R. Brown

Electricity & Atomic Physics

Electricity and Atomic Physics

Electricity and Atomic Physics

R. Brown MA(Cantab)
The Leys School
Cambridge

MACMILLAN

© R. Brown 1973

All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, without permission.

First published 1973

Published by
MACMILLAN EDUCATION LIMITED
Basingstoke and London
The Macmillan Company of Australia Pty Ltd
The Macmillan Company of Canadà Ltd Toronto
St Martin's Press Inc New York
Companies and representatives
throughout the world

Printed in Great Britain by
R. & R. CLARK LTD, EDINBURGH

Preface

This book is primarily intended for use in the sixth form, and its contents are very largely the material upon which sixth form examinations, whether at Advanced or at Scholarship level, are based.

Every teacher will be well aware that the Advanced level syllabus in Electricity and Modern Physics is very extensive, but that there is no need to include the whole of it in a successful sixth form course; the wide choice of questions provided in the examination papers recognises this. The textbook, however, is a different proposition, since it must include all that is relevant, and this inevitably makes it rather long. Especially is this so in Modern Physics, where the examination syllabus tends to include isolated topics, which cannot be treated in similar isolation in a book. With regard to the presentation of Electricity, I have tried to make selection relatively easy by the arrangement of chapters, which is to some extent unconventional, while Modern Physics has been put into a sequence of chapters, some of which, such as chapters 20, 25, and 26, could be omitted without altering the effect of the others.

The mathematical treatment of electrostatics and of electromagnetism are tempting subjects for the teacher, and much explored by the examiners. They are not, however, quite the same attractive propositions to the average sixth form newcomer, who tends to regard such matters as electric circuits, alternating currents, and transistors as being more realistic. Furthermore, electric and magnetic field theory leads up logically to the subject of electromagnetic radiation, and yet Maxwell's equations do not feature in school physics at all. Consequently, I have postponed this mathematical treatment until chapter 13 and onwards, taking a largely qualitative view of magnetism and electrostatics before that, and have first concentrated on the electric circuit, including the behaviour of capacitors and inductors and the treatment of alternating current. I have also placed chapters on semiconductors and on electronics earlier than they normally appear, partly in order to stress the practical importance of the capacitor and to remove any possible illusion that alternating current is important only when its frequency is 50 Hz, and partly because the differences between metals, insulators, and semiconductors might as well come early. It should be possible, however, to alter the sequence without too much difficulty; chapters 9–12 could, for example, be postponed until after chapter 19.

Only in a few instances does the text extend beyond the combined Advanced level syllabus of the different examining bodies, and usually only briefly. Such extra topics are:

5.18	Fuel cells
11.16	Thermoelectric effects in semiconductors
19.17	Superconductivity
20	Electromagnetic radiation, and relativity
23.8 and 9	Energy bands in solids

25.9–16 Nuclear reactors
26 Cosmic radiation and accelerators

I have omitted reference to the tangent magnetometer through lack of space, although I still find it useful in the laboratory. The vibration magnetometer, on the other hand, has been included, on the grounds that its application still extends outside the school laboratory, and that its basic action does not depend on the presence of the earth's magnetic field. Thermionic valves have been allotted a secondary place in favour of the junction diode and the transistor, and for this reason they have been treated rather superficially. Such devices as the klystron and the waveguide, although used in many schools, have been omitted because they are not examined, and to include them properly would be a lengthy process. There are certain noticeable omissions, too, in atomic physics, such as the determination of excitation and ionization potentials, and a treatment of orbitals and subshells, but these appear fully in the chemistry books. Thomson's experiments with positive rays have been left out, because the Bainbridge mass spectrograph would seem to be a more than adequate substitute. The last chapter, on cosmic rays and accelerators, has been included largely to call attention to the existence of this important field in physics. It could obviously be omitted by a pressed examinee. Perhaps the cyclotron should not have been left until this chapter, but this is where it logically belongs.

Electric and magnetic fields have been treated exclusively in terms of E and B, since it does not seem to me to be possible to define D and H, inside a school syllabus, in any way that can make them seem useful. H has in fact been referred to once (section 16.5) with reference to hysteresis curves, because of its conventional use when these are plotted.

I have more than willingly adopted the electromagnetic form of magnetic moment, as defined by Sommerfeld, and recommended in the A.S.E. report on SI units. In extending its application to the permanent magnet, however, and still more in including a brief treatment of the magnetic pole, I have deliberately gone beyond the limit implied in this report, having an urge to defend our recent past; when the magnetic pole enjoyed pre-eminence in examination papers until but a short time ago, it seems to me almost unethical to ban it completely now from a textbook. While endeavouring in all other respects to adopt the A.S.E. recommendations, I have not contrived to suppress the electronvolt; it does not seem to me feasible to do so at the present time because of its widespread use, and I have used this unit freely in the last three chapters of the book. The kilowatt hour makes an occasional appearance in problems, as it seems likely that the student will continue to meet this unit elsewhere; the name, Wb/m^2, has occasionally been allowed to remain in a problem, for the same reason, as alternative to the tesla.

Questions I am grateful to the examining bodies listed below, who have permitted me to reprint questions from past examination papers. In doing this, they have allowed me to extract problem parts only where I wished to do so, thus avoiding the tendency to repetition in demanding descriptive answers. It will be realised, therefore, that the majority of the questions, as reproduced here, represent extracts only.

I have not hesitated to include many problems set either by the Oxford and Cambridge Colleges or in the scholarship papers at Advanced level. The most important reason for this is their great value as a teaching weapon for those students who are prepared to wrestle with them; while not involving any appreciable increase of the syllabus, they make its application more instructive and more interesting. These questions are marked with an asterisk, and usually appear at the end of their section.

In setting out the questions I have tried to arrange them in topics, approximately

in the order followed by the text, and to place them in a sequence of increasing difficulty. I hope that this will aid the teacher in the rather tedious labour of selection, and that he will be relieved of the need to search interminably for acceptable questions elsewhere.

As far as answers are concerned, a rough policy has been to provide answers to alternate questions. However, this has not been followed rigorously: answers have not been given to the easier questions, or to those questions that would thereby be spoiled.

The sources of questions have been indicated as follows:

Oxford University Entrance Scholarships	OXF. SCH.
Cambridge University Entrance Scholarships	CAM. SCH.
Oxford and Cambridge Schools Examination Board	O & C
Oxford Delegacy of Local Examinations	O
University of Cambridge Local Examinations Syndicate	C
University of London School Examinations Department	L
Joint Matriculation Board	N
Southern Universities Joint Board	S
Associated Examining Board	A.E.B.

I am of course indebted to numerous individuals, not least to my colleagues at the Leys School, and I would like, in particular, to thank Mr. G. R. Noakes and Dr. J. W. Warren for their very thorough appraisals of earlier editions of the script, and for the very essential criticisms and suggestions which they made.

R. B.

The cover illustration shows a nuclear reactor's fuel element cooling pond at the Atomic Energy Research Establishment, Harwell, Berkshire. It is reproduced by courtesy of the United Kingdom Atomic Energy Authority.

Contents

1	Introduction	11
2	Potential and Energy	20
3	Resistance	33
4	The Potentiometer and the Bridge	77
5	Electrolysis and Cells	96
6	Basic Electrostatics	109
7	Basic Magnetism	142
8	Capacitors and Inductors in a Circuit	164
9	Alternating Current	209
10	AC Demonstrations – the Cathode Ray Oscilloscope	240
11	Insulators and Semiconductors. Thermocouples	253
12	Diodes and Triodes	269
13	Electric Field Equations	299
14	Magnetic Flux	329
15	Magnetizing Coils and Magnets	358
16	Magnetic Materials	383
17	The Absolute Determination of Current and Resistance	400
18	Current Measuring Instruments	410
19	Applications of Electromagnetism	423
20	Electromagnetic Radiation and the Special Theory of Relativity	444
21	Conduction by Ionized Gases. The Electron	458
22	The Quantum Theory. Photocells and the X-ray Tube	472
23	Electrons in the Atom	488
24	Radioactivity	506
25	Nuclear Reactions	538
26	Cosmic Radiation and Particle Accelerators	554
	Answers	566
	Index	571

Chapter 1

Introduction

1.1 Electric charge

If two ebonite rods are stroked on flannel, or on one's hair, and one of them is placed in a horizontal stirrup hanging from a thread, the second rod, brought near, repels the first. This force acts even when the two rods are some distance apart; it would still act if they were in a vacuum. If an ebonite rod is stroked, not on flannel, but on polythene sheet, its subsequent behaviour is different. Two rods so treated repel each other, just as two rods stroked on flannel do, but a rod stroked on polythene attracts one stroked on flannel.

An introduction to electricity almost inevitably begins with an account of this kind, because it describes all that we can really define electricity to be. We have no explanation why these forces act, and so we provide some names, and leave the matter at that. The ebonite rod stroked on flannel has, we say, negative electric charge; the rod stroked on polythene has positive charge. Like electric charges repel each other; unlike charges attract each other. These forces act at a distance, which we interpret by saying that an electric charge is surrounded by an electric field; when a second charge is placed in this electric field, it experiences a force. The name, 'electricity', is derived from the Greek word for amber, because amber was among the first substances to be used in simple experiments of this kind. The justification for labelling the charges positive and negative is the fact that, brought together, the effect of two equal but unlike charges is to cancel each other out, and in many cases the charge seems to have vanished permanently, as if the two had destroyed each other. Which type of charge is labelled positive, and which negative, was in the first place immaterial, but the choice has long since been made, and we are now consistent in accepting it.

There is no special magic, however, about materials such as ebonite, polythene, perspex; these are chosen for simple demonstrations in electrostatics merely because they happen to charge up readily and to be good insulators. Any material can, in fact, be charged up, provided that allowance is made for the possibility that it may be a conductor, and consequently prone to lose its charge again. Moreover it can be made negative or positive, as ebonite can, dependent on what is used to charge it; some materials, it is true, have a tendency to one particular sign, and it is difficult, for example, to charge polythene positively.

The fact is that every material, whether charged or not, already contains vast amounts of electricity; if it is electrically neutral, this is merely because the charge is there in exactly equal amounts of positive and negative. It is electrical force which holds the atoms together in a solid, and every atom itself consists of a positively charged nucleus with negative electrons moving round it, held to the nucleus by electrical attraction. Charging processes consist of a transfer of electrons; when ebonite acquires positive charge from polythene sheet, this is because it surrenders some electrons to the polythene, which consequently acquires equal negative charge.

1.2 The electron, and the unit of charge

The charge on one electron is the natural unit of electric charge, and it is used as such in atomic physics; indeed, modern research establishes even more firmly its fundamental nature. High energy physics reveals the existence, usually very brief, of a variety of subatomic particles, and these differ among each other in their masses. The electric charge, however, when the particle has any, is always the same, being equal to $\pm e$, where $-e$ is the same charge as that of an electron.

Such a unit, however, fundamental though it may be, is far too small to be taken as the basis of ordinary electrical measurements. The practical unit of charge is the coulomb (C), which will be properly defined presently (section 1.12).

The charge on one electron is given by $-e$, where

$$e = 1.6 \times 10^{-19}\ \text{C approximately,}$$

and its mass is m_e, where

$$m_e = 9.1 \times 10^{-31}\ \text{kg approximately.}$$

This is equal to only 1/1840 of the mass of the lightest atom, hydrogen, so that nearly all an atom's mass resides in its positively charged nucleus.

