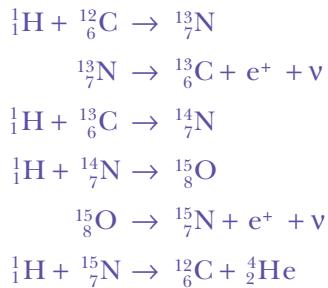
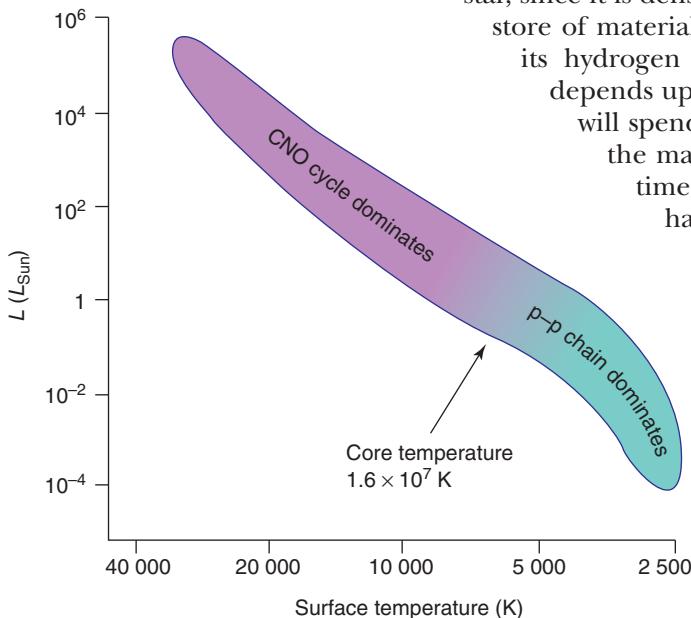


Figure 17.7 The CNO cycle

Both of the above mechanisms, the proton–proton chain and the CNO cycle, can occur simultaneously within a star. However, in less massive, cooler stars the p–p chain is dominant, while in more massive, hotter stars the CNO cycle dominates. This is represented in figure 17.8 on the following page.



The helium produced by hydrogen burning collects at the centre of the star, since it is denser than the hydrogen. Here it accumulates, building a store of material that will become the star's next energy source when its hydrogen supply finally runs down. How long this will take depends upon the mass of the star. A star of about one solar mass will spend approximately 10 billion years burning hydrogen on the main sequence, whereas a high mass star will have a lifetime of just a few million years. Many low mass stars have had lifetimes as long as the universe is old. (We discussed the relationship between a star's mass and its lifetime in the previous chapter.)

Whatever the star's mass, it does eventually run out of hydrogen fuel in its core. At this point, the star's life on the main sequence is over and it is about to become a red giant.

Figure 17.8 The lower end of the main sequence is populated by smaller, cooler stars in which the proton–proton chain dominates. In the more massive, hotter stars in the upper main sequence the CNO cycle dominates.

17.3 STAR LIFE AFTER THE MAIN SEQUENCE

A **red giant** is a star characterised by a helium-burning core surrounded by a hydrogen-burning shell. Let us look at how this structure develops, as represented in figure 17.9.

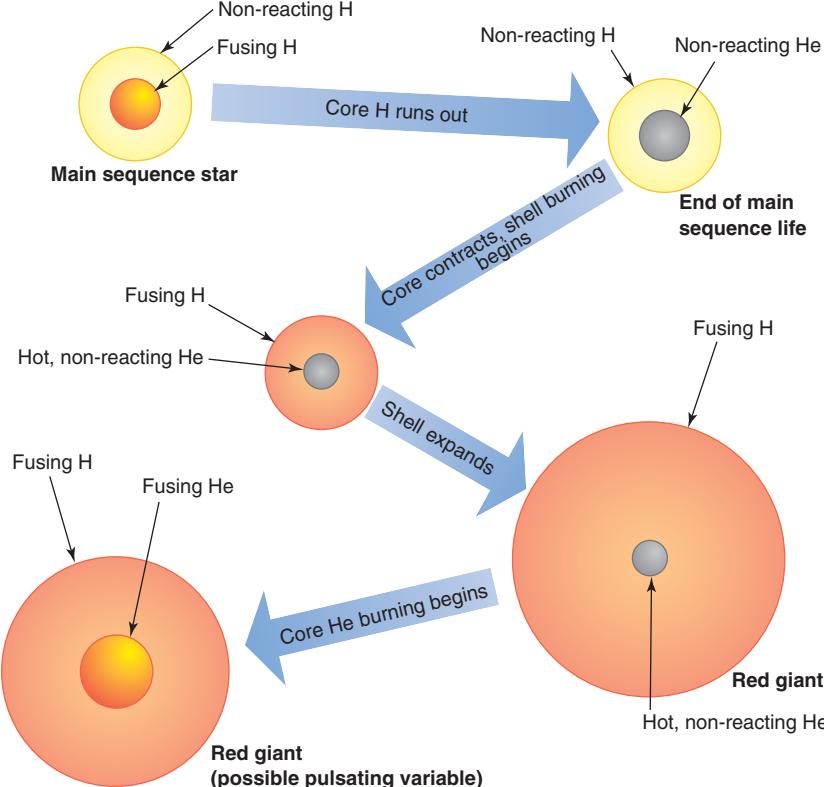


Figure 17.9 The developing layers within a star during the transition from main sequence to red giant

At the end of its main sequence life, a star's core is almost completely non-reacting helium. It is surrounded by a shell of unused hydrogen. With no radiation pressure to support the core, it begins to contract under gravity. This contraction heats up the core and also the shell surrounding it. Hydrogen fusion to helium begins in the shell, producing energy and increasing the luminosity of the star. Under its own radiation pressure, the shell begins to expand and this expansion causes the surface temperature to decrease. The shell expansion continues until it is enormous and the surface temperature is comparatively cool. The star is now a red giant and on an H–R diagram it has shifted up and to the right of its former main sequence position.

Meanwhile, the non-reacting helium core has been heating up. If the star is less than approximately 0.5 solar masses, it is doubtful that it will ever reach sufficient temperature for anything further to ignite. Such small mass stars are already near the end of their life.

If the star's mass is higher the core will reach a sufficient temperature for helium fusion to begin. High mass stars have hotter cores, which achieve this ignition smoothly. However, in the core of stars of intermediate mass the helium fusion begins very suddenly. This is referred to as a **helium flash**. The star adjusts to the new energy source by reducing its radius and luminosity slightly, moving down and to the left on the H–R diagram, towards the Cepheid instability strip. It is at this point that the hydrogen-burning shell can become sufficiently unstable to cause the star to pulsate as a periodic variable, driven by the changed radiation pressure within.

The transition of stars from the main sequence across to the red giants, as represented on an H–R diagram, has been summarised in figure 17.10. Note that very massive stars move across to the super giants rather than the red giants.

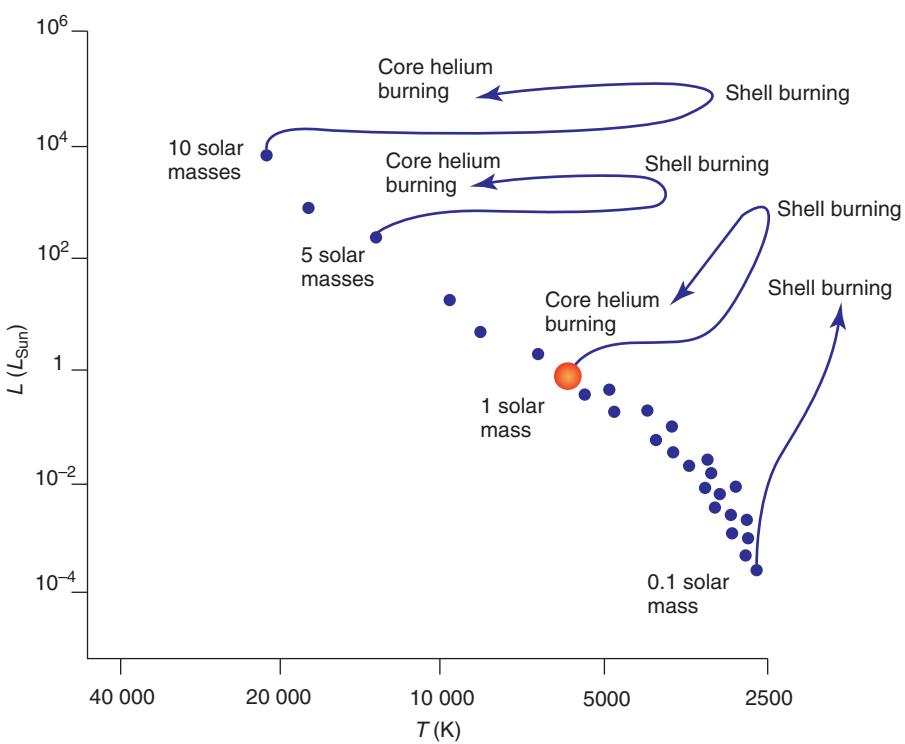


Figure 17.10 The transition from main sequence to giant, as represented on an H–R diagram

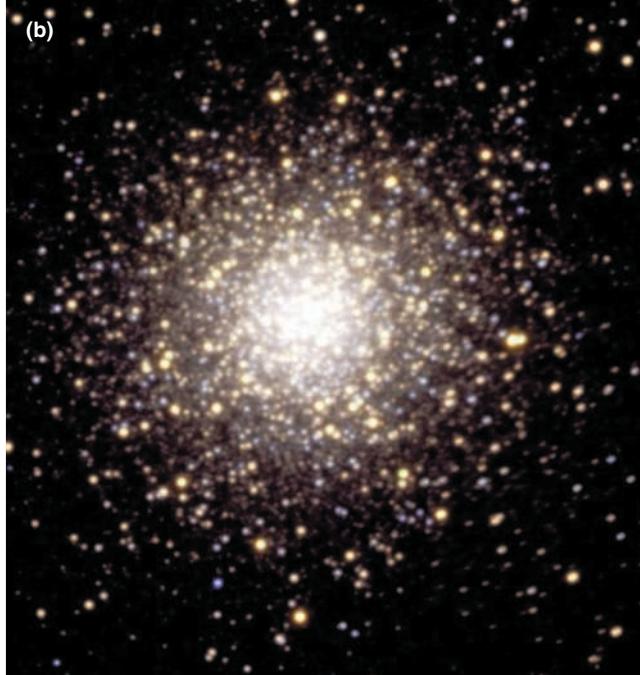
PHYSICS IN FOCUS

Evidence for the main sequence to red giant transition

Earlier in this chapter we discussed how protostars contract out of giant molecular clouds. These clouds have sufficient mass to form many hundreds of thousands of stars, and often stars will be formed in clusters. There are two distinct types of clusters that can be observed — open (or galactic) clusters and globular clusters. Examples of both are shown in figure 17.11.

One obvious difference between these two types is that globular clusters contain many more stars; however, a less obvious difference is that open clusters contain spectral class O and B stars, whereas globular clusters do not. Recall that these hot, massive stars have short lifetimes. This means that open clusters are younger than globular clusters.

Figure 17.11 (a) An open cluster (the Pleiades) and (b) a globular cluster (M3)



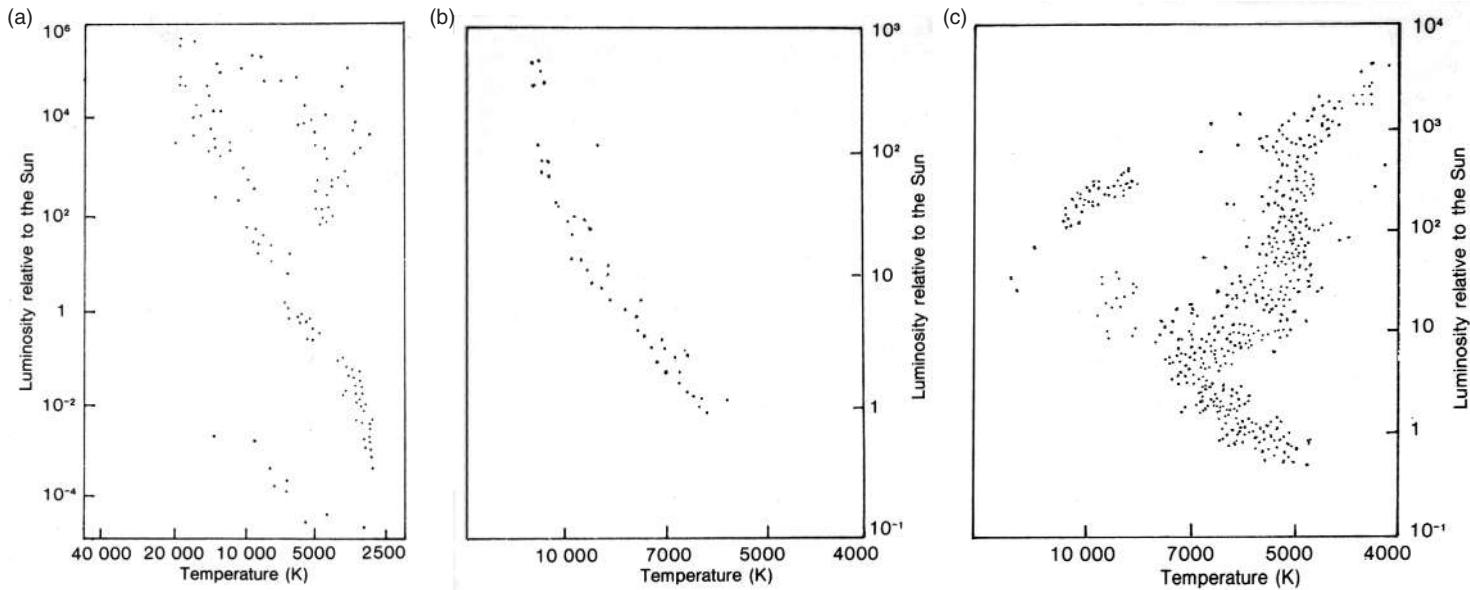


Figure 17.12 H–R diagrams of (a) the nearest and brightest stars, (b) an open cluster such as the Pleiades, and (c) a globular cluster such as M3. Notice that in (c) the top of the main sequence is missing as these stars have evolved and shifted to the red giant zone.

This difference is revealed also in an H–R plot of the stars within a cluster. Look first at figure 17.12(a), which is an H–R plot of a sampling of nearest and brightest stars. This represents a random sampling of star types and, not unexpectedly, each of the prominent star groups is represented. However, clusters are not a random sample, because all the stars within a cluster were formed at much the same time and so they are all of approximately the same age.

Look now at figure 17.12(b). If the stars within an open cluster are catalogued and plotted on an H–R diagram, it looks like this. We can see that they occupy almost the entire zero-age main sequence. If the same exercise is then performed with the stars within a globular cluster, it looks like figure 17.12(c), and this time the top of the main sequence is missing. However, there are now stars occupying the red giant region of the H–R diagram. This indicates that the missing stars have already moved on to become red giants and have shifted to the right.

The highest remaining point of the main sequence group is called the turn-off point. If this exercise is repeated for many clusters and a composite H–R diagram is constructed, it looks like figure 17.13. Notice that each cluster shows a different turn-off point, and this can be used to infer the age of a cluster.

As a cluster ages, the main sequence shortens from the top as the stars progressively evolve into red giants in order of their mass. The result is that the position of the turn-off point acts as an indication of a cluster's age. Using this method, the oldest clusters appear to be almost as old as the universe (12–15 billion years) while the Pleiades is estimated to be just 100 million years old.

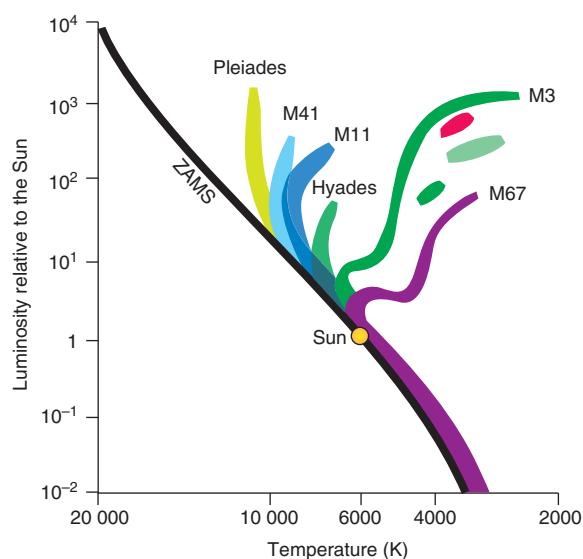


Figure 17.13 A composite H–R diagram of several clusters. Note that the lower the turn-off point, the older the cluster.

The **triple alpha reaction** is the process of helium fusion in the core of a red giant.

The triple alpha reaction

The fusion of helium in the core of a red giant star proceeds by the process known as the **triple alpha reaction**. You should recall that an alpha particle is a helium nucleus, so that in this reaction three helium nuclei combine to form a single carbon atom. The equation is as follows:



The carbon atom can easily fuse with another helium nuclei to form oxygen, using the following reaction:



Note that the triple alpha reaction produces just 10% of the energy per kilogram of fuel compared with hydrogen burning. The fuel is used up quickly so that the time a star spends as a red giant may be just 10–20% of its prior life as a main sequence star.

Post-helium burning

When the star has exhausted its supply of helium in the core, the fusion reactions cease there. The core is now largely composed of non-reacting carbon and oxygen, although hydrogen fusion is still going on in the shell. What happens from this point depends upon the star's mass.

A star of one solar mass is near the end of its energy-producing life. The still-fusing shell expands and becomes unstable, pulsating irregularly and shedding material, already in its death throes.

However, larger mass stars still have some life left in them. The non-reacting carbon contracts under gravity, heating up and igniting a helium-burning shell just below the hydrogen-burning shell. This new shell-burning causes the star to expand again, moving diagonally up and right on the H-R diagram. This helium-burning shell is turbulent and, unsupported from beneath like this, can make the star pulsate as a non-periodic variable.

If the star is larger than about five solar masses, then its core becomes hot enough to begin fusion of carbon to neon and magnesium, possibly starting quickly in a 'carbon flash'.

When the carbon supply is exhausted, a very massive star may proceed further. As each energy source runs out the core contracts under gravity and heats up. This ignites the element that has been produced by the shell immediately above it, creating a new shell of energy production (see figure 17.14). The core contracts further, heating sufficiently to ignite a new, heavier energy source (produced by previous fusions). Oxygen can be fused to silicon and sulfur, and the most massive stars are able to fuse these to form an iron core, but here the reactions must stop. This is because the fusion of iron, or any element heavier, consumes energy rather than producing it. Eventually, however massive the star, it finds itself at the end of its life, unable to initiate any new energy source.

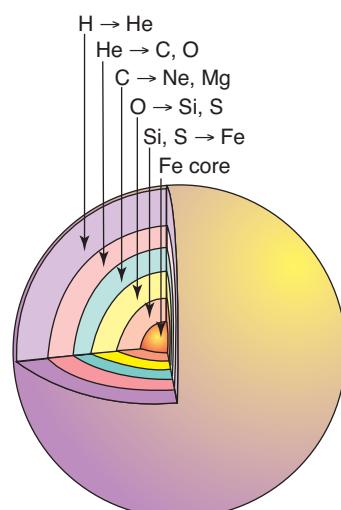


Figure 17.14 A very massive star can develop many layers of shell burning as it finds successively heavier core fuels each time the core contracts and heats up. When it develops an iron core, however, it can go no further.

PHYSICS IN FOCUS

The nucleosynthesis of heavy elements in stars

By now you may have wondered how heavier elements such as gold or uranium are manufactured. We have already seen that hydrogen, helium and lithium were created in the big bang, that main sequence stars fuse their hydrogen into helium, that red giants fuse their helium into carbon and oxygen, and that massive stars are able to continue to fuse these elements further to form heavier elements again. The most massive stars are able to fuse elements up to iron to produce energy, and no further. But iron is just number 26 on the periodic table, which contains over 100 different elements. So how are the elements heavier than iron created?

There are two different processes for the manufacture, or nucleosynthesis, of these

elements. Both processes require an input of energy and a supply of neutrons.

The first process is the slow capture of neutrons inside red giants that have achieved a helium-burning shell. The neutrons are captured by nuclei to form heavier ones. This slow process is capable of generating elements up to lead on the periodic table, including gold.

The second process is the fast capture of neutrons in a supernova explosion. In this environment there is sufficient energy available to allow the rapid formation of the elements heavier than lead, such as uranium.

The two processes complement each other to provide the wide range of elements found here on Earth and listed in the periodic table.

17.4 STAR DEATH

Sooner or later, a giant star develops a core of material that it cannot fuse (either because it cannot get it hot enough to fuse, or the core is iron which will not fuse to produce energy) surrounded by still-fusing shells. From this point the death of every star follows a similar pattern: the shells will be shed into space and the core will collapse under gravity. The way in which this happens depends upon the mass of the star, as represented in figure 17.15.

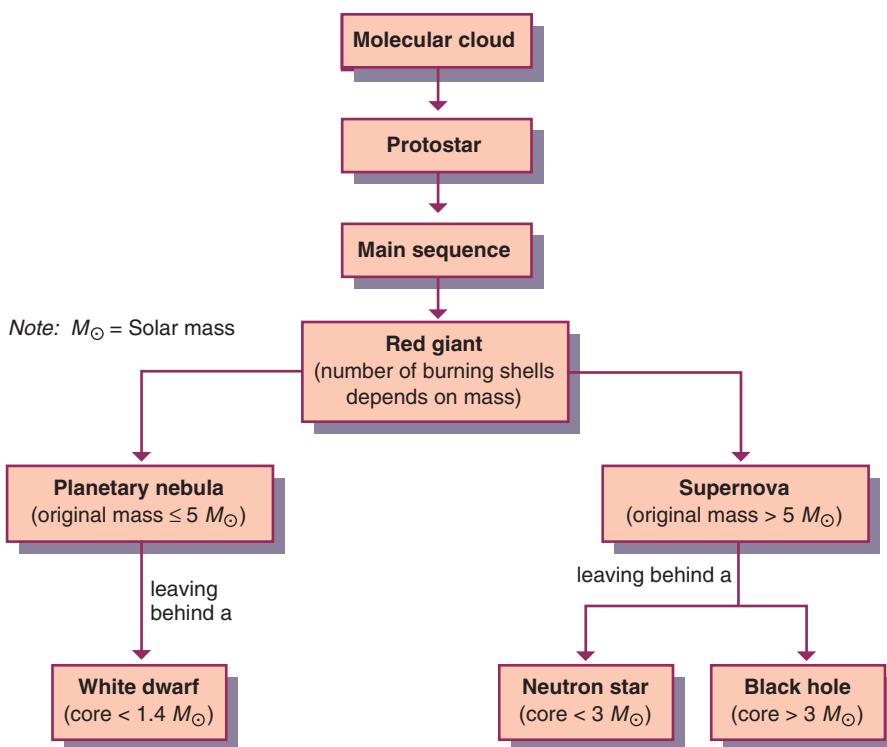


Figure 17.15 A schematic diagram showing the evolution of a star as a function of its mass. Note that all stars follow the same general pattern.

Stars of approximately five solar masses or less

The unsupported shells are unstable, producing bursts of energy known as thermal pulses, as well as extraordinarily high ‘superwinds’. These combine to blow material rapidly away from the star and disperse the shells until all that is left is the core. The dispersed material is initially in the form of an expanding shell-shaped nebula around the core, known as a **planetary nebula**. (An historical name since the nebula can look like a planet through a small telescope.) Seen from our perspective this shell looks very much like a ring, as the photograph of the Ring Nebula in figure 17.16 shows.

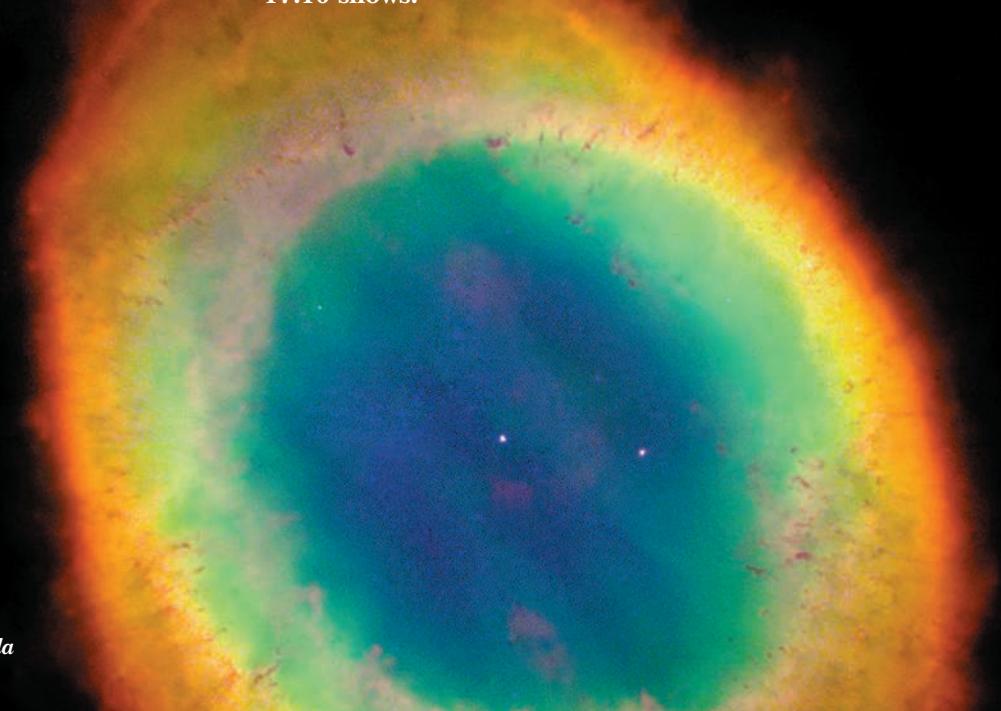


Figure 17.16
The Ring Nebula

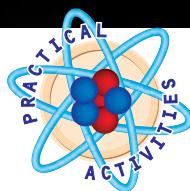
A **white dwarf** is a dense star made of degenerate matter. It is the end point of small- to medium-sized stars.

The **Chandrasekhar limit** (1.4 solar masses) is the greatest mass that a non-rotating white dwarf can have.

The core, which has a mass less than 1.4 solar masses, collapses under the force of gravity to form a white dwarf. This is a very dense star (about 10^9 kg/m^3), which means that a star the size of our Sun would crush down to the size of the Earth. The white dwarf is composed of ‘degenerate matter’ — a form of highly crushed matter. The pressure of the high-speed electrons within this matter (‘electron degeneracy pressure’) is all that is left to act against gravity and stabilise the star’s size.

The mass limit of 1.4 solar masses is an important one to note. Known as the **Chandrasekhar limit**, it is the greatest mass that a non-rotating core can have and still become a white dwarf. Rotation increases the limit somewhat. If the mass of the core is greater than this, even electron degeneracy pressure is not enough to hold back the gravitational collapse.

Eventually the planetary nebula disperses into space and the white dwarf simply cools down to form a stellar corpse known as a black dwarf.



17.1

Researching stellar objects

Stars of more than five solar masses

The shells of massive stars can develop multi-layered structures; however, they still experience the unstable pulsations and superwind of the less massive giants. The difference here, though, is the mass of the core, being greater than 1.4 solar masses. The core begins to collapse under the force of its own gravity, but this time electron degeneracy pressure will not stop it. Just how far this crush goes depends again on the mass of the core.

A **neutron star** is the extremely dense remnant of the core (1.4 to 3 solar masses) of a massive star. It is composed of neutron matter.

A **black hole** is the crushed remnant of the core (greater than 5 solar masses) of a very massive star. Theoretically, it is a point of zero volume and infinite density.

A **supernova** is a violent explosion of uncontrolled nuclear reactions that completely blows away the various layers of a massive star (original mass greater than five solar masses).

If the core is less than approximately three solar masses (but greater than 1.4 solar masses) then the matter is crushed to such an extent that electrons and protons are forced together to form a sea of neutrons, and it is the neutron degeneracy pressure that finally halts the collapse. The core has become a **neutron star**, with a density of approximately $10^{17} \text{ kg m}^{-3}$ and just 10 to 15 km in diameter.

If the core is greater than approximately three solar masses, nothing can stop its gravitational collapse to a **black hole**. The matter is crushed down to a point of infinite density, known as the singularity. The gravity of this point is so great that nothing, including light, can escape it from within a certain radius, called the ‘event horizon’. It is this characteristic that makes black holes black.

In both of these cases, the collapse of the core draws in the remaining gases of the shells of the star and they rebound from this implosion with a **supernova**. This is a violent explosion of uncontrolled nuclear reactions that completely blows away the material that was the various layers of the star, leaving behind just the highly dense core.

PHYSICS IN FOCUS

Pulsars

Neutron stars spin very rapidly, up to 600 revolutions per second. This is because the angular momentum of a comparatively slowly spinning giant’s core is conserved as it shrinks to a neutron star. This is analogous to a spinning ice skater who pulls in her arms to spin even faster.

In addition, these stars possess very strong magnetic fields. Such strong, quickly rotating magnetic fields result in the emission of a beam of electromagnetic radiation from each magnetic pole, as shown in figure 17.17. The beam traces out the surface of a cone as the star rotates. If the Earth happens to be intercepted by one of these beams, we will ‘see’ a very regular pulsation of radiation each time the beam swings past.

In 1967, while a research student at Cambridge University, British radio astronomer Jocelyn Bell (1943–) discovered just such a repetitive, pulsating source of radio waves for the first time. They were dubbed ‘pulsars’, and confirmation of their connection with neutron stars came the following year with the discovery of a pulsar at the centre of the Crab Nebula. This nebula, shown in figure 17.18 on the following page, is the remnant of a supernova observed by Chinese and Japanese astronomers in 1054.

There are currently over 500 pulsars known, with periods ranging from 1.54 milliseconds to 4 seconds. The radiation they emit can occur at radio, optical, X-ray as well as gamma ray wavelengths.

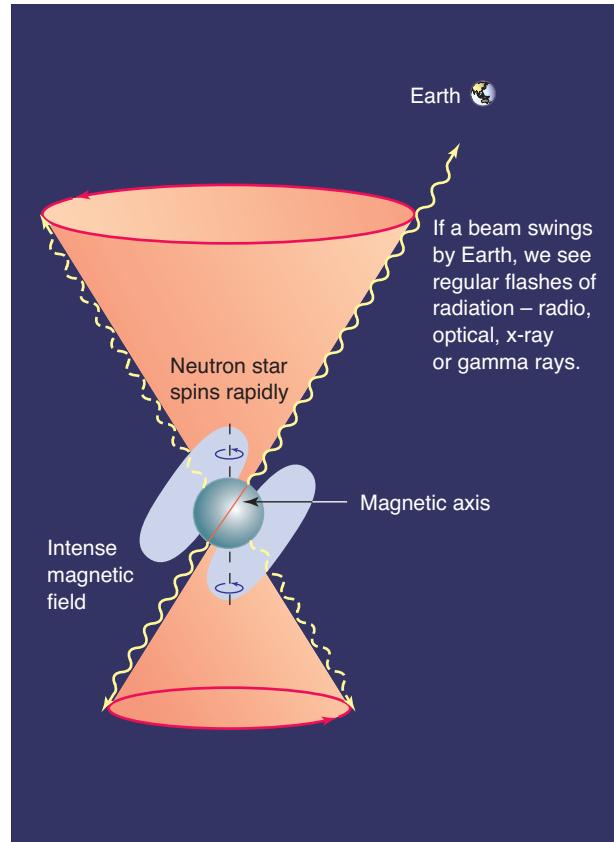


Figure 17.17 The strong magnetic fields of the quickly rotating neutron star produce beams of electromagnetic radiation from each magnetic pole that sweep through the galaxy like a lighthouse beacon.

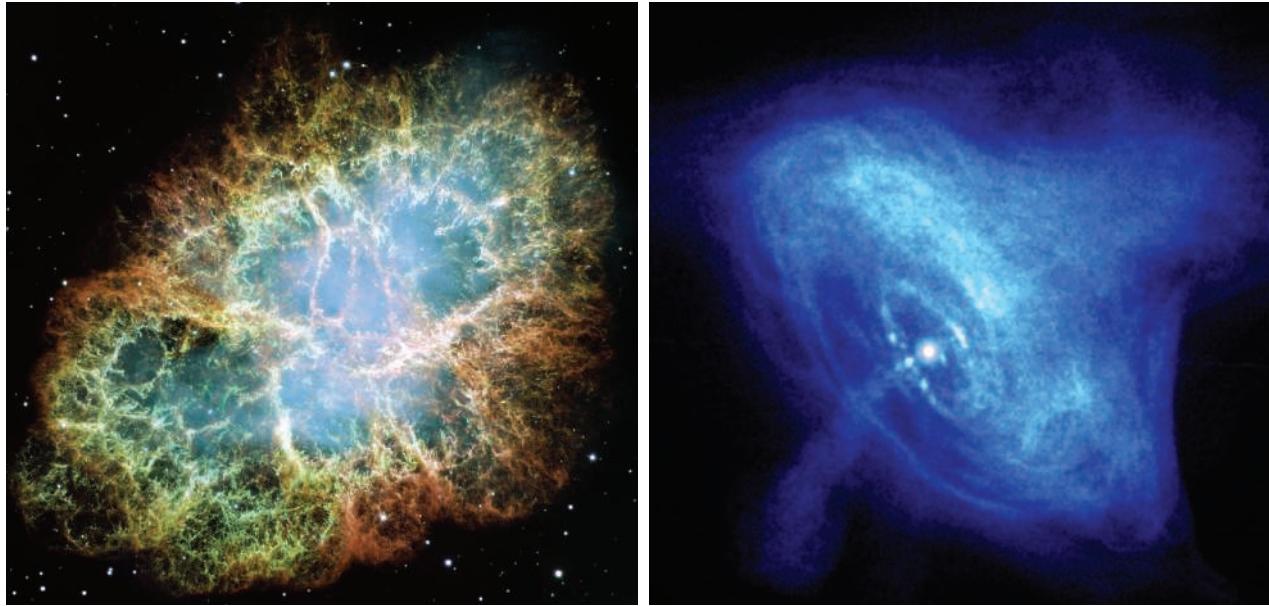


Figure 17.18 Two very different views of the Crab Nebula. This nebula is a remnant of a supernova observed in 1054. In 1968, a pulsar was observed in its centre, confirming that pulsars are neutron stars that happen to be sweeping their beam of radiation past our line of sight. The image on the left is an optical photograph taken by the Palomar Observatory; this is our usual view of the nebula. The image on the right is an X-ray photograph of the same nebula, taken by the Chandra X-ray Observatory shortly after it was placed in orbit. This previously unseen view clearly shows the powerhouse neutron star within.

Evolutionary tracks

In earlier sections we traced the paths followed by different stars on an H–R diagram, through their early and middle lives. Let us now complete the picture by including their deaths. Figure 17.19 presents this information for stars of approximately 0.1, 1, 5 and 10 solar masses.

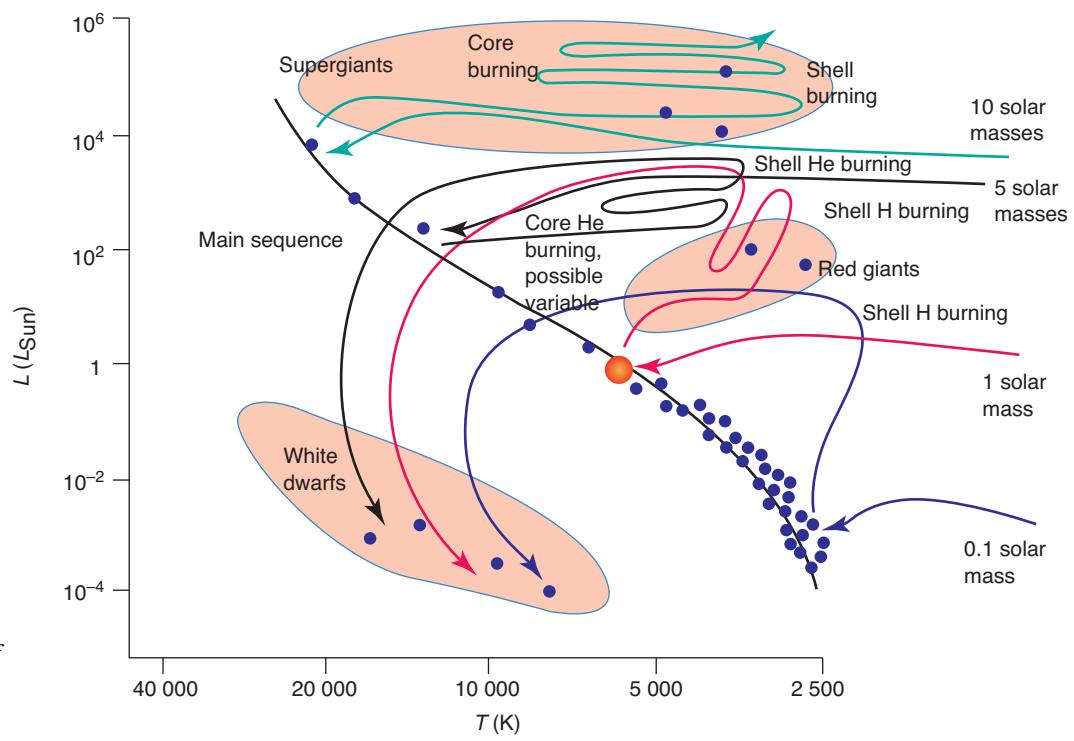


Figure 17.19
The evolutionary paths of several different stars on an H–R diagram

SUMMARY

- Interstellar space is filled with an uneven spread of gas and dust, sometimes forming clouds or nebulae. The gas is mostly hydrogen, and occurs in the form of neutral atoms, ions and molecules. The sparse dust is made up of icy grains.
- The dust is associated with molecular clouds — the grains act as the site of molecule formation.
- A molecular cloud that is sufficiently cool and massive will collapse under gravity to form a heated core called a protostar. This core is hidden from our view by the dust in the cloud.
- The protostar accretes some material from the cloud and then blows off the rest, revealing itself.
- When the star is hot enough to begin the fusion of hydrogen in its core, it becomes a main sequence star. This is the longest and most stable period of its lifetime.
- In cooler stars the dominant hydrogen-burning mechanism is the proton–proton chain; in hotter stars it is the CNO cycle.
- When core hydrogen runs out, the core contracts and heats up. The shell of previously unused hydrogen ignites and expands to turn the star into a red giant. The core then ignites to begin helium-burning. The star can become unstable and behave as a periodic (pulsating) variable.
- When the core fuel runs out, the core will contract and heat up. If the star is massive enough it will ignite a new layer of shell-burning as well as a new and heavier fuel in the core. This process can be repeated over and over but must stop when the core consists of iron.
- If the original star was five solar masses or less, it will shed its shells as a planetary nebula, while the core will be compacted into a white dwarf.
- If the original star was greater than five solar masses, it will shed its shells violently in a supernova. If the core was less than three solar masses, it will be compressed into a neutron star; if greater than this, it will be crushed to a black hole.
- A pulsar is a rotating neutron star that happens to swing its beam of electromagnetic radiation past us as it spins.

- The H–R diagrams of whole star clusters show a turn-off point that can be used to infer the age of the cluster. The lower the turn-off point, the older the cluster.

QUESTIONS

1. (a) Describe interstellar gas.
(b) Identify the types of substance that can be found within interstellar gas.
2. (a) Explain how ionised hydrogen is different from neutral hydrogen.
(b) If it is known that a region of hydrogen is ionised, what can be inferred about its temperature?
3. (a) Construct a list of molecules that can be found in the interstellar medium.
(b) What are ‘organic’ molecules, and do you think it strange that they should appear in the interstellar medium?
4. Describe the composition of an interstellar molecular cloud.
5. (a) Describe an interstellar dust grain.
(b) State where these grains are thought to originate.
6. Interstellar dust is closely associated with molecular clouds. Explain why.
7. The presence of dust within molecular clouds, the site of star formation, causes a seeing difficulty for astronomers. Describe this problem and the method used to overcome it.
8. In point form, outline the steps of formation of a star by gravitational collapse.
9. (a) Describe a protostar.
(b) Compare the luminosity, radius and core temperature of a protostar to the zero-age main sequence star it eventually forms.
10. Identify the event that marks the transition from protostar to zero-age main sequence star?
11. (a) Write a description of the proton–proton chain without using equations. Mention reactants and products.
(b) Repeat the exercise with the CNO cycle.
12. (a) Identify the hydrogen-burning mechanism that is dominant in cool stars.
(b) Identify the hydrogen-burning mechanism that dominates in hot stars.
(c) State the core temperature that marks the crossover between the reactions in (a) and (b).
(d) Can both reactions occur at the same time within a star? Explain.

13. (a) Explain the role of carbon in the CNO cycle.
(b) State why nitrogen and oxygen are mentioned in the name of this process.
14. (a) Identify the net reaction of hydrogen burning in the core of a main sequence star.
(b) Describe what happens to the helium produced by hydrogen burning.
15. In point form, summarise the steps in the transition of a star from main sequence to red giant.
16. Describe the two layers of reactions typical of a red giant.
17. (a) Write a description of the triple alpha reaction.
(b) Identify the temperature required to initiate this reaction.
18. Describe a helium flash, and the types of star experience it?
19. Identify the heaviest element able to be fused within the core of a star of:
 - (a) 0.1 solar mass
 - (b) 1 solar mass
 - (c) 10 solar masses
 - (d) 50 solar masses.
20. State when a giant is vulnerable to regular pulsations of luminosity.
21. Describe the structure of a massive star late in its giant stage.
22. Describe in general terms what becomes of:
 - (a) the various outer layers or shells of a star
 - (b) the coreduring the final stages of a star's life.
23. Discuss how the process of a star losing its outer shells depends upon its mass.
24. Describe the various final states for the core of a star and link each to the mass of the original star and the mass of the star's core.
25. (a) Construct a diagram that shows each step of a star's life, in a general form.
(b) Include on this diagram, the possible variations in the giant stage, noting mass with each variation.
(c) Include possible variations in the final stages, noting mass with each variation.
26. Construct a Hertzsprung–Russell diagram, including main sequence, red giants and white dwarfs. On this one diagram draw the evolutionary tracks of stars of 1, 5 and 10 solar masses, using a different colour for each path.
27. When pulsars were discovered, it was first thought that the pulsations were a communication from space. What feature of the signal do you think quickly dispelled this suspicion?
28. (a) Compare a pulsar and a neutron star.
(b) There must be many neutron stars in our galaxy, but only about 500 have been discovered. Explain why more pulsars have not been found. (*Hint:* It has to do with the radiation beam.)
29. Describe a cluster's turn-off point on an H–R diagram, and explain what it can tell us.
30. (a) It has been suggested that the Earth is partially supernova remnant. Discuss this contention.
(b) Bearing in mind that the solar system was formed from the same molecular cloud as the Sun, what does your answer to part (a) tell us about the Sun?



17.1 RESEARCHING STELLAR OBJECTS

Aim

To access up-to-date information on various stellar objects.

Apparatus

Internet access

Method

Use your computer to access on the internet some or all of the online astronomy databases listed below. Normally, each will have a specialty so that it is useful to sample a variety of sites. In this way, gather data, plus a picture (if possible), of two of each of the objects listed:

- main sequence stars
- variables
- binaries
- giants
- open clusters
- globular clusters
- supernovae
- white dwarfs/neutron stars/black holes
- nebulae (especially planetary)
- galaxies.

eBookplus

Weblinks:

[The Messier Database](#)

[MOST Supernova Remnant Catalogue \(MSC\)](#)

[A Catalogue of Galactic Supernova Remnants](#)

[The Double Star Library](#)

[The HIPPARCOS and Tycho Catalogues](#)

Try a general search engine or web crawler service if you still cannot find some information.



Chapter 18

The use of ultrasound in medicine

Chapter 19

Electromagnetic radiation as a diagnostic tool

Chapter 20

Radioactivity as a diagnostic tool

Chapter 21

Magnetic resonance imaging as a diagnostic tool

MEDICAL PHYSICS

CHAPTER 18

THE USE OF ULTRASOUND IN MEDICINE



Figure 18.1 Ultrasound is an important technique for medical diagnosis. This photograph shows a pregnant woman undergoing an ultrasound examination of her foetus.

Remember

Before beginning this chapter you should be able to:

- recall the features of waves, including speed, frequency and wavelength
- distinguish between transverse and longitudinal waves
- outline the properties of waves including reflection, refraction and scattering.

Key content

At the end of this chapter you should be able to:

- describe the properties and production of ultrasound
- describe the piezoelectric effect and the effect of an alternating potential difference on a piezoelectric crystal
- describe how acoustic impedance affects the behaviour of ultrasound
- calculate the acoustic impedance of a variety of materials
- calculate the amount of reflected ultrasound signal at various interfaces
- describe the situations where different types of ultrasound scans would be used
- describe the use of Doppler ultrasonics in detecting cardiac problems
- describe how ultrasound is used to measure bone density.

In this chapter we will look at the properties of ultrasound waves and how they are applied to medical imaging. Images of organs can be produced because ultrasound waves penetrate and interact with the body, as in the example of a pregnant woman undergoing an ultrasound (see figure 18.1). Movement such as blood flow in veins and arteries can also be measured using ultrasound. Ultrasound is one of the most frequently used imaging techniques in medical diagnosis.

18.1

WHAT TYPE OF SOUND IS ULTRASOUND?

Ultrasound is very high frequency sound. Ultrasound waves are sound waves that have a frequency above the range of human hearing, that is, greater than 20 000 hertz.

Ultrasound is very high frequency sound. Ultrasound waves are sound waves with frequency greater than that of normal human hearing. That is, the frequency is greater than 20 000 hertz.

You will recall that, in the Preliminary Course topic ‘The world communicates’, you learnt that sound waves are longitudinal waves and need a medium in which to travel. The particles of the medium oscillate back and forth in the same direction as the wave travels through the medium, producing a series of compressions and rarefactions. The compressions and rarefactions correspond to pressure differences in the medium. It is with reference to these compressions and rarefactions that the wavelength can be measured.

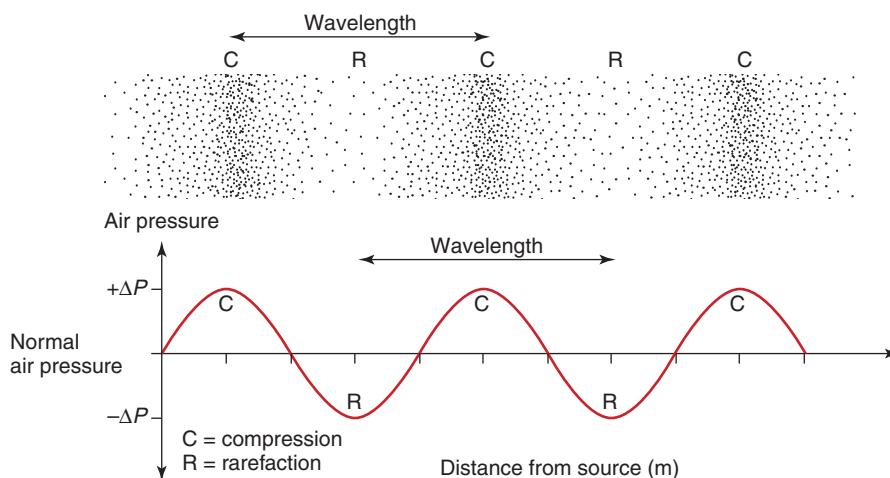


Figure 18.2 Wavelength is the distance between successive compressions or rarefactions of the longitudinal wave. The pressure in the medium is maximum at a compression and minimum at a rarefaction.

The amplitude of a wave is the maximum displacement of the particles on either side of the equilibrium position. The greater the amplitude of the wave, the greater the intensity of the wave and the greater the energy it is carrying.

The frequency (f) is the number of oscillations the particles make per second and is measured in hertz (Hz). The wavelength (λ) is the distance between two successive compressions or between two rarefactions. Frequency and wavelength are related by the equation

$$v = f\lambda$$

where v is the speed of the wave in the medium.

If the wavelength is measured in metres (m), the speed is in metres per second (m s^{-1}).

Ultrasound can be reflected, refracted, scattered and superimposed in the same way as audible sound or transverse waves such as light. These properties are important for the use of ultrasound in medical diagnosis.

Ultrasound and medical diagnosis

Ultrasound, in its applications in the field of medical diagnosis, is reflected off parts of the body, detected and analysed to produce an image. Structural images, not functional images are produced.

Ultrasound scanning provides a safe way of observing internal organs, as no damaging effects of ultrasound used in medical diagnosis are known. The frequency of the ultrasound determines the clarity of the image obtained.

Why isn't audible sound used for medical diagnosis?

In order to produce a clear image we must be able to distinguish between different internal parts of the body, some of which may be very close together. For example, we may want to detect the individual fingers of a foetus in the womb, or the parts of a heart valve. Audible sound will not distinguish between such close objects. A sound wave's ability to produce an image depends on the sound wave's wavelength and hence its frequency. Audible sound has a frequency below 20 000 Hz and a correspondingly poorer ability to produce images of small objects than higher frequency ultrasound. In producing an ultrasound image, echoes from two objects close together must be able to be detected. Generally, if the separation or size of the objects is less than the wavelength of the sound wave, the objects cannot be detected. (We say they cannot be resolved.)

The higher the frequency of the sound wave, the shorter the wavelength and the better the resolution that is possible. For example, in water the wavelength of sound at the limit of hearing (20 000 Hz) is 75 mm, whereas 1.2 MHz ultrasound has a wavelength of 1.2 mm, and 3.5 MHz ultrasound has a wavelength of 0.43 mm. This means 3.5 MHz ultrasound can be used to distinguish objects in water that are 0.43 mm apart. This is a much better resolution than that obtained by audible sound; even at the limit of hearing, objects in the water would have to be 75 mm apart to be distinguished.

What frequency will produce the clearest image?

We may ask why we do not use incredibly high frequencies with correspondingly small wavelengths so that we can detect very fine detail.

A compromise on the frequency used must be made because the absorption of the wave increases as the frequency increases. For example, a 50 MHz ultrasound does not penetrate nearly as well through tissue as a 10 MHz ultrasound (see table 18.1). Hence the frequency used must be suitable for the part of the body being analysed. A small organ on the surface of the body, such as the eye, can be examined with a much higher frequency than would be useful for examining deep abdominal organs. Medical diagnosis uses ultrasound in the range 1 MHz to tens of MHz.

PHYSICS FACT

The frequency range used for diagnosis varies depending on the part of the body that is being examined. Some examples are given in table 18.1.

Table 18.1 Ultrasound frequency for different parts of the body being imaged

FREQUENCY CHOSEN	PART OF BODY BEING EXAMINED	TYPICAL PENETRATION DEPTH	REASON FOR CHOOSING THIS FREQUENCY
50 MHz	Skin or areas reached through surgery, such as blood vessel walls and cartilage.	a few mm	The region must be close to the ultrasound as absorption of the ultrasound energy by the tissue is significant and hence it does not penetrate tissues effectively.
10–15 MHz	Eye	1 cm	The organ is small and on the surface of the body, so absorption of the ultrasound by tissues is not a problem.
4–10 MHz	Thyroid, carotid artery, breast	5 cm	These organs are quite close to the surface of the body, so absorption of the ultrasound is not a problem at this frequency range.
3–5 MHz	Liver, heart, other abdominal organs	10–20 cm	These organs, such as the heart, uterus and liver, are deep in the body. Ultrasound with a frequency that is too high will be absorbed before it reaches the organ.

18.2 USING ULTRASOUND TO DETECT STRUCTURE INSIDE THE BODY

A pulse of ultrasound is directed into the body and echoes from tissue boundaries are detected. Some of the energy of the pulses of ultrasound reflect off the junction between different substances in the body and some pass through. This is similar to what happens when a beam of light strikes the surface between water and glass — some light is reflected from the surface and some is transmitted. Because the surfaces inside the body are not usually flat, the reflection results in scattering of the ultrasound. The ultrasound that comes back to the detector is analysed.

The properties of body materials such as bone, skin and muscle are different and, as a result, sound propagates through them at different speeds and with different amounts of absorption.

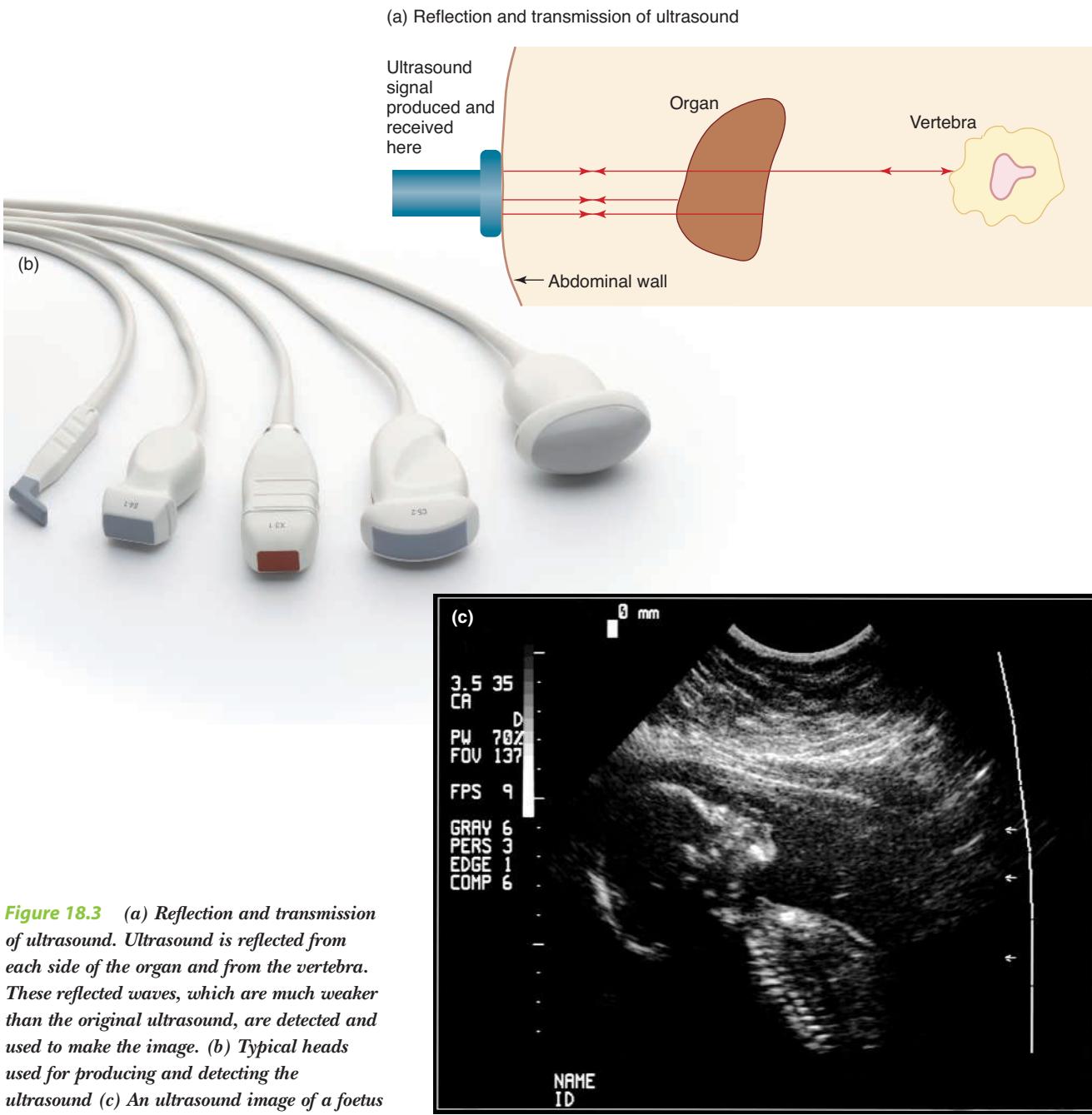


Figure 18.3 (a) Reflection and transmission of ultrasound. Ultrasound is reflected from each side of the organ and from the vertebra. These reflected waves, which are much weaker than the original ultrasound, are detected and used to make the image. (b) Typical heads used for producing and detecting the ultrasound (c) An ultrasound image of a foetus

Acoustic impedance

Acoustic impedance, Z , is a measure of how readily sound will pass through a material. It is measured in $\text{kg m}^{-2} \text{s}^{-1}$.

The extent to which body tissues transmit sound varies. The **acoustic impedance**, Z , of a material measures how readily sound will pass through a material and is defined by the formula

$$Z = \rho v$$

where

Z is acoustic impedance (in $\text{kg m}^{-2} \text{s}^{-1}$)

ρ is the density of the medium (kg m^{-3})

v is the velocity of sound in the material (m s^{-1}).

Table 18.2 shows various body materials and their properties from which their acoustic impedance can be calculated.

Table 18.2 Properties of selected body materials

MEDIUM	DENSITY (ρ) (kg m^{-3})	VELOCITY (v) (m s^{-1})
Air	1.3	330
Water	1000	1430
Eye		
aqueous humour	1000	1500
vitreous humour	1000	1520
lens	1140	1620
Soft tissue such as nerves (average)	1060	1540
Muscle (average)	1075	1590
Fat	952	1450
Liver	1050	1570
Brain	1025	1540
Blood	1060	1570
Bone	1400–1908	4080

SAMPLE PROBLEM**18.1****Ultrasound travelling through fat**

An ultrasound wave of 2.0 MHz is travelling through fat. Calculate:

- the wavelength of the ultrasound
- the acoustic impedance of the fat.

SOLUTION

- Using the wave equation,

$$v = f\lambda$$

$$1450 = 2.0 \times 10^6 \lambda$$

$$\lambda = 7.25 \times 10^{-4}$$

The wavelength of the ultrasound in fat is 7.25×10^{-4} m (0.725 mm).

- Using the acoustic impedance formula,

$$Z = \rho v$$

$$Z = 952 \times 1450$$

$$= 1.38 \times 10^6$$

The acoustic impedance of fat is $1.38 \times 10^6 \text{ kg m}^{-2} \text{ s}^{-1}$.

Reflection of ultrasound

The acoustic impedances of two adjoining tissues are used to calculate the intensity of the reflected pulse compared with the incoming one. The following formula enables us to determine the fraction of the intensity reflected at a surface between two media, such as bone and muscle:

$$\frac{I_r}{I_0} = \frac{(Z_2 - Z_1)^2}{(Z_2 + Z_1)^2}$$

where

I_r is the intensity of the pulse reflected back (W m^{-2})

I_0 is the intensity of the pulse incident on the surface (W m^{-2})

Z_1 and Z_2 are the acoustic impedances of media 1 and 2 ($\text{kg m}^{-2} \text{ s}^{-1}$).

PHYSICS FACT

Often, reflected ultrasound is very weak because, when the wave travels through tissue, much of the energy is absorbed and converted to heat. Why not send a very high intensity signal so that more energy can be reflected and detection of the reflected wave will be easier? The power levels for diagnostic medical ultrasound must be kept at very low intensity of 0.01–20 watts per square centimetre (W cm^{-2}) so that heating or destruction of tissue does not occur when energy is absorbed by tissues inside the body. Ultrasound at continuous power levels of about 1 W cm^{-2} is used to heat tissues for

therapy and, at power levels of 10^3 W cm^{-2} , may be used to destroy tissue. Ultrasound can be used to heat tissue deep in the body, and so produce relief from pain in the joints of sufferers of arthritis and stimulate blood flow to damaged tissues.

Recent advances have been reported in research into the treatment of prostate cancer. Investigations have shown that ultrasound of sufficient intensity can destroy cancer cells. Ultrasound has also been used to break up gallstones and kidney stones, avoiding the need for surgery, and to break up the lens of the eye during a cataract operation.

SAMPLE PROBLEM

18.2 Reflection of ultrasound

SOLUTION

Calculate the percentage of ultrasound that is reflected at the junction between fat and muscle, using information from table 18.2.

The acoustic impedance of fat is $1.38 \times 10^6 \text{ kg m}^{-2} \text{ s}^{-1}$ (from sample problem 18.1).

We need to calculate the acoustic impedance of muscle.

$$Z = \rho v$$

$$Z = 1075 \times 1590$$

$$\text{Acoustic impedance of muscle} = 1.70 \times 10^6 \text{ kg m}^{-2} \text{ s}^{-1}$$

Using the equation from page 346:

$$\begin{aligned} \frac{I_r}{I_0} &= \frac{(1.38 \times 10^6 - 1.70 \times 10^6)^2}{(1.38 \times 10^6 + 1.70 \times 10^6)^2} \\ &= 0.011 \end{aligned}$$

This is the ratio of reflected intensity to incident intensity. This ratio must be converted to a percentage ($0.011 \times 100\%$).

Hence 1.1% of the incident signal is reflected back.

If there is a very large difference in acoustic impedance between the two materials, a large fraction of the ultrasound will be reflected back. This is what happens when ultrasound is directed through the air to the skin. The very large difference in the acoustic impedance of air and skin means that most of the ultrasound will be reflected off the skin and will not enter the body. To minimise this reflection at the skin surface, a gel with an acoustic impedance similar to that of skin is placed on the skin. This gel excludes air from the space between the **ultrasound transducer** and the skin and provides an acoustic match. The signal can pass from the transducer through the skin without reflection.

In a pregnant woman, the bladder is between the outer surface of the body and the foetus. A pregnant woman must have a bladder full of urine in order to obtain a successful ultrasound of the foetus. The full bladder lifts the uterus to a good position for imaging and pushes the lower intestine out of the way of the signal. The lower intestine contains a lot of gas that would reflect the ultrasound and make the imaging of the foetus difficult.

An **ultrasound transducer** is a device for converting electrical energy to ultrasound energy or for converting ultrasound energy to electrical energy.

18.3 PRODUCING AND DETECTING ULTRASOUND: THE PIEZOELECTRIC EFFECT

The **piezoelectric effect** is the conversion of electrical energy to mechanical energy resulting in the change in shape of a piezoelectric crystal when it is subjected to a potential difference.

An ultrasound transducer produces ultrasound of a specific frequency and this same transducer is capable of detecting the reflected ultrasound. In order to produce ultrasound, a material must be made to vibrate at a very fast rate, of the order of 1.5×10^6 Hz. The material used is a piezoelectric crystal. If an electric field (potential difference) is applied across the crystal, the shape of the crystal changes. This behaviour of the crystal is called the **piezoelectric effect**. By reversing the potential difference repeatedly and rapidly, the crystal can be made to vibrate and produce frequencies in the ultrasound range. Such a crystal is part of the ultrasound transducer (see figure 18.4).

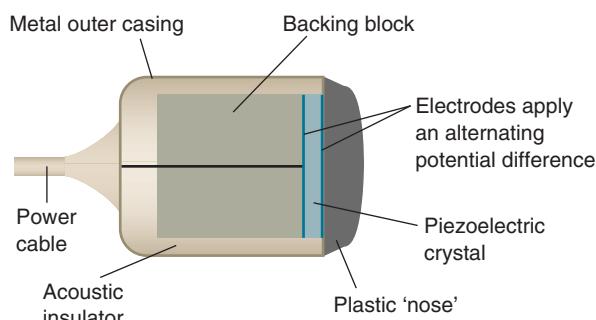


Figure 18.4 An ultrasound transducer

A **crystalline** substance is a solid in which the atoms or molecules are arranged in a regular pattern.

Naturally occurring **crystalline** substances exhibiting the piezoelectric effect include quartz, lithium sulfate and barium titanate. The most commonly used substance for a transducer for medical purposes is the artificial ceramic material, lead zirconate titanate (PZT).

Conducting material on either side of the crystal of the transducer permits the application of an alternating potential difference across the crystal. The backing block dampens the vibration. If there was no backing block, the crystal would continue to vibrate after the electric field was taken away, in the same way that a drum continues to vibrate after it has been struck. The backing block dampens the vibration of the crystal in the same way as placing a hand on a drum skin would stop the drum's vibrations.

When ultrasound strikes this crystal there is a variation in pressure felt at the crystal surface. (Recall from your Preliminary Course work in *Physics 1*, chapter 2 that compressions and rarefactions are detected as pressure differences in the medium through which the wave is travelling.) When a compression meets the surface, the crystal is compressed; when a rarefaction meets the surface, the crystal relaxes. The crystal will be changed in the same way as when the crystal was generating ultrasound. In this case the changing crystal will cause a changing electric field to be produced across the crystal. This changing electric field will vary in time with the frequency of the ultrasound and can be detected in the transducer. This means that the transducer can be used to detect ultrasound as well as to produce it. The voltage produced is greater when the amplitude of the ultrasound is greater. Hence the transducer provides information about the ultrasound wave intensity.

PHYSICS FACT

History of the use of ultrasound

The first device for the production of ultrasound was a pipe, produced in 1820 and called the Wollaston whistle. It was developed by the English scientist W. H. Wollaston, who was determining the limits of human hearing.

Modern medical transducers are based on the piezoelectric effect, which was discovered by Jacques and Pierre Curie in 1880. These improved as developments in electronics grew from the introduction of radar, just before and during World War II. At this stage, imaging based on the pulse-echo principle was possible.

A patent was filed in 1940 for an ultrasound device to detect flaws in metal structures. In 1948, Karl Dussik, who was trying to detect brain tumours, unsuccessfully attempted the first medical application of ultrasound. It was not until 1952 that the first successful medical use of pulse-echo imaging was described by J. J. Wild and J. M. Reid. Six years later, in 1958, the first commercial equipment for ultrasonic imaging appeared.

18.4 GATHERING AND USING INFORMATION IN AN ULTRASOUND SCAN

An ultrasound transducer detects reflected signals from different body structures. A computer then analyses the signal to obtain information about the location of the structures to produce an image. Different types of scans are chosen to suit particular purposes.

A-scans

An **A-scan** is a range-measuring system that records the time for an ultrasonic pulse to travel to an interface in the body and be reflected back.

An **A-scan** is a range-measuring system that records the time for an ultrasonic pulse to travel to an interface in the body and be reflected back. In an A-scan the ultrasound pulses are directed into the body in one line and the reflected signal is detected. The intensity of the reflected beams is plotted on a graph as a function of time. In this way the position of various features can be determined from the time lapse between sending the signal and receiving its echo and a knowledge of the speed of sound in the tissue. The intensity of the reflected beam provides information about the type of material through which the ultrasound is travelling.

An A-scan provides one-dimensional information about the location of the reflecting boundaries. Originally this type of scan was used to determine the midline position of the brain and detect any abnormalities there caused by tumours, because the midline would be displaced by a tumour. A-scans are no longer used for this as more sophisticated methods of imaging the brain have been developed. A-scans are still used in ophthalmology for the diagnosis of eye disease and for measurements of distances in the eye, where no image of the interior of the eye is needed (see figure 18.5).

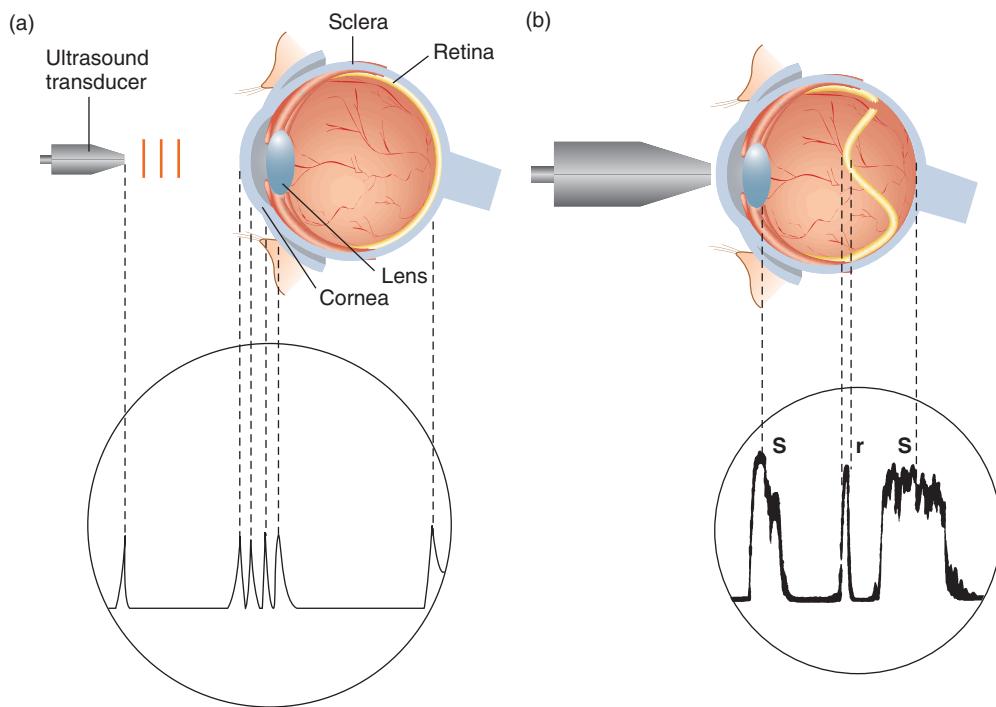


Figure 18.5 (a) The reflected ultrasound from parts of the eye displayed on an oscilloscope
(b) Ultrasound studies of a detached retina. The A-scan trace shows an echo (s) from the front of the eye, an echo (r) from the retina and an echo (s) from the back of the eye. In a normal eye the echo from the retina would blend with the echo from the back of the eye.

A **B-scan** displays the reflected ultrasound as a spot, the brightness of which is determined by the intensity of the ultrasound.

B-scans

In a **B-scan** the intensities of the reflected ultrasound are represented as spots of varying brightness, the brightest spot corresponding to the most intense reflected ultrasound. By moving the transducer probe, the body is viewed from a range of angles. A series of spots are obtained, each series corresponding to a different line through the body. These spots can give a 2-D picture of a cross-section through the body (see figure 18.6).

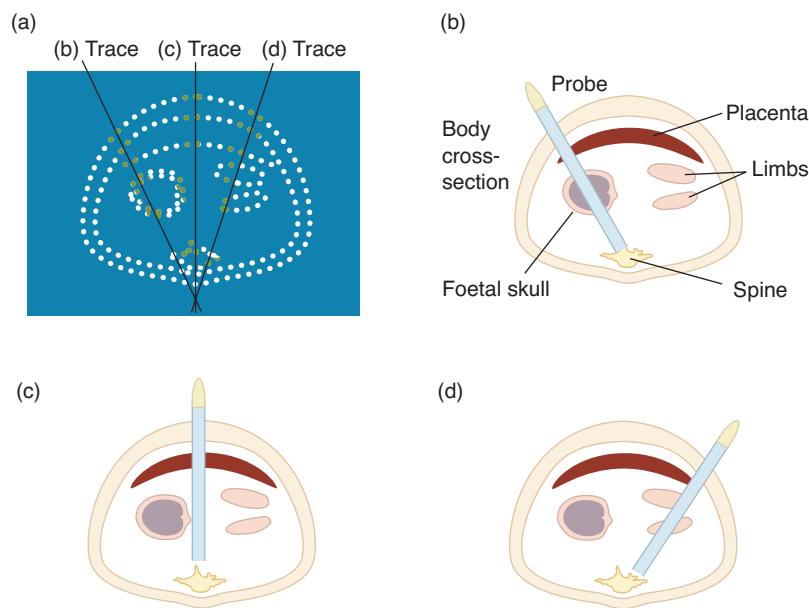


Figure 18.6 Building up a B-scan image of a foetus

Sector scans and phase scans

Sector scans are scans in the shape of a sector. They are made up from a series of B-scans.

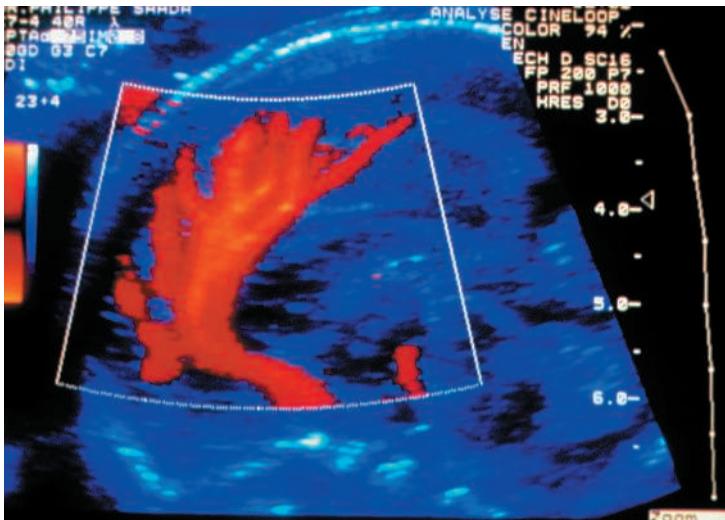


Figure 18.7 Sector scan of an infant's brain

Sector scans are scans of a fan-shaped section of the body. They are made of a number of B-scans, which build up an image of the sector in the body through a series of dots of varying intensities.

When this type of scanning was first used, a single transducer was rocked back and forth manually so that the ultrasound pulses would sweep across a sector of the body. This required skill and experience to achieve clear images and is now rarely used in hospital work. However, its advantage is that it requires a small entry 'window' into the body and is still valuable in imaging through a small space, such as the space between the bones of an infant's skull to obtain an image of the infant's brain, as shown in figure 18.7.

Modern scanning techniques use an array of transducers very close together in the one probe head. This enables very clear images to be produced and also allows the possibility of real-time scans (scans that are produced faster than 16 images per second and displayed on the monitor at that rate). Real-time scans allow movement to be monitored and so are used to examine, for example, foetal movement or heart movement.

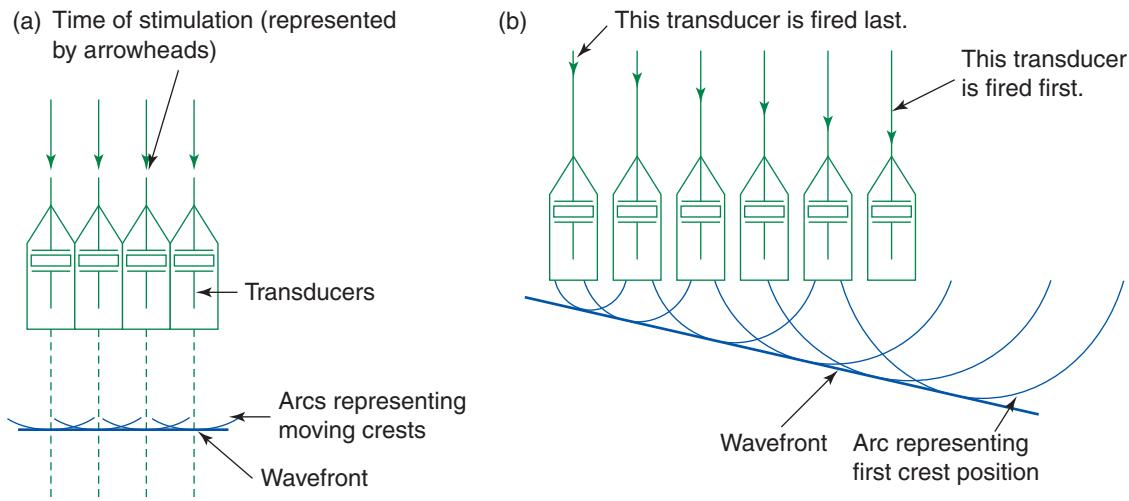


Figure 18.8 (a) When the transducers send their signals together the signals are in phase. The wavefront of the signal is parallel to the transducers. (b) When the transducers fire at different times the signal fired first will travel furthest. The waves at the probe surface will be out of phase with one another. The wavefront, which is a tangent to the crests of the advancing waves, will have changed direction compared with (a).

There may be as many as several hundred transducers in an array. These transducers may be fired simultaneously in which case they produce a wavefront parallel to the transducers. If the transducers are fired in close succession so that they are slightly out of phase with one another, they produce a wavefront that strikes the surface at an angle other than 90° . By changing the time between firing of the transducers, the phase difference of the waves from the transducers and hence the direction of the ultrasound beam can be altered. The beam can be swept

A **phase scan** is a scan produced using an array of transducers. The phase difference between the signals from each transducer may be varied to produce this scan.

from side to side, producing a scan over a wide arc. Improvements in transducer array construction and electronic processing have led to improvements in image quality. The firing is electronic and very accurate. By using this phased array of transducers, a **phase scan** is produced.

Ultrasound scanning using phase scans is the most common scanning technique used today.

PHYSICS FACT

Phase difference

The phase difference represents the relative positions of two waves compared with one another. If two identical waves are generated side by side at the same instant and travel through a medium, the crests of the waves are always together and there is no phase difference between the waves. If the crest of one wave is always in line with the trough of the other wave then the phase difference is half a wavelength. In other words, if the two identical waves are generated at slightly different times from one another, the first crest of one wave will always be ahead of the first crest of the second wave and the waves will be out of phase. (There is a phase difference.) The wavefront, or line representing the approaching waves, is the line that is a tangent to the crests of the waves.

If arcs represent the crests of the waves travelling through a medium, figure 18.9 shows waves where there is (a) no phase difference and (b) two distinct phase differences.

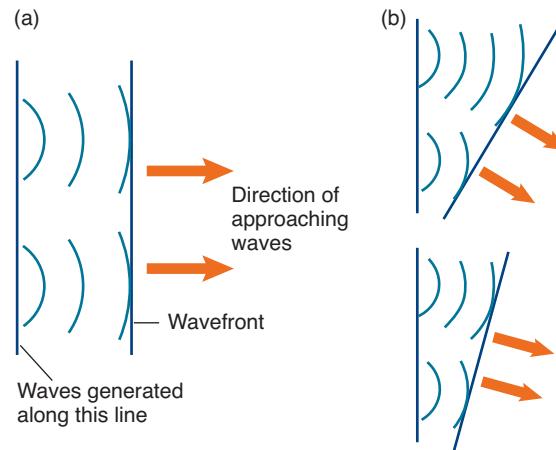


Figure 18.9 The size of the phase difference determines the direction in which waves propagate. (a) no phase difference (b) phase difference

PHYSICS IN FOCUS

Ultrasound and bone density

There are several techniques for measuring bone density. These vary in their usefulness for detecting the risk of osteoporosis, a disorder in which reduced bone density leads to brittle bones.

Normal X-rays are not suitable for the early detection of osteoporosis. This is because X-rays show changes due to loss of bone density only when approximately 30 per cent of bone has been lost. Diagnosis needs to be made earlier than this.

Ultrasound measurement of bone density is more effective and the technique is readily available, often through mobile units at pharmacies. The patient inserts a foot into a warm water bath and ultrasound waves are directed through the heel, as shown in figure 18.10.

The speed of the ultrasound through the bone and the ultrasound attenuation (degree of absorption) are measured. Normal bone has a

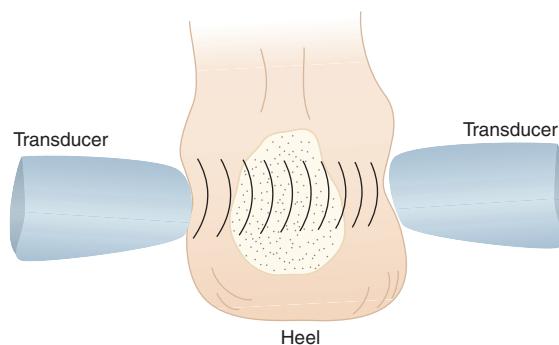


Figure 18.10 Transducers send and receive the ultrasound signal through the heel bone to provide data for the analysis of bone density.

higher speed of ultrasound and larger attenuation than osteoporotic bone. Speed and attenuation are combined to give an index from which an estimate of heel bone mineral density is reported.

(continued)

It is not possible by this method to measure sites of fractures that occur in people with osteoporosis, such as at the hip or spine. Although it is a cost-effective method of screening for bone density, ultrasound on its own does not predict the probability of fractures and so is not recommended for the diagnosing of osteoporosis. If an abnormal result is obtained from the bone density analysis, a DEXA examination should be undertaken to test for osteoporosis.

DEXA (Dual Energy X-ray Absorptiometry) using X-rays is regarded as the most reliable method of measuring bone density and can detect small changes only 6–12 months after a previous measurement. The density of the lumbar spine and left hip are usually measured. The DEXA procedure is different from a normal X-ray because low-energy X-rays are used. (The X-ray dose is so low that the radiographer can remain in the room with the patient.) People who are shown to have low bone density through ultrasound measurement are referred for a DEXA scan, because DEXA measures bone density with high accuracy and precision. More commonly, a person will be sent directly for a DEXA scan and not for ultrasound testing for bone density.

Further information about ultrasound measurement of bone density may be found from the Royal Adelaide Hospital website or using key words of ‘ultrasound bone mineral density’ in an internet search engine.

eBook plus

Weblink:

Ultrasound measurement of bone density

An interesting discussion of ultrasound and X-rays from the Health Report on ABC Radio National. This information is still reliable, although it was recorded in 1999.



Figure 18.11
A heel ultrasound scanner

18.5 USING ULTRASOUND TO EXAMINE BLOOD FLOW

The **Doppler effect** is the apparent change in frequency observed when there is relative movement between a source of a sound and an observer.

Ultrasound may be used to measure blood flow and hence detect problems such as threatened blockages in arteries. An understanding of the **Doppler effect** is needed to understand how ultrasound is used to measure blood flow.

The Doppler effect

You may have noticed the change in pitch when a car sounding its horn or an ambulance sounding its siren drives past. The sound becomes lower as the source of sound passes and moves away. The reason for this is shown in figure 18.12.

In figure 18.12(a) we see the sound waves generated by the horn. Sarah and Sam hear the exact frequency that the horn makes. In figure 18.12 (b) we see that the car is running into the sound waves it makes in front of it and moving away from the sound waves that travel out behind it. This means that the wavelength of the sound wave is shorter in front of the car and the frequency that is heard by Sarah is higher than normal. Behind the car, the wavelength of the sound wave is longer than normal and hence the frequency of the sound heard by Sam is lower.

The same effect occurs if we drive towards a stationary source of sound. We pass more waves per second when we travel towards the source compared with when we were stationary, and hence we hear a note of higher pitch than normal. Similarly, as we travel away from the source we meet fewer waves per second and hear a sound of lower pitch than normal.

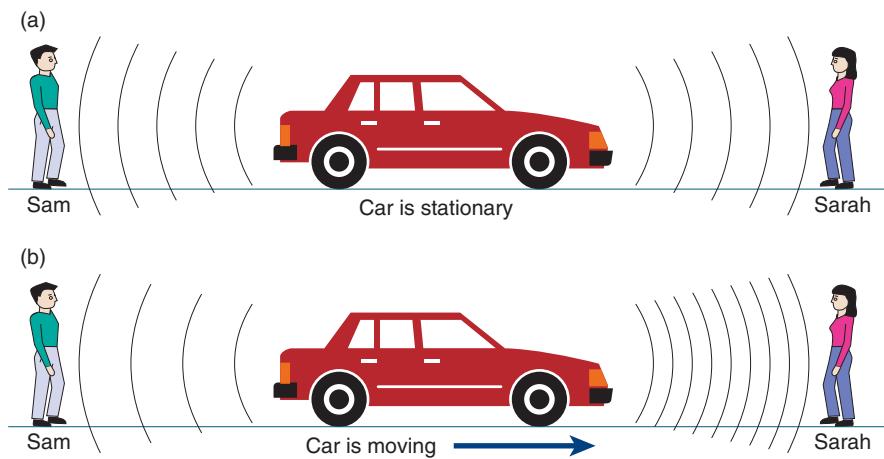


Figure 18.12 Change in frequency heard when source moves

This apparent change in frequency of a sound when there is relative movement between the source of a sound and the observer is called the Doppler effect.

Doppler ultrasound in practice

In Doppler ultrasound, the change in frequency is measured and analysed to give information about rate of blood flow in the body, and particularly through the heart.

An ultrasound is directed into the body and some of this ultrasound is reflected off blood cells moving with the blood. Due to the movement of the blood cells, the reflected ultrasound that is received by the transducer will have changed in frequency compared with the incoming signal.

In fact, the Doppler effect has to be taken into account twice. To illustrate this, imagine the blood is moving towards the transducer. The blood cells will receive a signal at a higher frequency than that given out by the transducer. These blood cells then act as a source when they reflect the signal. They reflect the higher frequency wave and then move into the wave at the same time, resulting in a further increase in frequency (see figure 18.13). This higher frequency is received by the transducer.

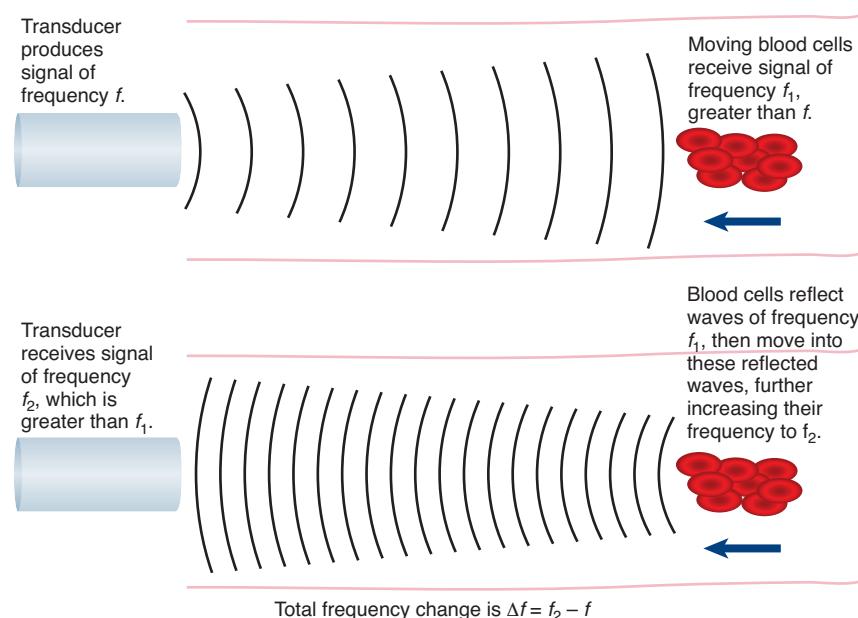


Figure 18.13 Double Doppler effect

For example, if the blood flow is 1 m s^{-1} and a 5 MHz ultrasound is used, the frequency *change* is approximately 3 kHz, which is in the *audible range*. Note that the frequency change is what is measured. An experienced practitioner can listen to the frequency change and make judgements about whether the flow is towards or away from the transducer and whether the blood flow rate is normal. Of course the signal can also be electronically analysed and displayed on a screen for examination.

The ultrasound reflected from internal tissues may pose a problem as these waves interfere with the echo ultrasound being analysed. Particular types of signals must be used to overcome this problem. These are discussed below.

Choosing the best ultrasound signal

If a continuous signal is used there must be separate transmitter and receiver transducers within the one head so that the wave being sent does not interfere with the one being received. The change in frequency can be detected through headphones or measured electronically. Using a reference frequency, the new frequency is measured above or below this reference frequency. The electronic device measures beats, equal to the difference in frequency between the transmitted and detected waves. The beats can be read from a display of amplitude of the signal against time. By this means, flow towards or away from the transducer can be determined and information about the blood flow at an instant can be determined. Note that the amount of Doppler shift also depends on the angle at which the beam strikes the moving blood.

The drawback of the continuous Doppler signal is that it does not convey clear information about deep blood vessels due to scattering and reflection from soft tissues encountered by the ultrasound as it penetrates the body. Although the continuous Doppler signal will work for blood vessels close to the skin, it is not suitable for studying the heart, which is deep in the body.

Most Doppler systems now use pulsed signals. These allow examination of blood vessels deep under the skin. The position of the blood vessel can be determined with a traditional B-scan. Then a pulsed signal is directed into the blood vessel. The time between pulses is determined by how deep the vessel is, as the return pulse must be received before the next pulse is sent. If a deeper blood vessel needs to be examined a longer period of time is waited before the next pulse is sent. The timing of the pulses is determined electronically and they are only microseconds apart.

Many modern Doppler instruments allow an image of the anatomy of the body and blood flow to be recorded in real time. This is called real-time, two-dimensional colour flow imaging. A pulse-echo system is used to obtain a phase scan of, for example, the heart. Pulsed Doppler techniques are used at the same time to obtain blood flow information. This information is colour-coded to indicate whether the blood is flowing towards or away from the transducer. Red, orange and yellow are used for blood flowing towards the transducer, the colour depending on the speed of the blood. Dark blue to light blue is used for blood flowing away from the transducer. The colour varies across the artery, showing that the blood flows more quickly in the centre of the artery.

To view Doppler ultrasound video images of blood flow through the heart, search the internet using key words such as 'Doppler colour flow imaging' 'Doppler ultrasound images' or 'Colour Doppler echocardiography'.

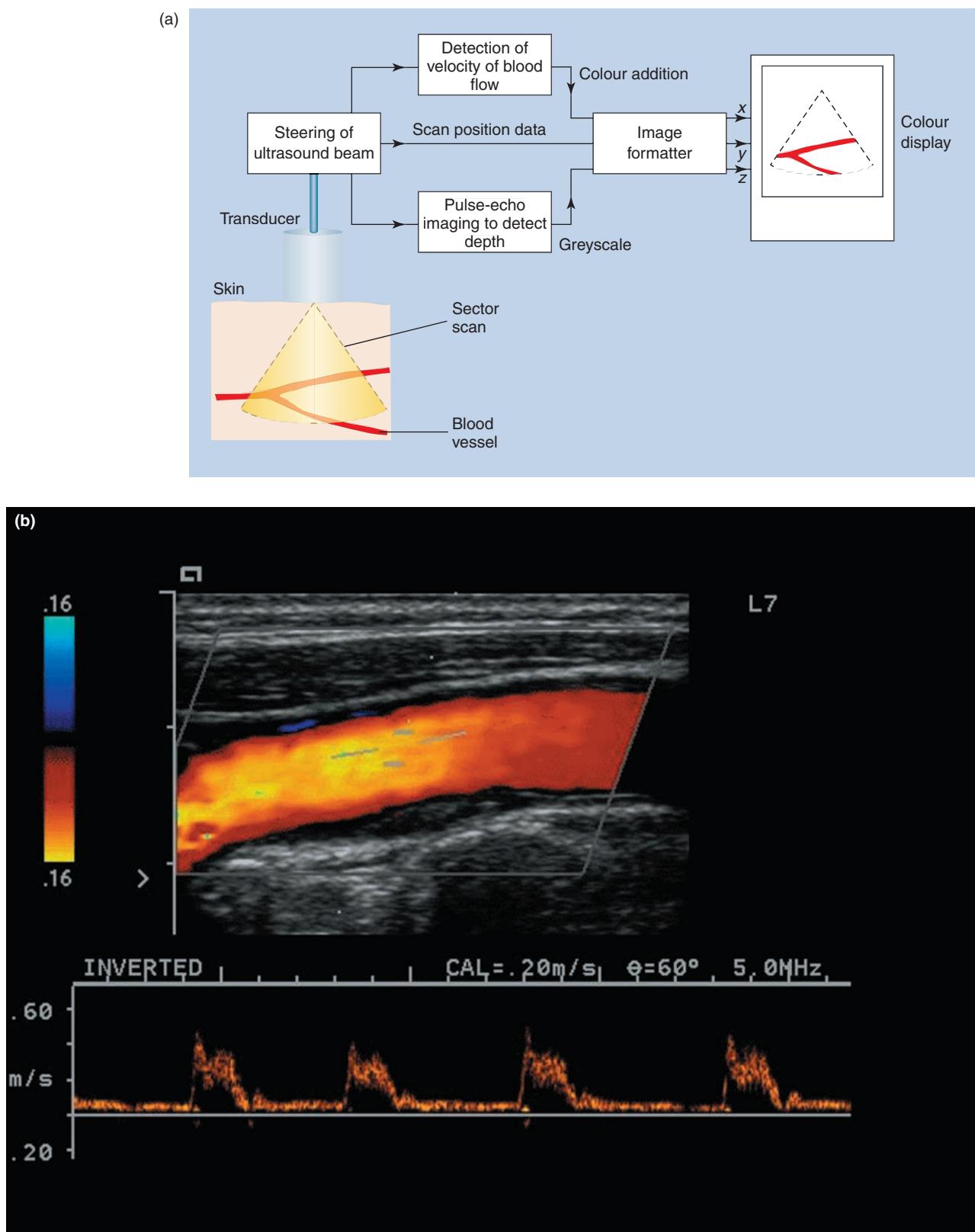


Figure 18.14 (a) Block diagram of a real-time, two-dimensional, colour flow imaging system. A phased array transducer produces a sector scan. Distance information is extracted by the pulse echo system, and velocities are determined from the size of the Doppler shift. From these two signals an overall display is made, in which the blood flow is colour coded. (b) Display of an ultrasound scan showing blood flow through a section of the carotid artery

Table 18.3 Medical uses for ultrasound scans

MEDICAL AREA AND ORGANS EXAMINED	HOW ULTRASOUND IS USED
Cardiology: the heart	A real-time ultrasound phase scan of the beating heart allows diagnosis of abnormal heart wall motion or disease or fluid accumulation in the region around the heart (an echocardiogram). When the scan is combined with Doppler colour flow imaging, the valves can be checked to see if they open and shut correctly and if they leak. The blood flow in the vessels and heart chambers can also be checked.
Endocrinology: the thyroid gland	Ultrasound phase scans are used to detect cysts, tumours or goitres.
Gynaecology: female reproductive organs	Ultrasound phase scans are used to detect blocked oviducts or ectopic pregnancy (foetus attached in the fallopian tubes, not the uterus).
Obstetrics: the foetus and uterus	Ultrasound scans are used to determine the position of the foetus, placenta and umbilical cord before birth (see figure 18.3(c), page 344). Multiple births can be detected. Ultrasound can be used during amniocentesis to guide the safe sampling of amniotic fluid.
Paediatrics: the infant	A sector scan can be used to image an infant's brain through the unclosed space in the skull.
Renal: the kidneys	Pulsed ultrasound phase scans are used to detect kidney tumours, kidney stones or blockages of the renal tubes.
Vascular: the arteries	Doppler ultrasound can be used to detect blood flow, particularly through the carotid artery, to assess the chance of a stroke. Blood flow through the umbilical cord can also be studied.

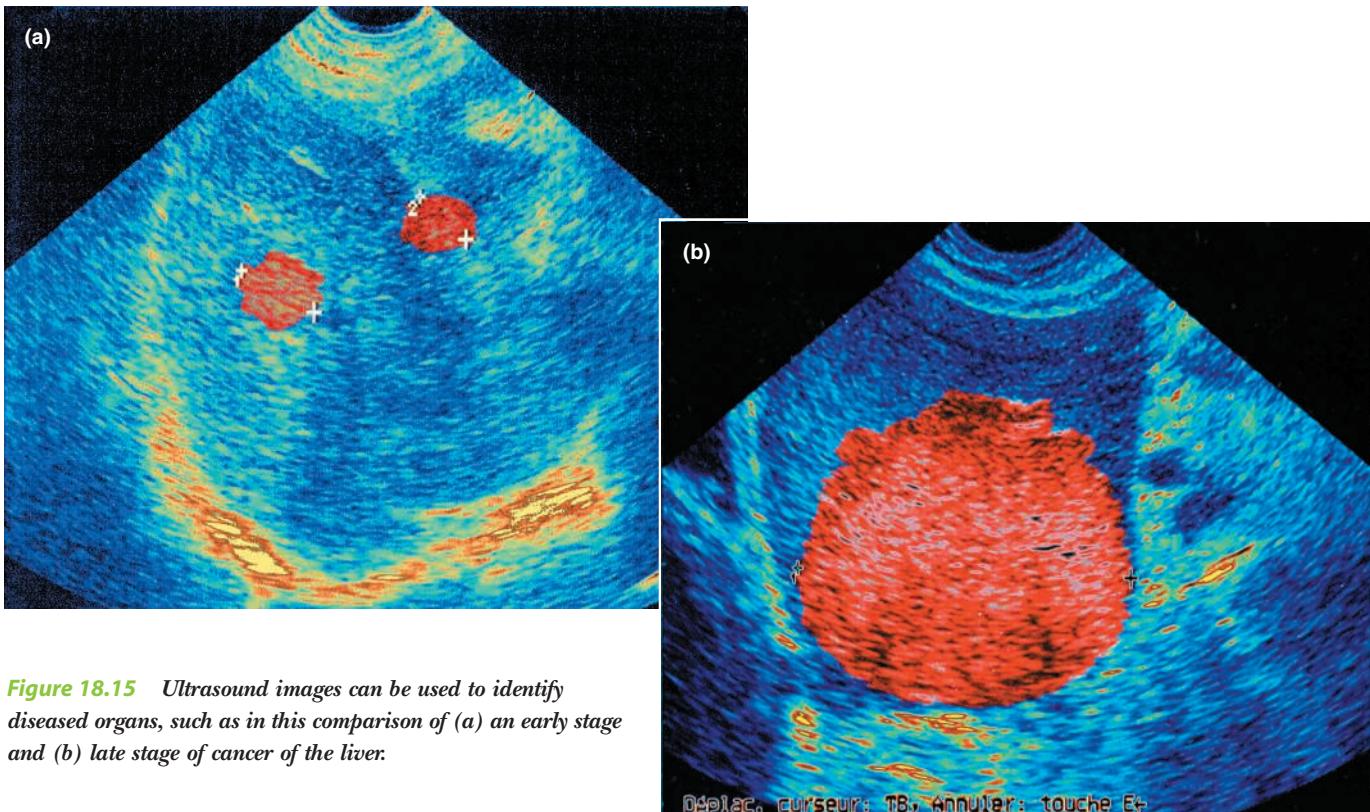


Figure 18.15 Ultrasound images can be used to identify diseased organs, such as in this comparison of (a) an early stage and (b) late stage of cancer of the liver.



Figure 18.16 Ultrasound is safe for use with babies and foetuses.

Table 18.4 Advantages and disadvantages of ultrasound for medical diagnosis

ADVANTAGES	DISADVANTAGES
<ul style="list-style-type: none"> • No damaging side-effects are known. • It is non-invasive. As no surgery is involved, there is no risk of infection. • It is more effective than conventional X-rays in producing images of soft tissues. • The equipment is relatively inexpensive. • The equipment is safe, portable and can be operated from a wall socket. • Real-time imaging is possible. • 'Keyhole' surgery can be carried out at sites close to the body surface by the use of real-time ultrasound imaging while operating. For example, a damaged part of an artery can be located using ultrasound and a balloon catheter can be inserted to help repair it through a small incision using local anaesthetic. 	<ul style="list-style-type: none"> • As the interface between gas and soft tissue reflects 99.9 per cent of the ultrasound energy that hits it, images cannot be obtained of structures that lie on, for example, the far side of the lungs or intestines. • It is difficult to obtain good ultrasound images of the brain of an adult, as most of the ultrasound is reflected from the tissue/bone interface. • Images are not as clear as those obtained by many other techniques.

SUMMARY

- Ultrasound is high frequency sound above the range of normal hearing.
- As the frequency of the wave increases, the absorption of the wave increases.
- As the energy of the wave increases, the heating effect on the tissues increases.
- A piezoelectric crystal can be made to vibrate and produce ultrasound by applying an alternating voltage. This is the principle behind generation of ultrasound in a transducer.
- Ultrasound can make a piezoelectric crystal vibrate and produce an alternating voltage. This is the principle that allows a transducer to detect ultrasound.
- Acoustic impedance (Z) can be calculated using $Z = \rho v$ where ρ is the density of the material and v is the velocity of the sound through the material. Acoustic impedance measures how readily sound passes through a material.
- At interfaces where the acoustic impedance changes, ultrasound is reflected and refracted.
- The ratio of the intensity of reflected signal (I_r) to initial signal (I_o) is given by

$$\frac{I_r}{I_o} = \frac{(Z_2 - Z_1)^2}{(Z_2 + Z_1)^2}.$$

- A-scans measure in one dimension and record depth of structure in the body. They are used mainly for scanning the eye.
- B-scans record intensity of reflection in two dimensions and build up an image of internal structure through a series of dots of varying intensities.
- Sector scans are scans in the shape of a sector, made up from a series of B-scans.
- Phase scans are scans produced using an array of transducers. There can be a phase difference between the signals from the transducers.
- Bone density can be investigated by measuring the speed of ultrasound and attenuation of ultrasound in the heel bone.
- The Doppler effect is the apparent change in frequency observed when there is relative movement between the source and an observer.

- Pulsed Doppler techniques are used to measure the speed of blood through blood vessels and through the heart.
- The Doppler effect can be used to detect heart problems such as abnormal heart wall motion, faulty valves and accumulation of fluid around the heart.

QUESTIONS

1. Contrast ultrasound with the sound used for normal hearing in humans.
2. Describe how ultrasound is transmitted through body material.
3. Describe what is meant by the piezoelectric effect.
4. Outline why a piezoelectric crystal can be made to produce and receive ultrasound waves.
5. Using the data from table 18.2 (page 345), calculate the acoustic impedance of muscle.
6. If the acoustic impedance of blood is $1.59 \times 10^6 \text{ kg m}^{-2} \text{ s}^{-1}$ and the velocity of sound through the blood is 1570 m s^{-1} , calculate the density of blood.
7. The density of air is 1.3 kg m^{-3} and its acoustic impedance is $429 \text{ kg m}^{-2} \text{ s}^{-1}$. Calculate the velocity of ultrasound through air.
8. (a) Using the data from table 18.2, calculate the acoustic impedance of:
 - (i) soft tissue
 - (ii) bone of density 1600 kg m^{-3} .
 (b) Calculate the ratio of the amplitudes of reflected and transmitted ultrasound travelling from soft tissue to the bone in part (a).
9. Using the data from table 18.2, calculate the fraction of incident ultrasound intensity reflected from a liver-muscle interface.
10. The value of the ratio of reflected intensity to incident intensity (I_r/I_o) for sound at various interfaces is found in the table on the opposite page. Use it to answer the following questions.
 - (a) Identify the tissue interface at which there is the most reflection.
 - (b) Identify the tissue interface at which there is the least reflection.
 - (c) Identify the tissue interface at which the greatest amount of absorption occurs.

- (d) If an ultrasound signal of intensity 60 mW cm^{-2} meets a fat-bone interface, calculate the intensity of the reflected signal.
- (e) (i) If the ultrasound signal striking a fat-muscle interface is 80 mW cm^{-2} , calculate how much energy travels into the muscle.
(ii) Describe what happens to this energy.

TISSUE OR MATERIAL INTERFACE	$\frac{I_r}{I_0}$
Air-water	0.999
Water-blood	0.0057
Brain-fat	0.0044
Fat-muscle	0.011
Fat-bone	0.029
Skin-air	0.999
Water-brain	0.024

11. If ultrasound travels towards the lungs, which are full of air, what will happen at the interface between the lung and the surrounding tissue? Justify your answer.
12. Using the information from table 18.2 (page 345), compare the percentage of ultrasound reflected at the junction between fat and liver with the percentage of ultrasound reflected at the junction between liver and fat.
13. A pregnant woman needs to have a bladder full of urine if she wishes to have a successful ultrasound of her baby. Explain why an empty bladder would make an ultrasound unsuccessful.
14. A low-intensity ultrasonic beam of 15 mW cm^{-2} is used to study the lens of the eye. Use the data in table 18.2 to calculate the intensity of the reflected beam if we assume the fluid in front of the lens is aqueous humour.
15. For the following question, assume the density of skin is 1010 kg m^{-3} and the velocity of sound through skin is 1540 m s^{-1} . A 1 MHz transducer requires the use of a gel on the skin to avoid acoustic mismatch at the skin-transducer interface.
- (a) Describe what would happen if air was between the transducer and the skin.

- (b) Calculate the optimum acoustic impedance of the gel and justify your answer.
- (c) Calculate the speed of the ultrasound in the gel if the gel is made of material of density 1200 kg m^{-3} .
16. Outline the difference between an ultrasonic A-scan and an ultrasonic B-scan.
17. Using a specific situation where A-scans are used in the body, outline why they are sometimes referred to as 'range finders'.
18. Figure 18.17 shows an oscilloscope display of pulse amplitude against time for an ultrasound A-scan through a person's abdomen.

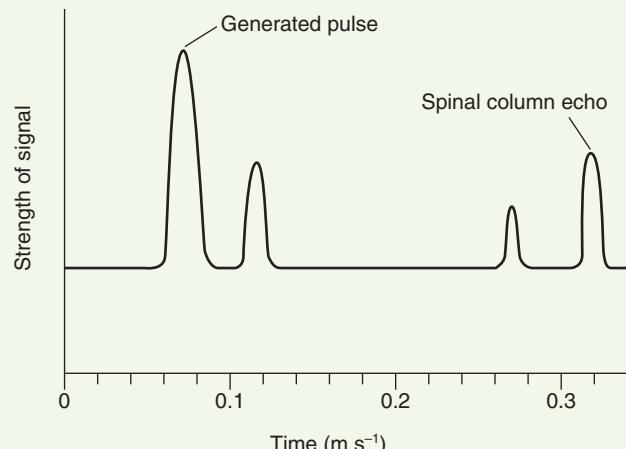


Figure 18.17 An oscilloscope display of pulse amplitude against time for an ultrasound A-scan

- (a) Explain how the spacings of the pulses are interpreted.
- (b) Give two reasons why the amplitude of the reflected pulses varies.
- (c) If the speed of ultrasound through water and soft tissue is approximately 1500 m s^{-1} , estimate the distance between the front of the patient's abdomen and the spinal column.
- (d) Name the type of scan now more commonly used for this part of the body.
19. A body structure at a depth of 350 mm is to be imaged using a B-scan. In order to obtain a clear image the reflected signal must be received before the next pulse is sent.
- (a) Assuming the sound speed is 1540 m s^{-1} in the body, calculate the minimum time between pulses that may be used to provide an unambiguous image.
- (b) Explain why a faster rate of pulse would produce an image that was not clear.
20. Use a table to discuss the value of A-scans, B-scans, sector scans and phase scans as ultrasound imaging techniques.

21. (a) Use a flow diagram to describe how ultrasound is used to obtain bone density information.
- (b) Outline its effectiveness in the diagnosis of osteoporosis. (You may wish to use the web sites and key words suggested on page 352).
22. Outline how ultrasound Doppler effect is used to monitor foetal heart rate.
23. Outline what is meant by 'real-time imaging' and assess its value in ultrasound diagnosis.
24. Carry out internet research, using the key words suggested on page 354, to observe and describe blood flow from the heart by observing ultrasound video images.
25. Collect at least two ultrasound images of body organs. One source of these images is the internet (search using key word 'ultrasound images'). Compare the images.

CHAPTER 19

ELECTROMAGNETIC RADIATION AS A DIAGNOSTIC TOOL



Figure 19.1 X-ray of the lungs showing damage to the right-hand side due to tuberculosis

Remember

Before beginning this chapter you should be able to:

- recall the features of X-rays, including speed, frequency and wavelength range
- outline what is meant by the photoelectric effect
- describe what is meant by critical angle and outline the conditions needed for total internal reflection.

Key content

At the end of this chapter you should be able to:

- describe how X-rays are produced
- compare the differences between 'soft' and 'hard' X-rays
- explain how a computed axial tomography (CT) scan is produced
- identify X-ray images of fractures and other body parts
- compare the information from a CT scan image with that provided by an X-ray image of the same body part
- describe when a CT scan would be a superior diagnostic tool to either X-rays or ultrasound
- explain how an endoscope works with particular reference to the transfer of light through the optic fibres
- observe images produced by an endoscope
- discuss the role of coherent and non-coherent bundles of optic fibres in an endoscope
- explain how an endoscope is used to observe internal organs and to assist in obtaining tissue samples.

In this chapter we will discuss the properties of electromagnetic radiation as they apply to medical diagnosis. X-rays are used frequently in areas of medicine and dentistry. It is likely that you or people you know have had X-rays at some time, for example, to check the development of teeth at the dentist or at a hospital for a suspected broken bone. (Figure 19.1 shows an X-ray of the lungs.) A more complex procedure is the CT scan. It also uses X-rays and may be used to obtain images of cross-sections through the body to enable diagnosis of such problems as brain disorders, ruptured spinal discs or damage to soft tissue in association with a bone fracture.

We will also consider the development of optic fibres and their use in endoscopes, which have allowed light to be used to examine the body internally and even to operate using keyhole surgery.

19.1 X-RAYS IN MEDICAL DIAGNOSIS

What are X-rays?

X-rays are electromagnetic waves of very high frequency and very short wavelength.

X-rays are part of the electromagnetic spectrum, discussed in the Preliminary Course (see *Physics 1*, chapter 3). X-rays are electromagnetic waves of very high frequency and very short wavelength, in the range 0.001 nm to 10 nm. Because of their high frequency, and hence high energy, they will penetrate flesh and may cause ionisation of atoms they encounter.

PHYSICS FACT

Effect of X-radiation on the body

If the intensity of X-radiation striking the body is great enough, it may be absorbed and cause electrons to be removed from atoms or molecules (ionisation). The effect may be harmful, which is why X-radiation is often referred to as ‘harmful ionising radiation’. One reaction that may occur is the ionisation of water molecules in the body and the subsequent formation of hydroxyl and hydrogen free radicals. (Free radicals are uncharged fragments of a molecule resulting from a covalent bond being broken. Each free radical has an unpaired electron and is highly reactive.) These free radicals may alter base structures and sequences in DNA in chromosomes, causing mutations.

The result may be somatic effects that affect only the person exposed to the radiation, or hereditary effects that affect the reproductive organs and may be passed on to the person’s children.

Radiation, which can cause damage to the body, includes alpha (α), beta (β) and gamma (γ) radiation as well as X-rays. The amount of radiation present is measured in units called sieverts (Sv). Dose limits that are considered to be safe are set by government bodies, therefore they vary from one country to another. For example, in Australia

the recommended limit for the general population is 1 mSv per year. This appears to be a conservative value as the limit for radiation workers is set at 20 mSv per year averaged over 5 consecutive years. These values are in addition to the background radiation from the Earth and from cosmic rays, which amounts to a value under 3 mSv. Approximate values for radiation from various sources are listed below.

Aircraft crew additional annual exposure due to cosmic rays	2000 μ Sv
Dental X-ray	<10 μ Sv
Chest X-ray	20 μ Sv
Pelvic X-ray	70 μ Sv
Mammogram	<4000 μ Sv
‘Barium meal’ X-ray	3000 μ Sv
CT scan of head	2000 μ Sv
CT scan of chest	8000 μ Sv

When X-rays pass through the body, energy is absorbed by the body tissue and the intensity of the beam is reduced. Denser material, such as bone, absorbs more X-radiation.

Production of X-rays

In the topic 'From ideas to implementation' (see chapter 10), you learnt that X-rays are emitted from a cathode ray tube when the cathode rays strike the glass of the tube. Similar principles are used to produce X-rays for medical diagnosis.

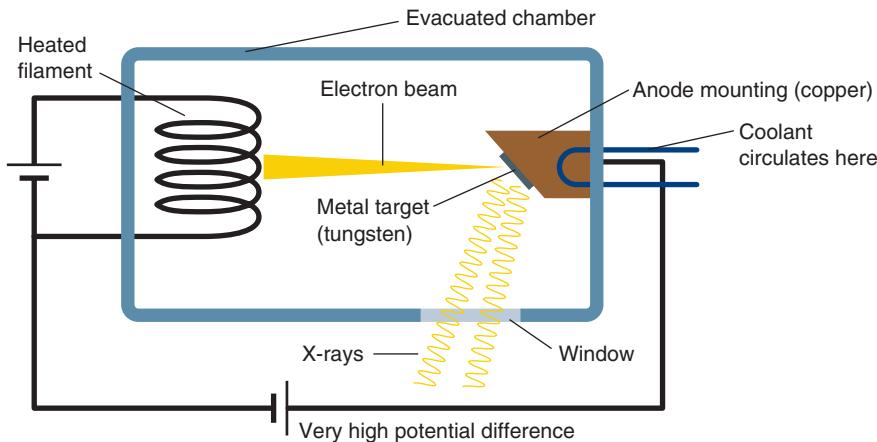


Figure 19.2 An X-ray tube

A cross-section of an X-ray tube is shown in figure 19.2. The tube is highly evacuated, and a very high potential difference, from 25 000 to 250 000 volts, is applied between the anode and cathode.

The cathode is a filament of wire through which a current is passed. Electrons are emitted from the hot filament and a metal focusing cup directs the electrons towards the anode. The very high potential difference between the cathode and anode accelerates the electrons to the anode. The anode is usually made of tungsten that can withstand the high temperatures generated. When the electrons strike the tungsten they are absorbed and some of their energy is converted to X-rays. By placing the tungsten target at an angle to the incoming electron beam, the X-rays that are emitted from the tungsten can be sent in a pre-determined direction.

Tungsten is usually used for the target as it has a very high melting point of about 3400°C and emits X-rays when struck by electrons. The production of X-rays is not very efficient as only about 1 per cent of the energy reaching the target is converted to X-rays, the rest being converted to heat. The heat generated in the target per second can be enough to heat a cup of water to boiling point in one second. Hence it is important to prevent the target from overheating or melting. Copper — a good conductor of heat — is used for the anode mountings, and oil circulating in the outer region near the anode helps the cooling by convection. Cooling can also occur by rotating the target at a rapid rate of approximately 3600 revolutions per minute, allowing the heat produced to be distributed over a large area.

Use and detection of X-rays

Since X-rays cannot be focused, the images from X-rays are shadows of objects placed in the beam. To obtain the sharpest image it is necessary to have an object that is as still as possible and illuminated by an X-ray beam of small cross-sectional area, with the detecting plate as close to the

object as possible. In this way, blurring of the image is minimised as the penumbra of the shadow is reduced. The surrounding material also affects the sharpness of the image as scattering of X-rays from surrounding tissue may occur. This is illustrated in figure 19.3, using light instead of X-rays, a foot to represent internal bone and cloudy water to represent surrounding body tissue.

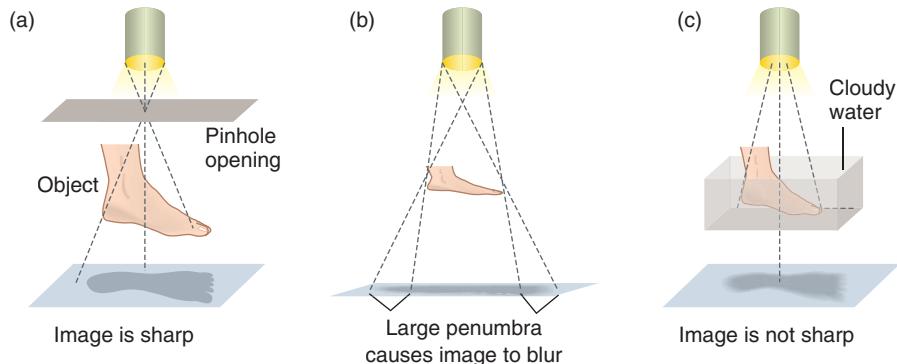


Figure 19.3 *Obtaining a sharp shadow image. (a) A narrow source produces a sharp shadow. (b) An extended source or large distance between object and screen results in a blurry image due to the large penumbra. (c) Cloudy water scatters light and produces a blurry image.*

A narrow beam of X-rays can be obtained by controlling the angle of the anode target. This technique allows the beam of electrons to strike the target over a reasonable area while significantly reducing the width of the X-ray beam, as illustrated in figure 19.4.

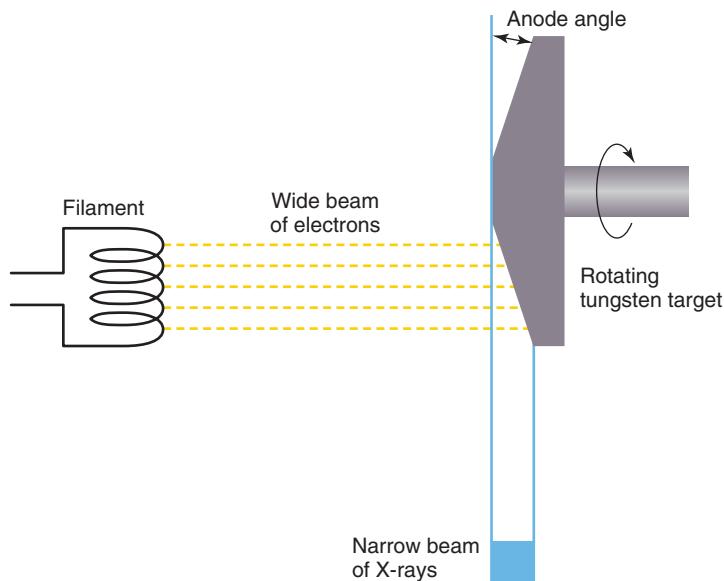


Figure 19.4 *Electrons hitting the target over a wide area produce a narrow beam of X-rays.*

The X-ray beam is directed at the part of the patient being imaged. Some tissues absorb X-rays very well and cast a shadow on the detecting screen. Bone is more dense than soft tissue and absorbs X-rays. Consequently bones produce a clear image when X-rayed.

X-rays may be detected on a photographic film or by an image intensifier. The photographic film is used when a record of the image is required. An image intensifier allows direct viewing of the X-ray image.

X-rays strike a phosphor screen that produces light. This light stimulates a photocathode to produce electrons that are accelerated to strike an output phosphor screen, producing more light than was originally generated and intensifying the image up to 1000 times. The image produced can be viewed directly by the eye, a movie camera or a TV camera. The viewing area can be altered while the X-ray process is occurring.

PHYSICS FACT

In shoe shops in the first half of the twentieth century, a fluorescent screen was used to observe the X-rays passing through a person's foot to see if toes were squashed by shoes that were too tight. These screens were banned from about 1960 because of the danger from exposure to scattered radiation — a danger that clearly outweighed any benefit in shoe fitting, as is obvious from the figure 19.5.

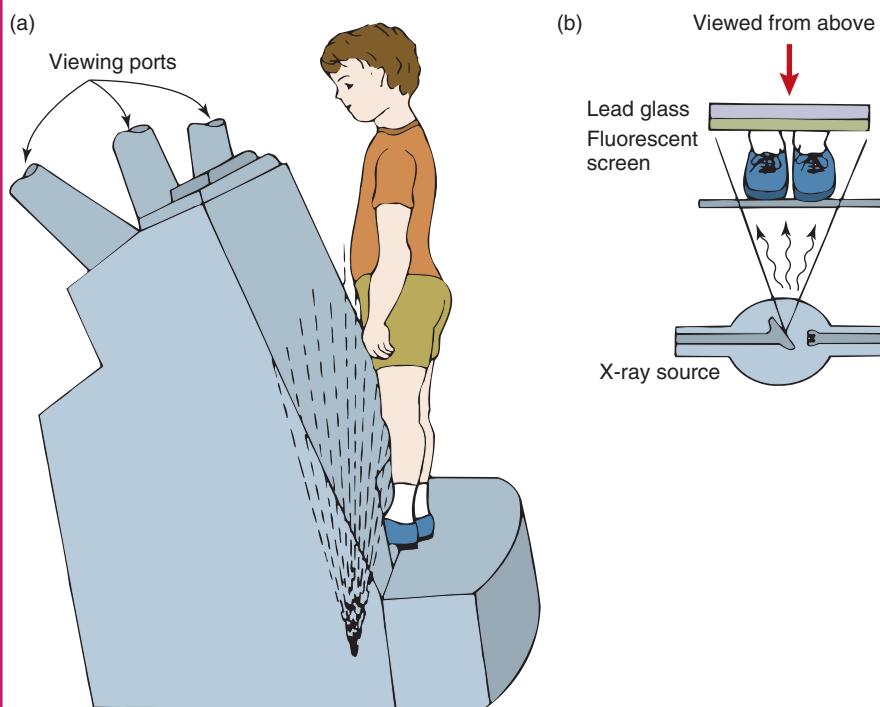


Figure 19.5 (a) Shoe fluoroscope showing a considerable amount of scattered X-radiation (dashed lines) striking other parts of the boy's body, for example, the reproductive organs. (b) X-rays are used to produce an image of the boy's feet.

A fluorescent screen did not produce a bright enough image to view easily. Rather than increasing the intensity of the X-ray beam, the fluorescent screen was replaced, in medical diagnosis, by the X-ray intensifying technique using phosphor screens and photocathodes.

Types of X-rays

As outlined earlier, X-rays are produced when electrons strike a target. There are two mechanisms by which X-rays are produced.

The first mechanism produces X-rays with a range of frequencies. The electrons are slowed down by the target atom and some of each electron's kinetic energy is converted to electromagnetic radiation corresponding to

X-radiation. The frequency of the X-radiation depends on the amount by which the electron has been slowed, or, in other words, the amount of braking that has occurred. This radiation is called the Bremsstrahlung radiation, which is German for ‘braking radiation’.

The second mechanism results in the ionisation of the atom and so the frequency of the X-radiation produced depends on the nature of the target atom. A series of frequencies that make the characteristic or line spectrum are produced. To produce X-radiation of a particular frequency, an electron is knocked out of an inner shell of an atom by the approaching electron. An outer shell electron takes the inner shell electron’s place, losing energy equal to the energy difference between the two shells. The energy difference between electrons in the two shells determines the frequency of the X-ray produced.

Figure 19.6 shows the X-rays produced by a typical target.

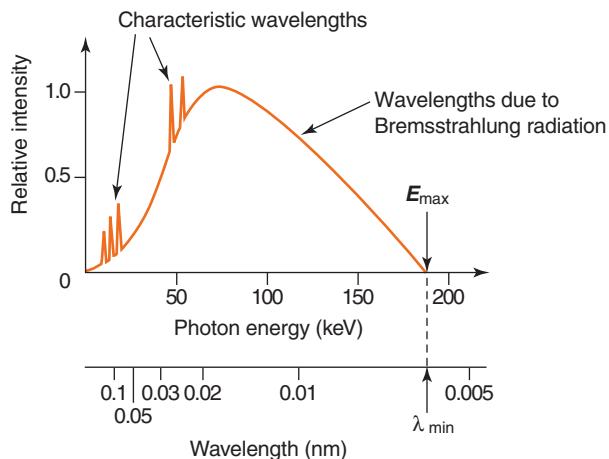


Figure 19.6 The energy of the X-ray photons (and the corresponding wavelength) plotted against the intensity of the X-rays

Hard and soft X-rays

A thin sheet of material placed in the path of the X-ray beam will act as a filter and absorb more low-energy photons than high-energy photons. The beam will now have more high-energy photons, which are more penetrating. The beam is said to be **hard**. By contrast, a beam of X-rays with lower energy photons is less penetrating and is said to be **soft**.

Note that the higher the energy of the photons, the higher their frequency and the shorter their wavelength.

Hard X-rays are preferred for imaging as they penetrate the body and are absorbed by material such as bone, allowing images of the bone to be observed. Soft X-rays are not useful for imaging as they will not penetrate the body. They expose the patient to additional useless and possibly harmful X-radiation.

Using conventional X-rays as a diagnostic tool

X-rays have a number of different effects on the tissues of the body, depending on the energy of the X-ray photons and the time of exposure to them. For diagnostic purposes, the optimal photon energy is around 30 keV, resulting in the best contrast between different tissues. At this energy the photoelectric effect dominates. This means the X-rays are absorbed by the tissues and electrons are released. The extent of the X-ray absorption depends on the cube of the number of protons in the nuclei of the atoms encountered. For example, bone, which has a high

Hard X-rays consist of high-energy photons and are more penetrating than soft X-rays, which have lower energy photons.

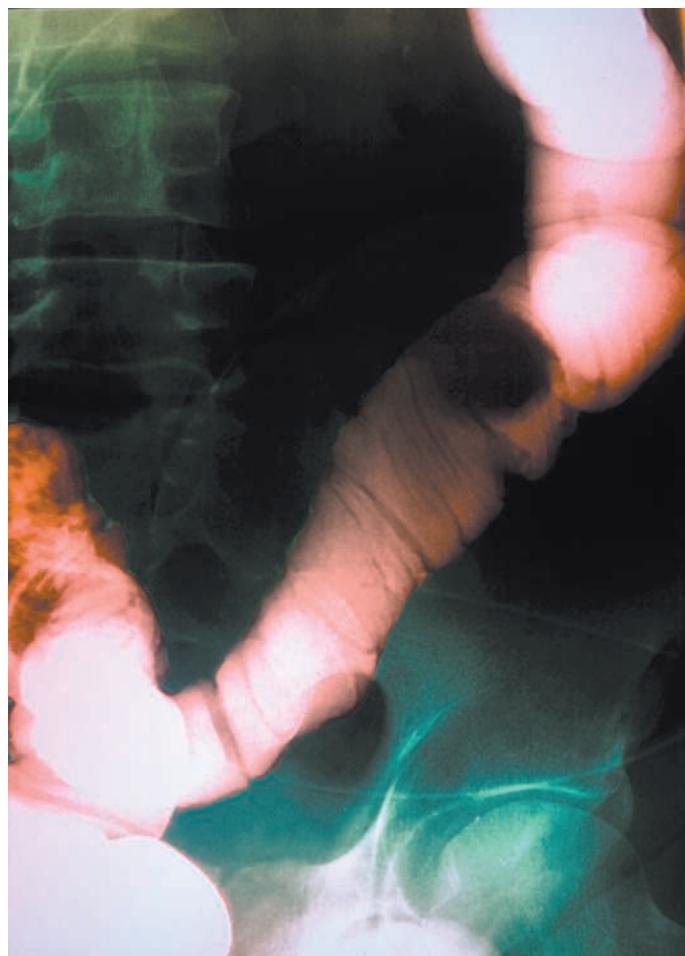
Attenuation of a signal like an X-ray beam is the reduction in intensity of the beam.

atomic density (high number of protons in the nuclei), will **attenuate** the beam about eleven times more than the surrounding tissue and hence produce a strong X-ray shadow, allowing a very good image of the bone to be obtained.

Atomic density values are high for bone, moderate for soft tissue and low for air. Hence the skeleton is imaged very well by X-rays.



Figure 19.7 X-ray showing a fracture of the radius and ulna in the forearm



Imaging parts of the body

To image soft tissue, an artificial contrast medium that absorbs X-rays readily may be introduced. Iodine in a compound is introduced into the bloodstream for investigations of the circulatory system. To X-ray the gastrointestinal tract, which is composed of soft tissue, a ‘barium meal’ consisting of a thick suspension of barium sulfate is swallowed by the patient or introduced into the intestines through the anus. The barium compound absorbs X-rays and gives a clear image, as shown in figure 19.8.

A chest X-ray is the most common way of detecting lung cancer or tuberculosis. The X-ray must be taken from several different directions to overcome the problem that the heart sometimes obstructs a clear view of the lungs.

The teeth and jaw are X-rayed to detect tooth decay as well as crowded teeth or wisdom teeth, before surgery or orthodontal treatment is recommended.

Someone who has swallowed a foreign object may be X-rayed to locate the position of the object.

Figure 19.8 Barium sulfate in the bowel shows as white contrast. Two narrowings, one due to a tumour, are shown. The black sections of the bowel are where the barium compound has coated the wall of the bowel, which is filled with gas.

An X-ray may be needed to determine whether a patient has a metallic implant before an MRI examination is ordered (see figure 19.9(a)).

For imaging the breast, which is an area of continuous soft tissue, careful choice of the X-ray beam and film detector provides high resolution (see figure 19.9(b)). Molybdenum targets in the tube and low voltage maximise photoelectric attenuation. High tube current and short exposure time minimise image blur due to movement by the patient.

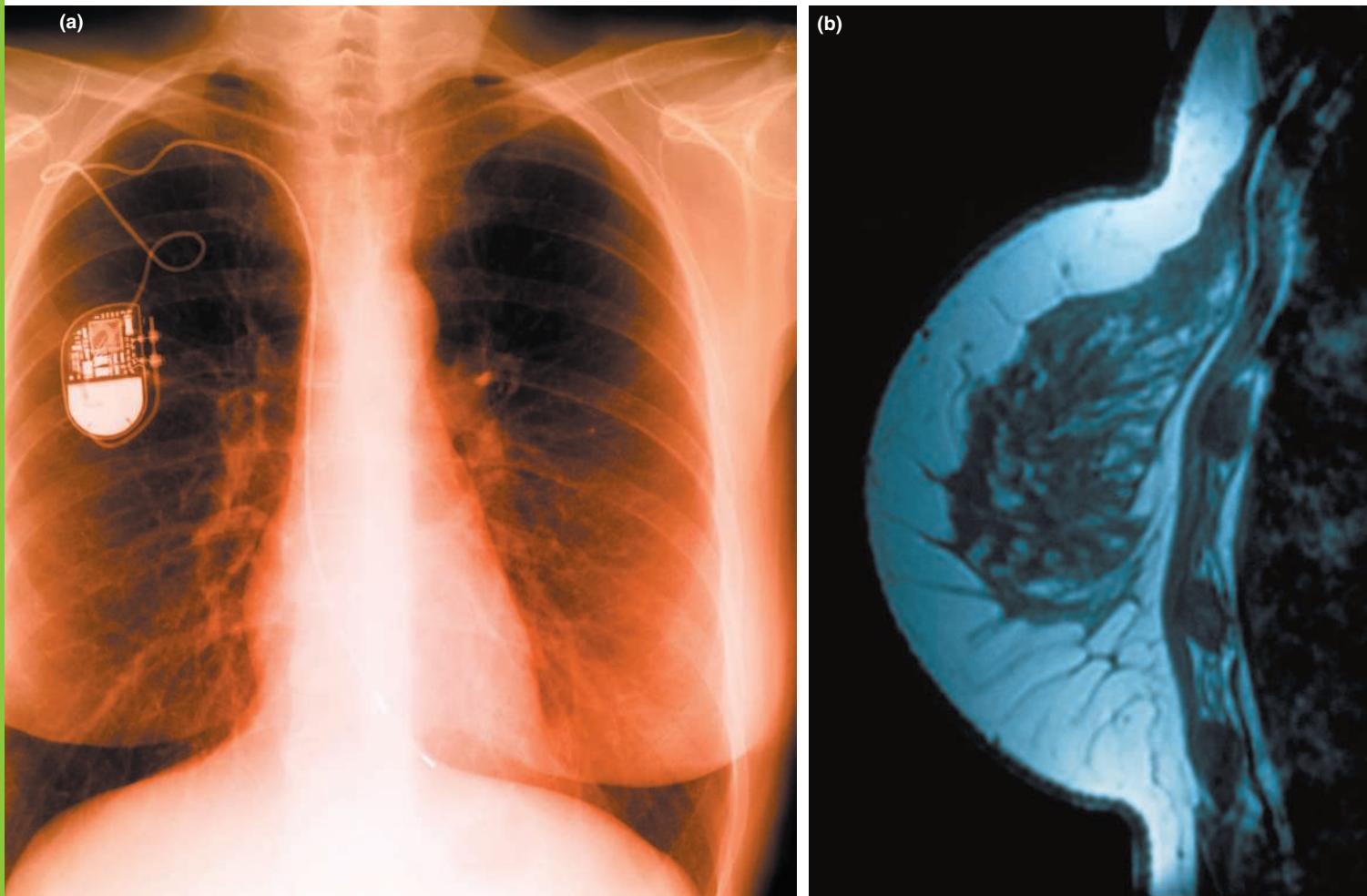


Figure 19.9 (a) An X-ray image showing a heart pacemaker (b) An X-ray image of a breast showing a tumour

As we will see in the next section, a better technique, called computed axial tomography (CT), is used for imaging soft tissue as it detects small differences in X-ray attenuation.

19.2 CT SCANS IN MEDICAL DIAGNOSIS

Computed axial tomography scanning (or CT scanning) uses X-rays to obtain an image of a cross-section through the body. Very slight differences in X-ray attenuation can be measured and so soft tissue can be accurately imaged.

PHYSICS FACT

Godfrey N. Hounsfield was born in England in 1919 and educated as an electrical engineer. He joined the medical systems section of the firm EMI in 1951 and his long career in medical research and engineering led to his invention of the computed axial tomography scanner, for which he earned the Nobel Prize for Medicine in 1979.

Tomography comes from the Greek word *tomos* meaning ‘slice’. The CT scanner produces an image of a slice through the object it is examining. Hounsfield analysed the data by computer, using a technique that was originally developed for use in astronomy.

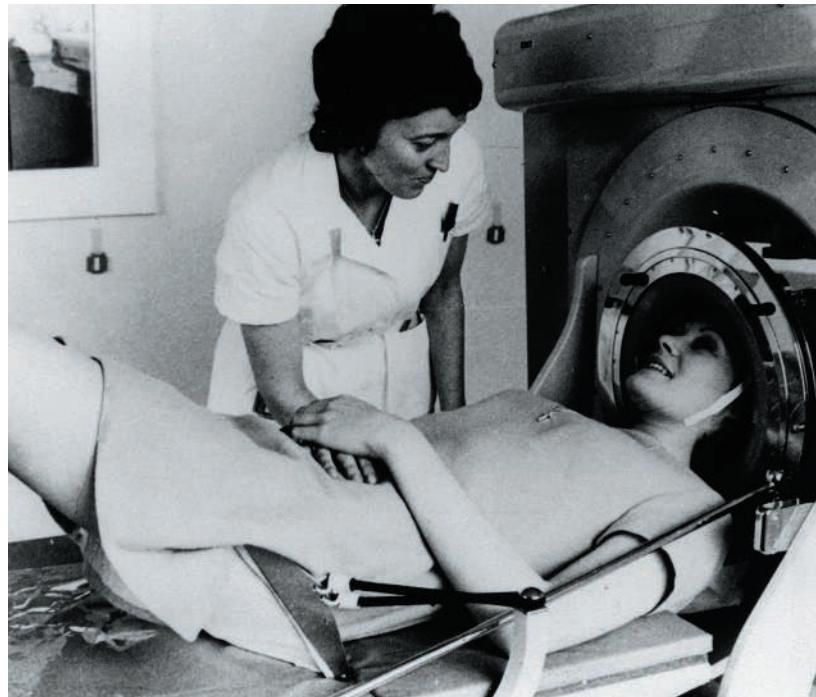


Figure 19.10 The first commercial CT scanner. The head is surrounded by a rubber bag filled with water to absorb scattered X-rays.

How is a CT scan produced?

A CT scanner consists of an X-ray tube that is rotated around the patient being imaged. The tube and detection mechanism are mounted on a frame called a gantry. The part of the patient’s body being scanned is positioned in a gap in the gantry. An image is obtained in the plane being examined. The patient, on a bed, is moved slowly through the gantry so that a series of images ‘slices’ through the body may be obtained.



Figure 19.11 A modern CT scanner showing the control console and gantry assembly

The X-ray source must produce a very narrow beam so that the path of the X-rays can be carefully controlled. To produce the narrow beam the tube voltages are high and consequently a lot of heat must be conducted away from the anode in the tube generating the X-rays. This requirement, coupled with the tube movement during scanning, results in the tubes failing and having to be replaced after a few months of use. The cost of such replacement is high.

The beam needs to be filtered to remove some soft X-rays which are not needed. This ensures that the beam is relatively homogeneous and the dose to the patient's skin is reduced. The X-rays are detected by an array of several hundred detectors. The newer detectors convert the X-radiation directly into electrical signals that go to multiple integrated-circuit amplifiers.

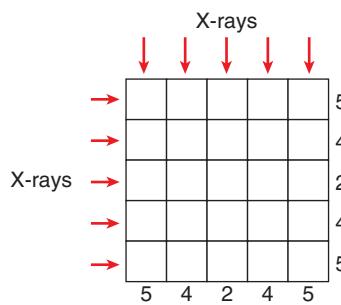
The patient is accurately positioned in the gantry so that a plane of the body can be scanned. A beam of X-rays is sent through the patient and detected on the other side. The beam is then rotated, usually 1° , and another beam transmitted and detected. This process is repeated until an angle of 180° has been swept out.

PHYSICS FACT

Early CT machines used a single X-ray beam and a single detector, and to rotate over the full 180° took about 20 minutes. This was because each beam scan took about 5 seconds followed by a rotation of 1° followed by a 1-second delay to allow the machine to stop vibrating. Recent improvements using several hundred stationary detectors and a rotating X-ray tube reduce the scan time to less than 2 seconds and the slice to 2 mm. In more recent machines, the table on which the patient lies is moved in a smooth, stepless motion with the tube assembly rotating continuously and these have resulted in a spiral scanning method in which image data is obtained faster than 5 images per second. By using paired detectors it is possible to scan two slices 1 cm apart simultaneously. The increased speed of data collection has made possible such studies as CT angiography (dynamic studies of the blood vessels of a beating heart). CT angiograms obtained in a 9-second scan are increasingly popular to diagnose blocked arteries in seemingly healthy people.

The data from the scan is collected, displayed and reconstructed using a powerful computer. The computer analyses the absorption of the X-rays at each measured point in the slice. For example, if X-ray beam absorption is measured at 160 distinct points along each scanning path and 1° increments in angle are used, approximately 29 000 distinct pieces of data about X-ray absorption are obtained. The reconstruction, which is explained in simplified form in figure 19.12, is the result of around one million computations. The image can be displayed on a TV screen or stored in the computer's memory and used with other data to produce an image in a different plane.

In recent years, full CT body scans have been advertised for those who want to detect problems before symptoms appear. The medical profession has criticised this opportunity, citing several reasons. People are exposed to unnecessary radiation, potential problems may not be detected and harmless abnormalities may be found. Hence people are given either false security or false alarms. (For internet information about full CT body scans, use 'CT scans' or 'full body scans' as key words in a search engine.)

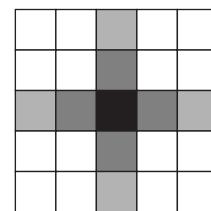


- (a) The attenuation is measured at many points (pixels) from different angles. (Here only 2 angles at 90° are recorded.)

attenuation = 5 + 5	10	9	7	9	10
	9	8	6	8	9
attenuation = 5 + 2	7	6	4	6	7
	9	8	6	8	9
	10	9	7	9	10

attenuation = 5 + 2

- (b) The total attenuation of the X-rays at each pixel is calculated.



- (c) A shade of grey is assigned to each pixel and from this the image is created. (As an example assign the darkest shade to the smallest number.)

Figure 19.12 Creating a CT scan image

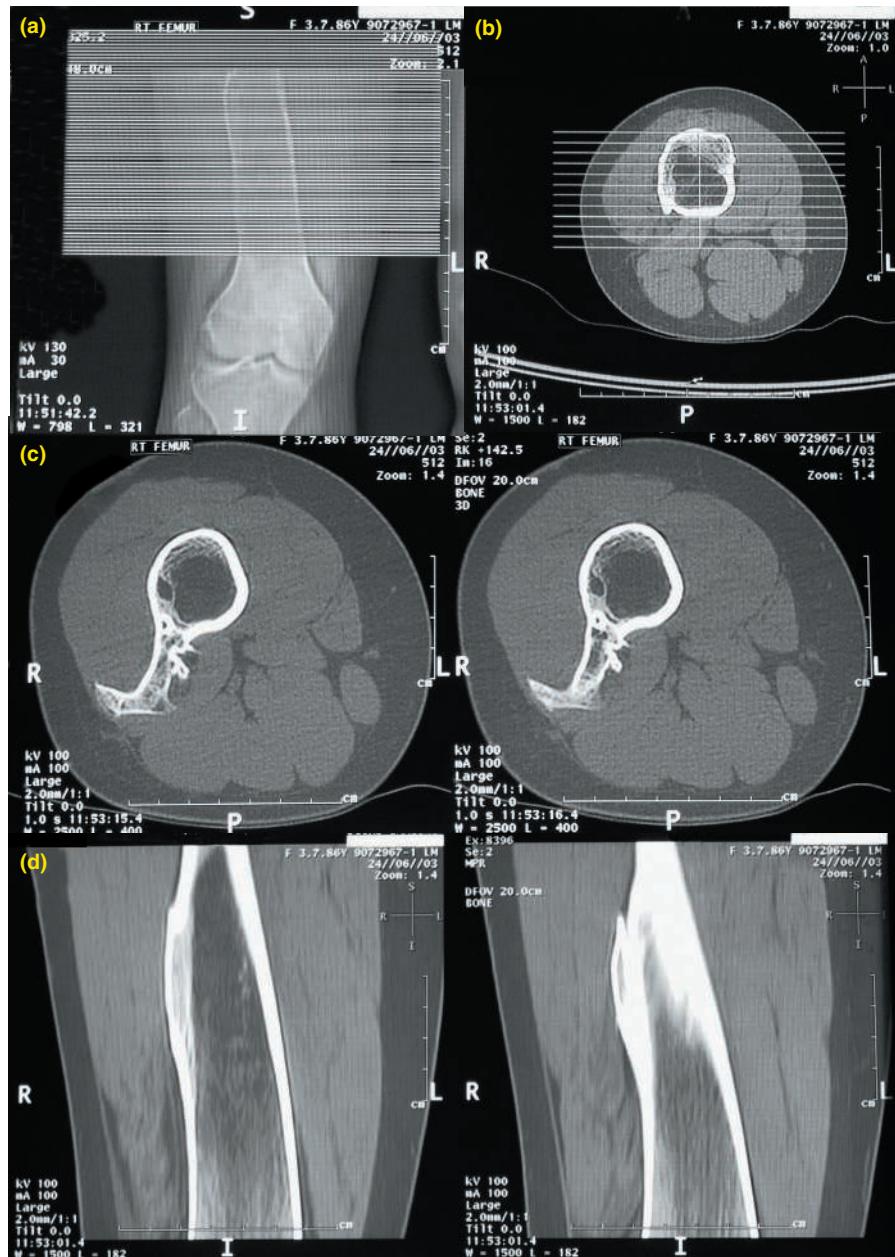


Figure 19.13 Images showing positions of (a) transverse and (b) longitudinal ‘slices’ of the upper leg (femur) taken by a CT scan (c) Two transverse slices through the femur showing positions of a tumour (d) Two longitudinal slices showing the same tumour

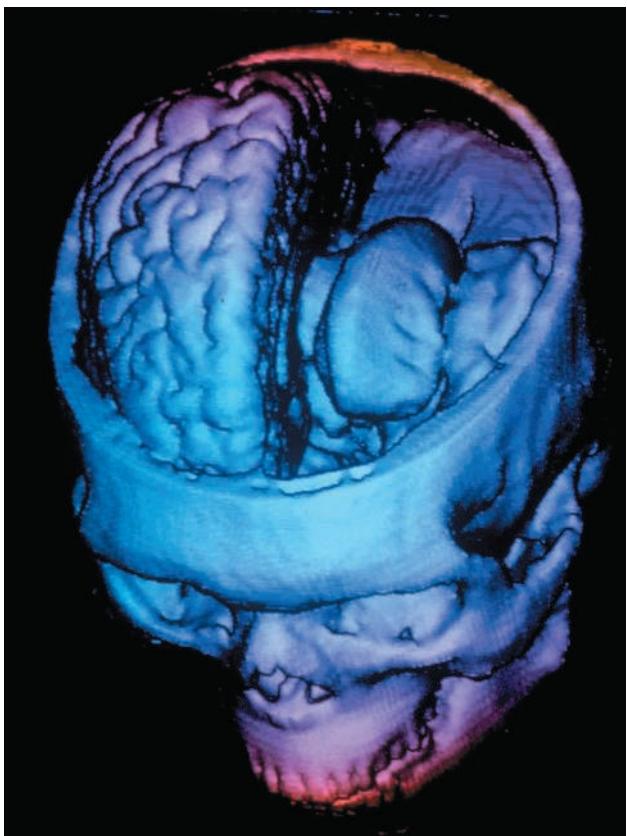


Figure 19.14 Using computer analysis, the data from images of 'slices' through the body can be combined to produce a three-dimensional image of the area under investigation.

CT scans as a diagnostic tool compared with X-rays and ultrasound

CT scans are more expensive than ultrasound and significantly more expensive than conventional X-rays. They are, however, a superior diagnostic tool to ultrasound and X-rays when fine detail is needed. This is the case when an image of the brain is required.

CT scans provide detail to distinguish between areas where the density difference is quite small even though a dense material shields the area. In the brain, the density range is only a few per cent but the bony skull is so dense that most of the X-rays are absorbed by the skull. A conventional X-ray will therefore provide an image of the dense skull rather than the brain tissue inside. However, by taking X-ray images from many angles in a CT scan, the material along the path of the X-ray beams can be distinguished clearly. Due to the method of obtaining and analysing the image it is possible to see behind bone using a CT scan. With ultrasound, imaging behind bone is not possible as the ultrasound signals are reflected strongly from a tissue-bone interface so they do not reach the tissue beyond the bone and cannot give information about this material. Hence ultrasound cannot be used to image the brain through the skull.

X-rays are valuable when there is high natural contrast between the tissues to be viewed. This means there is a large difference in proton number. The proton number is high for bone, moderate for soft tissue and low for air. Hence X-rays are good for diagnosing bone problems such as fractures, dislocation and arthritis. Conventional X-rays can also be used to image soft tissue if an artificial contrast medium is introduced, such as a barium meal if the digestive tract is being imaged (see page 367).

CT scans provide much better resolution of soft tissue than conventional X-rays. They can be used to investigate soft tissue damage due to bone fracture or ruptured spinal discs, or in other areas where a contrast medium cannot be easily introduced. CT scans are also used to scan the liver and kidneys to obtain resolution better than 1 mm. This means differences separated by 1 mm can be detected. Ultrasound cannot produce as clear an image as this.

CT scans are preferred for imaging the lungs (see figure 19.1, page 361). Although conventional X-rays give adequate routine lung screening, CT scans provide clearer detail. Ultrasound cannot image past air spaces in the lungs.

Ultrasound is the preferred imaging method for viewing blood flow, whereas CT scans are limited unless sophisticated CT angiography is used. X-rays need a contrast dye such as an iodine compound to image this area.

X-rays of sufficient intensity are a harmful ionising radiation whereas ultrasound does not produce any ionisation. Hence ultrasound is safe for use with foetuses. Ultrasound could in fact be used many times with the same patient without any harmful effects. The specific gain from a CT

scan or X-ray would need to be considered in each case before exposing the patient to the X-ray doses involved.

Often a cheaper, quicker and relatively portable imaging technique using X-rays or ultrasound can give an initial diagnosis that could lead to further testing for tissue damage or internal bleeding by ordering a CT scan.

19.3 ENDOSCOPES IN MEDICAL DIAGNOSIS

An **optical fibre** is a glass core surrounded by a cladding of lower refractive index. Light is transferred along the optical fibre by total internal reflection.

Endoscopes are optical instruments using light for looking inside the body to examine body organs, cavities and joints. Modern endoscopes use **optical fibres** to transfer light to and from the area being examined. Light is transferred along the optical fibre by total internal reflection — the same concept that you studied in relation to communication technologies in *Physics 1*, chapter 4.

PHYSICS FACT

The first endoscope was invented in the late nineteenth century. It was a straight, rigid metal tube illuminated by an oil lamp and it appears that sword swallowers were the first subjects for experimentation. The tale is told that one such recruit exclaimed, ‘I’ll swallow a sword anytime, but I’m damned if I’ll swallow a trumpet!’

In the twentieth century, Rudolf Schindler produced an endoscope with a semi-flexible tube using prisms and lenses to bend the light into an arc. This endoscope was still quite bulky and must have been uncomfortable. A real breakthrough came in the 1960s with the invention of the optical fibre. Endoscopes that were very flexible and of narrow diameter could be made, much to the relief of patients.

Optical fibres and transmission of light

An optical fibre is a glass core surrounded by a glass cladding of lower refractive index than the core. You will recall from your Preliminary Course studies that a critical angle exists for light travelling in the core. If this critical angle is reached or exceeded by light striking the core-cladding interface, the light is totally internally reflected and trapped in the core. By reflecting off the core-cladding interface, the light can travel along the core whether the core is bent or straight (see figure 19.15).

Usually optical fibres are grouped together in a bundle. For endoscopes the bundle of optical fibres may contain up to 10 000 fibres.

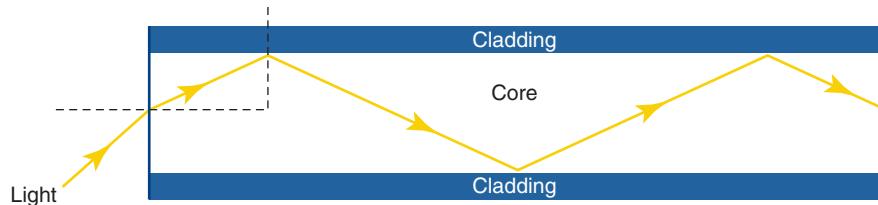
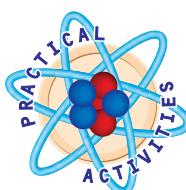


Figure 19.15 Cross-section of a glass fibre showing total internal reflection



19.1

Transfer of light by optic fibres

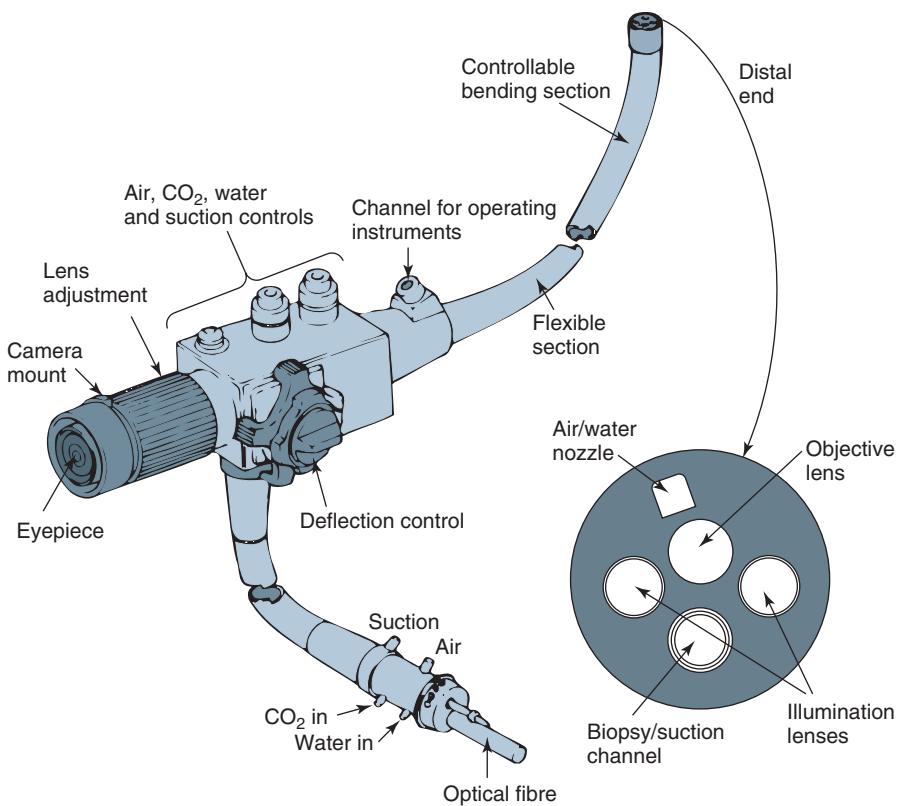


Figure 19.16 A fibre-optic endoscope

A **coherent** optic fibre bundle is one in which the optic fibres keep the same position relative to one another. A **non-coherent** optic fibre bundle is one in which the fibres are not kept in the same position relative to one another.

Structure and use of an endoscope

Figure 19.16 shows a diagram of a typical endoscope. The long flexible shaft containing the fibres is protected from damage by a helical steel band inside steel mesh. The outer coating is plastic to prevent chemical damage and to make the shaft waterproof and easy to move through the body. The shaft is about 10 mm in diameter and may be up to 2 m long.

The end inserted in the patient's body, known as the distal end, is able to bend in directions that are controlled from the viewing end by the operator. The distal end contains a lens to focus the image onto the end of the fibre bundle. The shaft of the endoscope contains a number of distinct parts as listed in table 19.1.

Table 19.1 Function of parts of the endoscope shaft

PART OF SHAFT	FUNCTION
Non-coherent optic fibre bundles (usually two)	To guide the light to the area to be examined
Coherent optic fibre bundle	To transmit the image back to the eyepiece for viewing
Water pipe	To wash the distal face of the endoscope to keep the optical section clear
Operations channel	To insert surgical instruments to perform specific tasks (see figure 19.17 for some surgical instruments)
Control cables	To operate the flexible end
Additional optional channel	To suck or pump in air or carbon dioxide

Once the image is sent to the viewing end of the endoscope it may be viewed by the operator directly or captured as a still photograph or video record (see figure 19.19, page 377). In this way an operation can be accurately controlled and also recorded for later study.

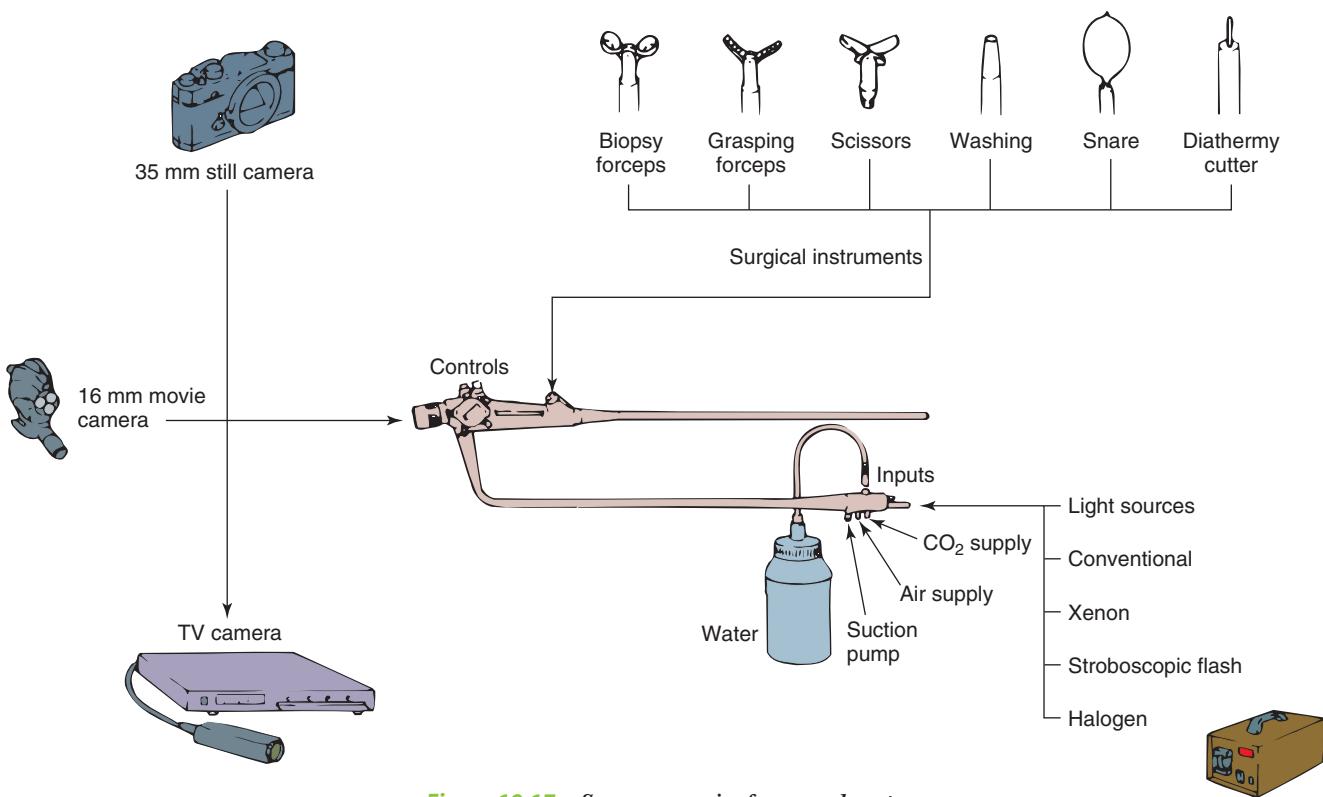


Figure 19.17 Some accessories for an endoscope

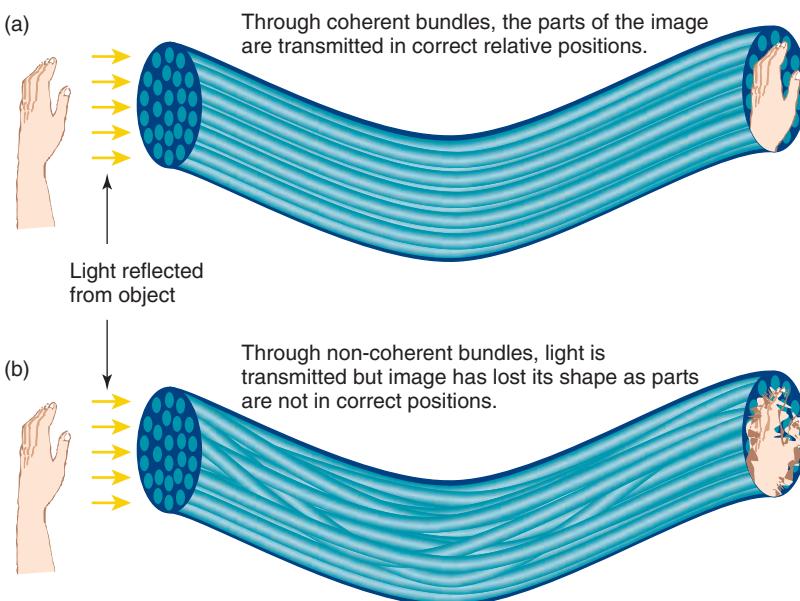
How do endoscopes work?

Endoscopes work because light can be transmitted along them to illuminate internal areas of the body and then the image of those areas can be carried back along the fibres of the endoscope.

Bundles of optic fibres are needed to carry the light to and from the inside of the body. These fibres operate by the principle of total internal reflection as indicated previously. The optic fibre bundles carrying light into the body can be non-coherent, and hence less expensive, as their role is only to illuminate the internal area. These non-coherent bundles can have thicker fibres, which are more efficient at transmitting light as there are fewer reflections along a given path length and hence fewer opportunities for the light to be lost through the sides of the fibre (see figure 19.18(b)).

The optic fibre bundle transmitting the image back along the shaft must be coherent so that when the light reaches the viewing area it will be in the same relative position as it was when it left the object being viewed (see figure 19.18(a)). It is an advantage to make these fibres as narrow as possible with the core to cladding ratio as large as possible to ensure that there are many beams of light transmitting the image. The image will therefore be very clear — in other words, it will have good resolution.

Figure 19.18 (a) A coherent bundle of optic fibres (b) A non-coherent bundle of optic fibres



Using an endoscope

The endoscope is inserted through a natural orifice in the body or through a small incision. It allows doctors to see inaccessible parts of the body and in some cases to carry out minor surgical procedures. Such a procedure, guided by use of an endoscope, is called keyhole surgery. Some of the uses of endoscopes are indicated in table 19.2.

Table 19.2 Uses of endoscopes

NAME OF PROCEDURE	PLACE OF INSERTION OF ENDOSCOPE	PURPOSE OF PROCEDURE
Arthroscopy	Through skin near joint	To examine joints and carry out repairs such as removal of torn cartilage
Bronchoscopy	Through bronchial tubes	To examine trachea and lungs to show problems such as inflammation, bronchitis, cancer and tuberculosis
Colonoscopy	Through the anus	To detect problems such as polyps, tumours, ulceration and inflammation in the colon and large intestine
Colposcopy	Through the vagina	To look for problems such as inflammation and cancer in the vagina and cervix (in females)
Cytoscopy	Through the urinary tract	To examine the bladder, urethra and opening of the prostate gland (in males)
Endoscope biopsy	Through a natural opening or through an incision	To remove specimens of tissue for examination and analysis by a pathologist
Gastroscopy	Through the mouth	To look for the source of problems such as bleeding from the lining of the oesophagus, stomach and duodenum
Laparoscopy	Through an incision in the abdominal wall	To examine abdominal organs including the stomach, liver and fallopian tubes (in females)

If a tumour is detected, a small sample may be taken for analysis, using the sampling implements in the endoscope. Such a sample is called a biopsy.

Polyps may be surgically removed using an endoscope with an attachment. Sometimes lasers are used in conjunction with endoscopes when surgery is being carried out, as lasers will cut without distorting the area around the incision and without causing bleeding.

Before the use of endoscopes, open surgery was needed to examine and treat internal organs. The procedure using endoscopes is less invasive and carries less risk, allowing the patient to recover quickly. It may be carried out in an outpatients department and so not require admission of the patient to hospital. The cost of health care is thus reduced.

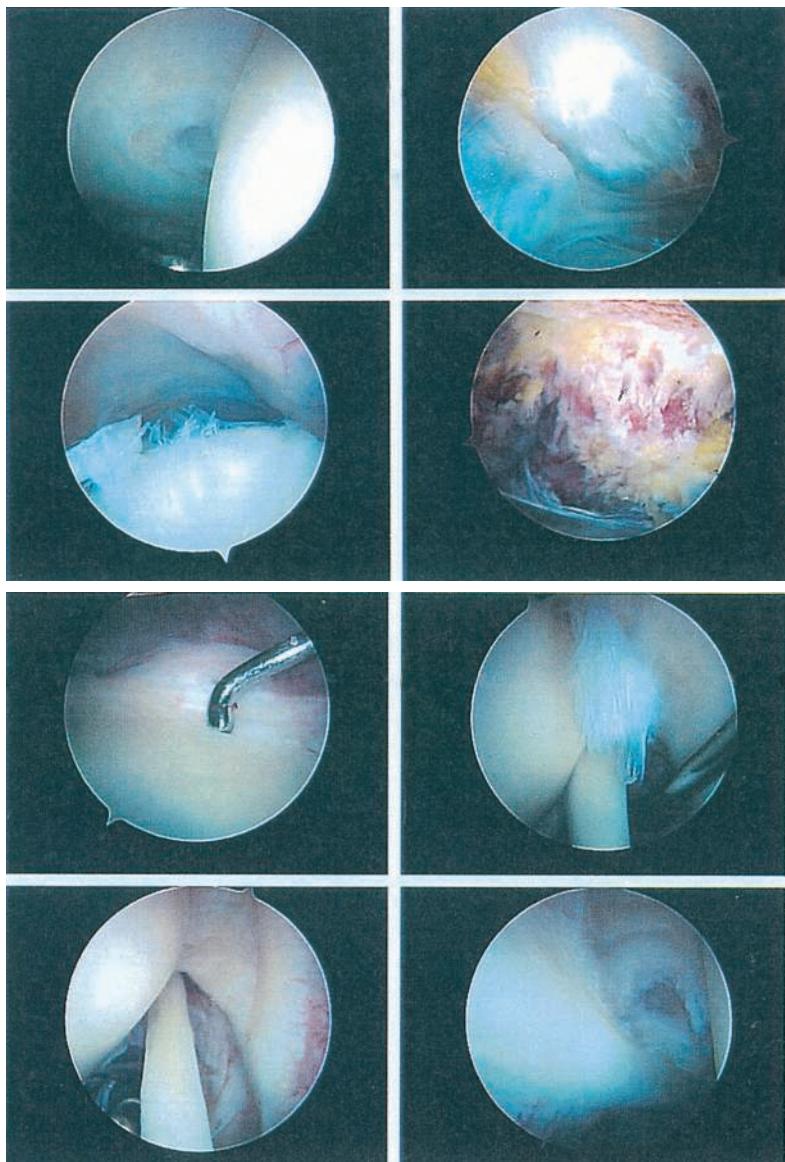


Figure 19.19 Images taken during an operation on a damaged shoulder; using an endoscope

SUMMARY

- X-rays are produced by the collision of electrons with a target material.
- Soft X-rays are less penetrating and have lower frequency than hard X-rays.
- A CT scan is produced by the computer analysis of the attenuation of X-rays moving around a slice of the body.
- A series of consecutive 2-D scans can be stored by the computer and combined to produce a 3-D image.
- CT scans can distinguish soft tissue with small differences in density and can produce an image of tissue behind bone.
- CT scans are expensive compared with conventional radiographs.
- CT scans can provide confirmation of suspected problems, such as tumours, damaged cartilage or internal bleeding.
- Endoscopes use optical fibres to transfer light to and from the inside of the body to enable internal structures to be seen.
- Non-coherent bundles of optical fibres transfer light to the inside of the body.
- An image is formed when light is transferred along coherent bundles of optical fibres from inside the body to the light receptor.

QUESTIONS

1. Outline the property of electrons that allows them to be focused using electric and magnetic fields but prevents X-rays from being focused.
2. X-rays used for diagnosis are generated with tube voltages of about 70 kV compared with several megavolts when X-rays are needed to destroy tissue.
 - (a) State what is meant by ‘diagnosis’.
 - (b) Outline the effect that X-rays produced with tube voltages of several megavolts will have on body tissue.
3. With the aid of a labelled diagram, give a description of the way in which X-rays are produced.
4. For a typical X-ray tube with a tungsten target:
 - (a) sketch a graph that shows how the intensity of the resulting X-radiation varies with photon energy
- (b) explain (i) the range of frequencies obtained and (ii) the sharp peaks on your graph
- (c) redraw the graph to show how it would be changed if some soft X-rays were removed by filtering.
5. (a) Outline how the attenuation of X-rays changes for different materials in the body.
(b) Describe and account for the appearance of an X-ray image of part of the body containing bone, muscle and air spaces.
6. State a difference between ultrasound and X-rays and outline why this difference is important for the way each is used.
7. Use a table to compare hard and soft X-rays.
8. This question refers to figure 19.15 on page 373. Assume the external medium is air.
 - (a) Explain what is meant by the ‘critical angle’ of a material.
 - (b) Outline why a critical angle is important in an optical fibre.
 - (c) Describe the function of the cladding in an optical fibre.
 - (d) State the relationship between the refractive index of the cladding and the refractive index of the core of the fibre.
9. (a) Describe how the fibres are positioned in a coherent bundle.
(b) Explain why a coherent bundle is necessary in an endoscope.
(c) Explain why the endoscope has to be used in conjunction with a powerful light source.
(d) Which properties of the fibre bundle affect the ability of the observer to see small details when using the instrument?
10. ‘The main principle behind the operation of an endoscope is the transfer of light to and from the internal organs of the body.’ Evaluate this statement.
11. Use a table to summarise situations in which CT scans are a superior diagnostic tool to X-rays or ultrasound. For each situation, outline why X-rays and ultrasound are inferior.
12. An endoscope is used to take a biopsy of a small tumour in the oesophagus, which leads from the mouth to the stomach. Explain how an endoscope can be used to do this.
13. Explain how an endoscope could be used to examine and repair a torn ligament inside the knee joint.