The conventional descriptions of electric charge and electric current were developed before the electron was identified, and for this reason, when we refer to a current flowing from A to B, this refers, unless the opposite is specified, to an effective flow of positive charge from A to B, although we are well aware that in most circumstances the reality is a flow of negative electrons in the opposite direction. This is the convention, however, which is almost universally observed, and arrows marking currents in circuits throughout this book will represent the direction of flow of positive charge.

1.3 Coulomb's inverse square law

When an electrical charge is sufficiently concentrated to be regarded as effectively a point charge, so that its sphere of influence spreads out symmetrically in three dimensions, its field strength is subject to an inverse square law, which amounts to saying that the force of attraction or repulsion between two such point charges may be expressed as

$$F = \alpha \frac{Q_1 Q_2}{x^2}, \tag{1.1}$$

where α is a constant term, and x is the distance separating the two charges. This law was originally formulated by Coulomb in the eighteenth century, and electrical forces between charges are often now referred to as Coulomb forces.

Putting the force F in newtons, the distance x in metres, and the charges Q_1, Q_2 in coulombs, experiment shows the value of α to be

$$\alpha = 9.0 \times 10^9 \text{N}\cdot\text{m}^2\cdot\text{C}^{-2}\ \text{approximately.} \tag{1.2}$$

It is interesting to compare this with the similar equation of gravitation, giving the force with which two concentrated masses attract each other, which is due simply to their mass;

$$F = G \frac{M_1 \cdot M_2}{x^2}, \tag{1.3}$$

where the value of G is

$$G = 6.67 \times 10^{-11}\ \text{N}\cdot\text{m}^2\cdot\text{kg}^{-2}. \tag{1.4}$$

Thus, unless the coulomb is an absurdly large unit in which to measure electrical charge, which it is not, electrical forces are, all things being equal, vastly more powerful than gravitational ones. The only reason that the earth holds us to its surface by the large force which is our weight, and stretches out its gravitational influence to hold the moon in its orbit, whereas there is no electrical contribution to these things at all, is the fact that whereas there are two kinds of electric charge, opposite in sign, there are not two kinds of mass. The gravitational effect of the accumulated mass of the whole earth is something very considerable. The balance of positive and negative charge in nature, on the other hand, is very finely kept, and a large electrical force is seldom experienced. When an ebonite rod parts with electrons to a sheet of polythene, it parts with only a very small proportion of the vast number of electrons in its atoms; if it were to part with many, the attraction between ebonite and polythene would make it impossible to separate the two.

1.4 An atom model

Since in discussing electrical currents we are discussing the flow of electrons, and these are subatomic particles, it is useful to have some form of atom model as a basis for discussion. This need not be in all respects consistent with up-to-date conceptions, because no model of the atom can be more than approximate to the truth. Indeed, the more accurate it becomes in detail, the more is it a mathematical interpretation rather than a physical one. For this reason, we do not discard an earlier model so long as it serves our limited purpose, and has some accuracy in resembling its successor; we accept it for the merit of simplicity.

The diagrams in Fig. 1.1 derive from the nuclear model of hydrogen, originally suggested in 1913 by Bohr, as a modification of the ideas of Rutherford.

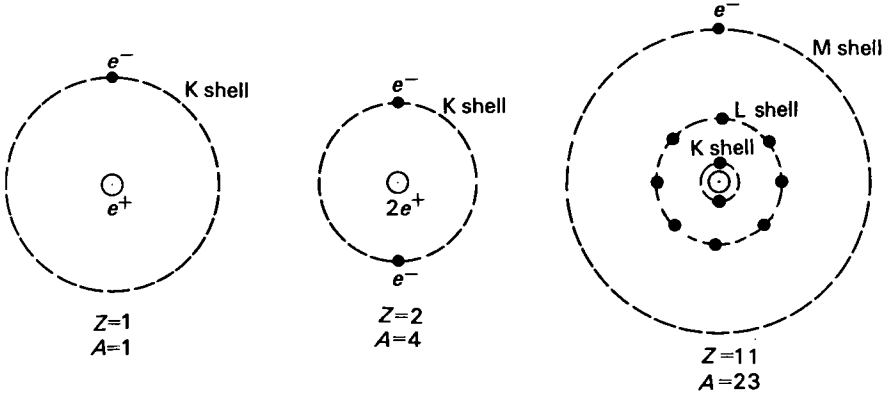


Fig. 1.1 Simple atom models Hydrogen Helium Sodium

These diagrams suggest that the nucleus of each atom acts as a kind of electrical planet, positively charged, with electrons orbiting round it, as a number of identical satellites, held in their orbits by the electrical attraction between positive and negative charges. These electrons are restricted in the energy which they may possess, and they inhabit successive electron shells, as indicated in the diagrams. Nucleus, electrons, and electron shells are all relevant to later models, as we shall later see, and so, with this simplicity, we can at present be satisfied.

1.5 The nucleus

The nucleus of an atom consists of protons and neutrons, and both these particles, for this reason, are classified as nucleons. The masses of both are very nearly the

same, both being nearly two thousand times as great as that of the electron. Thus nearly all the atom's mass is concentrated in its nucleus, which must be extremely dense, for it occupies only a minute fraction of the effective volume of the atom.

Whereas the neutron is electrically neutral, the proton has a positive charge which is exactly equal to that of the electron, but opposite in sign. Evidently there must be a powerful force operating inside the nucleus to hold the protons and neutrons together; otherwise wherever there is more than one proton, the electric force of repulsion between the protons would blow the nucleus apart. This cohesive force, which cannot be electrical in origin, since it must act on both protons and the electrically neutral neutrons, is termed a nuclear force. Evidently the nuclear force, while very powerful inside the nucleus, does not operate outside it, and is therefore an exceedingly short range force.

Atomic number Z

Because the only charged particles in the nucleus are the protons, it follows that the number of electrons which surround the nucleus of the neutral atom is equal to the number of protons within it, and it is this number which distinguishes one element from another, and gives each its position in the periodic table. This is called the atomic number Z of the element. For hydrogen, helium, and sodium, for example (Fig. 1.1), the atomic number is, respectively, $Z = 1$, $Z = 2$, and $Z = 11$. The hydrogen nucleus, in fact, consists of a single proton.

Nucleon number A

While the presence of a neutron in a nucleus affects the stability of the nucleus and adds to its mass, it contributes nothing to its charge, and so the total number of nucleons in the nucleus, protons and neutrons together, is not so significant as the number of protons alone. This total number is responsible for the mass of the nucleus, however, and is termed the nucleon number or mass number A. Thus, the sodium nucleus represented in Fig. 1.1, having nucleon number $A = 23$ and atomic number $Z = 11$, is a nucleus which contains 23 nucleons, of which 11 are protons and 12 are neutrons. This is represented in a nuclear equation by writing the symbol $^{23}_{11}Na$ for sodium.

1.6 Electron shells

The eleven electrons of the neutral sodium atom are pictured in the diagram as occupying places in three different shells, labelled K, L, M respectively. The K shell can accommodate no more than 2 electrons, and the L shell no more than 8, which leaves the eleventh electron no freedom to occupy a place in any shell which is closer to the nucleus than the M shell. This electron has more energy than the others, being further from the nucleus, and it is more easily freed from the influence of the nucleus. It is responsible for the chemical properties of sodium, and is referred to as the valency electron. It is also responsible for the conducting properties of the metal, and it is typical of a metal to find one, or sometimes more than one, electron occupying an incomplete outer shell.

Ions

If the valency electron deserts the sodium atom entirely, this leaves an atom, Na^+, which is positively charged, and this is termed a sodium ion. Another element might more readily gain an electron extra to its neutrality than lose one, and in doing so it becomes a negative ion, e.g. Cl^-. As a special case, a hydrogen ion, H^+, since the hydrogen has but a single electron to lose, is the bare hydrogen nucleus, i.e. a proton.

Inert gases

The element which comes before sodium in the periodic table is neon. This has all its ten electrons in completed K and L shells, and none left over for an uncompleted shell. Such an arrangement has exceptional stability, and neon has but few chemical properties. Helium, pictured in Fig. 1.1, is similar, having merely two electrons, which serve to complete its K shell. Helium is in one respect unique among the inert gases, because it takes only two electrons to complete the K shell. In all other cases the stable complement in the outer shell is eight, as it is with neon.

1.7 Crystalline structure

Most solids have a crystalline structure, which is to say that the atoms or ions have a regular arrangement in the solid, forming a crystal lattice, even if the separate crystals may be small, and the lattices imperfect. The forces which hold the atoms together are electrical attraction, which changes to repulsion if the atoms come so close together that their electron shells begin to overlap. Consequently, there is an equilibrium position for each atom in the lattice, as implied in Fig. 1.2. Each atom is supposed to be vibrating about this position, with energy which increases with temperature rise.

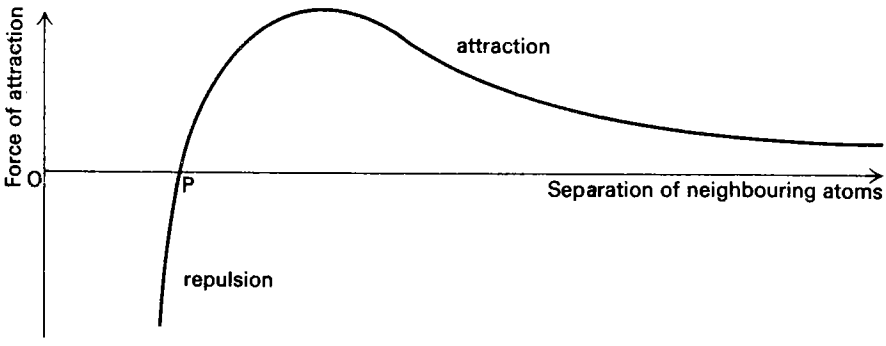


Fig. 1.2 The electrical force between neighbouring atoms or ions in a crystal lattice

It will be noted that the graph is not quite linear where it passes through the equilibrium point P, owing to the rapidity with which repulsion increases if the separation of the atoms is reduced. A result of this is that the mean separation of the vibrating atoms increases slightly when the energy of vibration increases. This causes slight irregularities in the crystal lattice, because not all the vibrating atoms have the same energy, and this will have some relevance when the causes of electrical resistance later come under discussion (section 3.12).

1.8 Metallic conduction

A finite-sized specimen of a metal is usually a collection of very small crystals, although single large crystals can be specially prepared. The overall effect upon a valency electron in the metal is that it experiences the attractions of neighbouring nucleii as well as that of its own atom. As a result, a large proportion of the valency electrons cease to be bound to any individual atom, although they are bound to the metal as a whole. They are free to wander at random through the metal, and are, in fact, in a condition of continuous random motion, colliding with each other and with ions in the lattice, rather like molecules in a gas. The atoms in the lattice are,

therefore, positively charged ions, which attract the electrons to the bulk metal, and in doing so are attracted by the electrons. It is this mutual force between the ions and the negatively charged 'electron gas', which fixes the ions in place.

Being free to wander through the metal in this way, the valency electrons may be termed conduction electrons, being available to act as charge carriers, i.e. to convey a current, when a directional drift is imposed on them by an electric field. It is the presence, in abundance, of these conduction electrons which makes a metal into a good electrical conductor.