14. Examine the image of the lungs taken by an X-ray (figure 19.1 on page 361) and a CT scan of the upper leg (figure 19.13 on page 371). Compare the information provided by each image.
15. Find at least three different X-rays images on the internet, using key words such as ‘X-ray image of fracture’, ‘mammogram’, ‘barium meal X-ray’, ‘lung X-ray’. Outline why X-rays have been successful in producing the image in each case.
16. From the internet or other sources, find at least three images of internal organs obtained by using an endoscope. If searching the Internet, use key words such as ‘laparoscopy images’ or ‘colonoscopy images’. For each image, describe the internal organ as viewed from the endoscope.
17. (a) Describe how a CT scan is obtained.
(b) Outline why improvement in computer technology is linked strongly with clearer CT scans.
18. From the internet or other sources, find at least three different images of the body obtained using CT scanning. For each image, describe the detail that is visible and outline how the image would be different if ultrasound or X-rays were used.



19.1 TRANSFER OF LIGHT BY OPTIC FIBRES

Aim

- (a) To demonstrate the transfer of light by optic fibres.
- (b) (Extension) To compare light transferred through optic fibres with light transferred through the air.

Apparatus

optic fibre bundle

light source

light probe and data logger (or lightmeter)

Theory

Light is transferred by total internal reflection along an optic fibre.

The intensity of light travelling through air decreases according to the inverse square law (see *Physics 1 Preliminary Course, 3rd edition*, chapter 3).

Method

1. (a) Shine the light in one end of the optic fibre bundle.
(b) Observe the light coming from the other end.
(c) Bend the bundle and again observe the light coming out the other end.
(d) By carrying out the experiment in a dark room, note whether light is lost through the sides of the optic fibre.
(e) Carry out step (d) with most of the bundle under water and note any changes to your observation. If there was a change, outline reasons based on refractive index.
(f) Record your results as labelled diagrams.
2. (a) Using a light probe and data logger, compare the intensity of the light coming out of the optic fibre with the light entering.
(b) Using the inverse square law, compare the light intensity coming through the probe, as measured in part (b), point 1 above, with that expected for light travelling the same distance through air.

Analysis

Relate your observation to the operation of an endoscope.

CHAPTER 20

RADIOACTIVITY AS A DIAGNOSTIC TOOL

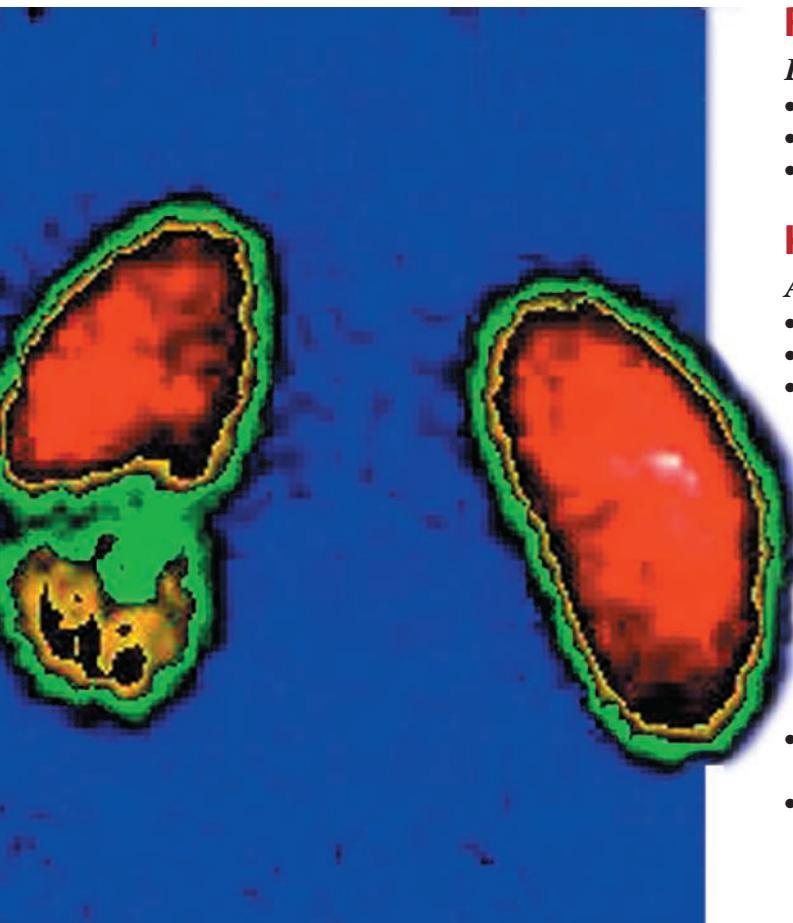


Figure 20.1 This image, taken using a radioactive tracer, shows cancer of the lower section of the left kidney

Remember

Before beginning this chapter you should be able to:

- recall the structure of the atom
- recall the nature of alpha, beta and gamma radiation
- recall the law of conservation of momentum.

Key content

At the end of this chapter you should be able to:

- outline properties of radioactive isotopes
- sketch in general terms what is meant by half-life
- recognise from a list and name radioisotopes used to obtain scans of organs
 - describe how radioisotopes are metabolised by the body so that they are found in the target organ
 - compare a bone scan with an X-ray
 - identify that positron emission occurs during the decay of certain radioactive nuclei
 - discuss the interaction of electrons and positrons to produce gamma rays in the context of positron emission tomography (PET)
- describe how positron emission tomography (PET) technique is used for medical diagnosis
- compare the scan of at least one healthy body organ with its diseased counterpart.

In this chapter the properties of radioisotopes will be applied to medical diagnosis. Cancer is often treated using ‘radiotherapy’. The idea of using a radioactive material to kill cancerous cells is well known. Less well known is the use of a radioactive material inside the body to diagnose disease. Use of radioactive material in the body may seem very risky because of the danger associated with radioactivity. In fact the use of radioisotopes and more recently PET to image organs and study their function has become a very common, effective and safe means of diagnosis. The image in figure 20.1 (page 381), taken using a radioactive tracer, shows a cancer in the patient’s left kidney.

20.1 RADIOACTIVITY AND THE USE OF RADIOISOTOPES

Diagnostic nuclear medicine is essentially a functional imaging technique involving the imaging of physiological processes. This is to be contrasted with imaging using X-rays, where only structural information is obtained.

For the purposes of medical diagnosis, radioactive substances may be introduced into the body and used to target areas of interest. The radiation produced is measured and used to determine the health of the organ or section of the body under investigation.

Properties of radioisotopes

Each element has a particular number of protons in the nucleus. However, the number of neutrons in each element may vary. The atoms of the same element with different numbers of neutrons are **isotopes** of the element.

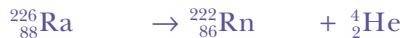
For example, hydrogen (H) has isotopes ^1H , ^2H and ^3H . All these isotopes have the same chemical properties because they have one proton in their nucleus and one orbiting electron. ^1H has no neutrons, ^2H has one neutron and ^3H has two.

All elements have more than one isotope; some occur naturally and some may be made artificially. Sometimes the isotope is unstable and is then known as a **radioactive isotope** or **radioisotope**. ^3H is the only radioactive isotope of hydrogen. Radioisotopes are unstable and will **decay** by emitting particles from their nucleus. Sometimes a new radioactive substance is produced and the decay continues until a stable isotope forms.

Types of radiation: alpha (α), beta (β) or gamma (γ)

In your studies of ‘The cosmic engine’ in the Preliminary Course, you learnt about emissions from radioactive material. The radiation from a radioactive isotope may consist of alpha (α) particles, beta (β) particles or gamma (γ) radiation.

Alpha particles consist of two protons and two neutrons held together. Alpha particles are helium nuclei (^4_2He). Their emission results in the mass number of the radioisotope decreasing by 4 and the atomic number decreasing by 2. The following is an example of decay by the emission of alpha particles from the parent nucleus radium-226:



Radon-222 is known as the daughter nucleus.

Beta particles are electrons. They are formed in the nucleus when a neutron changes to a proton and an electron. The emitted electron is called a beta particle. The following is an example of beta particle decay:



In beta decay, the atomic number of the new element increases by 1 and the mass number remains the same.

Gamma radiation is electromagnetic radiation of very short wavelength and very high energy. This means that it can readily penetrate the body. Gamma radiation is frequently produced in conjunction with alpha and beta particles. The decay of oxygen-19 in the example above produces gamma radiation as well as beta particles.

Usually the gamma radiation is emitted less than a microsecond after the emission of alpha or beta particles, but sometimes there is a delay if the daughter nucleus is left in an excited state, known as a **metastable** state. This is the case with technetium-99m, a very important radioisotope used in medical diagnosis as a gamma radiation emitter. (The 'm' in '99m' means this isotope is metastable.)

The characteristics of α , β and γ radiation are summarised in table 20.1.

A **metastable** nucleus is in an excited state for a period of time before decaying.

Table 20.1 Characteristics of α , β and γ radiation

	SYMBOL	REST MASS (kg)	RELATIVE CHARGE	NATURE	PENETRATION
Alpha	α	6.6×10^{-27}	+2	helium nucleus	Stopped by a sheet of paper or 7 cm of air
Beta	β	9.0×10^{-31}	-1	electron	Stopped in a few mm of tissue
Gamma	γ	0	0	electromagnetic radiation of wavelength shorter than an X-ray	Absorbed in many cm of tissue

Half-life of a radioactive isotope

Half-life is the time taken for half the radioactive material in a sample to decay.

Not all radioactive isotopes decay at the same rate. The rate of decay is measured by the **half-life** of the radioisotope. This is the time taken for half the radioactive material to decay.

Half-lives can vary significantly. Uranium-238 has a half-life of 4.5×10^9 years whereas polonium-218 has a half-life of only 1.5×10^{-4} seconds.

A radioisotope with a very long half-life is unsuitable for medical diagnosis as it lingers in the patient after all necessary measurements are taken. This can pose a danger to the patient and people in close contact because of the radiation emitted. On the other hand, if the half-life is too short, the radioisotope either loses its useful radiation before measurements can be taken or has to be administered in a dangerously large dose. Radioisotopes with half-lives ranging from several minutes to days are used for medical diagnosis.

The decay of a radioisotope can be plotted on a graph from which the half-life can be read. In the graph in figure 20.2, we see that there is initially 100 g of sodium-24. From the graph, the mass has halved to 50 g after 15 hours. The half-life ($T_{1/2}$) of sodium-24 is therefore 15 hours.

SAMPLE PROBLEM 20.1a**Radioactive decay of iodine-123**

A 20 mg sample of iodine-123 is to be used as a radioactive tracer in the body. The half-life of the iodine-123 is 13 hours.

- How long will it take for 17.5 mg to decay?
- Calculate how much iodine-123 will remain after 26 hours.

SOLUTION

(a) In 1 half-life, 10 mg of iodine-123 will decay. This will leave 10 mg iodine-123.

In the second half-life, 5 mg iodine-123 will decay, leaving 5 mg iodine-123.

In the third half-life, 2.5 mg iodine-123 will decay.

Altogether, 17.5 mg ($10 + 5 + 2.5$ mg) iodine will have decayed in 3 half-lives or 39 hours.

- 26 hours is 2 half-lives (2×13 hours).

After 1 half-life, 10 mg of iodine-123 will decay leaving 10 mg iodine-123.

After 2 half-lives, 5 mg iodine-123 will decay leaving 5 mg iodine-123.

5 mg iodine-123 will remain after 26 hours.

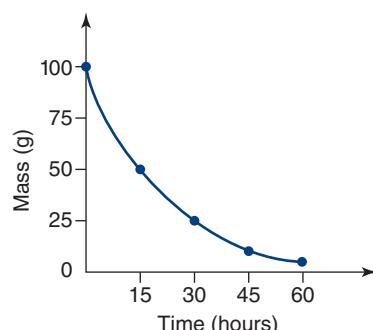


Figure 20.2 The radioactive decay of sodium-24

SAMPLE PROBLEM 20.1b**Radioactive decay**

A radioisotope sample has a half-life of 10.0 minutes.

- Calculate the time it will take the activity to drop from 8.0 MBq (mega becquerels) to 4.0 MBq.
- Calculate the time it will take for its activity to be 1.0 MBq.

SOLUTION

(a) When half the sample has decayed the activity will also halve. This assumes that the atoms formed are not radioactive. Hence the time needed to reduce the activity to 4.0 MBq is one half-life, or 10.0 minutes.

(b) Halving the activity each half-life means 3 half-lives have passed before the activity is 1.0 MBq. The time taken is 30.0 minutes.

PHYSICS FACT

Determining the size of the dose of radioisotope for medical diagnosis

When a radioisotope is introduced into the body, other factors in addition to the half-life of the radioisotope need to be considered. The radioisotope is removed from the patient's body by processes such as respiration, urination and defecation. Also, some patients metabolise the chemical to which the radioisotope is attached more quickly than others, so it is important that the characteristics of the particular patient are considered when dosages are being determined.

The half-life must be long enough to allow useful readings to be taken after the radioisotope

has been taken up by the targeted organ. For example, calculations show that, although iodine-123 has a half-life of 60 days, half the administered radioisotope will be removed from an average patient's body in 16 days.

Generally, if the radioisotope remains in the patient's body for a long period of time, the half-life of the radioisotope should be comparable to the time taken to carry out the investigation, to minimise the dose to the patient. When the radioisotope is excreted in about the same time as is needed for the investigation, a longer half-life radioisotope can be safely used.

PHYSICS IN FOCUS

Producing radioisotopes: the medical cyclotron

The effectiveness of nuclear medicine for diagnosis of disease has depended on the ability to:

- produce radioisotopes
- detect the gamma radiation produced.

The production of radioisotopes became possible with the development of the cyclotron by E. O. Lawrence in 1931. From the mid 1940s, a range of radioisotopes was available from the United States and later from the United Kingdom.

Most of the radioisotopes used in nuclear medicine in Australia are made by ANSTO at two major facilities: the OPAL nuclear research reactor, at Lucas Heights in the south of Sydney, and the National Medical Cyclotron, at Royal Prince Alfred Hospital in Sydney. Some radioisotopes are imported from South Africa and Canada. Besides the ANSTO facilities, there are also hospital-based cyclotrons in Melbourne, Brisbane and Perth. Cyclotrons are needed to make radioisotopes for positron emission tomography (PET), a diagnostic technique discussed later in this chapter.

In late 2007 ANSTO announced a partnership with Siemens Medical Solutions to build two new state-of-the-art cyclotrons. These will supply hospitals with more of the isotopes needed for PET, increasing the availability of PET scanning.



Figure 20.3 The National Medical Cyclotron, located near Royal Prince Alfred Hospital in Sydney

Metabolising of radioactive isotopes by the body

Radioisotopes that emit alpha particles are not used in the diagnosis of disease because the alpha particles cause damaging ionisation inside the body.

Beta particles travel further than alpha particles before they are absorbed but their ionisation damage is much less. They are used in therapy but not in diagnosis of disease.

The most useful radioisotopes for nuclear medicine are those that emit gamma radiation only. Technetium-99m and iodine-131 are two such isotopes. A gamma-emitting radioisotope inside the body can be detected outside the body because gamma radiation is very penetrating.

Common radioisotopes used in medical diagnosis are listed in table 20.2 on the following page.

The radioisotope is chosen on the basis of its ability to target the organ to be studied. First, the radioisotope needs to be chemically attached to a compound that would normally be metabolised by the organ of interest. When this compound is chemically attached ('labelled') with the radioisotope, it is called a **radiopharmaceutical**. For example, glucose is a compound that is readily absorbed by the brain. Hence glucose is labelled to become a radiopharmaceutical for imaging brain function.

A **radiopharmaceutical** is a compound that has been labelled with a radioisotope.

Table 20.2 Radioisotopes used in medical diagnosis

Radioisotope	Production site	Half-life	Function
Chromium-51	Nuclear reactor	27.70 days	Used to label red blood cells and measure gastro-intestinal protein loss.
Gallium-67	Cyclotron	3.26 days	Used to detect tumours and infections.
Molybdenum-99	Nuclear reactor	65.94 hours	Used as the 'parent' in a generator to produce technetium-99m, which is the most widely used isotope in nuclear medicine.
Technetium-99m	'Milked' from molybdenum-99	6 hours	Used to investigate bone metabolism and locate bone disease; assess thyroid function; study liver disease and disorders of its blood supply; monitor cardiac output, blood volume and circulation clots; monitor blood flow in lungs; assess blood and urine flow in kidneys and bladder; investigate brain blood flow and function; estimate total body plasma and blood count.
Iodine-123	Cyclotron	13 hours	Used to monitor thyroid function, evaluate thyroid gland size and detect dysfunction of the adrenal gland. Also used to assess stroke damage.
Iodine-131	Nuclear reactor	8 days	Used to diagnose and treat various diseases associated with the thyroid gland. Used in the diagnosis of the adrenal medullary. Used for imaging some endocrine tumours.
Thallium-201	Cyclotron	3.05 days	Used to detect the location of damaged heart muscles.

PHYSICS IN FOCUS

Radioisotopes emitting gamma radiation

Both iodine-123 and technetium-99m are valuable radioisotopes because they decay by the emission of gamma radiation only.

Iodine-123 is more expensive than iodine-131, an emitter of beta and gamma radiation. Iodine-131 has been used in the investigation of the thyroid gland. However, it emits beta radiation, leading to larger radiation doses than desirable. Also, the energy of the gamma radiation produced from iodine-131 is very high, resulting in poor image quality when detected by the gamma camera. (The operation of a gamma camera is discussed later in this chapter, page 388). The half-life of 8 days for iodine-131 is relatively long, resulting in exposure of the patient to radiation well after the testing has been carried out. By contrast, iodine-123 has a half-life of 13 hours, also concentrates in the thyroid gland and emits gamma rays of energy that can be detected clearly by the gamma camera.

Technetium-99m has a half-life of only 6 hours so it must be produced in the hospital where it is to be used. A purpose-built generator system is used to obtain the technetium-99m when needed. The generator contains the 'parent' isotope, molybdenum-99, which decays to the metastable 'daughter' isotope technetium-99m. The technetium-99m is flushed from the molybdenum using a saline solution. The flushing is called elution. The molybdenum remains in the generator as it is chemically attached to a central column. The technetium-99m is said to be 'milked' from the molybdenum. This operation usually happens daily, allowing the technetium sufficient time to build up. Because the molybdenum has a half-life of approximately 66 hours it must be replaced weekly as by that time the rate of production of technetium is too low to be of value.

Technetium-99m has the added advantage that it readily attaches to different compounds to form radioactive tracers.

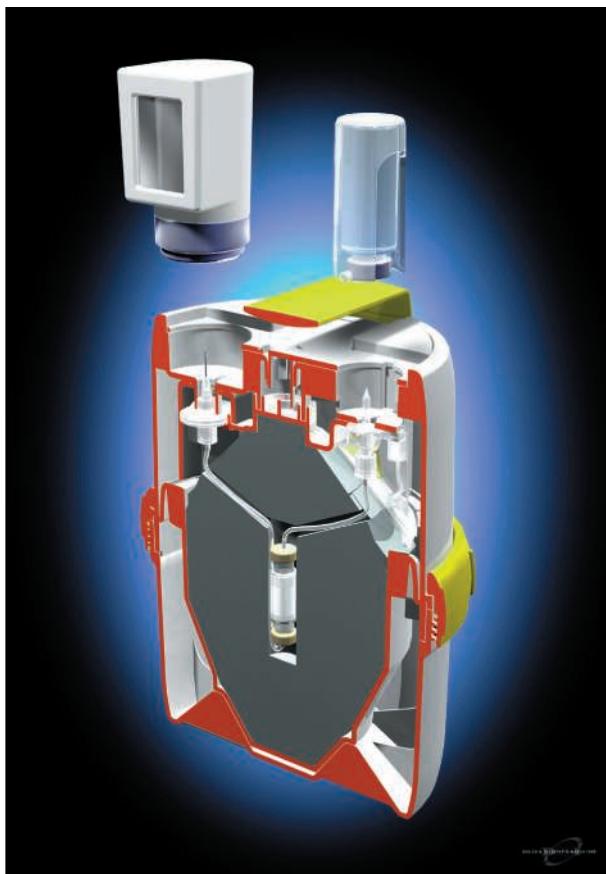


Figure 20.4 A cross-section through a typical technetium generator used in hospitals to generate technetium-99m

Figure 20.4 shows a cross-section through a technetium generator and figure 20.5 shows a technician preparing radiopharmaceuticals for tracer studies in a hospital nuclear medicine department.



Figure 20.5 The preparation of labelled compounds for tracer studies

Using radioactive isotopes to target body organs

As indicated in table 20.2 (page 386), particular radioisotopes are chosen to target particular organs. The radiopharmaceutical is injected into the bloodstream, inhaled or taken orally, and its passage through the body is traced by measuring the radiation it emits.

Sometimes an image is taken after a period that may be up to several hours. The radioisotope has accumulated in the target organ, so this image measures the amount of radiation emitted from different organs and shows where the radioisotope has accumulated.

In other situations, a series of images is taken over a period of time starting from when the radioisotope is first introduced. This type of investigation shows the distribution of the radioisotope and rate of absorption or excretion by various organs. The images may be taken over a few minutes for a heart or lung study or over a period up to half an hour for a kidney or bladder investigation.

In analysing the images, radiologists identify ‘hot spots’ with a higher than normal concentration of radioisotope and ‘cold spots’ showing a lack of radioisotope. These areas often indicate disease (see figure 20.6).

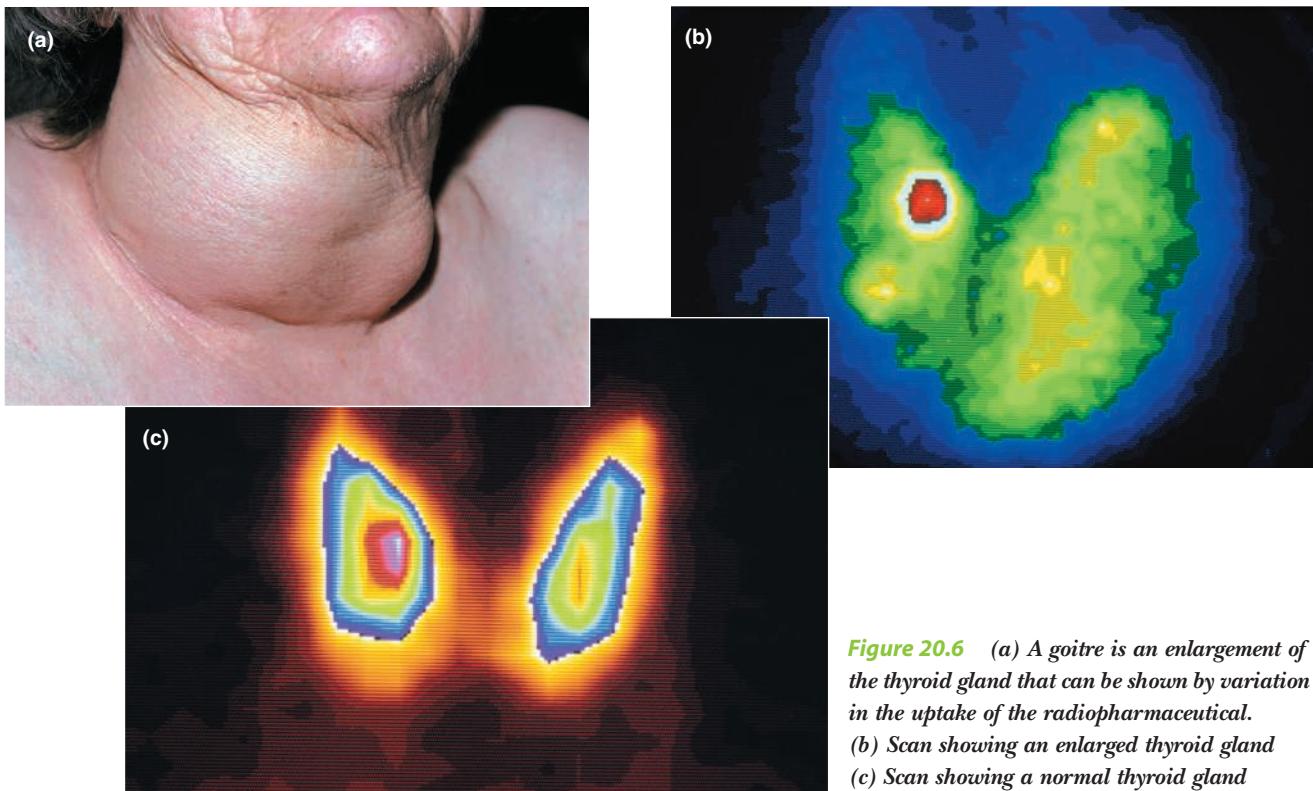


Figure 20.6 (a) A goitre is an enlargement of the thyroid gland that can be shown by variation in the uptake of the radiopharmaceutical.
 (b) Scan showing an enlarged thyroid gland
 (c) Scan showing a normal thyroid gland

Obtaining an image

The image is obtained by measuring the amount of gamma radiation coming out of the patient’s body using a gamma camera. The gamma camera is a stationary imaging system that collects gamma radiation over a large area. It converts the gamma rays into light flashes (scintillations) which are transformed into amplified electrical signals. These are then analysed and processed to form an image.

A gamma camera is shown in figure 20.7. The three main sections are the collimator, the sodium iodide crystal and the phototubes. Gamma rays travelling at right angles to the sodium iodide crystal enter the camera through a lead collimator. Usually the collimator is a circular slab of lead with many holes perpendicular to the face (see figure 20.7(c)). Gamma rays striking the crystal from other angles would degrade the image and so are blocked by the lead collimator.

The radiation detector is a single sodium iodide crystal 30 to 40 cm in diameter and 1.2 cm thick. An array of photomultiplier tubes is arranged in a hexagonal pattern at the rear of the detector.

When a gamma ray enters the sodium iodide crystal, the light from the resulting scintillation spreads through the crystal and each photomultiplier tube receives some of the total light. The fraction of the total light seen by each tube depends on how close that tube is to the original point of entry of the gamma ray. The resulting electrical pulses from each photomultiplier tube are decoded and converted to signals to be displayed on a computer screen. The image showing the gamma ray output from the organ is constructed from all the gamma rays detected.

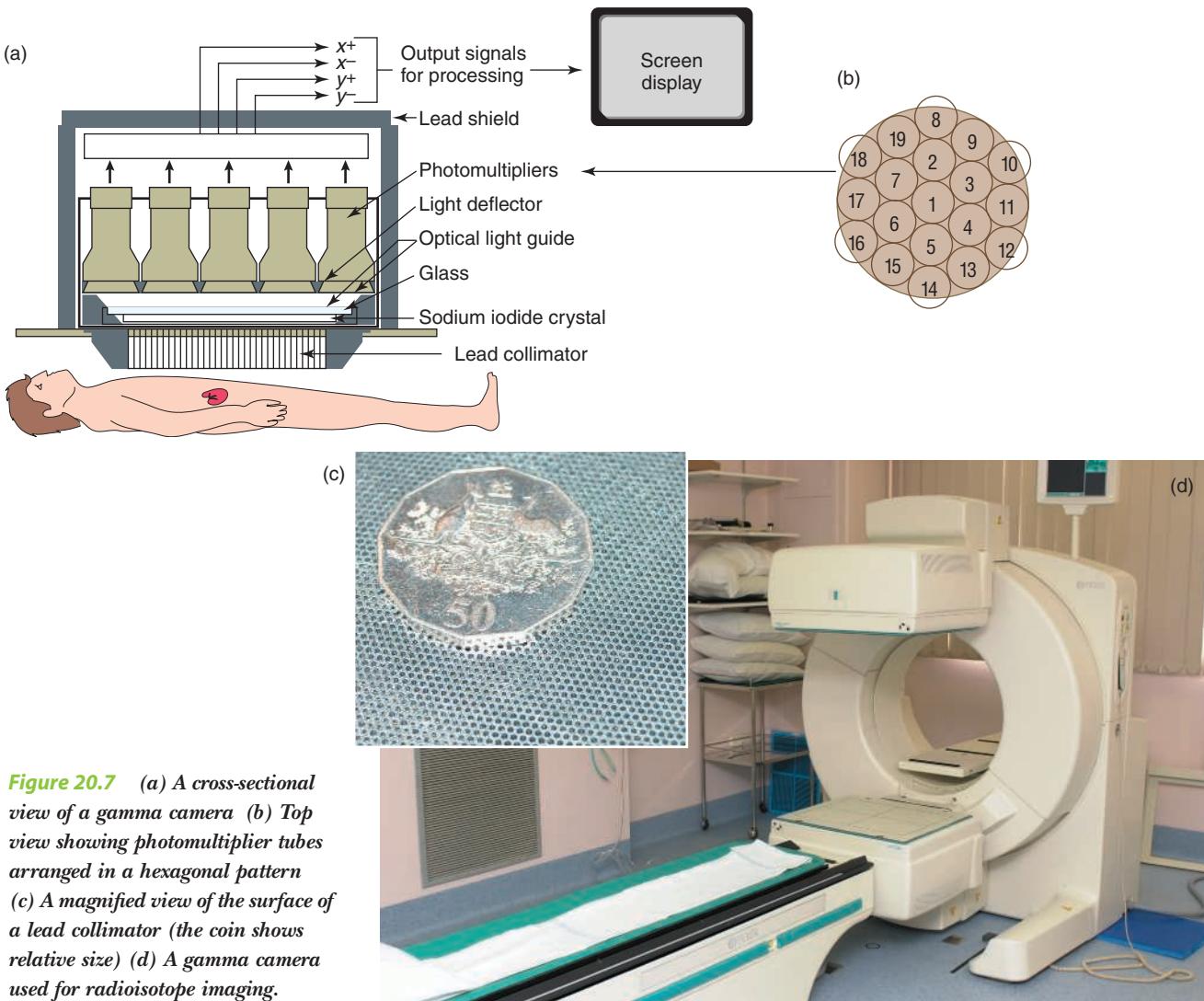


Figure 20.7 (a) A cross-sectional view of a gamma camera (b) Top view showing photomultiplier tubes arranged in a hexagonal pattern (c) A magnified view of the surface of a lead collimator (the coin shows relative size) (d) A gamma camera used for radioisotope imaging.

Medical applications

Thyroid investigations

The thyroid gland metabolises iodine. A drink of a dilute solution of sodium iodide tagged with iodine-123 is administered and its accumulation

is measured from 10 minutes to 48 hours after being administered. An image of the goitre may be obtained as in figure 20.6 (page 388), or the uptake of the isotope may be graphed and compared with a standard as in figure 20.8.

Thyroid investigations now commonly use technetium-99m, which is also taken up by the thyroid but is more readily released than iodine.

The heart

Human serum albumen is labelled with technetium-99m and injected into the patient to measure the efficiency of the heart as a pump. The passage of the radiopharmaceutical is monitored through the heart chambers. Thallium-201 as part of thallium chloride is injected and monitored to assess damage caused by a stroke or to measure the effect on the heart of exercise or drugs (see figure 20.9).

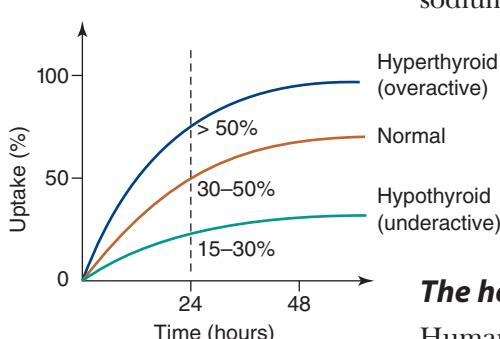
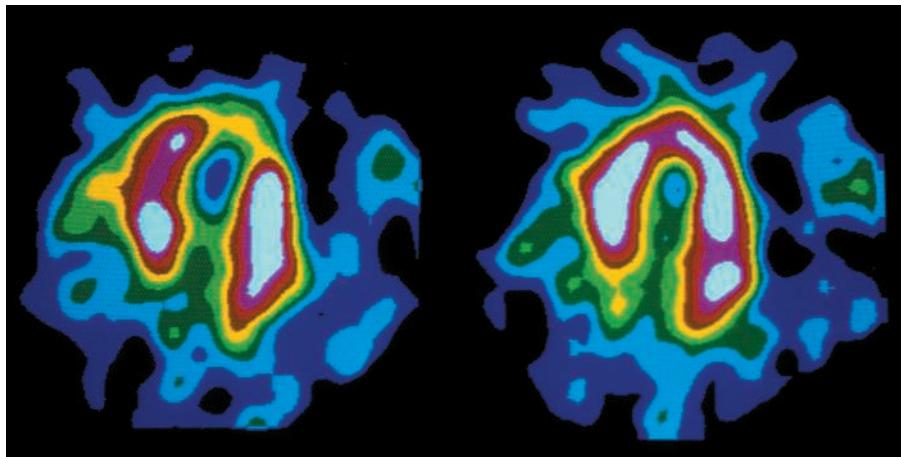


Figure 20.8 Uptake of iodine-123 by the thyroid gland

Figure 20.9 Performance of heart muscle using thallium-201

A series of images produces ‘slices’ through a chamber in the heart. The left image shows a ‘hole’ in the muscle during exercise, probably indicating a blockage to that part of the muscle. A ‘hole’ is seen when there is no gamma radiation from that area, and shows up as no blue or red. The right image, taken during resting of the patient, shows the muscle is alive because the ‘hole’ has disappeared, indicating blood is flowing. The blockage can possibly be cleared, leading to recovery.



Bones, lungs and brain

Technetium-99m is used in imaging the bones, lungs and brain.

Polyphosphate ions are labelled with technetium-99m and injected, accumulating in bone within an hour. The image shows the function of the bone. Areas of increased blood flow show up as ‘hot spots’. Such areas are frequently associated with disease. Bone imaging often shows up bone tumours and stress fractures earlier than standard X-rays, which only show the structure of the skeleton (see figure 20.10).

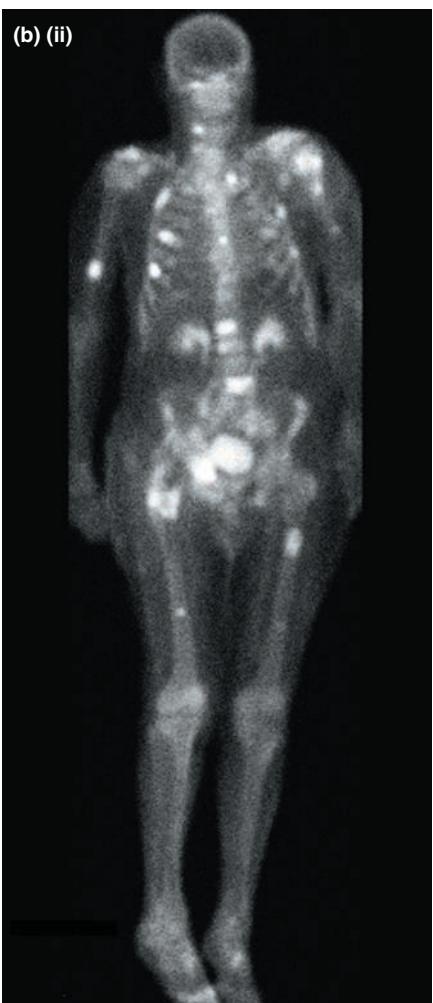


Figure 20.10 (a) An X-ray of a broken leg (b) A bone scan showing (i) a healthy skeleton and (ii) a skeleton with tumours. (Note the white spot on the right arm in each bone scan shows where the isotope was injected.)

Brain studies using technetium-99m as a tracer measure blood flow through the brain, allowing dementia and stroke damage to be identified (see figure 20.11).

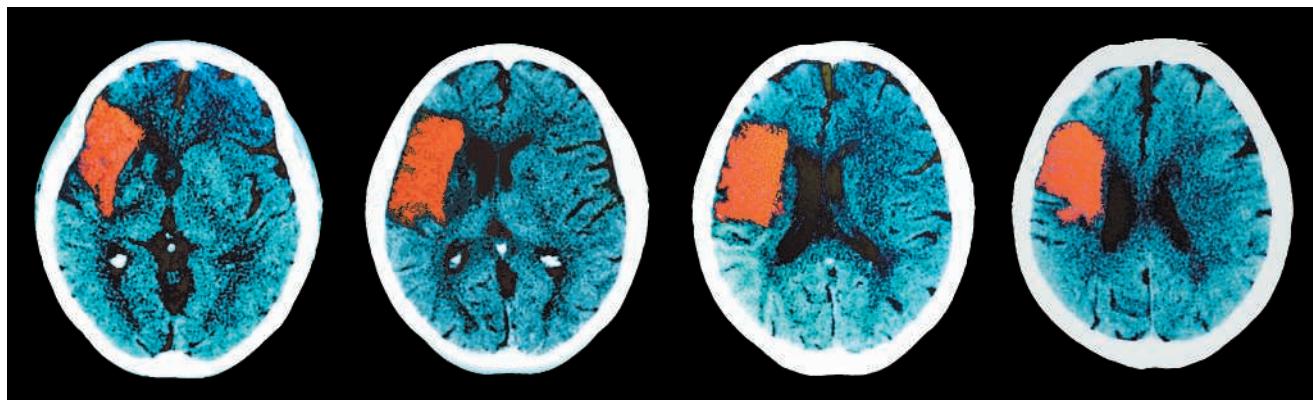


Figure 20.11 Images of ‘slices’ through the brain show areas of reduced activity due to stroke damage (shown in red).

To study the lungs, technetium-99m attached to albumen is coagulated, mixed with saline and injected into the veins in the arm. It becomes trapped in the fine capillaries in the lung and allows a map to be made of the functioning capillaries. Any blockage in the lung, perhaps due to a clot, shows as a region without any radioactive tracer. This blood flow study is called a perfusion study.

To enable the health of the airways to be studied, the patient inhales an aerosol labelled with technetium-99m. This ventilation study shows, over about half an hour, ‘cold spots’ where the radioisotope has not accumulated because the airway is blocked (see figure 20.12).

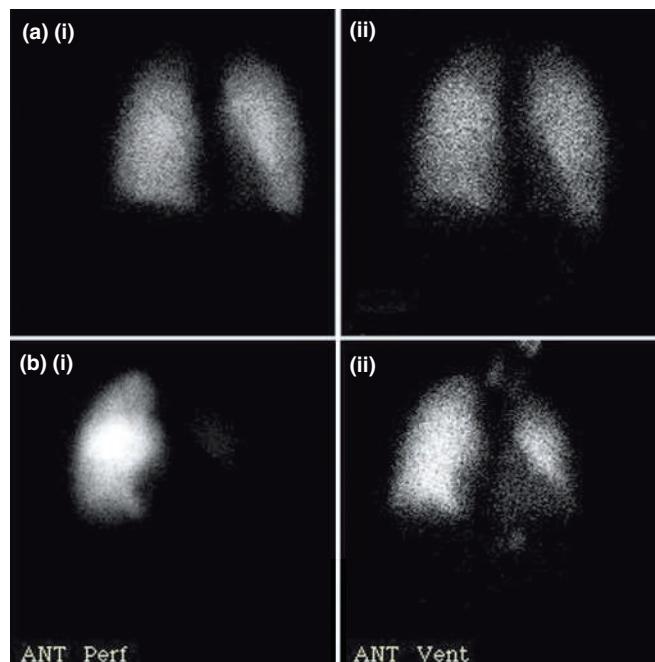


Figure 20.12 Lung studies
(a) A normal perfusion study and ventilation study of the lungs
(b) Front view of lung scans of a patient with a blockage in the left pulmonary artery: (i) the perfusion scan shows no blood flow to the left lung; (ii) the ventilation scan shows both lungs as the airway is not blocked.

To determine the volume of blood in the body, a measured quantity of a radioisotope is administered, and after a period of time a sample of blood is taken. If the activity of the tracer in the blood is measured, the dilution of the tracer and hence the volume of blood in the body can be calculated. This procedure, known as dilution analysis, is valuable in investigating disorders such as anaemia, assessing stroke damage and monitoring blood loss as a result of an accident.

PHYSICS FACT

Radioactivity and safety issues

In a hospital, the general public, medical teams and patients must be protected from overexposure to radioactive material. Strict guidelines are implemented to control and monitor exposure to radiation.

Areas where work is carried out with ionising radiation are clearly marked as controlled areas with limited access. Equipment is checked regularly to make sure it does not leak radioactive material. Personnel distance themselves from radioactive material where possible and wear monitors to measure their exposure to radioactive sources. These monitors are checked regularly.



Figure 20.13 A radiation warning sign. The trefoil is the internationally recognised symbol for radiation.

20.2 POSITRON EMISSION TOMOGRAPHY (PET)

Positron emission tomography, known as PET, is used to diagnose and monitor brain disorders, investigate heart and lung functioning and detect the location and spread of tumours. Using particular radio-pharmaceuticals, a cross-sectional image through an organ can be obtained or a region of the body can be imaged, allowing the function of an area to be determined. A PET image of the brain shows the patient's responses to factors such as noise, illumination and changes in mental concentration.

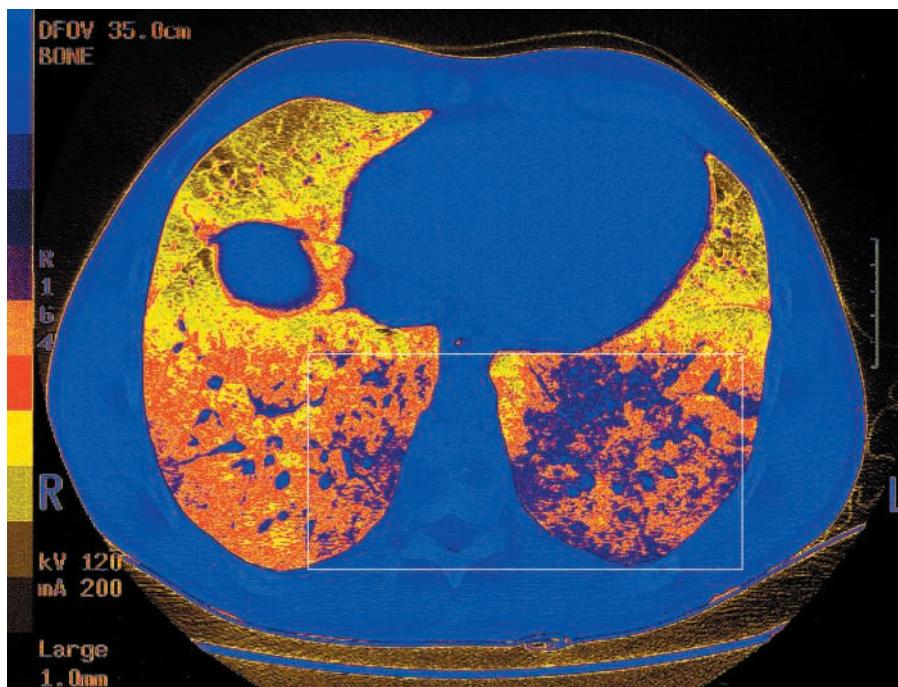


Figure 20.14 A PET scan showing a build-up of fluid in the lungs (pulmonary edema)

Positron-electron interactions

Positrons are positively charged beta particles formed when a proton disintegrates to form a neutron and a positron. A positron is identical to an electron except that its charge is positive instead of negative.

MeV stands for million electron volts and is the energy gained by one electron accelerating through a potential difference of one million volts.

Certain radioisotopes decay by the emission of **positrons**. Positrons are positively charged beta particles. That is, they are identical to electrons except that they have positive charge instead of negative charge. Positrons are formed when a proton disintegrates to form a neutron and a positron. Radioisotopes that are deficient in neutrons often decay in this way. For example, carbon-11 decays to boron-11, emitting a positron (β^+) as shown below.



When a positron meets an electron they ‘annihilate’ each other, converting their combined energy and mass into two gamma rays (γ rays). The energy of each of these gamma rays is 0.51 **MeV** (million electron volts) and they travel in opposite directions, as momentum is conserved in the interaction. This process is sometimes called ‘pair annihilation’.

How a PET scan is obtained

To obtain a PET scan, a suitable pharmaceutical is labelled with a positron-emitting radioisotope. The radiopharmaceutical is usually injected into the patient but sometimes the chemical is inhaled. After a short period of time the radiopharmaceutical has accumulated in particular areas of the body and begun to decay by the emission of positrons. These positrons travel a short distance, of the order of a few millimetres, before they encounter electrons in the body. Pair annihilation takes place and two gamma ray photons are produced. The gamma photons travel in opposite directions from the site of annihilation and emerge from the body where they are detected by gamma cameras.

Modified gamma cameras surround the patient in the section being scanned. The cameras do not have multihole collimators so that gamma photons from all angles can be detected. Pairs of gamma photons travelling in opposite directions are detected and their relative intensity measured. By taking measurements of pairs of gamma photons from all angles and correlating these measurements with known attenuation coefficients for gamma rays passing through tissue, the position of the decaying radioisotope can be approximately determined. In this way, an image is produced showing where radioisotopes accumulate. It takes about half a million gamma ray pairs to produce a useful image.

A PET imaging system detecting emissions from a region of the brain is illustrated in figure 20.15.

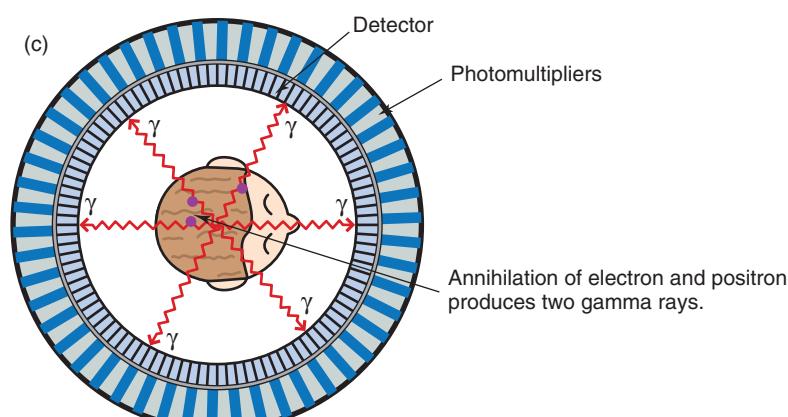
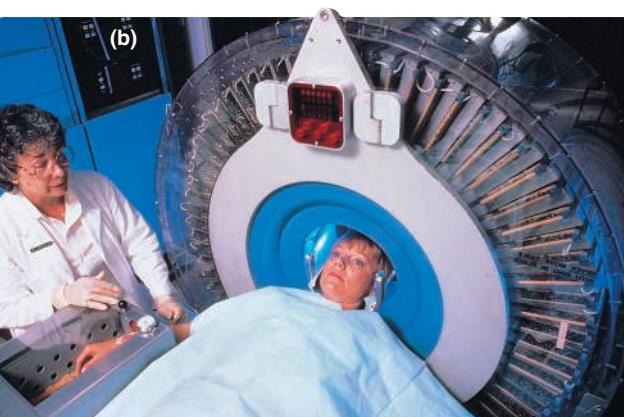
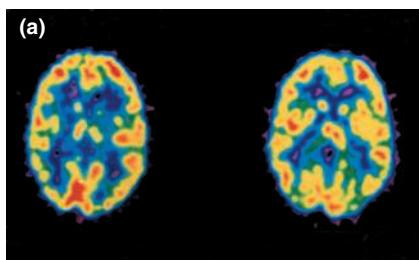


Figure 20.15 (a) Some PET images of brain activity (b) A patient undergoing a PET scan of their brain (c) Cross-section showing pairs of gamma rays travelling in opposite directions and reaching detectors

Isotopes used in PET

Common isotopes used in PET are listed in table 20.3.

Table 20.3 Common isotopes used in PET

RADIOISOTOPE	SYMBOL	HALF-LIFE
Carbon-11	$^{11}_6\text{C}$	20.4 minutes
Nitrogen-13	$^{13}_7\text{N}$	10.0 minutes
Oxygen-15	$^{15}_8\text{O}$	2.13 minutes
Fluorine-18	$^{18}_9\text{F}$	109.8 minutes

To image brain function, the patient may inhale a small quantity of carbon monoxide (CO) made with carbon-11 because CO quickly accumulates in the brain. Oxygen-15 can be used to label oxygen to study blood volume in the brain or to label water to study blood flow in the brain. Fluorine-18 or carbon-11 are used in research on the brain related to Parkinson's disease and schizophrenia.

In all these studies it is important that the amount of radiopharmaceutical administered is large enough to obtain a good image in the time needed for the procedure but small enough to minimise patient exposure to radiation.

As can be seen from table 20.3, the half-lives of isotopes suitable for PET are very small. The isotopes must be created on the day of use and, except for fluorine-18, must be made at the site of use. A cyclotron is needed for their production (see the box on page 385). This poses a serious limitation, as most hospitals cannot afford the cost of an on-site cyclotron and facility for producing radiopharmaceuticals. In New South Wales, the cyclotron near the Royal Prince Alfred Hospital is used to produce isotopes for the PET facilities at that hospital (see figure 20.3 on page 385). Fortunately the longer half-life of fluorine-18 means tracers can be labelled with fluorine-18 and shipped to nearby hospitals.

20.3 IMAGING METHODS WORKING TOGETHER

Medical imaging to obtain both functional and structural images is often needed for adequate diagnosis. CT scans are used to obtain structural images. (MRI, which is discussed in chapter 21, is also generally used for structural images, although it can be used to obtain functional images in certain parts of the body.) Radioisotopes, on the other hand, allow functional information to be gathered from which the site of a tumour can be determined. A nuclear medicine image may show tumours but not very much normal tissue. Hence it may be difficult to determine the position of the tumour relative to other structures. If a CT or MRI scan is obtained at the same time, the location of the tumour can be established precisely. In fact, scanning devices combining CT with PET are available already, and small-scale systems combining MRI with PET have been developed for use in veterinary hospitals.

PHYSICS FACT

PET has found important applications in the studies of brain function and metabolism. As early as 1902, P. Ehrlich suggested that there was a barrier that prevented many molecules in the bloodstream from gaining access to the brain. It was called the blood-brain barrier (BBB). This model suggested that some molecules could penetrate the barrier and some could not. Obviously it was important that certain molecules cross this barrier for healthy brain function. β -D-glucose is one such molecule. It is often referred to as the brain's 'fuel'.

β -D-glucose may be suitably labelled so that PET scans of the brain's function can be made. Fluorine-18 is commonly used for this labelling. Fluorine-18 replaces a hydrogen atom on glucose to produce 2-fluoro-2-deoxy-D-glucose (FDG). Healthy brain function correlates with regions where glucose is found in known concentrations. However, tumours require more oxygen and therefore more glucose. As a result, tumours show up in a PET scan due to the accumulation of glucose in the tumour.

A compound containing technetium-99m is also of the correct size to cross the BBB. Technetium-99m is used to image the brain in a process called single photon emission computed tomography (SPECT). SPECT also enables a cross-sectional image of the brain to be made by rotating the gamma camera around the head, but the position of the decaying radioisotope is not as clearly identified as with PET.

Evaluating imaging using radioisotopes

The advantages and disadvantages of using radioisotopes as a diagnostic tool in medicine can be summarised as follows.

Advantages:

- Body function can be assessed through examining particular organs and flow rates of blood and water.
- Stress fractures can be identified early by detecting increased activity of bone cells.
- Volume of blood and water can be measured using dilution analysis.
- Whole body scanning will allow disease of the skeleton to be assessed.

Disadvantages:

- Resolution of the images is poor compared with other methods of imaging.
- Some radiation risk exists due to use of radioisotopes.

- The radioisotope is usually injected, so this is an invasive procedure.
- Radioactive waste needs to be disposed of with care.
- Costs are greater than for ultrasound or X-rays (and are similar to MRI).

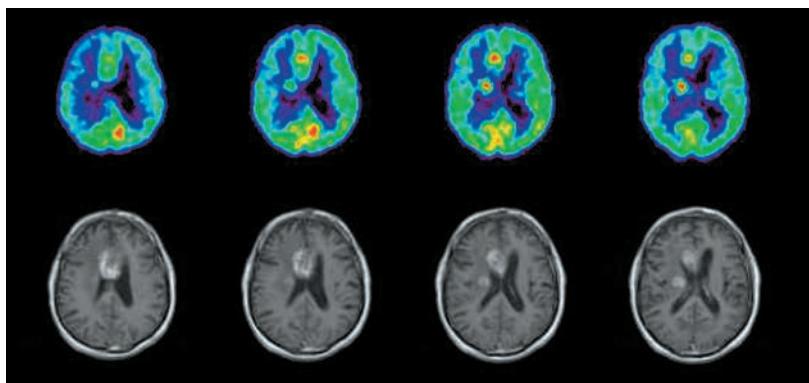


Figure 20.16 Comparison of PET (top) and MRI scans (bottom) of a patient with a brain tumour

SUMMARY

- Radiopharmaceuticals are taken up by particular organs in the body.
- Gamma radiation from the radioisotope is detected and used to make an image of the organ.
- The rate at which the radioisotope accumulates in the target organ indicates the health of the organ.
- The half-life of the radioisotope and length of time needed for the procedure must be considered when choosing an appropriate radioisotope.
- PET uses radioisotopes that are positron emitters.
- Positrons and electrons annihilate in the body, producing two gamma rays.
- Detecting the position where the gamma rays originate enables the position of the positron-emitter to be mapped.
- PET scans indicate the biochemistry, metabolism and function of a particular area.
- PET scans are used for studying the brain and heart, detecting cancers at an early stage and monitoring cancers during treatment.

QUESTIONS

1. Define the following terms:
 - (a) radioisotope
 - (b) radioactive decay
 - (c) emissions from radioactive nuclei.
2. (a) Using an example, outline what is meant by half-life of a radioisotope.
(b) Using the example given in (a), outline how the exposure to radioactive emissions could be decreased.
3. Using a specific example of a radioisotope, describe how it accumulates in the target organ.
4. A particular isotope has a half-life of 100 days. Discuss the suitability of this isotope for use in medical diagnosis.
5. Describe the problems associated with using a radioisotope of very short half-life for medical diagnosis.
6. A sample of a particular radioisotope has a half-life of 2.0 minutes.
 - (a) Calculate the time it will take the activity to drop from 4.0 MBq (mega becquerels) to 1.0 MBq.

- (b) Calculate the time it will take for its activity to be 0.25 MBq.

7. The function of the lungs can be studied using a radioactive gas. The choices are krypton-81m or xenon-133 and their properties are in the table below.

ISOTOPE	EMISSION PRODUCTS	HALF-LIFE
Krypton-81m	γ	13 seconds
Xenon-133	β, γ	5.3 days

Evaluate the claim that 'Xenon should be used in preference to krypton for investigations of lung function'.

- (a) Choose two specific radioactive isotopes used in medical diagnosis and outline where they would be used in the body. Justify your answer.
(b) Explain why α -emitting radioisotopes are not used for medical imaging.
9. State two factors, other than its emissions, that affect the choice of a radioisotope for a tracer study.
10. Carbon-11 has a half-life of 20 minutes and bromine-75 has a half-life of 100 minutes. If samples of these isotopes initially have the same activity, show on the same graph how their activities vary with time.
11. Identify a radioactive tracer study in which the tracer:
 - (a) mixes with the substance under investigation
 - (b) is accumulated in the organ of interest.
12. Use a flow diagram to outline the steps in obtaining technetium-99m from its parent isotope.
13. Explain why technetium-99m is such an ideal radioisotope for medical imaging.
14. Describe how a radioisotope of your choice is used in a PET investigation. In your answer you should name the isotope and state what radiation is emitted and how it is monitored. You should describe what measurements are made and how they are used to obtain a result. You should also mention any precautions or safety procedures.
15. (a) Describe what is meant by a positron.
(b) Identify how positrons may be obtained.

- (c) Identify issues associated with positron-electron interaction and describe how this interaction is used in medical diagnosis.
16. Examine figure 20.1 (page 381), which shows a study of kidneys. Compare the diseased kidney with the healthy one and outline reasons for the observed differences in the images.
17. Figure 20.12 (page 391) shows two different types of studies of lungs.
(a) Contrast the studies.
(b) Relate the type of study to the disease diagnosed.
18. Figure 20.10 (page 390) shows an X-ray of a leg and a bone scan of the body.
(a) Compare the X-ray image with the bone scan.
(b) Explain why there are differences in the images obtained.
19. Using the internet or other sources:
(a) find a scanned image of at least two healthy body parts or organs (the image should have been obtained using radio-isotopes)
(b) find a scanned image of the diseased counterpart of the body parts or organs
(c) compare the images and outline why the differences are obvious in the images.
(In a search engine, use phrases such as ‘radiopharmaceuticals’, ‘nuclear medicine images’, ‘PET images’, ‘bone scans’ or ‘brain scans’ as key words. Go to a particular hospital web site and search for ‘nuclear medicine department’.)

CHAPTER 21

MAGNETIC RESONANCE IMAGING AS A DIAGNOSTIC TOOL



Figure 21.1 An image of a brain taken using MRI, showing multiple sclerosis plaques (shown as black patches)

Remember

Before beginning this chapter you should be able to:

- recall what is meant by radio frequency electromagnetic radiation
- recall that there is a magnetic field associated with a charge moving in a circle
- recall what is meant by a superconductor.

Key content

At the end of this chapter you should be able to:

- describe how the net spin of protons and neutrons in the nucleus is produced
- explain that nuclei with net spin produce a magnetic field and this influences their response to an external magnetic field
- describe the response of nuclei to a strong magnetic field
- relate frequency of precession to the composition of the nuclei and the strength of the applied magnetic field
- discuss the effect of subjecting precessing nuclei to pulses of radio waves
- explain that the amplitude of the radio signal emitted by the nuclei as they relax increases as the number of nuclei present increases
- contrast the relaxation time between tissue containing hydrogen-bound water molecules and tissue containing other molecules
- compare MRI scans of healthy and damaged tissue
- explain why MRI scans can be used to distinguish between grey and white matter in the brain, to identify areas of high blood flow and to detect cancerous tissue
- identify the function of the following parts of MRI equipment: electromagnet, radio frequency oscillator, radio receiver and computer
- compare advantages and disadvantages of X-ray images, CT scans, PET scans and MRI scans
- assess the impact of medical applications of physics on society.

In this chapter you will learn how the properties of magnetic fields of nuclei are used to produce images for medical diagnosis. Magnetic resonance imaging (MRI) uses strong magnetic fields and the magnetic properties of nuclei in the body to obtain clear images of the brain, spinal cord and soft tissues such as muscle, tendons, cartilage and joints. It produces excellent spatial resolution and hence fine detail between different tissues can be detected in the areas that are imaged.

Since 1977 when the first whole-body magnetic resonance image was produced, MRI has developed at a remarkable pace and is now an indispensable imaging technique. It poses minimum risk to the patient and can provide accurate information about organ function and biochemistry as well as body anatomy.

21.1 THE PATIENT AND THE IMAGE USING MRI



Figure 21.2 This metal chair was not secured in the room with the MRI machine and, due to the strong magnetism, became stuck in the machine.

If you are a patient undergoing an MRI scan, you are placed on a bed that moves on a gantry inside a region where the magnetic field is very strong. Before the MRI machine is turned on, all metal objects must be removed from the patient and from the room because the magnet is so strong that loose magnetic material can become missiles (see figure 21.2). Eddy currents are induced in any nearby metallic material when the machine is operating because there are changing magnetic fields in the MRI machine. If you have a pacemaker or other metallic implant you will not be permitted to undergo MRI.

The following sequence of steps occurs while you are in the strong magnetic field:

1. A pulse of electromagnetic radiation in the radio frequency range is sent into your body.
2. This pulse is turned off.
3. Nuclei in your body produce a radio frequency pulse as a result of the pulse that was sent into your body.
4. The radio waves emitted are analysed, amplified and processed by a computer to form part of the image that is displayed on a screen.
5. The process is repeated rapidly with many radio frequency pulses while the whole area of interest is scanned.

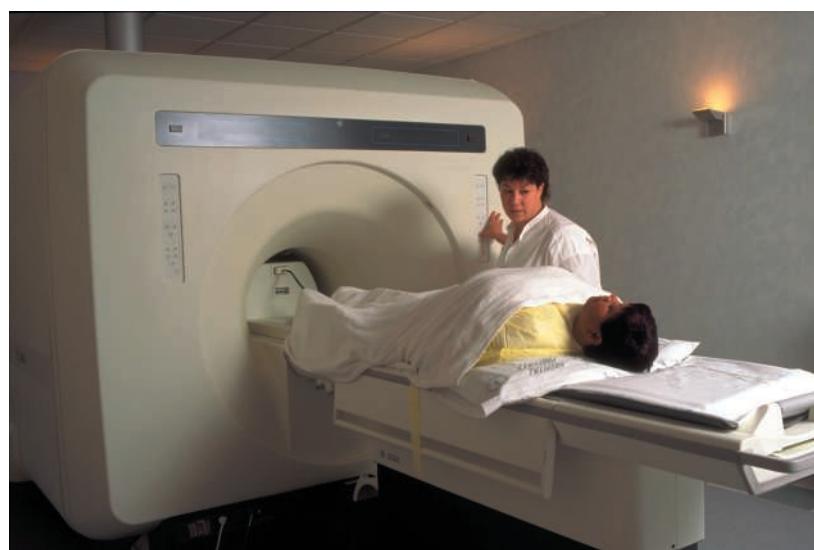


Figure 21.3 A patient having a scan using MRI

To understand MRI it is necessary to understand:

- how magnetic effects in the body arise
- how external magnets and then electromagnetic radiation can interact with the nuclei in the body
- how data are collected from the radio waves produced to create an image.

Magnets in the body

Our bodies are made of a relatively small variety of elements combined into a large variety of compounds. Hydrogen is the most commonly imaged element in MRI because it is abundant in the body and gives the strongest signal when subjected to the MRI process that we will outline. About one-tenth of the mass of our body is hydrogen and of this approximately 70 per cent is in water molecules, 20 per cent is in fats and a small amount is in proteins.

It is possible to use other nuclei to produce an image. These include carbon-13, fluorine-19, sodium-23 and phosphorus-31. However, they are not as abundant in the body as hydrogen and the signal from them is not as strong.

All the nuclei used in magnetic resonance imaging have a **net spin**. Net spin is a concept used in quantum mechanics and is described in more detail in the Physics fact on page 401. If the nucleus of an atom has a net spin, it may behave as a small magnet (see figure 21.4).

For this discussion the nucleus of a hydrogen atom will be considered, because hydrogen is the most commonly imaged element in MRI. The nucleus of a hydrogen atom is a proton and has a net spin. Individual protons behave as small magnets. The magnetic field associated with the protons in hydrogen atoms will be randomly orientated until an external strong magnetic field is applied.

Protons become aligned due to the interaction of their magnetic field with an external magnetic field. They align themselves either in the direction of the external magnetic field (parallel) or in the opposite direction to the external magnetic field (anti-parallel). (The parallel and anti-parallel result follows from quantum mechanics and is beyond the scope of this discussion.) Each of these orientations corresponds to a slightly different energy state and this fact is important in the MRI process. The lower energy state corresponds to parallel alignment and slightly more of the protons are found in this state when an external magnetic field is applied.

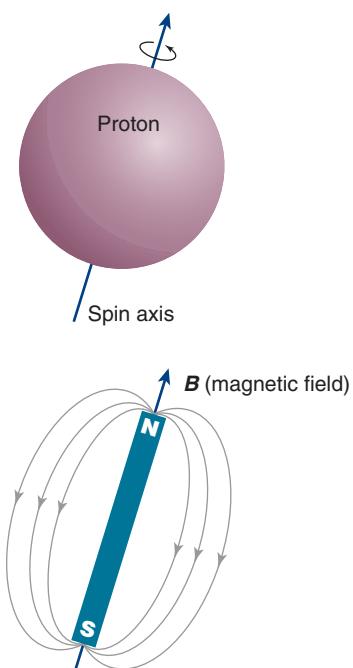


Figure 21.4 The magnetic field associated with a hydrogen proton is like that around a bar magnet.

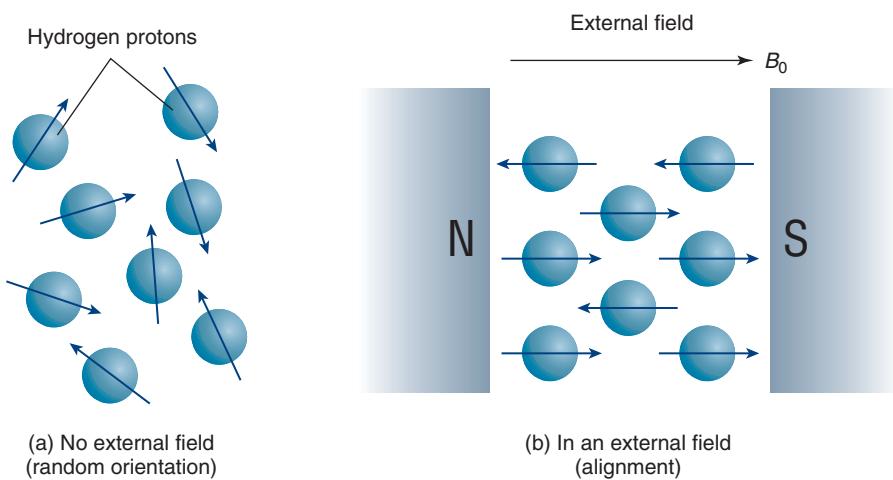


Figure 21.5 Alignment of protons in an external magnetic field

PHYSICS FACT

Net spin of a nucleus

You will recall from the Preliminary Course (*Physics 1 Preliminary Course, 3rd edition*, chapter 11) that linear momentum is possessed by objects moving in a straight line. Spinning objects and orbiting objects have momentum called angular momentum. In classical mechanics, angular momentum has a direction perpendicular to the plane in which the object is spinning or orbiting. If the object is moving clockwise, when viewed from above, the direction of the angular momentum is down. If the object is moving anticlockwise, the direction of the angular momentum is up.

In atomic physics, electrons can exist in different energy levels around the nucleus. Using quantum mechanics, each electron can be thought of as having a spin angular momentum and an orbital angular momentum. The total angular momentum is a combination of the spin angular momentum and the orbital angular momentum. (This is often thought of as an electron that is spinning and orbiting the nucleus, although the electrons are not actually spinning or orbiting the nucleus like a satellite around the Earth. The angular momentum of the electron is a concept from quantum mechanics.)

The nucleus of an atom is composed of positively charged protons and neutral neutrons — collectively called nucleons. The model that deals with nuclear spin is called the Nuclear Shell Model. Each nucleon has an energy level associated with it, just like the energy levels of electrons. Each nucleon has a total angular momentum which is a combination of the spin angular momentum and the orbital angular momentum. The equation from quantum mechanics that gives the total angular momentum is:

$$J = \frac{Ih}{2\pi}$$

where

J = total angular momentum

h = Planck's constant

I = net spin.

The net spin can be zero, multiples of $\frac{1}{2}$, or whole numbers.

To calculate the net spin for nuclei we need to follow some rules from quantum mechanics. These are:

- There can be a maximum of two protons or two neutrons in any energy level in the nucleus. A proton cannot be paired with a neutron.
- The pair of neutrons will have opposite spins to one another, called spin-up and spin-down. The same rule applies to a pair of protons. It should be noted that spin is a concept from quantum mechanics and there is no evidence that the protons and neutrons are actually spinning.

In general:

1. If the mass number is even but the atomic number is odd, the net nuclear spin is a whole number.
2. If the mass number and the atomic number are both even, the net nuclear spin is zero.
3. If the mass number is odd, the net nuclear spin is a multiple of $\frac{1}{2}$.

It is very difficult to calculate the net spin for a nucleus without using quantum mechanics at a level beyond this course. Usually the net spin of a nucleus is looked up in a table to determine whether the spin is non-zero. The nucleus would behave as a small magnet if the spin was non-zero. Table 21.1 shows the net spin for some nuclei.

Table 21.1 Net spin for selected nuclei

NUCLEUS	NET SPIN
Hydrogen-1	$\frac{1}{2}$
Helium-4	0
Carbon-13	$\frac{1}{2}$
Fluorine-19	$\frac{1}{2}$
Sodium-23	$\frac{3}{2}$
Aluminium-27	$\frac{5}{2}$
Phosphorus-31	$\frac{1}{2}$

Although any nucleus with a non-zero net spin can respond to an external magnetic field and could be used in MRI, it is the nucleus of hydrogen, ${}_1^1H$, that is used most frequently, because hydrogen is a very common component of chemicals in the human body.

PHYSICS FACT

Energy levels of protons

We are familiar with the idea that electrons orbiting an atom can have discrete energy levels. (We say their energy is quantised.) If an electron jumps from a high energy level to a lower energy level it releases a bundle of energy, known as a quantum, equivalent to the difference in energy of the two levels. This energy is often measured in the convenient units of electron volts (eV) where $1\text{ eV} = 1.6 \times 10^{-19}$ joules.

The energy released in these electron transitions is of the order of 10 eV and often corresponds to a frequency in the visible electromagnetic radiation range. This is what happens when particular elements are excited and then return to a lower energy level, emitting their own characteristic colours and

allowing the element to be identified (this is discussed in the Astrophysics option, chapter 15, pages 282–284).

A similar absorption or release of energy occurs when hydrogen protons, having been subjected to a strong magnetic field, move between parallel and anti-parallel orientations in the hydrogen atom. The energy difference between these two levels is only about 0.2 μeV and corresponds to electromagnetic radiation of frequency about 42.5 MHz. This is in the radio frequency range. Hence by directing the correct radio frequency waves at the protons, they can be made to absorb energy and change from being parallel to being anti-parallel to the external magnetic field.

21.2 THE MRI MACHINE: EFFECT ON ATOMS IN THE PATIENT

An MRI machine must provide:

- a strong magnetic field
- additional weaker varying magnetic fields, called gradient magnetic fields
- pulses of radio frequency waves
- detectors of the radio frequency waves
- computers to analyse the signals received.

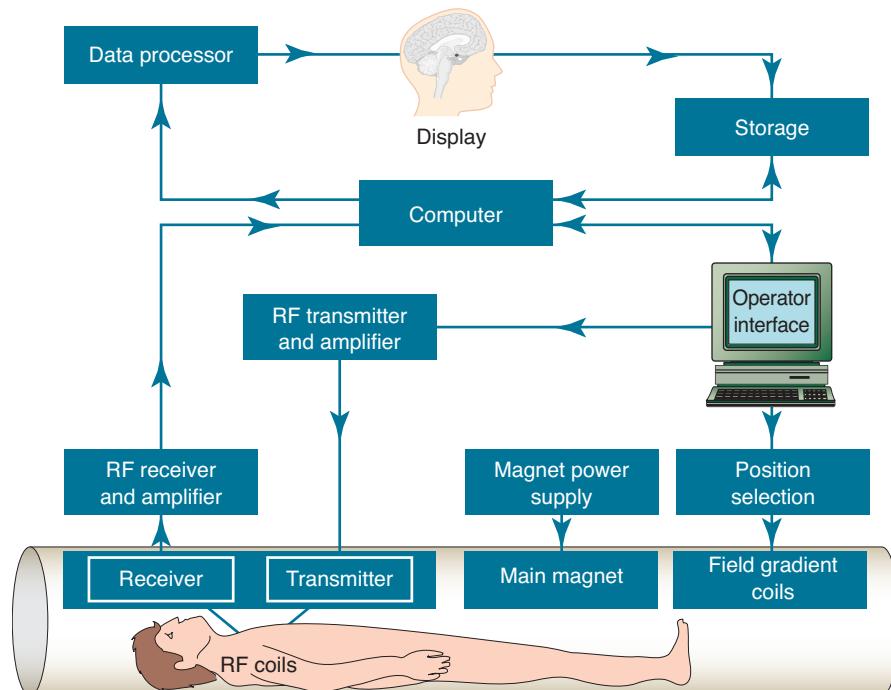


Figure 21.6 A block diagram of an MRI machine

Effects on the orientation of nuclei of applying a strong magnetic field

The patient undergoing MRI lies on a bed in a cylindrical bore called a gantry and is subjected to a uniform magnetic field of strength between 0.2 and 2 teslas (T). The strength of this field can be appreciated when we realise that it is more than 10 000 times the strength of the Earth's magnetic field. A magnetic field with a strength of approximately 2 T results in an image of better resolution than that obtained when the field is 1 T or less.

The magnetic field causes protons in the hydrogen atoms in the patient's body to try to line up either parallel or anti-parallel to the external field.

PHYSICS FACT

The strong external magnetic field

You may ask how such a strong magnetic field can be produced by a machine. There are three possible ways such a magnet could be produced.

1. Permanent magnets

The permanent magnets are usually made of an alloy of aluminium, cobalt and nickel (alnico). Due to their large weight, only 0.2 T field strength can be achieved. Even then the magnet weighs about 80 tonnes! They need no power supply, the field does not extend significantly beyond the magnet and the running costs are low. However, the magnet cannot be switched off and the scanning time is long, producing an image of only reasonable quality. In figure 21.7 we can see that the external magnetic field does not exist along the patient's body.

2. Electromagnets

Electromagnets are created by passing a direct current through a coil of copper wire. Magnetic field strengths no greater than 0.5 T are generated due to the significant loss of energy through heat in the wires. For example, to generate a magnetic field of 0.15 T, over 60 per cent of the energy put in is converted to heat in the wires. Up to 150 litres of water per minute must be pumped through the system to remove this heat. The running cost is high as a large power supply is needed. An advantage of having a power supply is that it can be switched on and off. The magnetic field extends beyond the coils so shielding is necessary. The weight of the magnet is only about 2 tonnes and installation costs are relatively low. Although the scan time is long, the image quality is good.

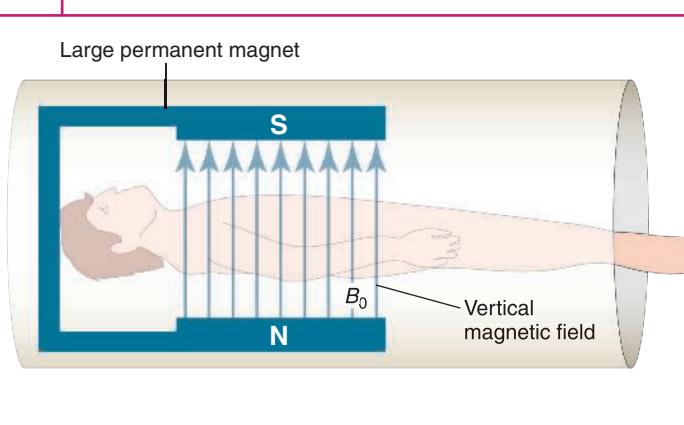


Figure 21.7 Magnetic field from a permanent magnet

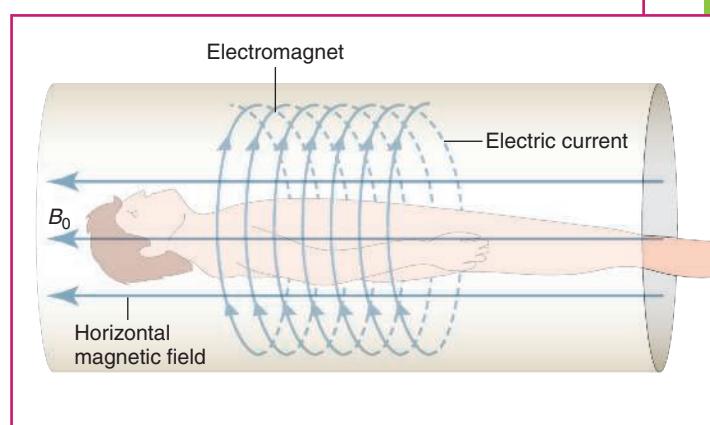


Figure 21.8 Magnetic field from an electromagnet

(continued)

3. Superconducting magnets

These magnets use a coil of superconducting material, which is cooled with liquid helium to the critical temperature. This means that huge currents can flow with very little power input and magnetic fields up to 2.5 T can be obtained. Shielding is needed because the large magnetic field extends outside the coil. Cost of installation is high for this magnet, which weighs about 6 tonnes. The maintenance of the coolants also adds to the running costs. To shut the magnet down the coolant must be slowly drained. The use of this magnet allows short scan times with excellent image quality. The limited supply of helium as a coolant has fuelled research to find suitable superconducting material that becomes a superconductor at the temperature of liquid nitrogen or even at higher temperatures.

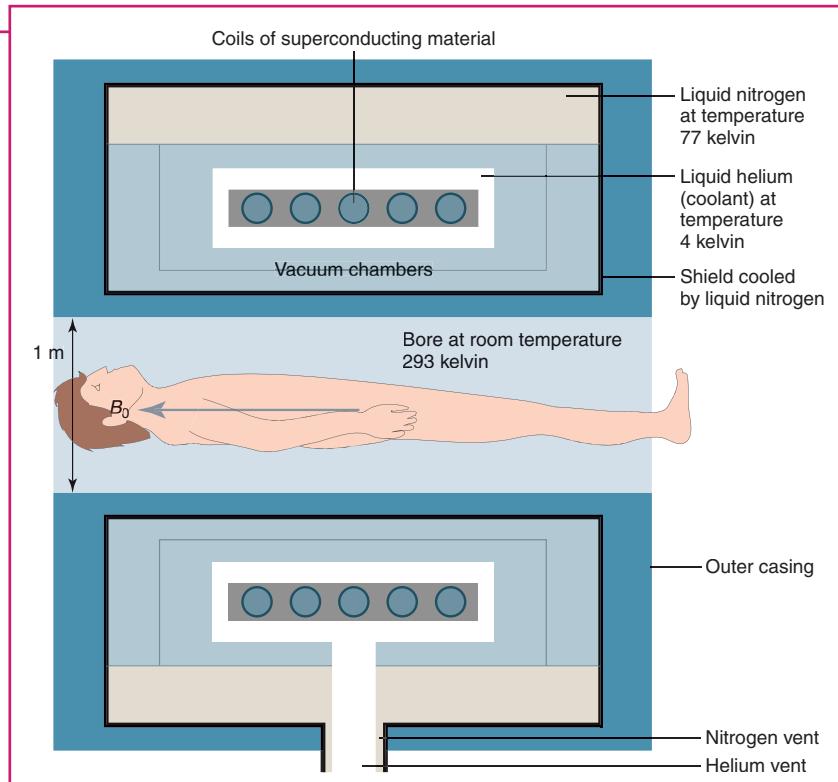


Figure 21.9 A superconducting magnet shown in cross-section. The external magnetic field runs along the length of the patient's body.

Precession

When the nuclei with net spin change their magnetic field orientation in response to the external magnetic field, they do not remain in a steady position along the external magnetic field, but rather precess around the direction of the magnetic field.

To understand **precession**, consider a spinning top. It stays in the upright position while it is spinning rapidly. However, if it slows down and starts to tip over due to gravity, or if you tilt it off its vertical orientation and allow the force of gravity to try to tip it over, the spinning top starts to wobble on its axis, tracing out a conical path. This motion, called precession, is similar to the motion of the nuclei in response to the force of the external magnetic field (see figure 21.10).

The frequency with which a nucleus precesses in a given magnetic field is called the **Larmor frequency**. The Larmor frequency is different for different nuclei in the same magnetic field, as illustrated in table 21.2. We can use the Larmor frequency in a given magnetic field to identify an element.

Table 21.2 Larmor frequency of nuclei

NUCLEUS	LARMOR FREQUENCY IN A 1.0 T MAGNETIC FIELD (MHz)
Hydrogen-1	42.57
Carbon-13	10.70
Phosphorus-31	17.24

Precession is the movement, in a conical path, of the axis of a spinning object.

The **Larmor frequency** is the frequency with which a nucleus precesses about its spin axis, in response to the force due to an external magnetic field.

The Larmor frequency depends on the strength of the external magnetic field. The Larmor frequency is $42.57B_0$ MHz for hydrogen protons, where B_0 is the strength of the external magnetic field. This equation shows that the Larmor frequency is proportional to the strength of the applied external magnetic field. If the external field strength is 1 T, the Larmor frequency is 42.57 MHz. The frequency 42.57 MHz is in the radio frequency range, a fact that is critical for the MRI process because radio frequency signals are made to interact with the precessing hydrogen protons.

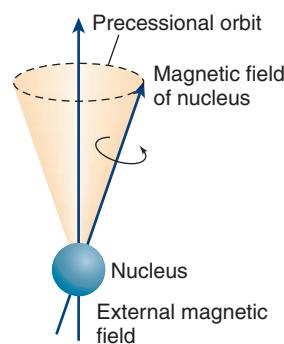


Figure 21.10 Precession of a nucleus due to the application of an external magnetic field

SAMPLE PROBLEM

21.1

Calculating the Larmor frequency

Calculate the Larmor frequency for hydrogen protons in a magnetic field of strength 1.6 T.

SOLUTION

For hydrogen protons:

$$\text{Larmor frequency} = 42.57B_0 \text{ MHz}$$

$$\text{where } B_0 = 1.6 \text{ T}$$

$$\therefore \text{Larmor frequency} = 42.57 \times 1.6 \text{ MHz} \\ = 68.11 \text{ MHz}$$

Application of radio frequency pulses

To **resonate** means to absorb energy when an applied frequency matches the natural frequency of an object.

With the patient in the strong magnetic field, so that the protons in the hydrogen atoms are precessing around the direction of the external magnetic field, a pulse of radio frequency electromagnetic radiation is beamed into the patient. The frequency is chosen to correspond exactly with the Larmor frequency. This is the frequency of precession of the protons. The protons will **resonate** with the radio frequency, and so absorb its energy, move to a higher energy level and precess in phase with one another. A radio oscillator produces pulses of a precise frequency. The radio oscillator can change the frequency of the signal to match different Larmor frequencies.

The amount of energy absorbed is very small, corresponding to the small energy difference between the parallel and anti-parallel precessing protons. (Recall from the Physics fact on page 402 that this energy is in the radio frequency range.) When the pulse is switched off, the protons release the absorbed energy. The intensity and duration of the energy signal released is analysed, enabling part of an image of a ‘slice’ through the patient’s body to be obtained. (The complex steps from releasing the energy to creating an image are discussed later in the chapter.)

You may be wondering how a single slice can be imaged when the Larmor frequency is the same for all the protons that are precessing from one end of the patient’s body to the other. Ability to distinguish between ‘slices’ of the patient is achieved by changing the strength of the strong field slightly and uniformly along the length of the patient’s body by the use of gradient coils. In this way, the Larmor frequency changes along the patient’s body. Then the radio frequency oscillator can change the frequency of the pulse of the signal to match the particular Larmor frequency of the chosen ‘slice’ of the body.

A **gradient magnetic field** is one that changes by small known increments throughout the region of the field.

The gradient magnetic field

A **gradient magnetic field**, which changes by small uniform amounts, is applied along the length of the patient's body. This field adds to the external field, varying it by no more than 1 per cent (see figure 21.11). This means that the Larmor frequency changes along the length of the patient's body.