1.9 The electric circuit

The amount of charge stored in a charged conductor is insufficient to do more than produce a small transient current while the object is being discharged. For this reason, electrical discoveries before the nineteenth century were restricted to the fields of magnetism and electrostatics, and the study of electric currents had to wait until the first electric cell was invented by Volta in 1800, and the principle of the dynamo was discovered by Faraday in 1834. These are the electrical counterparts of a pump in a water circulating system, being capable of maintaining a continuous current of conduction electrons round a closed circuit. We note that a mechanical pump does work, and has to be supplied continuously with energy. It is the same with the cell and the dynamo, the source of energy being respectively chemical and mechanical.

When a continuous electric current flows, the strength of the current I is a measure of the quantity of charge which is passing any point in the circuit in unit time. If charge Q passes the point in time t, the current is I, where

$$I = \frac{Q}{t} \tag{1.5}$$

If Q is measured in coulombs, and t in seconds, the current is in amperes. As a practical matter, however, it is the size of the ampere which is fixed first by definition, and the coulomb becomes the secondary unit, being defined by writing equation 1.5 in the form

$$Q = It \tag{1.6}$$

and the coulomb is the ampere second.

Before we can define the ampere, some brief description of the magnetic effect of an electric current is necessary, because it is by its magnetic effect that a current is usually measured, and by its magnetic effect that it is defined.

1.10 The magnetic field due to a current

When a conductor carries current, it is surrounded by a magnetic field, e.g. it deflects a magnetic compass placed near by – an effect which was first observed by Oersted in 1819. The magnetic field is not caused by a stationary charge, as the electric field is, but appears only when there is a current. Also the presence of the magnetic field can be revealed by a force acting on a magnet, or on a moving charge, but there is no force on a stationary charge, such as the electric field produces.

The lines of magnetic flux, drawn to map a magnetic field, as, for example, those shown in Fig. 1.3, mark the direction in which a compass points. The lines of flux in the diagram show the magnetic field set up by current flowing in a long straight conductor.

If a second straight conductor, also carrying current, is placed parallel to the

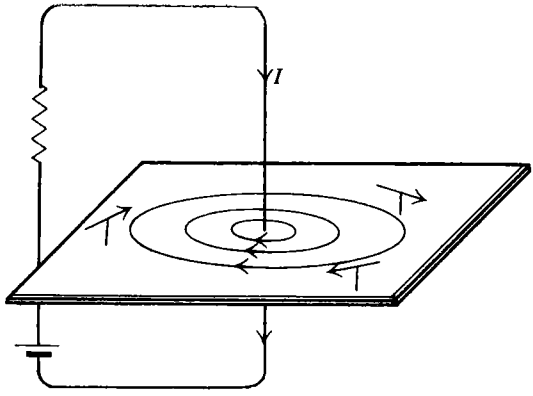


Fig. 1.3 The magnetic field due to current in a straight conductor (The effect of the earth's field has been neglected)

first, so that each conductor is perpendicular to the field of the other, there is a mutual force between the two conductors. If the currents both flow in the same direction, there is a force of attraction, and if one of the two currents is reversed, the force becomes repulsion.

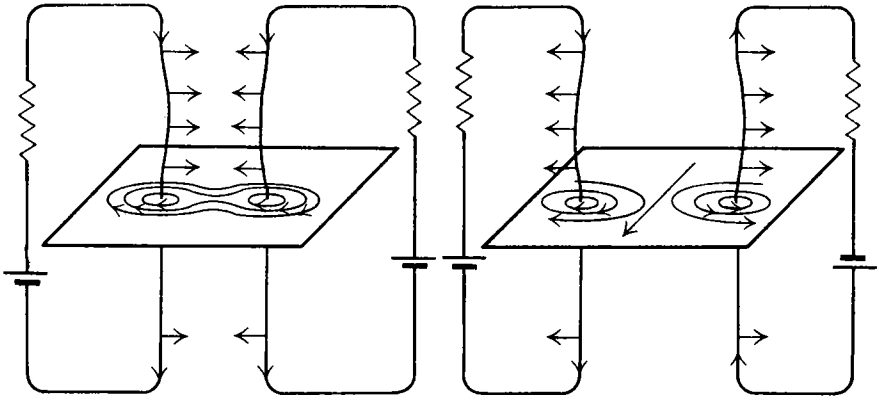


Fig. 1.4 The magnetic interaction between two parallel conductors which carry current (a) Attraction (b) Repulsion

1.11 The definition of the ampere

When two parallel conductors are made very long, the force F which acts between them is given by the equation

$$F = \beta 2 I_1 I_2 \frac{l}{r}, \tag{1.7}$$

where r = the distance separating the two conductors, and F = the force which acts on length l of one conductor.

The practical unit of current, the ampere (A), is defined on the basis of this equation in such a way that the constant β takes the *exact* value, 10^{-7},

i.e. $$\beta = 10^{-7}\ \mathrm{N \cdot A^{-2}} \text{ exactly.} \tag{1.8}$$

The formal definition of the ampere is as follows:

If a current of one ampere flows in each of two very long parallel conductors, of negligible cross-section, held one metre apart, in vacuo, the force of attraction

between one conductor and one metre length of the other conductor is 2.0×10^{-7} newtons.

It must be admitted that this definition is not suited to direct practical application, since an accurate apparatus cannot be constructed to comply with the specified conditions. It will be seen later, however (chapter 17), that the definition is consistent with equations which fit more practical arrangements. The condition, 'in vacuo', need not, in practice, be insisted upon, provided that all ferromagnetic materials, such as iron, are kept at a distance.

1.12 The definition of the coulomb

This definition of the ampere leaves the unit of charge, the coulomb, in the subsidiary place, defined by equation 1.6. But the choice has been made on practical grounds, as providing the best basis for measurement. This does not alter the theoretical pre-eminence of electric charge; an electric current is a flow of electric charges.

The formal definition of the coulomb is:

The coulomb is the practical unit of electric charge. When a current of one ampere flows in a circuit, a charge of one coulomb passes any point in the circuit in every second.

1.13 A note on the values of the constants α and β

It will be noted, on comparing the electrostatic equation 1.1

$$F = \alpha \frac{Q_1 Q_2}{x^2} \tag{1.1}$$

with the electromagnetic equation (1.7)

$$F = \beta 2 I_1 I_2 \frac{l}{r} \tag{1.7}$$

that, using SI units, the constant β is allotted the exact value, 10^{-7}, by definition of the ampere, whereas the value of the constant α can be determined only by experiment. The fact that its value in SI units, $\alpha = 9.00 \times 10^9$, is so close to being a simple number is fortuitous, and has not been arranged by definition.

1.14 SI units

In the selection of a system of electrical units, it is necessary to single out only one electrical quantity which must rest on an arbitrary definition. By accepting the established units of mass, length, and time, and adding to them this one electrical unit, it becomes possible to build up a systematic series of units for all electrical quantities.

The practical system adopted in electricity uses the MKS units of length, mass, and time, the joule being the accepted unit of electrical energy. The ampere having been selected as the additional fundamental electrical unit, the system becomes the MKSA system. This is a part of the universal system of units which is now being adopted for use in all branches of physics, known as the Système International. The fundamental SI units are the metre, kilogram, second, ampere, candela, and kelvin, the last two of which are units of luminous intensity and of temperature respectively.

QUESTIONS

the electrostatic constant α	$= 9.0 \times 10^9\ \mathrm{N \cdot m^2/C^2}$
the charge on one electron	$= -e$
the charge on one proton	$= +e$
where e	$= 1.6 \times 10^{-19}\ \mathrm{C}$
the mass of an electron m_e	$= 9.1 \times 10^{-31}\ \mathrm{kg}$
the mass of a proton m_p	$= 1.67 \times 10^{-27}\ \mathrm{kg}$
the constant of gravitation G	$= 6.67 \times 10^{-11}\ \mathrm{N \cdot m^2/kg^2}$

1.1 (*a*) The nucleus of the helium atom contains two neutrons and two protons; its radius is believed to be of the order of 10^{-15} m. Estimate the minimum force with which their electrical charge must cause the two protons inside the nucleus to repel each other. On what basis is their continued presence together in the nucleus accounted for?

(*b*) The Bohr model assumes that the hydrogen atom consists of a single proton as its nucleus, with an electron moving round the nucleus in a circular orbit of radius r. Find the ratio of the electrostatic force acting between the two particles to the gravitational force of attraction between them. (OX. SCH.)

1.2 Two small copper spheres, each of mass 10 g, are placed with their centres 20 cm apart. Supposing that it were possible to remove one electron from every millionth atom in one sphere, and to transfer all these electrons to the other sphere, calculate the force which would act between the two spheres as a result.

Make the simplifying assumption that the charge on each sphere acts as if it were concentrated at the centre.

(63.6 g of copper contains 6×10^{23} atoms.)

1.3 Two small spheres are charged up by transferring electrons from one sphere to the other. When they are placed with their centres 5 cm apart, the force of attraction between the two spheres, due to their charge, is 10^{-3} N. Calculate the charge, in coulombs, on each sphere. How many electrons have been transferred?

Supposing that the two spheres are now connected together by a conducting wire, so that the electric charge flows back again, taking 10^{-6} s altogether, what is the average value of the current which flows during this time?

1.4 When an electric cell is connected across two copper plates immersed in a solution of copper sulphate, electrolysis takes place, and copper ions are transferred from one copper plate to the other. The copper ion lacks two of the electrons which belong to the neutral copper atom, so that it is positively charged. Assuming that there are 6×10^{23} copper atoms in 1 mole, i.e. in 63.6 g, calculate the total charge which has been carried through the solution by the copper ions by the time that 1 g of copper has been transferred from one plate to the other. Such a charge, stored on one of the plates at the expense of the other, would exert an enormous force. Explain why no such force materializes.

1.5 Suppose that a man, whose mass is 70 kg, is standing on an insulating platform at a point on the earth's surface. Now suppose that it were possible to remove charge $+Q$ from the man, and to position this charge at the point on the other side of the earth, diametrically opposite to the point at which the man is standing. Calculate the size of the charge Q, if the resulting electrical force on the man were to have the effect of doubling his apparent weight. How many electrons would have been transferred?

(The earth's radius $= 6.4 \times 10^6$ m;

$g = 10\ \mathrm{m/s^2}$)

Chapter 2

Potential and energy

2.1 Potential and potential difference

Figure 2.1 shows two arrangements which are, in some respects, analogous.

In the first case, the pair of cylindrical vessels labelled B has a larger capacity than the separate vessel A, with the result that, if equal amounts of water are poured into both A and B, the level rises higher in A. It is this difference in levels which governs the direction of the water flow when the tap in the connecting pipe is turned on. The water flows from the higher level to the lower, until the two have become equalized. The actual amount of water originally contained by each vessel is relatively unimportant here; system B might actually contain more water than A in the first place, and yet the flow will be from A into B because the level in A is the higher.