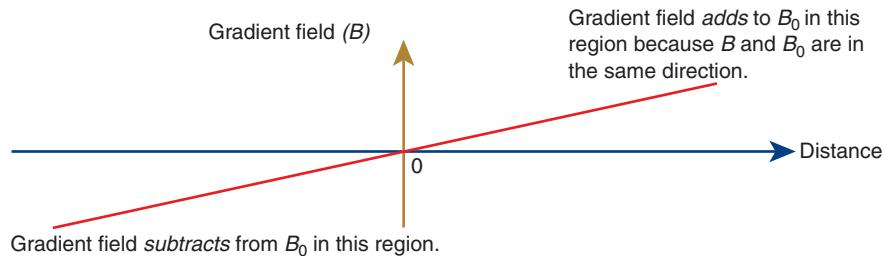


Figure 21.11 A gradient magnetic field changes by small amounts over a distance.

The gradient magnetic field can be generated from oppositely wound coils that have the same current flowing but vary uniformly in how tightly they are wound. In figure 21.12, the direction of the current in the right-hand coil produces a magnetic field which adds to the external field. The field is stronger at the far right-hand end of the coil because the coils are wound more tightly. The way the coils are wound on the left-hand side results in a field which is opposite in direction to the external field and so reduces the external field. The tightness with which the coils are wound changes uniformly and so the strength of the field produced changes uniformly.

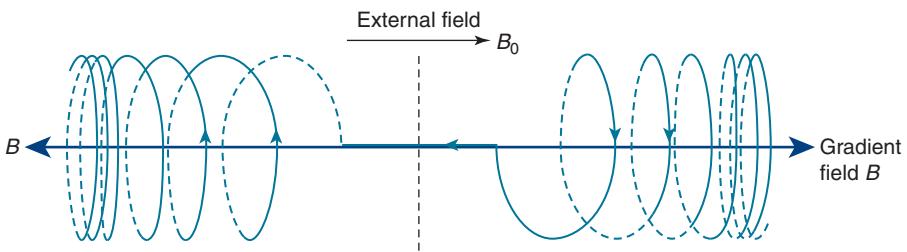


Figure 21.12 A gradient field generated by coils of wire

SAMPLE PROBLEM

21.2

Determining the gradient magnetic field

A small child of height 1.0 m is placed in the magnetic field (1.5 T) of an MRI machine. A gradient magnetic field is applied to the external field. This gradient field increases the external field by 0.005 T for every 10 cm between the midpoint of the child and her head, and decreases the external field in the same way between the midpoint of the child and her feet.

Determine:

- the total external magnetic field strength at the child's chin, which is 20 cm below the top of her head
- the total external magnetic field strength at the bottom of her feet.

SOLUTION

- The child's chin is 30 cm above her waist. As the field is increased by 0.005 T for every 10 cm moved towards her head,

$$\begin{aligned}\text{external field strength} &= 1.5 + 3 \times 0.005 \text{ T} \\ &= 1.515 \text{ T}\end{aligned}$$

- The child's feet are 50 cm below her midpoint. As the field is decreased by 0.005 T for every 10 cm moved below her midpoint,

$$\begin{aligned}\text{external field strength} &= 1.5 - 5 \times 0.005 \text{ T} \\ &= 1.475 \text{ T}\end{aligned}$$

Removal of the radio frequency pulses

When the radio frequency pulse is stopped, the protons release their absorbed energy. This energy released is the same as that absorbed from the radio frequency pulse. The energy is detected by a radio receiver, then the signal is sent to computers to be analysed.

How can this radio frequency signal be analysed to identify where in the slice the signal is coming from and to distinguish one type of hydrogen compound from another? We will look at these questions in the following sections.

Localising the signal within a slice

In order to determine exactly where particular signals originate, two further small gradient fields in the plane of the slice are applied.

In one direction the gradient field modifies the phase of the precessing protons very slightly. That is, it makes the protons slightly out of step with one another as they precess.

In the other direction, at right angles to the first, the gradient field alters the frequency of the precessing protons. It is common for the gradient fields in the two directions to divide the 'slice' into 256×256 **voxels**. The response from protons in each voxel can be determined from the very complicated returning signal (see figure 21.13). This signal is analysed mathematically by a process called Fourier transformation and each voxel is given an intensity value from which an image can be constructed. The gradient fields need to be switched on and off rapidly, allowing the pulses to be repeated at a fast rate. A typical rate at which the radio frequency signal is switched on and off is 63 million times per second for an MRI machine using a magnetic field of 1.5 T. Every time a pulse is sent in and then switched off, new information about the composition of particular voxels is obtained.

A **voxel** is a small volume. It is part of a 'slice' through the body.

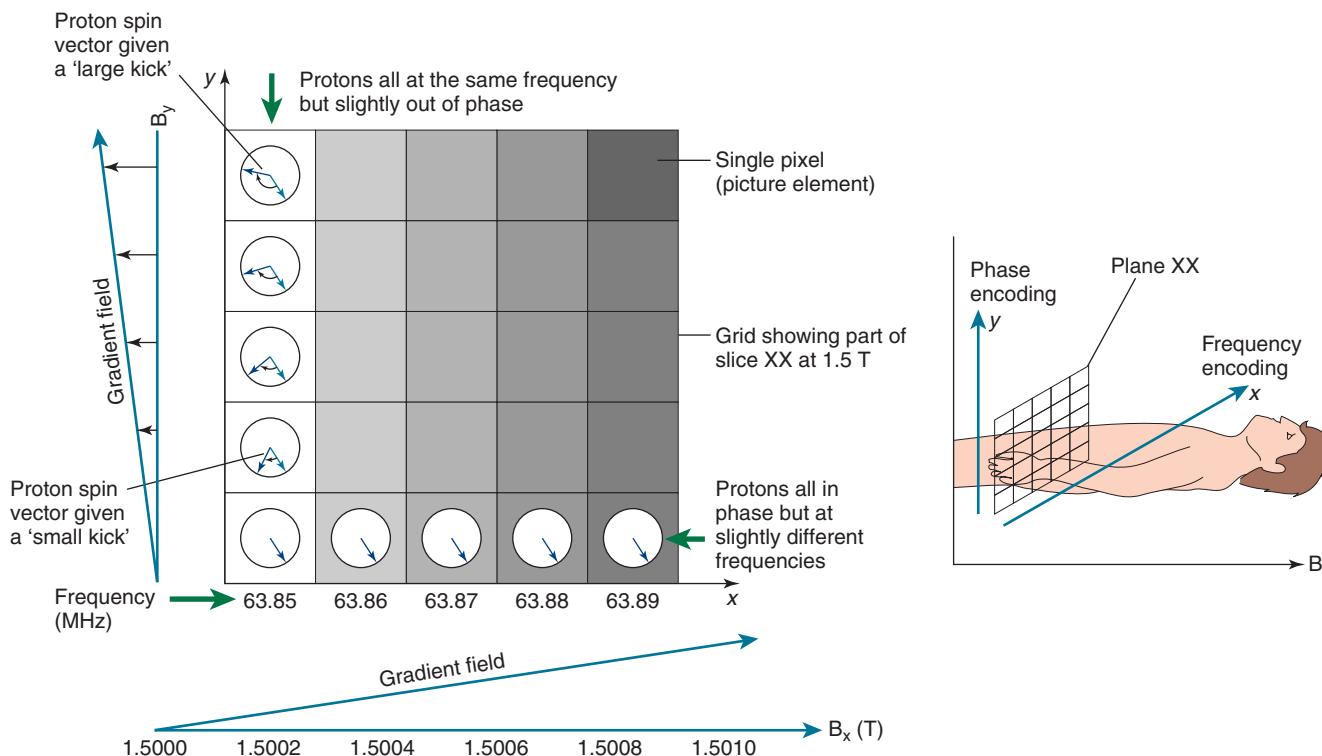


Figure 21.13 Localising the signal within a single slice

Distinguishing one type of hydrogen compound from another — factors influencing the strength of the signal returned from the patient

If the signal strength returned is large, the area on the image is bright; and if the signal is small, the area is dark. The signal is influenced by many factors. However, the three most important factors are:

- proton density
- type of tissue
- rate of radio frequency pulse (the rate will enable contrast to be changed). The first two factors are properties of the material being imaged. The third factor is imposed on the material in order to change the contrast in the image.

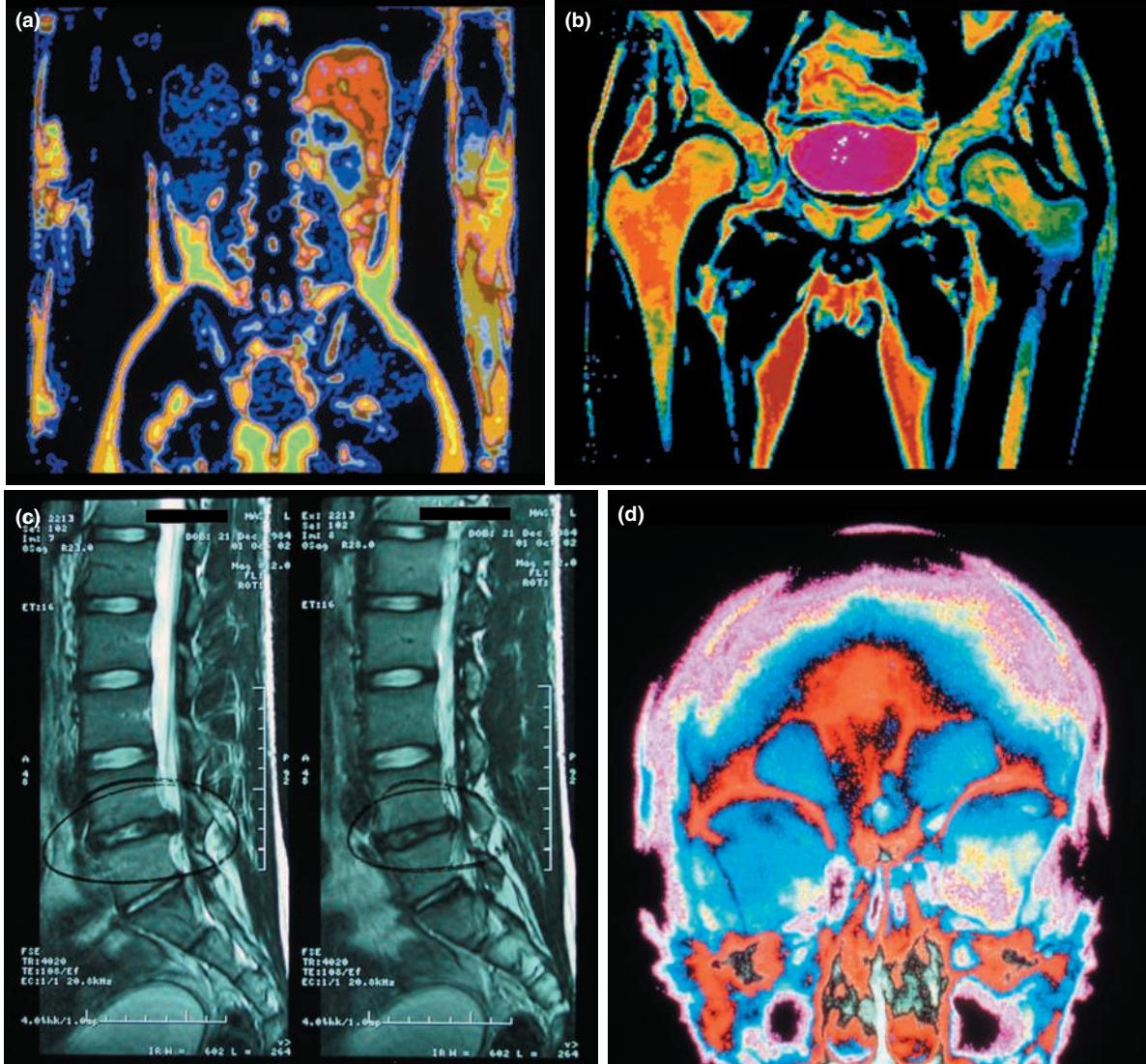


Figure 21.14 Images using MRI and showing clear contrast of soft tissue (a) MRI of the lungs, showing cancer on the right lung (in orange) (b) MRI of a normal pelvis (c) A patient's spinal column showing damaged discs (circled) (d) MRI image of a normal skull

The greater the density of hydrogen protons, the larger the signal and the brighter the image. Air and outer bone have no hydrogen and so appear dark in an MRI image. Cerebrospinal fluid in the brain and spinal column has a large amount of water that is not bound to other molecules. The hydrogen protons in the water are quite mobile, therefore water produces a strong signal and shows up bright on an image. Soft tissue also appears quite bright. However, it is difficult to differentiate between sections of the soft tissue, unless relaxation time, discussed below, is taken into account.

The type of tissue determines how easily the protons can release their energy to neighbouring nuclei. By examining the time it takes to 'relax' or reach the original energy state, the type of tissue can be identified.

By changing the rate at which the radio frequency pulse is sent into the patient, we can make the image of one type of tissue brighter or less bright and hence change the **contrast** in the image, and identify what type of tissue we are looking at.

Contrast refers to the brightness difference between parts of the image.

Relaxation refers to the precessing nuclei moving back to their original energy state.

Measuring relaxation time and changing the contrast in images

When the radio frequency pulse is removed, the protons move back to their original energy level. This process is called **relaxation** and the time it takes to do that is called the relaxation time.

There are two relaxation times that can be measured, and the measurements can be used to emphasise different aspects of the image. The resulting different images are said to be T_1 weighted or T_2 weighted.

For the relaxation time called T_1 , relaxation energy is transferred from precessing protons to the surrounding molecules. If the molecules have a natural frequency of oscillation close to the Larmor frequency, energy is more rapidly released. The T_1 value is small. This is true for larger molecules or bound water molecules, which move more slowly than free water molecules. Fat, liver and spleen tissues have a short T_1 . An image emphasising small T_1 values is said to be T_1 weighted.

You may be wondering how the T_1 image can be made. It depends on the rate at which the radio frequency pulses, mentioned in the previous section, are switched on and off. If the radio frequency pulses are repeated quickly before all the protons have had time to relax, those with a short relaxation time will keep absorbing and releasing energy and appear bright. (Fat molecules, which are large, will appear bright in this case.)

For the relaxation time called T_2 , we are measuring the time for protons to go out of phase with one another by exchanging energy with one another. T_2 is short for solids and larger molecules which are found in tendons, muscle and liver, and long for watery tissues. Images recording the long T_2 relaxation time are said to be T_2 weighted.

Once again, the pulse rate of the radio frequency signals is important in obtaining T_2 weighted images. If the radio frequency pulses are repeated more slowly, watery tissues will have had time to relax before the next pulse is sent in and so they will appear bright. Table 21.3 on the following page summarises the features involved in the different weighted images, and figure 21.15 shows some examples of images.

Table 21.3 Types of images and their features

TYPE OF IMAGE	APPEARANCE OF ORGANS	USEFULNESS
T ₁ weighted	Fat and larger molecules are bright. Water is dark.	For body structure. Excellent for soft tissue detail.
T ₂ weighted	Watery tissues and diseased tissues bright. Tendons, muscle and liver are dark.	Preferred for investigating diseased areas.
Proton density weighted images	Urine and cerebrospinal fluid are highlighted because the density of water (protons) is high.	Preferred for showing diseased organs.

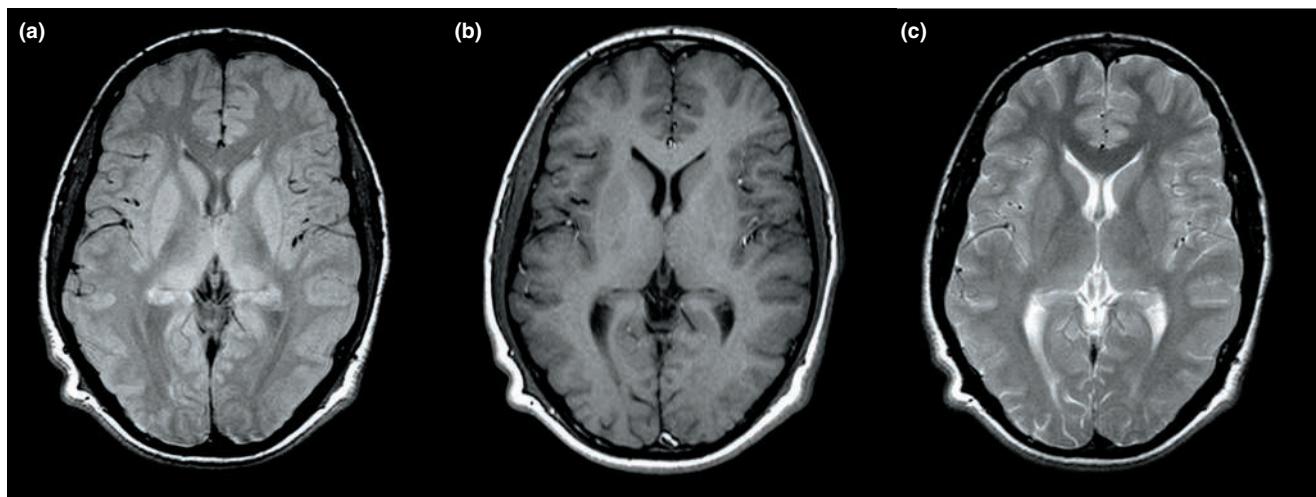


Figure 21.15 The same section through the brain showing images that are (a) proton density weighted, (b) T₁ weighted and (c) T₂ weighted.

Each time the pulse is switched on and off, the gradient fields used to locate signals within a ‘slice’ are also switched on and off. The rapid switching of the fields causes the loud noises heard by the patient during the scan.

If the manipulation of contrast is inadequate, artificial contrast agents may be injected intravenously or taken orally. Agents that travel through the blood are used for highlighting tissues with a large number of blood vessels, such as tumours, or for imaging blood vessels themselves. The contrast agents have the potential to make the signal stronger from specific tissues or even from regions in which specific genes are being expressed. When genes are being expressed, they are producing certain proteins that can be detected using MRI. Genetic research is already using MRI in this way.

21.3 MEDICAL USES OF MRI

MRI is considered to be the best diagnostic imaging technique for structural resolution and contrast. It depicts soft tissue so well that it is the preferred choice for imaging the brain and spine, where it is able to show suspected tumours and slipped discs. It is useful for imaging areas with large amounts of water as these areas have many hydrogen nuclei. Cancerous tumours contain different amounts of water from normal tissue

or are surrounded by watery tissue. They can be distinguished in an MRI scan because of the different brightness due to different proton density.

Grey matter in the brain and spinal cord contains hydrogen bound in a different way from that in white matter. As a result, the relaxation times for hydrogen protons are different in grey matter and white matter. This means they are able to be distinguished in an MRI scan. The fact that nerve cells of grey matter can be distinguished from those of white matter can be used in the diagnosis of multiple sclerosis.

A clear image of the brain and spinal cord can be made without the skull or spine interfering, because bone contains no hydrogen and will not show up in an MRI scan.

PHYSICS FACT

Cardiac MRI allows investigation of congenital abnormalities and coronary heart disease to be carried out. Improvements in the speed of MRI have made abdominal imaging possible. Early MRI machines took 10 minutes to scan 24 'slices' of the body and this can now be done in under 1 second. Injection of a contrast agent into the blood, combined with rapid imaging techniques, now allows blood flow in the kidneys to be examined and narrowing of the arteries due to fatty plaques to be seen.

Functional MRI allows parts of the brain to be investigated while changes are taking place. There is an increased flow of oxygenated blood to areas that are stimulated. Knowledge of the magnetic properties of oxygenated blood allows parts of the brain involved in processing sensory data or motor tasks to be identified and studied (see figure 21.16). Parts of the brain may be able to be studied prior to surgery.

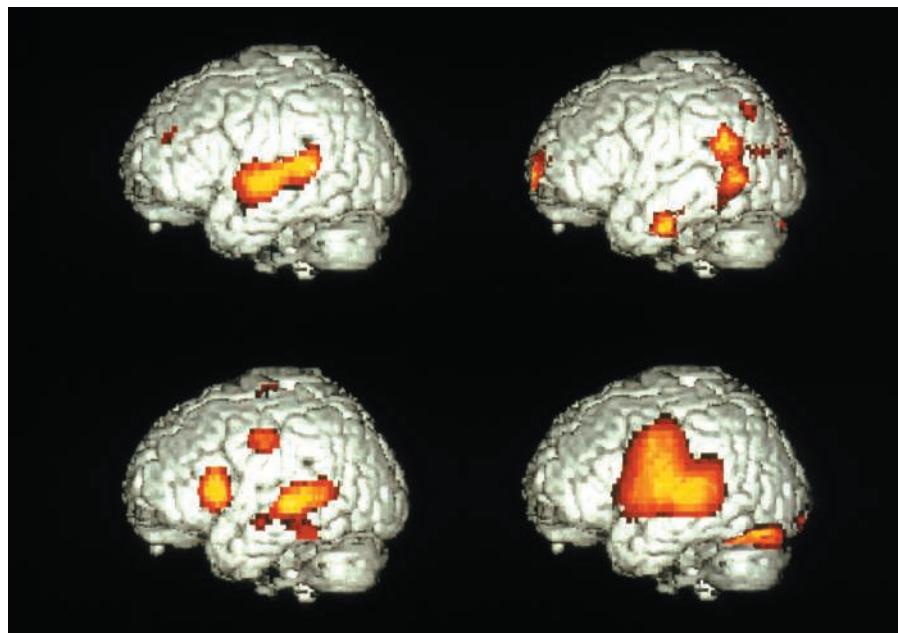


Figure 21.16 Brain activation showing increased blood flow due to a simple motor task

With the development of special non-metallic tools and more open magnet geometries for MRI machines, minimally invasive procedures and open surgery can be performed inside the MRI scanner.

21.4 COMPARISON OF THE MAIN IMAGING TECHNIQUES

Table 21.4 provides some comparison between imaging techniques. Improvements are being made in the machines used in all these imaging methods, and students are advised to search the internet for the latest advances. At the time of printing, the white boxes represent the preferred method for imaging the organ or tissue indicated.

Table 21.4 Comparing imaging techniques

	ULTRASOUND	X-RAYS	CT	NUCLEAR	MRI
Cost of machine (capital cost)	Moderately expensive	Least expensive	Quite expensive	Quite expensive	Very expensive
Mobility of machines	Portable machines commonly used	Small portable machines available	Fixed machines	Fixed machines	Very few mobile machines
Spatial resolution (ability to see fine detail)	1–5 mm	0.1 mm	0.25 mm	5–15 mm	0.3–1 mm
Time for examination	Moderate	Very fast	Moderate	May be long, depending on tracer and procedure.	Relatively long but some procedures are now quite short
Comfort and safety	No known hazards	Small dose of ionising radiation	Usually higher dose of ionising radiation than for X-rays	Moderate dose of ionising radiation from radioisotopes	Some claustrophobia from lying inside the bore containing the magnetic field. Patients with metallic implants cannot be scanned.
Imaging soft tissue of abdomen	Excellent, especially for obstetric cases, as it is safe and real-time imaging is possible (see page 344)	Image poor — needs contrast medium	Good for whole abdomen scan	Good for growth of tumours and functional study of liver and kidneys (see page 381)	Good resolution for specific areas e.g. kidneys
Imaging soft tissue of joints	Reasonable if bone can be bypassed	Poor contrast	Good — preferred to MRI when extra bone detail is needed (see page 414)	Poor resolution but good for functional information	Excellent for studying muscles, tendons and cartilage (see page 408)

	ULTRASOUND	X-RAYS	CT	NUCLEAR	MRI
Imaging heart and circulation	Excellent for structure and using Doppler technique for blood flow (see page 355)	Contrast medium is needed	Limited use with digital imaging techniques	Good for blood flow studies	Good resolution and ability to measure blood flow
Imaging chest	Poor as air-tissue boundary reflects sound waves	Adequate for routine lung screening (see page 361)	Better detail than X-rays	Good for functional studies of blood and air flow	Not good for imaging air spaces
Imaging brain and spinal cord region	Poor as bone-tissue boundary blocks sound waves	Limited use as bone blocks most waves	Good and preferred to MRI for details of bone of spine (see page 414)	PET scans are useful for showing function	Excellent for giving good contrast between tissues
Imaging bone	Poor as waves are blocked by bone	Gives very good resolution (see page 367)	Good when more complicated structures must be viewed	Good for whole body bone cancer and early diagnosis of stress fractures	Signal is weak so of limited use.

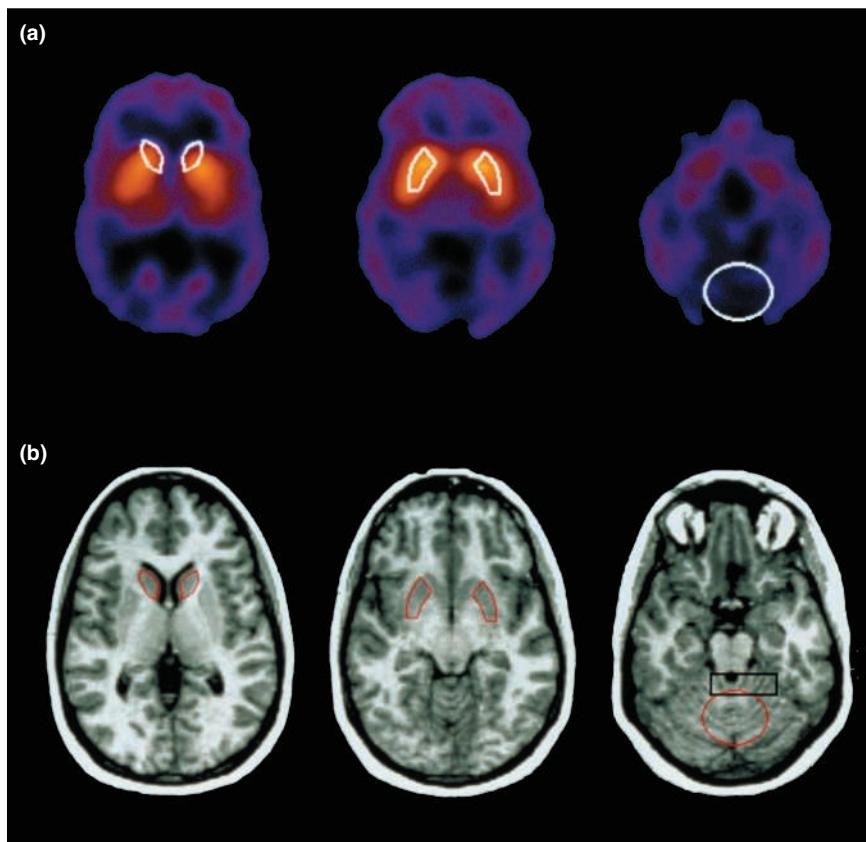


Figure 21.17 Comparison of (a) PET and (b) MRI scans of the brain

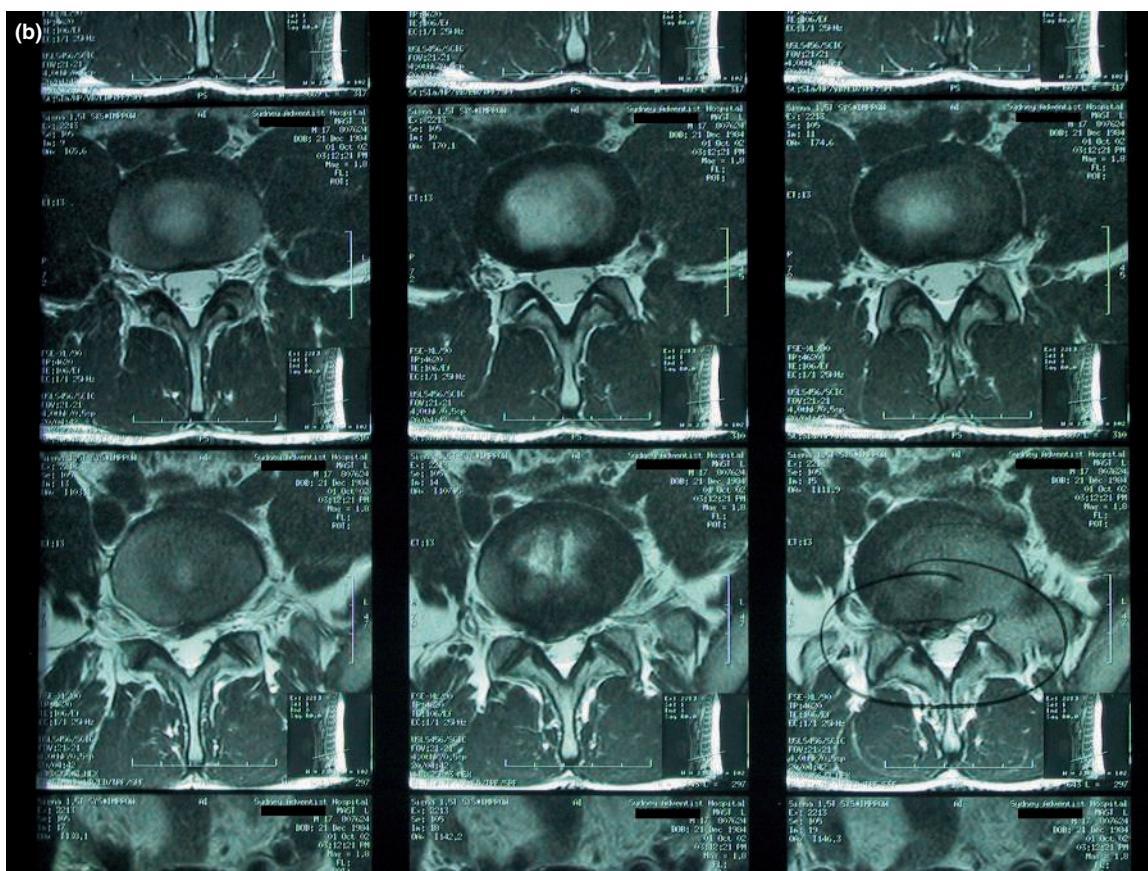
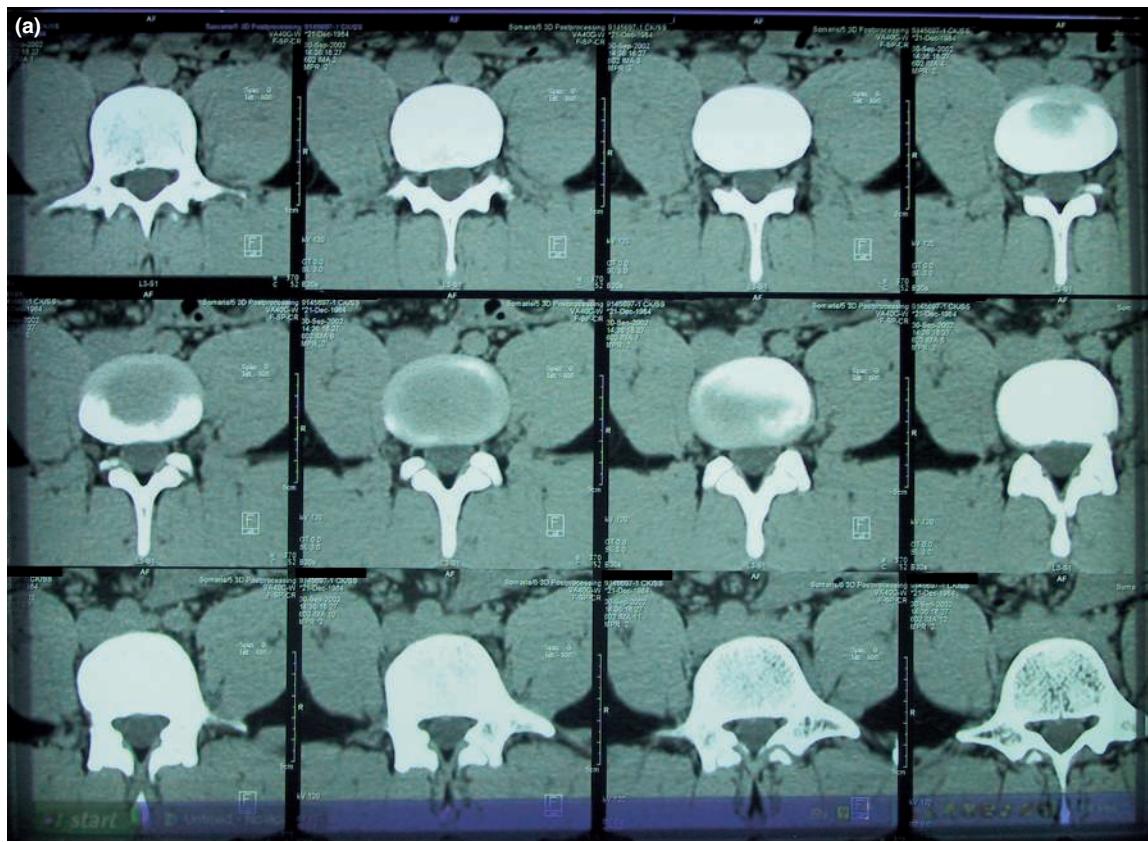


Figure 21.18 (a) CT scan of the lower disc in the spine (b) MRI scan of the same area of discs of the spine

SUMMARY

- Nuclei with net spin align themselves in an external magnetic field and precess about the field direction with a frequency dependent on the strength of the field.
- Subjecting precessing nuclei to pulses of radio waves at the Larmor frequency causes some nuclei to move from a low to a high energy state.
- When the high energy nuclei relax, they produce a signal, the intensity of which is related to the number of nuclei present.
- The relaxation signal allows information about the abundance of the atom and its bonding with neighbouring atoms to be determined.
- By changing the rate at which the radio frequency pulses are switched on and off, different aspects of the image can be emphasised.
- The MRI machine consists of a strong magnet and at least three other varying magnets, a radio frequency oscillator, a radio receiver, and a computer.
- An MRI scan detects soft tissue clearly, allowing cancerous tissue and areas of high blood flow to be detected and grey and white matter in the brain to be distinguished.

QUESTIONS

1. Compare the types of strong magnets that may be used for the MRI machine. (See the Physics fact on magnets, page 403.) Use a table for your answer.
2. (a) Outline what is meant by:
 - (i) proton spin
 - (ii) proton precession.
 (b) Describe how an external magnetic field influences a hydrogen proton.
3. (a) Outline two reasons why the hydrogen nucleus is imaged more than other nuclei in MRI.

(b) When protons are in a strong magnetic field they can occupy one of two possible energy states. Describe these energy states and state which is the higher of the two.

(c) Describe in what sense the MRI technique is nuclear.
4. (a) Explain the function of the magnetic gradient fields used in MRI.

- (b) How many magnetic gradient fields are used in an MRI machine?
- (c) A patient of height 1.8 m is positioned horizontally in a steady, uniform, horizontal magnetic field of flux density 1.0 T, running from her feet to her head. If a gradient field of 8.0 mT m^{-1} is applied horizontally from her feet to her head (as in figure 21.11, page 406), calculate the magnetic flux density 1.4 m from her feet.
5. Outline why gradient magnetic fields are needed to image a slice through the body.
6. Explain why artificial metal joints between bones and metal fillings in teeth are a problem with MRI.
7. Use figure 21.15 (page 410) to answer these questions.
 - (a) Explain why the cerebrospinal fluid that surrounds the brain may look brighter than brain tissue on a MRI image.
 - (b) Predict whether grey or white matter in the brain would look darker on an MRI image. Justify your answer.
8. (a) Compare the superconducting external magnetic field used in MRI with the Earth's magnetic field.

(b) (i) What are superconducting magnets and why do most MRI scanners use superconducting magnets?

 (ii) Are superconducting magnets electromagnets? Justify your answer.

9. Describe two different pieces of information that can be obtained from the relaxation of the protons in the nucleus of a hydrogen atom.
 10. 'Medical physics has produced a wide range of harmless techniques that avoid the use of invasive surgery.' Evaluate this statement.
 11. Examine the photograph in figure 21.14(c) (page 408). From this MRI image, compare the healthy and damaged discs.
 12. Research and then explain why MRI scans can be used to:
 - (a) detect cancerous tissue
 - (b) identify areas of high blood flow
 - (c) distinguish between grey and white matter in the brain.
- To help with your research use a search engine and key words such as 'MRI and cancer', 'blood flow and MRI', 'grey and white matter and MRI', 'relaxation time and MRI', ' T_1 and T_2 weighted images', 'proton density and MRI'.

13. Compare the advantages and disadvantages of X-ray scans, CT scans, PET scans and MRI scans. Use your own research and information gathered from chapters 18–21 of this book. For example, figure 21.18 on page 414 compares CT and MRI scans of the spine.
 14. Assess the impact of medical applications of physics on society. Plan your answer to this question before you begin to write a full answer. In your final answer make sure you address the meaning of the verb ‘assess’. Ideas for planning are given below.
 - (a) Make a list of medical applications of physics. Be more specific than simply listing, for example, ultrasound. List particular medical applications of ultrasound.
 - (b) Consider the impacts on society that are important. They might include:
 - (i) cost, affecting the time a patient is away from work or the time of treatment or place of treatment
 - (ii) cost of equipment, which might affect the health budget
- (iii) ability to diagnose early, leading to effect on the recovery rate of patients. The effect on people closely associated with the patient should be considered.
- (iv) type of diagnosis and treatment available. Include to what extent the procedure is invasive.
- Resources to use in planning could include:
- newspaper articles about medical applications relevant to the areas you have studied
 - relevant journal articles such as those found in *New Scientist*
 - relatives or friends who can recall medical diagnoses and treatment available 20, 40 and 60 years ago.
- Finally, link the impacts on society with the medical applications you have identified.
- In your presentation of the answer to the question, begin with a statement of your assessment of the impact. In the following paragraphs, link the applications and the evidence to support your assessment. Conclude with a summary statement of your assessment based on the evidence that you have provided.

FROM QUANTA TO QUARKS

Chapter 22

The atomic models of Rutherford and Bohr

Chapter 23

Development of quantum mechanics

Chapter 24

Probing the nucleus

Chapter 25

Nuclear fission and other uses of nuclear physics

Chapter 26

Quarks and the Standard Model of particle physics

CHAPTER 22

THE ATOMIC MODELS OF RUTHERFORD AND BOHR



Figure 22.1 Photograph of ‘aurora australis’, the southern lights. The stars’ trails indicate that the photograph is a time exposure of several minutes. In an aurora, atoms of the gases in the upper atmosphere emit radiation after being excited by interactions with charged particles from the Sun.

Remember

Before beginning this chapter, you should be able to:

- recall the discovery of the electron by J. J. Thomson
- outline Thomson’s ‘plum pudding’ model of the atom
- recall the contributions of Planck and Einstein to the development of the quantum model of light (photons)
- state the relationship between the energy and frequency of a photon ($E = hf$).

Key content

At the end of this chapter you should be able to:

- discuss the main features of the Rutherford model of the atom and identify difficulties with this model
- understand the role that the hydrogen spectrum played in leading Bohr to formulate his model of the atom
- discuss the contribution of Planck to the concept of quantised energy
- state Bohr’s postulates
- understand that with Bohr’s postulates superimposed on the Rutherford atom, it is possible to derive a theoretical equation for the hydrogen spectrum that is in agreement with Balmer’s empirical equation
- solve problems using $\frac{1}{\lambda} = R \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$
- discuss the limitations of the Rutherford–Bohr model of the hydrogen atom.

In the late nineteenth century, many physicists believed that the answers had been found to all the major questions in physics. Electricity, magnetism, light, mechanics, cosmology, gravity — all, they claimed, could be understood using the theories of Newton and Maxwell, which we now refer to as ‘classical’ theories. Many chemists thought similarly about chemistry. They were sure that with elements, each with its own indivisible atom, and the discovery of the periodic table, there was little left to discover. There were minor problems in physics but it seemed likely that these would soon be explained in terms of the existing theories. Even as some discoveries of new phenomena occurred, there still seemed no doubt that classical physics would explain all.

However, discoveries made from 1895 onwards eventually saw the demise of classical physics. Some aspects of classical physics were found to be inadequate and were replaced with a theory that became known as **quantum theory**.

The ideas that led to quantum theory flew in the face of accepted science. Although there were still groups of scientists who denied the existence of atoms, the late nineteenth century saw the atom become generally accepted as a small, indivisible chunk of matter. This was to be challenged by the work of J. J. Thomson and Ernest Rutherford. After fighting for the existence of atoms, many scientists regarded it as heresy when Thomson proposed that electrons were constituents of atoms. Rutherford proposed a nuclear atom and then Bohr looked at introducing ideas of quantum theory to atomic structure. As quantum theory developed, aspects of it were so strange that even some of the most famous physicists were not happy to apply it but found it the only possible way to explain their observations.

Perhaps the most amazing thing is that technology based on quantum theory works. It has given physics and chemistry a firm scientific base. We will study the findings of some of the most significant physicists in these early stages of understanding, and particularly the work of Rutherford and Bohr.

22.1

THE RUTHERFORD MODEL OF THE ATOM

In 1895, Ernest Rutherford (1871–1937), a New Zealand born physicist, went to work with Joseph John (J. J.) Thomson (1856–1940) at the Cavendish Laboratory at Cambridge University in England. As we saw in chapter 10 (pages 180–185), J. J. Thomson had identified the electron as a component of the atom in 1897. The model of the atom changed from the small indestructible sphere of Dalton to the ‘plum pudding’ model of Thomson. Negatively charged electrons were considered to be distributed throughout a sphere of positive charge. Ernest Rutherford and John Townsend helped Thomson with his work that led to this discovery of the electron, although it was Thomson who designed and performed the crucial experiment.

The first alpha particle scattering experiment

In 1898, Rutherford moved to McGill University in Montreal where he investigated radioactivity. While there, Rutherford had a difference of opinion with Henri Becquerel (1852–1908), who had discovered radioactivity in 1896. Rutherford had shown that alpha particles were deflected

Quantum refers to a quantity or an amount (from the Latin word *quantum* meaning ‘how much’). In ‘classical physics’ an object could possess any amount of energy. In **quantum theory** objects could possess only certain discrete amounts of energy. Instead of being ‘continuous’, energy was available only in ‘packets’.

‘Anyone who is not shocked by quantum theory has not understood it.’

Niels Bohr (1885–1962)

‘I don’t like it, and I’m sorry I ever had anything to do with it.’

Erwin Schrödinger (1887–1961)

Rutherford’s early work at the Cavendish Laboratory also included wireless signalling and at one time he held the world record for distance communication of about a kilometre.

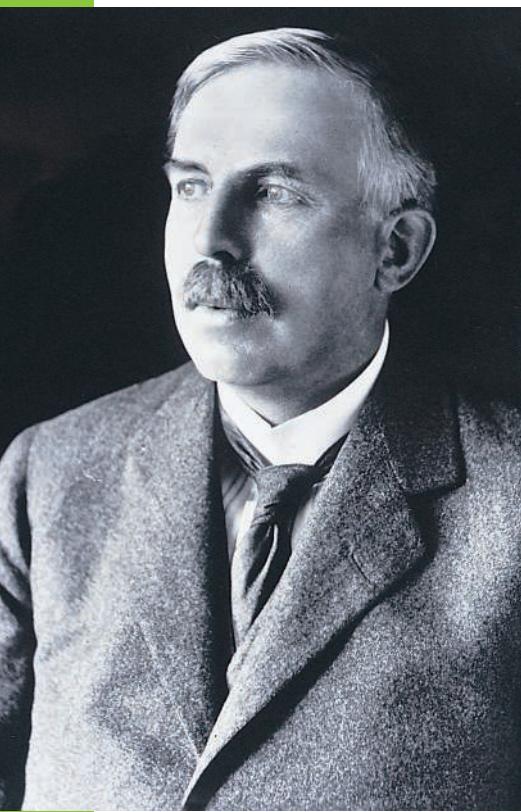


Figure 22.2 Ernest Rutherford (Lord Rutherford of Nelson), was awarded the Nobel Prize for Chemistry in 1908. The award that year was a matter of intrigue. In an attempt to award Nobel prizes to two atomists (Planck and Rutherford) in the same year, Dr Arrhenius, Director of the Nobel Institute for Physical Chemistry, arranged for Rutherford to be nominated for the chemistry prize and for that prize to be determined before the physics prize. In the end, Planck did not win the physics prize that year, but was eventually awarded it ten years later.

by magnetic fields. Becquerel studied the passage of alpha particles through magnetic fields and believed (incorrectly) that he had observed that the radius of curvature of the alpha particles increased as they moved greater distances through the magnetic field. He believed that the alpha particles increased in mass when passing through air and the increase in mass was responsible for the increase in radius. He did not like this idea but preferred it to the other alternative — that alpha particles increase their velocity as they pass through air (see ‘Becquerel’s predicament’ below).

Rutherford demonstrated that alpha particles slow down as they collide with air molecules. He repeated Becquerel’s experiment but with alpha particles passing through a magnetic field in air and also in a vacuum. He found that the beam of alpha particles was wider in air than in a vacuum. He observed that when he put a thin mica sheet in the beam it deflected the alpha particles by up to two degrees. Rutherford calculated that a relatively large electric field must be present in the mica sheet to deflect the alpha particles. As far as is known, he did not speculate about the origin of that electric field.

PHYSICS FACT

Becquerel’s predicament

A charged particle moving in a circular path in a uniform magnetic field will experience a centripetal force of magnitude $F_c = \frac{mv^2}{r}$,

which is provided by the magnetic force of magnitude $F = qvB$. When these equations are combined, an expression for the radius of the path can be determined as $r = \frac{mv}{qB}$.

We can see Becquerel’s predicament. He believed that r was increasing and hence either m or v (or both) would have to increase as q and B were constant. Although he did not like the idea of an increasing mass he preferred it to an increasing velocity and defended it very strongly when Rutherford challenged him.

The real problem was the photographic detection method Becquerel was using. The radius actually decreased.

Geiger and Marsden’s alpha particle scattering experiment

Rutherford did nothing more with alpha particle scattering until 1907 when he moved to Manchester, England. There he inherited Dr Hans Geiger (1882–1942), a German physicist, as his assistant. Rutherford returned to his investigations of the scattering of alpha particles, this time by very thin metal foils. Rutherford suggested to Ernest Marsden (1889–1970), an undergraduate student being trained in radioactive detection techniques by Geiger, that Marsden could determine whether alpha particles were directly reflected from a metal surface. Marsden observed that a very small fraction of the alpha particles were reflected from a thin gold foil. (About 1 in 8000 alpha particles were deflected at an angle greater than 90°.) Geiger and Marsden published these results in 1909. They used a very simple apparatus with a thin conical tube containing ‘radium emanation’, which we now know as radon, as their source of alpha particles (see figure 22.3).

Rutherford and Geiger also developed the Rutherford–Geiger detector — later improved by Geiger and Müller and known as the GM tube or, more commonly, a ‘Geiger counter’.

A **scintillation** is a flash of light observed on a scintillation screen. Another example of scintillation is electrons striking the screen of a cathode ray oscilloscope. The screen produces many scintillations when it is struck by electrons. Of course, the continuous beam of electrons produces a constant glow, not individual flashes as would be observed when alpha particles hit such a screen.

Figure 22.3 Drawing of the apparatus used by Geiger and Marsden, from their original paper published in 1909. AB is the conical tube sealed at the end with a mica sheet. P represents a lead shield that prevents the alpha particles from travelling directly to the scintillation screen, S. RR represents the thin metal foil and M the low power microscope through which the scintillations were observed.

Geiger and Rutherford had confirmed that each **scintillation** (flash of light observed on the scintillation screen) was produced by an alpha particle and that all of the alpha particles produced a scintillation. Geiger and Marsden reported the number of scintillations observed per minute for a number of different metals. They also investigated the number of scintillations observed per minute for different thicknesses of gold foil. The radon gas was at low pressure in the conical tube but the experiment was performed in air. The alpha particles from the conical tube did not form a well-defined beam and, while Geiger and Marsden were able to detect that alpha particles had been deflected through large angles, their simple apparatus was not able to detect a significant change in the number of particles deflected through different angles. (The apparatus was later refined to permit the measurement of angles and the experiment was also performed in an evacuated chamber; see figure 22.4.)

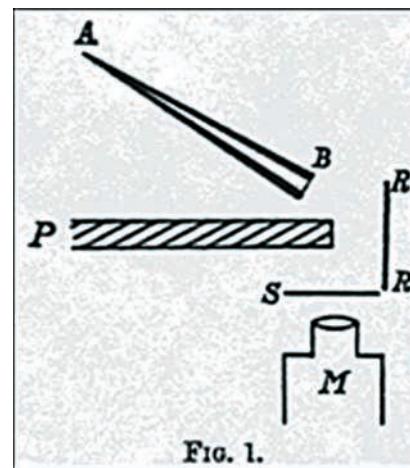


Fig. 1.

PHYSICS FACT

Several years later, during World War I, Hans Geiger and Ernest Marsden found themselves on opposite sides of the same sector of the front line in France. In 1915, Marsden had been appointed Professor of Physics at Victoria College in Wellington, New Zealand. Marsden joined the New Zealand army as a signals officer and returned to the fighting in France where he won the Military Cross. While in France he received a letter from Geiger, congratulating him on his appointment. Rutherford also kept in touch with Geiger and other German scientists by sending letters via mutual friends.

In one of his last lectures, Ernest Rutherford described his reaction to Marsden’s discovery of deflection of alpha particles through large angles as ‘the most incredible event that has ever happened to me in my life. It was almost as incredible as if you had fired a fifteen-inch shell at a piece of tissue-paper and it came back and hit you’.

The nuclear atom

It was two years after the publication of the paper by Geiger and Marsden that Rutherford explained the result by proposing a nuclear atom. Rutherford informed Geiger that he knew what the atom looked like. Using his nuclear model, he had worked out the relative numbers of alpha particles that would be scattered through different angles and Geiger began a series of careful experiments (see figure 22.4 on the following page).

Figure 22.4 The apparatus used to study alpha particle scattering in 1911. In this version, the microscope and scintillation screen can be rotated to observe the alpha particles at different angles. Polonium was used as the alpha particle source, the metal foil used was gold and the apparatus was evacuated.

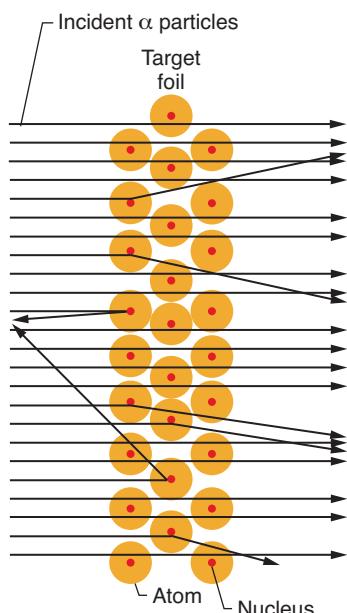
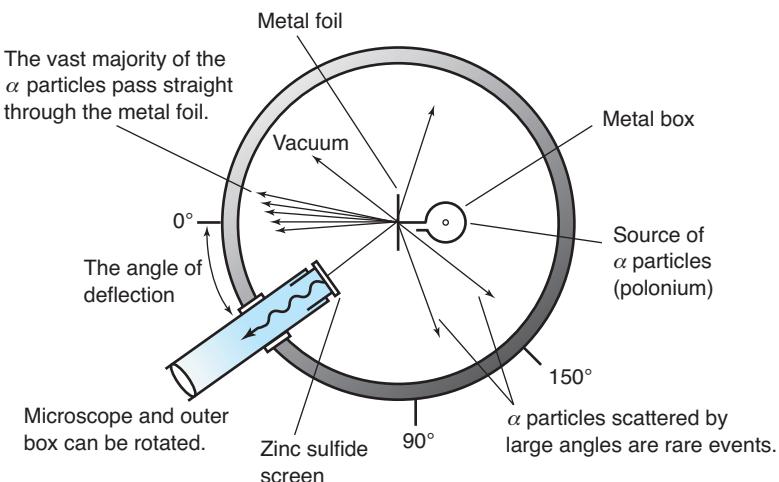


Figure 22.5 Deflection of alpha particles

Geiger made observations of the number of alpha particles scattered through different angles and his results confirmed the predictions of Rutherford. The distribution of the scattered alpha particles confirmed that the scattering was caused by an inverse square force, which Rutherford took to be the electrostatic force.

As the thickness of the gold foil was increased, the number of alpha particles scattered first increased but then remained constant. This confirmed that the alpha particles were in fact interacting with the atoms of gold. The scattering of alpha particles through small angles could be accounted for by the alpha particles undergoing a large number of interactions with different atoms, each interaction contributing a small amount to the total scattering. Even scattering at about 90° could be accounted for by multiple scattering. However, the probability of multiple scattering producing deflections of more than 90° (see figure 22.5) was so small that Rutherford concluded that the deflection must be due to the encounter of the alpha particle with a single atom.

Once it was established that the deflection was due to an encounter with a single atom, Rutherford showed that the charge that caused the deflection must be concentrated in a region about 10 000 times smaller than the radius of the atom. The alpha particles, which were known to have a velocity of about $1.6 \times 10^7 \text{ m s}^{-1}$, would penetrate to within $3 \times 10^{-12} \text{ cm}$ of the centre of the atom before being turned back. He concluded that most of the mass and positive charge of the atom must be concentrated in a very small nucleus. Rutherford's model of the atom is shown in figure 22.6.

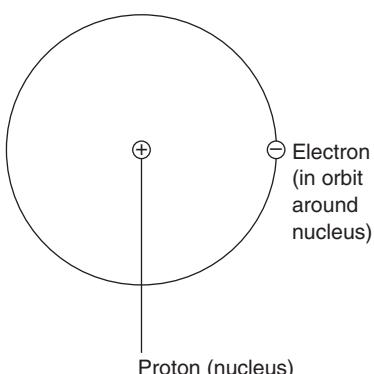


Figure 22.6 Diagram of the Rutherford model of the atom of hydrogen

PHYSICS FACT

A scale model of a hydrogen atom?

The radius of a hydrogen atom (in its first excited state — see page 433) is about $2.1 \times 10^{-10} \text{ m}$. The radius of a proton is about 0.85 femtometres ($0.85 \times 10^{-15} \text{ m}$). Physicists sometimes call this unit a fermi, named after Enrico Fermi (see chapter 24, page 462–463). The ratio of the radius of this atom to the radius of its nucleus is about 2.5×10^5 . This would make it very difficult to construct an accurate scale model of an atom in your laboratory. If your laboratory was 10 m across and this represented the diameter of the atom, the diameter of the nucleus would have to be $4 \times 10^{-5} \text{ m}$ or 40 microns in diameter.

Electrons in the Rutherford atom

When Rutherford published his paper ‘The scattering of alpha particles by matter and the structure of the atom’ (1911) he stated: ‘It will be shown that the main deductions from the theory are independent of whether the central charge is supposed to be positive or negative. For convenience, the sign will be assumed to be positive. The question of the stability of the atom proposed need not be considered at this stage, for this will obviously depend upon the minute structure of the atom, and on the motion of the constituent parts ...’

Rutherford knew that if the electrons were in orbit around the nucleus, they would be accelerating. They would be expected to be emitting electromagnetic radiation in accordance with the theories of Maxwell and, of course, the atom would be unstable.

22.2

BOHR'S MODEL OF THE ATOM

Before moving to Manchester to work with Rutherford, Bohr had worked for a short time at the Cavendish Laboratory under J. J. Thomson. Bohr and Thomson did not get along. At their first meeting, Bohr informed Thomson that one of Thomson's equations was wrong. There were several similar incidents and Bohr later recalled his disappointment that Thomson was not interested to learn that his work was incorrect. Bohr did acknowledge that his own lack of knowledge of the English language contributed to the failure of the two men to hit it off.

Niels Bohr (1885–1962), a Danish physicist, was one of eleven Nobel prize winners who were trained by Rutherford. One of Bohr's first contributions was to predict that a hydrogen atom would contain only one electron outside the positively charged nucleus. (At the same time, others predicted that one-electron atoms could not exist.)

Planck, Einstein and 'quantised energy'

Bohr attempted to apply the new quantum ideas of Planck and Einstein to the model of the hydrogen atom. As we saw in chapter 11 (page 201), Planck had managed to find an equation that solved the problem of the ‘ultraviolet catastrophe’ that troubled the theory of black-body radiation. Planck held a traditionalist's view of physics and was opposed to the statistical processes of Boltzmann. After attempting to explain his black-body equation, Planck reluctantly tried to derive it using the methods of Boltzmann. This involved dividing the energies up into small amounts and eventually should have finished with an integration in which all the energies would have been added together and would have experienced the problem of an infinite energy. However, before that final step, Planck realised that he had reached his equation for black-body radiation and therefore did not ‘complete’ the process. Einstein later showed that the problem of infinities will occur in any process where ‘classical’ theories and quantum theories are linked.

Planck interpreted his result as meaning that the ‘atomic oscillators’ that produced the radiation could vibrate only with certain discrete amounts of energy. These discrete amounts of energy were called **quanta**.

A **quantum** (plural: **quanta**) of energy can be considered to be the smallest amount of energy possible in a given situation. Planck's atomic oscillators could oscillate only with certain precise amounts of energy.



Figure 22.7 Niels Bohr

A **photon** is the quantum of electromagnetic radiation which exhibits both a particle and wave nature.

An **empirical equation** is one that has no theoretical basis but can be used to calculate correct values. Kepler's Third Law, $T^2 \propto R^3$, which you encountered in 'The Cosmic Engine', is another example of an empirical equation.

Einstein later extended this idea to the radiation itself being quantised. Einstein's 'quanta of light' were later named '**photons**' by Gilbert Lewis (1875–1946).

Bohr uses quantum theory to explain the spectrum of hydrogen

Bohr knew that, somehow, atoms must produce radiation that formed a characteristic spectrum for each element (see Physics in focus on 'The spectra of gases'). Bohr realised that the 'atomic oscillators' of Planck were probably electrons in the atom. The Rutherford model failed to provide any information about the radius of the atom or the orbital frequencies of the electrons. Bohr attempted to introduce the quantum ideas of Planck to the atom, but at first failed.

Early in 1913, Bohr was introduced to Balmer's equation (see below) for the wavelengths of the spectral lines of hydrogen and it 'made everything clear to him'. After seeing this equation, Bohr realised how electrons were arranged in the hydrogen atom and also how quantum ideas could be introduced to the atom.

PHYSICS FACT

Balmer's equation

Johann Jakob Balmer (1825–1898) completed a PhD in mathematics in 1849. He became a teacher at a girls' school in Basel, Switzerland, and had a desire to 'grasp the harmonic relationships of nature and art numerically'. Anders Angström (1814–1874) had measured the wavelengths of four of the spectral lines of hydrogen (now known as the Balmer series). Balmer found an equation that enabled him to calculate the wavelengths of these and, he believed, the infinite number of spectral lines emitted by hydrogen.

Balmer's equation was $\lambda = b \left(\frac{n^2}{n_f^2 - 2^2} \right)$ and the

constant b was found empirically to be 364.56 nm.

Janne Rydberg (1854–1919) modified Balmer's equation for wavelength to produce the familiar equation:

$$\frac{1}{\lambda} = R_H \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$$

where

λ = wavelength of the emitted radiation
 R_H = Rydberg's constant ($R_H = 1.097 \times 10^7 \text{ m}^{-1}$)
 n_f and n_i are integers.

The wavelengths of the visible lines of hydrogen correspond to $n_f = 2$ and $n_i = 3, 4, 5$ or 6. Of course, this is an **empirical equation** (Balmer played around with numbers until he arrived at something that worked).

Sometimes the equation $\frac{1}{\lambda} = R_H \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$ is

known as the Rydberg equation. Sometimes it is called Balmer's equation. Rydberg had attempted to find his own equation for the spectral lines of hydrogen. He was unsuccessful and, as his contribution was to modify Balmer's equation, we will continue to refer to it as the Balmer equation.



Figure 22.8 Johann Jakob Balmer

Calculating the wavelengths of hydrogen spectral lines

Calculate the wavelength of the visible spectral line of hydrogen with the longest wavelength.

SOLUTION

From Balmer's equation $\frac{1}{\lambda} = R_H \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$, we can see that the longest wavelength will occur when the term $\left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$ is smallest.

As the visible spectral lines correspond to $n_f = 2$, the smallest value will be when $n_i = 3$.

$$\begin{aligned}\frac{1}{\lambda} &= R_H \left(\frac{1}{2^2} - \frac{1}{3^2} \right) \\ &= 1.097 \times 10^7 \left(\frac{1}{2^2} - \frac{1}{3^2} \right) \\ &= 1.524 \times 10^6 \\ \lambda &= 6.562 \times 10^{-7} \text{ m}\end{aligned}$$

The wavelength is 6.562×10^{-7} m. This is the wavelength of the red line in the hydrogen spectrum in figure 22.9.

PHYSICS IN FOCUS*The spectra of gases*

There are three types of emission spectra: continuous spectra, bright-line spectra and band spectra. Continuous spectra are produced by incandescent objects, bright-line spectra are produced by excited gases and band spectra are produced by excited molecules. We will consider the bright-line emission spectra of excited gases and also the absorption spectra of cool gases (as shown in figure 22.9).

Spectral lines are produced as images of the slit that is an essential component of any spectroscope. After passing through the slit, the different wavelengths of light are diffracted by different amounts by a grating or dispersed by a prism by different amounts. Hence, the images of the slit corresponding to the different wavelengths are separated. When the slit is very narrow, closely spaced lines can be resolved (distinguished from one another). If the slit is wider, more light is admitted, but at the expense of the resolution.

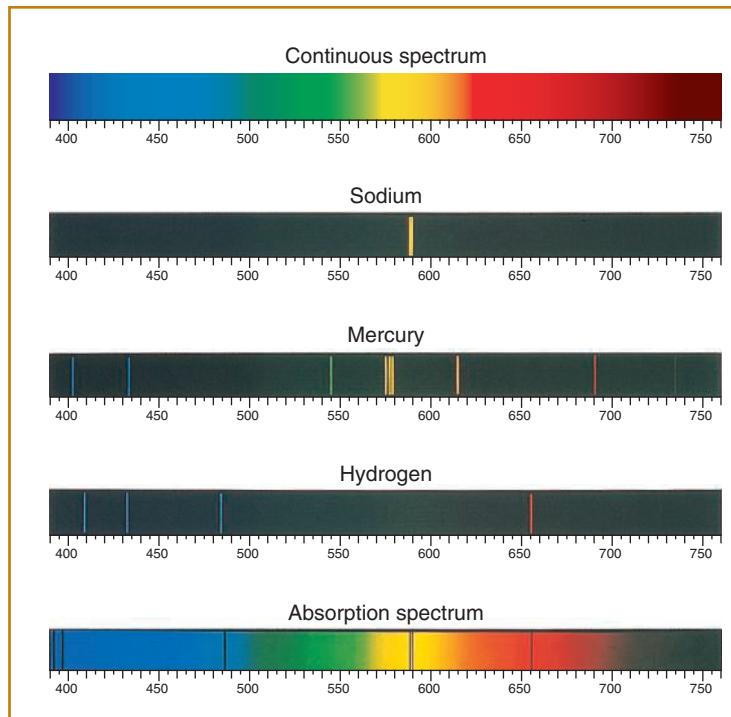


Figure 22.9 A continuous white light spectrum, the emission spectra of excited atoms of the elements sodium, mercury and hydrogen, and an absorption spectrum. The red line in the hydrogen spectrum is known as the H_α line, and the other lines as H_β , H_γ and H_δ respectively.

(continued)

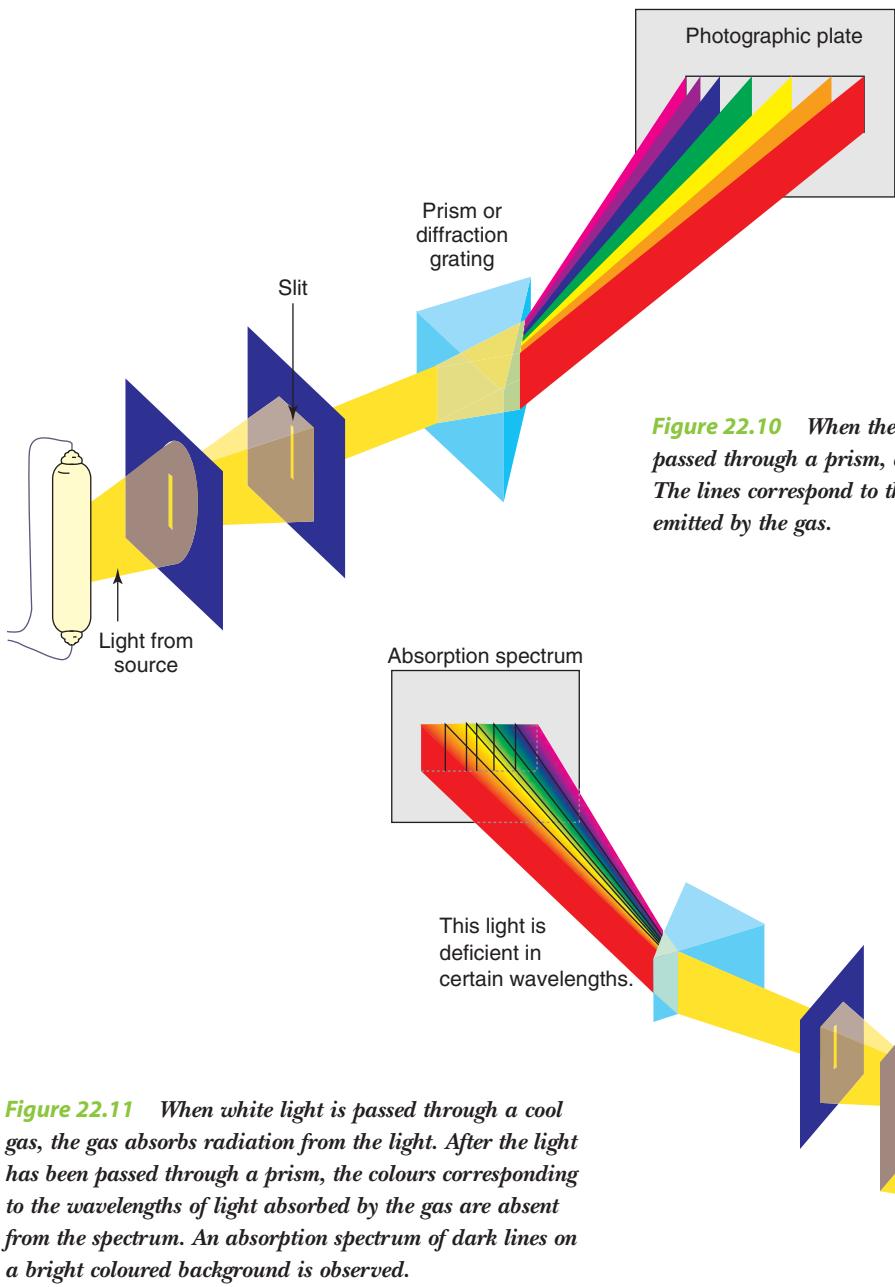


Figure 22.11 When white light is passed through a cool gas, the gas absorbs radiation from the light. After the light has been passed through a prism, the colours corresponding to the wavelengths of light absorbed by the gas are absent from the spectrum. An absorption spectrum of dark lines on a bright coloured background is observed.

An **emission spectrum** is a series of brightly coloured lines on a dark background that is produced when light from an excited gas is viewed through a spectroscope.

An **absorption spectrum** is a series of dark lines on a coloured background that is produced when white light is passed through a cool gas and viewed through a spectroscope.

An **emission spectrum** (see figure 22.10) is produced when a gas is excited. A gas can be excited by heating it or by passing an electrical discharge through it. The emission spectrum is a series of narrow coloured lines on a dark background. Each element has its own characteristic spectrum and this can be used to identify the gas.

An **absorption spectrum** (see figure 22.11) is produced when white light is passed through a cool gas. The atoms in the gas absorb energy from the white light. The atoms will then re-emit the energy that was absorbed. The energy will be emitted as light and it will be emitted in random directions. Therefore, the transmitted beam of light will be deficient in light at those energies or wavelengths. When this light is analysed, it will show a continuous spectrum of the white light with a series of narrow dark lines across it.



22.1

The spectrum of hydrogen

eBook plus

Weblink:
Spectra

PHYSICS IN FOCUS

Observing spectra using a simple spectroscope

Small, hand-held, direct vision spectrosopes can be used to examine spectra. However, if accurate measurements of wavelength are needed, a spectrometer that incorporates a collimator, prism or diffraction grating and telescope (see figure 22.12) is required. The method for using such a spectroscope is given in practical activity 22.1 (page 437).

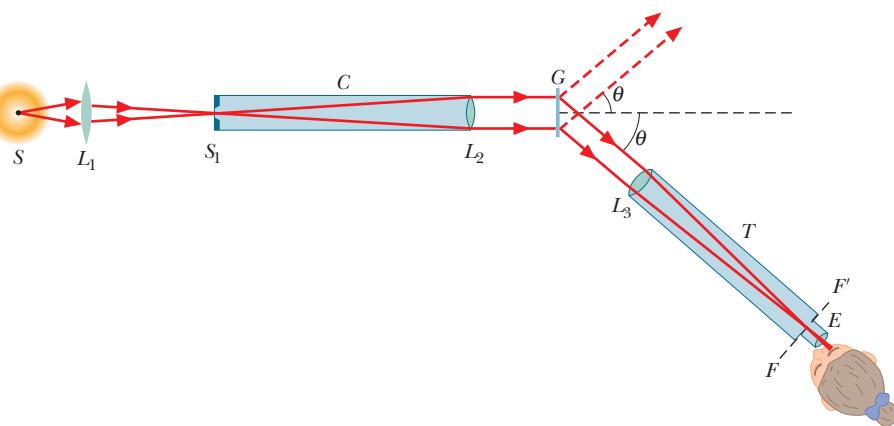


Figure 22.12 A simple spectroscope or spectrometer. The light enters the collimator, C, which focuses parallel light rays onto the prism or grating, G. The telescope, T, which has been focused for parallel light rays, is used to observe the dispersed or diffracted light. The telescope can be rotated and accurate measurements can be made of the angle through which the light has been deviated. This enables the wavelengths of the spectral lines to be calculated.

A spectacular example of emission from excited atoms occurs in the production of an aurora (figure 22.1). In this case the colours are produced by excited atoms or ions present in the atmosphere above about 60 km. Neutral oxygen atoms can produce pink and green colours, nitrogen molecules produce a red-violet colour and nitrogen ions can produce blue-violet. The atoms or ions are excited by interactions with charged particles from the Sun, usually after intense solar activity. Auroras do not usually extend very far north but occasionally they are visible from New South Wales.

22.3 BOHR'S POSTULATES

While Bohr believed that he knew the arrangement of electrons, he could not explain why the electrons were arranged in this way. Bohr published three papers between April and August 1913. In these papers, known as the great trilogy, he started with the problem of electrons in the Rutherford model and pointed out that the accelerating electrons must

lose energy by radiation and collapse into the nucleus. He then applied quantum theory to the atom. He generally assumed that the orbits of the electron were circular.

Bohr was awarded the Nobel Prize for Physics in 1922 and in his Nobel lecture stated in reference to Rutherford's discovery of the nucleus:

'This discovery made it quite clear that by classical conceptions alone it was quite impossible to understand the most essential properties of atoms. One was therefore led to seek for a formulation of the principles of the quantum theory that could immediately account for the stability in atomic structure and the radiation sent out from atoms, of which the observed properties bear witness. Such a formulation I proposed [1913] in the form of two postulates.'

Bohr continued with rather lengthy statements of his postulates. Simpler statements are:

1. Electrons in an atom exist in '**stationary states**' in which they possess an unexplainable stability. Any permanent change in their motion must consist of a complete transition from one stationary state to another.
2. In contradiction to the classical electromagnetic theory, no radiation is emitted from an atom in a stationary state. A transition between two stationary states will be accompanied by emission or absorption of electromagnetic radiation (a photon). The frequency, f , of this photon is given by the relation:

$$hf = E_1 - E_2$$

where

h = Planck's constant

E_1 and E_2 = values of the energy of two stationary states that form the initial and final states of the atom.

The **angular momentum**, L , of a point mass, m , which is in circular motion of radius, r , with velocity, v , is given by:

$$L = mvr$$

Angular momentum is the rotational equivalent to linear momentum and is an important quantity in rotational motion. (It follows a similar conservation principle to linear momentum.)

Many famous physicists had addressed the problem of electrons being in non-uniform motion without radiating energy. This had become important after Thomson had discovered the electron and was not just associated with the Rutherford model.

Bohr then introduced what is generally known as his quantisation condition and is sometimes called his third postulate.

An electron in a stationary state has an **angular momentum** that is an integral multiple of $\frac{h}{2\pi}$ (Planck's constant divided by 2π).

Bohr actually proposed that the kinetic energy of an electron was $\frac{n}{2h^2}$ but this reduces to the quantisation condition given if the orbits are circular.

In his first postulate, Bohr put forward one of the most audacious hypotheses ever proposed in physics by predicting that electrons exist in states in which they do not radiate energy. The second postulate involves the quantum of energy being emitted or absorbed when an electron jumps from one stationary state to another and hence explains the origin of spectral lines. The quantisation condition is really an intuitive guess.

Using these postulates together with the energy of electrons calculated from 'classical' physics applied to the Rutherford model, it is possible to derive a theoretical equation for the wavelengths of the spectral lines of hydrogen. It is a great success of the Bohr model that this theoretical equation is the same as the empirical equation of Balmer.

22.4 MATHEMATICS OF THE RUTHERFORD AND BOHR MODELS

In the following sections we will derive an expression for the classical energy of the Rutherford hydrogen atom and then impose Bohr's postulates on that atom. This will enable us to calculate the energies of the stationary states of the hydrogen atom and then calculate the change in energy of an electron involved in a transition between two stationary states. Finally, this change in energy will enable us to calculate the frequency (or wavelength) of the spectral lines of hydrogen.

The 'classical' energy of the Rutherford hydrogen atom

When you studied the escape velocity of an object fired from the Earth, you found the total energy of the object was the sum of its kinetic energy and its gravitational potential energy. When this total energy was zero, the object had just enough energy to escape from the Earth. If the total energy was negative, the object was unable to escape the Earth.

In a similar way we can calculate the total energy of a proton and electron. This time it is the sum of the kinetic energy and the electrical potential energy. The zero point will be when the electron has just enough energy to escape from the proton.

Kinetic energy of electron:

$$E_k = \frac{1}{2} m_e v^2.$$

The electron is held in orbit around the proton by the electrical force of magnitude:

$$F = \frac{k q_e^2}{r^2}$$

where

q_e = magnitude of the charge on the proton and electron (1.602×10^{-19} C).

We know that this electrical force provides the centripetal force of magnitude:

$$F_c = \frac{m_e v^2}{r}$$

$$F_c = F_E$$

$$\frac{m_e v^2}{r} = \frac{k q_e^2}{r^2}$$

$$\frac{1}{2} \frac{m_e v^2}{r} = \frac{1}{2} \frac{k q_e^2}{r^2}$$

$$\frac{1}{2} m_e v^2 = \frac{1}{2} \frac{k q_e^2}{r}$$

$$E_k = \frac{1}{2} \frac{k q_e^2}{r}.$$

The potential energy of the electron is given by:

$$E_p = -\frac{k q_e^2}{r}.$$

The total energy is the sum of the kinetic and potential energies.

$$\begin{aligned}\text{Total energy} &= E_k + E_p \\ &= \frac{1}{2} \frac{k q_e^2}{r} - \frac{k q_e^2}{r} \\ &= -\frac{1}{2} \frac{k q_e^2}{r}\end{aligned}$$

This is the total ‘classical’ energy of Rutherford’s hydrogen atom.

Radii of the ‘stationary states’ of the Bohr hydrogen atom

When Bohr’s quantisation condition is applied to the ‘classical’ hydrogen atom, the electron is restricted to stationary states in which the angular momentum of the electron is an integer multiple of Planck’s constant, divided by 2π .

$$\begin{aligned}\text{Angular momentum} &= \frac{n\hbar}{2\pi} \\ m_e v r &= \frac{n\hbar}{2\pi}\end{aligned}$$

The value of n for each stationary state or orbit of the Bohr atom is called the **principal quantum number** of that stationary state or orbit.

In this equation, n is an integer, known as the **principal quantum number**. We can obtain an expression for the radius of the stationary states corresponding to each value of the integer, n :

$$\begin{aligned}m_e v r &= \frac{n\hbar}{2\pi} \\ r &= \frac{n\hbar}{2\pi m_e v} \\ r^2 &= \frac{n^2 \hbar^2}{4\pi^2 m_e^2 v^2}.\end{aligned}$$

From the earlier equation, $\frac{m_e v^2}{r} = \frac{k q_e^2}{r^2}$, we can obtain an expression for v^2 :

$$v^2 = \frac{k q_e^2}{m_e r}.$$

Substituting this gives:

$$\begin{aligned}r^2 &= \frac{n^2 \hbar^2}{4\pi^2 m_e^2 \frac{k q_e^2}{m_e r}} \\ r_n &= \frac{n^2 \hbar^2}{4\pi^2 m_e k q_e^2}\end{aligned}$$

where

r_n = the radius of the stationary state corresponding to the integer n .

The radius of the stationary state corresponding to $n = 1$ will be:

$$\begin{aligned}r_1 &= \frac{1^2 \hbar^2}{4\pi^2 m_e k q_e^2} \\ &= \frac{\hbar^2}{4\pi^2 m_e k q_e^2}.\end{aligned}$$

We can combine the expressions for r_n and r_1 to give $r_n = n^2 r_1$.

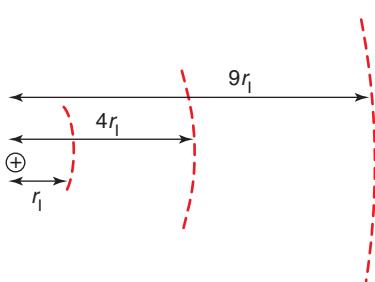


Figure 22.13 The relative radii of the orbits of an electron in different stationary states in a hydrogen atom

Energies of the 'stationary states' of the Bohr atom

If we now return to the classical energy of the Rutherford hydrogen atom (energy = $-\frac{1}{2} \frac{kq_e^2}{r}$) and impose the restriction that the only possible energies correspond to values of radius given by $r_n = \frac{n^2 h^2}{4\pi^2 m_e k q_e^2}$, we can calculate the value of these energy states:

$$\begin{aligned} E_n &= -\frac{1}{2} \frac{\frac{kq_e^2}{n^2 h^2}}{4\pi^2 m_e k q_e^2} \\ &= -\frac{1}{2} \frac{4\pi^2 k^2 m_e q_e^4}{n^2 h^2} \\ &= -\frac{1}{n^2} \left(\frac{2\pi^2 k^2 m_e q_e^4}{h^2} \right). \end{aligned}$$

Again we can see that:

$$E_1 = -\left(\frac{2\pi^2 k^2 m_e q_e^4}{h^2} \right)$$

and hence

$$E_n = \frac{1}{n^2} E_1$$

remembering that E_1 has a negative value.

SAMPLE PROBLEM

22.2

Calculating the energies of electrons in the hydrogen atom

Given that the energy of an electron in the first stationary state of hydrogen is $E_1 = -2.179 \times 10^{-18}$ J, determine the energy in electron volts (eV) of an electron in the following stationary states of the hydrogen atom:

- (i) the first stationary state ($n = 1$)
- (ii) the second stationary state ($n = 2$)
- (iii) the tenth stationary state ($n = 10$).

SOLUTION

- (i) We have been given this energy in joules so it is only a matter of converting to electron volts:

$$\begin{aligned} 1 \text{ eV} &= 1.602 \times 10^{-19} \text{ J} \\ 2.179 \times 10^{-18} \text{ J} &= \frac{2.179 \times 10^{-18}}{1.602 \times 10^{-19}} \text{ eV} \\ &= 13.60 \text{ eV}. \end{aligned}$$

The energy of the first stationary state is -13.6 eV.

- (ii) The energy of an electron in the second stationary state, for which $n = 2$ is given by:

$$\begin{aligned} E_n &= \frac{1}{n^2} E_1 \\ E_2 &= \frac{1}{2^2} E_1 \\ &= \frac{-13.6}{4} \\ &= -3.4 \text{ eV}. \end{aligned}$$

- (iii) The energy of an electron in the tenth stationary state, for which $n = 10$ is given by:

$$\begin{aligned}E_n &= \frac{1}{n^2} E_1 \\E_{10} &= \frac{1}{10^2} E_1 \\&= \frac{-13.6}{100} \\&= -1.36 \times 10^{-1} \text{ eV.}\end{aligned}$$

Electron volt

When an electron gains energy as it is accelerated across a potential difference of V volts, its gain in energy is given by $W = q_e V$.

When the electron is accelerated across a potential difference of 1.0 V, it will gain energy equal to $1.602 \times 10^{-19} \times 1.0 = 1.602 \times 10^{-19} \text{ J}$.

The gain in energy of an electron accelerated across a potential difference of 1.0 V is also called 1.0 electron volts (eV). $1.0 \text{ eV} = 1.602 \times 10^{-19} \text{ J}$.

Theoretical expression for wavelengths of the spectral lines of hydrogen

We are able to combine the expression for the energies of the stationary states with Bohr's second postulate to derive an expression for the energy differences between stationary states and, hence, the energies of the photons that may be emitted or absorbed by hydrogen.

We will consider the emission of a photon as an electron jumps from a higher energy initial state, E_i , to a lower energy final state, E_f .

The change in energy of the electron is:

$$\begin{aligned}\Delta E &= E_i - E_f \\&= \frac{1}{n_i^2} E_1 - \frac{1}{n_f^2} E_1 \\&= E_1 \left(\frac{1}{n_i^2} - \frac{1}{n_f^2} \right).\end{aligned}$$

This is the energy of the emitted photon, hf .

We can now derive an expression for the frequency and wavelength of the photon.

$$hf = E_1 \left(\frac{1}{n_i^2} - \frac{1}{n_f^2} \right)$$

$$f = \frac{-E_1}{h} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$$

$$\frac{c}{\lambda} = \frac{-E_1}{h} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$$

$$\frac{1}{\lambda} = \frac{-E_1}{hc} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$$

This equation is of the same form as Balmer's equation. If the value of $\frac{-E_1}{hc}$ is calculated, it agrees with the value of the Rydberg constant in Balmer's equation. (Remember that E_1 is a negative quantity and, hence, $-E_1$ is positive.)

Balmer's equation is an empirical equation (see page 424). A theoretical equation derived from Bohr's model of the atom now agrees with the empirical equation. This is a major achievement and offers very strong support for the Bohr model.

Emission of photons from a hydrogen atom

- (a) Given that the energy of the first stationary state of hydrogen is -13.60 eV , calculate the energy of the fourth stationary state of the hydrogen atom.
- (b) Use this information to calculate the frequency of the photon emitted when an electron undergoes a transition from the state $n = 4$ to the state $n = 1$.
- (c) Calculate the wavelength of the radiation emitted.

SOLUTION

- (a) The energy of the fourth stationary state is:

$$\begin{aligned}E_n &= \frac{1}{n^2} E_1 \\E_4 &= \frac{1}{4^2} E_1 \\&= \frac{-13.6}{16} \\&= -0.85\text{ eV.}\end{aligned}$$

- (b) The energy emitted by the photon will be:

$$13.60\text{ eV} - 0.85\text{ eV} = 12.75\text{ eV.}$$

$$\text{Energy of photon} = 12.75 \times 1.602 \times 10^{-19} \text{ J}$$

$$\begin{aligned}f &= \frac{E}{h} \\&= \frac{12.75 \times 1.602 \times 10^{-19}}{6.626 \times 10^{-34}} \\&= 3.083 \times 10^{15} \text{ Hz}\end{aligned}$$

- (c) The wavelength will be calculated from:

$$\begin{aligned}\lambda &= \frac{c}{f} \\&= \frac{3.00 \times 10^8}{3.083 \times 10^{15}} \\&= 9.73 \times 10^{-8} \text{ m.}\end{aligned}$$

(Of course the wavelength could have been calculated directly from Balmer's equation.)