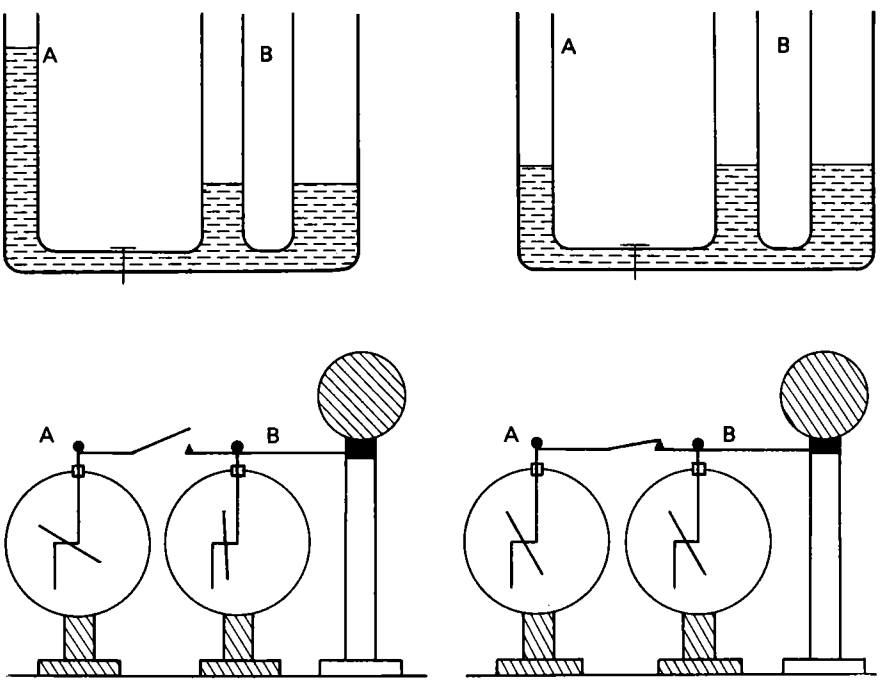


Fig. 2.1 An analogy between water level and electric potential (a) A and B hold equal amounts of water (b) Water flows from A to B to equalize the water levels (c) A and B hold equal amounts of electric charge (d) Charge flows from A to B to equalize the potentials

The electrical arrangement suggested in the second pair of diagrams behaves in an analogous way. The right-hand electroscope, B, is connected to a large insulated metal sphere, put there to increase the amount of electric charge which the system

can easily store, whereas electroscope A is isolated. If equal amounts of charge are given to A and B, the needle of electroscope A rises higher than that of B. If the two systems are now connected together by a conducting wire, charge flows from A to B, until the two electroscopes register equal deflections. It is evident that the direction in which charge flows does not depend primarily on the relative amounts of charge held by the two systems, but on some condition analogous to the difference in water levels of the first experiment, and this was indicated by the needles of the electroscopes. This is described by saying that the electric potential of A was initially higher than that of B, and this is what the electroscopes registered. Charge flows from A to B as long as there is a potential difference between them, the potential of A being the higher.

The practical unit in which potential and potential difference are measured is the volt (V). The description given above, however, is not precise enough to serve as a basis of definition. A proper definition involves energy.

2.2 Energy and power

There is no problem which is special to electricity in defining energy, or a unit in which to measure it. Since energy can be converted from one form to another, we may take the definition, and the unit, from mechanics. The unit used in practical electricity is the joule (J).

Energy – the joule

If W units of energy are converted to mechanical form, the work available can move a body through a distance s against an opposing force F where

$$W = Fs. \tag{2.1}$$

Putting F in newtons, and s in metres, the energy W is in joules,

$$1\text{J} = 1\text{N}\cdot\text{m}.$$

Power – the watt

Power is the rate of converting energy, e.g. the rate of doing work.

If energy is converted from one form to another at the rate of W units of energy in time t, then the power is P, where

$$P = \frac{W}{t}. \tag{2.2}$$

Putting W in joules, and t in seconds, the power P is in watts (W), 1W = 1J/s. The kilowatt hour (kWh) is a unit of energy, often used commercially in preference to the joule to measure large amounts. It does not properly belong in the SI system, however, because the second and not the hour is the SI unit of time. Writing equation 2.2 in the form

$$W = Pt, \tag{2.2a}$$

if P is in kilowatts, and t is in hours, then the energy W is in kilowatt hours, and 1 kWh = 3.6×10^6J. Similarly of course, 1J = 1W·s.

2.3 The definitions of potential difference and potential

Potential difference and its practical unit, the volt, can now be properly defined. First, we return to the water analogy, with the arrangement shown in Fig. 2.2.

Figure 2.2 shows two water tanks. Tank A contains weight G of water, raised

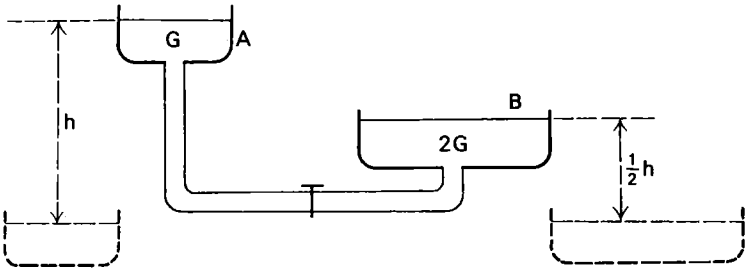


Fig. 2.2

to a height h above the ground. Tank B contains twice as much water, $2G$, but this is at half the height, $\frac{1}{2}h$.

The potential energy of the water W is the same for both tanks, viz.

$$W = G \cdot h = 2G \cdot \tfrac{1}{2}h. \tag{2.3}$$

It is neither this energy, nor the weight of water, which governs the direction of the water flow when the tap in the connecting pipe is turned on. What matters here is the difference in water levels $(h - \frac{1}{2}h)$.

It is something electrically analogous to this difference in levels which we require to define in a way which lends itself to practical interpretation. The method of achieving this can be suggested by noting that the heights h, $\frac{1}{2}h$ of the water in the tanks are equal to ratios of the energy possessed by the water to its weight,

i.e.

$$h = \frac{W}{G}, \qquad \frac{h}{2} = \frac{W}{2G}. \tag{2.4}$$

Electric potential is defined on a similar basis, as follows.

Potential difference

The potential difference V between two points is equal to the amount of electrical energy which is converted to other forms of energy, for every unit of charge which is transferred from the point at higher potential to the point at lower potential,

i.e.

$$V = \frac{W}{Q}. \tag{2.5}$$

[More strictly, this equation should be written in the form

$$V = \frac{\delta W}{\delta Q} \tag{2.5a}$$

for the reason that transfer of an appreciable quantity of charge would, in some circumstances, result in a change in the potential difference V just as the water levels in Fig. 2.2 would be affected by the flow of an appreciable quantity of water.]

The abbreviation p.d. will be used throughout the rest of this book.

The volt

Using SI units, W is measured in joules, and the charge Q in coulombs. The unit of p.d. V is then called the volt,

i.e. $$1\text{V} = 1\text{J/C}.$$

Potential

The electric potential at any one point is the p.d. between that point and another point which is arbitrarily chosen as being at zero potential.

In theoretical work, this zero is usually taken to be the potential at a place which is infinitely distant from any electric charge. In practice, however, the earth is more often chosen instead, and any conductor, such as a water pipe, which makes good connection with the earth is at zero potential.

A cell may be used to raise the potential at a point above zero, or to put it at negative potential, as in the examples given in Fig. 2.3.

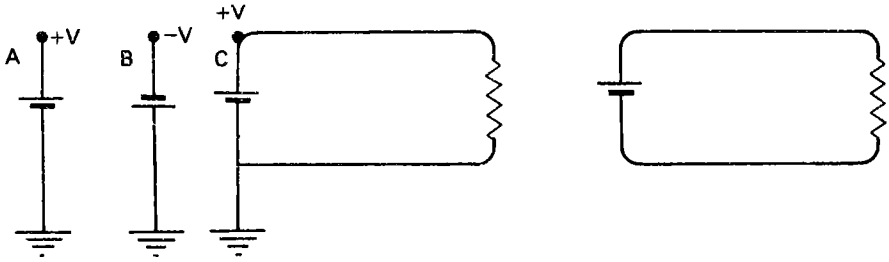


Fig. 2.3 Earth connections (a) The potential of A is $+V$ (b) The potential of B is $-V$ (c) The potential of C is $+V$ (d) This is a 'floating' circuit; it has no reference to earth potential

2.4 Potential difference and power

In Fig. 2.4 the water analogy is applied to the case of a continuous circulation of water. In diagram (*a*), a pump is being used to create and maintain an artificial difference in water levels, the available energy then being used to drive a turbine. (It is of interest to note that the early steam engines were used to operate water pumps, and the energy was originally applied to machinery by making artificial water falls in this way.) The pump maintains a continuous circulation of water. When the water has lost its energy to the turbine in falling, the pump raises it again.

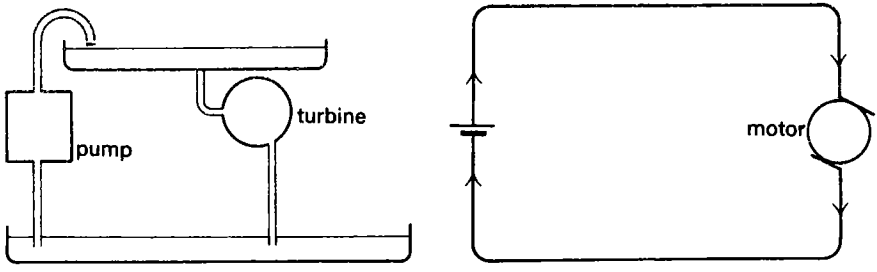


Fig. 2.4 A continuous circuit (a) (b)

In a similar way, in the electric circuit of diagram (*b*), the cell maintains the circulation of electric charge. When the charge has lost its energy to the electric motor, and fallen in potential in consequence, the charge flows through the cell again, where its energy supply is replenished, and it is raised again in potential.

In the case, such as this, of a continuous current, it is convenient to put the definition of p.d. in terms of power P and current I rather than referring to energy and charge. Since power P is the rate of converting energy,

i.e.
$$P = \frac{W}{t}, \tag{2.2}$$

and current I is the charge passing any point in a circuit in unit time,

i.e.
$$I = \frac{Q}{t}. \tag{2.6}$$

we have, quoting equation 2.5,
$$V = \frac{W}{Q}, \tag{2.5}$$

whence
$$V = \frac{P}{I}, \tag{2.7}$$

which is, more usually, quoted in the form
$$P = VI, \tag{2.8}$$

where, using SI units: P is measured in watts, V in volts, I in amperes.

[It is emphasised that equation 2.8, using SI units, expresses a definition of the volt. It does not define the watt, which is the name given to the joule per second.]

2.5 Electromotive force

In the circuit of Fig. 2.4(*b*), both the cell and the motor are performing energy conversions; from chemical form to electrical in the cell, and from electrical to mechanical form (and some heat) in the motor. From the electrical standpoint, however, the function of the cell is a special one because, in providing electrical energy, the cell is the source of the p.d. across the motor and of the current. The energy conversion in the cell for every unit of charge which flows through it is consequently given a special name, which is applied to all sources, and this is electromotive force, or e.m.f. Thus:

The e.m.f. of a cell, or of any other electrical source, is equal to the electrical energy provided by the cell for every unit of charge which flows through it.

The SI unit of e.m.f. is the same as that of p.d., i.e. the volt, because the definitions of both e.m.f. and p.d. are correctly expressed by equation 2.5.

[This definition will be put rather better in section 3.21.]

Back e.m.f.