The hydrogen atom explained

We are now able to calculate the wavelengths of the many spectral lines of the hydrogen atom. The original series of spectral lines was known as the Balmer series and contained the four spectral lines in the visible region of the spectrum. These lines correspond to electron jumps to the second lowest energy state, or first excited state, ($n = 2$) of the hydrogen atom.

The wavelengths of the spectral lines in other series can be calculated using Bohr's equation and are shown in figure 22.14 on the following page. The Paschen series of infra-red lines had already been discovered but other series of lines were found later and their wavelengths were in agreement with Bohr's theory. The series of lines in the ultraviolet and infra-red, named after their discoverers, are:

- Lyman series, discovered in 1916. These were ultraviolet lines with transitions to the **ground state** ($n = 1$).
- Paschen series, discovered in 1908. These were infra-red lines with transitions to the second **excited state** ($n = 3$).
- Brackett series, discovered in 1922. These were infra-red lines with transitions to the third excited state ($n = 4$).
- Pfund series, discovered in 1924. These were infra-red lines with transitions to the fourth excited state ($n = 5$).

An electron has the lowest possible amount of energy when it is in the **ground state**.

If it exists in a stationary state in which it has more energy, it is said to be in an **excited state**.

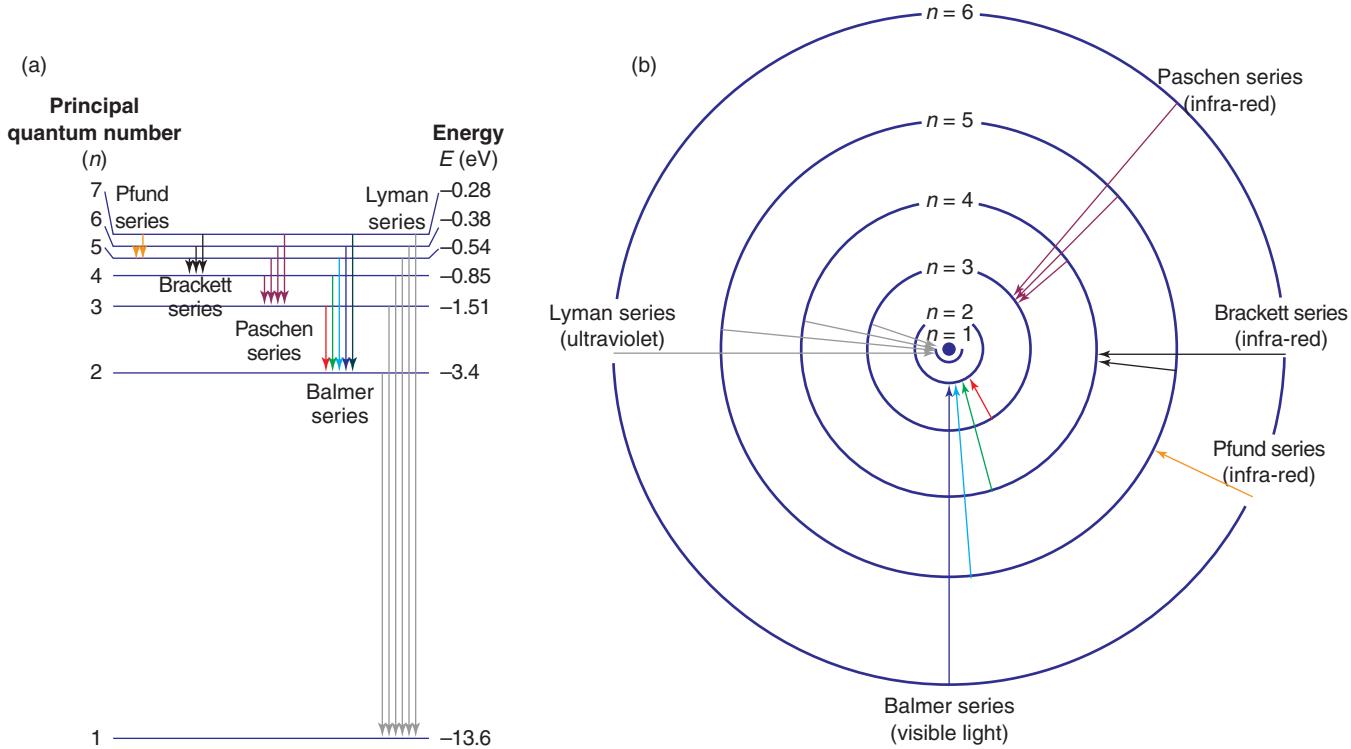


Figure 22.14 a) Atomic energy level view of the spectral series of hydrogen (b) Electron orbit view of the spectral series of hydrogen. Note that the radii of the orbits of the electrons are not to scale.

22.5 LIMITATIONS OF THE BOHR MODEL OF THE ATOM

The Bohr model takes the first step to introduce quantum theory to the hydrogen atom but it is only a first step. The model has the following limitations:

- it is not possible to calculate the wavelengths of the spectral lines of all other atoms
- the Bohr model works reasonably well for atoms with one electron in their outer shell but does not work for any of the others
- examination of spectra shows that the spectral lines are not of equal intensity but the Bohr model does not explain why some electron transitions would be favoured over others
- careful observations with better instruments showed that there were other lines known as the hyperfine lines. There must be some splitting of the energy levels of the Bohr atom but the Bohr model cannot account for this.
- when a gas is excited while in a magnetic field, the emission spectrum produced shows a splitting of the spectral lines (called the Zeeman effect). Again, the Bohr model cannot account for this.
- finally, the Bohr model is a mixture of classical physics and quantum physics and this, in itself, is a problem.

In the next chapter we will examine some of the observations that supported ideas from Bohr's model but also see that there was a need to break completely from classical physics and move to a new theory — quantum mechanics.

SUMMARY

- The scattering of alpha particles through large angles by very thin gold foils led Rutherford to propose that an atom consisted of a very small, dense, positively charged nucleus. Electrons were in orbit about the nucleus at distances very large compared to the dimensions of the nucleus.
- A major problem with the Rutherford model was that it did not account for any properties of the electrons in the atom, in particular how the electrons could be accelerating without emitting electromagnetic radiation.
- Bohr extended the Rutherford model by formulating two postulates that enabled him to apply the quantum ideas of Planck and Einstein to the Rutherford atom.
- Bohr's postulates enabled him to describe an atom in which electrons existed in stable 'stationary states' where they did not emit electromagnetic radiation. The transition of an electron from one stationary state to another would be accompanied by the emission or absorption of a quantum of electromagnetic radiation or a photon.
- Using his model of the atom, Bohr was able to derive a theoretical expression for the wavelengths of the spectral lines of hydrogen which was in agreement with Balmer's empirical formula.
- While successful in explaining the wavelengths of the spectral lines in the hydrogen spectrum, Bohr's model failed to account for the relative intensities of the lines, the existence of the hyperfine structure of the lines or for the splitting of spectral lines when the excited gas was in a magnetic field. Bohr's model was also a strange mixture of classical physics and quantum physics.

QUESTIONS

- Use Balmer's equation to calculate the wavelength of the radiation emitted from an excited hydrogen atom when an electron undergoes a transition from the state $n = 5$ to:
 - the state $n = 1$
 - the state $n = 2$
 - the state $n = 3$.
- (a) Calculate the wavelengths of the lines of the Balmer series corresponding to transitions from the states $n = 8$, $n = 10$, $n = 12$.
- (b) What trend do you notice in the wavelengths as the value of n increases?
- The radius of the orbit of an electron in the ground state of the hydrogen atom is 5.3×10^{-11} m. Calculate the radius of the orbit of an electron when it is in each of the following states:
 - the state $n = 2$
 - the state $n = 3$
 - the state $n = 4$.
- (a) State which photon, red or blue, has the higher frequency.
 (b) State which photon, red or blue, has the longer wavelength.
 (c) State which photon, red or blue, has the higher energy.
- If the atoms in a sample of hydrogen were all in the state $n = 5$, how many different spectral lines could possibly be produced by the gas as the electrons returned to the ground state?
- Given that $E_1 = -13.6$ eV, $E_2 = -3.40$ eV, $E_3 = -1.51$ eV, $E_4 = -0.85$ eV, $E_5 = -0.54$ eV, calculate the wavelengths of:
 - the first two lines in the Lyman series
 - the first two lines in the Balmer Series
 - the first two lines in the Paschen series.
- (a) What is the wavelength of the longest wavelength spectral line of the Pfund series?
 (b) What is the wavelength of the shortest wavelength line of the Pfund series?
- The 'series limit' is the term applied to the shortest wavelength spectral line in each of the spectral series of hydrogen.
 - What value of n_i would be used to calculate the wavelength of the series limit?
 - Calculate the series limit for the Lyman, Balmer and Paschen series of hydrogen.
 - How many electron volts of energy would be carried by a photon corresponding to the series limit of the Lyman series?
- Figure 22.15 is an energy level diagram for energies of the stationary states in atoms of a gas, Q.
 - (i) Determine the energy of the photon emitted when an electron in the state $n = 3$ undergoes a transition to the state $n = 2$.
 (ii) Determine the frequency and wavelength of this photon.

- (b) Determine the wavelength of the photon absorbed by this gas when an electron undergoes a transition from the state $n = 1$ to the state $n = 4$.

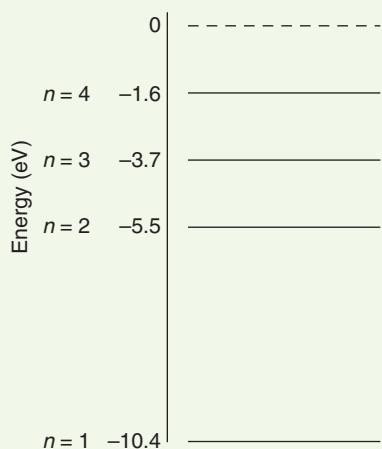


Figure 22.15 The energy level diagram for the gas, Q

10. The emission spectrum of a particular gas has eight bright lines in the visible region as shown in figure 22.16. The absorption spectrum of the same gas has only three lines in the visible region as shown.
- Explain why each of the absorption lines corresponds to one of the emission lines.
 - Explain why there is not a corresponding absorption line for five of the emission lines.

11. An absorption spectrum is produced when the atoms in a cool gas absorb energy from white light passing through the gas. These excited atoms then re-emit the energy and return to low energy states. How can this re-emission occur but there still be dark lines in the absorption spectrum?

- Balmer predicted accurately the wavelengths of the visible spectral lines and invisible spectral lines of hydrogen that had not been detected. Bohr did the same about thirty years later. Explain why Bohr's prediction is considered more important than that of Balmer.
- What evidence supports the idea that the electron energies in the hydrogen atom are discrete?
- If electrons in hydrogen atoms obeyed the rules of classical mechanics instead of those of quantum mechanics, would the hydrogen atoms produce a line spectrum or a continuous spectrum? Explain your answer.
- Explain why each element has its own characteristic spectrum.
- Two spectral lines of hydrogen have frequencies of 2.7×10^{14} Hz (infra-red) and 4.6×10^{14} Hz (red).
 - Explain how you could use this information to determine the frequency of a higher frequency spectral line of hydrogen.
 - Calculate the frequency of that line.

Emission spectrum

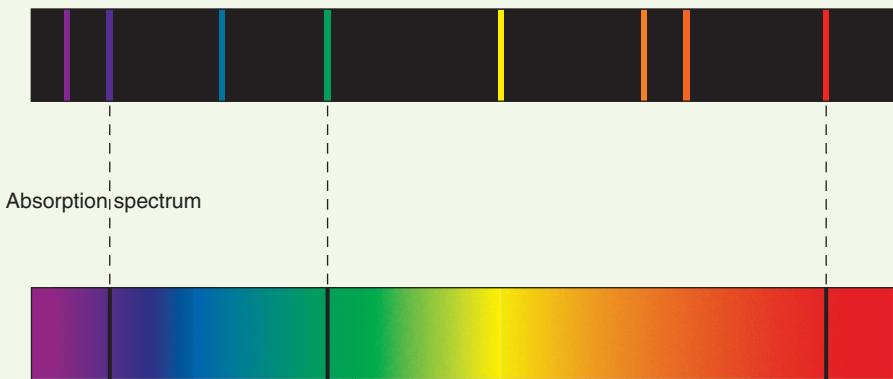


Figure 22.16 The emission and absorption spectra for a particular gas



22.1 THE SPECTRUM OF HYDROGEN

Aim

To observe the spectral lines of hydrogen, measure their wavelengths and compare these values with the theoretical values.

Apparatus

hydrogen spectral tube and power supply spectroscope

Theory

According to the Bohr model of the hydrogen atom, when an electron jumps from a higher energy state to a lower energy state, it will emit a photon. When an electron jumps to the state $n = 2$ from any of the states from $n = 3$ to $n = 6$, the emitted photon will be in the visible region of the spectrum.

A spectroscope can be used to measure the deviation of the spectral lines. The wavelengths of the spectral lines can then be calculated.

The wavelengths of the spectral lines will be given by:

$$\lambda = \frac{d \sin \theta}{n}$$

where

λ = wavelength

θ = angle of deviation

d = distance between lines on the grating

n = order of spectra.

The theoretical values of the wavelengths, based on Bohr's theory of the hydrogen atom, can be calculated after the energies of the states $n = 2$ to $n = 6$ have been calculated.

The energy of the ground state, $n = 1$ is -13.6 eV . The energies of the other states are given by

$$E_n = \frac{E_1}{n^2}.$$

Method

In this experiment, the hydrogen spectral tube is switched on and the radiation viewed through a spectroscope.

Setting up the hydrogen spectral tube

Different types of spectral tube and power supply may be used but we will describe a special power

supply that is designed for spectral tubes. The spectral tube can be clamped in place on a vertical metal rod mounted on top of the power supply. (The rod is maintained at Earth potential and is safe to touch when the power supply is switched on.)

Some hydrogen spectral tubes are very faint and this makes measurement of the spectral lines very difficult. Hydrogen spectral tubes should probably be replaced fairly regularly as they tend to become fainter over time.

The spectroscope

The spectroscope consists of two tubes, one of which can be rotated around a small central table. One tube, the fixed one, is a collimator and the moveable one is a telescope. A small prism or a diffraction grating can be mounted on the small table. Figure 22.12 (page 427) shows a diagram of a spectroscope.

There is an adjustable narrow slit at the front of the collimator. The collimator is set up to shine parallel rays of light onto the diffraction grating or prism. (For the remainder of this practical activity, we will assume that a diffraction grating is being used. We will assume that the information about the number of lines per metre is provided. It is possible to calibrate a grating, and a procedure to do this is included in the last section of the method.)

The light that passes through the diffraction grating deviates through an angle that depends on the wavelength of the light and the number of lines per metre ruled on the diffraction grating.

The telescope is rotated around the table and the image of the narrow slit is observed at different angles for the different wavelengths of light. These angles can be measured, usually with the help of a vernier scale fixed to the telescope.

Setting up the spectroscope

Setting up the spectroscope involves two parts; adjusting the telescope for parallel light rays and then adjusting the collimator to produce parallel light rays.

There should be fine cross-wires visible in the eyepiece of the telescope. These cross-wires should be in sharp focus and an adjustment of the eyepiece in its holder may have to be made if they are not sharply focused.

The telescope should be pointed at a distant object and the focus adjusted using the objective lens of the telescope, lens L_3 in figure 22.12, until the image of the distant object is sharply focused. (In fact any object outside should be far enough away.) The telescope is then aligned with the collimator and the lens on the collimator, lens L_2 in figure 22.12, is adjusted until the slit is seen sharply focused.

A light source, possibly a brighter spectral tube than the hydrogen tube, can now be set up in front of the slit and the slit width adjusted until narrow spectral lines can be viewed when the telescope is rotated to the appropriate position.

By clamping the telescope and then using the fine adjustment, it should be possible to align the cross-wires visible in the eyepiece with the spectral line. A measurement can then be made.

Reading a vernier scale

A vernier scale has ten lines on the moveable scale in the space of nine lines on the fixed scale. This enables an extra decimal place to be determined. This extra digit corresponds to the position of the line on the vernier, moveable scale that aligns with any one of the lines on the fixed scale.

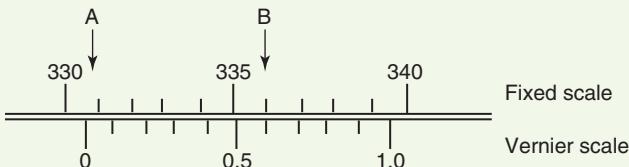


Figure 22.17 Reading a vernier scale. The position of the zero mark on the vernier scale, indicated by arrow A, is just less than 331. The line on the vernier scale that matches a line on the fixed scale is 0.6, as indicated by arrow B. Therefore, the reading is 330.6°.

Measuring the wavelengths of the spectral lines of hydrogen

The lines will probably be quite faint and it will probably be necessary to have the apparatus in a darkened room to observe the lines clearly. The most difficult part is aligning the spectral lines with the cross-wires. If the room is completely dark, it will be impossible to see the cross-wires. A small amount of field illumination is necessary to be able to see the cross-wires.

There should be no problem with making the measurement for the straight through position. There should be sufficient light coming directly through the slit to make locating the image of the slit on the cross-wires quite easy.

Record this value and then record the reading of as many of the spectral lines as possible. (If it is possible to measure any of the spectral lines of the second order spectrum it is worth doing so.)

Calibration of diffraction grating

If necessary, the diffraction grating could be calibrated using a sodium vapour spectral tube. Set up this tube and observe the angle to the very bright orange line in the first order spectrum of sodium

($n = 1$). This line is really a double line, the wavelengths of the lines being 589.0 nm and 589.6 nm.

You can use the information in the equation $\lambda = \frac{d \sin \theta}{n}$ to calculate d .

Results

Record your results in a table similar to the table below and calculate the wavelengths of the spectral lines.

The number of lines per centimetre or perhaps even the number of lines per inch is probably supplied with the diffraction grating. It will be necessary to convert this to lines per metre and d is the inverse of this value.

Record the reading of the straight through position θ_0 .

SPECTRAL LINE COLOUR	POSITION θ	ANGLE $\theta - \theta_0$	ORDER OF SPECTRA (n)	WAVELENGTH
Faint violet			1	
Violet			1	
Blue-Green			1	
Red			1	
			2	
			2	
			2	
			2	

Analysis

1. The energy of the ground state of hydrogen is $E_1 = -13.6$ eV.

The energies of the other states are given by

$$E_n = \frac{E_1}{n^2}.$$

Determine the energy, in electron volts, of the energy states $n = 2, 3, 4, 5$, and 6 .

2. Draw an energy level diagram and calculate the energies (in electron volts) of photons emitted when an electron jumps to the $n = 2$ state from each of the four higher energy states.
3. Convert these values from electron volts to joules and calculate the wavelengths of these photons.

Use:

$$E = hf = h \frac{c}{\lambda}$$

$$\lambda = \frac{hc}{E}$$

where

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$$

$$h = 6.602 \times 10^{-34} \text{ J s.}$$

4. Compare the values of the wavelengths calculated above with the values determined from the measurements of the angles.

Questions

1. How accurate do you consider your determination of the wavelengths of the spectral lines? Aside from any difficulty with aligning the spectral lines with the cross-wires, you are restricted to measuring the angle to the nearest 0.1° . Consider how a change in angle of 0.1° will alter your calculations.
2. Taking into account the expected accuracy of your observations, do you consider that your results are in agreement with the theoretical values of the wavelengths of these four spectral lines of hydrogen?

CHAPTER 23

DEVELOPMENT OF QUANTUM MECHANICS

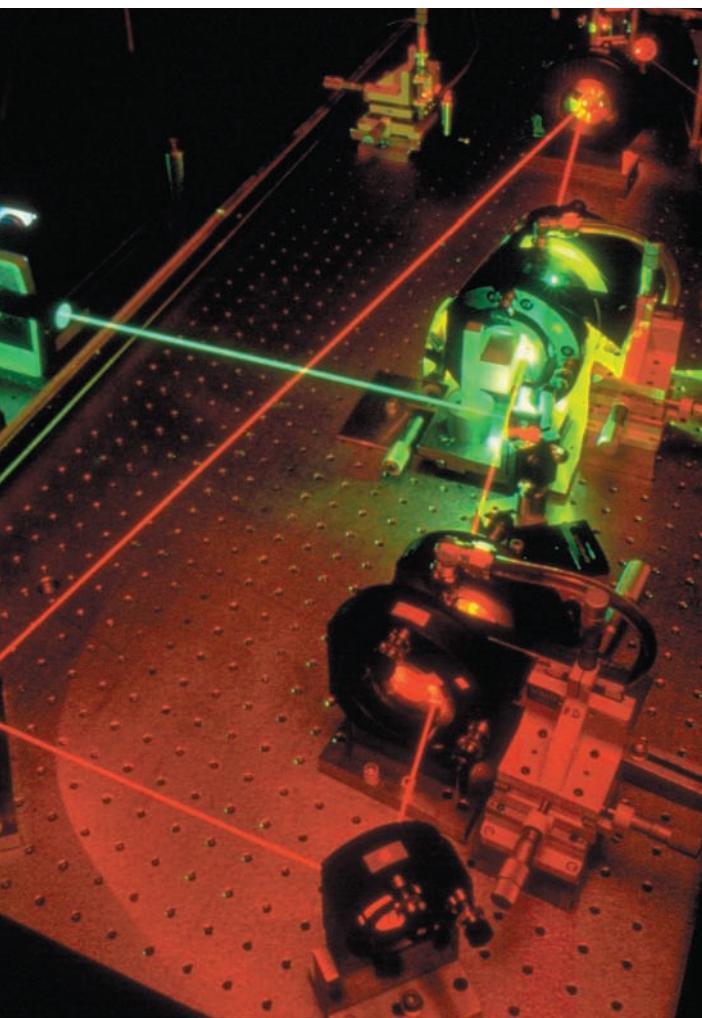


Figure 23.1 Neodymium, YAG (yttrium aluminum garnet), argon and dye lasers. Lasers are used in a very wide range of fields including communication, medicine, measurement, holography, entertainment and scientific research. Common devices such as CD players and supermarket scanners use laser technology. These are just two of the many applications in use today that are directly related to the theory of quantum mechanics.

Remember

Before beginning this chapter, you should be able to:

- recall Planck and Einstein's contributions to the early development of quantum theory
- recall the features of Bohr's model of the atom in which he introduced the ideas of quantum theory to atomic structure
- realise that there were difficulties with the Bohr model, not the least of these being that it mixed classical and quantum physics
- recall that photons of light possessed the nature of both waves and particles.

Key content

At the end of this chapter you should be able to:

- recognise that diffraction is a wave phenomenon and that a diffraction pattern is produced by the interference of diffracted waves
- describe de Broglie's proposal that matter has a wave nature as well as a particle nature
- solve problems using $\lambda = \frac{h}{mv}$
- describe the impact of de Broglie's proposal
- describe the experimental evidence provided by Davisson and Germer confirming the wave nature of electrons
- use de Broglie's matter waves to explain the stability of the stationary states of the hydrogen atom
- assess the contributions of Heisenberg and Pauli to the development of a quantum mechanical model of the atom.

*I like relativity and quantum theories
because I don't understand them
and they make me feel as if space shifted
about like a swan that can't settle,
refusing to sit still and be measured;
and as if the atom were an impulsive thing
always changing its mind.*

—D. H. Lawrence

As we saw at the end of the previous chapter, Bohr had taken the first steps in applying quantum ideas to atomic structure but there were problems associated with his model of the atom. Despite these problems, Bohr's model, which reached its peak in 1922, was able to explain the periodic table and make accurate predictions about the properties of then undiscovered elements (see page 447).

In the 1920s, there was still a perceived problem with the nature of light. In 1924, Einstein wrote: 'There are therefore two theories of light, both indispensable and — as one must admit today despite twenty years of tremendous effort on the part of theoretical physicists — without any logical connection.' (Reference from an article by Einstein in *Berliner Tageblatt*, 20 April 1924, quoted in Abraham Pais, *Inward Bound*.)

When Einstein made reference to the wave theory of light and the particle theory of light being without any logical connection, he was unaware of the predictions of Louis de Broglie that particle and wave natures were inextricably linked. Einstein was soon called on to make comment on de Broglie's doctoral thesis. In it, de Broglie predicted that not only did light have a dual wave and particle nature, but particles also had a wave nature. Einstein was impressed with de Broglie's 'crazy idea'.

Other famous physicists expressed discontent with the state of physics in the early 1920s. In 1924, Max Born wrote, 'At the most we possess only a few unclear hints'; and in 1925, Wolfgang Pauli wrote, 'Physics at the moment is very muddled.' The important breakthrough was supplied by Werner Heisenberg who devised his theory of matrix mechanics, later to be known as quantum mechanics.

Before we can study the wave nature of matter as predicted by de Broglie, we must first study diffraction, a phenomenon exhibited by waves, which was important in detecting the wave nature of particles.

23.1

DIFFRACTION

Diffraction of light occurs when light passes through a very finely ruled grating, or when it is reflected from a surface with fine lines ruled across it. It also occurs when light passes a barrier or passes through a small opening (see figure 23.2). It is not easy to observe because the dimensions of the barrier or opening must be comparable to the wavelength of light.

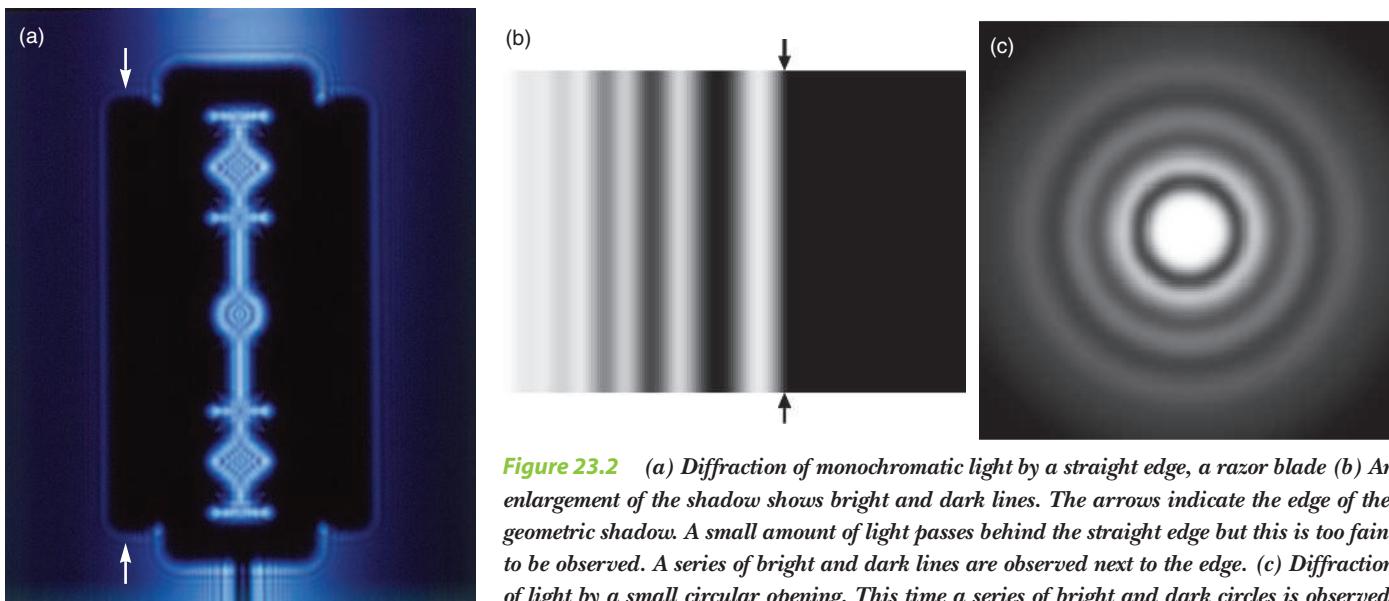


Figure 23.2 (a) Diffraction of monochromatic light by a straight edge, a razor blade (b) An enlargement of the shadow shows bright and dark lines. The arrows indicate the edge of the geometric shadow. A small amount of light passes behind the straight edge but this is too faint to be observed. A series of bright and dark lines are observed next to the edge. (c) Diffraction of light by a small circular opening. This time a series of bright and dark circles is observed.

An explanation of diffraction

A **wavefront** is either the crest or trough of a wave. The wavefront is perpendicular to the direction of the velocity of the wave.

In the seventeenth century, Christian Huygens proposed that light was a wave. He proposed what has become known as Huygens' Principle which states 'Every point on a **wavefront** may be considered to act as a source of circular secondary wavelets that travel in the direction of the wave. The new wavefront will be tangential to the secondary wavelets.' Huygen's Principle is shown in figure 23.3.

This principle can be used to derive the laws of reflection and refraction. It also helps to explain diffraction.

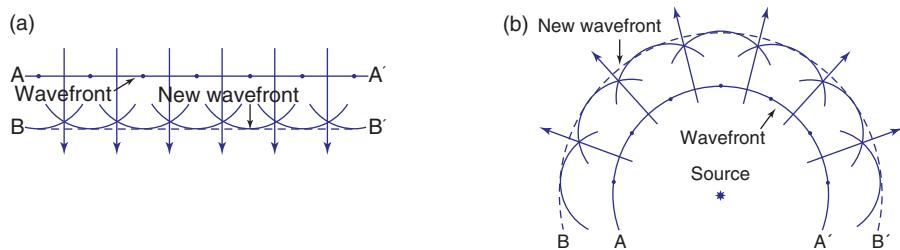
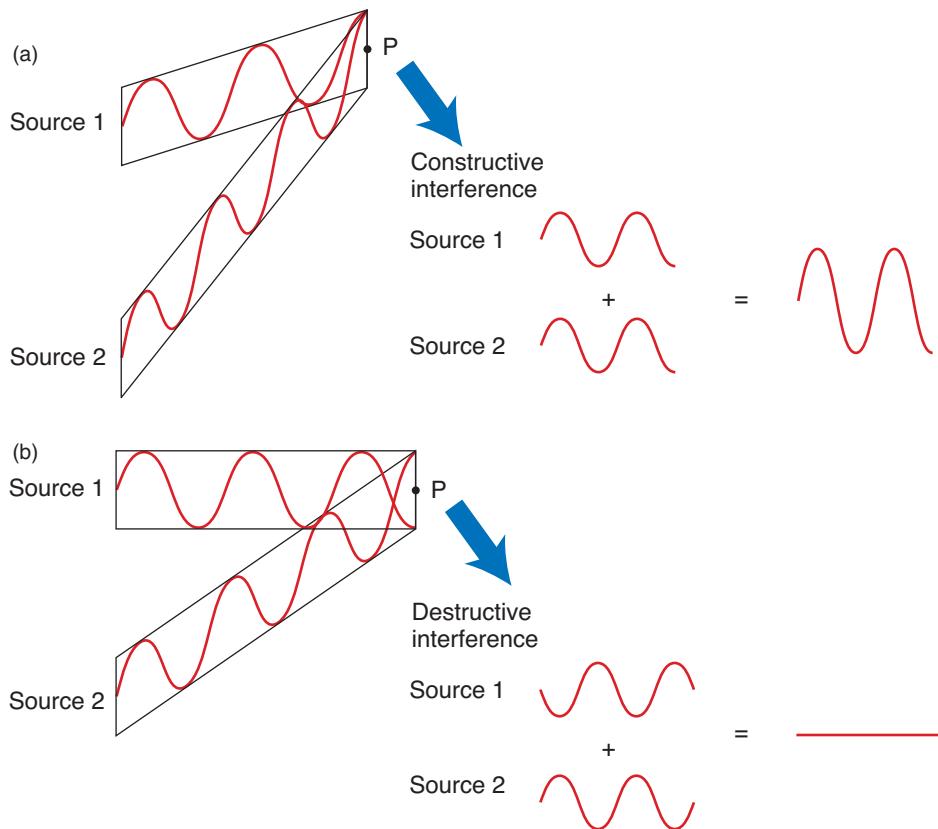


Figure 23.3 (a) Huygens' Principle explains the propagation of a plane wave. Each point acts as a point source and a new plane wave is formed. AA' is the original wavefront and the new wavefront is BB'. (b) Each point on the curved wavefront acts a point source and a new curved wavefront is formed.

If we use Huygen's Principle to explain the propagation of a wave (the formation of a new wavefront), it is necessary that all the point sources contribute to the production of the new wavefront. When a barrier blocks part of the wave, not all the point sources will be able to contribute to the new wavefront.

Figure 23.4 Interference of two identical waves (a) Where the crest of one wave meets the crest of another wave, constructive interference occurs and the resultant displacement will be twice that of one wave. (b) Where a crest from one source meets a trough from the other source, destructive interference occurs and there will be no displacement.



Interference of light was first demonstrated by Thomas Young in the early nineteenth century (see figure 23.5). He used two narrow slits as light sources and produced a pattern of bright and dark lines on a screen. The bright lines occurred at positions where waves met in phase (trough met trough or crest met crest), and the dark lines where the waves met out of phase (crest met trough).

If we consider a point on the new wavefront, some of the wavelets that were blocked would have interfered *destructively* with the wavelets that reach that point. Others would have interfered *constructively*. This effect is shown in figure 23.4. If the net effect was that more of the blocked wavelets would have interfered destructively than constructively, that point on the new wavefront will be of greater intensity than the incident wavefront. As we move further across the pattern, the relative amounts will change and we will observe successive regions that are brighter and then fainter than the incident light. The intensities gradually become closer to that of the incident light.

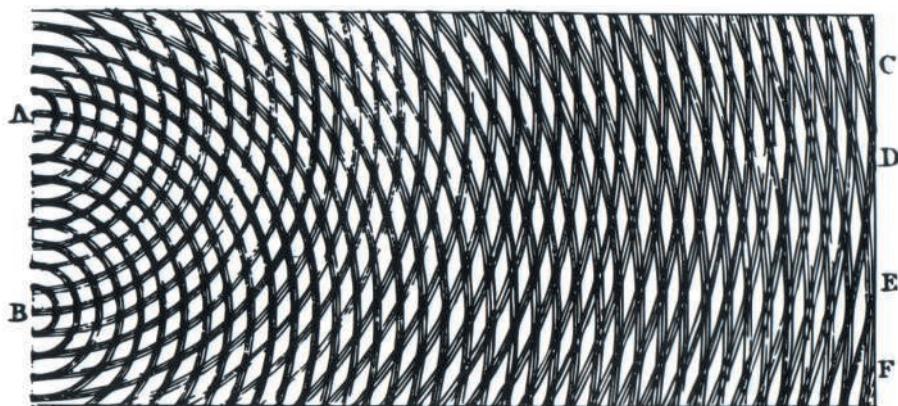


Figure 23.5 Thomas Young's original drawing of a two-source (double-slit) interference pattern. A and B are the point sources. The dark circles represent the wave crests and the troughs are the white spaces in between crests. Destructive interference occurs on the screen at C, D, E, and F where crest meets trough.

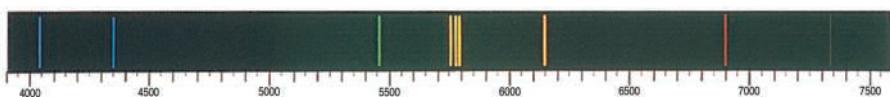
We can explain the light and dark regions of a diffraction pattern in terms of the constructive and destructive interference of wavelets from the point sources.

Figure 23.6 The pattern produced when light from a mercury vapour lamp is shone through a transmission diffraction grating

If light waves maintain a constant phase relationship, they are said to be **coherent**.

If we observe a straight edge or a narrow slit, we will notice bright and dark straight lines. If it is a circular aperture, we will see a series of bright and dark rings.

A transmission diffraction grating is a transparent material that has many fine lines ruled across it. (A grating may have many thousands of lines per centimetre.) These lines can be considered to be breaking the wavefront into point sources. The interference of light from these many point sources produces a diffraction pattern (see figure 23.6).



A reflection diffraction grating is a reflecting surface with many lines ruled across it. The 'reflected' light is not reflected with an angle of incidence equal to the angle of reflection because the gaps between the lines act as point sources.

If a laser, which is a source of **coherent** light, is shone onto part of the scale of a metal ruler at an angle of incidence close to 90° , the 'reflected' light will produce a series of bright spots, a diffraction pattern. This effect is shown in figure 23.7.

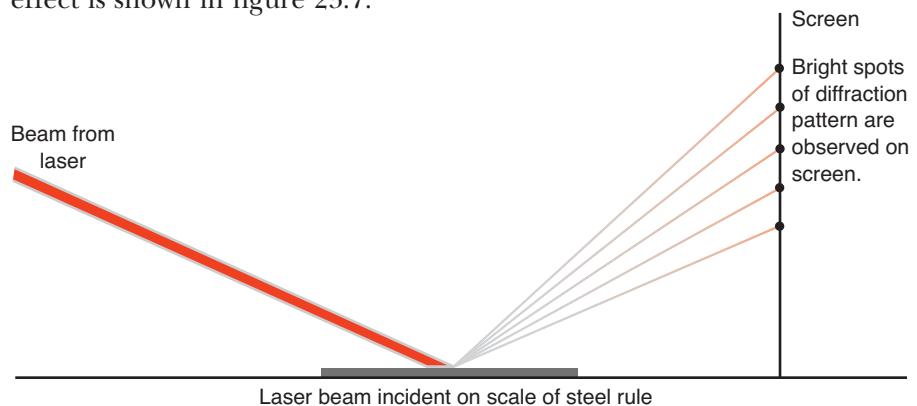


Figure 23.7 A diffraction pattern can be observed when light from a laser is reflected from the scale on a metal ruler.

23.2 STEPS TOWARDS A COMPLETE QUANTUM THEORY MODEL OF THE ATOM

At the end of chapter 22 we saw that Bohr's model of the atom did not explain certain aspects of the spectrum of hydrogen. It could not account for hyperfine spectral lines or the fact that spectral lines were split if the gas was placed in a magnetic field. However, perhaps the most puzzling aspect of the model was how Bohr could throw out some features of classical physics while retaining others and how he could impose quantum theory features on an otherwise classical model.

Louis de Broglie's proposal that particles had a wave nature

The correct pronunciation of de Broglie is 'de broy'.

Louis de Broglie (1892–1987) was a French nobleman who had the misfortune to have his studies of physics interrupted in 1913 by what should have been a short period of compulsory military service. Because of World War I, his military service continued until 1919.

He did, however, return to his studies and, in 1923, published three short papers on light quanta. He then prepared his doctoral thesis. In it he argued that the fact that nobody had managed to perform an experiment that settled once and for all whether light was a wave or a particle was because the two kinds of behaviour are inextricably linked.

The expressions for the energy and momentum of light quanta:

$$E = hf \text{ and } p = \frac{hf}{c}$$

have quantities that are properties of particles on the left-hand side and quantities that are properties of waves on the right-hand side.

De Broglie made the bold proposal that all particles must have a wave nature as well as a particle nature.

Electrons had been thought of as well-behaved particles except for the fact that they occupied distinct energy states in the hydrogen atom. These energy states were associated with integers. De Broglie was aware of other phenomena in physics that were associated with integers. These included the interference of waves and the vibration of standing waves. He stated: 'This fact suggested to me the idea that electrons, too, could not be regarded simply as corpuscles, but that periodicity must be assigned to them.'

His work was not just idle speculation. His great achievement was to take this idea and develop it mathematically. He described how matter waves ought to behave and suggested ways that they could be observed.



Figure 23.8 Prince Louis Victor de Broglie, who was awarded the Nobel Prize in 1929 for his discovery of the wave nature of electrons

$$E = hf \text{ and } E = mc^2$$

$$mc^2 = hf$$

$$mc = \frac{hf}{c}$$

$$\therefore p = \frac{hf}{c}$$

The wavelength of a photon was Planck's constant divided by its momentum and de Broglie proposed that, similarly, the wavelength of a moving particle would be Planck's constant divided by its momentum. Therefore, photon momentum would be:

$$p = \frac{hf}{c}$$

$$= \frac{h}{\lambda}$$

$$\lambda = \frac{h}{p}$$

The de Broglie wavelength of a particle $\lambda = \frac{h}{mv}$.

The examiners of de Broglie's thesis liked his mathematics but did not believe that it had a physical significance. When de Broglie was questioned about this he disagreed and claimed that it should be possible to observe the wave nature of a beam of electrons **diffracted** from the surface of a crystal. The examiners accepted de Broglie's thesis and were influenced by Einstein's comment 'I believe it is a first feeble ray of light on this worst of our physics enigmas'.

Most other developments in physics in the past had occurred when a theory was developed to explain an observation. Here, de Broglie did the opposite. He made a prediction based on his theoretical work and suggested the observations that would support his theory.

In fact, de Broglie had initiated the revolution in which Heisenberg, Schrödinger, Dirac, Born, Pauli and others developed a detailed theory called **quantum mechanics**. Even before experimental evidence of the wave nature of electrons had been observed, the development of quantum mechanics was well on its way (see Physics in focus, page 449).

Quantum mechanics is a complete theory, not a mixture of classical and quantum ideas as had been used previously by Bohr. The 'old' quantum-theory model of the atom (which still retained some classical physics) reached its peak in about 1922 but was then replaced by quantum mechanics, a completely quantum theory.

In quantum mechanics, particles have both a wave and a particle nature and the rules of mechanics that are obeyed on a macroscopic scale are not obeyed. The uncertainty principle and wave-particle duality lie at the heart of quantum mechanics.

SAMPLE PROBLEM

23.1

SOLUTION

Determining the wavelength of a moving electron

Calculate the wavelength of an electron moving with a velocity of $5.00 \times 10^5 \text{ m s}^{-1}$ and compare this value to the wavelength of visible light. (At this velocity, relativistic effects can be ignored.)

$$\begin{aligned}\lambda &= \frac{h}{mv} \\ &= \frac{6.63 \times 10^{-34}}{9.11 \times 10^{-31} \times 5.00 \times 10^5} \\ &= 1.46 \times 10^{-9} \text{ m}\end{aligned}$$

The shortest wavelength of light (violet) is about 400 nm. The wavelength of the electron moving at $5.00 \times 10^5 \text{ m s}^{-1}$ is about 300 times less than the wavelength of violet light.

Confirmation of de Broglie's matter waves

We shall see in the Physics in focus section on page 449 that Max Born played an important role in working with Heisenberg to develop quantum mechanics. Born made a major contribution to the mathematics of the theory which he felt was generally overlooked.

In 1922 and 1923, Clinton Davisson (1881–1958) and Charles Kunsman studied the strange behaviour of electrons scattered from the surface of crystals. They explained their scattering results as being caused by the structure of the atoms that were bombarded by the electrons. After de Broglie's prediction of matter waves, Walther Elsasser, a 21-year-old student of Max Born, published a brief note explaining the results of these experiments in terms of the wave nature of the scattered electrons. This was not appreciated or accepted by Davisson and Kunsman.

In 1927, Davisson, working this time with Lester Germer (1896–1971), studied the surface of a piece of nickel by examining the scattering of electrons from it (see figure 23.9 below). The surface of the nickel consisted of many microscopic crystals bonded together at random orientations and it was expected that even the smoothest possible surface would appear rough to the electrons.

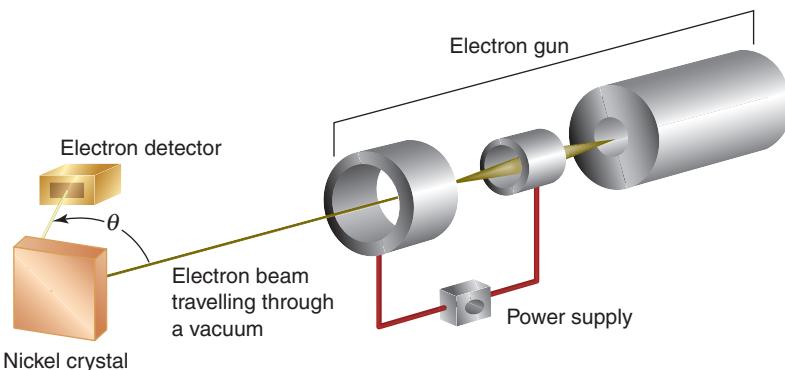


Figure 23.9 The experiment of Davisson and Germer. Electrons are accelerated in the electron gun and fired at the surface of a nickel crystal. As the angle to the electron detector (θ) was changed, the diffraction pattern was observed.

During the course of their experiment, an accident occurred and air entered the vacuum chamber. An oxide film formed on the metal surface. In an attempt to remove the oxide film, Davisson and Germer heated the metal to a temperature just below its melting point. They did not know this, but it had the effect of annealing the metal. Large single crystal regions, which were larger than the width of their electron beam, were produced.

The results were now very different. Davisson and Germer were familiar with X-ray diffraction and with de Broglie's theory of matter waves. They recognised that the electrons were being diffracted. As diffraction is a property of waves and not particles, they had established that electrons had a wave nature as well as a particle nature.

G. P. Thomson (1892–1975), a son of J. J. Thomson, made a similar discovery in England about the same time. If Davisson had accepted the idea of Walther Elsasser, he might have received the Nobel prize by himself. As it was he shared it with G. P. Thomson in 1937.

Bohr's electron orbits explained

When de Broglie developed the idea of matter waves, he had believed that the orbits of the electron in the hydrogen atom were something like standing waves. This idea is depicted in figure 23.10.

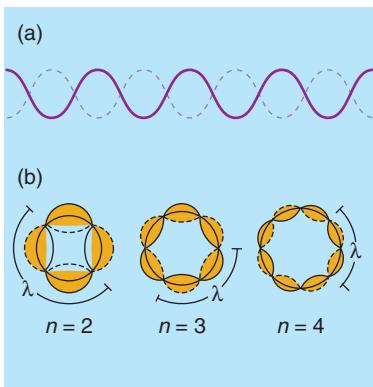


Figure 23.10 (a) A standing wave in a string (b) Circular standing waves

The condition for a standing wave to be formed on a string fixed at each end, is that the length of the string must be an integral number of half wavelengths.

If we consider an electron as setting up a standing wave pattern as it orbits around a nucleus, there must be an integral number of wavelengths in that pattern.

If the circumference is taken as $2\pi r$ then there are n wavelengths in the circumference, $n\lambda = 2\pi r$.

The de Broglie wavelength is:

$$\lambda = \frac{h}{mv} \quad \text{and} \quad n\lambda = 2\pi r$$

$$\text{So: } \frac{nh}{mv} = 2\pi r$$

$$mvr = \frac{nh}{2\pi}.$$

This is Bohr's quantisation condition that angular momentum can exist only in integer multiples of $\frac{h}{2\pi}$. The quantised electron orbits of Bohr can be explained by de Broglie's proposal that particles have a wave nature and that their wavelength is $\lambda = \frac{h}{mv}$.

The Bohr model revisited

We have just seen that we can use the wave nature of electrons to account for the stability of the electron orbits in the Bohr model. This also provides a reason for the quantisation condition used by Bohr in 1913.

The Bohr model had other significant successes. In 1913, H.G. Moseley had shown in his work on the emission of X-rays from various substances that the wavelengths of the emitted X-rays were in agreement with the values predicted by Bohr's theory.

Also in 1913, Johannes Stark showed that when a gas was placed in an electric field, the spectral lines were split. Bohr was able to show that this could be explained by his theory.

In 1914, two German scientists, James Franck and Gustav Hertz, found that when mercury atoms were bombarded by electrons, the atoms would not absorb energy below a certain critical value from the incident electrons. Bohr immediately realised that this could be interpreted in terms of his theory, but Franck and Hertz continued for some time to provide their own incorrect interpretation.

In 1922, Bohr developed an explanation of the atomic structure that underlies the regularities of the periodic table. He considered that atoms are built up of shells of orbiting electrons with the shells being filled, in the case of uranium, by 2, 8, 18, 32, 18, 8 and 6 electrons. Uranium is the naturally occurring element with the highest atomic number. However, Bohr could not explain why these were the numbers of electrons in the shells.

Chemists had successfully predicted the properties of previously unknown elements using the periodic table, but Bohr's variation predicted that element 72, when discovered, would not belong to the rare earth group as chemists predicted but rather be a valency four metal similar to zirconium. Georg de Hevesy and Dirk Coster discovered the new element, hafnium, and the night before Bohr was awarded his Nobel Prize they relayed to him the information that it had the properties he

The periodic table is reproduced in Appendix 2, page 528.

had predicted. This was the high point of the ‘old’ quantum theory. (The term ‘old’ is applied to quantum theory before 1925.)

The wave nature of electrons may have helped with an understanding of the electron orbits, but it did not help with any of the other difficulties we have experienced with the Bohr model.

A new quantum theory

The Physics in focus section on pages 449–451 provides an outline of some of the steps that were taken in replacing the old quantum theory of Bohr with the new theory of quantum mechanics.

The complete quantum theory came about after breakthroughs by Werner Heisenberg (1901–1976) and Erwin Schrödinger (1887–1961). These scientists, independently in 1925 and 1926, discovered different forms of the same theory. (Heisenberg’s matrix mechanics and Schrödinger’s wave mechanics were later shown to be equivalent.) Heisenberg introduced the uncertainty principle and Bohr completed the theory with his principle of complementarity.

By the time of the Solvay Conference of October 1927, the old quantum theory had been replaced. At this conference, Schrödinger presented a paper on his wave function theory but he declined to discuss the interpretation of the wave functions (which Born interpreted as being related to the probability of finding an electron in a certain location). The theory is now called quantum mechanics, and Bohr’s ideas — along with Heisenberg’s uncertainty principle and Born’s probability interpretation — became known as the Copenhagen interpretation. At the Solvay Conference, Einstein raised his first public objections to quantum mechanics and he was to continue to debate with Bohr this interpretation of quantum mechanics. Einstein never accepted that quantum mechanics was a ‘complete’ theory and the Copenhagen interpretation is still considered obscure by some physicists today. It is no wonder that Bohr made his famous statement, ‘Anyone who is not shocked by quantum theory has not understood it.’

We have not even scratched the surface of quantum mechanics. We have seen that there were major problems with the ideas of the original quantum theory and that, in the process of overcoming those problems, a new theory was developed that required a modification of our ideas about the physical world.

In this strange new theory there is no such thing as a particle or a wave but rather there is a wave-particle duality, and making an accurate observation of one property means that another property cannot be measured accurately. We have not studied the mathematics of either the matrix mechanics of Heisenberg or the wave mechanics of Schrödinger, and we have not studied the probability interpretation of Born. We are therefore not in a position to see why quantum theory and, in particular, the Copenhagen interpretation is as shocking as Bohr suggests.

A deeper study of quantum mechanics leads to an atomic world that is fuzzy and nebulous and in which, according to Bohr and the Copenhagen interpretation, nothing actually exists until it is observed. The clockwork world of Newton becomes a world of quantum uncertainty where nothing is predictable. Even worse, events can occur without having a cause and quantum particles can suddenly pop into existence.

Yet, despite all the problems of interpretation, quantum mechanics is an incredibly successful theory. Consider the following facts:

- Quantum mechanics helps us explain and control the properties of metals, insulators, semiconductors and superconductors.



Figure 23.11 Werner Heisenberg (1901–1976)

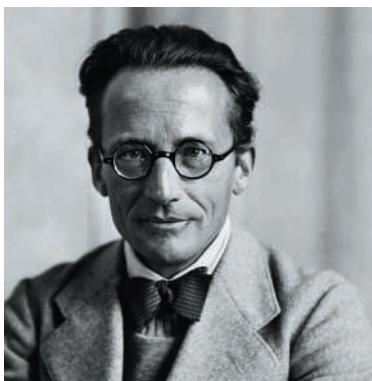


Figure 23.12 Erwin Schrödinger (1887–1961)

- The inventors of the transistor acknowledge the part quantum theory played in their discovery. That discovery led to the development of ever more powerful computers and microcomputers and a revolution in communications and information technology.
- Lasers and masers are quantum devices.
- Quantum mechanics explains the structure of the atom and nucleus as well as mechanical and thermal properties of solids.
- Quantum mechanics gave chemistry a firm base and explained chemical bonding. The new areas of molecular biology and genetic engineering have also arisen from quantum chemistry.
- In astrophysics, the processes that occur in stars can be explained by quantum mechanics, and even our theories regarding such exotic objects as black holes are based on quantum mechanics. There has even been the suggestion that our universe began as a ‘quantum fluctuation’. Quantum mechanics is a theory that has changed the world and our view of it.

PHYSICS IN FOCUS

The development of quantum mechanics

Many physicists contributed to the development of the theory of quantum mechanics. Many arguments on the interpretation of the meaning of quantum mechanics took place in the twentieth century and continue today. Some of the major contributors are mentioned below with a brief statement about their contribution. Research on the internet will reveal more about their lives and contributions.

Heisenberg and Bohr

Werner Heisenberg (1901–1976) heard Niels Bohr (1885–1962) lecture on the periodic table when Bohr visited Germany. Heisenberg was impressed with Bohr, although he believed that Bohr did not understand the reason why his theories were correct. Bohr was also impressed with Heisenberg, who had objected to one of Bohr’s statements. Bohr liked to identify smart people who were not afraid to speak up. In a similar way he had picked out Richard Feynman when he first met him at Los Alamos during the Manhattan Project — the project to develop the atomic bomb (see chapter 25, page 481).

Bohr invited Heisenberg to Copenhagen and they took a fresh look at quantum theory. Heisenberg thought Bohr’s electron orbits were fanciful and Bohr suggested to Heisenberg that he should forget about electron orbits around the atom. Heisenberg decided to reject a mechanical model completely and to look for patterns in numbers — in other words, to develop a completely mathematical model.

Heisenberg develops quantum mechanics

In May 1925, Heisenberg suffered from very bad hayfever and took a holiday for two weeks. He went to Heligoland where there was little pollen and he soon recovered. While there, he developed his mathematical theory of quantum mechanics. In it he had arrays of numbers that Max Born (1882–1970) later realised were matrices. In the next few months, Heisenberg worked with Born and Pascual Jordan (1902–1980) to develop what he called ‘a coherent mathematical framework, one that promised to embrace all the multifarious aspects of atomic physics’. At the time, this was referred to as matrix mechanics.



Figure 23.13 Max Born (1882–1970)

(continued)

Pauli applies quantum mechanic to hydrogen

Wolfgang Pauli (1900–1958), like Heisenberg, was a student of Arnold Sommerfeld at Munich. Pauli was a friend of Heisenberg and had influenced his path into atomic physics. Pauli took Heisenberg's quantum mechanics and applied it, with difficulty, to the hydrogen atom. He was able to derive Balmer's equation and Rydberg's constant using quantum mechanics.

Bohr had done the same thing in 1913 using his inconsistent assumptions of classical physics and quantum theory. Bohr was delighted with the work of Pauli and the success of the new quantum mechanics.

Pauli and his exclusion principle

The first three 'quantum numbers' are the principal quantum number n , from the Bohr model, the angular momentum quantum number l , and the magnetic quantum number m .

Pauli used Bohr's idea of shells of electrons and, in 1925, realised that if he introduced a fourth quantum number, he could explain the maximum number of electrons in each shell. The fourth quantum number was associated with 'spin'. (For more on spin see page 465.) The maximum number of electrons in each shell corresponded to the number of different sets of quantum numbers available for each shell. Pauli's exclusion principle states that no two electrons can have the same set of quantum numbers.

Pauli's exclusion principle provided the reason for electrons in atoms being arranged in shells with the maximum number of electrons being 2, 8, 18, 32, 18, 8 from the first to sixth shell.

Schrödinger — a different approach

De Broglie's work on the wave nature of particles might not have received wide publicity if it had not been brought to the attention of Einstein. In 1925, Erwin Schrödinger (1887–1961) read a comment of Einstein on de Broglie's work that referred to it as more than a mere analogy. Schrödinger then set about trying to restore some of the familiar concepts of waves to quantum theory. He eventually derived equations



Figure 23.14 Wolfgang Pauli (1900–1958)

that looked like the equations used to describe real waves, and it seemed that he had managed to bring quantum ideas back towards a much more comfortable formulation associated in some way with classical physics.

Heisenberg did not like Schrödinger's approach. Heisenberg did not see how continuous waves could be used to describe the discontinuous behaviour of an electron jumping from one state to another. Many papers started appearing on Schrödinger's wave mechanics and very few on Heisenberg's matrix mechanics. Schrödinger did not like Heisenberg's non-visual interpretation or the use of matrices. Most physicists of the time preferred Schrödinger's approach to Heisenberg's approach.

However, it was not long before Schrödinger demonstrated that the two different approaches were simply different versions of the same thing. He showed that Heisenberg's matrices could be generated in Schrödinger's theory and Schrödinger's waves could be produced from Heisenberg's matrices.

Schrödinger later spent time with Bohr and was most disappointed to find that his 'waves' were not real waves at all. Max Born showed that they were associated with the probability of finding an electron at a particular location.

Heisenberg and uncertainty

In late 1926, Heisenberg showed that uncertainty is an inherent property of quantum mechanics. He showed that there are pairs of quantities that cannot be determined simultaneously. If we know the accurate position of a particle, say an electron, then you cannot know its momentum accurately. If you determine its momentum accurately, you cannot specify its position accurately.

This is represented by Heisenberg's uncertainty principle:

$$\Delta x \times \Delta p \geq \frac{h}{2\pi}$$

where

Δx and Δp = the inherent uncertainties in position and momentum

h = Planck's constant.

Bohr and the principle of complementarity

Bohr had a problem with Heisenberg's uncertainty principle. It was based on wave-particle dualism and also with the fact that an observation of an atomic system would disturb the system. A slightly oversimplified version of Bohr's principle of complementarity that addresses this is that, when you make an observation of the particle nature of something, it still has a wave nature but you do not see the wave nature during that observation or vice versa.

The Dirac equation

Paul Dirac (1902–1984) extended quantum mechanics and derived the Dirac equation, which added relativity to quantum theory. It predicted correctly the spin of electrons (which is a relativistic effect). It also predicted the existence of a particle similar to an electron but with a positive charge. The anti-electron or positron was observed by Carl Anderson in 1932.

Dirac discovered that the equations of quantum mechanics have the same structure as the equations of classical physics and that the equations of classical physics can be obtained from quantum mechanics by using very large quantum numbers or setting Planck's constant to zero.



Figure 23.15

Paul Dirac
(1902–1984)

SUMMARY

- Bohr's model of the atom used both the ideas of classical and quantum physics. This was the fundamental problem of his model.
- Diffraction is a wave phenomenon and a diffraction pattern is produced when diffracted waves interfere.
- Louis de Broglie proposed that the wave and particle behaviour of light were inextricably linked. He then predicted that matter must also have both a particle and a wave nature. The wavelength of particles would be equal to Planck's constant divided by the momentum of the particle.
- Einstein considered that de Broglie's theory might be the first step in explaining the problem of the apparent dual nature of light.
- De Broglie proposed that electrons would undergo diffraction. Davisson and Germer observed diffraction of electrons reflected from the surface of a crystal and supplied the evidence for the wave nature of electrons.
- Heisenberg proposed a mathematical model rather than a mechanical model as the basis for his theory of quantum mechanics.
- Pauli applied the ideas of quantum mechanics to the hydrogen atom and was able to derive Balmer's equation and the value of the Rydberg constant.
- Pauli introduced a fourth quantum number and used his exclusion principle to explain why there was a maximum number of electrons possible in each shell of electrons.

QUESTIONS

1. Outline the ways in which the quantum mechanics developed by Heisenberg improved on the Bohr model that had successfully predicted the wavelengths of the spectral lines of hydrogen.

2. Explain why Bohr was said to be delighted after Pauli used a new approach to calculate the wavelengths of the spectral lines of hydrogen.
3. When de Broglie was examined for his PhD, his thesis was first thought by his examiners to bear little relationship to reality.
 - (a) What did de Broglie predict that made it seem unrelated to reality?
 - (b) What did de Broglie suggest could be observed to support his prediction?
 - (c) How did the work of Davisson and Germer provide evidence to support de Broglie?
4. If a proton and an electron are travelling with equal velocities, state which has the longer de Broglie wavelength.
5. If one electron travels twice as fast as another electron, state which one has the greater wavelength.
6. (a) If an electron travelling at $1.0 \times 10^4 \text{ m s}^{-1}$ was accelerated to $2.0 \times 10^4 \text{ m s}^{-1}$, calculate the ratio of its new wavelength to its original wavelength.
(b) If an electron travelling at $1.0 \times 10^8 \text{ m s}^{-1}$ was accelerated to $2.0 \times 10^8 \text{ m s}^{-1}$, would it change its wavelength by the same amount as the electron in part (a)? Explain your answer.
7. (a) Calculate the de Broglie wavelength of an electron in a television set that hits the screen with a velocity one-tenth of the velocity of light.
(b) With what velocity would you roll a ball of mass 0.1 kg if it is to have the same de Broglie wavelength as the electron in part (a)?
8. A neutron emitted when a uranium-235 nucleus undergoes fission may have an energy of about 1 MeV. A 'thermal' neutron that would be captured by a uranium-235 nucleus in a nuclear reactor would have an energy of about 0.02 MeV.
 - (a) Calculate the wavelength of a 1 MeV neutron.
 - (b) Calculate the wavelength of a 0.02 MeV neutron.

CHAPTER 24 PROBING THE NUCLEUS



Figure 24.1 A wide-angle view of the Super-Kamiokande detector 1000 m below ground in Japan as it was nearing completion. The view is from the bottom of the 42 m high, 39 m diameter tank. The top and walls of the tank at this stage were covered with about 9000 light-sensitive phototubes. On completion it was filled with 50 000 tonnes of ultra-pure water. In 1998, results from this detector indicated neutrino oscillation and hence, neutrino mass.

Remember

Before beginning this chapter, you should be able to:

- recall the features of the model of the atom in terms of the arrangement and properties of protons, neutrons and electrons
- describe the nature of alpha, beta and gamma rays and recall their properties in terms of ionising power, penetrating power and deflection by magnetic and electric fields.
- recall the relationship between energy and mass, $E = mc^2$.

Key content

At the end of this chapter you should be able to:

- define the term ‘transmutation’ and describe the transmutations involved in naturally occurring radioactivity
- describe Chadwick’s discovery of the neutron and the part played by conservation laws in this discovery
- contrast the properties of protons and neutrons and describe them as nucleons
- describe the problems associated with the energy distribution of electrons emitted in beta decay that led Pauli to predict the existence of the neutrino
- describe the properties of the strong nuclear force and realise that over very short ranges it is much stronger than the electrostatic and gravitational forces between nucleons
- explain the concept of mass defect (using $E = mc^2$) and be able to calculate the mass defect of nuclei and the energy associated with nuclear reactions.

24.1 DISCOVERIES PRE-DATING THE NUCLEUS

In chapter 22 we studied the development of the nuclear atom. In this chapter, we will study the development of the theories of the nucleus itself. Some important discoveries and investigations were made before the discovery of the nucleus and we will start by examining those investigations.

Discovery and early investigations of radioactivity

A **phosphorescent** substance absorbs radiation of one wavelength and then emits radiation of a different wavelength over a period of time. The hands of some analogue watches are coated with a phosphorescent substance to enable them to be seen in the dark.

Henri Becquerel discovered radioactivity in 1896 when he was studying the radiation emitted from **phosphorescent** substances that had previously been exposed to sunlight. Becquerel found by accident that a salt of uranium, potassium–uranyl sulfate, continuously emitted radiation regardless of whether or not it had been exposed to sunlight. This radiation penetrated matter, passing through black paper (opaque to light) and causing a photographic plate to become darkened. It seemed to be similar in nature to X-rays, which had been recently discovered by Wilhelm Roentgen (1845–1923).

In 1898, Rutherford showed that there were two components (alpha and beta rays) of the radiation discovered by Becquerel, and in 1900 Paul Villard (1860–1934) discovered the third component (gamma rays). You encountered alpha, beta and gamma rays in the preliminary course unit ‘The Cosmic Engine’. The properties of the radiation are reviewed in ‘Physics in focus’ below.

PHYSICS FACT

The radiation discovered by Becquerel

Uranium is an emitter of alpha particles, which have a low penetrating power and would have been stopped by the black paper. What really caused the photographic plate to be darkened was the beta particles emitted by the thorium produced by the alpha particle emission from uranium.

PHYSICS IN FOCUS

Review of the properties and identities of alpha, beta and gamma radiation

The properties of alpha, beta and gamma radiation can be summarised as follows.

Penetrating power

Figure 24.2 shows that the penetrating power is lowest for alpha particles, which can be stopped by a sheet of paper or a few centimetres of air. Beta particles will be stopped by many metres of air or a sheet of aluminium about a centimetre thick. Gamma rays may pass through a few centimetres of lead or many metres of concrete before being stopped.

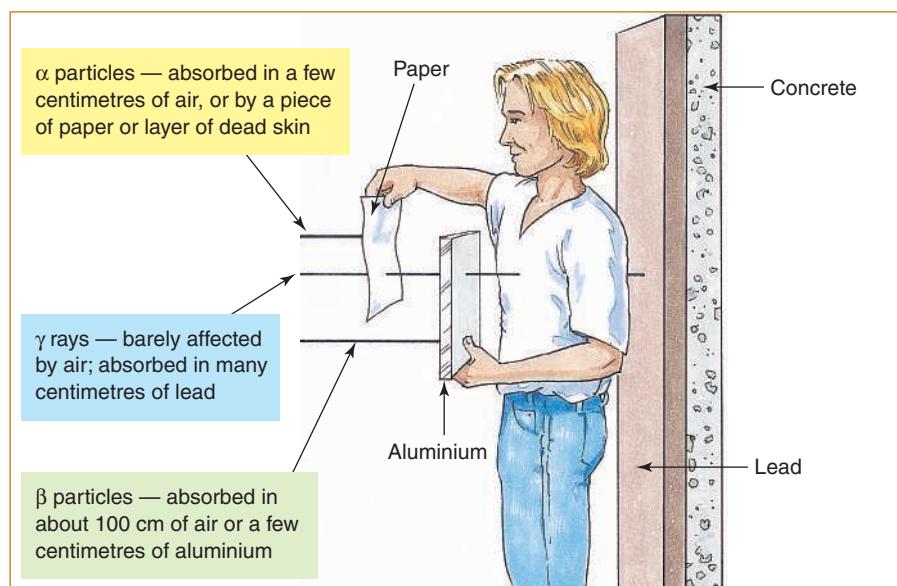


Figure 24.2 The relative penetrating powers of alpha, beta, and gamma radiation

Ionising power

As might be expected, the ionising power is the inverse of the penetrating power. Alpha particles interact most strongly with matter and hence have a low penetrating power and high ionising power. The ionising power of beta particles is lower and that of gamma rays is very low (see figure 24.3).

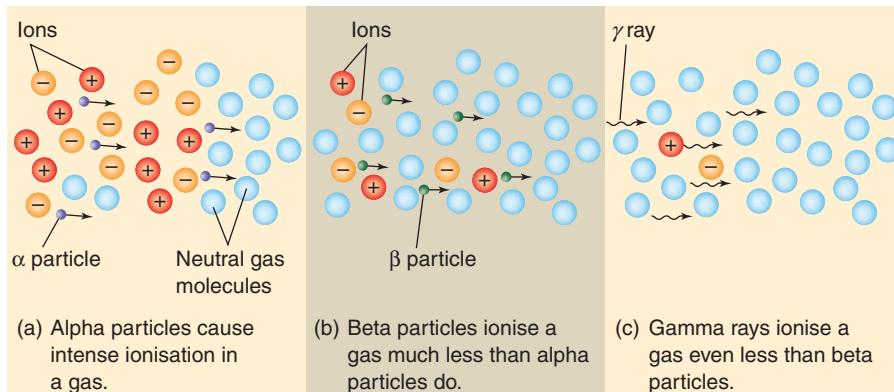


Figure 24.3 Ionisation caused by radioactive emissions passing through gas

Deflection by a magnetic field

The paths of the different types of radiation through a magnetic field proved to be harder to observe, but once detected they indicated that alpha particles were positively charged, beta particles were negatively charged and gamma rays were neutral.

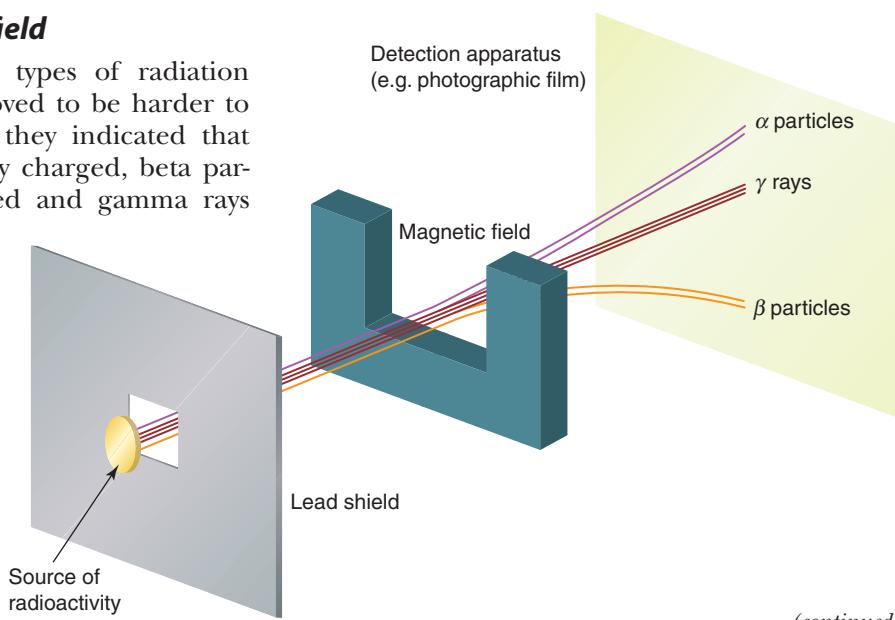


Figure 24.4 The paths of alpha, beta and gamma rays through magnetic fields

(continued)

Table 24.1 Summary of properties and identities of alpha, beta and gamma radiation

Type of radiation	Penetrating power	Ionising power	Path through magnetic field	Nature of radiation (in today's terms)
Alpha	Very low	Very high	Curved path of positive charge	Helium nucleus
Beta	High	Moderate	Curved path of negative charge	Electron
Gamma	Extremely high	Very low	Not deflected	High energy photon

PHYSICS FACT

Detection of radioactivity

One of the earliest methods of detecting radioactivity used a radioactive source's ionising power. As shown in figure 24.5, if a radioactive source was brought near to a charged electroscope, the electroscope would discharge. It did not matter whether the radioactive source emitted alpha particles or beta particles or what the sign of the charge was on the electroscope. The electroscope was discharged because it attracted ions of opposite charge to the electroscope.

The radioactive particles produced both positively and negatively charged ions when passing through air and hence the electroscope was discharged.

Apart from the scintillation method used by Rutherford and his co-workers, the other methods used to detect radioactivity relied on the ionising power of the radiation. This is one of the reasons why detecting the neutral particles, neutrons and neutrinos, proved so difficult.

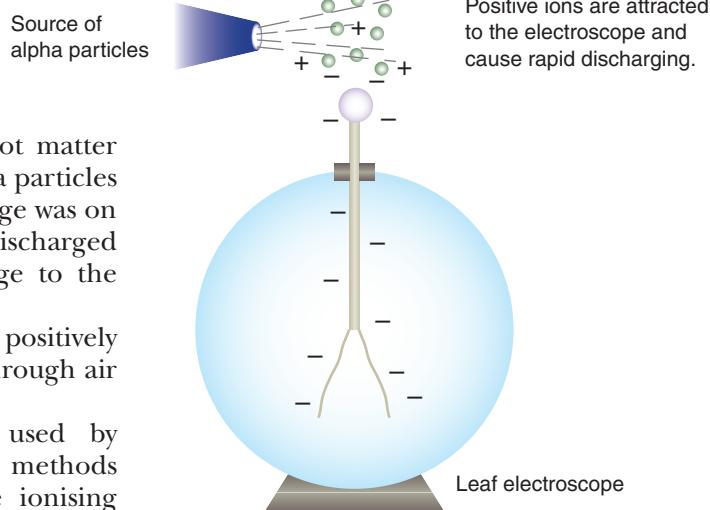


Figure 24.5 Discharging an electroscope with a radioactive source

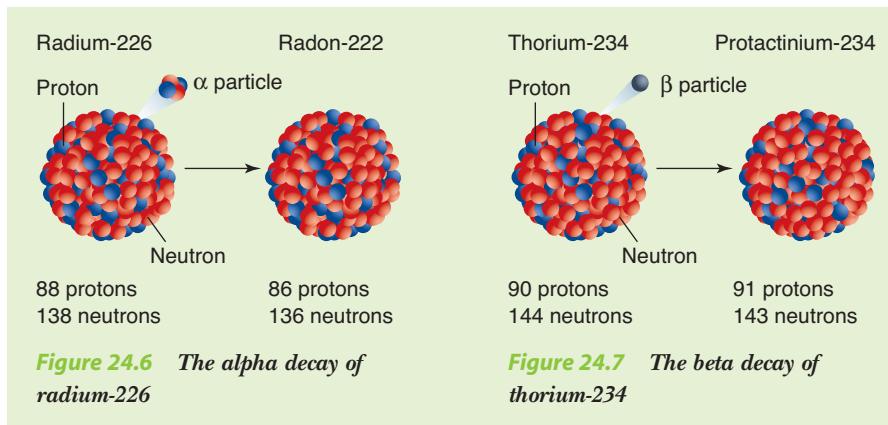
Naturally occurring radioactivity explained

In 1902, Rutherford and the English chemist Frederick Soddy (1877–1956) published a paper proposing that the emission of radioactivity was the result of 'radioactive transformation'. When a radioactive atom emitted an alpha particle or beta particle, the atom split into two. The alpha or beta particle was emitted and what remained was a heavy leftover part that was chemically and physically different from the parent atom (see figures 24.6 and 24.7).

This 'transformation', 'disintegration', 'decay' or '**transmutation**' was responsible for turning one element into another.

After the discovery of the nucleus, the transmutation was identified as the emission of alpha or beta particles *from the nucleus*.

When a radioactive atom emitted an alpha particle or a beta particle, an atom of a new element was produced. This process by which a new daughter element was formed from a parent element was termed **transmutation**.



PHYSICS FACT

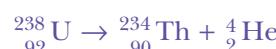
Writing nuclear equations

The formulas for nuclei are written in the form ${}^A_Z X$ where X is the symbol for the element, A is the mass number (number of protons plus neutrons) and Z is the atomic number (number of protons).

The term **nuclide** is used to denote a nucleus characterised by particular values of Z and A. If a group of nuclides share the same atomic number but have different mass numbers, they are referred to as **isotopes** of that element.

In any nuclear reaction, the sum of the mass numbers before the reaction must be equal to the sum of the mass numbers after the reaction. The sum of the atomic numbers before the reaction must likewise be equal to the sum of the atomic numbers after the reaction. This can be slightly complicated by the fact that if a beta decay is involved, the electron is assigned an atomic number of negative 1. It may be necessary to look up a periodic table to determine the element formed if not all the information is supplied.

The equations for the transmutations associated with some common examples of alpha decay and beta decay are:



We can see that alpha decay reduces the atomic number by two and the mass number by four, and beta decay increases the atomic number by one and leaves the mass number unchanged.

The term **nuclide** refers to a particular nucleus with certain values of Z (atomic number) and A (mass number).

An **isotope** is a nuclide that has the same number of protons but different numbers of neutrons.

We will see later that another particle, an antineutrino, is also emitted during beta decay. As we are dealing with the transmutations observed and explained in the early 1900s we will omit the antineutrino at present.

SAMPLE PROBLEM

24.1

Radioactive decay

The decay series starting with uranium-238 proceeds by alpha decay and beta decay until the stable isotope of lead-206 is reached.

- (a) How many alpha decays are involved in this series?
- (b) How many beta decays are involved in this series?

SOLUTION

- (a) As alpha decay is the only decay that reduces the mass number, and each alpha decay causes a decrease in mass number of four, there must be $\frac{238 - 206}{4} = 8$ alpha decays.

- (b) The atomic number of uranium is 92 and that of lead is 82. As there are eight alpha decays, these would reduce the atomic number by 16. However, it decreased only by 10. Hence, there must have been six beta decays, each one increasing the atomic number by one.

24.2 DISCOVERY OF THE NEUTRON

After the discovery of the nucleus, it seemed logical to assume that the nucleus contained protons and electrons. It was possible to explain radioactive transmutations in terms of emission of alpha and beta particles from the nucleus of protons and electrons. However, there were major problems with the idea of a nucleus containing these constituents.

In 1920, Rutherford proposed that a neutral particle, with mass comparable to that of a proton, must be another constituent of the nucleus. He named this particle a neutron. In future research he and James Chadwick (1891–1974) continued to look out for any result that would suggest the existence of such a particle.

Experiments involving artificially induced radioactivity

The radioactivity that we have encountered so far has been associated with natural alpha, beta and gamma emitters. Rutherford was the first to use alpha particles to produce nuclear reactions.

The first artificially induced transmutation

In 1919, Rutherford bombarded nitrogen gas with alpha particles from bismuth-214. A positively charged particle which was more penetrative than an alpha particle was produced. This particle was identified as a proton.

What had occurred, as shown in figure 24.8, was that the alpha particle had combined with the nitrogen nucleus and a proton had been emitted. The alpha particles from the bismuth-214 source were able to approach the nucleus very closely and occasionally make contact with it. The equation for this reaction is:



PHYSICS FACT

Alpha-particle-induced nuclear reactions

When the first alpha particle scattering experiments were performed, low-energy alpha particles were used and those that approached a gold nucleus (containing 79 protons) were strongly repelled. In the alpha-particle-induced reaction with nitrogen, the alpha particles had a much higher energy than those used in the early experiments and there was only a weak repelling force from a nitrogen nucleus that contained only 14 protons. An energetic alpha particle was able to make contact with the nitrogen nucleus.

Various writers have commented that Rutherford was fortunate that he did not use a source of very powerful alpha particles when he performed his first alpha particle scattering experiments!

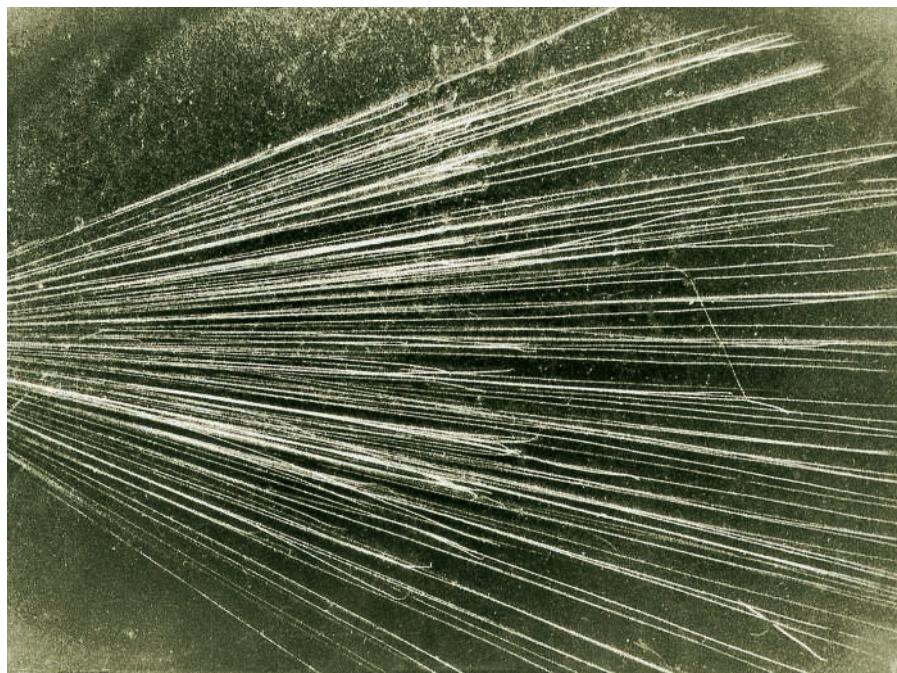
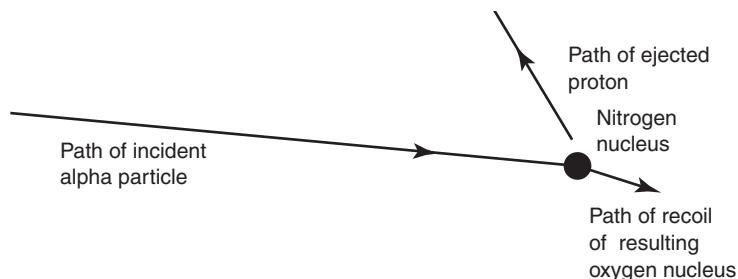


Figure 24.8 The photograph shows alpha particle tracks through a cloud chamber filled with nitrogen gas. The diagram helps to identify an event where a nitrogen nucleus has been struck by an alpha particle. A proton is ejected upwards and the resulting oxygen nucleus recoils downwards.

We have seen that in atomic physics it was convenient to measure energy in electron-volts. In nuclear physics, energy is usually measured in million electron volts (MeV). The energies associated with nuclear processes are very much larger than those associated with atomic processes.



An artificially induced radioactivity

In 1930, Bothe and Becker (in Germany) fired alpha particles at beryllium and found that a highly penetrating radiation was produced. The radiation seemed to be similar to gamma rays (high-energy photons) but it was much more highly penetrating than the gamma rays previously observed. It was found to have an energy of about 10 MeV, again much higher than that previously observed for gamma rays.

In France, Frédéric Joliot (1900–1958) and his wife Irène Curie (1897–1956) (daughter of Marie Curie), studied this mysterious radiation and let it fall on a block of paraffin. Paraffin is a hydrocarbon very rich in hydrogen atoms. They found that the radiation knocked protons (hydrogen nuclei) from the paraffin (see figure 24.9). The energy of the protons was about 5 MeV. Of course, now that charged particles (protons) were involved, it was much easier to determine their properties. They also found that many more protons than expected were emitted from the paraffin. If gamma rays had been responsible, their very high penetrating power would have resulted in fewer interactions with protons.

The high energy of the protons (5 MeV) was a problem because applying the conservation of energy and conservation of momentum to the collision between a gamma ray and a proton yielded a value for the incident gamma ray of at least 50 MeV. This was a major dilemma because the energy of the incident alpha particles was only about 5 MeV. In other words, if this was the correct interpretation, there had to have been a tenfold increase in energy in the interaction!

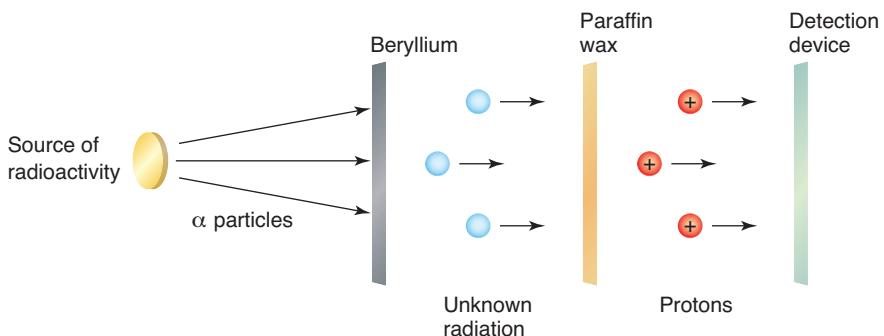


Figure 24.9 The reaction of alpha particles with beryllium produced a mysterious radiation that knocked protons out of paraffin.