A further distinction can be made between e.m.f. and p.d., which concerns reversibility of the energy conversion. If a more powerful source were inserted in the circuit of Fig. 2.4 (*b*), acting in the opposite direction round the circuit, this would reverse the direction of the current. The direction of the p.d. across the motor would consequently be reversed, but its energy conversion would not; it would continue to act as a motor, converting electrical energy to mechanical form. This would not, however, be the effect on the cell, the e.m.f. of which would continue to act in the same direction as before; because the current is reversed, however, energy would now be converted in the cell from electrical form to chemical, i.e. the energy conversion would be reversed. When an e.m.f. acts against the direction of the current in this way, it is often referred to as a back e.m.f. It is the special feature of an e.m.f., which distinguishes it from a p.d., that reversal of the current results in reversal of the energy conversion.

2.6 High and low voltage power

The power P accepted by a load is equal to the product of the current I through the load and the p.d. V across it. This is equation 2.8

$$P = VI \tag{2.8}$$

It follows that a small current, taken through a large potential drop, can release energy at the same rate as a larger current, for which the potential drop is smaller.

This point is illustrated in the circuits of Fig 2.5, where four torch bulbs are made to give the same amount of light in three different arrangements.

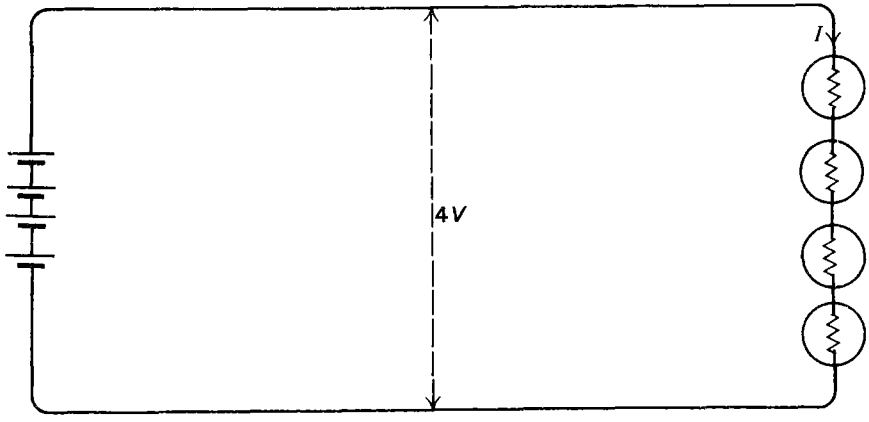


Fig. 2.5 (a)

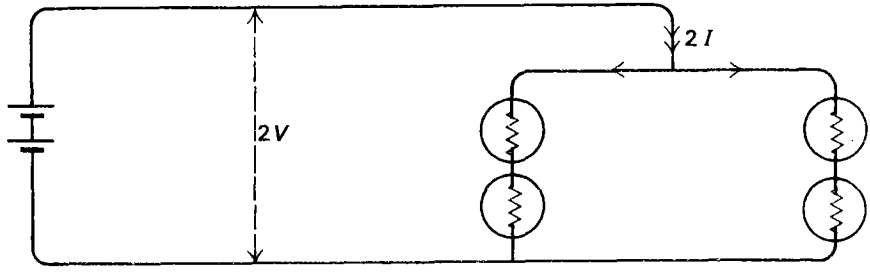


Fig. 2.5 (b)

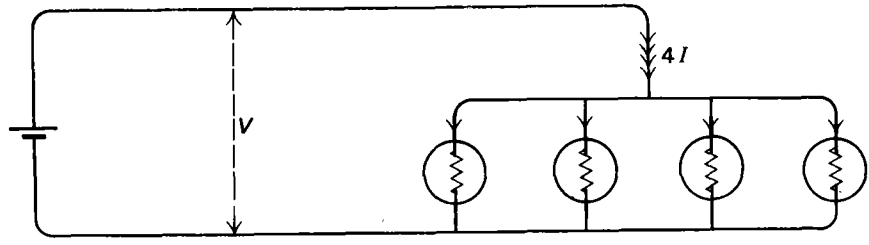


Fig. 2.5 (c)

Fig. 2.5 Three arrangements of four bulbs which take the same power (a) $P = 4V.I$ (b) $P = 2V.2I$ (c) $P = V.4I$

Various analogies can be drawn from mechanics, including that suggested by Fig. 2.2, where a large weight of water at a low level stores the same amount of potential energy as half the weight of water at twice the height.

2.7 Current and p.d. distribution in a circuit

Figure 2.6 illustrates the basic principles relating to current and p.d. in a network. The current distribution is consistent with conservation of electric charge. Charge

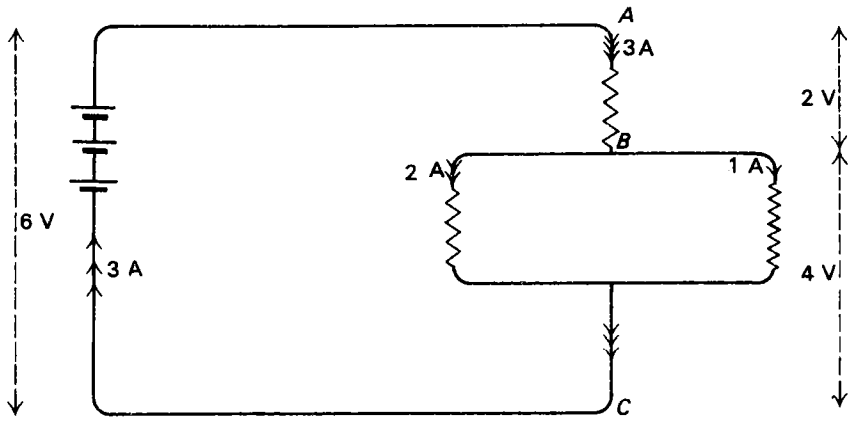


Fig. 2.6 Distribution of currents in a network

cannot appear from nowhere, nor can it disappear, or accumulate at a point. Hence, when the circuit branches, the current divides, being shared between the branches, e.g. in the example of the diagram,

$$3 = 2 + 1.$$

This principle will be stated formally in the next chapter, as Kirchhoff's first law.

Potential difference may be treated in terms of potential drop, analogous to fall in level of a flow of water. The battery in the diagram puts a p.d. across the network equal to 6 V. If the potential drop in AB is 2 V, this leaves a further potential drop, $(6 - 2) = 4$ V, to appear across BC. This is so, whether the left-hand branch is being considered, or the right-hand branch; similarly, if two paths on a hillside were to branch apart at a point B, and come together again at a point C, both paths would take the traveller down through the same height h.

This treatment of p.d. is an aspect of conservation of energy. Every coulomb of charge flowing through the 6 V battery acquires 6 J of electrical energy. Since the potential drop in AB is 2 V, 2 J of energy are lost in the path AB, leaving a further 4 J to be lost by each coulomb of charge before reaching C, whether it follows the left-hand branch or the right. This principle also will be stated formally in the next chapter, as Kirchhoff's second law.

To account for the whole of the energy taken from the battery, the energy dissipated in all branches of the network must be allowed for: e.g. referring to the diagram,

the power supplied by the battery $= 6 \times 3 = 18$ W;
the power used in AB $= 2 \times 3 = 6$ W;
the power used in the left-hand branch of BC $= 4 \times 2 = 8$ W;]
the power used in the right-hand branch of BC $= 4 \times 1 = 4$ W;]
$\therefore$ the total power used = the total power supplied.

2.8 The acceleration of electrons in a vacuum tube

When electric charges suffer a potential drop while flowing round a circuit, what happens to their energy depends, of course, on the nature of the circuit, and the simplest case is when they are crossing the vacuum of a device such as a cathode ray tube or an X-ray tube. The electrons which convey the current in these tubes are emitted into the vacuum by a hot metal surface, the cathode, and attracted towards a second electrode, the anode, because a large p.d. is applied between the electrodes, the anode being made positive relative to the cathode (which is what the names anode, cathode are intended to imply). In the case of the X-ray tube, of which a sketch is given in the next diagram, the value of the applied p.d. may be 5.10^4 V or more.

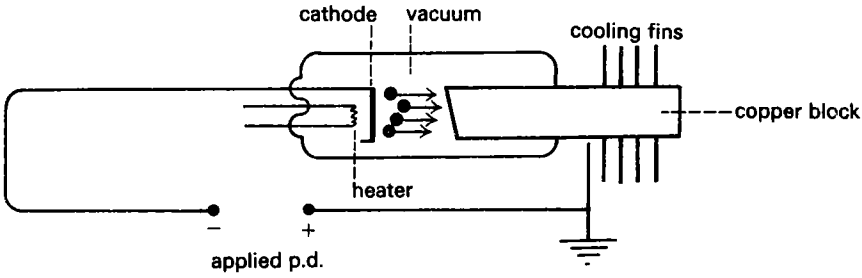


Fig. 2.7 An X-ray tube

Since there is nothing in the vacuum to offer resistance to the passage of the electrons, the electrical energy is not converted to heat, but to kinetic energy, instead. Heat is produced only where the electrons are brought to rest, and in the X-ray tube this is when they strike the anode. [The intention in the X-ray tube is to convert this energy into electromagnetic waves of very short wavelength, i.e. X-rays, but a large proportion of it turns into heat instead. This is the reason for the arrangements shown in the diagram for heat dissipation.]

The kinetic energy gained by the electrons is easily deduced. The case is analogous to that of a freely falling ball; the electron, with charge e, has fallen through a potential drop V. The loss of electrical energy is eV. Hence, the kinetic energy acquired, E_k, is

$$E_k = eV. \tag{2.9}$$

Because the electron has negative charge, the potential drop referred to above means a drop in negative potential; the electrons 'fall' towards the electrode, i.e. the anode, which has the higher positive potential.

2.9 The drift speed of charge carriers in a conductor

When electrons or any other charge carriers flow in a conductor as a result of an applied p.d., they do not have the same easy passage as electrons do in a vacuum tube. The atoms of the conductor which lie in their path have the effect of arresting their progress, with the result that they repeatedly gain a small speed, only to lose it again. The net effect is that the electrons drift through the conductor, with a mean speed v, which is usually quite slow, and has the same value at all parts of the conductor.

The energy, VQ, is converted to heat, distributed throughout the conductor.

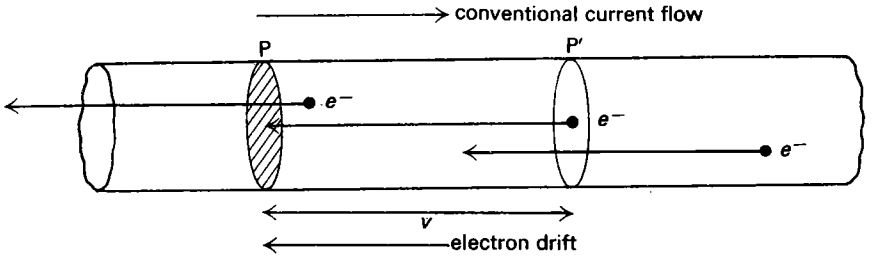


Fig. 2.8

2.10 The drift speed of charge carriers; its relation to current

The current I in a conductor is equal to the charge which passes any point in unit time. If the current is, for example, 6 A, this means that 6 C are passing any point in the circuit in every second.

This statement conveys no information as to how fast the individual charge carriers, e.g. electrons, are moving. The connection between the current I and the mean speed v will now be derived.

Figure 2.8 represents a section of a conductor, in which free electrons are supposed to be drifting from right to left with a mean speed v. Because electrons are negative charges, this represents a conventional flow of current from left to right.

Three electrons are shown in the diagram, one close to the plane P, one at distance v from P, and one at a distance from P which is greater than v. Since we have assumed that, on average, all the electrons are moving a distance v to the left in unit time, it is evident that, in the course of the next unit of time, the first electron will have crossed through the area A at P, the second electron will have just reached P, and the third will not yet have travelled far enough. All the electrons which will pass through P in the next unit of time are contained between the two planes P and P′, where the distance PP′ is v.

Suppose there are n conduction electrons in unit length of the conductor. Then there are nv of these electrons between P and P′. This is a charge equal to nve, where the charge on one electron is e. Since, as we have shown, these are the charges which will pass through P in unit time, the equation giving the current I is

$$I = nve, \tag{2.10}$$

where n = the number of conduction electrons in unit length of conductor, v = the mean drift speed, and e = the charge on one electron. Alternatively, equation 2.10 may be written

$$I = NAve, \tag{2.11}$$

where A = the area of cross-section of the conductor and N = the number of conduction electrons in unit volume. The conduction electrons are, of course, in random motion all the time. This random motion is not concerned in the above equation, only the mean drift speed in one direction, which is superimposed on it, when there is a current.

In some cases, such as that of electrolysis, or in some semiconductors, the charge carriers are not necessarily electrons, and the current is a result of a flow of positive charge in one direction, and negative charge in the opposite direction. To cover such cases, the general form of equation 2.11 is

$$I = N_1 A v_1 Q_1 + N_2 A v_2 Q_2. \tag{2.12}$$

In a metal, the concentration N of conduction electrons is so large that the speed v is surprisingly small, even when the current is large. If, for example, a current of 2.0 A is flowing in a copper wire of 1 mm diameter, the mean electron drift speed is about 0.2 mm/s.

2.11 The speed of a signal

The drift speed v of the charge carriers in a conductor has no relation with the speed at which a signal is conveyed along the conductor.

According to our common experience, if the lamp L in Fig. 2.9 is served by straight conductors, it lights up as soon as the switch S is made. Any apparent delay is usually due to the time taken by the filament to heat up. The lamp does not have to wait for an individual electron to travel all the way to it from the switch. There is a whole succession of conduction electrons, all the way round the circuit, ready to start moving as soon as the switch is made.

Yet the signal does take time to travel, although its speed is very fast indeed. The electron drift at A and A′ in Fig. 2.9 starts as soon as the switch is made.

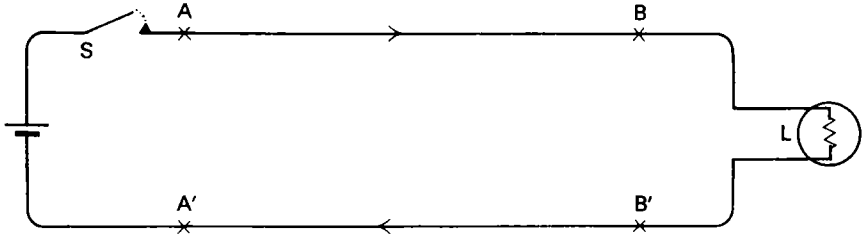


Fig. 2.9

The electron drift at B and B′ starts a very small fraction of a second later. The speed with which the signal travels, using air-spaced conductors, is 3×10^8 m/s, which is the same as the speed of light.

If a lamp were 3.10^5 km away from the switch, it would begin to light up after a delay of 1 s. Such a negligible effect does not seem worth considering. The result might become noticeable, however, if the switch were a vibrator, with a repetition rate of, say, 3.10^7 contacts every second. If the lamp were 5 m from the switch, it would then be receiving current whenever the switch was off, and no current when the switch was on. This is not a practical suggestion, of course, in terms of ordinary switches. But an electric oscillator with frequency 3.10^7 s^{-1}, i.e. 30 MHz, has the effect of a reversing switch with this frequency. It is with such high frequency oscillators that the speed of the signal becomes an important factor.

QUESTIONS

For an electron

$$g = 10 \text{ m/s}^2$$
$$e = 1.6 \times 10^{-19} \text{ C}$$
$$\frac{e}{m} = 1.76 \times 10^{11} \text{ C/kg}$$

2.1 A waterfall is 300 m high, and 200 kg of water in the fall pass through a turbine in every second. If the power of this water is used to generate electricity, and the overall efficiency is 80%, calculate the value of the output current, if the p.d. is 240 V.

If the same power were generated, instead, at a p.d. of 4800 V, what would the value of the current be in this case?

2.2 Two electric motors both produce a mechanical power of 1500 W, and the overall efficiency of the energy conversion in each case is 90%. If the input p.d. of one motor is 240 V, and that of the other is 100 V, calculate the current taken by each motor. Calculate also the speed with which each motor could raise a 20 kg weight.

2.3 A small dynamo is energized for a short period of time by using a falling 0.5 kg weight to turn its armature. If the dynamo, during this period, provides sufficient power to light a 2.5 V, 0.3 A torch bulb, and if the energy conversion is 75% efficient, estimate the speed at which the weight is falling, assuming this to be constant.

2.4 A metal wire has 5.0×10^{28} conduction electrons per cubic metre, and its area of cross-section is 2.0 mm^2. If these electrons, under the influence of an applied p.d., have a mean directional drift velocity of 0.30 mm/s, what is the current flowing in the wire?

2.5 A beam of electrons is accelerated in an evacuated cathode ray tube by applying a p.d. of 1000 V between two electrodes. Deduce the velocity attained by the electrons.

If the current in the tube is 0.10 mA, how many electrons flow through the tube in each second, and at what rate does the tube take energy from the supply?

2.6 A beam of protons is accelerated in an evacuated tube by applying a p.d. of 4 kV between two electrodes. Calculate the velocity attained by the protons.

If helium nucleii, He^{2+}, were accelerated through the same p.d. instead, what would the velocity be in this case?

(The charge/mass ratio for a proton $= 9.6 \times 10^7$ C/kg.
The nucleon number of helium = 4.)

2.7

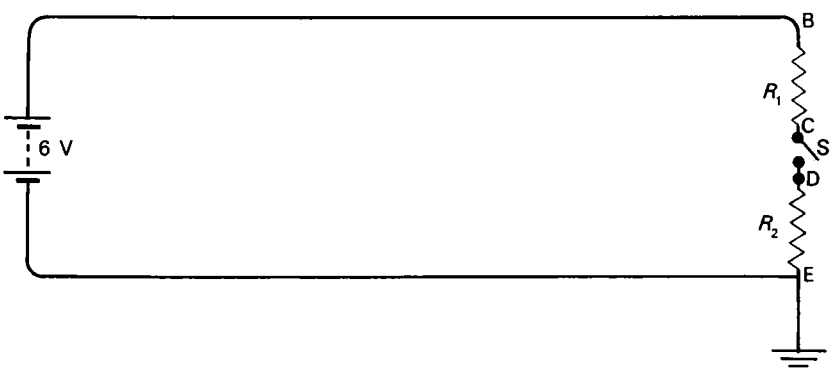


Fig. e 1

Whether switch S is open or closed, the p.d. across BE is 6.0 V. When the switch S is closed, the p.d. across $R_2 = 2.0$ V.

(*a*) When the switch is closed, deduce the value of the p.d., (i) across BC, (ii) across CD, (iii) across CE.

(*b*) When the switch is open what is the p.d. now, (i) across BC, (ii) across CD, (iii) across CE?

2.8

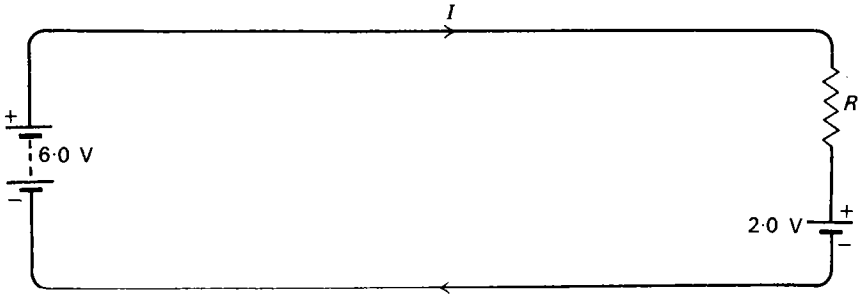


Fig. e 2

If the p.d.s across the two batteries in this circuit, as shown, are 6.0 V and 2.0 V respectively, what is the p.d. across the resistor R?

Assuming that the value of the current I is 2.0 A, what power is being taken from the 6.0 V battery? How much of this power is producing heat in the resistor? Suggest what happens to the remainder.

2.9

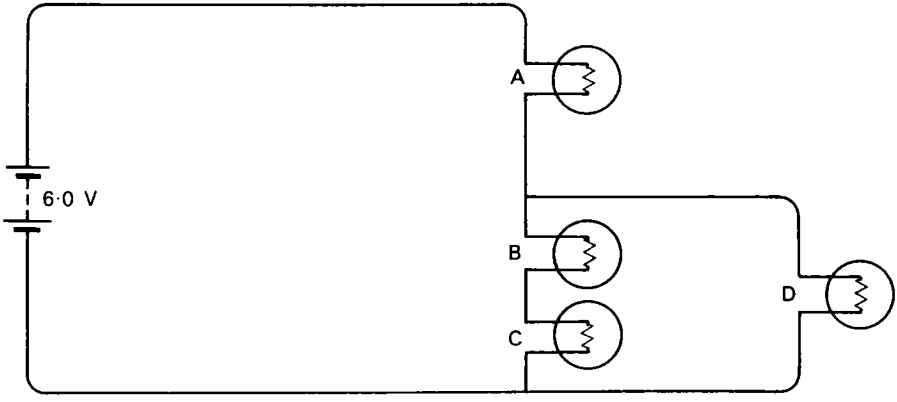


Fig. e 3

A, B, C, D are all exactly similar torch bulbs.

(*a*) Which bulb is glowing the most brightly? Why?

(*b*) If bulb D were disconnected, what change would you expect to occur in the brilliance of each of the others?

(*c*) If bulb D were reconnected, and bulb B were disconnected instead, what effect would you expect this to have on each of the others?

(*d*) When all the bulbs are connected, the p.d. across C is 1.2 V, and the current through it is 0.2 A, while the current through bulb D is 0.4 A. Deduce the values of the p.d.s across bulbs A, B, and D respectively. Deduce also the power taken by each bulb and the power supplied by the battery. Show that your answer is consistent with conservation of energy.

2.10 Once the speed of an accelerated particle becomes so great that it is more than (say) one tenth of the speed of light, the mass of the particle begins to increase appreciably, and some of the equations of Newtonian dynamics no longer apply.

Calculate the p.d. which will accelerate (*a*) an electron, (*b*) a proton to the velocity v, where v is one tenth of the speed of light.

(The speed of light $= 3.0 \times 10^8$ m/s.
The charge/mass ratio for a proton is 9.6×10^7 C/kg.)