PHYSICS FACT

Rutherford's prediction of the neutron

In his Bakerian lecture of 1920, Rutherford had suggested that 'it may be possible for an electron to combine much more closely with the hydrogen nucleus than is the case in the ordinary hydrogen atom'. He later used the term neutron. It is worth noting that Rutherford's conjecture about the existence of the neutron had not received wide publication and it had not been read by either Joliot or his wife. Some years later Joliot commented on the fact that he had not read Rutherford's Bakerian lecture and that, had he done so, it was possible or probable that he and his wife would have identified the neutron before Chadwick.

Chadwick identifies the neutron

Just over two weeks after reading the paper of the Joliot-Curies, James Chadwick (1891–1974) had completed his work and submitted a paper on 'Possible existence of a neutron' (1932). In that time Chadwick applied conservation of energy and conservation of momentum to the interaction of a neutral particle (of mass similar to that of a proton) with a proton. Chadwick made measurements of the recoil of nuclei of hydrogen and nitrogen after interactions with his proposed neutron. The measurements were difficult but led to the mass of a neutron being calculated to be 1.15 times that of a proton.

At this time (1932), there was doubt expressed about whether or not the conservation laws of classical physics would apply to nuclear processes. Some leading physicists were adamant that they would but others, including Bohr, thought otherwise. In fact it was 1936 before Bohr dropped his ideas of non-conservation of energy.

As Chadwick's neutron identification depended on the conservation laws and there was doubt expressed about them at the time, he concluded his paper 'Up to the present, all the evidence is in favour of the neutron ... [unless] the conservation of energy and momentum be relinquished at some point'.

The nuclear equation for the reaction of alpha particles with beryllium is:



PHYSICS FACT

Problems with electrons and protons in close association

It is worth noting that there are major difficulties with the concept of electrons and protons in close association either as a single particle (the neutron) or generally in the nuclei of atoms. The masses of atoms could not be explained in terms of numbers of protons and electrons. Another problem involved the de Broglie wavelength of an electron. How could an electron with an energy of a few MeV be confined to a region with a radius of 5×10^{-15} m when its de Broglie wavelength was large compared to this radius?

These difficulties were overlooked at the time because, after all, an alpha particle seemed to be 4 protons and 2 electrons combined very tightly together.

There were no naturally occurring neutron emitters but now, with a high-energy alpha particle source (such as polonium) and some beryllium, it was possible to produce neutrons and conduct neutron scattering experiments.

24.3 DISCOVERY OF THE NEUTRINO

The discovery of the neutron had gone a long way to help explain the nucleus but problems with beta decay remained. Eventually, the only way to solve these problems was to predict the existence of another neutral particle, the neutrino.

Puzzles and problems of beta decay

There are good reasons why an electron cannot possibly be confined to a nucleus (see Physics fact above), and beta decay had been a problem since its discovery. Attempts to explain beta decay in a similar manner to alpha decay were doomed to failure.

All alpha particles emitted from a particular radioactive species had the same energy, but beta particles seemed to be emitted with a range of energies. There was considerable debate as to whether the beta particles had a continuous or line spectrum.

Detection methods proved to be confusing and photographic methods of detection, which tended to favour line spectra, were not sufficiently sensitive to determine the continuous spectrum.

Prior to World War I, James Chadwick had used Geiger's 'point counter' to detect beta particles that had been deflected by a magnetic field. He detected beta particles with a continuous range of radii indicating that they had been emitted with a wide spread of energies. Figure 24.10 is a graph of these energies.

Some very competent physicists, including Otto Hahn (1879–1968) and Lise Meitner (1878–1968), had claimed to observe many lines in the beta particle spectrum.

In fact, for many years Hahn and Meitner continued to cling to their theory that all beta particles were emitted from the nucleus with the same energy but that this energy was modified as the beta particles left the atom.

How could one beta decay be associated with emission of a certain amount of energy from a nucleus but another beta decay from a similar nucleus be associated with a different amount of energy? After all, both

As we saw in section 22.1, the radius of the path of a charged particle through a magnetic field is given by

$$r = \frac{mv}{qB}$$

Beta particles, all with the same charge to mass ratio, had paths of different radii when travelling through a uniform magnetic field (see figure 24.11). Therefore, they must have had different velocities and, hence, been emitted with different energies.

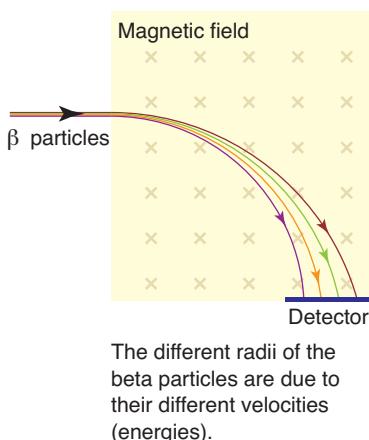


Figure 24.11 The variation of radius of the beta particles as they travelled through a magnetic field indicated that they had a wide range of energies.

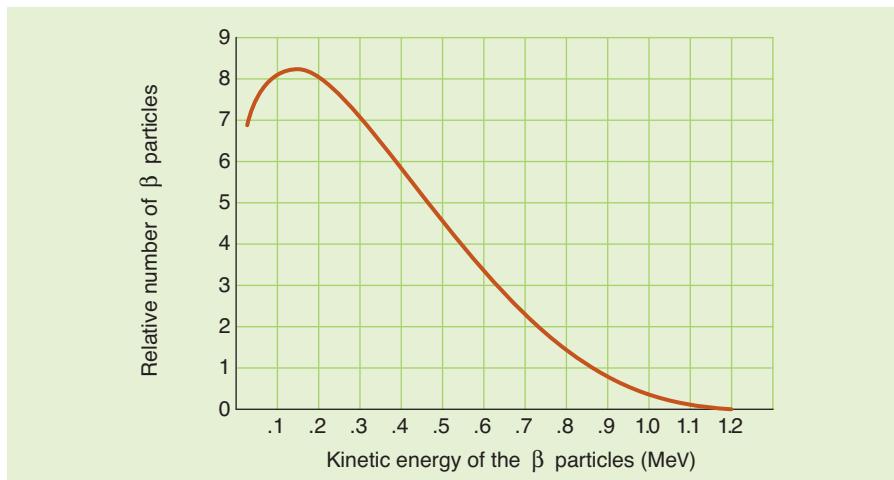


Figure 24.10 The distribution of energy of beta particles

decays produced the same new nucleus. It is not difficult to see why Hahn and Meitner had developed their theory about the same energy on emission or why Bohr continued to doubt that conservation of energy applied to nuclear processes. Bohr's view in particular could be called desperate and, initially, so was Pauli's solution.

In an attempt to resolve the paradoxes involving beta decay, in 1931 Wolfgang Pauli (1900–1958) took the bold step of predicting that there must be another sub-atomic particle.

Pauli himself later stated, 'In June 1931, on the occasion of a conference in Pasadena, I proposed the following interpretation: the conservation laws remain valid, the expulsion of beta particles being accompanied by a very penetrating radiation of neutral particles, which has not yet been observed.'

This was prior to Chadwick's discovery of the neutron and Pauli referred to his predicted particle as a 'neutron'. It was later renamed 'neutrino' by Enrico Fermi (1901–1954) to avoid any confusion with the neutron of Rutherford and Chadwick.

PHYSICS FACT

Pauli and the neutrino

Pauli was reluctant at first to speak about the neutrino. Fermi invited him to speak at a conference in Rome but Pauli was still very cautious about it and would speak privately only to Fermi.

Pauli told astronomer Walter Baade, 'Today I have done the worst thing for a theoretical physicist. I have invented something which can never be detected experimentally.' Baade offered to bet a crate of champagne that the particle would be detected and Pauli accepted. Pauli could never win the bet as there was no time limit specified but in the mid-1950s the bet was paid.

Fermi explains beta decay

By late 1933, Fermi had completed his famous paper on beta decay. In it, he outlined the problems of the continuous spectrum of beta particles, and electrons being present in the nucleus. He then stated that he would explain beta decay in terms of Pauli's suggestion of the emission of a

lightweight neutral particle, the neutrino, along with the electron, and also in terms of the suggestion of Werner Heisenberg (1901–1976) that the nucleus contained only ‘heavy’ particles, protons and neutrons.

Fermi also proposed that the number of electrons and neutrinos was not constant. Electrons and neutrinos could be created or could disappear just like photons. He accounted for the process in which a neutron was transformed into a proton with emission of an electron and a neutrino. He also dealt with more advanced quantum aspects and was able to produce the shape of the spectrum associated with beta decay. He did this for three different values of the mass of the neutrino and found that the shape that most closely resembled the observed shape was for a neutrino mass of zero, or very close to zero.

By the 1950s the decay of other subatomic particles had been found to be similar in character to beta decay, and this led to the idea that a new force was associated with these decays. The force became known as the weak nuclear force or weak force and joined the gravitational force, the electromagnetic force and the strong nuclear force as a fundamental force of nature.

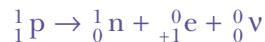
Fermi’s neutrino is now called an antineutrino. (In terms of some of its other properties it is more sensible to classify it as an antiparticle rather than a particle.)

There are now three different simple types of beta decay. They are beta-minus, beta-plus and electron capture. Figure 24.12 shows diagrams representing beta-minus and beta-plus decay and the equations are given below.

Beta-minus decay:

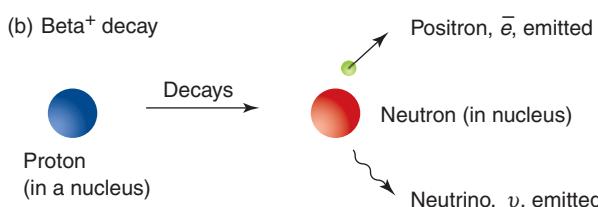
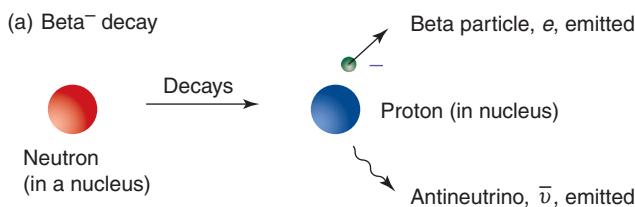


Beta-plus decay:



Anti-particles are written with a bar above the symbol.

Unless stated otherwise, the term beta decay will be used to refer to beta-minus decay.



In 1979, Sheldon Glashow (1932–), Abdus Salam (1926–1996) and Steven Weinberg (1933–) were awarded the Nobel prize for showing that the weak nuclear force and the electromagnetic force could be viewed as different aspects of a single force, called the electroweak force. This reduced the number of fundamental forces from four to three, the other two being gravity and the strong nuclear force that we will encounter later in this chapter.

In 1984, Carlo Rubbia (1934–) and Simon van der Meer (1925–) were awarded the Nobel prize for their experimental verification of the theory of the electroweak force.

Figure 24.12 Diagrams representing (a) beta⁻ decay
(b) beta⁺ decay

When Fermi tried to publish his paper in the prestigious journal ‘Nature’, it was rejected as being too far removed from reality. It remains one of the classic papers of physics.

Detection of neutrinos

When Pauli predicted that the neutrino would never be observed, he imagined that it had a mass similar to the mass of an electron. After Fermi had predicted that it had a lower or zero mass, the detection would have seemed even more remote.

PHYSICS FACT

Interaction of neutrinos with matter

If you hold out your hand as if to catch something coming from the direction of the Sun, about 10^{13} neutrinos pass through it every second. If you did the same thing at night, with the Earth between you and the Sun, again about 10^{13} neutrinos would pass through it every second. The chance of a neutrino interacting with matter is extremely small.

After Fermi's paper on beta decay, Hans Bethe, a US physicist born in Germany in 1906, and Rudolf Peierls (1907–1995) calculated that a neutrino could travel through about 1000 light-years of water before it would be absorbed. They were certain that neutrinos would never be detected by inverse beta decay. Peierls commented about 50 years later that they had not allowed for the existence of nuclear reactors, or the ingenuity of experimental physicists.

In 1953, Cowan and Reines built a detector that was the forerunner of some of the detectors used today. Their detector contained a tank of liquid that emitted scintillations after gamma rays passed through the tank. Photomultiplier tubes around the tank detected light (the scintillations) emitted by the liquid. They hoped to observe scintillations caused by gamma rays produced by the annihilation of a positron and an electron, and also the gamma rays emitted after a nucleus had gained a neutron.

As we have already seen, the process of beta decay produces anti-neutrinos, and Cowan and Reines used the antineutrinos produced in beta decays occurring in a nuclear reactor.

Their experiment relied on the process of inverse beta decay in which a proton interacted with an antineutrino and produced a neutron and a positron.



The results from this detector led them to believe that they had probably detected events produced by antineutrinos. This was confirmed in 1956 when they built an improved version of the detector and counted about three events per hour. This indicated inverse beta decay.

PHYSICS FACT

Neutrinos from SN1987A

On 23 February 1987, various neutrino detectors registered an increase in neutrinos. The increase was small but significant (only 12 neutrino events were detected as an estimated 10^{16} neutrinos passed through the Japanese neutrino detector, the Super-Kamiokande). Some hours later, SN1987A, the first supernova visible to the naked eye for about 400 years, was observed in the Large Magellanic Cloud, a distance of about 50 kpc (kiloparsecs) from Earth. (One parsec is approximately three light years.)

Properties of neutrinos

As we have already seen, neutrinos have an incredibly high penetrating power and only very rarely interact with matter. Neutrinos have other properties, for example:

- they are neutral
- they have either zero or an extremely small mass
- they travel at the speed of light
- they possess both momentum and energy (and carry away from a beta decay the momentum and energy that was previously seen to be missing after a beta decay)
- they have an intrinsic spin (see Physics fact below).

PHYSICS FACT

The concept of 'spin'

Although spin is not dealt with in this course, it is a particularly important concept in quantum mechanics. An electron in the Bohr atom has angular momentum because it is in orbit about the nucleus. It has another component of angular momentum that can be considered to be due to it rotating on its axis. It can be thought of as similar to the Earth in orbit around the Sun and the Earth rotating on its axis as it does so. As spin is really a relativistic effect, this analogy breaks down but it serves as a starting point.

The spin of an orbiting electron in an atom is quantised and gives the fourth quantum number. (The first three are the principal quantum number, n , from the Bohr model, the angular momentum quantum number, l , and the magnetic quantum number, m .)

Similarly, nucleons have an intrinsic angular momentum or spin, and this is also quantised.

PHYSICS IN FOCUS

Recent discoveries related to neutrinos

There have been many mysteries about neutrinos, some of which have been solved by recent discoveries. In chapter 26 we will study the ‘standard model of particle physics’. Perhaps further study of neutrinos will help take physics beyond the standard model. Two neutrino detectors have figured in these important discoveries:

- the Super-Kamiokande (SK) detector in Japan
- the Sudbury Neutrino Observatory (SNO) detector in Canada.

In 1998, observations made with the Super-Kamiokande neutrino detector indicated that neutrinos could change from one type to another. (There are three types of neutrino in the standard model: the electron neutrino, the muon neutrino and the tau neutrino.) Neutrinos had previously been thought to have zero mass but, if they oscillate from one type to another, they must possess some mass, perhaps as small as one millionth of the mass of an electron. The ‘K2K long baseline neutrino oscillation experiment’ indicated that neutrinos do change from

one type to another. Information on this experiment can be found at <http://neutrino.kek.jp/>.

A particular puzzle was the fact that less than half the predicted solar neutrinos were detected on Earth. The Sun produces electron neutrinos and the early neutrino detectors could detect only electron neutrinos. Researchers at the Sudbury Neutrino Observatory in Canada have now confirmed that the electron neutrinos make up one-third of the total number and that the two other types (muon and tau neutrinos) account for the total number in agreement with models of the processes occurring in the solar core. (It is possible to detect all three types of neutrino at SNO.) This not only explains the solar neutrino problem but also confirms the fact that some of the electron neutrinos produced in the Sun have changed into muon or tau neutrinos before they reach the Earth.

In December 2002, more evidence for neutrino oscillation was presented by a group of researchers in Japan and the United States. Their

(continued)

experiments using the Kamioka Liquid Scintillator Neutrino Detector (KamLAND) in Japan have indicated that antineutrinos of one type can change into antineutrinos of another type.

Super-Kamiokande (SK) detector

The Super-Kamiokande neutrino detector (see figure 24.1, page 453) suffered a major disaster in 2001 when it was being refilled after cleaning and maintenance. One of the photomultiplier tubes imploded when the tank was about three-quarters full. The shock waves from the implosion destroyed most of the photomultiplier tubes that were underwater at that time (about 6700 of the total number of 11146 photomultiplier tubes).

The detector was rebuilt in two stages. First the remaining photomultiplier tubes were rearranged (and protected from the effects of another implosion, should one occur) to enable observations to continue; then new photomultiplier tubes were installed. The detector was returned to its full potential in June 2006. It is now referred to as Super-Kamiokande III.

It was very expensive to fully repair the SK detector as each photomultiplier tube costs about US\$3000. However, the cost has been justified because of the results achieved from SK:

‘It is this reviewer’s opinion that SK has to be regarded as an astonishing success. SK produced more physics than any other experiment in particle physics in the last 10 years and a large fraction of such physics has turned out to be unexpected. Indeed it is fair to say that the only clear indications we have of physics beyond the standard model come today from neutrino oscillations and SK is the single most important experiment behind them.’ (A member of SAGENAP quoted in Department of Energy, National Science Foundation, *Report of Scientific Assessment Group on Experimental Non-Accelerator Physics (SAGENAP)*, 12–14 March 2002)

Sudbury Neutrino Observatory (SNO)

The detector at the Sudbury Neutrino Observatory is similar to the Super-Kamiokande detector but is considerably smaller, holding only 1000 tonnes of water with about 10 000 light sensors surrounding it. The key feature of the SNO detector is that it is used with ultra-pure heavy water, or heavy water containing salt. When used with heavy water, electron neutrinos can be detected. When salt is added to the heavy water, all three types of neutrino can be detected. It was this feature that enabled the solar neutrino problem to be solved.

It is worth noting that the detection rate at SNO is about one neutrino per hour and that four years of data were required to produce the first meaningful results from SNO.

2002 Nobel Prize in Physics

Raymond Davis of the University of Pennsylvania and Masatoshi Koshiba of Tokyo University shared one half of the 2002 Nobel Prize for their pioneering work in neutrino detection.

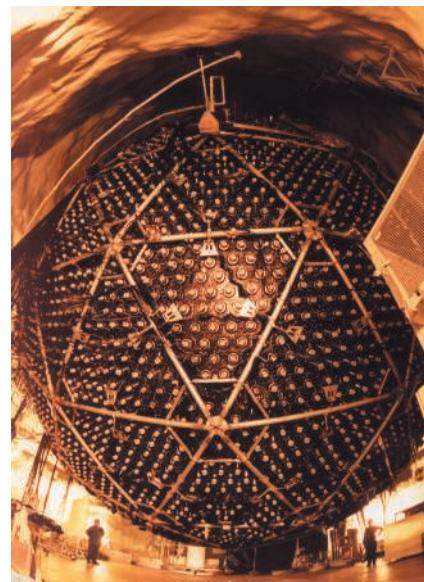


Figure 24.13
Outside view of the photomultiplier tube support structure at the Sudbury Neutrino Observatory

24.4 THE STRONG NUCLEAR FORCE

Both the gravitational and electrostatic forces are inverse square forces so both should become large at the small separation of nucleons in a nucleus. The force of gravity will provide an attractive force between proton–proton, proton–neutron and neutron–neutron, but there will be an electrostatic repulsion between pairs of protons. Another force, the strong nuclear force, is present in the nucleus and holds nucleons together.

Relative strengths of gravitational and electrostatic forces between nucleons

The magnitude of the gravitational force between two masses is given by

$$F_G = \frac{G m_1 m_2}{r^2}$$
 and the magnitude of the electrostatic force between two

$$\text{charges is given by } F_E = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2}.$$

We can see that the ratio of the gravitational force, F_G , to the electrostatic force, F_E , is given by:

$$\begin{aligned}\frac{F_G}{F_E} &= \frac{G m_1 m_2}{r^2} \times \frac{4\pi\epsilon_0 r^2}{q_1 q_2} \\ &= \frac{G m_p^2 \times 4\pi\epsilon_0}{q_p^2}.\end{aligned}$$

Using this equation with the following data:

$$\text{Mass of proton, } m_p = 1.673 \times 10^{-27} \text{ kg}$$

$$\text{Mass of neutron, } m_n = 1.675 \times 10^{-27} \text{ kg}$$

$$\text{Charge on proton, } q_p = 1.602 \times 10^{-19} \text{ C}$$

$$\text{Universal gravitation constant, } G = 6.673 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$$

$$\text{Permittivity of free space, } \epsilon_0 = 8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$$

shows that the gravitational force is smaller than the electrostatic force by a factor of 8.1×10^{-37} .

Clearly, the attractive force of gravity between nucleons is so small as to be insignificant when compared to the electrostatic repulsion between protons. There must be another force present in the nucleus to hold the nucleons together. That force is called the strong nuclear force.

Properties of the strong nuclear force

The properties of the strong nuclear force include:

- an independence of charge and a similar force between proton–proton, neutron–neutron and proton–neutron when electrostatic forces are ignored
- a very strong attractive force, much stronger than the electrostatic repulsion between protons. (At very short distances, much less than the diameter of a nucleon, it changes from attraction to repulsion — see figure 24.14.)
- a very short range force acting over a distance of only about 10^{-15} m. Every proton in a nucleus repels every other proton but the strong nuclear force exists only between a nucleon and its nearest neighbours. This is indicated by the almost uniform density of nuclear matter and also by the nearly uniform binding energy per nucleon.

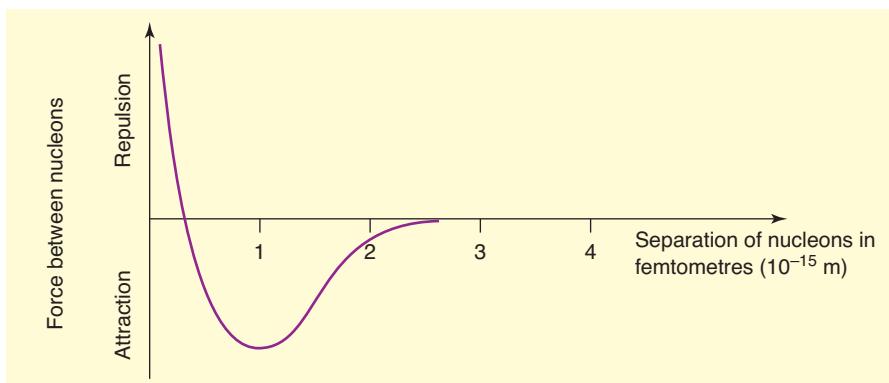


Figure 24.14 The graph shows how the strong nuclear force between two nucleons varies with the separation of the nucleons.

This ‘old’ strong force, which is well described by the exchange of pions, is now seen as a consequence of the complexities of the processes involving interactions between quarks and their messenger particles called gluons.

- a favouring of the binding of pairs of nucleons with opposite spins (see note on spin) and pairs of pairs with each pair having a total spin of zero. (This helps account for the exceptional stability of two protons and two neutrons in an alpha particle.)

The fundamental forces so far, with the exception of gravity, have been shown to involve the exchange of particles called ‘exchange’ particles or ‘messenger’ particles. In electromagnetism, the exchange particle is a photon, which is massless, and the electromagnetic force extends to infinity. The strong force as described above is carried by the pi meson or pion which is about 273 times heavier than an electron. The large mass is associated with a very short range force.

24.5

MASS DEFECT AND BINDING ENERGY OF THE NUCLEUS

The energy associated with a nuclear process is incredibly large compared to that associated with an atomic process. You have already encountered the equivalence of mass and energy and we now apply this concept to nuclear processes.

Mass defect

The key to the large energy involved in nuclear reactions is the fact that mass and energy are equivalent and are linked by Einstein’s relationship, $E = mc^2$. The other important fact is that the mass of any nucleus is *not the sum* of the masses of its constituent protons and neutrons. The difference between the mass of a nucleus and the total mass of its constituent nucleons is called the **mass defect** of the nucleus. Rather than define mass in kilograms, it is usual to use atomic mass units for the masses of nuclei. The conversion factor is:

$$1 \text{ atomic mass unit, } u = 1.661 \times 10^{-27} \text{ kg.}$$

The masses of protons, neutrons and electrons in atomic mass units are:

$$\text{mass of a proton, } m_p = 1.007\,276 \text{ u}$$

$$\text{mass of a neutron, } m_n = 1.008\,665 \text{ u}$$

$$\text{mass of an electron, } m_e = 0.000\,548\,580 \text{ u.}$$

The mass of a deuterium atom, an atom of the isotope of hydrogen, with a neutron as well as a proton in its nucleus, is 2.014 102 u. Therefore, the mass of a deuterium nucleus is $2.014\,102 - 0.000\,549 = 2.013\,553$ u (the mass of the atom – the mass of the electron).

The total mass of an isolated proton and an isolated neutron would be $1.007\,276 + 1.008\,665 = 2.015\,941$ u.

If this proton and neutron combined to form a deuterium nucleus, they would have to lose $2.015\,941 - 2.013\,553 = 0.002\,388$ u.

The mass defect of deuterium is 0.002 388 u and if a proton and a neutron combined, energy equivalent to a mass of 0.002 388 u would be released.

If more nucleons could be added to build bigger nuclei, energy would be released and the total mass defect would increase.

Binding energy

If we now tried to do just the opposite, that is, to split our deuterium nucleus into an isolated proton and neutron, we would find that it was not possible. There would not be sufficient mass for an isolated proton

The **mass defect** of a nucleus is the difference between the mass of the constituent nucleons and the mass of the nucleus.

It is possible to convert the mass defect in atomic mass units to a mass in kilograms and then use $E = mc^2$ to find the energy in joules that would be released. This energy in joules can then be converted to an energy in MeV. However, it is much easier to use the standard conversion factor where the energy equivalent of a mass of 1 u is 931.5 MeV. On data sheets this is stated as $1 \text{ u} = 931.5 \frac{\text{MeV}}{c^2}$.

The **binding energy** of a nucleus is the energy equivalent of the mass defect of the nucleus. It is the energy that would have to be provided and converted to mass to enable all the nucleons in a nucleus to be separated from each other.

The **average binding energy per nucleon** is the total binding energy of a nucleus divided by the number of nucleons in the nucleus. It is a measure of the stability of the nucleus.

and neutron to exist. If we really wanted to accomplish the separation, we would have to provide the missing mass, the mass defect of the deuterium nucleus. Somehow, we would have to supply energy to the deuterium nucleus and have that energy converted into mass. The exact amount of energy that would have to be converted into mass would be the energy equivalent of the mass defect. We call this energy the **binding energy** of the deuterium nucleus. The mass defect of the deuterium nucleus was $0.002\ 388\ \text{u}$. The equivalent energy is $0.002\ 388 \times 931.5 = 2.224\ \text{MeV}$. The binding energy of a deuterium nucleus is 2.224 MeV.

If a proton and neutron combined to form a deuterium nucleus, 2.224 MeV of energy would be released:



If we wanted to split a deuterium nucleus into an isolated proton and neutron, we would have to supply 2.224 MeV of energy that could be converted into mass.

We saw in the previous section that as the number of nucleons in a nucleus increased, so did the mass defect. This means that the total binding energy of the nucleus must also have increased. The total binding energy of a nucleus must be related to the stability of that nucleus, but it is difficult to obtain useful information from the total binding energy.

However, the stability of the nucleus is indicated by the **average binding energy per nucleon**. This gives a measure of how strongly an average nucleon is bound to a particular nucleus. The graph of average binding energy per nucleon against mass number (figure 24.15) shows that the most stable nuclei have a mass number of about 50 to 60. The most stable nucleus is, in fact, iron-56.

We can see from figure 24.15 that if we were able to join together light nuclei, we would produce nuclei with a higher average binding energy per nucleon and, hence, energy would be released. This is the process of nuclear fusion. Some atomic masses for light nuclides are given in table 24.2.

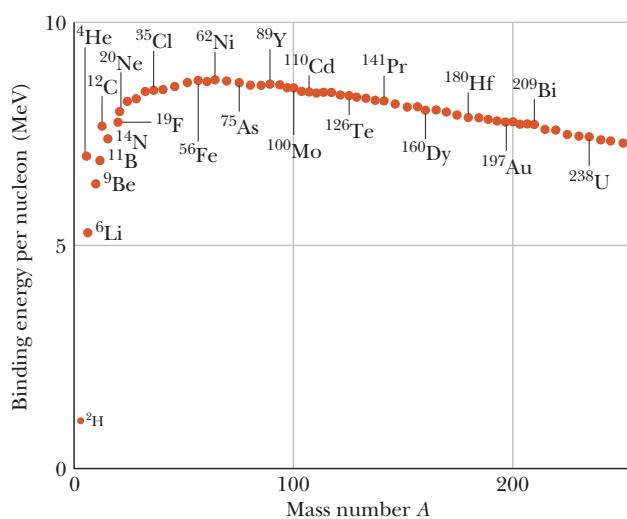


Figure 24.15 A graph of average binding energy per nucleon plotted against mass number

Also, if we were able to take a large mass number nucleus and split it in two, we would produce two new nuclei with higher average binding energy per nucleon than the original nucleus. Again energy would be released. This is the process of nuclear fission. We will study this process in chapter 25.

Table 24.2 Atomic masses for some light nuclides

ELEMENT AND ISOTOPE	NEUTRON NUMBER, N	ATOMIC NUMBER, Z	ATOMIC MASS (u)	MASS NUMBER, A
Hydrogen (${}^1_1\text{H}$)	0	1	1.007 825	1
Deuterium (${}^2_1\text{H}$)	1	1	2.014 102	2
Tritium (${}^3_1\text{H}$)	2	1	3.016 049	3
Helium (${}^3_2\text{He}$)	1	2	3.016 029	3
Helium (${}^4_2\text{He}$)	2	2	4.002 603	4
Lithium (${}^6_3\text{Li}$)	3	3	6.015 121	6
Lithium (${}^7_3\text{Li}$)	4	3	7.016 003	7
Beryllium (${}^9_4\text{Be}$)	5	4	9.012 182	9
Boron (${}^{10}_5\text{B}$)	5	5	10.012 937	10
Boron (${}^{11}_5\text{B}$)	6	5	11.009 305	11
Carbon (${}^{12}_6\text{C}$)	6	6	12.000 000	12
Carbon (${}^{13}_6\text{C}$)	7	6	13.003 355	13
Nitrogen (${}^{14}_7\text{N}$)	7	7	14.003 074	14
Nitrogen (${}^{15}_7\text{N}$)	8	7	15.000 109	15
Oxygen (${}^{16}_8\text{O}$)	8	8	15.994 915	16
Oxygen (${}^{17}_8\text{O}$)	9	8	16.999 131	17
Oxygen (${}^{18}_8\text{O}$)	10	8	17.999 160	18

SAMPLE PROBLEM

24.2

Determining mass defect and binding energy

The mass of a helium atom is 4.002 603 u.

- Calculate the mass defect of the helium nucleus.
- Calculate the total binding energy of the helium nucleus.
- Calculate the average binding energy per nucleon of helium.

SOLUTION

- The total mass of the constituents of a helium atom (two protons, two neutrons and two electrons) is:

$$2 (1.007 276 + 1.008 665 + 0.000 549) = 4.032 980 \text{ u.}$$

$$\begin{aligned} \text{Mass defect} &= 4.032 980 - 4.002 603 \\ &= 0.030 377 \text{ u} \end{aligned}$$

(b)

$$\begin{aligned} \text{Binding energy} &= \text{Mass defect} \times 931.5 \\ &= 28.30 \text{ MeV} \end{aligned}$$

$$\begin{aligned} (\text{c}) \quad \text{Average binding energy per nucleon} &= \frac{28.30}{4} \\ &= 7.08 \text{ MeV per nucleon} \end{aligned}$$

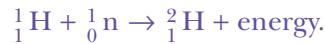
We use atomic masses when performing calculations to do with nuclear reactions (even if the reaction involves particles such as alpha particles that do not contain electrons). By doing so, we ensure that we have accounted for the mass of the same number of electrons on each side of the equation.

SAMPLE PROBLEM

24.3

Energy change in nuclear reactions

When we considered the formation of a deuterium nucleus from a proton and neutron, we were really dealing with the nuclear reaction:

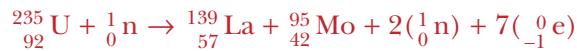


We know that energy was released because the product nucleus had less mass than the two reacting nucleons.

We can treat any nuclear reaction in the same way. If the mass of the products is less than the mass of the reacting nuclei, energy will be released. The energy released will be the energy equivalent to the decrease in mass.

Calculating the energy released in nuclear fission

A possible fission reaction for uranium-235 is given below. Find the energy (in MeV) released when one uranium-235 nucleus undergoes such a fission.



Atomic masses:

$${}^{139}\text{La} = 138.8061 \text{ u}$$

$${}^{95}\text{Mo} = 94.9057 \text{ u}$$

$${}^{235}\text{U} = 235.0439 \text{ u}$$

SOLUTION

$$\begin{aligned}\text{Total mass of reactants} &= 235.0439 + 1.008665 \\ &= 236.0526 \text{ u} \text{ (to four decimal places)}\end{aligned}$$

$$\begin{aligned}\text{Total mass of products} &= 138.8061 + 94.9057 + 2 \times 1.008665 + 7 \times 0.000549 \\ &= 235.7330 \text{ u} \text{ (to four decimal places)}\end{aligned}$$

$$\begin{aligned}\text{Decrease in mass} &= 236.0526 - 235.7330 \\ &= 0.3196 \text{ u}\end{aligned}$$

$$\begin{aligned}\text{Energy released} &= 0.3196 \times 931.5 \\ &= 297.7 \text{ MeV}\end{aligned}$$

SUMMARY

- Radioactivity was discovered in 1896. There were found to be three different components of the radiation and these were termed alpha, beta and gamma.
- It was found that transmutation occurred when an atom emitted an alpha or beta particle and an atom of a new element was formed.
- In alpha decay, the mass number decreased by four and the atomic number decreased by two. In beta decay, the mass number did not change but the atomic number increased by one.
- The properties of the nucleus could not be explained by assuming it contained protons and electrons and the existence of another nuclear particle, the neutron, was predicted. Chadwick discovered this particle in 1932.
- The spectrum of energies of beta particles emitted by nuclei seemed to indicate that neither energy nor momentum was conserved in the emission of beta particles. Pauli overcame this problem by predicting the existence of another neutral particle, the neutrino. Fermi used this prediction to explain beta decay.
- The emission of a beta particle involves overcoming the weak nuclear force.
- The gravitational and electrostatic forces cannot hold the nucleons in a nucleus together and another force which is incredibly strong over a very short range, the strong nuclear force, holds nucleons together.
- The mass of a nucleus is less than the masses of its constituent nucleons. This difference in mass is known as the mass defect and the energy equivalent to this mass is called the binding energy of the nucleus.
- The energy equivalent to the change in mass in a nuclear reaction is emitted or absorbed in the reaction.

QUESTIONS

1. Even though radioactive decay was discovered before the nucleus was identified, it was not thought to be a chemical reaction. State the properties of radioactive decay that exclude it from being a chemical reaction.

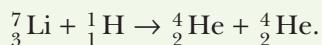
2. State the numbers of protons and neutrons in each of the following nuclei:
 - (i) aluminium-27
 - (ii) lead-208
 - (iii) radon-220
 - (iv) polonium-218
 - (v) uranium-238.
3. What changes in atomic number and mass number result from the emission of:
 - (a) an alpha particle?
 - (b) a beta particle?
 - (c) a gamma ray?
4. Write nuclear reactions for the following decays:
 - (a) polonium-214 emits an alpha particle
 - (b) thallium-210 emits a beta particle.
5. If nuclei do not contain electrons, how can beta decay be accounted for?
6. Complete the following nuclear equations:
 - (a) $^{212}_{?} \text{Pb} \rightarrow ^{212}_{?} \text{Bi} + ?$
 - (b) $^{212}_{?} \text{Bi} \rightarrow ? + ^0_{-1} \text{e}$
 - (c) $? \rightarrow ^{208}_{?} \text{Pb} + ^4_2 \text{He}$.
7. Complete the following nuclear equations:
 - (a) $^{27}_{13} \text{Al} + ^2_1 \text{H} \rightarrow ? + ^1_0 \text{n}$
 - (b) $^{10}_{5} \text{B} + ^4_2 \text{He} \rightarrow ? + ^1_1 \text{H}$
 - (c) $^{27}_{13} \text{Al} + ^4_2 \text{He} \rightarrow ? + ^1_1 \text{H}$
 - (d) $? + ^4_2 \text{He} \rightarrow ^{35}_{17} \text{Cl} + ^1_1 \text{H}$
 - (e) $^{9}_{4} \text{Be} + ^1_1 \text{H} \rightarrow ? + ^4_2 \text{He}$
 - (f) $^{22}_{11} \text{Na} + ^4_2 \text{He} \rightarrow ? + ^1_1 \text{H}$.
8. What property of neutrons makes them particularly useful for producing nuclear reactions?
9. Why was a particle that it was initially thought would be impossible to detect predicted to be involved in beta decay?
10. (a) Why don't the repulsive electrostatic forces between protons cause the protons in a nucleus to fly apart?
(b) Why would the electrostatic forces between protons have a greater chance of making a large nucleus rather than a light one break apart?
11. If it was possible for a nucleus of carbon-12 to absorb a neutron into its nucleus and form carbon-13, calculate the amount of energy that would be released.

The atomic masses are:

$$\text{carbon-12} \quad 12.000\,000 \text{ u}$$

$$\text{carbon-13} \quad 13.003\,354 \text{ u.}$$

12. When a proton is fired at a lithium nucleus, a nuclear reaction in which two alpha particles are produced may occur:

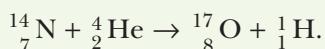


Ignoring any initial kinetic energy of the lithium atom and the proton, calculate the total kinetic energy of the two alpha particles after the reaction.

The atomic mass of ${}^7_3\text{Li}$ is 7.016 003 u.

Note: even though we are dealing with a proton and two alpha particles in this reaction, we still use the atomic masses of hydrogen and helium. By doing so, we ensure that we have accounted for the mass of the same number of electrons on each side of the equation.

13. The first artificial nuclear transmutation which Rutherford performed in 1919 was the reaction between an alpha particle and a nitrogen nucleus. The alpha particles used in the experiment were emitted from bismuth-214.



- (a) Use the masses provided to determine the change in energy (in MeV) that occurred in the reaction.

- (b) You should have noticed that energy was absorbed in the reaction. What was the source of this energy?

The atomic masses are:

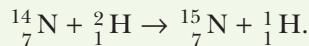
$${}^{14}_7\text{N} = 14.003\,074 \text{ u}$$

$${}^{17}_8\text{O} = 16.999\,131 \text{ u}$$

$${}^4_2\text{He} = 4.002\,603 \text{ u}$$

$${}^1_1\text{H} = 1.007\,825 \text{ u.}$$

14. Calculate the amount of energy released in the following nuclear reaction:



The atomic masses are:

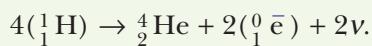
$${}^{14}_7\text{N} = 14.003\,074 \text{ u}$$

$${}^2_1\text{H} = 2.014\,102 \text{ u}$$

$${}^{15}_7\text{N} = 15.000\,108 \text{ u}$$

$${}^1_1\text{H} = 1.007\,825 \text{ u.}$$

15. The process of nuclear fusion occurring in the Sun involves a number of steps but can be summarised in the equation:



How much energy is released in this process?

16. Is total binding energy or average binding energy per nucleon a better indicator of the stability of a nucleus? Explain your answer.

CHAPTER 25

NUCLEAR FISSION AND OTHER USES OF NUCLEAR PHYSICS



Figure 25.1 An atomic bomb test. The first atomic test took place at Alamogordo in New Mexico at 5:29:45 am, 16 July 1945. William Laurence, the official journalist and only ‘outsider’ present at the test stated, ‘It was like the grand finale of a mighty symphony of the elements, fascinating and terrifying, uplifting and crushing, ominous, devastating, full of great promise and great forebodings’.

Remember

Before beginning this chapter, you should be able to:

- recall that the strong nuclear force between nucleons is a very strong but very short range force
- recall that energy will be released in a nuclear reaction if there is a decrease in mass in that reaction
- recall that fission of a heavy nucleus releases energy.

Key content

At the end of this chapter you should be able to:

- describe Fermi’s attempts to produce transuranic elements and explain why the interpretation of his observations changed after the discovery of nuclear fission
- describe Fermi’s first demonstration of a nuclear chain reaction in an atomic pile in 1942
- compare the requirements for a controlled and an uncontrolled nuclear chain reaction
- explain the basic principles of a fission reactor
- describe the use of a named isotope in each of the fields of medicine, engineering and agriculture
- explain, by referring to the properties of neutrons, why neutron scattering is used as a probe
- assess the significance of the Manhattan Project to society.

25.1 ENERGY FROM THE NUCLEUS

In 1903, Rutherford was already aware of the vast amount of energy available from radioactivity. He concluded a lecture by commenting that, based on Pierre Curie's experiments, each gram of radium gave out sufficient energy in its lifetime to raise 500 tonnes a mile high. In 1916, he commented on the possibility of a nuclear bomb: 'Fortunately at the present time we had not found out a method of so dealing with these forces, and personally I am hopeful we should not discover it until man was living at peace with his neighbour.'

In 1933, Rutherford made a famous statement that suggested that it would be impossible to obtain energy from the nucleus, 'It is a very poor and inefficient way of producing energy, and anyone who looked for a source of power in the transformation of atoms was talking moonshine.' Perhaps he had changed his mind from his earlier views or perhaps he hoped that it would not be possible.

PHYSICS FACT

A chain reaction

Leo Szilard (1898–1964) recalled reading a report of Rutherford's 1933 statement in *The Times*. After reading it he walked through London, stopped at a red light on the corner of Southampton Row and wondered if Rutherford might be wrong. He thought about firing neutrons, not alpha particles at a nucleus and realised as the light turned green 'that if we could find an element which is split by neutrons and which would emit *two* neutrons when it absorbs *one* neutron, such an element, if assembled in sufficiently large mass, could sustain a nuclear chain reaction.'

PHYSICS IN FOCUS

Sir Mark Oliphant

Sir Mark Oliphant (1901–2000), one of Australia's great scientists, worked with Rutherford at the Cavendish Laboratory for the ten years before Rutherford's death and described it as the most wonderful time of his life. Sir Mark recalled that in about 1934 or 1935, while Rutherford was absent from the Cavendish Laboratory, he had performed some experiments. The aim of these experiments was to see if it was possible to get a net gain of energy by modifying an experiment in which deuterium atoms were bombarded with accelerated deuterium nuclei. He obtained a negative result but when Rutherford returned and Sir Mark informed him of the experiment, Rutherford was at first very angry. Sir Mark suggested to Rutherford's biographer, John Campbell, that perhaps Rutherford was aware of the enormous energy available and had hoped that energy would never be able to be efficiently extracted from the nucleus.



Figure 25.2 Sir Mark Oliphant

(continued)

Sir Mark Oliphant was later a member of the Manhattan Project that developed the atomic bomb. It is ironic that Sir Mark worked on the development of the fission bomb and had discovered the reaction that occurs in a fusion bomb as he was opposed to the use of nuclear weapons.

Sir Mark Oliphant recalled of his work with Rutherford:

In this work, which we did together, we were able to discover two new kinds of atomic species; one was hydrogen of mass 3 (tritium) unknown until that time, and the other helium of mass 3, also unknown. These new atoms were produced as a result of atomic transformations induced by our ion beam hitting targets of lithium, beryllium and other materials. Incidentally, at the same time, we were able to show that heavy hydrogen nuclei, that is to say the cores of heavy hydrogen atoms, could be made to react with one another to produce a good deal of energy and new kinds of atoms. This particular reaction, which we discovered at this time, is the basic reaction in the so-called hydrogen bomb.

Sir Mark modestly did not state that the work leading to these discoveries was possible because

he had constructed a particle accelerator that was able to accelerate charged particles to the energies necessary to induce the reactions. (The reactions could be induced with earlier accelerators but his accelerator achieved higher reaction rates.)

The reaction induced between ‘the cores of heavy hydrogen atoms’ produced helium-3 and a neutron.



In 1934, Sir Mark duplicated his accelerator and operated it while Rutherford gave the first public display of a nuclear fusion reaction to a Friday night meeting of the Royal Institution in London.

After the war, Sir Mark returned to the University of Birmingham in England but in 1950 moved back to Australia as the Director of the Research School of Physical Sciences at the newly established Australian National University. He established the Australian Academy of Science and was its first president (1954–1956).

He retired from ANU in 1967 and was appointed to a five-year term as State Governor of South Australia in 1971. He retired to Canberra in 1976 and continued to promote science and technology and to foster the development of Australian science until his death in July 2000.

25.2 THE DISCOVERY OF NUCLEAR FISSION

After the discovery of the neutron, Enrico Fermi (1901–1954) and his co-workers in Rome led the world in neutron physics. They set out to study the neutron bombardment of as many elements as possible. With heavy elements there was often a delayed emission of a beta particle, which resulted in the production of an element with a higher atomic number. This happened with uranium (see figure 25.3) and suggested that an element with an atomic number greater than 92 was formed.

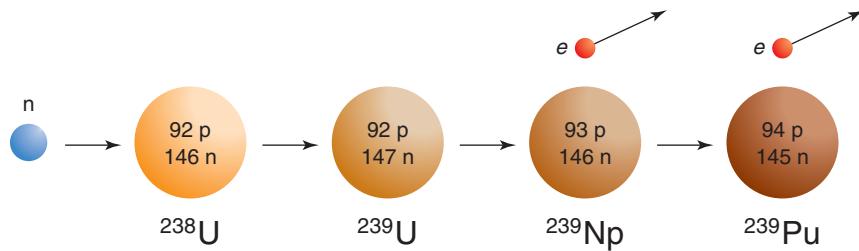


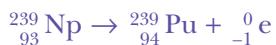
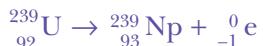
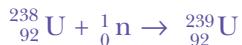
Figure 25.3 Bombardment of uranium with neutrons may produce transuranic elements.

PHYSICS FACT

Transuranic elements

Half-life is the time taken for half the radioactive nuclei in a sample to decay. If we exclude the activity of daughter nuclei, it is the time taken for the activity of a particular sample to drop to half its initial value.

It is possible to identify the products of neutron bombardment by measuring the **half-lives** of the radiation it emits. In 1934, Fermi reported that an activity with a half-life of 13 minutes was not due to any known isotope of any element and must be due to an element with an atomic number greater than 92. Elements with an atomic number greater than 92 are known as transuranic elements.



It is possible for a uranium-238 nucleus to capture a neutron and then form plutonium-239 after emitting two beta-particles.

(This reaction does occur and it is how plutonium can be produced for use in reactors or weapons. However, after the discovery of nuclear fission, it was realised that radiation originally attributed to transuranic elements was really associated with isotopes of much lighter elements.)

Fermi's neutron bombardment of uranium

We now know that when Fermi and his associates bombarded uranium with neutrons, they must have produced nuclear fission. The production of transuranic elements would also have occurred but the observation of radioactivity that Fermi took as evidence for the production of transuranic elements was certainly due to the production of fission products.

In his biography of Fermi, Emilio Segré (1905–1989), a co-worker and also a Nobel prize winner, comments: ‘As is well known, the discovery of fission required a reappraisal of the work on the radioactivities induced in uranium. It has occasionally been said that Fermi’s was the only Nobel Prize awarded for an unsubstantiated discovery — on the assumption that discovery of the transuranic elements was the reason for the prize. It is clear from the presentation statement, however, that this was not the case’ (see the next section, page 478).

The first observations that confirmed nuclear fission were made by Otto Frisch using an ionisation chamber to detect the particles emitted (see page 478). Interestingly, Fermi and his group performed a similar experiment in early 1935 when trying to detect alpha radiation from their supposed transuranic elements. They were able to detect beta radiation but were unable to detect the alpha radiation they thought should be emitted.

Segré recalls that they reasoned that the alpha particles emitted from the transuranic elements would be more energetic than those emitted from uranium. They set up a sample of uranium in front of an ionisation chamber and irradiated the uranium with neutrons from a source surrounded by paraffin. They did not want to detect the natural alpha radiation from the uranium, so they covered the uranium with aluminium foil. They hoped this would stop the alpha particles emitted from the uranium but would not stop the higher energy alpha particles they expected to be emitted from their new elements. They did not manage to detect anything at all and Segré recalls that the aluminium foil prevented

them from observing the large ionisation pulses associated with fission. He also states that even if they had observed these pulses, it is impossible to say if they would have interpreted them correctly.

Slow neutrons are more effective than fast neutrons

During the course of their neutron bombardment experiments, Fermi's co-workers, Edoardo Amaldi (1908–1989) and Bruno Pontecorvo (1913–1993) discovered that the same experiment yielded different results when performed in different parts of the same room. Most noticeably, the activity was much greater after substances had been irradiated with neutrons on a wooden table rather than on a marble table. The results did not make sense.

Fermi set out to investigate this strange phenomenon and planned to use a block of lead between the neutron source and the target. He had it carefully machined, not his usual custom, and then at the last minute changed his mind and substituted a rough piece of paraffin. The result was a spectacular increase in the intensity of the activation.

Fermi had discovered that slow neutrons were much better at irradiation than fast neutrons. Neutrons did not need to have a large energy to closely approach a nucleus because there was no electrostatic repulsion of the neutron. Neutrons travelling at low speeds spent more time in the vicinity of the nucleus and hence had a better chance of being captured. (This is associated with the de Broglie wavelength of the neutrons. The slower neutrons have a much longer wavelength and hence have a much greater possibility of capture by a nucleus.) The action of the paraffin is similar to that of a moderator in a nuclear reactor (see page 486).

Fermi was awarded the Nobel prize in 1938 for ‘...your discovery of new radioactive substances belonging to the entire race of elements and for the discovery you made in the course of the selective power of slow neutrons.’ The winning of the Nobel prize was a most fortuitous event for Fermi. Fermi and his Jewish wife, Laura, were granted permission to travel to Sweden to accept the prize but then did not return to Italy and fled to the USA. Many other Jewish physicists had escaped from Germany.

Near the end of 1938 Lise Meitner and her nephew Otto Frisch, two others who had fled, realised that earlier experiments indicated the fission of uranium with the potential release of a vast amount of energy.

Meitner and Frisch identify fission

Before fleeing Germany, Lise Meitner (1878–1968) had been working with Otto Hahn (1879–1968) and Fritz Strassmann (1902–1980). They studied the substances into which the heaviest elements transmuted after neutron bombardment. While on a Christmas holiday in Sweden with her nephew Otto Frisch (1904–1979), Meitner received a letter from Hahn. Hahn wrote that he had identified barium rather than radium as one of the products of bombardment of uranium with slow neutrons.

Meitner replied to Hahn, ‘Your radium results are very amazing. A process that works with slow neutrons and leads to barium! ... To me for the time being the hypothesis of such an extensive burst seems very difficult to accept, but we have experienced so many surprises in nuclear physics that one cannot say without hesitation about anything: “It’s impossible”.’ When Frisch arrived they discussed Hahn’s letter. Later Frisch summarised the discussion. They noted that there was insufficient

energy present to chip enough protons and alpha particles off a uranium nucleus to produce barium, and no other particles had been observed. It was impossible for a nucleus to be cleaved across. Then they began to consider Bohr's liquid drop model of the nucleus. They considered that if the nucleus was like a liquid drop, it could become unstable and split in two. The two parts would be forced apart by the electrostatic repulsion between them. (The strong nuclear force has already been overcome once the two parts are separated.) This process is shown in figure 25.4. They calculated that the two parts would be forced apart at extremely high velocities corresponding to an energy of about 200 MeV.

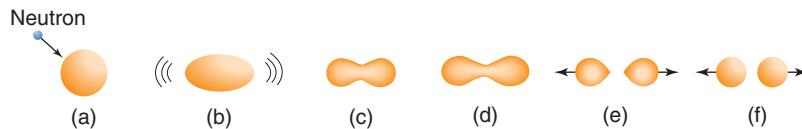


Figure 25.4 Stages in a fission process (a) A ^{235}U nucleus absorbs a thermal neutron. (b) A ^{236}U nucleus with excess energy is formed and oscillates. (c) The oscillations may cause a neck to form but the strong nuclear force is still in control. (d) As the neck narrows, the electrostatic repulsion begins to overcome the strong nuclear force. (e) The electrostatic force causes the neck to break. (f) Now that the two fragments have been separated, there is no attraction but only the electrostatic repulsion which forces the two fragments apart at high velocity.

Meitner was able to calculate the energy with the use of the mass defects of the nuclei that she had memorised. She calculated that if the uranium nucleus did indeed split in two, the mass of the products would be approximately equal to the mass of the uranium nucleus minus one fifth the mass of a proton. Using $E = mc^2$ they calculated that about 200 MeV would be released.

Their calculations confirmed the fission of uranium and explained the origin of the barium. When the uranium nucleus split in two, barium was one of the many elements that could be formed.

The fission reaction that produces barium is:



Another example of one of the many fission reactions that may occur is:



The first observations of fission

Frisch returned to Copenhagen and informed Bohr of their theory just as Bohr was about to travel to the USA. A few days later, Frisch performed the first experiment that actually confirmed the fission reaction. Uranium was bombarded with neutrons and the particles emitted were passed into an ionisation chamber. The highly energetic nuclei that were the products of the fission of uranium were easily detected. The output from the ionisation chamber was viewed on an oscilloscope.

News of fission reached the USA before Frisch and Meitner had published their paper and many others performed experiments similar to that of Frisch.

In Chicago, Herbert Anderson, a graduate student, and Fermi set out to perform such an experiment. Their later recollections differed but it seems that Fermi went to a conference in Washington and did not actually observe the fission.

PHYSICS FACT

Fission

Otto Frisch demonstrated his fission reaction to many people. One was an American biologist, William Arnold, who was studying in Copenhagen. Frisch asked Arnold what biologists called it when a bacterium split in two and Arnold replied that it was binary fission. Frisch asked if he could use the term fission by itself and hence the biological term associated with reproduction of cells became the term used in physics for the process in the most destructive weapon that has ever been used in war.

25.3

THE DEVELOPMENT OF THE ATOM BOMB

eBookplus

Weblinks:

The Quebec Agreement
Manhattan Project files
Einstein's letter to
President Roosevelt

Although nuclear fission was recognised as a reality, the feasibility of an atom bomb was still doubtful. Leo Szilard's idea of a chain reaction was the key but there was doubt about the mass of uranium required to make a bomb. Szilard was sure that it would be possible and lobbied the US government to proceed. In 1939 he approached Albert Einstein who wrote a letter to President Roosevelt outlining his belief that America should be actively researching the possibility of a nuclear bomb. This letter has since become famous.

Others had also realised that a chain reaction might be possible. Preliminary research had also taken place in England. Just a few hours before Japan entered the war, the US government made a commitment to expand its nuclear research. In 1942, the US government made an agreement to work with the British government. The 'Manhattan Project', where an atomic bomb was planned to be designed and constructed, was commenced. Britain supplied the information it already possessed but it was unclear how much information would be shared after the war.

PHYSICS FACT

Early ideas of an atomic bomb

Otto Frisch returned to the University of Birmingham where he worked with Mark Oliphant and another expatriate German Jew, Rudolf Peierls (1907–1995). In 1940, Oliphant worked on the top secret radar development, but Frisch and Peierls were both technically enemy aliens and were banned from the radar work.

Frisch and Peierls, who later worked with Oliphant on the Manhattan Project, did their own work on the atomic bomb. They considered that the possibility of using pure uranium-235 as the fuel had been overlooked and concluded that as little as 1 kg of uranium-235 would be suitable for a bomb.

About 40 kg of uranium was eventually used and it was about 89% U-235.

The Manhattan Project

It was thought that there were two different pathways to an atomic bomb, using uranium-235 as fuel or using plutonium-239 as fuel. (The plutonium nucleus contains two protons and two neutrons more than a uranium-235 nucleus.)

The problem was to produce sufficient of each of these. Not knowing which would be better, the United States Government decided to proceed with both methods. Vast plants were constructed to produce pure U-235 from natural uranium and to produce Pu-239 from the neutron bombardment of U-238.

eBook plus

Weblink:

Events at Stagg Field: The first atomic pile

An eyewitness account of the events at Stagg Field on 2 December 1942, by Corbin Allardice and Edward R. Trapnell.

The first nuclear reactor

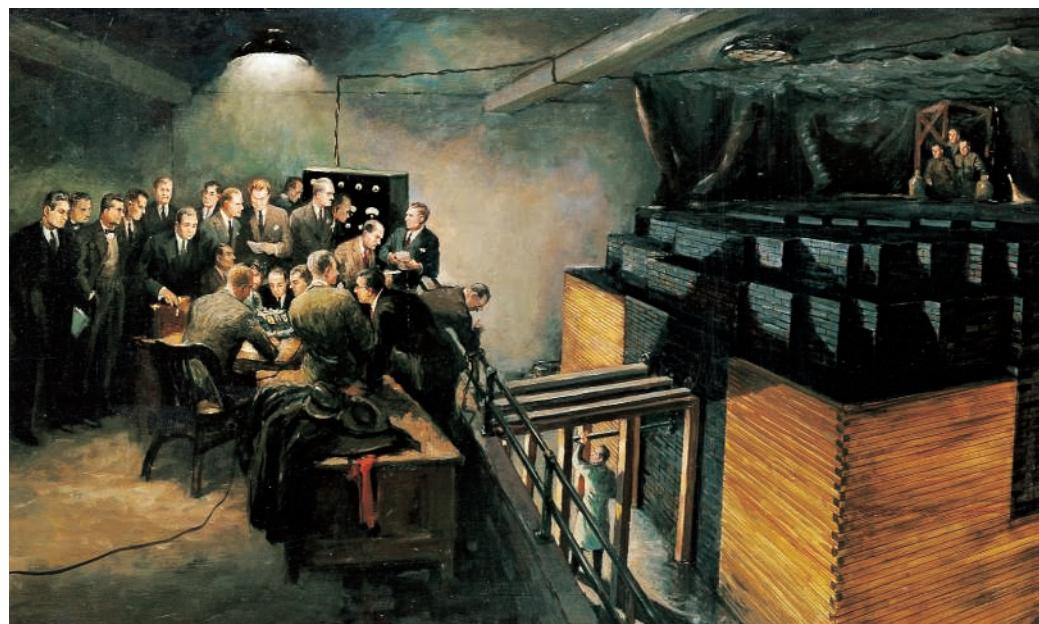
The first man-made nuclear reactor, or atomic pile as it was then known, was built in a squash court under a grandstand at Stagg Field in Chicago. Enrico Fermi was the head of the group which designed and constructed the reactor. During its design and construction, Fermi and his group solved many of the problems associated with nuclear reactors. They took out patents on 'neutronic reactors' but, after the war, assigned these patents to the US government without compensation.

The aim was to see if it was possible to obtain a neutron multiplication factor greater than one. This would mean that a chain reaction could occur. If a chain reaction did occur, it would create a means of producing plutonium for use as a fuel in atomic bombs.

Fermi realised that impurities in the uranium and the moderator would result in the capture of some neutrons. This, in turn, would prevent them from causing fissions. The question was whether sufficient neutrons would produce fissions to enable a chain reaction to occur. (Even before the success of this reactor, construction of full-scale reactors designed to produce plutonium had commenced. These reactors were constructed at Hanford in the state of Washington.)

Fermi's atomic pile contained 50 tonnes of natural uranium in the form of 22 000 slugs. These were dispersed throughout 400 tonnes of graphite which had been machined into 40 000 graphite bricks. Graphite was used as the moderator (see page 486) because at that time it was the only material that was available in sufficient quantity and of the required degree of purity.

Figure 25.5 The first nuclear reactor. Few photographs were taken but paintings have been reproduced many times. Enrico Fermi slowly withdrew the control rods and the radiation produced was measured by Geiger counters and plotted on a chart recorder.



The test of the pile occurred on 2 December 1942 and is shown as a sketch in figure 25.5 on the previous page. At 9.45 am they started slowly withdrawing the neutron-absorbing control rods. After each six- or perhaps 12-inch step, Fermi performed calculations on his slide rule and predicted where the trace on the chart recorder would level off. At about 11.45 am the automatic safety rod, which had been set at too low a level, was triggered. Fermi called a break for lunch and the process was resumed at 2.00 pm. At 3.25 pm Fermi predicted that the trace would now not level off and a few minutes later reported that the reaction was self-sustaining. Twenty-eight minutes later the control rods were inserted and the reaction was stopped.

Research at Los Alamos

The theoretical work on the atomic bomb was done at Los Alamos in New Mexico where a new secret laboratory was built on the site of a boys ranch school about 100 km from Santa Fe.

The greatest scientists in the free world gathered at Los Alamos. They all realised the terrible potential of the weapon that they were designing; however, they were initially driven by the fear that German scientists, and there were still famous atomic scientists in Germany, would build the atomic bomb and give Hitler incredible power.

Once the project commenced, there was the same excitement that accompanies any new scientific work, in this case magnified by the presence of so many great scientific minds.

The problem was to assemble a mass greater than the **critical mass** in a very short period of time. Different methods were used for the different fuels. The plutonium bomb had normal explosives packed around a sphere of plutonium. The plutonium was slightly less than a critical mass. The explosives forced the sphere of plutonium into a smaller sphere of super-critical mass (see figure 25.6). (The increase in density caused by the compression from detonating the explosives resulted in the formation of a super-critical mass.)

The smallest amount of fuel necessary to sustain a chain reaction is called the critical size or **critical mass**. As the size increases, the volume to surface area ratio increases and a smaller proportion of neutrons are lost from the fuel.

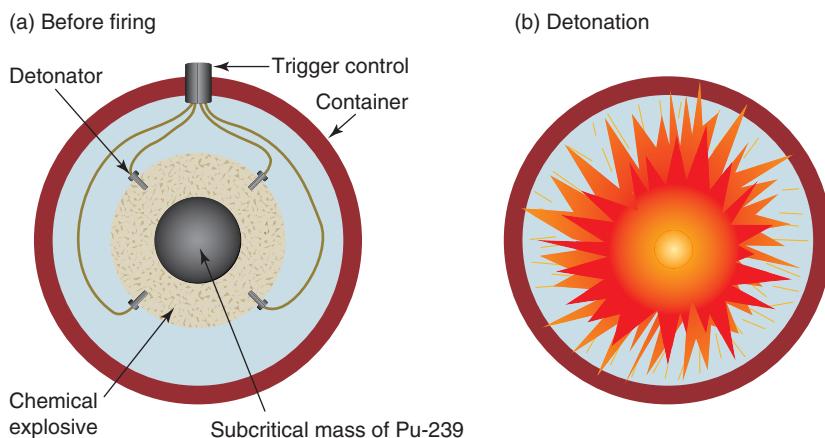
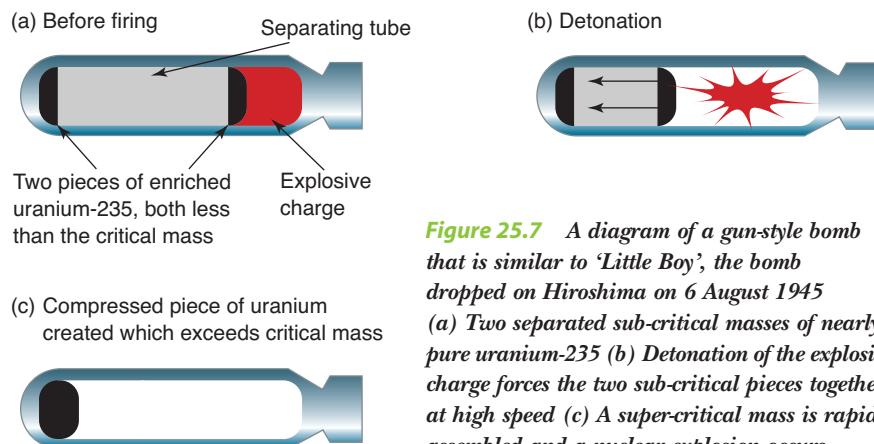


Figure 25.6 A diagram of an implosion bomb that is similar to 'Fat Man', the bomb dropped on Nagasaki on 10 August 1945 (a) A slightly sub-critical mass of plutonium is surrounded by a shell of explosives (b) Detonation of the explosive compresses the plutonium, producing a super-critical mass and then a nuclear explosion.

The uranium bomb was simpler with two sub-critical masses of uranium being fired into each other to assemble the super-critical mass (see figure 25.7). This was known as the gun method.

The implosion method for plutonium was tried in the first test of an atomic bomb, the Trinity test at Alamogordo in New Mexico. It was not thought necessary to test the gun method because it used a simple method of assembling the super-critical mass. Therefore, the uranium bomb was dropped without a test.



Heisenberg's recollection of the 1941 meeting was conveyed to Robert Jungk and printed in Jungk's book Brighter than a Thousand Suns: A Personal History of the Atomic Scientists. It is rather different from Bohr's recollection. Letters exist in the Bohr archives that Bohr wrote to Heisenberg but did not send. One letter was found in the pages of Bohr's own copy of Jungk's book. It was reported that Bohr took exception to Heisenberg's description of their meeting. In Jungk's book, Heisenberg suggests that he proposed that physicists should not work on such a weapon and that he would see that German physicists would not build an atomic bomb if Bohr would use his influence to stop Allied physicists from building one. The conversation was spoken in guarded terms and it is possible that neither really understood what the other was referring to. Michael Frayn's play Copenhagen deals with this meeting of Heisenberg and Bohr. The letters, which were to be released in 2012 on the fiftieth anniversary of the death of Bohr, were released in 2002 because of the intense interest created by the play. As can be seen, Bohr made a number of attempts to write to Heisenberg but did not actually send him a letter.

Figure 25.7 A diagram of a gun-style bomb that is similar to 'Little Boy', the bomb dropped on Hiroshima on 6 August 1945

(a) Two separated sub-critical masses of nearly pure uranium-235 (b) Detonation of the explosive charge forces the two sub-critical pieces together at high speed (c) A super-critical mass is rapidly assembled and a nuclear explosion occurs.

Speculation remains as to the state of the German atomic bomb project. Werner Heisenberg, who was in charge of the German project, actually met Niels Bohr in Copenhagen in 1941 before Bohr fled to the USA. Apparently, Bohr believed that Heisenberg did not realise that a small critical mass was possible if U-235 was used. Bohr thought that Heisenberg was building something more like a nuclear reactor although he did gain the impression that there had been a major German nuclear effort.

After the end of the war with Germany, ten of the top German scientists who were thought to have made contributions to nuclear weapon research were taken to England and held in Farm Hall for many months. Their conversations were bugged and transcripts indicate the surprise of the scientists after the atomic bomb was dropped on Hiroshima. (In fact they thought at first that it was simply a propaganda trick.) The transcripts then indicate that the German scientists came up with their own version, or *Lesart*, describing their view of atomic weapons during the war. This version was that the German scientists had not wanted the atomic bomb either because it was impossible to achieve during the expected duration of the war or because they simply did not want to create an atomic bomb.

Views of some Manhattan Project physicists

Work on the atomic bomb did not stop after Germany was defeated. It is not hard to see that the Manhattan Project scientists wanted to see the conclusion of their work.

Leo Szilard saw no need to continue after the defeat of Germany and in an attempt to stop the bomb being dropped, organised a petition to President Truman. (It was Szilard who had played an important part in starting the bomb project and had lobbied Einstein to convince President Roosevelt of the need for the project.)

After the successful test of the first plutonium bomb at Trinity in 1945, many scientists did not want to see it used as a weapon. There were attempts to stop it from being used. Some wanted to invite Japanese leaders to a demonstration of the weapon. However, it was too late for that; the politicians now had control.

We have seen previously (page 476) that Sir Mark Oliphant was strongly opposed to the use of nuclear weapons. This was a fairly common view of Manhattan Project scientists. An interesting view of many of the scientists was that the atomic bomb would see the end to war as it would be too horrible to even contemplate a nuclear war. Others were of the view that the secrets should be shared. Little did they know that the Russians already knew the important details, as the result of

eBook plus

Weblink:
Documents from the
Niels Bohr Archive

espionage and the cooperation of some Los Alamos scientists, and that a nuclear arms race was about to begin.

Richard Feynman (1918–1988) recalled that after the Trinity test there were parties, but one man, Bob Wilson, who was responsible for Feynman becoming involved with the project, was just moping around. Wilson said that they had done a terrible thing. Feynman later said, ‘You see, what happened to me — what happened to the rest of us — is that we started for a good reason, then when you’re working very hard to accomplish something and it’s a pleasure, it’s excitement. And you stop thinking, you know; you just stop. Bob Wilson was the only one who was still thinking about it at that moment.’

One of the Manhattan Project physicists, Nobel prize winner Hans Bethe (1906–2005), wrote an open letter in 1994 calling for scientists to cease work on the production of weapons of mass destruction.

Open letter from Hans Bethe

As the Director of the Theoretical Division at Los Alamos, I participated at the most senior level in the World War II Manhattan Project that produced the first atomic weapons.

Now, at age 88, I am one of the few remaining such senior persons alive. Looking back at the half century since that time, I feel the most intense relief that these weapons have not been used since World War II, mixed with the horror that tens of thousands of such weapons have been built since that time — one hundred times more than any of us at Los Alamos could ever have imagined.

Today we are rightly in an era of disarmament and dismantlement of nuclear weapons. But in some countries nuclear weapons development still continues. Whether and when the various nations of the world can agree to stop this is uncertain. But individual scientists can still influence this process by withholding their skills.

Accordingly, I call on all scientists in all countries to cease and desist from work creating, developing, improving and manufacturing further nuclear weapons; and, for that matter, other weapons of potential mass destruction such as chemical and biological weapons.

Hans A. Bethe

25.4 NUCLEAR FISSION REACTORS

After the release of energy by an atomic bomb had been achieved, fission reactors followed. The principles of a fission reactor are similar to those of an atomic bomb except that the release of energy is controlled.

In an atomic bomb the release of energy has to occur in an extremely short time. All the **fissile** nuclei present must capture neutrons and undergo fission. As many as possible of the neutrons produced in each fission must produce further fissions. Because of the very short time involved, slow neutrons are useless in a bomb. The fuel would be blown apart before the neutrons could be slowed or captured.

Neutrons in a nuclear reactor

In a nuclear reactor that has reached its desired power level, one neutron from each fission must produce another fission to maintain the reaction at a steady rate.

We will examine the principles of the operation of fission reactors that use uranium as their fuel.

If we consider the interaction between a neutron and a nucleus, the neutron is either scattered, with the loss of some energy, or captured

A **fissile** nucleus is a nucleus that may undergo fission.

(see figure 25.8). If it is captured by a fissile nucleus there is a possibility of fission. Of course, another alternative is that the neutron may simply escape from the fuel.

Natural uranium has 0.7% fissile U-235 and the remainder is non-fissile U-238. If the percentage of U-235 is increased by the process of enrichment, the amount of uranium needed as fuel can be reduced. Similarly, if enriched uranium is used, the problem of the loss of neutrons is reduced and less efficient moderators can be used.

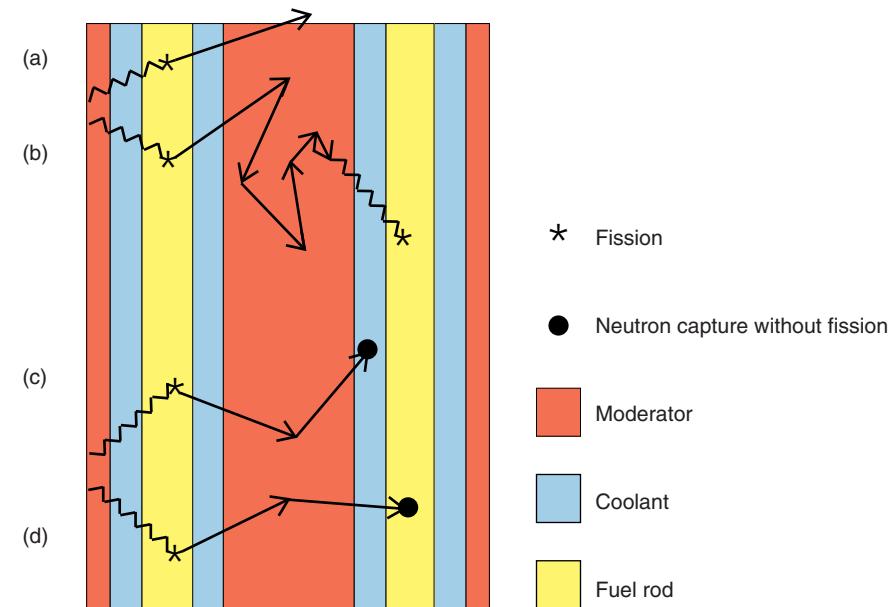


Figure 25.8 Neutrons in a nuclear reactor (a) Some neutrons escape (b) High-energy neutrons released during one fission should collide with nuclei in the moderator, losing much of their energy as they do so. When their energy has been greatly reduced, they may be captured by a fissile nucleus and produce another fission. The zig-zag line in the last part of the path before fission indicates a thermal neutron. (c) Some neutrons are captured by non-fissile nuclei. (d) Some neutrons are captured by a fissile nucleus but do not produce fission.

The problem of leakage of neutrons

Neutrons are very highly penetrative and may travel a large distance through matter before being captured, perhaps escaping from the fuel. As the size of the fuel is increased, the number of neutrons that escape will decrease, and when the critical mass or critical size is reached the problem of leakage is overcome.

The problem of the energy of neutrons

Fission is induced most efficiently by low-speed or ‘thermal’ neutrons. The fast neutrons produced in fissions should be slowed by a moderator to enhance the probability of producing further fission.

The capture of neutrons

The fact that low-speed neutrons rather than high-speed neutrons are more likely to be captured by a uranium-235 nucleus is associated with the wavelength of the neutron. A slow or ‘thermal’ neutron has an energy of about 0.025 eV and hence has a de Broglie wavelength of about 1.6×10^{-11} m. It may still interact with a nucleus even if it passes this far from it.

A 1.0 MeV neutron, however, has a de Broglie wavelength of only 2.5×10^{-15} m and hence must make a very close approach to the nucleus if it is to interact with it.

Robert Serber, who prepared the *Los Alamos Primer*, compared this to an archer shooting at a target. It is as if for fast arrows, the target has a diameter of one metre but for slow arrows, the target has magically expanded to 13 metres in diameter!



An unfortunate result of slowing the neutrons is that they must be slowed through an energy range where they are likely to be captured but then do not produce fission. Capture of neutrons with an energy in this range can be reduced by not mixing the fuel and moderator uniformly. The uranium fuel is present as uranium oxide in the fuel rods. A neutron produced in a fission in a fuel rod should slow while in the moderator and then re-enter another fuel rod where it may produce another fission.

PHYSICS FACT

Moderators

A moderator should contain nuclei with a low mass. If a neutron undergoes a collision with a mass much larger than itself, it will rebound elastically from the mass (imagine a ping pong ball bouncing off a bowling ball). Momentum and energy are conserved in the collision and the ping pong ball rebounds with almost no change in energy. However, if an object collides with another object of similar mass, it will pass all, or almost all of its energy on to the second object. This suggests that the best way to slow down neutrons is to have them collide with protons. Unfortunately, however, there is a high probability of the proton capturing the neutron. The next best way is for the neutron to

collide with a deuterium nucleus. The chance of capture, and hence forming tritium, ${}^3_1\text{H}$, is very low. While this sounds ideal and is used in some reactors, there is the drawback that heavy water, also known as deuterium oxide, is expensive to produce. Carbon, in the form of graphite, is an alternative and, while not as efficient as heavy water at slowing neutrons, it is economically viable.

Some nuclear reactors use water as a moderator and accept the fact that some neutrons will be lost as they are captured by the hydrogen nuclei (protons). Those reactors must use enriched uranium to compensate for this loss of neutrons.

Coolant

As we have seen, the products of a fission are fired apart with extremely high kinetic energies. This kinetic energy is transferred to the atoms and molecules in the reactor core as thermal energy. A coolant is used to extract this thermal energy. If enriched uranium is used as the fuel, it is possible to use ordinary water (under pressure) as the coolant. The absorption of neutrons by the hydrogen nuclei in the water is compensated for by increasing the percentage of U-235.

If heavy water is used as the moderator, it also performs the role of the coolant. A pressurised water reactor (PWR) uses water under high pressure as its coolant. A boiling water reactor (BWR) uses water, still under pressure, but not enough to stop it from boiling.

A high temperature gas-cooled reactor (HTGR) uses helium gas as its coolant.

Control rods

Control rods are made of a neutron-absorbing material such as cadmium or boron and can be raised from the reactor core to increase the rate of the reactor, or lowered into the reactor to decrease the rate or shut down the reactor.

A reactor is critical when one neutron from each fission produces another fission. Reactors are designed to be supercritical but are maintained at the critical level by use of the control rods.

eBookplus

Weblink:

Nuclear control rods

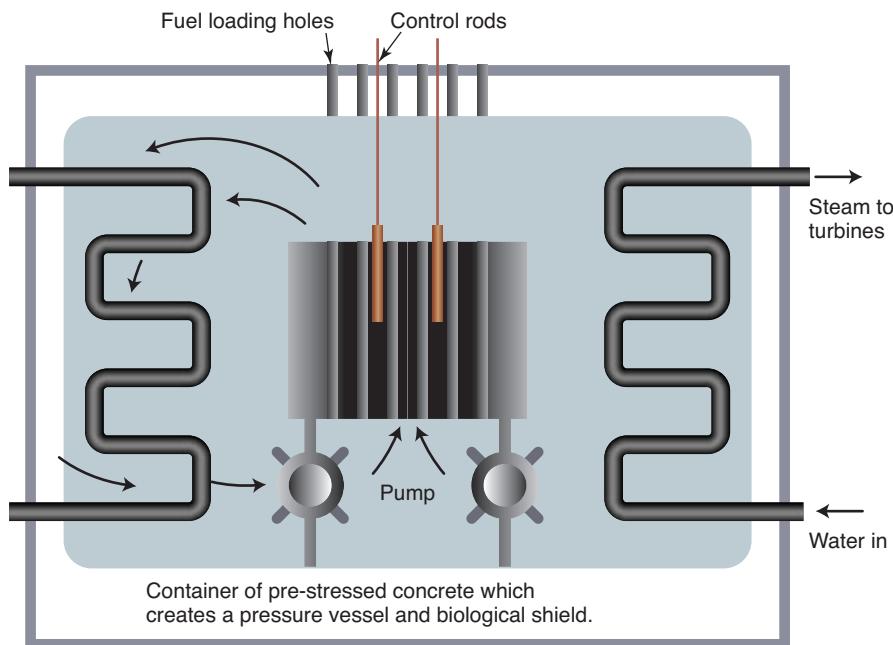


Figure 25.9 An advanced gas-cooled reactor (AGR). This reactor has a graphite moderator, uses enriched uranium oxide as fuel and is cooled by carbon dioxide gas that is pumped through the core. The core would contain about 120 tonnes of fuel and 1500 tonnes of graphite.

Producing electricity

The coolant that passes through the reactor core passes through a heat exchange unit where it heats coolant from another circuit. (As a safety precaution, the coolant from the core would usually not pass outside the main reactor building.)

The coolant from the second circuit would carry the thermal energy to a boiler where it would heat water to produce steam to drive a turbine to produce electricity (see figure 25.10 below).

Safety of nuclear reactors will always be of major concern, but theoretically a nuclear reactor should be very safe. In an emergency, control rods should shut down the reactor in a very short period of time. Some people would argue that despite the problem of disposing of long half-life radioactive wastes from a fission reactor, nuclear reactors are a far more environmentally safe means of producing electricity than using power stations that are fed by carbon dioxide-producing fossil fuels.

However, there is a stigma attached to nuclear processes.

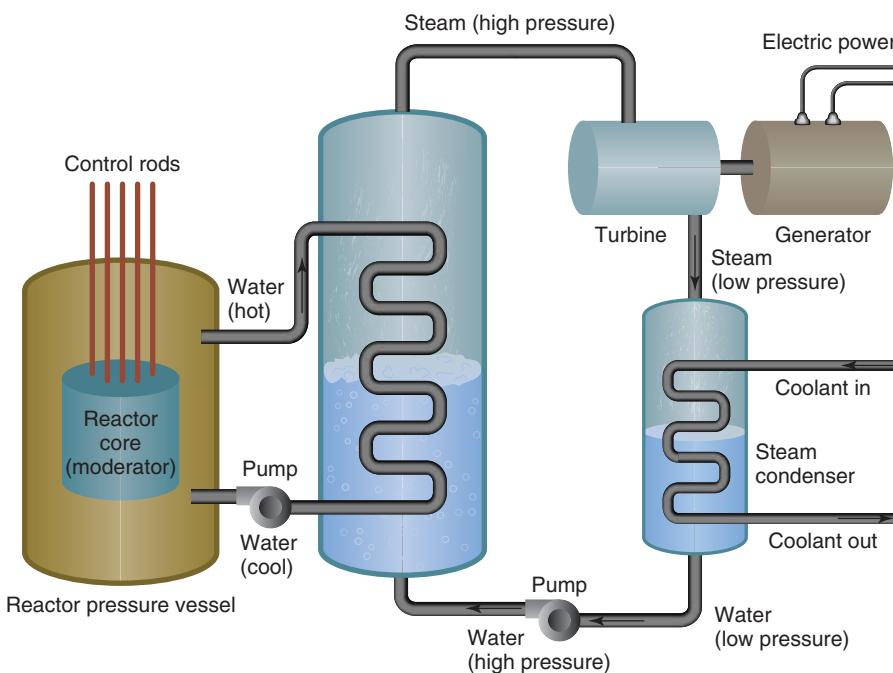


Figure 25.10 A nuclear power station. Hot water under pressure from the reactor boils water in the steam generator and this steam drives a turbine, which in turn drives an electricity generator.

Radioactive waste products

Even if nuclear reactors were completely safe, there would still be a problem associated with the radioactive waste products from the reactor.

If the fuel rods remain in the reactor core for about three years, only a small proportion of the uranium nuclei present will have undergone fission when the fuel rods are removed. The fuel rods will contain uranium-235, uranium-238, and some plutonium-239 (formed when uranium-238 absorbs neutrons). There will be other radioactive waste products that must be stored safely for long periods.

Discharged fuel rods are usually stored under water at the reactor site for a few months and then transported to a reprocessing plant. They are probably stored again and then the fuel is dissolved in nitric acid and separated into unused uranium, plutonium and other wastes. While large volumes of low-activity waste are produced and stored until the activity reaches a safe level, relatively small volumes of high-activity waste must be stored safely for long periods.

Several methods have been used. A vitrification process is used where the radioactive wastes are incorporated into borosilicate glass for immobilisation. This process has been developed in France over a long period of time and is in use at reprocessing plants in the UK and France.

An Australian invention, Synroc, which was first developed at the Australian National University in Canberra, may offer significant advantages over the vitrification process in long-term performance and, possibly, overall cost savings.

eBook plus

Weblink:
Synroc

PHYSICS IN FOCUS

Chernobyl

There have been several nuclear accidents involving nuclear reactors. The worst occurred at Chernobyl on 26 April 1986 when a reactor and building caught fire and a large amount of radioactive material escaped to the atmosphere.

The reactor was a type RBMK 1000 which was graphite moderated and was cooled by boiling water. The design was such that the neutrons were fully moderated by graphite. The reactivity was reduced because the water in the coolant tubes absorbed neutrons. (The neutrons combined with protons to produce deuterium.) The reactor was designed to operate with a mixture of steam and water in coolant tubes. A ‘steam void’ captures fewer neutrons than a similar volume of water.

When the accident occurred the reactor was being run in far from normal conditions. The output had been reduced to test whether the inertia of a steam turbine, when isolated from both the driving steam supply from the reactor and from the electricity grid, would be sufficient to run reactor pumps for a short period. This was designed to improve reactor safety procedures. (A few weeks earlier, a similar test of reactor safety had

been completed at a research reactor in the USA. In that test, the reactor had shut itself down.)

The aim of the test was to reduce the reactor power to 1000 MW and shut off the steam supply to one of the turbogenerators. Measurements would then be taken to see how long the inertia of the turbogenerator would provide enough electricity to drive four of the eight water-coolant-circulating pumps. The other four pumps would be controlled from the electricity grid in the normal manner.

A number of the safety features were overridden or bypassed to enable the test to go ahead. These included the emergency core-cooling system being rendered inoperable and the local control of the automatic reactor power control rods being disconnected. The power was reduced but it was impossible to stabilise at 1000 MW and the power fell to about 30 MW. The tubes were full of water (no steam) and this reduced reactivity as the maximum number of hydrogen nuclei were available to absorb neutrons.

In an attempt to increase reactor power, almost all auto and manual control rods were raised as far as possible and this increased

reactor power to 200 MW. Raising all control rods was against standard operating instructions and would have activated a safety trip reinserting control rods but this safety trip mechanism had been rendered inoperable for the test.

At 200 MW the steam was shut off to one of the turbogenerators. The automatic safety system designed to shut down the reactor in event of steam supply failure had been shorted out for the test. As the turbogenerator ran down, the speed of the pumps feeding the water coolant circuit fell and steam production in the pressure tubes increased. This increased the reactivity of the system. (With more steam and less water in the coolant tubes, the absorption of neutrons decreased.)

The reactor power increased from 200 MW to over 500 MW in three seconds and continued to rise exponentially!

The operator attempted to close down by inserting all control rods but it took 10 seconds to insert control rods and shut down. As the control rods were inserted, the reaction continued and was concentrated at the bottom of the core.

In this time, the pressure tubes had become void of water and the superheated steam interacted with the zirconium fuel cladding, releasing fuel and fission products. The superheated steam also interacted with the zirconium pressure tubes producing hydrogen and rupturing the tubes.

There seem to have been two explosions. The first was when the steam and fuel interacted, the explosion breaching the reactor building. The second was caused by the interaction of the hydrogen (produced in the steam zirconium interaction) combining with carbon monoxide and the air that had entered the building after the first explosion. The incoming air ignited the exposed graphite moderator, which was still near its normal reactor temperature of 770°C.

The flames were eventually extinguished by sand dropped from helicopter flights over the reactor building.

There was no nuclear explosion. The explosions were entirely chemical in nature. There had previously been safety concerns with this type of reactor and changes had been made to address those concerns. Those changes were ignored in the test.

PHYSICS FACT

The number of naturally occurring elements

A common claim in school science textbooks is that there are 92 naturally occurring elements. In fact the elements technetium (element 43) and promethium (element 61) are not naturally occurring elements on Earth.

Many texts claim that uranium is the heaviest naturally occurring element but that too is incorrect.

In 1972, an estimated two tonnes of plutonium was located in the bed of the Okla River in the Republic of Gabon when uranium deposits were being mined. The existence of the plutonium has been explained by considering that a ‘natural

fission reactor’ produced the plutonium in prehistoric times. The percentage of uranium-235 in natural uranium would have been higher (uranium-235 has a shorter half-life than uranium-238) and water flowing over the uranium would have acted as a moderator. Other long half-life isotopes of elements that could have been produced in the fission have also been identified, further supporting the natural reactor idea.

This leaves us with the current thinking that there are 91 elements occurring naturally on Earth.

25.5 MEDICAL AND INDUSTRIAL APPLICATIONS OF RADIOISOTOPES

Radioisotopes are used in a wide variety of ways in areas such as medical imaging and treatment, preservation of food, measuring and testing of materials and inspection of metal and welds. Some of the properties of the radioisotopes discussed below are summarised in table 25.1 (page 491).

Nuclear medicine

It has been estimated that in 1995, over 250 000 Australians underwent procedures that involved the use of radiopharmaceuticals. Exploratory surgery has become less common as new diagnostic techniques, many of which are based on nuclear medicine, become readily available.

X-rays have long been used to examine the structure of the body, but techniques associated with nuclear medicine are able to provide information on the functions of the body.

Radioisotopes carried in the blood can help doctors detect clogged arteries or check the function of the circulatory system. Some chemicals collect in specific organs or tissues. Radioactive tracers that concentrate in an organ or tissue enable an image of that organ or tissue to be formed.

More information on the use of radioisotopes in medicine can be found in the 'Medical Physics' module, pages 382–395.

The radioisotopes used have short half-lives, which is an advantage for the patient but means that they cannot be stored for very long in the hospital. Technetium-99m is a very commonly used radioisotope. It has a half-life of six hours and is produced through the decay of the radioisotope molybdenum-99 which is formed in nuclear reactors. Gallium-67 and thallium-201 are other commonly used radioisotopes. Gallium-67 is used in the detection and localisation of tumours and thallium-201 is used in the diagnosis of coronary artery disease and other heart conditions.

Radioisotopes are also used in therapeutic applications. When living tissue is exposed to high levels of radiation, the cells may be destroyed or damaged in a way that stops them from reproducing. Radioisotopes such as cobalt-60 are used to destroy malignant tumours. Many types of cancer are treated by radiation therapy.

Positron Emission Tomography (PET)

Positron Emission Tomography is a non-invasive means of producing diagnostic images. The patient is usually injected with a metabolically active tracer, a molecule that will be used by the body, which contains a positron-emitting isotope such as carbon-11, nitrogen-13, oxygen-15 or fluorine-18. Glucose labelled with carbon-11, which has a half-life of 20 minutes, can be used to study the brain.

The positron-emitting isotopes are prepared by bombarding the appropriate elements with protons in a cyclotron. (A cyclotron is a type of particle accelerator.) Carbon-11 can be formed when nitrogen-14 is bombarded with protons. This results in the emission of an alpha-particle:



In PET, after a positron is emitted, it will combine with and annihilate an electron, usually after travelling less than a millimetre. This produces two gamma rays that travel in opposite directions (see figure 25.11).

When two gamma rays are detected simultaneously by detectors on opposite sides of the patient, they must have been emitted from the line joining the detectors.

After about half a million such events have been recorded, a computer is used to perform a tomographic reconstruction that can be either two-dimensional or three-dimensional if multiple sections have been taken. See chapter 20, pages 392–394 for more information and illustrations about PET.

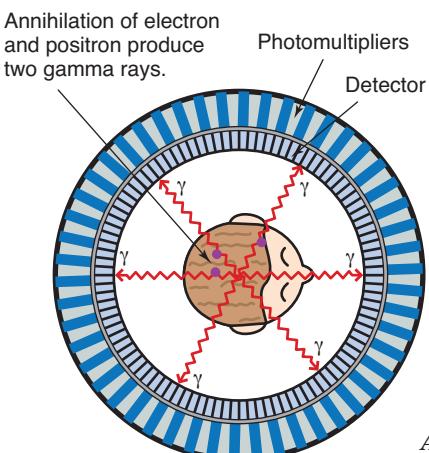


Figure 25.11 In PET, when an annihilation of a positron and an electron occurs, two photons that travel in opposite directions are produced. As there is a complete ring of detectors around the patient, the photons trigger detectors on opposite sides simultaneously. After about half a million events, an image is constructed by computer.

Industrial and agricultural applications

Industrial radiography is used to inspect metal parts and welds for defects. Radiation from iridium-192 or cobalt-60 is beamed at an object and the radiation that passes through is recorded on special photographic film. More radiation will pass through cracks or flaws and hence these can be detected. An example of the use of this process is the radiographing of jet engine turbine blades to ensure that the internal cooling passages have been manufactured correctly.

Radioisotopes in gauges are used to monitor and control the thickness of sheet metals, textiles, metal foils, paper and photographic film. The amount of radiation passing through the material depends on its thickness and density. Radiation is passed through the material as it is processed and the amount of radiation detected indicates whether or not the material is of the correct thickness.

Other uses of radioactive gauges include:

- monitoring roads, buildings and bridges
- use in the exploration for oil, gas and minerals
- detecting explosives in luggage at airports.

A very common application of a radioisotope is the use of americium-241 in smoke detectors. (Americium is another transuranic element. It has an atomic number of 95 and is produced by the decay of plutonium-241.)

Agricultural uses of radioisotopes include the use of tracers in plant nutrients. If phosphorus-32 or nitrogen-15 is used in fertilisers, data about the rate at which the plant takes up the fertiliser can be obtained. This, in turn, can yield information that will allow more efficient use of the fertiliser.

The Sterile Insect Technique (SIT) involves irradiating male insects reared in a laboratory to sterilise them. They are then released in large numbers in infected areas. They mate with females but no offspring are produced.

Table 25.1 Properties of some radioisotopes

NAME	EMISSION	HALF-LIFE
Phosphorus-32	beta (1710 keV) with range of 6 m in air	14.3 days
Cobalt-60	beta (318 keV) and gamma (1333 keV)	5.3 years
Molybdenum-99	beta	67 hours
Technetium-99m	gamma (140 keV)	6.03 hours
Iridium-192	beta (672 keV) and gamma (468 keV)	73.83 days
Thallium-201	gamma (135 and 167 keV) and photons (68 to 80 keV)	73 hours
Americium-241	alpha	432.7 years

Using radiation to preserve food is still a controversial use of radioisotopes. A search of the internet will yield a number of sites vehemently opposed to food irradiation. About forty countries have now approved the irradiation of food.

Irradiation is most useful in the areas of:

- preservation
- sterilisation
- controlling airborne diseases
- controlling sprouting, ripening and insect damage.

Not all foods can be irradiated; some fruits become soft and dairy products develop an unpleasant taste.

25.6

NEUTRON SCATTERING

eBookplus

Weblink:

Neutron scattering
at ANSTO

Neutron scattering has become an important tool in many fields of study. In Australia, neutron scattering investigations are carried out at the Australian Nuclear Science and Technology Organisation, ANSTO, at Lucas Heights in Sydney.

The main tools used to detect scattered neutrons are diffractometers and spectrometers. Diffractometers are used to determine atomic and molecular structure when there has been elastic scattering of neutrons. Spectrometers are used when neutrons have been inelastically scattered and information about quantities associated with atomic motion or energy is required.

Neutron scattering has been used for research in fields such as geology, environmental science, biology and biotechnology, engineering, materials science, physics and chemistry.

X-rays are more intense and more common than neutron sources but there are areas where neutrons have an advantage over X-rays. Some of these advantages are listed below.

- The neutron has a wave nature. The de Broglie wavelength of a thermal neutron is comparable to the spacing between atoms in molecules. Neutrons scattered from an atomic lattice will therefore produce interference patterns.
- The neutron has a magnetic moment, which makes it an ideal tool for studying magnetic structures and materials.
- Neutrons have an energy similar to the vibrational energy of atoms in solids and liquids. This enables neutrons to be used to study the motion of atoms in molecules in detail.
- Neutrons can be used to study materials without causing destruction.
- Neutrons interact strongly with nuclei. The strength of the interaction varies for different nuclei, which makes it possible to study isotopes of light elements.

The disadvantage of neutron scattering is that a nuclear reactor is required to produce the neutrons.

The HIFAR reactor at ANSTO was replaced with a new reactor, OPAL, in 2007. There were seven instruments used in neutron scattering investigations when HIFAR was operating, and this was increased to nine when OPAL started. However, all has not been plain sailing with OPAL. OPAL first went critical in August 2006 and was operating at full power in November 2006. HIFAR was permanently shut down in January 2007. OPAL was officially opened by the then Prime Minister John Howard in April 2007 but experienced problems in July 2007. Loose fuel plates enabled water to seep into the heavy water in the reactor, so it was shut down. It remained shut down for the rest of 2007 and was to be restarted in 2008.

SUMMARY

- As early as 1903 it was realised by Rutherford that a vast amount of energy was associated with processes involving radioactivity.
- During the 1930s some physicists realised that there was the possibility of a chain reaction which would produce an uncontrolled release of energy from the nucleus.
- Fermi and his group conducted research in which neutrons bombarded heavy elements. When slow or ‘thermal’ neutrons were used, new isotopes, some of which were thought to be of transuranic elements, were produced.
- The element barium was identified as being present after uranium was bombarded with slow neutrons. This was interpreted by Meitner and Frisch as being evidence of the fission of a nucleus of uranium.
- Refugee physicists from Germany were very concerned about Germany developing nuclear weapons and lobbied the US government to undertake research into the possible development of nuclear weapons.
- The Manhattan Project was commenced. It was to produce the fuel for both a uranium bomb and a plutonium bomb, and to complete the design of an atomic bomb.
- During World War II, the greatest physicists in the free world worked on the Manhattan Project. Work on the atomic bomb continued after the war with Germany ended and the first atomic bomb was designed and constructed at Los Alamos. It was tested successfully in July 1945.
- Against the wishes of many of those physicists, atomic bombs were dropped on the Japanese cities of Hiroshima and Nagasaki.
- Nuclear reactors, which used thermal neutrons rather than the fast neutrons used in weapons, were developed and used for generation of electricity.
- In nuclear reactors, control rods are used to control the rate at which fission reactions occur. A controlled chain reaction in which, on average, one neutron from each fission produces another fission, releases energy which ultimately is used to generate electricity.
- The debate about the safety and environmentally sensitive aspects of nuclear power production has continued unabated. Nuclear

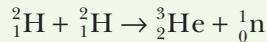
accidents such as the one that occurred at Chernobyl have provided ammunition for those opposed to nuclear reactors.

- There are now many uses for radioisotopes, some of which are produced in nuclear reactors and some which are produced in particle accelerators (such as a cyclotron).
- Medical imaging techniques that require radioisotopes are providing new diagnostic techniques. These techniques are replacing much of the exploratory surgery that used to be carried out.
- There are also industrial and agricultural techniques that use radiation from radioisotopes.
- Neutron scattering provides another very useful tool for examining the properties of materials.

QUESTIONS

- Which has a greater mass, a uranium nucleus before fission or the products after fission? Explain.
- What is the isotope formed when uranium-238 captures a neutron?
 - The nucleus of this isotope is unstable because it has an excess of neutrons. What happens to it to allow it to restore stability?
 - What is the isotope eventually formed?
- List the things that can happen to a neutron produced in a fission of uranium-235.
- State some reasons why a chain reaction does not occur in a natural deposit of uranium.
- Why was there such a large mass of graphite (about 400 tons) in the first atomic reactor?
- The strong nuclear force between adjacent protons in a nucleus is very much greater than the electrostatic repulsion between the protons. However, Otto Frisch believed that the electrostatic force was the force that accelerated the product nuclei formed in a fission. Explain.
- Explain how the chain reaction within a nuclear reactor is maintained at a steady level.
- Near the end of a three year period in a nuclear reactor, most of the energy released from a uranium fuel rod comes from the fission of plutonium. Explain.
- The RBMK nuclear reactor at Chernobyl was cooled by boiling water and was designed to operate with a mixture of liquid water and steam in the coolant tubes.

- (a) Explain how the rate of fissions taking place would be affected by filling the coolant tubes with liquid water (no steam). (This happened before the Chernobyl accident.)
- (b) Explain how suddenly reducing the flow of water through the coolant tubes would contribute to the nuclear disaster. (Why did the rate of fissions increase so dramatically when the water in the coolant tubes started to boil?)
10. The discovery of nuclear fission came after it was confirmed that barium, and not radium, was really present in samples of uranium that had been bombarded with neutrons. Suggest why it was so difficult to distinguish between very small amounts of radium and barium. (A periodic table may be helpful.)
11. Rutherford estimated that there was sufficient energy in one gram of radium to raise 500 tonnes a mile high. Presumably, one gram of uranium would contain similar energy, perhaps enough to raise 500 tonnes, 1.6 km high. The energy output from the first uranium bomb was about 20 kilotonnes. The mass of uranium in that bomb was about 40 kg. (1 kilotonne is equal to 4.2×10^{12} J.) How accurate was Rutherford's 1903 estimate?
12. Calculate the energy released in the reaction of two deuterium nuclei as demonstrated in Rutherford and Oliphant's experiment.



Masses:

$${}^2_1\text{H} = 2.014\ 102\ \text{u}$$

$${}^3_2\text{He} = 3.016\ 029\ \text{u}$$

$${}^1_0\text{n} = 1.008\ 665\ \text{u}$$

13. Fluorine-18 is a positron-emitting isotope used in PET that can be prepared in a cyclotron by bombarding oxygen-18 with a proton.
- (a) Complete the nuclear reaction:
- $${}^{18}_8\text{O} + {}^1_1\text{H} \rightarrow {}^{18}_9\text{F} + ?$$
- (b) What other very highly penetrating particle would be produced in this reaction?
14. (a) Calculate the energy (in keV) of each of the gamma rays produced in PET when a positron and an electron annihilate each other. (Any kinetic energy associated with the positron and electron is so small that it can be neglected.)
- (b) Explain why the two gamma rays that are detected must travel in opposite directions. (The mass of an electron and the mass of a positron are identical and each is equal to 0.000 548 580 u.)
15. In 1994, Bertram N. Brockhouse and Clifford G. Shull were awarded the Nobel Prize for Physics for their work on developing neutron scattering as an investigative tool. List the properties of neutrons that make them such an important tool for investigating properties of matter in such a wide variety of fields.

CHAPTER 26

QUARKS AND THE STANDARD MODEL OF PARTICLE PHYSICS



Figure 26.1 An aerial view of the Fermi National Accelerator Laboratory (Fermilab) at Batavia, Illinois in the USA. The main accelerator, 6.28 km in circumference, is clearly visible as the circle in the top half of the photograph. The circle in the lower half is the main injector ring. Some of the other accelerator and storage rings are just visible near the main building on the very left. Many important discoveries have been made at Fermilab, with possibly the greatest being the discovery of the top quark in 1995.

Remember

Before beginning this chapter, you should be able to:

- recall how charged particles interact and move within electric and magnetic fields
- recall the methods used by the pioneers of atomic and nuclear research to detect ionising radiation
- define the properties of the strong nuclear force that binds nucleons together in a nucleus.

Key content

At the end of this chapter you should be able to:

- describe how a cloud chamber can be used to detect charged particles
- identify why particle accelerators are needed to probe the structure of matter
- identify the contribution that particle accelerators make to our understanding of the structure of matter
- recall the key features and components of the Standard Model.

26.1 INSTRUMENTS USED BY PARTICLE PHYSICISTS

Many of the most important discoveries in physics made between fifty and one hundred years ago were made, sometimes by accident, by a single physicist working with a very simple apparatus. We have seen that Fermi was awarded a Nobel prize for the work completed after he put a rough piece of paraffin in the path of the neutrons that he was using to irradiate a sample. We have also noted (chapter 22) the basic apparatus used by Marsden in his desktop alpha-particle scattering experiment. Rutherford and his co-workers used simple scintillation detectors to observe the scattering of alpha particles. This method of detection enabled them to count the alpha particles. However, despite the significant results achieved with such simple apparatus, better detectors that would provide information such as the charge and energy of the particles were required.

These earlier physicists used alpha-particle sources that were naturally occurring alpha-particle emitters. Some of these produced alpha particles of much higher energy than others. When alpha particles were used to induce artificial radioactivity, it soon became apparent that particles with higher energies still would be more useful.

The quest for better detectors saw the use of the cloud chamber and the development of the bubble chamber. In recent times, these were superseded by larger and more complex multicomponent detectors. Examples of these are the detectors used at the high-energy accelerator facilities, such as CERN, Fermilab and Brookhaven.

The quest for higher energy particles saw the development of a variety of particle accelerators. The higher energy particles from the particle accelerators were used to bombard nuclei and produce a wide variety of new particles.

As a result, the days of simple experiments are long since gone and now discoveries and advances in the field of particle physics require very expensive equipment and perhaps many hundreds of physicists working together on a single project. The ‘Physics in focus’ section on the discovery of the top quark (pages 510–511) provides an example of how modern research is carried out.

We will look at the design and use of some of the particle detectors and particle accelerators that have been used throughout the twentieth century.

Particle detectors

Both the cloud chamber and bubble chamber were very useful particle detectors. In the following section, we will see how they were used to detect particles, and examine some of the discoveries made using them.

Cloud chambers

The cloud chamber was invented by C. T. R. Wilson before the end of the nineteenth century, but not used to detect particles until about 1910. It remained in use until about 1960.

A cloud chamber contains a supersaturated vapour (see the note at left). As ionising radiation passes through the vapour, fine droplets of vapour form on the ions produced by the radiation. This leaves a visible vapour trail showing the path of the particle. If the chamber is in a magnetic field, the path of a charged particle will be curved, with the direction of the curve indicating the charge of the particle.

There is usually a maximum concentration of a vapour that can be present in air. (Humidity is the amount of water vapour in the air. It is expressed as a percentage of the maximum amount of water vapour that can be held by the air.) In a supersaturated vapour, the amount is greater than 100%. This may seem strange; however, some vapours can condense only if there are particles, such as dust or ions, on which condensation can commence. The passage of an ionising particle through a cloud chamber produces the ions on which the vapour can condense.



26.1

Cloud chambers

There are two main types of cloud chamber: the expansion cloud chamber and the diffusion cloud chamber. Small versions are available for use in school laboratories. There are various liquids that could produce a supersaturated vapour but propan-2-ol has been found to work well in the cloud chambers used in school laboratories. A small amount of the alcohol is usually soaked into a felt ring or disc and evaporation provides the supersaturated vapour. (The features and peculiarities of the different types of cloud chamber are covered in Practical activity 26.1, page 518.)

PHYSICS FACT

Discoveries made with a cloud chamber

In 1919, Rutherford demonstrated the reaction of an alpha particle with a nitrogen nucleus. In 1932, P. M. S. Blackett (1897–1974) demonstrated the same reaction in a cloud chamber (see figure 24.8, page 459).

In 1933, Carl D. Anderson (1905–1991) observed a track in a cloud chamber that was made by a particle similar to an electron but with a positive charge, hence discovering the anti-electron or positron.

Bubble chambers

In 1952, Donald Glaser invented the bubble chamber. It is claimed that the observation of bubbles in glasses of beer played a significant part in the invention.

The bubble chamber has a similar principle of operation to the cloud chamber except that the bubble chamber contains a superheated liquid. (A superheated liquid exists in the liquid state at a temperature above its normal boiling point.) Propane and pentane were used in early bubble chambers and hydrogen in later ones. When ionising radiation passes through the liquid, localised boiling occurs on the ions and leaves a trail of bubbles. Bubble chambers were much better detectors than cloud chambers because of the greater density of the substance in the chamber. A 10 cm bubble chamber was approximately equivalent to a 10 m cloud chamber. Figure 26.2 shows a bubble chamber at CERN that was dismantled in 1984 after being used for over six million photographs.

Modern detectors

Detectors in use at large nuclear research facilities, such as CERN, Fermilab and Brookhaven, are now larger and more complex than bubble chambers. In typical high-energy experiments performed at these facilities, multicomponent detectors are used to record what may be millions of events and to store them on computer for later analysis. The Collider Detector at Fermilab is shown in figure 26.10 (see page 511).

The function of a detector is to record the trajectory, energy and momentum of the particles produced in a collision ‘event’. If two beams of particles, perhaps protons and antiprotons, with similar



Figure 26.2 The 3.7 m bubble chamber at CERN. Before being dismantled in 1984, this bubble chamber was used for over six million photographs.

eBookplus

Weblinks:

European Organisation for Nuclear Research (CERN)
 Fermi National Accelerator Laboratory
 Brookhaven National Laboratory
 Stanford Linear Accelerator Center

eBookplus

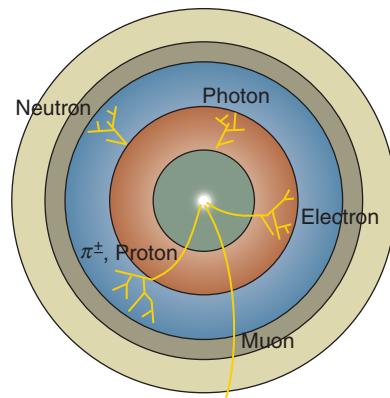
Weblinks:

Hands on CERN
 SLAC virtual visitor centre

energies collide head-on, the particles produced could travel in any direction and a large cylindrical detector would be used. Such a detector would commonly have four different regions. There would be an inner tracking chamber surrounded in turn by an electromagnetic calorimeter, then a hadronic calorimeter and finally a muon chamber (see figure 26.3). The inner tracking chamber contains a gas, and as the charged particles produced in the collision event traverse this chamber, they produce ions. The ions may be collected on thin metal wires and produce a small electrical pulse. Once the presence of the ions has been detected, the tracks of the particles that produced the ions can be deduced. Many very short-lived particles do not leave tracks but they may decay into particles that do.

The calorimeters are made of dense materials that absorb the energy of the particles interleaved with sensitive detector materials. The different materials are segmented and it is possible to determine where a particle was finally absorbed. The electromagnetic calorimeter is optimised to measure the energy and positions of electrons and photons that interact via the electromagnetic force. The hadronic calorimeter is optimised to measure the energy and positions of hadrons which interact via the strong force (see page 504 for a description of hadrons).

Only muons (and neutrinos) are able to pass through the two inner calorimeters. Any charged particle that reaches the outermost calorimeter must be a muon. Neutrinos, of course, continue without interacting with any part of the detector. The passage of different types of particle through a detector is shown in figure 26.4.



- Beam pipe (centre)
- Tracking chamber
- E-M calorimeter
- Hadron calorimeter
- Magnetised iron
- Muon chambers

Figure 26.3 A simplified end-on view of a cylindrical detector that might be used for a colliding beam experiment

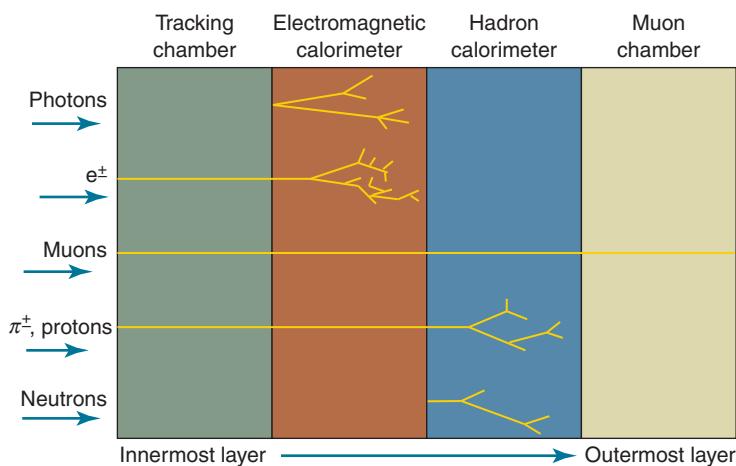


Figure 26.4 The passage of particles through the different sections of a multicomponent detector

Particle accelerators

A very simple accelerator is the electron gun in the tube of a television set. Electrons are accelerated across a large potential difference and then directed at the screen of the television set. Some of the early particle accelerators were similar to this in that they were single-stage electrostatic accelerators. Higher energies were possible when the particles were accelerated many times, such as in a linear accelerator. The development of cyclic accelerators such as the cyclotron saw large energies possible with smaller devices but, as we will see, modern accelerators have become very large and complex devices.

The first particle accelerators

An early particle accelerator that was used to accelerate protons to 770 keV was developed by John D. Cockcroft and Ernest T. S. Walton at the Cavendish Laboratory in 1932. It was an electrostatic machine that gave the protons a single high energy ‘kick’. Another electrostatic accelerator is the Van de Graaff generator that you have probably encountered. Large versions were capable of reaching 1.5 MeV or higher.

These accelerators have since been improved as it became apparent that many more important discoveries could be made with particles of a higher energy, and new accelerators were developed. However, despite these new accelerators, the earlier accelerators are still found to be useful. It is interesting to note that at Fermilab, the initial step or pre-acceleration is provided by a Cockcroft–Walton accelerator and at Brookhaven, the initial acceleration of the Relativistic Heavy Ion Collider (RHIC) is provided by tandem Van de Graaff accelerators.

Linear accelerators

The most famous linear accelerator is at the Stanford Linear Accelerator Center (SLAC). Charged particles are fired through a three-kilometre-long evacuated tube. The charged particles pass through one cylindrical electrode and are then accelerated by an electric field as they pass through a gap before encountering another electrode. This process is repeated and the particles increase their energy. Of course, the alternating accelerating potential has to keep in step with the particles and this requires the cylindrical electrodes to become longer and longer (see figure 26.5 on the next page). Eventually it becomes impractical to add extra stages to a linear accelerator. At SLAC, electrons were accelerated to a velocity very close to that of light.

Cyclotrons

Like a linear accelerator, a cyclotron is able to give a charged particle many ‘kicks’ as it passes through the electric fields between the ‘dees’ of the cyclotron (see figure 26.6 on the following page). Again the particles move through an evacuated region. The whole apparatus lies between the poles of a large magnet. Therefore, the particles move in circular paths, with the radii of the paths increasing each time the particle gains energy as it passes through the gap between the dees. When the particles reach the limit of the magnetic field they are deflected into a target. Very high energy cyclotrons are not possible for a number of reasons. Eventually size would become prohibitive and also, as the particles reached very high velocities, the relativistic increase in mass would mean that the particles would become out of step with the applied alternating potential.

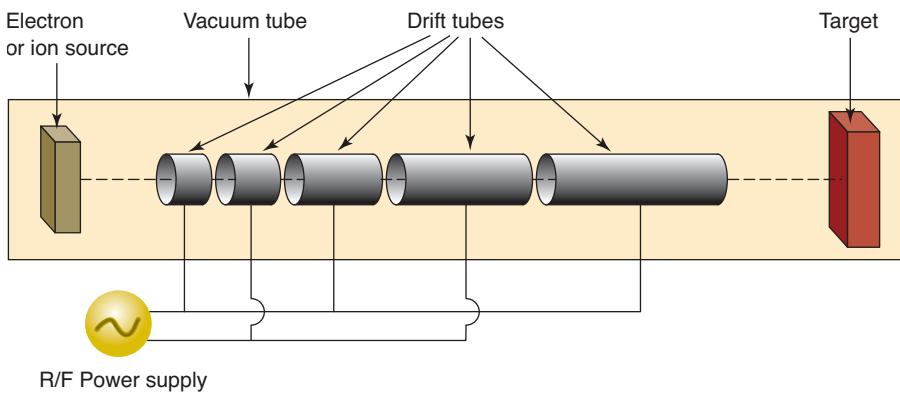


Figure 26.5 A linear accelerator. Charged particles are accelerated when they pass through the gaps between the cylindrical electrodes, or drift tubes. It is necessary to keep the charged particles in step with the radiofrequency high-voltage applied to the electrodes. Hence, the drift tubes have to become progressively longer. The SLAC, which was built in 1967, is three kilometres long and accelerates electrons to 20 GeV. After modifications were completed in 1987, the SLAC was able to produce 50 GeV electrons.

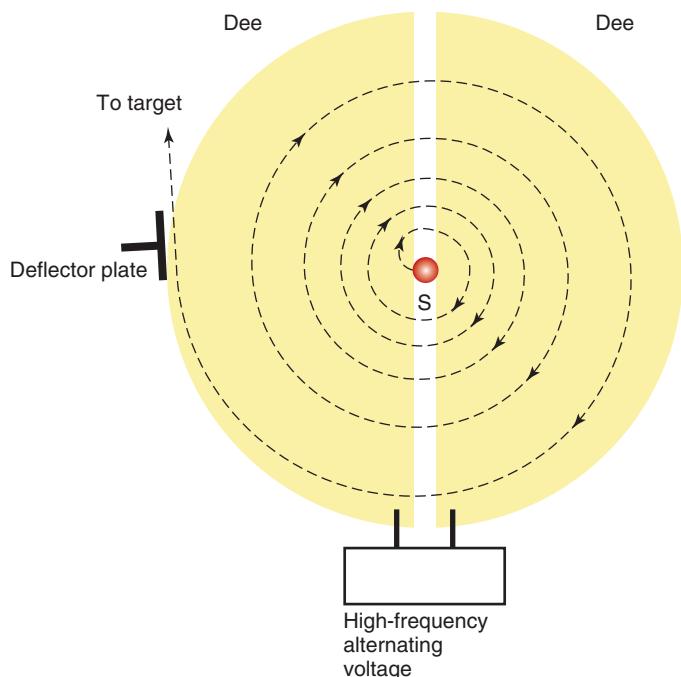


Figure 26.6 A top view of a cyclotron. Positively charged particles from the ion source, S, travel in semicircular arcs as they pass through the 'dees'. They are accelerated by a high voltage as they pass from one dee to the other and as their velocity increases, so does the radius of their path. Finally, near the outside of the dees, they are deflected towards the target. The dees are hollow cylinders of non-magnetic metal and the apparatus is in an evacuated container between the poles of a powerful magnet. (Note that for positively charged particles to travel clockwise as shown, the magnetic field must be directed out of the page.)

Synchrotrons

The main accelerators today are synchrotrons. Synchrotrons keep the particles in a path of constant radius. As the particles gain energy, the magnetic field is increased to maintain the same path. Many powerful magnets are required around this path. The particles move through a small-diameter, evacuated tube that forms a large-diameter ring.

Weblinks:

- [Australian synchrotron](#)
- [Nova: Science in the news](#)
- [Berkeley Lab's Advanced Light Source](#)

On each circuit around the ring, the particles pass through regions where an applied radio frequency provides the ‘kick’ to increase their energy. This radio frequency provides an electric field in a direction such that it produces the ‘kick’ that increases the energy of the particles. The radio frequency increases as the particles increase in energy and take shorter and shorter periods to complete their orbits. A disadvantage of a synchrotron is that a ‘batch’ of particles must complete their journey through the accelerator before another batch can enter. However, the advantages of the synchrotron, in terms of energy that can be achieved, far outweighs this disadvantage.

The dimensions and statistics of the large accelerators are impressive. The main accelerator, the Tevatron at Fermilab, has a circumference of 6.28 km. The original accelerator was able to accelerate protons to about 200 GeV. The magnets used in this accelerator are the light blue and distant red sections in the middle of figure 26.7. When higher energies were required, another accelerator was built below the first. It uses superconducting magnets to steer the particles around the ring, and because these do not heat up, they can be left on for very long periods. Another accelerator in a separate ring now accelerates protons to 150 GeV, at which point they are transferred to the new accelerator and accelerated further to about 1000 GeV (1 TeV).

The Large Electron Positron collider (LEP) at CERN occupied a tunnel of 27 km circumference straddling the Swiss–French border. In 2000 it was shut down, and construction of its replacement, the Large Hadron Collider (LHC) began. The LHC was scheduled to commence operation in 2005 but that was delayed until 2007. Then in 2007 a problem with one of the magnet support structures caused a further delay until 2008. The LHC has 1232 superconducting magnets, each 15 m long, around 85% of its circumference. The magnets were supplied by Fermilab. These magnets will be powered by superconducting cables carrying currents of 12 000 amps and will be cooled by liquid helium to -271°C . The LHC should be able to accelerate protons to 7 TeV and collide these protons with other protons travelling in the opposite direction, also with an energy of 7 TeV. It should be able to collide heavy ions, such as lead, with a total energy of 1250 TeV, about thirty times that of the RHIC at Brookhaven. (For more on the LHP and RHIC, see ‘Recent developments in accelerator research’, page 512.)



Figure 26.7 Part of the 6.28 km-circumference Fermilab Tevatron accelerator. The yellow and red sections on the floor are parts of the new accelerator, which can accelerate protons to nearly 1000 GeV.

Fixed target collisions versus colliders

In collisions between high-energy particles and a target, conservation of momentum and conservation of energy naturally apply. The momentum of the products of any interaction must be the same as the momentum of the incident particle. This means that the products of the interaction have considerable kinetic energy and, hence, only a small amount of the kinetic energy of the incident particles is available for the production of new particles. When a 400 GeV proton from an accelerator is fired at a target and then interacts with a particle in the target, the amount of energy available for producing new particles is only 27 GeV. If the accelerator produces 1000 GeV protons, the energy available for producing new particles will be only 42 GeV.

In the late 1970s, this problem was overcome by producing interactions where the total momentum was very small. Instead of firing a high energy particle into a stationary target, accelerators were modified to accelerate protons and antiprotons to the same high energies. Carlo Rubbia (1934–), at CERN, was the driving force behind the necessary modifications being made to produce a collider. Although it was not an easy task at the time, antiprotons were produced in large numbers, injected into the accelerator and accelerated simultaneously with the protons. The antiprotons travelled around the accelerator in the opposite direction to the protons. When they reached their maximum energy the beams of protons and antiprotons were deflected to intersect.

The total momentum of the proton and antiproton before interaction is close to zero and, hence, the total momentum of the products of any interaction has the same near zero value. Therefore, if the protons and antiprotons were each accelerated to about 400 GeV, the total energy available for the production of new particles should be 800 GeV. (In fact, the 400 GeV accelerator at CERN was able to accelerate the protons and antiprotons to about 260 GeV and this gave a total energy of 520 GeV.)

Today, the highest energy accelerators are colliders, but there are still many experiments performed with fixed target accelerators.

Fixed target accelerators can produce very large numbers of interactions as the high-energy particles are fired into a dense target. They also have the advantage of using secondary beams of particles that have been formed by the interactions in the primary collision of the original high-energy particles with the target. These secondary beams of particles can be used to strike other targets and perform other experiments, or even to form tertiary beams for other experiments. More than 15 different experiments were run simultaneously using the fixed target accelerator at Fermilab. Fixed target accelerators also have the advantage of a wide range of target materials.

Although colliders have the obvious advantage of higher energies, they produce a far smaller number of interactions and they are limited by the choice of colliding objects (which have generally been electrons and positrons or protons and antiprotons). This situation is changing as the RHIC (Relativistic Heavy Ion Collider) came on line in 2000 and the LHC (Large Hadron Collider) is now being developed at CERN (page 512).

The Tevatron at Fermilab can be used for fixed-target research or converted to form a storage-ring collider using protons and antiprotons. As a collider, the Tevatron can accelerate the protons and antiprotons to 980 GeV for a total energy of 1960 GeV.

26.2

THE STANDARD MODEL OF PARTICLE PHYSICS

eBook plus

Weblink:

The Particle Adventure

This is an excellent web site on particle physics and the Standard Model.

In the previous sections we encountered the discovery of just a few of the many sub-atomic particles. In the next section we will encounter some more. Before 1970, well over 200 different particles had been discovered. We note that some particles were predicted by new theories that were being developed. With the discovery of so many sub-atomic particles, a search was begun for a structure that would help to explain the existence of the particles and perhaps even identify some truly ‘fundamental’ particles. The existence of quarks as fundamental particles was predicted and the concept of quarks gained acceptance when the first quarks were detected.

A model called the ‘Standard Model’ of particle physics was developed. The Standard Model is a mathematical description of all known particles and the forces between them. It enables us to explain all the behaviour of these particles. A very close interplay between experimental and theoretical physics prompted the development of the Standard Model. There are still problems with some aspects of it, and as the 1999 Nobel prize winner Gerard ‘t Hooft (1946–) has said, ‘We do admit that the model is not absolutely perfect... however, the degree of perfection reached is quite impressive’.

New particles

In the early 1930s, the proton and neutron had been identified as the constituents of the nucleus, and electrons were known to be in orbit around the nucleus. With this knowledge the constituents of matter seemed to have been identified. In 1933, the situation became a little clouded when Carl Anderson discovered the positron and when Pauli predicted the existence of the neutrino.

It was soon realised that more and more particles were awaiting discovery.

In 1936, Anderson and Seth Neddermeyer observed tracks in a cloud chamber that did not match that of electrons or protons. The tracks were produced by cosmic rays and were too thin to be made by protons. The particles that made them penetrated thick lead plates that would stop electrons. It seemed that the tracks were made by particles that had a mass between that of an electron and a proton. In fact the mass was shown to be about $100 \frac{\text{MeV}}{c^2}$. A new particle had been discovered. This particle was originally called a mesotron but is now called a muon. (Rather than use grams or kilograms for mass, nuclear physicists usually use MeV or more correctly $\frac{\text{MeV}}{c^2}$ as their measure of mass. The masses of some common particles are given in table 26.1.)

The muon at first appeared to be the particle that Hideki Yukawa (1907–1981) had proposed in 1935 to explain the strong nuclear force. He had predicted a particle with a mass about 200 times that of an electron. However, there were problems with this idea and after the development of a very fine-grained photographic emulsion in 1947, another particle, the pion, with a mass of 140 MeV was discovered. The pion was the particle predicted by Yukawa. In fact, the pion disintegrated into a muon and a neutrino.

Table 26.1 Particle masses

PARTICLES	MASS ($\frac{\text{MeV}}{c^2}$)
Proton (p)	938.3
Neutron (n)	939.6
Electron (e)	0.511
Muon (μ)	105.7
Pion (π^+)	139.6

Initially, the classification of particles was done by mass: leptons (light), mesons (intermediate) and baryons (heavy).

The muon, being a particle of intermediate mass, was originally called a mu-meson. This classification system has been changed and particles are now classified in terms of their interaction. The muon is not a meson but a member of the group called leptons (see page 504).

Particles are now classified as hadrons and leptons. Particles which experience the strong force are called hadrons and are named after the Greek word for strong. Particles that do not experience the strong force are called leptons.

Hadrons

Hadrons are particles that experience the strong nuclear force. Mesons and baryons are both hadrons.

Baryons

Baryons are hadrons that have half-integer spin (and are fermions). Examples are the proton and neutron. Some other baryons are included in table 26.1 on page 503.

Mesons

Mesons are hadrons that have zero or integer spins. Some of the mesons with zero spin are included in table 26.2 on page 507.

Leptons

Leptons are particles which do not experience the strong nuclear force. They are all fermions with half-integer spin. An electron is a lepton.

Fermions

Fermions are particles that have half-integer spins. They obey the Pauli exclusion principle.

Bosons

Bosons are particles that have either integer or zero spin. They do not obey the Pauli exclusion principle. Bosons are force-carrying particles.

eBookplus

Weblink:
Bose-Einstein
condensation

Gell-Mann had discovered that many particles could be organised into families of eight or ten. He believed that his theory was related to the concept of group theory from mathematics. The particles were graphed in terms of certain quantum numbers that were given such exotic names as 'isotopic spin' and 'strangeness'.

Developments leading to the Standard Model

By the early 1960s, about 100 particles had been discovered using new accelerators and improved particle detectors. Attempts were made to organise these particles and perhaps find an underlying structure.

The organisation of elements (as shown in the periodic table, see Appendix 2, page 528) was understood when the details of atomic structure, in terms of a nucleus and orbiting electrons, had been discovered. Physicists at this time wondered whether a pattern would also be discovered for hadrons. When Enrico Fermi was asked about the names of some of these particles he made his famous response, 'If I could remember the names of all these particles I would have been a botanist'.

While we sympathise with Fermi's view, we have to look at some of the terms that are collectively assigned to different groups of particles. **Hadrons** are particles (including **mesons** and **baryons**) that experience the strong force. **Leptons** are particles that do not experience the strong force. All leptons have half-integer spin and are called **fermions**. They obey the Pauli exclusion principle.

Some hadrons are fermions, having a half-integer spin, and some are **bosons** with either integer or zero spin. Bosons do not obey the Pauli exclusion principle.

PHYSICS FACT

The Pauli exclusion principle forbids two fermions from existing in exactly the same quantum state. (The arrangement of electrons in atoms is a reflection of the Pauli exclusion principle. Electrons, which are fermions, cannot accumulate in the lowest energy state because they cannot exist in the same quantum state.)

However, it is a different situation with bosons.

In 1995, physicists at Boulder, Colorado, managed to produce a *Bose-Einstein condensate* in which about 2000 rubidium-87 atoms were confined to a single quantum state of approximately zero energy. The rubidium-87 atoms are bosons which do not obey the Pauli exclusion principle, so it is possible to have a large number in the same quantum state.

The Eightfold Way

In 1961, Murray Gell-Mann (1929–) in the USA and Yuval Ne'eman (1925–2006), an Israeli theorist in England, independently discovered a method of organising particles. Gell-Mann called the method, perhaps a little irreverently, by the Buddhist term 'The Eightfold Way'.

The theory suggested that there was a missing particle. It was called the Ω^- (omega minus). In 1963, a search for this particle was started using the bubble chamber at Brookhaven. The bubble chamber was about two metres in diameter and contained liquid hydrogen. Every few seconds, a burst of kaons (K-mesons), collided with protons (the nuclei of the atoms of liquid hydrogen) in the bubble chamber. This produced a spray of particles that, it was hoped, would include the Ω^- . Eventually, in photograph number 50 321, an event indicating the existence of the Ω^- was discovered. This photograph is shown in figure 26.8 along with a sketch that shows the particles involved in the interaction.

This confirmed the Eightfold Way's organisation of particles, but the reason for this organisation was still unknown.

Quarks

In 1964, Murray Gell-Mann and George Zweig (1937–), both from the California Institute of Technology, but working independently, proposed that there were three fundamental particles that were the constituents of hadrons. Gell-Mann named these particles ‘quarks’.

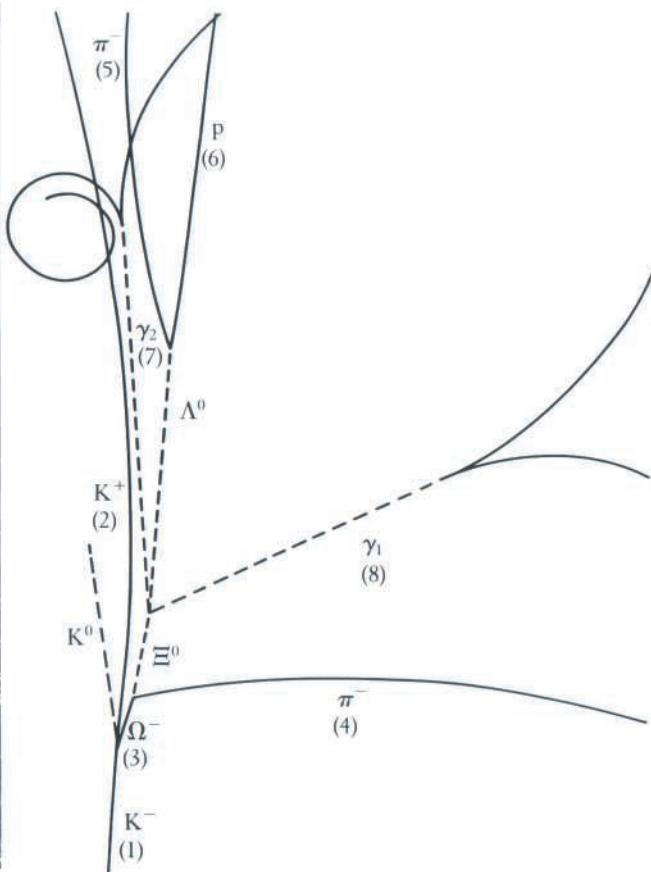
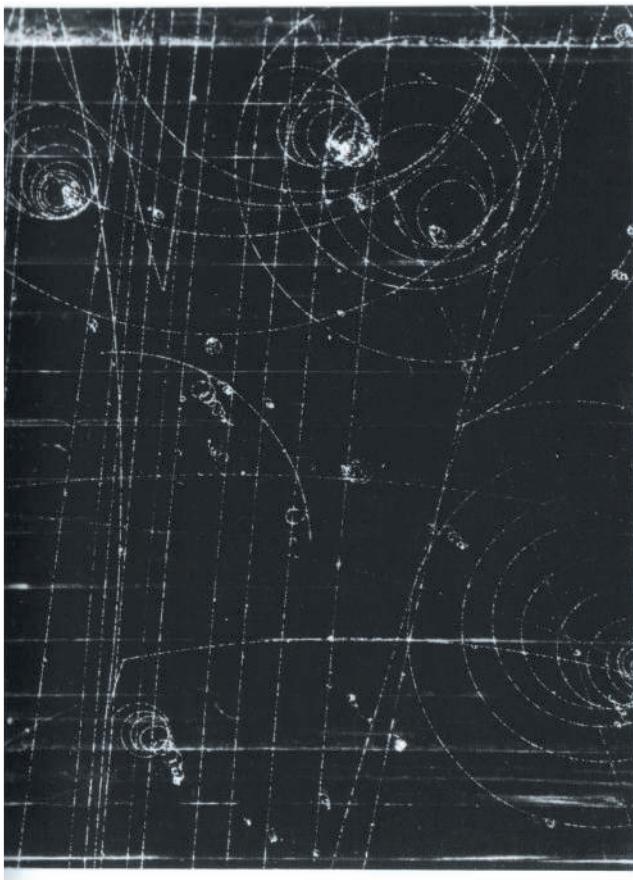


Figure 26.8 The bubble chamber photograph taken at Brookhaven National Laboratory that shows the path of the Ω^- particle. The diagram at the right identifies the particles responsible for the various trails. Dotted lines show the paths of particles not visible in the photograph. The incident K^- particle (1) collided with a positron at (3). The short tail at (3) was produced by the Ω^- before it decayed to the π^- and ultimately a number of other particles of which some left trails in the photograph.

George Zweig named his particles ‘aces’ while Murray Gell-Mann named his ‘quarks’.

Gell-Mann had in mind the sound ‘kwork’ rhyming with squawk or pork. He then came across the quote ‘Three quarks for Muster Mark’ from Finnegan’s Wake by James Joyce. He decided that he could spell the name of the particle ‘quark’ but pronounce it to rhyme with squawk. Gell-Mann tried to reason why Joyce may have intended ‘quark’ to sound as he wanted but from the quote it seems that it should rhyme with ‘Mark’.

Gell-Mann’s name stuck while Zweig’s was lost. Both pronunciations of quark are accepted today.

The idea for the existence of quarks came from the arrangement of baryons and mesons in the patterns of the ‘Eightfold Way’.

The idea of quarks simplified the structure of hadrons. It was believed that each baryon was composed of three quarks and that each meson was composed of two quarks. However, there was still doubt about the reality of quarks.

In the late 1960s and early 1970s, experiments conducted at SLAC (Stanford Linear Accelerator Center) provided convincing evidence of the existence of quarks. These experiments involved bouncing high-energy electrons off protons in an attempt to discover the structure of protons. In some ways this was using a similar principle to Rutherford’s scattering experiment in which he bounced alpha particles off atoms to find the structure of atoms. However, the energies involved in Rutherford’s scattering experiments were much smaller.

In his book *The God Particle*, Leon Lederman (1922–), Nobel prize winner and a past director of Fermilab, presented the following extended analogy for many puzzles in physics, particularly particle physics. He wrote of a World Cup soccer match attended by some intelligent extraterrestrial beings from the planet he calls Twilo. He describes these beings as similar to humans, except that they cannot see objects with the sharp juxtapositions of black and white, like zebras or soccer balls.

The Twiloans watch the soccer game and at first are totally mystified at seeing people running around and doing many strange things for apparently no reason, and also the reaction of the crowd at certain events. They make charts of what is happening but nothing makes sense until one of them suggests that there is an invisible ball involved. He has noticed that there is a slight hemispherical bulge in the net of the goal just before the referee blows his whistle, the crowd cheers madly and a point is added to the score. Once this idea of a ball that is invisible to Twiloans is accepted, all the previous charts remain valid but now there is a meaning to all the events.

Quarks cannot be observed directly and a theory has been developed that predicts that they cannot exist outside hadrons. However, like the invisible soccer ball, other observations make sense only if the presence of quarks is accepted.

Unification of the electromagnetic force and the weak nuclear force

The unification of the electromagnetic force and the weak nuclear force resulted from breakthroughs by Steven Weinberg (1933–), working in the USA, and Abdus Salam (1926–), working independently in England. They extended the idea of a force-carrying particle to the weak nuclear force and argued that the weak and electromagnetic forces were really the same thing.

The force-carrying particle, or boson, for the electromagnetic force was the photon. Weinberg and Salam proposed that there would be three force-carrying bosons called W^+ , W^- and Z^0 , for the weak interactions. These bosons were termed *intermediate vector bosons*. These bosons were very heavy, about one hundred times the mass of the proton. Ultimately however, the electromagnetic and weak force were manifestations of the same force that they called the electroweak force.

It was the search for these intermediate vector bosons that led Carlo Rubbia to convert the 400 GeV accelerator at CERN to a collider (page 502). The total energy of 520 GeV from the collisions was then sufficient to enable the discovery of the W particles. The theory predicted the mass of the 'W's to be about 80 GeV.

In 1983, Rubbia worked with Simon Van der Meer (1925–) and a group of about 130 physicists and provided the final evidence of the electroweak force with the discovery of the Z^0 boson. In 1984, Rubbia and Van der Meer shared the Nobel prize for the discovery of the W and Z particles.

Particles of the Standard Model

The particles of the Standard Model are quarks and leptons and are shown in table 26.2 on the following page. Today it is accepted that there are six flavours of quarks and six flavours of leptons. The description, 'flavours', just means different types of quarks or leptons. Quarks and leptons can be divided neatly into groups called 'generations'. All the visible matter in the universe is composed of first-generation quarks and leptons, the up and down quarks, and electrons.

eBookplus

Weblink:

The Standard Model

Table 26.2 The particles of the Standard Model

GENERATION	LEPTONS				QUARKS			
	NAME	SYMBOL	REST MASS (MeV)	ELECTRIC CHARGE	NAME	SYMBOL	REST MASS (MeV)	ELECTRIC CHARGE
I	Electron neutrino	ν_e	≈ 0	0	Up	u	≈ 5	$+\frac{2}{3}$
	Electron	e^-	0.511	-1	Down	d	≈ 7	$-\frac{1}{3}$
II	Muon neutrino	ν_μ	≈ 0	0	Charm	c	1500	$+\frac{2}{3}$
	Muon	μ^-	105.7	-1	Strange	s	≈ 150	$-\frac{1}{3}$
III	Tau neutrino	ν_τ	< 35	0	Top	t	170 000	$+\frac{2}{3}$
	Tau	τ^-	1784	-1	Bottom	b	≈ 5000	$-\frac{1}{3}$

Quarks

Gell-Mann and Zweig first proposed that hadrons were composed of only three quarks. It was predicted that there were three different types of quarks and these were called up, down and strange. Later it was necessary to add more quarks and these became charm, discovered in 1974, bottom, discovered in 1976, and top, discovered in 1995.

In the strange language of particle physics, these types of quarks became known as ‘flavours’. There are six different flavours of quarks, up, u , down, d , strange, s , charm, c , top, t , and bottom, b . The top and bottom quarks were sometimes referred to as ‘truth’ and ‘beauty’ but top and bottom are now the accepted names. See ‘Physics in focus’ (pages 510–511) for an account of the discovery of the top quark.

The particles of the Standard Model are given in table 26.2 above. Quarks possess charges that are either $+\frac{2}{3}$ or $-\frac{1}{3}$ of the charge on an electron.

A proton is composed of two up quarks, each of charge $+\frac{2}{3}$, and one down quark with a charge of $-\frac{1}{3}$, giving the proton a charge of +1.

A neutron is also composed of up and down quarks, but one up and two down quarks are required to produce a neutral particle.

Other combinations of quarks that form baryons (composed of three quarks) and mesons (composed of a quark and an antiquark) are listed in table 26.3 on the following page.

Leptons

The first known leptons were the electron, the muon and neutrinos. Leptons are regarded as being fundamental particles.

In 1961, the alternating gradient synchrotron at Brookhaven was used to bombard a beryllium target with 15 GeV protons. Among the products were pions which then decayed into muons and neutrinos. A steel barrier over 10 m thick was used to filter out all particles except for the neutrinos. Studies of this beam of neutrinos showed that in the unlikely event that they did interact with matter, it was always muons and not electrons that were associated with the interactions. Therefore, it was concluded that there were in fact two types, or flavours, of

neutrinos — electron neutrinos and muon neutrinos. In 1976, the Stanford Positron Electron Annihilation Ring (SPEAR) provided evidence of another lepton, named tau.

It was predicted that another neutrino, the tau neutrino, must be the sixth lepton of the Standard Model. Even before there was any suggestion of its discovery, it was accepted that the tau neutrino was the sixth lepton in the Standard Model. In July 2000, an international collaboration of scientists at Fermilab announced the first direct evidence for the tau neutrino.

Table 26.3 Some hadrons and their properties

The composition of baryons and mesons. The particles in the table are composed of combinations of two or three quarks given the names up, *u*, down, *d*, strange, *s* and charm *c*. In the case of mesons, it is a combination of a quark and an antiquark. Baryons are composed of three quarks.

PARTICLE	MASS MeV c^2	CHARGE RATIO $\frac{Q}{e}$	SPIN	QUARK CONTENT
Mesons				
π^0	135.0	0	0	$u\bar{u}, d\bar{d}$
π^+	139.6	+1	0	$u\bar{d}$
π^-	139.6	-1	0	$\bar{u}d$
K^+	493.7	+1	0	$u\bar{s}$
K^-	193.7	-1	0	$\bar{u}s$
η^0	547.5	0	0	$u\bar{u}, d\bar{d}, s\bar{s}$
Baryons				
p	938.3	+1	$\frac{1}{2}$	uud
n	939.6	0	$\frac{1}{2}$	udd
Λ^0	1116	0	$\frac{1}{2}$	uds
Σ^+	1189	+1	$\frac{1}{2}$	uus
Σ^0	1193	0	$\frac{1}{2}$	uds
E^-	1197	-1	$\frac{1}{2}$	dds
Ξ^0	1315	0	$\frac{1}{2}$	uss
Ξ^-	1321	-1	$\frac{1}{2}$	dss
Δ^{++}	1231	+2	$\frac{1}{2}$	uuu
Ω^-	1672	-1	$\frac{1}{2}$	<mathsss< math=""></mathsss<>
Λ_c^+	2285	+1	$\frac{1}{2}$	udc

PHYSICS IN FOCUS

Boson force-carriers in the Standard Model

There are four fundamental forces, the electromagnetic force, the weak nuclear force, the strong nuclear force and the force of gravity. The Standard Model, at present, describes three of these forces, the electromagnetic force and weak nuclear force (which are unified in the electroweak force) and the strong nuclear force, in terms of force-carrying particles called bosons.

As we saw on page 506, there are four boson force-carriers that have been experimentally identified with the electroweak force. These are the W^+ , W^- , Z^0 and photon. There are eight gluons that have been identified that are associated with the strong force. The force of gravity is not yet included but a boson, the graviton, has been predicted to be the force-carrier for the force of gravity. However, it has not yet been discovered. In the next sections we will encounter another feature of quarks — the colour charge. We will see that colour charge is associated with gluons and the strong force.

Colour — another property of quarks

Quarks are fermions, which are particles that have a spin of $\frac{1}{2}$. Fermions obey the Pauli exclusion principle, so it seems that it should not be possible for baryons to consist of three particles all in the same quantum state. The Ω^- consists of three strange quarks, apparently with two of these in identical quantum states. This appears to violate the Pauli exclusion principle. Two quarks can exist in an identical quantum state because they can have spin in the opposite directions, called spin up or spin down. If a third quark is added, it must also have spin up or spin down and hence the Pauli exclusion principle is apparently violated.

This difficulty was overcome by proposing that each quark has three different varieties. These varieties were called ‘colour’ but this, of course, has nothing to do with real colour. It is just another whimsical name applied to quarks.

The colours are usually called red, green and blue and it was assumed that the Pauli exclusion principle applied differently to each colour.

This then means that the Ω^- consists of a red s , a green s , and a blue s , so there is no longer a problem with the Pauli exclusion principle.

While individual quarks have colour, any particle composed of quarks must have no net colour. A combination of a red, a green and a blue quark results in a particle lacking in colour. Baryons always contain one quark of each colour and, hence, are colourless.

Although quarks have colour, antiquarks have anti-colour. Mesons, which are a quark–antiquark pair, are colourless because they have a quark for colour and a corresponding quark for anticolour.

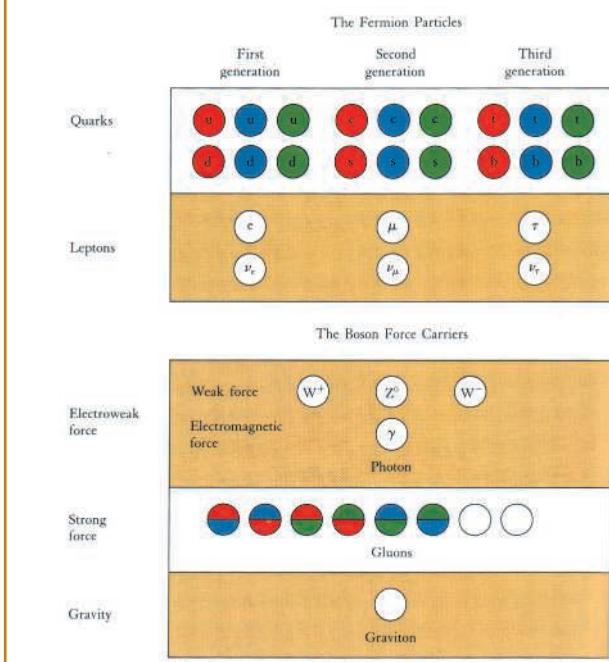
Gluons and the strong force

When we encountered the strong nuclear force in chapter 24 (page 466), we noted that it involved the exchange of pions between nucleons. In the Standard Model, the boson force-carriers for the strong force are gluons that are exchanged between quarks. (The role of mesons in the force between nucleons is really a more complex and secondary example of the strong force.)

Gluons have no mass but they carry the ‘colour charge’. They actually carry a colour and an anticolour. We are used to positive and negative charges. With colour charge we are dealing with something similar to positive and negative charge, but with three varieties rather than two. The three different varieties are called the colour charges.

Table 26.4 The particles and boson force-carriers of the Standard Model.

Gluons have no mass but carry the colour ‘charge’. They carry colour and an anticolour. While the first pair of gluons shown appear to be both red blue, one is red antibleue and the other is blue antired. It is not easy to represent anticolours in a diagram!



(continued)

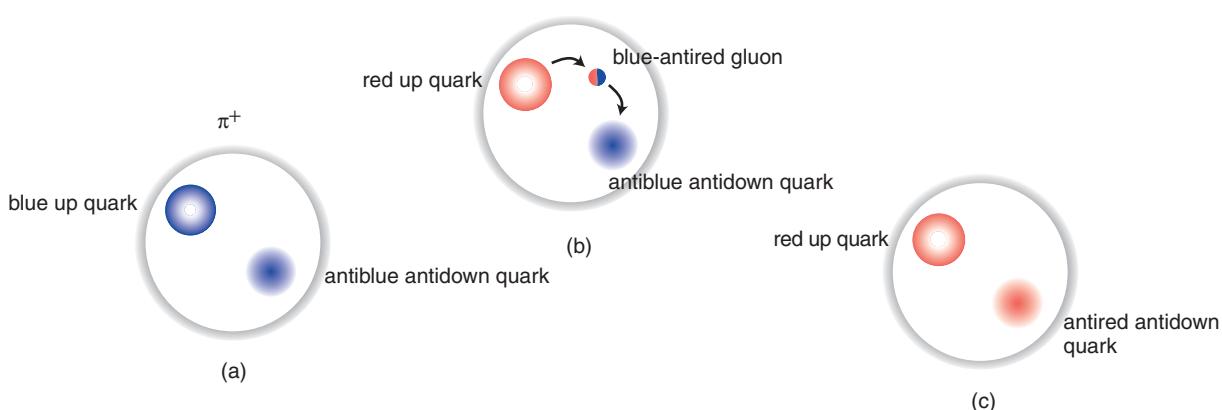


Figure 26.9 (a) The structure of π^+ , which contains an up quark and an antidown quark. In this diagram the up quark is blue and the antidown quark is antiblue. (b) The blue up quark emits a blue-antired gluon leaving a red up quark. (c) The antiblue antidown quark absorbs the blue-antired gluon and hence forms an antired antidown quark. The π^+ now has changed to having a red up quark and an antired antidown quark.

The study of the interaction of light and electrons is called quantum electrodynamics (QED). The study of the colour force and interactions involving gluons is called quantum chromodynamics (QCD).

Figure 26.9 shows the exchange of a gluon between the up and antidown quarks of a π^+ (a positively charged pion). The π^+ can be red-antired, green-antigreen or blue-antiblue. It

remains a π^+ when gluons are exchanged between the up and antidown quarks, even though the colours of the quark and anti-quark change. As shown in (b) and (c) the colour of the blue up quark and antiblue antidown quark will be changed to red and antired when the blue up quark emits a blue antired gluon that is absorbed by the antiblue antidown quark.

PHYSICS IN FOCUS

The discovery of the top quark

The bottom quark was discovered in experiments conducted at Fermilab in 1977. It had a mass of about 5 GeV. After this discovery it was predicted that another quark, the top quark, must exist and it was thought that it would have a mass between 10 and 30 GeV. The Standard Model predicted many of the properties of this undiscovered quark, but did not limit its mass.

By 1988, experiments at CERN had not discovered the top quark and it was concluded that its mass must be greater than 41 GeV.

The Fermilab collider had been activated in 1985, and in 1988 and 1989 the CDF (Collider Detector at Fermilab) group were in intense competition with the group at CERN. An energy of 77 GeV was reached without the top quark being detected. Leon Lederman, the director of Fermilab during the 1980s, decided that some local competition would be a positive move. Another group, DØ (pronounced ‘dee zero’), was created

and began recording data in 1992. CDF and DØ were international collaborations of more than 400 physicists each. They also included large numbers of engineers and technical staff.

The CDF and DØ collaborations constructed enormous, complicated instruments to try to detect the ‘signature’ of the top as it passed through the detector. The two groups had different approaches but expected that if one group found any evidence of the top, the other should be able to find supporting evidence. If a top and anti-top were produced, they would decay almost instantly into a W and a bottom quark. Hence, the top and anti-top would produce two Ws and a bottom and antibottom. Unfortunately, neither the W nor the bottom or antibottom could be observed directly. What was observed was a ‘jet’, a directed beam of particles that travelled in roughly the same direction as the original top quark.

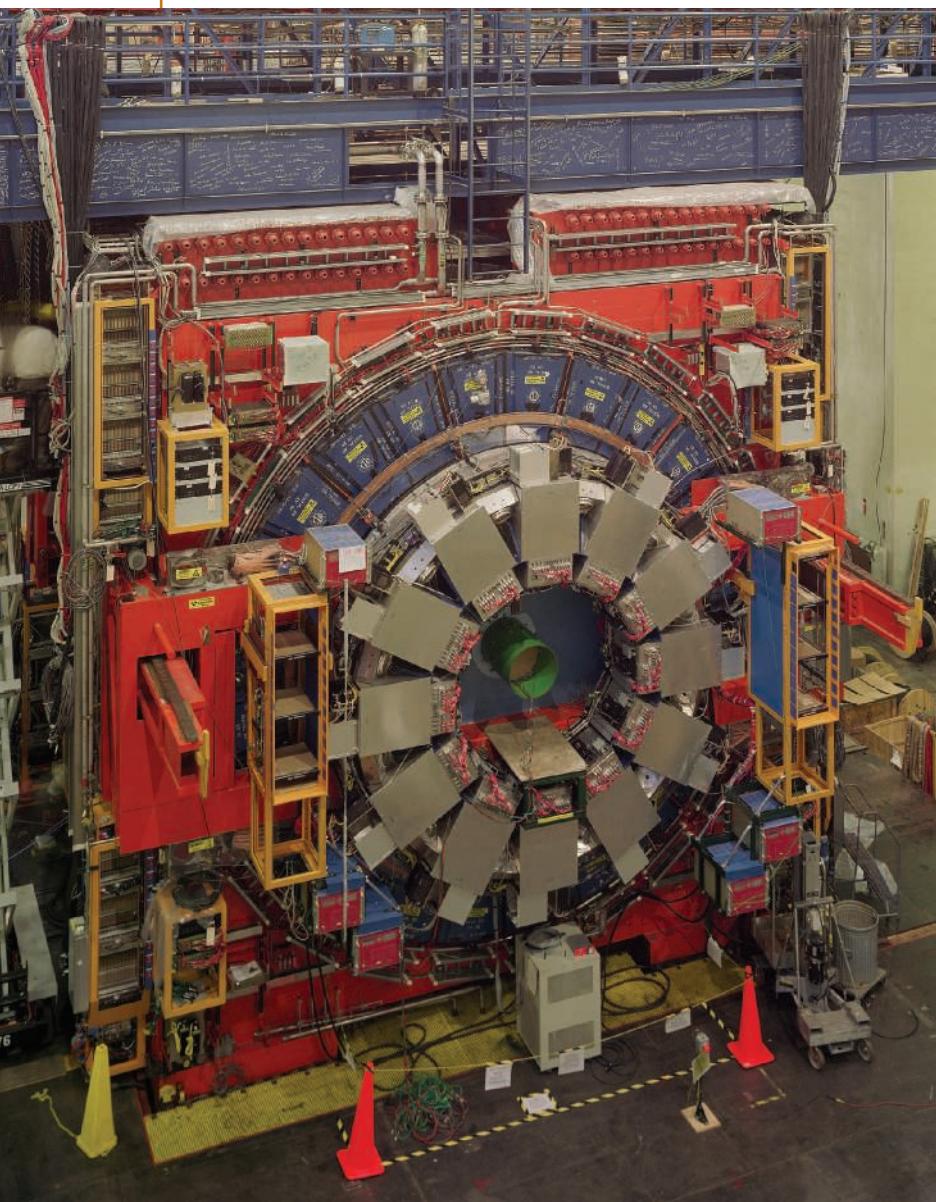


Figure 26.10 The Collider Detector at Fermilab (CDF). The detector, which is three storeys high, is shown with the modules of the central calorimeter moved to the sides. The detector provides 70 000 channels of data to computers.

In 1994, the CDF group had isolated 12 events that may have involved a top–antitop pair. There was a possibility that the 12 events were really just caused by background events that gave the same signature as the top–antitop. However, the group estimated that 5.7 of such background events were to be expected but that the probability of all the events being background events was less than 1 in 400.

The CDF group then estimated the total energy of the particles in the jets, and also the associated leptons, and found that all of the energies clustered in a small range around 175 GeV. If it had been background events, a wider spread of energies would have been expected.

CDF then had to write a paper to satisfy its 400 members. This did not prove to be an easy task. The paper was submitted in April 1994.

The DØ group had focused their research on the search for a lighter particle and had little to support CDF. DØ performed a re-analysis based on the higher mass and found promising results. When the final presentation was made in March 1995, both CDF and DØ showed overwhelming evidence for the top quark. More events had been detected and the probability that background events could explain the data had been reduced to less than 1 in 500 000.

How very different this is from the work performed in the Cavendish Laboratory during the first half of the twentieth century!

The Standard Model — today and beyond

The Standard Model is a great achievement and a large number of experiments have confirmed, sometimes to incredible precision, the predictions of the Standard Model. However, it is acknowledged that there are serious flaws with it. Some of those are listed below and on the following page.

- It is incompatible with Einstein's general theory of relativity. Therefore, unification of the forces cannot involve the force of gravity. The Standard Model is a quantum–mechanical model while general relativity is not.

- The Standard Model provides no reason for the numbers of particles. Why are there six quarks and six leptons? Is it coincidence or is there an underlying reason?
- Is there some underlying truly fundamental particle such as a ‘leptoquark’? This might explain why three leptons have electrical charges of one unit and quarks have electrical charges of $+\frac{2}{3}$ or $-\frac{1}{3}$ of this unit of electrical charge.
- The Standard Model does not have a mechanism to generate the observed masses of particles. (This objection may disappear if the Higgs boson is detected.)

PHYSICS FACT

The Higgs boson

A difficulty with the electroweak unification theory was that the W and Z bosons are massless at some energies but need to acquire mass at lower energies. At high temperatures and energies, the W and Z are similar to a photon, the other force carrier of the electroweak force which has no rest mass.

At lower energies, the W and Z need to acquire mass, while the photon remains with zero rest mass. Peter Higgs (1929–), a Scottish physicist, was among those who proposed a mechanism for providing mass for these particles. He proposed the existence of a new field now called the Higgs field. A particle called the Higgs particle or boson is associated with this field. At low temperatures, space will be filled with Higgs particles. The W and Z interact with the Higgs particles and do not travel through empty space at the velocity of light. They have acquired an effective mass through their interaction with the Higgs particles. The photon does not interact with the Higgs particles and so continues to travel at the

velocity of light. (Note that empty space being filled with Higgs particles is in some ways similar to the idea of an ether filling all space.)

At higher temperature and energies, the interactions of the Higgs particles are such that they do not fill space and the W and Z can pass through space at the speed of light. They are no longer slowed down and hence, have no mass.

The Higgs field can also account for other quarks and leptons having mass. In fact, it may answer the question of what we really mean by mass. (It is a mechanism for providing mass but cannot accurately account for exactly how much mass these particles have.) The mass of the Higgs boson is not known but some physicists think that they are on the verge of discovering the particle.

The search for the Higgs boson has been a massive task and Peter Higgs himself has stated ‘When I consider the huge sums going for this, the lifetimes spent in the search, I can’t help but think: “Good heavens, what have I done?”’

Recent developments in accelerator research

eBook plus

Weblinks:

[The Relativistic Heavy Ion Collider \(RHIC\)](#)
[RHIC animations and multimedia](#)
[The Large Hadron Collider \(LHC\)](#)

In June 2000, the Relativistic Heavy Ion Collider (RHIC) commenced operation at Brookhaven. Beams of gold ions are accelerated to 99.995% of the velocity of light. Two beams travel in opposite directions and are then deflected to intersect. There has been speculation that there may be sufficient energy to produce a quark-gluon plasma.

The Large Electron Positron collider (LEP) at CERN was scheduled to end its 11-year life at the end of September 2000. However, its life was extended for an extra two months after some events were detected that may have been decays associated with the Higgs boson. They did not, however, provide evidence of the Higgs boson.

The LEP is being replaced by the Large Hadron Collider (LHC), which will occupy the same 27 km circumference tunnel. The LHC, originally designed to come on line in 2005, then delayed to 2007, has been further delayed to 2008. The Higgs boson has proved elusive; it is beyond

the reach of existing particle accelerators, but researchers hope that the LHC will be able to detect it.

Future directions of research

In the United States, the National Academy of Sciences has set up a special committee to assess key questions about the nature of the universe. They have identified eleven key questions, which they hope will either be answered in the next decade or that we should be thinking of answering in the following decades.

Their eleven questions are:

- What is dark matter?
- What are the masses of neutrinos, and how have they shaped the evolution of the universe?
- What is the nature of dark energy?
- Are protons unstable?
- How did the universe begin?
- Are there new states of matter at exceedingly high densities and temperatures?
- Is a new theory of matter and light needed at the highest energies?
- How were the elements from iron to uranium made?
- Are there additional space-time dimensions?
- Did Einstein have the last word on gravity?
- How do cosmic accelerators work and what are they accelerating?

Further information is available at www.nationalacademies.org/bpa.

Most of these questions involve the close linking of particle physics and cosmology. Perhaps the new accelerators currently being developed will be able to shed light on some of these questions.

Source: Connecting Quarks with the Cosmos: Eleven Science Questions for the New Century, 8 January 2001, Committee on the Physics of the Universe

Table 26.5 Timeline of events in quantum and nuclear physics

1885	Balmer's Equation for spectral line of hydrogen
1895	C. T. R. Wilson begins developing cloud chamber. Wilhelm Röntgen discovers X-rays.
1896	Henri Becquerel discovers radioactivity.
1897	J. J. Thompson discovers the electron.
1898	Rutherford discovers that radioactivity consists of alpha and beta radiation.
1900	Villard discovers radioactivity also consists of gamma rays. Planck develops quantum theory. Rydberg modifies Balmer's equation for spectral lines of hydrogen.
1902	Transformation theory of radioactivity
1903	Hydrogen atom thought to contain about a thousand electrons
1905	Einstein proposes light quantum. Einstein introduces special relativity. Einstein introduces $E = mc^2$.
1906	Rutherford discovers alpha particle scattering. Beta particle spectrum detected
1908	Death of Becquerel Paschen series (infra-red) of spectral lines of hydrogen detected

(continued)

1909	Geiger and Marsden publish results of Marsden's experiment investigating the deflection of alpha particle scattering by thin metal foils. Rutherford and Royds identify alpha particles as doubly charged helium ions.
1911	Ernest Rutherford predicts the nuclear atom (based on his interpretation of the results of Geiger and Marsden).
1912	Cosmic radiation is discovered during manned balloon flights.
1913	Bohr publishes three papers that include his postulates that form the basis of the Bohr model of the atom. Millikan determines the charge on the electron.
1914	First detection of continuous beta spectrum
1916	Lyman series (ultraviolet) of spectral lines of hydrogen detected.
1919	Rutherford identifies the proton as a constituent of nuclei.
1920	Rutherford predicts the existence of the neutron.
1922	Brackett series (infra-red) of spectral lines of hydrogen detected
1924	De Broglie introduces particle-wave duality. Pfund series (infra-red) of spectral lines of hydrogen detected
1925	Pauli proposes his exclusion principle. Heisenberg's first paper on quantum theory is published. Uhlenbeck and Goudsmit discover spin. Born, Heisenberg and Jordan's paper on matrix mechanics is published.
1926	Equation of hydrogen spectrum derived from matrix mechanics. Schrödinger's first paper on wave mechanics is published. G. N. Lewis uses the name 'photon' for a light quantum.
1927	Davisson and Germer discover electron diffraction. Heisenberg introduces the uncertainty principle. Bohr introduces his idea of complementarity.
1928	Dirac presents his equation.
1931	Lawrence invents the cyclotron. Pauli predicts the existence of the particle now known as the neutrino. Dirac proposes the existence of the positron.
1932	Chadwick discovers the neutron. Anderson discovers the positron.
1933	Fermi publishes the theory of beta decay. Szilard has the idea of chain reaction.
1934	Fermi discovers that slow neutrons are much better than fast neutrons when irradiating elements. Discovery of beta ⁺ radioactivity
1936	Discovery of meson, later called muon, in cosmic rays
1937	Death of Rutherford
1938	Hahn and Strassman discover the presence of barium after bombarding uranium with slow neutrons.
1939	Meitner and Frisch realise that nuclear fission is taking place in the experiments of Hahn and Strassman.

1941	Plutonium discovered Term 'nucleon' introduced
1942	The Manhattan Project commences. Fermi produces controlled fission in an atomic pile in Chicago.
1945	First atomic bomb test Two atomic bombs are dropped on Japanese cities Hiroshima and Nagasaki.
1946	Term 'lepton' introduced
1953	First bubble chamber pictures taken
1954	Foundation of CERN Death of Fermi
1955	Discovery of the antiproton Death of Einstein
1956	Cowan and Reines detect antineutrino
1958	Death of Pauli
1961	Gell-Mann proposes 'Eightfold Way'. Death of Schrödinger
1962	Discovery of the muon neutrino Death of Bohr
1964	Hypothesis that all hadrons are composed of three quarks (and antiquarks) Introduction of fourth quark
1976	Death of Heisenberg
1977	Discovery of fifth quark, bottom
1983	Discovery of the W and Z bosons
1984	Death of Dirac
1986	Chernobyl nuclear accident
1995	Discovery of sixth quark, top
2000	Detection of tau neutrino

SUMMARY

- The method that physicists have used to discover information on the most basic structure of matter is to bombard the matter with high-energy particles. The particles produced in these interactions are then detected and analysed. This process has led to a quest for particles of higher and higher energies.
- Cloud chambers were a very useful detector in the first half of the twentieth century but were replaced by bubble chambers which were much more efficient detectors. These, in turn, have been replaced by multi-component detectors that include tracking chambers and calorimeters.
- With the invention of new types of particle accelerators the energies to which charged particles can be accelerated in particle accelerators have increased dramatically. The original one-stage electrostatic accelerators such as the Cockcroft–Walton machine and the Van de Graaff generator are still used as the pre-accelerators in some of today's biggest accelerators.
- While linear accelerators and cyclotrons were once very important accelerators, the main accelerators today are synchrotrons in which particles are accelerated around a constant radius path.
- Early accelerators fired high-energy particles at fixed targets; however, conservation of momentum and energy restricts the amount of energy that can be used in the production of new particles. Much more energy is available for particle production when similar mass particles, travelling in opposite directions (and hence having little total momentum) collide. A recently constructed collider is the RHIC at Brookhaven. The LHC is an even more energetic collider being built at CERN, and it will be completed by 2005.
- By 1960, hundreds of different types of particles had been detected and there seemed to be little pattern to link all the particles. Murray Gell-Mann and Yuval Ne'eman discovered that there was an underlying organisation of all these particles which Gell-Mann called the 'Eightfold Way'. This organisation method prompted the prediction of the existence of another particle. The actual detection of the particle confirmed this method of organisation.

- Gell-Mann and Zweig believed that there was an underlying structure and that each baryon was made up of three truly fundamental particles, named quarks by Gell-Mann. They believed that mesons were composed of a quark and an anti-quark.
- This led to the development of the 'Standard Model' in which all matter can be formed from twelve fermions — six flavours of leptons and six flavours of quarks. These particles interact with each other via the electroweak force, the strong force and the force of gravity. These interactions occur through force-carrying bosons. Eight gluons carry the strong force or colour force between quarks. The force between leptons is carried by the three intermediate vector bosons called W^+ , W^- , Z^0 , and the photon. The graviton, the conjectured force carrier of the gravitational force, has not been detected.
- The Standard Model may have been a great scientific achievement; however, it has serious flaws. It is incompatible with the general theory of relativity, Einstein's theory of gravity. It cannot explain why there are six leptons and six quarks. It does not have a mechanism to explain the mass of particles, but perhaps the detection of the Higgs boson will alleviate this difficulty.
- The construction of more powerful accelerators such as the colliders RHIC and LHC might enable physicists to produce a QGP (quark-gluon plasma). If they are able to do this it would produce matter in the form that dominated the universe one millionth of a second after the big bang. The LHC may have sufficient energy to detect the Higgs boson.