2.11 In the circuit shown, R is the resistance of a heater in a well-lagged constant flow calorimeter. When the resistor X is adjusted so that the galvanometer G

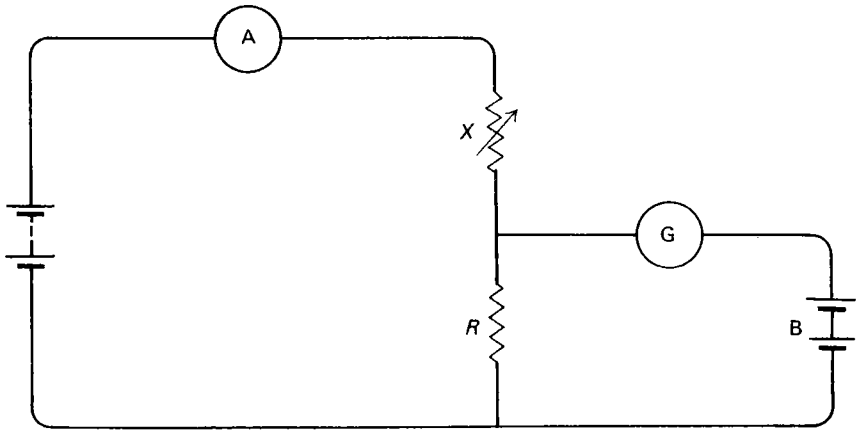


Fig. e 4

shows no deflection, the ammeter reads 5.5 A. In 60 s, 50 g of water flow through the calorimeter, and the steady input and output temperatures are 12.0 °C and 18.0 °C respectively. Calculate the e.m.f. E of the battery B.

(The specific heat capacity of water $= 4.2 \times 10^3$ J/kg·K.) (C)

2.12 In a commercial X-ray tube, a current of 15 mA of electrons strikes a target which is at a potential of 100 kV positive with respect to the filament. What flow of water is required to keep the target cool, assuming that all the heat is conducted away by the water, and that its rise in temperature should not be greater than 30 °C?

(You may assume that the entire kinetic energy of the electrons is converted to heat in the target.) (O & C)

Chapter 3

Resistance

3.1 Resistance and conductance

When flow of current in a conducting device involves the irreversible conversion of electrical energy into heat, the device is said to have electrical resistance. A conductor, such as a coil of metal wire, which has been made up with the object of offering resistance to current, is a resistor.

The electrical resistance R of a conductor is defined as being the ratio of the p.d. V applied across it to the current I which flows in it,

i.e.
$$R = \frac{V}{I}. \tag{3.1}$$

Electrical conductance G is the reciprocal of resistance,

i.e.
$$G = \frac{1}{R} = \frac{I}{V}. \tag{3.2}$$

The practical units of resistance and conductance, i.e. the values obtained by putting potential difference V in volts and current I in amperes, are respectively, the ohm (Ω), and the ohm^{-1}.

It is worth noting the use of word endings with respect to these definitions, because they involve a convention which has now become general in electrical theory. Thus:

resistor is the name given to a particular electrical device. Other examples are conductor, capacitor.

resistance is a property of the device. Similarly, conductance, capacitance.

Also, resistivity, to be defined shortly (section 3.8), is a property of a material; e.g. the resistivity of copper is a property of copper, from which the resistance of any chosen specimen of copper wire can be deduced.

3.2 The heat released in a resistor

The rate at which energy is converted into heat in a resistor is given by equation 2.9, which we will re-label as equation 3.3,

$$P = VI. \tag{3.3}$$

Inserting the value $V = IR$ from equation 3.1, this becomes

$$P = I^2R, \tag{3.4}$$

where, using SI units, P is the rate at which heat is released, measured in watts.

Equation 3.4 provides an alternative definition of resistance, as being the ratio of the rate at which heat is released to the square of the current.

3.3 Linear and non-linear conductors

There is nothing in the above definitions and equations to imply that the resistance of any particular conductor is constant, i.e. that the current I is actually proportional to the p.d. V and, of course, there are many cases where it is not, two examples of which are given in Fig. 3.1. If the resistance of a particular device does turn out to be constant, it is called a linear conductor, because a graph of I against V is a straight line.

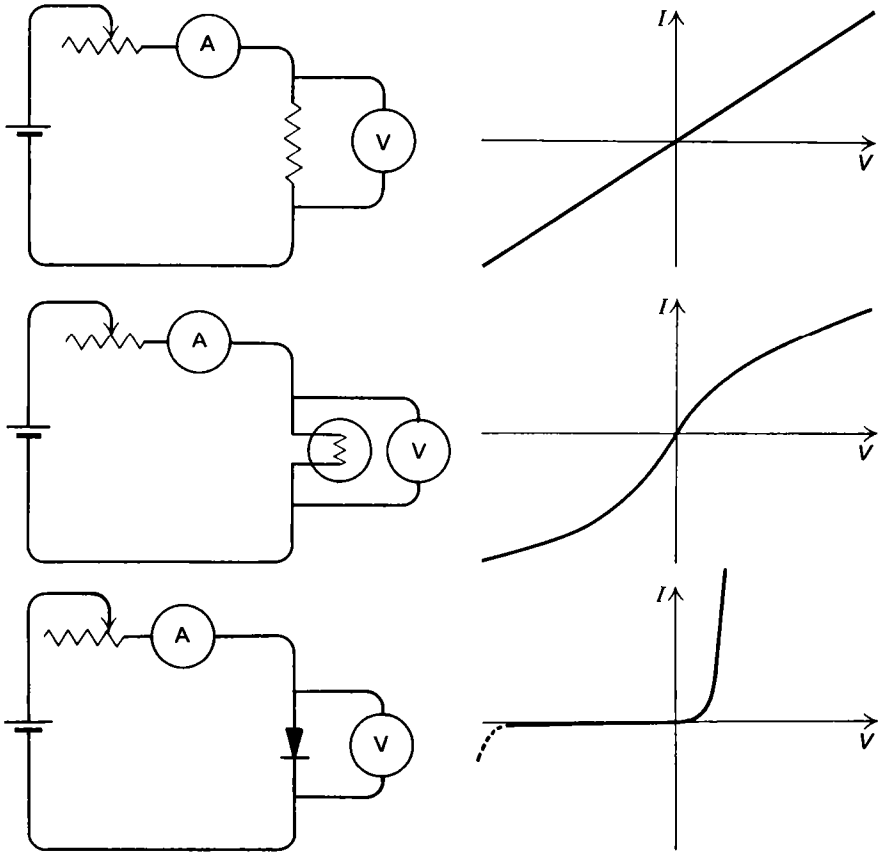


Fig. 3.1 Linear and non-linear conductors, (a) A metal wire at constant temperature is **a linear conductor**. Temperature variation makes little difference to the behaviour of a standard resistor made from special alloys (manganin or constantan) (b) A tungsten lamp filament is **non-linear**, due to the effect of changing temperature (c) A germanium or silicon diode is **non-linear** (In practice, in this case, the voltmeter might be better connected across both ammeter and diode)

Referring to the three cases shown in the diagram, a metal wire is a linear conductor, provided that its temperature is kept constant. For some alloys, from which standard resistors are made, even a quite considerable temperature change matters very little. A tungsten filament lamp, however, and a semiconductor diode, are non-linear. The temperature of the tungsten filament increases rapidly with increasing current, and this causes its resistance to go up. The diode offers low resistance to current flowing in the forward direction, but its resistance in the reverse direction (provided that the reverse p.d. is not too large) is very high indeed.

3.4 Ohm's law

This law, first stated by G. S. Ohm in 1826, refers specifically to a metallic conductor:

> 'The current flowing in a metallic conductor is proportional to the potential difference applied across it, provided that the temperature is constant.'

Alternatively, provided that resistance R is first defined, this may be more briefly stated, in the form:

> 'The resistance R of a metallic conductor is constant, provided that its temperature is kept constant.'

Clearly, by definition, any linear conductor is one which obeys Ohm's law.

It is stressed that, if we quote equation 3.1, we are giving the definition of resistance; we are not quoting Ohm's law.

3.5 Joule's law

This law, also, refers specifically to a metallic conductor. It follows directly from equation 3.4, for cases when the resistance is constant.

> 'The rate at which heat is released in a metallic conductor is proportional to the square of the current, provided that the temperature is kept constant.'

Ohm's law and Joule's law are important in practice because metallic conductors are in such frequent use – particularly so because with alloys a variation in temperature has but small effect. But they are based on experimental evidence, applying to a special class of materials, and lack the universal validity of equations 3.1 and 3.4, which are expressions of definitions.

The remainder of this chapter will largely be concerned with linear conductors.

3.6 Series and parallel connections

Series connections

When a number of resistors is connected in series, the current I is common to all, and the p.d. across the system is the sum of the p.d.s across the individual resistors:

$$V = V_1 + V_2 + \ldots V_n$$
$$= I(R_1 + R_2 + \ldots R_n),$$

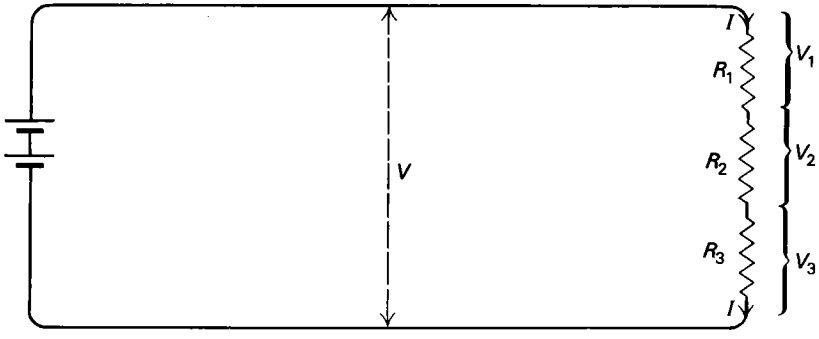


Fig. 3.2 Resistors in series and in parallel (a) Series Connection

whence the resistance R of the single equivalent resistor is

$$R = \frac{V}{I} = R_1 + R_2 + \ldots R_n. \tag{3.5}$$

This is, of course, the principle of the resistance box, in which standard resistances are made up by adding resistances put in series.

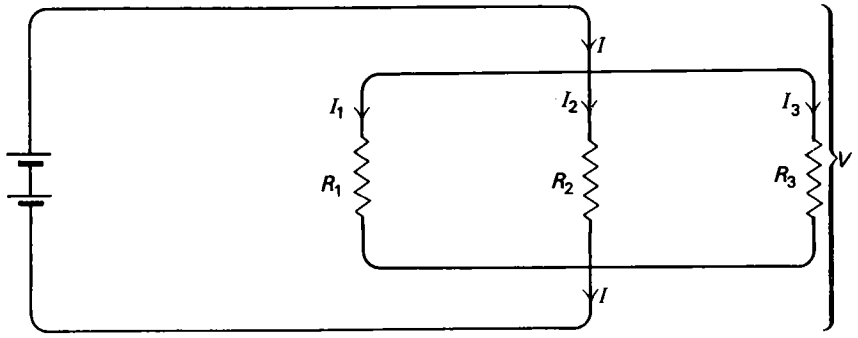


Fig. 3.2 Resistors in series and in parallel (b) Parallel connection

Parallel connections

When the resistors are connected in parallel, there is a single potential difference V common to all of them and the resultant current through the system is the sum of the currents through the individual resistors:

$$\begin{aligned} I &= I_1 + I_2 + \ldots I_n \\ &= V\left(\frac{1}{R_1} + \frac{1}{R_2} + \ldots + \frac{1}{R_n}\right), \end{aligned}$$

whence the conductance $G = 1/R$ of the single equivalent resistor is the sum of the individual conductances:

$$G = \frac{1}{R} = \frac{I}{V} = \frac{1}{R_1} + \frac{1}{R_2} + \ldots + \frac{1}{R_n}. \tag{3.6}$$

3.7 Connection of a variable resistor

The use of a variable resistor to control the current taken by a circuit is such common practice that it is worth a brief investigation. The obvious thing to do is to place the variable resistor in series with the load R_L. The current I is then given by the equation

$$I = \frac{V}{R + R_L}. \tag{3.7}$$

In practice, there are two drawbacks to this simple arrangement, resulting from the non-linear variation of current with the value of R. At the high current end of the range, the variable resistor becomes a very coarse control, the variation of I with R being rapid. At the other extreme, there is a minimum value for the current, corresponding to the effect of the whole of resistor R; to vary the current smoothly all the way down to zero would require a resistor of infinite size. Some methods of overcoming one or other of these defects are shown in Fig. 3.4.