QUESTIONS

- In 1930, Ernest Lawrence constructed the first cyclotron. It was approximately 12.5 cm in diameter and accelerated hydrogen ions in a magnetic field of 1.27 T between the poles of a magnet 10 cm in diameter. The accelerating voltage applied to the dees was 2000 V and Lawrence determined that the hydrogen ions had been accelerated to an energy of 80 000 eV.
 - How many times had the hydrogen ions experienced the 2000 V accelerating potential and how many orbits of the cyclotron did they complete?
 - Determine the velocity of an 80 000 eV hydrogen ion (proton).

- (c) A charged particle moving in a circular path in a magnetic field has a centripetal force of magnitude $F_c = \frac{mv^2}{r}$ provided by the magnetic force $F_B = qvB$. Determine the radius of an 80 000 eV proton in a magnetic field of 1.22 T.
2. The Stanford Linear Accelerator (SLAC) was used to accelerate electrons to very high velocities. Suggest a reason why a linear accelerator was preferable to a cyclic accelerator, such as a cyclotron, as a device to produce very high energy electrons. (*Hint:* you may wish to think back to the Rutherford model of the atom and the electrons in orbit around the nucleus.)
3. The latest high-energy particle accelerators that have been constructed have been colliders, or at least have been able to function as colliders as well as fixed-target devices.
- (a) Consider a collider where a beam of 200 GeV protons interacts with a beam of 200 GeV antiprotons and an accelerator where a beam of 400 GeV protons were fired at a fixed target. With reference to momentum and energy, explain why there would be much more energy available for formation of new particles in the collider than in the fixed target accelerator.
- (b) Obviously the difference in available energy shows that a collider has a significant advantage over a fixed-target accelerator. However, there are some advantages that a fixed target accelerator has over a collider. List some of these advantages.
4. (a) A proton is composed of two up quarks and one down quark, and a neutron is composed of one up quark and two down quarks. Show that a proton will have a charge of plus one unit and a neutron will be neutral.
- (b) An antiproton would be expected to have a charge opposite to that of a proton. Identify the quarks (or antiquarks) in an antiproton and show that it will have a charge of minus one unit.
- (c) A neutron has no charge but still has an anti-particle. Identify the quarks in an antineutron and show that an antineutron is neutral.
5. In 1983, particle physicists in the United States proposed that a new accelerator be constructed. This new accelerator was called the Superconducting Supercollider (SSC) and was to be approximately 86 kilometres in circumference. It was planned to be approximately 60 metres below the ground surrounding the city of Waxahachie, Texas. The total cost was estimated to be eight billion dollars.
- Construction began in 1990, but the project was cancelled in 1994 when it was about 20% complete. (A search on the internet will yield information about the project and reasons for and against stopping the construction.)
- Prepare material that could be used in a debate, either to support or to oppose the expenditure of such a large amount of money on such a project.
6. The Large Electron–Positron collider (LEP) at CERN was planned to be shut down on 11 September 2000. (The LEP is to be replaced by the Large Hadron Collider (LHC) which may start in 2008.) Researchers using the LEP thought that they were on the verge of discovering the Higgs boson and requested that the LEP continue. Its life was extended until 2 November 2000. By that time researchers thought that they were even closer to the discovery of the Higgs. They believed that they had reduced the possibility of their results being due to statistical fluctuations to less than two parts in 1000. They believed that in 2001 they would be able to increase the energy of the LEP and complete a run which would reduce the uncertainty to three parts in 10 million.
- Researchers were shocked when they were not granted a further extension in the life of the LEP and it was finally shut down on 2 November 2000.
- There is still considerable debate about the decision not to extend the life of the LEP. Much of the information is available on the internet.
- Research this decision and decide if you support the closure or if you think that the life of the LEP should have been extended.



26.1 CLOUD CHAMBERS

Aim

To observe the tracks produced by charged particles passing through a cloud chamber.

Safety

Tongs or forceps should be used when handling radioactive materials. They should not be handled with bare hands. The sources encased in perspex are probably not suitable for use in cloud chambers.

Some cloud chambers have sources already mounted. Special care should be taken with others that rely on a radioactive salt such as thorium oxide. Gloves should be worn if it is necessary to open a bottle containing thorium oxide powder.

You should wash your hands after handling any radioactive sources.

If using a diffusion cloud chamber, gloves should be worn when handling dry ice and special care taken when breaking and or crushing the dry ice.

Introduction and theory

As there are two different types of cloud chamber available for use in schools, we will describe how each can be used. The principle of operation is the same for both types as they rely on the production of a supersaturated vapour (see page 496). As ionising radiation passes through the supersaturated vapour, droplets condense on the ions formed and reveal the path of the radiation.

Diffusion cloud chamber

The diffusion cloud chamber was developed in 1939 by Dr Alexander Langsdorf, Jnr at the University of California at Berkeley. A diffusion cloud chamber can operate continuously, apparently for hours at a time instead of the few seconds of an expansion type chamber, but the small devices used in schools are unable to maintain the correct conditions for much longer than fifteen to twenty minutes. In these cloud chambers, alcohol from the top of the chamber evaporates and then diffuses downwards towards the base of the chamber. The base of the chamber is

cooled with dry ice (beneath the base) and as the alcohol vapour diffuses downwards, the vapour becomes supersaturated. There should be a region in the chamber that maintains a supersaturated vapour and the trails of charged particles through this region can be observed.

Expansion cloud chamber

This was the original type of cloud chamber and was invented by C. T. R. Wilson in 1911 to detect radiation. In such a cloud chamber, a piston below the chamber was lowered to reduce the pressure and cool the gas in the chamber. This produced a supersaturated vapour. In the expansion cloud chamber made in Australia by IEC, the expansion is performed by a bicycle pump (with the washer reversed to extract air from the chamber).



Figure 26.11 A Wilson cloud chamber

Reduction of the pressure causes the vapour to become supersaturated and hopefully trails appear. A voltage applied to a metal ring inside the chamber sweeps the ions from the chamber before another expansion or pressure reduction can be performed. A disadvantage of such a chamber is that it is then necessary to wait perhaps a minute for the alcohol to evaporate into the chamber before repeating the process. The small school version can usually be used without this wait. Once the correct conditions have been attained, withdrawing the handle of the bicycle pump can usually be done every few seconds.

Part A: Observing tracks in a diffusion cloud chamber

Apparatus

cloud chamber (which probably has a built-in radioactive source)
woollen cloth
alcohol (Propan-2-ol also known as iso-propyl alcohol recommended but ethyl alcohol should work)
dry ice
light source (possibly a microscope lamp)

Method

We will assume that the cloud chamber has a built-in source.

Remove the perspex top and moisten the felt ring with a few drops of alcohol. Turn the chamber upside-down and unscrew the base and remove the foam pad. Place some small pieces of dry ice over the black metal plate. These will be held against the metal by the foam when the base is screwed back on.

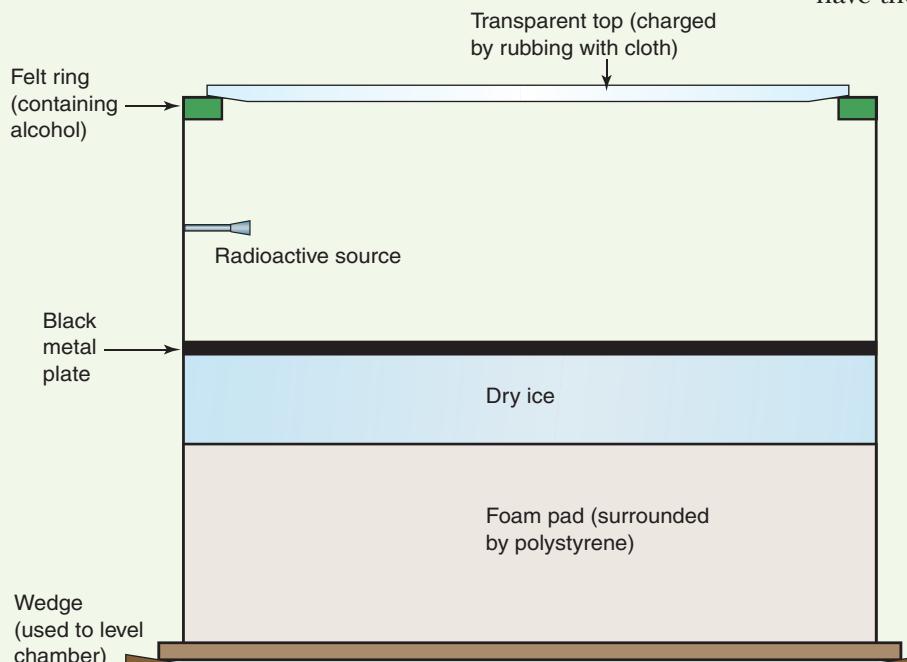


Figure 26.12 A diffusion cloud chamber

Place the chamber right side up on small wedges (probably provided) and level it as carefully as possible. If the metal plate is not horizontal, convection currents can hinder the production of tracks. Replace the perspex top and rub it gently with the woollen cloth to charge it. (Charging the top of the chamber produces an electric field that

will remove dust and ions and this should assist in the production of tracks in the chamber.)

After a few minutes, tracks should become visible. If a light is shone horizontally through the side of the chamber, the visibility of the tracks may be enhanced.

The chamber should continue to operate while the temperature difference between the top and bottom of the chamber region is maintained. If the trails become hard to see, recharging the perspex top may help.

It is sufficient to set up a cloud chamber and observe the tracks of alpha particles through the chamber.

If the cloud chamber works particularly well you may be able observe other tracks that do not come from the source. (Cloud chambers played an important role in the discovery of cosmic rays.)

Part B: Observing tracks in an expansion cloud chamber

The IEC Wilson's Expansion Cloud Chamber is described below. Usually expansion cloud chambers have the disadvantage that they provide a brief view

of the tracks when expansion occurs and then there is a delay before the expansion can be repeated. However, with this particular cloud chamber, expansions can be performed every few seconds.

Apparatus

Wilson's Expansion Cloud Chamber including modified bicycle pump, radioactive source
alcohol (propan-2-ol, also known as iso-propyl alcohol, is recommended but ethyl alcohol should work.)
light source (possibly a microscope lamp)
high-voltage DC power supply (at least 300 V)

Method

Preparing the radioactive source

The cloud chamber is supplied with a small quantity of thorium oxide. The thorium oxide can be used to prepare a point radioactive source (which can be screwed into the side of the chamber) or radon 220 gas, produced from the decay of thorium, can be used as the source. (If you search for information on the decay of thorium, you will find that radon 220 is produced part of the way along the decay series.)

Special care must be taken if it is necessary to transfer thorium oxide into the squeeze bottle that is used to inject the radon into the cloud chamber. Your teacher will probably do this for you. (The apparatus works well with about 10 grams of thorium oxide in the squeeze bottle.)

Setting up the expansion cloud chamber

Make sure that the base of the cloud chamber is level. Remove and clean the glass top. Pour a few millilitres of propan-2-ol onto the black metal disc at the bottom of the chamber (2 to 3 millilitres works well). Replace the top and gently tighten the screws to produce a seal.

Attach the hose from the squeeze bottle containing the thorium oxide to the fitting on the side of the chamber.

Connect the terminals on the side and base of the chamber to a high-voltage power supply. (The instructions say up to 600 volts can be used but excellent tracks were observed with a voltage of 200 volts. This high voltage is essential for the production of tracks.)

Set up the microscope lamp so that it shines horizontally into the chamber. (It should not be too close to the chamber as it is essential that the chamber is not heated.)

Release the Mohr clip on the hose connecting the squeeze bottle to the chamber, squeeze the bottle gently once and then replace the clip. (As the purpose of this is to inject some radon 220 gas into the chamber, it is easier to do this if the bicycle pump is disconnected which means that the chamber is not sealed.)

Reconnect the bicycle pump, and quickly and smoothly withdraw the handle of the pump.

Tracks should be visible throughout the chamber. The radon gas should have spread throughout the chamber and as it decays by alpha particle emission, the short tracks produced by alpha particles should be visible throughout the chamber.

The trails disappear quickly but repeating within a few seconds should produce another set of trails. Occasionally much longer trails, which are probably thinner than those of the alpha particles will be observed. What particles are likely to produce these trails and what might be their origin?

The number of tracks produced will gradually decrease. (Check the half-life of radon 220 and you will see why.) Eventually it will be necessary to squeeze another puff of gas into the chamber.

Further investigations

It is an achievement to set up a cloud chamber and observe the trails. If you find that you are able to get your cloud chamber working very well without any difficulty you might like to try to investigate further.

You could try different sources and compare the tracks produced by different type of radiation. (This might be impractical with a diffusion cloud chamber but might well be possible with the expansion cloud chamber.)

You could try to produce a magnetic field in the chamber and see if it is possible to deflect the particles in the magnetic field.

GLOSSARY

A

A-scan: a range-measuring system that records the time for an ultrasonic pulse to travel to an interface in the body and be reflected back

absolute magnitude, M : the magnitude that a star would have if it were viewed from a standard distance of 10 parsecs

absorption spectrum: a series of dark lines on a coloured background that is produced when white light is passed through a cool gas and viewed through a spectroscope

active optics: a slow feedback system to correct sagging or other deformities in the primary mirror of large modern reflector telescopes

acoustic impedance, Z : a measure of how readily sound will pass through a material. It is measured in $\text{kg m}^{-2} \text{s}^{-1}$.

adaptive optics: use a fast feedback system to attempt to correct for effects of atmospheric turbulence

aether: the proposed medium for light and other electromagnetic waves, before it was realised that these waveforms do not need a medium in order to travel

angular momentum, L : of a point mass, m , which is in circular motion of radius, r , with velocity, v , is given by: $L = m v r$. Angular momentum is the rotational equivalent to linear momentum and is an important quantity in rotational motion. (It follows a similar conservation principle to linear momentum.)

annual parallax, p : half the angle through which a nearby star appears to shift against the backdrop of distant stars, over a particular six-month period

apparent magnitude, m : the magnitude given to a star as viewed from Earth

armature: a frame around which a coil of wire is wound, which rotates in a motor's magnetic field

astrometry: the careful measurement of a celestial object's position, and changes of position, to a high order of accuracy

attenuation: the reduction in intensity of a signal

average binding energy per nucleon: the total binding energy of a nucleus divided by the number of nucleons in the nucleus. It is a measure of the stability of the nucleus.

B

B-scan: displays the reflected ultrasound as a spot, the brightness of which is determined by the intensity of the ultrasound

back emf: an electromagnetic force that opposes the main current flow in a circuit. When the coil of a motor rotates, a back emf is induced in the coil due to its motion in the external magnetic field.

baryons: hadrons that have half-integer spin (and are fermions). Examples are the proton and neutron. Some other baryons are included in table 26.1.

binding energy: the energy equivalent of the mass defect of the nucleus. It is the energy that would have to be provided and converted to mass to enable all the nucleons in a nucleus to be separated from each other.

black hole: the crushed remnant of the core (greater than 5 solar masses) of a very massive star. Theoretically, it is a point of zero volume and infinite density.

bosons: particles that have either integer or zero spin.

They do not obey the Pauli exclusion principle. Bosons are force-carrying particles.

brushes: conductors that make electrical contact with the moving split metal ring of a commutator

C

cathode/anode: cathode rays are now known to be streams of electrons emitted within an evacuated tube from a cathode (negative electrode) to an anode (positive electrode). They were first observed in discharge tubes.

cathode ray tube or discharge tube: a sealed glass tube from which most of the air is removed by vacuum pump. A beam of electrons travels from the cathode to the anode and can be deflected by electrical and/or magnetic fields.

centripetal acceleration: always present in uniform circular motion. It is associated with centripetal force and is also directed towards the centre of the circle.

centripetal force: the force that acts to maintain circular motion and is directed towards the centre of the circle

Chandrasekhar limit: (1.4 solar masses) the greatest mass that a non-rotating white dwarf can have

CNO (carbon–nitrogen–oxygen) cycle: the hydrogen fusion mechanism that dominates in hotter main sequence stars

coherent: when there is a constant phase difference between light waves; that is, the peaks line up and the troughs line up. Also refers to an optic fibre bundle in which the optic fibres keep the same position relative to one another.

colour index: the difference between a star's photographic magnitude, B , and its visual magnitude, V

commutator: a device for reversing the direction of a current flowing through an electric circuit, for example, the coil of a motor

contrast: refers to the brightness difference between parts of the image.

covalent bond: a strong chemical bond formed between atoms by the sharing of electrons in the valence band

critical mass: smallest amount of fuel necessary to sustain a chain reaction. As the size increases, the volume to surface area ratio increases and a smaller proportion of neutrons are lost from the fuel.

crystal: a naturally occurring solid with a regular polyhedral shape. All crystals of the same substance grow so that they have the same angles between their faces. The atoms that make up the crystal have a regular arrangement called a crystal lattice.

crystalline: a solid in which the atoms or molecules are arranged in a regular pattern

D

diffraction: refers to the spreading out of light waves around the edge of an object or when light passes through a small aperture

diffraction grating: a device consisting of a large number of slits, used to produce a spectrum

diode: a device that contains only two electrodes
distance modulus: equal to (apparent magnitude – absolute magnitude). It is directly related to the distance of a star from Earth.

dopant: a tiny amount of an impurity that is placed in an otherwise pure crystal lattice to alter its electrical properties

Doppler effect: the apparent change in frequency observed when there is relative movement between a source of a sound and an observer

drift velocity: the average velocity of electrons in a conductor under the influence of an electric field

E

eddy current: a circular or whirling current induced in a conductor that is stationary in a changing magnetic field, or that is moving through a magnetic field. They resemble the eddies or swirls left in the water after a boat has gone by.

electromagnetic induction: the generation of an emf and/or electric current through the use of a magnetic field

emission spectrum: a series of brightly coloured lines on a dark background that is produced when light from an excited gas is viewed through a spectroscope

empirical equation: one that has no theoretical basis but can be used to calculate correct values. Kepler's Third Law, $T^2 \propto R^3$, which you encountered in 'The Cosmic Engine', is another example of an empirical equation.

escape velocity: the initial velocity required by a projectile to rise vertically and just escape the gravitational field of a planet

excited state: when an electron exists in a stationary state in which it has more energy

F

fermions: particles that have half-integer spins. They obey the Pauli exclusion principle.

field vector: a single vector that describes the strength and direction of a uniform vector field. For a gravitational field, the field vector is \mathbf{g} .

fissile: a nucleus that may undergo fission

fluorescence: the emission of light from a material when it is exposed to streams of particles or external radiation

flux: from the Latin word *fluo* meaning 'flow'. Flux is a state of flowing or movement. In physics, flux is the rate of flow of a fluid, radiation or particles.

G

galvanometer: an instrument for detecting small electrical currents

geostationary orbit: an altitude at which the period of the orbit precisely matches that of the Earth. This corresponds to an altitude of approximately 35 800 km.

gradient magnetic field: a magnetic field that changes by small known increments throughout the region of the field

gravitational field: a field within which any mass will experience a gravitational force. The field has both strength and direction.

gravitational potential energy: E_p , the energy of a mass due to its position within a gravitational field. On a large scale, gravitational potential energy is defined as the work done to move an object from infinity (or some point very far away) to a point within a gravitational field.

ground state: the state an electron is in when it has the lowest possible amount of energy

H

hadrons: particles that experience the strong nuclear force. Mesons and baryons are both hadrons.

half-life: the time taken for half the radioactive nuclei in a sample to decay. If we exclude the activity of daughter nuclei, it is the time taken for the activity of a particular sample to drop to half its initial value.

hard X-rays: consist of high-energy photons and are more penetrating than soft X-rays, which have lower energy photons

heliosphere: the zone around the solar system dominated by the Sun's magnetic field and solar wind. It is bound by the heliopause, approximately 100 AU from the Sun.

helium flash: the sudden onset of helium fusion in the core of a new red giant

I

incandescent: bright or glowing. Like black bodies, most substances become incandescent when they become hot enough.

induction: a process where one object with magnetic or electrical properties can produce the same properties in another object without making physical contact

induction motor: an AC machine in which torque is produced by the interaction of a rotating magnetic field produced by the stator and currents induced in the rotor

inertial frame of reference: a non-accelerated environment. Only steady motion or no motion is allowed. A non-inertial frame of reference experiences acceleration.

integrated circuit (IC): an electronic circuit in which all the components, such as transistors, diodes, resistors, capacitors and connections, are made in or on a single piece of semiconductor, such as a silicon chip

interference: the interaction of two or more waves — producing regions of maximum amplitude (constructive interference) and zero amplitude (destructive interference). The Michelson–Morley experiment used the interference of light in an attempt to measure the movement of the Earth through the aether.

interferometry: a technique used to combine the data from several elements of an antenna array in order to achieve a higher resolution

interstellar dust: made of grains of silicates and ices in a core and mantle structure, just one micrometre across

interstellar gas: occurs as regions of neutral atoms, ions or molecules. It is mostly hydrogen.

interstellar medium: consists of gas and dust

ionisation blackout: a period of no communication with a spacecraft due to a surrounding layer of ionised atoms forming in the heat of re-entry

isotope: a nuclide that has the same number of protons but different numbers of neutrons compared to another nuclide of the same element

L

Larmor frequency: the frequency with which a nucleus precesses about its spin axis, in response to the force due to an external magnetic field

length contraction: the shortening of an object in the direction of its motion as observed from a reference frame in relative motion

leptons: particles which do not experience the strong nuclear force. They are all fermions with half-integer spin. An electron is a lepton.

light-year: the distance travelled through space in one year by light or other electro magnetic wave. It corresponds to a distance of 0.3066 parsecs or 9.4605×10^{12} km.

low Earth orbit: an orbit higher than 250 km and lower than 1000 km

M

magnetic flux, Φ_B : the amount of magnetic field passing through a given area. In the SI system, Φ_B is measured in weber (Wb).

magnetic flux density: the strength of a magnetic field, B . In the SI system, B is measured in tesla (T) or weber per square metre (Wb m^{-2}).

magnetosphere: the region around a planet in which the planet's magnetic field exerts an influence

main sequence star: characterised by the fusion of hydrogen to helium in its core

mass defect: the difference between the mass of the constituent nucleons of a nucleus and the mass of the nucleus

mass dilation: the increase in the mass of an object as observed from a reference frame in relative motion

maxima: refers to points on an interference pattern where the peaks of each set of waves coincide. This produces a bright spot when light is used and is a point of constructive interference.

medium: the material through which a wave travels

mesons: hadrons that have zero or integer spins. Some of the mesons with zero spin are included in table 26.2.

metastable: a nucleus in an excited state for a period of time before decaying

MeV: a million electron volts — the energy gained by one electron accelerating through a potential difference of one million volts

minima: refers to points on an interference pattern where peaks of one wave coincide with troughs of the other. This produces a dark spot and is a point of destructive interference.

motor effect: the action of a force experienced by a current-carrying conductor in an external magnetic field

N

net spin: a property of a nucleus. If a nucleus has a net spin it behaves as a tiny magnet.

neutron star: the extremely dense remnant of the core (1.4 to 3 solar masses) of a massive star. It is composed of neutron matter.

non-coherent: refers to an optic fibre bundle in which the fibres are not kept in the same position relative to one another

nuclide: refers to a particular nucleus with certain values of Z (atomic number) and A (mass number)

O

optical fibre: a glass core surrounded by a cladding of lower refractive index. Light is transferred along the optical fibre by total internal reflection.

P

parallax: the apparent shift in position of a close object against a distant background due to a change in position of the observer

parsec: a parallax-second — the distance that corresponds to an annual parallax of 1 second of arc

period, T: the time taken to complete one orbit

phase scan: a scan produced using an array of transducers. The phase difference between the signals from each transducer may be varied to produce this scan.

phosphorescent: a substance that absorbs radiation of one wavelength and then emits radiation of a different wavelength over a period of time. The hands of some analogue watches are coated with a phosphorescent substance to enable them to be seen in the dark.

photocell: a device that uses the photoelectric effect.

These devices include photovoltaic cells and solar cells which convert electromagnetic energy, such as sunlight, into electrical energy.

photoconductive cell: or photo-resistor, uses the fact that electrical resistance is affected by light falling onto it

photoelectric effect: the name given to the release of electrons from a metal surface exposed to electromagnetic radiation. For example, when a clean surface of sodium metal is exposed to ultraviolet light, electrons are liberated from the surface.

photometry: the measurement of the brightness of a source of light or other radiation

photon: a quantum (or discrete packet) of electromagnetic radiation. It can be thought of as an elementary particle with zero rest mass and charge, travelling at the speed of light.

piezoelectric effect: the conversion of electrical energy to mechanical energy resulting in the change in shape of a piezoelectric crystal when it is subjected to a potential difference

planetary nebula: a shell-shaped cloud of gas that is the blown-away outer layers of a star

positrons: positively charged beta particles formed when a proton disintegrates to form a neutron and a positron. A positron is identical to an electron except that its charge is positive instead of negative.

precession: the movement, in a conical path, of the axis of a spinning object

principal quantum number: the value of n for each stationary state or orbit of the Bohr atom

projectile: any object launched into the air

proton-proton (p-p) chain: the hydrogen fusion mechanism that is first to occur in main sequence stars

protostar: a new star before it begins to produce any nuclear energy in its core

Q

quantum (plural: quanta): can be considered to be the smallest amount of energy possible in a given situation. Planck's atomic oscillators could oscillate only with certain precise amounts of energy.

quantum mechanics: the name given to a set of physical laws that apply to objects the size of atoms or smaller. The concepts of wave–particle duality and uncertainty lie at the heart of quantum mechanics.

quantum theory: based on quantity or amounts (from the Latin word *quantum* meaning 'how much'). In 'classical physics' an object could possess any amount of energy. In quantum theory objects could possess only certain discrete amounts of energy. Instead of being 'continuous', energy is available only in 'packets'.

R

radioactive decay: the emission of particles from the nucleus of a radioactive element

radioactive isotope or **radioisotope:** an isotope that is unstable and will emit particles from the nucleus until it becomes stable

radiopharmaceutical: a compound that has been labelled with a radioisotope

radio telescope: a large dish or array aimed at the sky that detects radio waves arriving from space. The signal is fed to computers that are able to compile the information into an image.

red giant: a star characterised by a helium-burning core surrounded by a hydrogen-burning shell

relaxation: refers to precessing nuclei moving back to their original energy state

resolution: the ability to distinguish closely spaced points as separate points. The resolution limit is the smallest separation of points that can be distinguished as distinct.

resonate: to absorb energy when an applied frequency matches the natural frequency of an object

rest energy: the energy equivalent of a stationary object's mass, measured within the object's rest frame

rest frame: the frame of reference within which a measured event occurs or a measured object lies at rest

right-hand grip rule: used to find the direction of a magnetic field around a straight current-carrying conductor.

right-hand push rule: (also called the right-hand palm rule) used to find the direction of the force acting on a moving charged particle or current-carrying conductor in an external magnetic field

rotor: the rotating part of an electrical rotating machine

S

scintillation: a flash of light observed on a scintillation screen. Another example of scintillation is electrons striking the screen of a cathode ray oscilloscope. The screen produces many scintillations when it is struck by electrons. Of course, the continuous beam of electrons

produces a constant glow, not individual flashes as would be observed when alpha particles hit such a screen.

sector scan: a scan in the shape of a sector, made from a series of B-scans

seeing: refers to the twinkling and blurring of a star's light due to atmospheric distortion

semiconductor: a material in which resistance decreases as it rises. Its resistivity lies between that of a conductor and an insulator.

slingshot effect: or planetary swing-by, is a manoeuvre used with space probes to pick up speed and proceed on to another target

slip speed: the difference between the speed of the rotating magnetic field and the speed of the rotor

soft X-rays: X-rays consisting of low-energy photons

solenoid: consists of a coil of wire wound uniformly into a cylinder

space–time: a single four-dimensional concept that considers space and time as being bound together

spectroscope: a device used to spread a light into its spectrum. It can be attached to the eyepiece of a telescope to examine the spectra of starlight.

spectroscopic parallax: a method of using the H–R diagram and the distance modulus formula to determine the approximate distance of a star

speed of light: $3.0 \times 10^8 \text{ m s}^{-1}$, or approximately 173 million km h⁻¹. It is the theoretical maximum velocity in our universe.

split metal ring: the two-piece conducting metal surface of a commutator. Each part is connected to the coil.

squirrel-cage rotor: an assembly of parallel conductors and short-circuiting end rings in the shape of a cylindrical squirrel cage

stationary state: the state an electron is in when it orbits the nucleus without emitting any electromagnetic radiation

stator: the non-rotating magnetic part of a motor

stellar spectroscopy: the examination of the spectra of stars in order to learn more about their composition, surface temperature, velocity, density, etc.

step-down transformer: provides an output voltage that is less than the input voltage

step-up transformer: provides an output voltage that is greater than the input voltage

stroboscope: a light that produces quick flashes at regular (usually small) time periods

supernova: a violent explosion of uncontrolled nuclear reactions that completely blows away the various layers of a massive star (original mass greater than five solar masses)

T

terminal: the free end of a cell or battery to which a connection is made to the rest of a circuit

theoretical resolution: a telescope's ability to distinguish two close objects as separate images. It is measured as an angle.

thrust: the force delivered to a rocket by its engines

time dilation: the slowing down of events as observed from a reference frame in relative motion

torque: the turning effect of a force. It is the product of the tangential component of the force and the distance the force is applied from the axis of rotation.

trajectory: the path that a projectile follows during its flight
transfer orbit: an orbit used to manoeuvre a satellite from one orbit to another

transformer: a magnetic circuit with two multi-turn coils wound onto a common core

transistor: a tiny switch that changes the size or direction of electric current as a result of very small changes in the voltage across it. Transistors are used in sound amplifiers and in a wide range of electronic devices. Today, a single chip of silicon can hold many microscopic transistors and is called an integrated circuit.

transmutation: when a radioactive atom emits an alpha particle or a beta particle and an atom of a new element is produced. A new daughter element is formed from a parent element.

trigonometric parallax: a method of using trigonometry to solve the triangle formed by parallax to determine distance

triple alpha reaction: the process of helium fusion in the core of a red giant

U

uniform circular motion: circular motion with a uniform orbital speed

universal motor: a series-wound motor that may be operated on either AC or DC electricity

ultrasound: very high frequency sound. Ultrasound waves are sound waves that have a frequency above the range of human hearing, that is, greater than 20 000 hertz.

ultrasound transducer: a device for converting electrical energy to ultrasound energy or for converting ultrasound energy to electrical energy

V

valence band: the energy band in a solid in which the outermost electrons are found

valve: a thermionic device in which two or more electrodes are enclosed in a glass tube. The name comes from the rectifying property of the device; that is, the current flows in only one direction.

vector: any quantity that has both magnitude and direction. Force is one example.

visual magnitude: refers to magnitude as judged by eye, or more accurately by a photometer fitted with a yellow-green filter

voltage: the electrical pressure between two points that is capable of producing a flow of current between the points when they are connected by a closed circuit

voxel: a small volume, part of a ‘slice’ through the body

W

wavefront: either the crest or trough of a wave. The wavefront is perpendicular to the direction of the velocity of the wave.

weight: the force on a mass due to the gravitational field of a large celestial body, such as the Earth

white dwarf: a dense star made of degenerate matter. It is the end point of small- to medium-sized stars.

work function: the energy required to release the electron from the surface of a particular material

X

X-rays: electromagnetic waves of very high frequency and very short wavelength

Z

zero-age main sequence (ZAMS): a plot of the main sequence using only zero-age stars

APPENDIX 1: Formulae and data sheet

DATA SHEET

Numerical values of several constants

Charge on the electron, q_e	-1.602×10^{-19} C
Mass of electron, m_e	9.109×10^{-31} kg
Mass of neutron, m_n	1.675×10^{-27} kg
Mass of proton, m_p	1.673×10^{-27} kg
Speed of sound in air	340 m s ⁻¹
Earth's gravitational acceleration, g	9.8 m s ⁻²
Speed of light, c	3.00×10^8 m s ⁻¹
Magnetic force constant $\left(k \equiv \frac{\mu_0}{2\pi} \right)$	2.0×10^{-7} N A ⁻²
Universal gravitational constant, G	6.67×10^{-11} N m ² kg ⁻²
Mass of Earth	6.0×10^{24} kg
Planck's constant, h	6.626×10^{-34} J s
Rydberg's constant, R _H (hydrogen)	1.097×10^7 m ⁻¹
Atomic mass unit, u	1.661×10^{-27} kg $931.5 \frac{\text{MeV}}{c^2}$
1 eV	1.602×10^{-19} J
Density of water, ρ	1.00×10^3 kg m ⁻³
Specific heat capacity of water	4.18×10^3 J kg ⁻¹ K ⁻¹

FORMULAE SHEET

PRELIMINARY COURSE

The world communicates

$$v = f\lambda$$

$$I \propto \frac{1}{d^2}$$

$$\frac{v_1}{v_2} = \frac{\sin i}{\sin r}$$

Electrical energy in the home

$$E = \frac{F}{q}$$

$$R = \frac{V}{I}$$

$$P = VI$$

$$\text{Energy} = VIt$$

Moving about

$$v_{\text{av}} = \frac{\Delta r}{\Delta t}$$

$$a_{\text{av}} = \frac{\Delta v}{\Delta t} = \frac{v - u}{t}$$

$$\Sigma F = ma$$

$$F = \frac{mv^2}{r}$$

$$E_k = \frac{1}{2}mv^2$$

$$W = Fs$$

$$p = mv$$

$$\text{Impulse} = Ft$$

The cosmic engine

$$\text{Brightness} = \frac{\text{luminosity}}{4\pi r^2}$$

$$\lambda_{\text{max}} T = W$$

$$v = H_0 D$$

HSC COURSE

Space

$$E_p = -G \frac{m_1 m_2}{r}$$

$$\mathbf{F} = mg$$

$$v_x^2 = u_x^2$$

$$v = u + at$$

$$v_y^2 = u_y^2 + 2a_y \Delta y$$

$$\Delta x = u_x t$$

$$\Delta y = u_y t + \frac{1}{2} a_y t^2$$

$$\frac{r^3}{T^2} = \frac{GM}{4\pi^2}$$

$$F = \frac{G m_1 m_2}{d^2}$$

$$E = mc^2$$

$$L_v = L_0 \sqrt{1 - \frac{v^2}{c^2}}$$

$$t_v = \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$m_v = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Motors and generators

$$\frac{F}{l} = k \frac{I_1 I_2}{d}$$

$$\mathbf{F} = BIl \sin \theta$$

$$\tau = Fd$$

$$\tau = nBIA \cos \theta$$

$$\frac{V_p}{V_s} = \frac{n_p}{n_s}$$

From ideas to implementation

$$\mathbf{F} = qv\mathbf{B} \sin \theta$$

$$E = \frac{V}{d}$$

$$E = hf$$

$$c = f\lambda$$

Astrophysics

$$d = \frac{1}{p}$$

$$M = m - 5 \log\left(\frac{d}{10}\right)$$

$$\frac{I_A}{I_B} = 100^{\frac{(m_B - m_A)}{5}}$$

$$m_1 + m_2 = \frac{4\pi^2 r^3}{G T^2}$$

Medical physics

$$Z = \rho v$$

$$\frac{I_r}{I_0} = \frac{[Z_2 - Z_1]^2}{[Z_2 + Z_1]^2}$$

From quanta to quarks

$$\frac{1}{\lambda} = R_H \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$$

$$\lambda = \frac{h}{mv}$$

The age of silicon

$$A_0 = \frac{V_{\text{out}}}{V_{\text{in}}}$$

$$\frac{V_{\text{out}}}{V_{\text{in}}} = -\frac{R_f}{R_i}$$

APPENDIX 2: Periodic table

* Lanthanide series

- For elements with no stable nuclides, the mass of the longest living isotope is given in square brackets.
- The atomic weights of Np and Tc are given for the isotones ^{237}Np and ^{99}Tc .

APPENDIX 3: Key words for examination questions

HSC syllabus documents and examination questions use the following key words that state what students are expected to be able to do.

Account	Account for: state reasons for, report on. Give an account of: narrate a series of events or transactions
Analyse	Identify components and the relationship between them; draw out and relate implications
Apply	Use, utilise, employ in a particular situation
Appreciate	Make a judgement about the value of
Assess	Make a judgement of value, quality, outcomes, results or size
Calculate	Ascertain/determine from given facts, figures or information
Clarify	Make clear or plain
Classify	Arrange or include in classes/categories
Compare	Show how things are similar or different
Construct	Make; build; put together items or arguments
Contrast	Show how things are different or opposite
Critically (analyse/ evaluate)	Add a degree or level of accuracy, depth, knowledge and understanding, logic, questioning, reflection and quality to (analysis/evaluation)
Deduce	Draw conclusions
Define	State meaning and identify essential qualities
Demonstrate	Show by example
Describe	Provide characteristics and features
Discuss	Identify issues and provide points for and/or against
Distinguish	Recognise or note/indicate as being distinct or different from; note differences between
Evaluate	Make a judgement based on criteria; determine the value of
Examine	Inquire into
Explain	Relate cause and effect; make the relationships between things evident; provide why and/or how
Extract	Choose relevant and/or appropriate details
Extrapolate	Infer from what is known
Identify	Recognise and name
Interpret	Draw meaning from

Investigate	Plan, inquire into and draw conclusions about
Justify	Support an argument or conclusion
Outline	Sketch in general terms; indicate the main features of
Predict	Suggest what may happen based on available information
Propose	Put forward (for example a point of view, idea, argument, suggestion) for consideration or action
Recall	Present remembered ideas, facts or experiences
Recommend	Provide reasons in favour
Recount	Retell a series of events
Summarise	Express, concisely, the relevant details
Synthesise	Put together various elements to make a whole

© Board of Studies NSW, 2003

ANSWERS TO NUMERICAL QUESTIONS

CHAPTER 1

g on surface (m s⁻²)	Weight of 80 kg person there (N)
3.7	296
8.9	712
1.8	143
1.3	101

5. (a) 0.124
 (b) 0.515
 (c) 0.904
 (d) 0.466
 11. $-8.59 \times 10^9 \text{ J}$
 12. (a) $-7.4 \times 10^{30} \text{ J}$
 (b) $-3.24 \times 10^{35} \text{ J}$

CHAPTER 2

8. 3.14 m
 9. (a) 2.39 m
 (b) 6.14 m
 10. (a) 6.2 s
 (b) 108.5 m
 11. Yes
 12. (a) 56 500 m
 (b) 57 400 m
 (c) 56 500 m
 13. 3390 m
 14. 115 000 m
 17. Mercury: 4250 m s^{-1} ; Venus: $10\ 400 \text{ m s}^{-1}$;
 Io: 2550 m s^{-1} ; Callisto: 2470 m s^{-1}
 18. (a) $a = 60 \text{ m s}^{-2}$, $g = 7.1$
 (b) $a = 69.6 \text{ m s}^{-2}$, $g = 8.1$
 19. (a) $a = 2.7 \text{ m s}^{-2}$, $g = 1.3$
 (b) $a = 83 \text{ m s}^{-2}$, $g = 8.5$

CHAPTER 3

2. $F = 31 \text{ N}$, $a = 78 \text{ m s}^{-2}$
 3. 19 400 N
 5. (a) $28\ 400 \text{ km h}^{-1}$
 (b) 85 min
 (c) 2.44 N
 6. (a) $28\ 050 \text{ km h}^{-1}$
 (b) 88.1 min
 (c) 9.25 m s^{-2} towards Earth's centre
 (d) 1020 000 N
 7. Mercury: .244 Earth years; Venus: .619 Earth years; Mars: 1.89 Earth years; Jupiter: 11.9 Earth years; Saturn: 29.4 Earth years
 10. Low Earth: 360.0, 1.53, 7686
 Geostationary: 35 800, 23.93, 3070

CHAPTER 4

4. (a) $3.59 \times 10^{21} \text{ N}$
 (b) $4.17 \times 10^{23} \text{ N}$
 5. $1.48 \times 10^{-10} \text{ N}$
 6. (a) 710 N
 (b) 650 N
 7. (a) Satellite:
 orbital velocity, 7721 m s^{-1}
 centripetal force, $1.21 \times 10^4 \text{ N}$
 gravitational force, $1.21 \times 10^4 \text{ N}$
 (b) Venus:
 orbital velocity, $3.52 \times 10^4 \text{ m s}^{-1}$
 centripetal force, $5.6 \times 10^{22} \text{ N}$
 gravitational force, $5.5 \times 10^{22} \text{ N}$
 (c) Callisto:
 orbital velocity, 8186 m s^{-1}
 centripetal force, $3.9 \times 10^{21} \text{ N}$
 gravitational force, $3.9 \times 10^{21} \text{ N}$

CHAPTER 5

Star	Distance (light-years)	Distance (parsecs)	Distance (km)
Canopus	75	23	7.1×10^{14}
Rigel	900	276	8.5×10^{15}
Arcturus	32.6	10	3.1×10^{14}
Hadar	3.26	100	3.1×10^{15}

15. (a) 0.745c (b) 53.4 m
 16. 3479.99999998 km
 17. 0.99c
 18. 0.99999994c
 19. Pluto: 15 min; Proxima Centauri: 69 days;
 Sirius: 141 days; Alpha Crucis: 23.36 years;
 Andromeda: 100 717 years
 20. (a) 0.866c (b) $1 \times 10^5 \text{ kg}$
 (c) 2 s
 21. (a) 28.6 m
 (b) 28 min 59 s
 (c) $3.14 \times 10^5 \text{ kg}$
 25. (a) $1.506 \times 10^{-10} \text{ J}$
 (b) $5.9819 \times 10^{-10} \text{ J}$
 (c) $1.7939 \times 10^{-9} \text{ J}$
 (d) $4.5 \times 10^{11} \text{ J}$
 (e) $9 \times 10^{16} \text{ J}$

CHAPTER 6

11. (a) 6.8×10^{-3} N, down the page
(b) 1.5×10^{-4} N, out of the page
12. (a) 12.5 T
13. (a) 6.0×10^{-2} N
14. 1.8×10^{-2} N
15. (a) 1.2×10^{-5} N
16. (a) 1.3×10^{-5} N
(b) 3.2×10^{-6} N
(c) 5.7×10^{-7} N
19. (b) 0.34 N
(c) 2.7×10^{-2} m²
(d) 3.6×10^{-2} N m
20. (a) 0.98 N, upwards

CHAPTER 7

3. (a) 3.0 Wb
(b) 2.3×10^{-2} Wb
(c) 6.0×10^{-6} Wb
(d) 0
11. (a) 1.4×10^{-3} Wb
12. (a) 6.3×10^{-3} Wb
(b) Would be 25 times greater
15. (a) 48 A
(b) 0.6 A
16. (a) 24 A
(b) 220 V

CHAPTER 8

11. (b) 64
12. 16 V
13. (a) 2.0 V -4; 6.0 V -12
(b) 2.5 A
14. (a) 400 V
(b) 200 W
(c) 200 W
(d) 10 A
15. (b) 3 (increase)
16. (a) 26
(b) 26 A
17. 0.020
18. (a) 1.3 V
(c) 0.8 A
20. (a) 0.24 A
(b) 0.096 V
(c) 500 kV
(d) 2.3×10^{-2} W
21. (a) 7.0×10^{-2}
(b) 220 MW
(c) 6.7×10^2 A
22. (a) 5.0 A
(b) 400 W
(c) 3.9×10^3 V

CHAPTER 10

1. 2.4×10^{-13} N
2. 8.6×10^{-16} N
3. 1.7×10^{-13} N
4. 1.9×10^{17} m s⁻¹
5. (a) 4.0×10^3 V m⁻¹ left
(b) 6.4×10^{-16} N right
(c) 6.4×10^{-16} N left
(e) 3.2×10^{-17} J each
8. 2.0×10^{-5} N
11. 4.8×10^{-19} C
12. (a) 2.00×10^2 V m⁻²
(b) 0.40 N
13. 0.488 N
21. (a) 3.00×10^4 V m⁻¹
(b) 3.00×10^6 m s⁻¹
(c) 4.8×10^{15} N

CHAPTER 11

1. (a) 6.0×10^{14} Hz
(b) 4.0×10^{-19} J
(c) 2.5×10^{14}
2. (a) 5.6 V
(b) 3.7×10^{-18} J
(c) 5.6×10^{15} Hz
6. (c) 4.2×10^{14} Hz
(d) 6.6×10^{-34} J s
(e) 2.8×10^{-19} J
9. 3.07×10^{-19} J
10. 55
11. 1.33 $\left(\frac{\text{number of red photons per second}}{\text{number of blue photons per second}} \right)$
12. (a) 2.6×10^{-19} J
(b) 2.5 V
(c) 6.9×10^{-19} J
14. (a) 2.86 m
(b) 6.95×10^{-26} J
17. 8.0×10^{-15} J

CHAPTER 12

1. 1.5×10^{-9} m

CHAPTER 13

5. (a) 0.014 W
(b) 3.0×10^{-5} V
6. 0.15 A

CHAPTER 14

9. (a) 2.1 arcsec
(b) 2.1 arcsec
(c) 1.1 arcsec
(d) 0.53 arcsec
(e) 0.035 arcsec
(f) 0.013 arcsec

10. (a) 0.53 arcsec
 (b) 0.63 arcsec
 (c) 0.74 arcsec
11. (a) 1.3×10^5 arcsec
 (b) 630 arcsec
 (c) 210 arcsec
 (d) 90 arcsec
 (e) 32 arcsec
 (f) 6.3 arcsec
 (g) 0.42 arcsec
12. (a) 4.2 arcsec
 (b) 4200 arcsec
 (c) 4.2×10^6 arcsec
13. (a) $m = 40 \times$ magnification, $R = 0.46$ arcsec
 (b) $m = 100 \times$ magnification, $R = 0.46$ arcsec
14. (a) $1.3 \times 10^4 \text{ m}^2$
 (b) 0.6 arcsec
15. (a) 1128 m
 (b) 0.02 arcsec
16. (a) 0.007 arcsec
 (b) 0.002 arcsec

CHAPTER 15

1.

	km	AU	l-y	pc
1 km =	1	6.685×10^{-9}	1.057×10^{-13}	3.2408×10^{-14}
1 AU =	1.49×10^8	1	1.5813×10^{-5}	4.848×10^{-6}
1 light year =	9.4605×10^{12}	6.324×104	1	0.3066
1 parsec =	3.086×10^{12}	206 265	3.2616	1

4. (a) 44.05 pc
 (b) 98.04 pc
 (c) 19.96 pc
 (d) 28.49 pc
 (e) 5.144 pc
 (f) 190 pc (to 2 significant figures)
 (g) 11.2 pc
 (h) 1.82 pc
 (i) 160 pc
 (j) 130 pc
5. (a) 0.0633 arcsec
 (b) 0.097 arcsec
 (c) 0.00422 arcsec

- (d) 0.38 arcsec
 (e) 0.0134 arcsec
 (f) 0.00763 arcsec
 (g) 0.0104 arcsec
 (h) 0.0775 arcsec
 (i) 0.001 arcsec
 (j) 0.13 arcsec
6. (a) 51.5 light-years
 (b) 33.6 light-years
 (c) 773 light-years
 (d) 8.5 light-years
 (e) 243 light-years
 (f) 427.6 light-years
 (g) 313 light years
 (h) 42.1 light-years
 (i) 3200 light-years
 (j) 25 light-years

13. (a) 3000 K
 (b) 8000 K
 (c) 6000 K
14. (a) 7.25×10^3 K
 (b) 9.37×10^{26} W
26. (a) ≈ 4.2
 (b) ≈ 2.8
 (c) ≈ 3700
 (d) ≈ 2
 (e) ≈ 65
 (f) ≈ 56
27. $\approx 2.5 \times 10^{28}$
30. (a) ≈ 96
 (b) ≈ 98 pc

31.

Star	<i>m</i>	<i>M</i>	<i>d</i>
Rigel	0.18	-6.69	237
Bellatrix	1.64	-2.72	74.5
Capella	0.07	-0.48	12.9
Sirius	-1.44	1.45	2.64
Deneb	1.25	-8.73	991
Altair	0.75	2.2	5.14
Achernar	0.45	-2.77	44.1
Spica	0.98	-3.55	81

32.

Star	Parallax (mas)	Distance (pc)	m	M
Fomalhaut	130.08	7.69	1.17	1.74
Vega	128.93	7.76	0.03	0.58
Canopus	10.43	95.9	-0.62	-5.5
Betelgeuse	7.63	131	0.45	-5.1
Rigel Kent	742.12	1.35	-0.01	4.3

35. Fomalhaut: 10 pc; Vega: 8 pc

36. 75 pc

41. Based on the colour index, Aldebaran is a red star of spectral class K with a surface temperature of approximately 3500 K.

42. Based on the colour index, Spica is a blue-white star of spectral class B with a surface temperature of approximately 15 000 K.

CHAPTER 16

5.

Total mass of system (kg)	Total mass of system (solar masses)
4.87×10^{30}	2.45
1.16×10^{31}	5.81
3.54×10^{30}	1.78
5.15×10^{31}	25.9
9.17×10^{31}	46.1
6.80×10^{30}	3.42
1.61×10^{31}	8.10
2.53×10^{31}	12.7
3.78×10^{30}	1.90
4.08×10^{30}	2.05

6. (a) 1.05×10^{32} kg
 (b) 7.16×10^9 m

CHAPTER 18

5. 1.71×10^6 kg m $^{-2}$ s $^{-1}$
 6. 1.01×10^3 kg m $^{-3}$ (1.01 g cm $^{-3}$)

7. 330 m s $^{-1}$ 8. (a) (i) 1.63×10^6 kg m $^{-2}$ s $^{-1}$ (ii) 6.53×10^6 kg m $^{-2}$ s $^{-1}$

(b) 3:2

9. 0.000319

10. (d) 1.74 mW cm $^{-2}$ (e) 79.12 mW cm $^{-2}$ 14. 0.16 mW cm $^{-2}$ 15. (b) 1.56×10^6 kg m $^{-2}$ s $^{-1}$ (c) 1300 m s $^{-1}$

18. (c) 18 cm

19. (a) 4.5×10^{-4} s**CHAPTER 20**

6. (a) 4.0 minutes

- (b) 8.0 minutes

CHAPTER 21

4. (c) 1.004 T

CHAPTER 22

1. (a)
- 9.496×10^{-8}
- m

- (b)
- 4.341×10^{-7}
- m

- (c)
- 1.282×10^{-6}
- m

2. (a)
- 3.889×10^{-7}
- m,
- 3.798×10^{-7}
- m,
-
- 3.751×10^{-7}
- m

3. (a)
- 2.1×10^{-10}
- m

- (b)
- 4.8×10^{-10}
- m

- (c)
- 8.5×10^{-10}
- m

5. 10

6. (a)
- 1.22×10^{-7}
- m,
- 1.03×10^{-7}
- m

- (b)
- 6.57×10^{-7}
- m,
- 4.87×10^{-7}
- m

- (c)
- 1.88×10^{-6}
- m,
- 1.28×10^{-6}
- m

7. (a)
- 7.65×10^{-6}
- m

- (b)
- 2.30×10^{-6}
- m

8. (a)
- ∞

- (b)
- 9.12×10^{-8}
- m,
- 3.65×10^{-7}
- m,
- 8.22×10^{-7}
- m

- (c) 13.6 eV

9. (a) (i) 1.8 eV
(ii) 4.35×10^{14} Hz, 6.89×10^{-7} m
(b) 1.41×10^{-7} m
16. (b) 7.3×10^{14} Hz

CHAPTER 23

7. (a) 2.43×10^{-11} m
(b) 2.73×10^{-22} m s⁻¹
8. (a) 2.86×10^{-14} m
(b) 2.02×10^{-13} m

CHAPTER 24

11. 4.95 MeV
12. 17.3 MeV
13. (a) 1.19 MeV absorbed
14. 8.6 MeV
15. 25.7 MeV

CHAPTER 25

12. 3.27 MeV
14. (a) 511 keV

CHAPTER 26

1. (a) 40 times or 20 orbits
(b) 3.9×10^6 m s⁻¹
(c) 3.2 cm

INDEX

- A-scans (ultrasound) 348–9
absolute magnitude 292
absorption spectra 284–5, 303, 425, 426
AC electric motors
 energy transformations and transfers 169–70
 induction motors 165–9
 main features 164
 universal motor 164–5
AC electricity
 household use 155
 versus DC 147–8
AC generators 142–3
AC induction motors 165–9, 172
 operation 168–9
 power 169
 slip speed 169
 squirrel-cage rotor 167–8
 stator of three-phase 166–7
 structure 166–8
acceleration, rocket lift-off 24–7
acceleration due to gravity 3–4, 6, 65
 pendulum determination 11
 variations 4–5
 weight values in the solar system 12
acceleration equations 15–16
acoustic impedance 344–5
active optics 267
adaptive optics 267–9
aether model 72–4, 75
agricultural uses of radioisotopes 491–2
air resistance 22–3
alpha particles 382
 deflection by magnetic field 455
 penetrating power 454
 properties 383, 456
alpha particles scattering experiments 458
 Geiger and Marsden 420–1, 422
 Rutherford and Bequerel 419–20
Anglo-Australian Telescope 259, 264, 268
angular momentum 428, 465
annual parallax 276, 277
 precision 302–3
anode 175
antineutrino 463, 464
antiprotons 502
apparent magnitude 291
armature 5
artificially induced radioactivity 458, 459–60
artificially induced transmutations 458
astrometric binaries 310
astrometric satellites 278
astrometry 275–8
atomic bomb development 480–4
atomic masses, light nuclides 470
atomic models
 Bohr's 423–7, 427–8, 430–1, 432, 434
 quantum theory steps 444–51
 Rutherford's 419–23, 429–30
attenuation of a signal 367
Australian Telescope Compact Array 272–3
average binding energy per nucleon 469

B-scans (ultrasound) 349
back emf in motors, and Lenz's Law 129–30

Balmer's equation 424
band structures 231
 doping, and 219–20
 semiconductors, in 216–19
 solids, in 213
baryons 504, 505, 508
Becquerel, Henri 419, 420, 454
Becquerel's predicament 420
beta decay
 Fermi explanation 462–3
 problems of 461–2
beta particles 383
 deflection by a magnetic field 455
 distribution of energy 462
 penetrating power 454
 properties 383, 456
Bethe, Hans 484
binary stars 306
 astrometric binaries 310
 eclipsing binaries 308–9, 318–19
 mass-luminosity relationship 311
 spectroscopic binaries 309–10, 319
 visual binaries 306–8
binding energy 469–70
black body radiation 199–201, 281–2
black hole 334
blood flow measurement by ultrasound 352–5
Bohr, Niels 449
 periodic table explanation 447–8
 principle of complementarity 448, 451
 views on atomic bomb 483
Bohr equation 432
Bohr's model of the atom 283, 423–7, 433
 de Broglie explanation of Bohr's electron orbits 446–7
 energies of 'stationary states' 431–2
 limitations 434
 mathematics of 429–34
 postulates 427–8
 quantum theory to explain hydrogen spectrum 424
 radii of 'stationary states', hydrogen atom 430
 'stationary states' of electrons 428
bone density and ultrasound 351–2
bone imaging 390
Born, Max 446, 448, 449
bosons 504, 509–10, 512
Bragg's experiment 238–9
Bragg's Law 238, 239
Bragg's X-ray diffraction studies 237–8
brain, imaging studies 390–1, 410, 411
breathalyzers 208
Bremsstrahlung radiation 366
brightness
 measurement 289
 stars 290–1
brightness ratios, stars 290–1
brushes 111
BSC theory 243
bubble chambers 497

carbon–nitrogen–oxygen (CNO) cycle 326–7
cathode 175
cathode ray oscilloscope (CRO) 187–8
cathode ray tubes 175, 176
 component parts 186

- cathode rays
 applications 186–8
 charge-to-mass ratio 183
 discovery 175–6
 electric field effects 177–82
 magnetic field effects 182
 Thomson's experiments 180, 183, 184–5
 waves or particles 184–5
- causality, principle of 85
- centripetal acceleration 40
- centripetal force 39, 41
- Chadwick, James 460–1
 identification of neutron 460–1
- Chandra X-Ray Observatory 268, 269
- charge-to-mass ratio of cathode rays 183
- Chernobyl nuclear accident 488–9
- classical physics 202
 photoelectric effect 204–5
- cloud chambers 496–7, 518–20
- CNO cycle 326–7
- coherent circular waves 233
- coherent light 443
- coherent optic fibre bundle 374, 375
- coiled conductor
 induced currents in 126, 137
 using a moving magnet in 125
- colliders 502, 512
- colour filters 303–4
- colour index, stars 297
- colour magnitudes, stars 296
- colour measurement, stars 295
- colour television 186–7
- commutators 109, 111
- Compton Gamma Ray Observatory 268, 269
- computed axial tomography *see* CT scans
- conductors 213–16, 239
 resistivity 217
- continuous spectra 280–2, 303, 425
- contrast (image) 409
- Coolidge X-ray tube 235
- Cooper pairs 244–5
- covalent bonding 218
- critical mass 482
- Crookes, William 184
- crystal lattice structure of metals 239–40
- crystalline substances 347
- crystals, X-ray diffraction 236, 236–8
- CT scans 368–73
 diagnostic tool, as 372–3
 production 369–71
- Curie, Irène 459
- current-carrying conductor *see also* parallel current-carrying conductor
 magnetic field 103–4, 120, 131, 137–8
 magnitude of the force on 104–5
 right-hand push rule 104
- cyclotrons 385, 499–500
- Davission, Clinton 446
- DC electric motors 109–14
 anatomy 109–10
 calculating torque of a coil 113–14
 changing speed 112
 commutators 111
 magnetic field 112
 model 121
 operation 110–11
- DC electricity, versus AC 147–8
- DC generators 144–5
- de Broglie, Louis
 explanation for Bohr's electron orbits 446–7
 matter waves 444–6
 wave model of electrons 215–16, 441
- de Broglie wavelength 444–6
- de Forest, Lee 221
- deflecting plates 186
- density, stars 289
- de-orbiting 50
 de-orbit manoeuvre 51
- DEXA (Dual Energy X-ray Absorptiometry) 352
- diffraction 235, 441, 445
 electrons 446
 explanation 442–3
 X-rays 235–8
- diffraction grating 235, 236, 443
- diffusion cloud chamber 518
- diodes 220, 221, 222
- Dirac equation 451
- discharge tubes 175, 176, 191
 everyday uses 176
- distance modulus 292–3
- dopant 217
- doping, and band structures 219–20
- Doppler effect 288, 352–3
- Doppler ultrasound
 blood flow measurement 352–5
 choosing the best signal 354–5
 practice, in 353–4
- Earth's gravitational field 3–12
 review 10
- Earth's rotational motion 31–2
- eclipsing binaries 308–9, 318–19
- eddy currents
 heat losses in transformers 151
 magnetic fields, and 131–2
 switching devices, in 132
- Edison, Thomas 147–8, 221
- Eightfold Way 504
- Einstein, Albert 75–6, 85, 91, 205, 206, 423, 424, 441
 photoelectric equation 206
 theory of relativity 76
- electric chair 147–8
- electric field strength 178
- electric fields, effect on cathode rays 177–82
- electric motors, DC 109–14, 121
- electric power generating stations 146
- electrical resistance
 low temperature effects 241–2
 superconductors, in 246, 254
- electricity
 AC/DC 147–8, 155
 society, and 156
- electricity production, nuclear fission reactor 487
- electromagnetic braking 132
- electromagnetic force, unification of 506
- electromagnetic induction 123, 126–7
- electromagnetic levitation 249
- electromagnetic spectrum 195
 atmospheric absorption 258–60
 components 258
- electromagnetic waves 185, 236
 Maxwell's theory 194–5
- electromagnets 103, 140, 403
- electron gun 186

electronics, superconductor applications 247–8
electrons 184, 419, 503 *see also* cathode rays
charge 181
de Broglie wave model 215–16, 441
diffraction 446
excited state 433
ground state 433
magnetic field effects 182
positron interactions 393
protons in close proximity, and 461
Rutherford atomic model, in 423
spin 465
stationary states 428
superconducting state, in 243
electrostatic forces, nucleons 467
elementary charge 181
elements, naturally occurring 489
elliptical orbits 45–7
emission spectra 282–4, 303, 425, 426
empirical equation 424
endoscopes
medical diagnosis, in 373–7
operation 375
structure 374–5
usage 376–7
energy, and mass 88–9
energy bands 213, 214, 224
energy transformations and transfers 169–70
escape velocity 23–4
exclusion principle (Pauli) 450, 504
expansion cloud chamber 518
extrinsic semiconductor materials 217
extrinsic semiconductors 219–20
extrinsic variables 312

Faraday, Michael 123
electromagnetic induction 123, 126–7
first experiments 123–4
iron ring experiment 124–5
motor effect 103–4
using a moving magnet 125–6
Faraday's Law of Induction 127, 149
fault current limiter (FCL) 247
Fermi, Enrico 504
explanation of beta decay 462–3
neutron bombardment of uranium 476, 477–8
fermions 504, 509
field vector g 3–4
fissile nucleus 484
fixed target accelerators 502
fluorescence 175
frequency 341
ultrasound 342–3
Frisch, Otto 478–9, 480

g forces 27–30
decelerating 53–4
variations during rocket launch 30–1

Galileo's telescopes 257
galvanometer 114, 123
gamma camera 388–9
gamma radiation 383, 386–7, 456
deflection by magnetic field 455
ionising power 455
penetrating power 454–5
gases, spectra 425–6
Geiger counter 421

Geiger, Hans 420–1, 422
Gell-Mann, Murray 504, 505, 507
generators 140–5
AC 142–3
current direction 143–4
DC 144–5
hand-operated, output 160
magnetic flux and emf variation 141–2
power stations 146
geostationary orbit 47
geosynchronous orbit 46
germanium, for semiconductors 218–19
Germer, Lester 446
gluons 509–10
gradient magnetic field 406
gravitational attraction, and satellite motion 62–4
gravitational collapse 322–4
gravitational field vector g 3–4
variations
altitude, with 4–5
geographical location, with 4
planetary body, with 4
gravitational fields 3–5, 65–6
weight and 6
gravitational forces, nucleons 467
gravitational potential energy 7–9

hadrons 504, 508
Hahn, Otto 478
hair dryer 170
half-life 383–4, 477
Hallwachs, Willhelm 203
hard X-rays 366
heart muscle, imaging studies 389–90
heavy elements synthesis, stars 332
Heisenberg, Werner 448, 451
uncertainty principle 450
work on German atomic bomb project 483
helium flash 328
Henry, Joseph 123
Hertz's experiments with radio waves 196–8
Hertzsprung–Russell diagrams 287, 324, 328, 330, 335
Higgs boson 512
HIPPARCOS Catalogue 278, 302
Hounsfield, Godfrey N. 369
Hubble Deep Field 256
Hubble Space Telescope (HST) 260, 268
Huygens' Principle 442
hydrogen atom 422
'classical' energy 429–30
energies of 'stationary states' 431–2
quantum mechanics perspective 450
radii of 'stationary states' 430
spectral lines explanation 432–4
hydrogen fusion mechanisms (main sequence stars) 324–5
carbon–nitrogen–oxygen cycle 326–7
proton–proton chain 325
hydrogen protons
external magnetic field, in 400, 403
Larmor frequency 405
hydrogen spectrum 424, 437–9
quantum ideas 424–5
theoretical expression for wavelengths 432–3

incandescent light 280
induced currents
coiled conductor, in 126, 137

- direction 138
 linked coils 137–8
 induction 126
 induction heating 133
 industrial uses of radioisotopes 491
 inertial frames of reference 74–5
 insulating transmission lines 155
 insulators 213–16
 integrated circuits (ICs) 225, 227–9
 interference 233–5, 442–3
 interferometry 265–6
 interstellar dust 321, 322
 interstellar gas 321
 interstellar medium 321–4
 intrinsic semiconductor materials 217
 intrinsic semiconductors 219
 intrinsic variables 313
 iodine-123 386, 389
 iodine-131 386
 ionisation blackout 50
 ionising power, radiation 455
 isotopes 382, 457 *see also* radioisotopes
- Joliot, Frédéric 459
 Josephson junction 246
- kaons 504
 Keck telescopes 269
 Kelvin scale of temperature 241
 Kepler's Law of Periods 41–2, 43, 46, 60, 62, 307, 309
 constant derivation 61
 Kunzman, Charles 446
- Langrangian point 49
 large-scale integrated circuits (LSI) 227
 Larmor frequency 404–5
 lattice structures 218
 doping effects 219–20
 metals 239–40
 Law of Conservation of Momentum 25, 68
 Law of Universal Gravitation 3–4, 42, 61–5, 70
 Lenard, Philipp von 203–4
 length, relativity of 81–4
 length contraction 84
 lenses
 light-gathering ability 272
 magnification, and 262–3
 Lenz's Law 128–30, 144
 Principle of Conservation of Energy, and 129
 production of back emf in motors, and 129–30
- leptons 504, 506, 507–8
 lift-off (rockets) 24–32
 light transmission by optical fibres 373, 380
 lightning protection 154, 179
 linear accelerators 499
 Los Alamos Laboratory 482–3
 loudspeakers 115
 low altitude polar orbit 49
 Low Earth orbit 49
 luminosity classes, stars 287
 lungs, imaging studies 390, 391
- maglev trains 248–9
 magnetic field lines 101–3
 magnetic fields
 cathode ray effects 182
 charged particles in 102, 131
- current-carrying conductor 103–5, 120, 131, 137–8
 current-carrying solenoid, around 102
 DC electric motors 112
 direction around a solenoid 103
 eddy currents, and 131–2
 effect on orientation of nuclei 403, 404–5
 hydrogen protons, and 400, 403
 radioactive emissions deflection by 455
 review 101–3
 rotating coils in 128
 superconductors, and 245
 magnetic flux 126–7
 variation in generator coil 141–2
 magnetic flux density 126
 magnetic resonance imaging *see* MRI
 magnitude of stars 290–2
 main sequence stars 324
 carbon–nitrogen–oxygen (CNO) cycle 326–7
 hydrogen ‘burning’ 324–5
 proton–proton chain 325
 transition to red giants 329–30
- Manhattan Project 480, 481–4
 first nuclear reactor 481–2
 physicists' views 483–4
 research at Los Alamos 482–3
- Marconi's radio wave experiments 198
 Marsden, Ernest 420–1, 496
- mass
 energy, and 88–9
 relativity of 85–9
- mass defect 468–71
 mass dilation 87
 mass energy 89
 mass-luminosity relationship 311
 matter waves (de Broglie) 444–6
 confirmation of 446
- maxima 234
 Maxwell's theory of electromagnetic waves 194–5
 medical cyclotron 385
 medical diagnosis
 CT scan use 368–73
 endoscopy use 373–7
 MRI use 248, 399–500, 410–11
 PET scans 392–4, 395, 490
 radioisotope use 382, 384, 386, 387–91 489–90
 SQUID (Superconducting Quantum Interference Device) 248
 superconducting magnets use 248
 ultrasound use 342–6, 351–2
 X-ray use 366–8
- medical imaging
 combined techniques 394–5
 comparison of techniques 412–14
- medium 68
 Meissner effect 245, 253–4
 Meitner, Lise 478–9
 mesons 504, 505, 508
 metal lattice 214
 metals
 crystal lattice structure 239–40
 superconductors 242
- metastable nucleus 383
 metre, definition 77
 MeV 393, 459
 Michelson–Morley experiment 72–4, 233
 modelling 96–7
- Millikan's oil drop experiment 181–2

- minima 234
 motor effect 103–5, 120
 MRI
 image and the patient 399–400
 magnets in the body 400
 medical uses 410–11
 MRI machine
 application of radio frequency pulses 405–6
 contrast in images 409, 410
 distinguishing one type of hydrogen compound from another 408–9
 effect on atoms in the patient 402–10
 effect on nuclei orientation in strong magnetic field 403–5
 precession 404–5
 relaxation time, measuring 409–10
 removal of radio frequency pulses 407–9
 muons 503
- n-type semiconductors 219–20, 222
 naming stars 311
 NASA's 'Great Observatories Program' 268–9
 naturally occurring elements 489
 naturally occurring radioactivity 456–7
 net spin of a nucleus 400, 401
 neutrinos 503
 detection 463–4, 466
 discovery of 461–6
 interaction with matter 464
 properties 464–5
 recent discoveries 465–6
 types of 508
 neutron scattering 492
 neutron stars 334
 neutrons
 discovery 458–61
 in a nuclear reactor 484–6
 slow and fast 478
 Newton's Law of Universal Gravitation 3–4, 42, 61–5
 Newton's Second Law of Motion 3, 6, 26, 40
 Newton's Third Law of Motion 25
 non-coherent optic fibre bundles 374, 375
 non-inertial frames of reference 71, 97–8
 non-periodic variables 313
 npn transistors 225
 NSW electrical distribution system 153–4
 nuclear atom (Rutherford model) 422
 nuclear equations 457
 nuclear fission
 discovery 476–80
 first observations 479
 Meitner and Frisch experiments 478–9
 nuclear fission reactor 484–9
 Chernobyl accident 488–9
 control rods 486
 coolant 486
 electricity production 487
 moderators 486
 neutrons in 484–6
 radioactive waste products 488
 nuclear medicine 382, 385, 386, 387–91, 490
 nuclear physics, timeline 513–15
 nuclear power station 487
 nuclear reactions, energy change 471
 nuclear reactor, first 481–2
 nucleons
 gravitational and electrostatic forces 467
 strong nuclear force 466–8, 468
- nucleus
 binding energy 469–70
 energy from 475–6
 Larmor frequency 404–5
 mass defect 468–71
 net spin 400, 401
 precession 404, 405
 nuclides 457
 atomic masses 470
- Oliphant, Sir Mark 475–6, 480
 optical fibres
 endoscopes, in 374, 375
 light transmission 373, 380
 optics
 active 267
 adaptive 267–9
 orbit
 elliptical 45–7
 types of 47–9
 orbital decay 49–50
 orbital energy 44–5
 orbital motion 39–50
 orbital velocity 42–4, 70
- p–n junction 222, 223–4
 p-type semiconductors 219–20, 222
 parallactic ellipse 277–8
 parallax 275
 annual 276, 277, 302–3
 spectroscopic 293–5
 trigonometric 275
 parallel current-carrying conductor
 forces between 105–7, 120–1
 magnitude of the force 106–7
 parallel plates, electric field between 177–9
 parsec (parallax-second) 276–8
 particle accelerators 250, 499–502, 512–13
 particle detectors 496–8
 modern detectors 497–8
 particle masses 503
 particle physics
 and cosmology 513
 new particles 503
 Standard Model 504–6
 timeline 513–15
 Pauli, Wolfgang
 exclusion principle 450, 504
 prediction of neutrino 462, 503
 quantum mechanics to hydrogen, application of 450
 Peierls, Rudolf 480
 pendulum, to determine g 11
 penetrating power, radiation 454–5
 period-luminosity relationship 315
 periodic table, Bohr's explanation 447–8
 periodic variables 313–14
 periods, Law of 41–2
 permanent magnets 103, 140, 403
 PET scans 392, 395, 490
 isotopes used 394
 operation 393–4
 phase difference 351
 phase scans (ultrasound) 350–1
 phosphorescent substances 454
 photocells 207
 photoconductive cells 208
 photocopier machine 180
 photoelectric effect 197, 202–8

- applications 207–8
 explanation 204–5
 photoelectric equation 206
 photoelectric photometry 298
 photographic photometry 298
 photometry 289–98
 photons 201, 203–4, 283, 424
 phototubes 208
 photovoltaic cells 207–8, 228–9
 photovoltaic effect 207
 piezoelectric effect 347
 pions 503
 Planck, Max 200–1, 206, 283, 423, 424
 Planck's constant 201
 Planck's equation 201
 planetary swing-by 66–9
 plutonium bomb 482
 pnp transistors 225
 Pogson scale 290
 pointed conductors 179
 positron emission tomography *see* PET scans
 positrons 393, 503
 post-helium burning 331
 potential difference 127–8
 moving charge through 178
 power 149
 AC induction motor 169
 power distribution 151–4
 NSW 153–4
 transformers to reduce power loss 152–3
 power generation, superconductor use 246–7
 power station generators 146
 power storage, superconductor use 247
 power transmission lines *see* transmission lines
 precession 404, 405
 principle of complementarity 448, 451
 Principle of Conservation of Energy 169
 Lenz's Law, and 129
 transformers, and 149–51
 principle of relativity 74–5
 projectile motion 14–23
 modelling 35
 projectiles 14
 acceleration equations 15–16
 air resistance 22–3
 combined vertical and horizontal motions 19
 horizontal motion 17–19
 maximum height 21
 range 22
 trajectory 15
 trip time 22
 velocity 20–1
 vertical motion 16–17
 proton–proton (p–p) chain 325
 protons 503
 antiprotons, and 502
 electrons in close proximity, and 461
 energy levels 402
 external magnetic fields, in 400
 resonation 405
 protostar 323
 pulsars 334
 quanta 423
 quantum chromodynamics (QCD) 510
 quantum electrodynamics (QED) 510
 quantum mechanics 445, 448–9
 development 449–51
- quantum physics, timeline 513–15
 quantum theory 201, 202, 419, 423–4, 448–9
 hydrogen spectrum 424–5
 model of the atom 444–51
 quarks 505–6, 507
 colour properties 509
 discovery of top quark 510–
- radiant energy 289
 radiation
 properties 383, 454–6
 types of 382–3
 radio aerials, operation 199
 radio frequency pulses
 application 405–6
 removal 407–9
 radio telescopes 260, 264, 265
 radio waves
 carrier waves and superimposed signal 221
 frequencies and 198
 Hertz's experiments with 196–8
 Marconi's experiments 198
 producing and transmitting 211
 radioactive decay 382, 383–4, 457–8
 radioactive waste products 488
 radioactivity
 artificially induced 458, 459–60
 detection 456
 early investigations 454
 naturally occurring 456–7
 safety issues 392
 radioisotopes
 advantages/disadvantages 395
 body organs, targeting 387–9
 emitting gamma radiation 386–7
 half-life 383–4
 industrial and agricultural applications 491–2
 medical diagnosis 382, 384, 386, 387–91, 395, 489–90
 metabolising by the body 385
 PET scans 392–4, 395, 490
 production 385
 properties 382, 491
 radiopharmaceuticals 385, 390
 red giants 327–31
 main sequence transition to, evidence 329–30
 post-helium burning 331
 triple alpha reaction 331
 re-entry (spacecraft) 50–3
 decelerating *g* forces 53–4
 extreme heat 51–3
 ionisation blackout 54
 reaching the surface 54–5
 reflecting telescopes 261–2
 reflection of ultrasound 345–6
 refracting telescopes 261
 relativistic space flight 89–91
 relativity *see also* special relativity
 length, of 81–4
 mass, of 85–9
 principle of 74–5
 simultaneity, of 77–8
 theory of 72, 81
 time, of 78–81
 resistance *see* electrical resistance
 resonance (protons) 405
 rest energy 89
 rest frame 79

- right-hand grip rule 102, 103, 128, 131
 right-hand push rule 104, 131, 143
 rocket science pioneers 32
 rockets
 - Apollo 10 launch 36–7
 - Earth's motion, effect on launch 31–2
 - g* forces 27–30
 - lift-off 24–7
 - thrust and acceleration 26–7
 - variations in acceleration and *g* 30–1
 Röentgen, Wilhelm 185, 235, 454
 rotational velocity, stars 288–9
 rotor 140
 Rutherford, Ernest 185, 454, 496
 - alpha particle scattering experiments 419–21
 - artificially induced transmutation 458
 - energy from the nucleus 475, 476
 - nuclear atom 421–2
 - prediction of the neutron 458, 460
 Rutherford model of the atom 419–23
 - 'classical' energy of hydrogen atom 429–30
 - electrons in 423
 - mathematics of 429–34
 satellite motion, and gravitational attraction 62–4
 satellites
 - orbital decay 49–50
 - orbital velocity 42–4
 - periods of 41–2
 - types of orbits 47–9
 S-Cam, the 264, 280
 Schrödinger, Erwin, wave function theory 448, 450
 scintillations 421
 sector scans (ultrasound) 350–1
 seeing 261, 278
 semiconductors 213–16
 - applications 228–9
 - band structures 213
 - doping and band structure 219–20
 - making 218–19
 - resistivity 217–18
 Shockley, William 225, 231
 silicon
 - doping effect on lattice structure 219–20
 - lattice structure 218
 - semiconductors, for 218–19
 silicon chips 227–8
 simple harmonic motion 11
 simultaneity, relativity of 77–8
 singularity 334
 slingshot effect 66–9, 70
 slip speed 169
 sodium chloride 215, 236
 - crystal structure 237
 soft X-rays 366
 solar cells 207–8, 228–9
 solar system weight values and *g* 12
 solenoid
 - determining poles of 103
 - magnetic field around 102
 solid state devices 222–3, 227–8
 - versus thermionic devices 224–5
 sound waves 341–2
 space exploration 32
 space shuttle
 - engines 26
 - re-entry 51
 space–time continuum 76
 spacecraft
 - re-entry 50–5
 - slingshot effect 66–9, 70
 special relativity
 - consequences 77–92
 - constant speed of light 75–6
 - inertial frames of reference 74–5
 - space–time continuum 76
 spectra
 - absorption 284–5, 303, 425, 426
 - continuous 280–2, 303, 425
 - emission 282–4, 303, 425, 426
 - gases, of 425–6
 - making 279–80
 - observing with spectroscope 427
 spectral analysis, starlight 285–6
 spectral classes, stars 286
 spectrophotometer 280
 spectroscope 279, 427
 spectroscopic binaries 309–10, 319
 spectroscopic parallax 293–5
 spectroscopy 279–89
 - speed of light 75–6, 195
 - faster than 85
 - spin, electrons 465
 - Spitzer Space Telescope 268, 322
 - split-metal ring 111
 - split-ring commutator 111
 - Square Kilometre Array (SKA) 266
 - SQUID (Superconducting Quantum Interference Device) 248
 - squirrel-cage rotor 167–8
 Standard Model (particle physics) 503
 - boson force-carriers 509–10
 - developments leading to 504–8
 - particles 506–11
 - today and beyond 511–13
 Stanford Linear Accelerator Center (SLAC) 505
 star birth
 - gravitational collapse 322–4
 - interstellar medium 321–4
 star death 332–5
 star life
 - main sequence, after 327–31
 - main sequence stars 324–7
 starlight, spectral analysis 285–6
 stars
 - absolute magnitude 292
 - absorption spectra 285
 - apparent magnitude 291
 - binary 306–11
 - class L stars 286
 - colour index 297
 - colour magnitudes 296–7
 - colour measurement 295
 - data 302
 - density 289
 - distance modulus 292–3
 - evolutionary tracks 335
 - heavy elements synthesis 332
 - luminosity classes 287
 - magnitudes 290–92
 - mass–luminosity relationship 311
 - measuring brightness and luminosity 289
 - naming 311
 - period–luminosity relationship 315

- rotational velocity 288–9
 spectral classes 286
 spectroscopic parallax 293–5
 temperature 288
 translational velocity 288
 variable 312–15
 stars of five solar masses or less 333
 stars of more than five solar masses 333–4
 stationary states
 electrons 428
 energies, Bohr hydrogen atom 431–2
 radii, Bohr hydrogen atom 430
 stator 109, 140
 three-phase induction motor 166–7
 Stefan's Law 282
 stellar birth 321–4
 stellar object research 338
 stellar spectroscopy 286
 step-down transformer 149
 step-up transformer 149
 stroboscope 15
 strong magnetic fields 403–4
 strong nuclear force 466–8
 gluons and 509–10
 properties 467–8
 Sudbury Neutrino Observatory 466
 Super-Kamiokande neutrino detector 466
 superconducting magnets 404
 superconductivity 240–42
 applications 246–50
 BCS theory 243–4
 explanation 243–50
 timeline 250
 superconductors
 critical temperatures 242
 levitation and Meissner effect 245, 253–4
 magnetic field effects 245
 resistance, and 246, 254
 temperature changes 244, 253
 tunnelling effect 246
 superluminal velocities 85
 supernovae 334, 464
 supersaturated vapour 496
 switching devices, eddy currents in 132
 synchrotrons 500–1

 technetium-99m 386–7, 389, 390, 391
 telescopes 259–60
 advanced telescope technology 268–9
 Galileo's 257
 improving performance 265–9
 performance 262–4
 reflecting 261–2
 refracting 261
 theoretical resolution 263–4
 television 186–7
 temperature, stars 288
 terminals 142
 thallium-201 389–90
 theoretical resolution of telescopes 263–4
 theory of relativity 72, 85
 thermionic devices 220–1
 versus solid state devices 224–5
 Thomson, J. J. 180, 183, 184–5, 203
 ‘plum pudding’ model of the atom 419
 three-phase power generation 146
 thrust 24, 26
 thyroid investigations 389

 time, relativity of 78–81
 time dilation 79–81
 torque 107–8
 coil in DC motor, calculating 113–14
 total internal reflection 373
 transformers 148–51
 AC input and output voltage 161
 eddy current heat losses 151
 household use 155
 Principle of Conservation of Energy, and 149–51
 reducing transmission line power loss 152–3
 simple 160–1
 transistors 225–6, 231
 translational velocity, stars 288
 transmission lines
 insulating 155
 power losses 152–3, 162
 protection from lightning 154
 superconductor use 246
 transmutations 456–7
 artificially induced 458
 transuranic elements 476–7
 trigonometric parallax 275
 triodes 221
 triple alpha reaction 331
 twins paradox 91–2

 UVB system 296–7
 ultrasound 341–57
 advantages/disadvantages 357
 blood flow, measurement of 352–5
 bone density, and 351–2
 comparison with X-rays and CAT scans for diagnosis 372–3
 detecting structure inside body 343–6
 history of use 348
 medical diagnosis, and 342–6, 357
 piezoelectric effect 347
 reflection 344, 345–6
 transmission 344
 type of sound 341–2
 ultrasound scans 348–9
 A-scans 348–9
 B-scans 349
 medical uses 356
 phase scans 350–1
 sector scans 350–1
 ultrasound transducer 346, 350
 ultraviolet catastrophe 200
 uncertainty principle (Heisenberg) 450
 uniform circular motion 39–41, 56, 58–9
 uniform electric fields 177–9
 universal motor 164–5
 uranium bomb 482

 valence bands 214
 valves 220
 vapour, supersaturated 496
 variable stars 312
 variables
 Cepheids 315
 extrinsic variables 312
 intrinsic variables 313
 non-periodic variables 313
 period-luminosity relationship 315
 periodic variables 313–14
 RR Lyrae 315

- vector 61
vector field 3
velocity
 projectiles 20–1
 superluminal 85
Very Large Array (VLA) 265–6
visual binaries 306–8
visual magnitude 296
Von Laue's diffraction experiment 236–7
voxels 407
- wave equation 341–2
wavefront 442
wavelength 341
weak nuclear force 506
weight 6
Westinghouse, George 147–8
Wien's Law 281
Wilson cloud chamber 518
work function 205
- X-radiation
 effect on the body 362–3, 365
 frequency 366
X-ray diffraction 235–8
 Bragg's experiment 238–9
X-rays 236
 comparison to CT scans and ultrasound for
 diagnosis 372–3
 CT scans, use in 368–71
 definition 362
 diagnostic tool, as 366–8
 discovery and application 185
 imaging parts of the body 367–8, 390
 production 363
 types of 365–6
 use and detection 363–5
- Young's 'double slit' experiment 234, 235, 442, 443
zero-age main sequence (ZAMS) 323