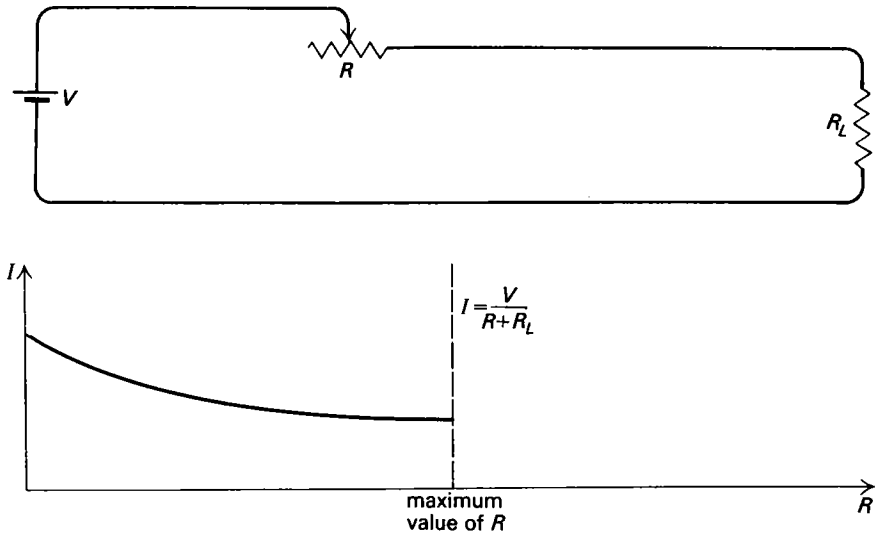


Fig. 3.3 Control of current by a series resistor (a) (b)

Coarse and fine control

A high resistance put in parallel with a low one is a fine control, because the low value resistor takes most of the current. When they are put in series, however, it is the low value resistor which is the fine control, because the high resistance takes most of the potential drop.

A potential divider

The variable resistor in Fig. 3.4(*b*) is placed across the supply, and a fraction of the p.d. then appears across the load. This arrangement has the advantage that the current can be varied gradually right down to zero.

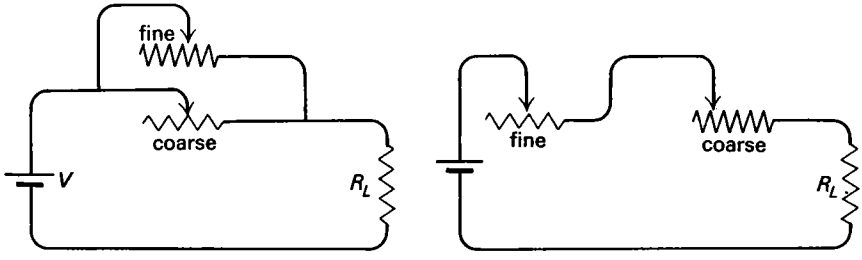


Fig. 3.4 (a)

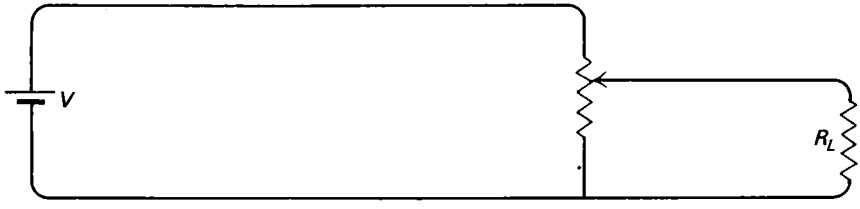


Fig. 3.4 (b)

Graded resistors

The non-linear behaviour of a simple variable resistor may be compensated by using a bank of resistors in series, suitably graded in value, as in Fig. 3.4(*c*).

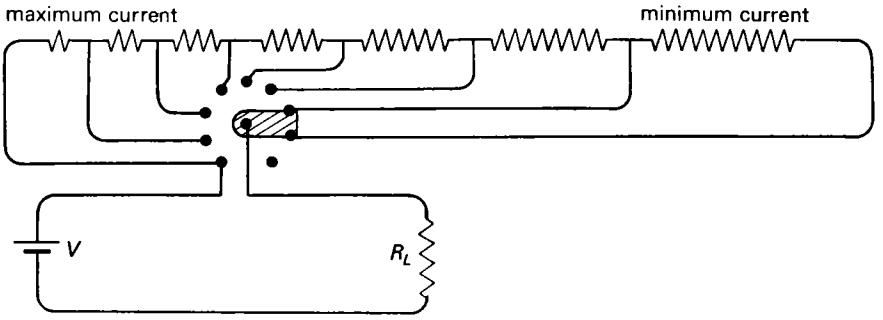


Fig. 3.4 (c)

Fig. 3.4 Improved arrangements for control of current (a) Coarse and fine control arrangements (b) A potential divider (c) Graded resistors

3.8 Resistivity

It follows from equations 3.5 and 3.6 that the resistance R of a uniform wire (provided that it obeys Ohm's law) is directly proportional to its length l and in inverse proportion to its cross-sectional area A. We can write

$$R = \rho \frac{l}{A} \tag{3.8}$$

where ρ is a property of the material from which the wire is made. This is the resistivity of the material. Its value varies with temperature in most cases.

The dimensions of resistivity are seen, from equation 3.8, to be those of resistance × length. The value for a material is commonly quoted in ohm metres, e.g. the resistivity of copper at 18 °C is about $1.6 \times 10^{-8}\ \Omega \cdot \text{m}$.

The reciprocal of resistivity, $1/\rho$, is the conductivity of the material. The practical unit is (ohm metre)$^{-1}$.

Perhaps it is worth stressing that resistivity is not resistance of unit volume. The resistance of a wire is not proportional to its volume, for whereas increase in length increases the resistance, increase in cross-sectional area reduces it.

3.9 The variation of resistance with temperature

In reviewing the effect of temperature change, we can classify solid materials into four groups: pure metals, alloys, semiconductors, and insulators.

The resistivity of a metallic element increases if the temperature is raised, being very approximately proportional to the Kelvin temperature. Some metals, e.g. lead (but not copper), lose their resistance entirely at extremely low temperatures a few degrees above the absolute zero. In this condition they are said to be superconducting.

The effect of adding foreign atoms to the pure metal is to raise its resistance in a way which is not very temperature dependent. Consequently, the resistivity of an alloy is higher than that of the pure metal but varies less with temperature. Even a small amount of impurity has an appreciable effect on resistance, so that the copper used for making conducting cables must be very pure.

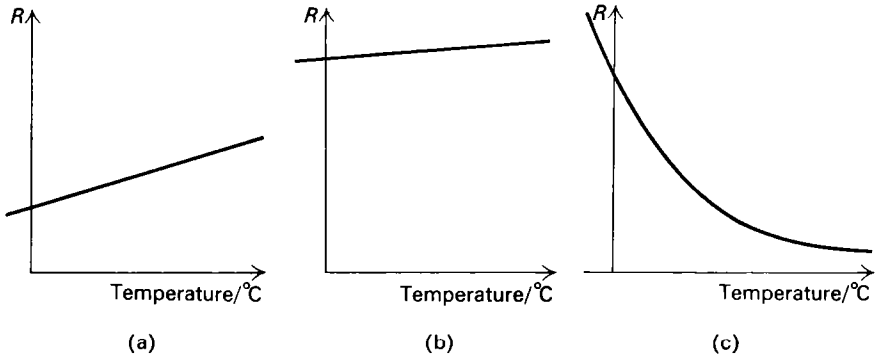


Fig. 3.5 Variation of resistance with temperature (a) A pure metal (b) An alloy (c) A semiconductor (a thermistor)

If a non-metallic solid conducts at all, its resistivity is usually much higher than that of a metal or an alloy, and it is classed as a semiconductor. In the general case of a semiconductor, a rise in temperature often results in an increase in the concentration, N, of free charge carriers, with the result that the resistivity falls with temperature rise, in many cases very markedly. Only occasional reference will be made in this chapter to semiconductors, which will be discussed more fully in chapter 11.

3.10 The temperature coefficient of resistance

In the case of a metal, the variation of resistance with temperature can be accurately represented by the equation

$$R = R_0(1 + \alpha\theta + \beta\theta^2) \tag{3.9}$$

where R_0 = resistance at 0 °C,
R = resistance at temperature θ,
θ = temperature, measured in °C,

and α, β are constants.

The variation is very nearly a linear one, as illustrated in the graphs in the diagram, which means that the constant β is very much smaller than α and the term $\beta\theta^2$ may be neglected, to a first approximation; in this case, the equation takes the simpler form:

$$R = R_0(1 + \alpha\theta) \tag{3.10}$$

or

$$\alpha = \frac{R - R_0}{R_0\theta} \tag{3.11}$$

The constant α, defined by these equations, is called the temperature coefficient of resistance, i.e. the fractional increase in resistance per unit rise in temperature, taking the original value R_0 at 0 °C. It may, however, be inconvenient, in practice, actually to take the temperature down to 0 °C, in which case the resistance values R_1, R_2 may be measured, instead, at temperatures θ_1 and θ_2. Writing out equation 3.11 twice, and eliminating R_0, gives

$$\frac{R_2}{R_1} = \frac{1 + \alpha\theta_2}{1 + \alpha\theta_1}. \tag{3.12}$$

3.11 Practical aspects of temperature effects

The increase in resistance of a metal with temperature rise is put to practical advantage in the resistance thermometer. This can be a very accurate instrument,

if allowance is made for the slight departure from linearity by taking into account the effect of the constant β in equation 3.9. One advantage of such an instrument is the large range of temperatures for which it can be used.

Special semiconducting devices known as thermistors are also useful for temperature measurement, the advantage being high sensitivity, since the resistance of a thermistor falls rapidly when the temperature rises. The scale, in this case, is far from linear.

More often, however, variation in resistance with temperature is likely to be a disadvantage. This is obviously the case when a standard resistor is required, or an accurate shunt is designed for use in an ammeter. In standard resistors, an alloy is chosen which has a very low temperature coefficient, e.g. manganin, or constantan. Manganin is rather the better of the two, having some extra advantages: when it is joined to copper, no e.m.f. appears at the junction, as is apt to happen with most metals; also, the resistance of a manganin wire remains accurately constant for an indefinite time.

Two other devices concerned with temperature rise are electric heater elements, and lamp filaments. In neither case is exact resistance value important, so that variation with temperature is irrelevant. What matters in the case of the heating element is that it must not oxidise, since it is no part of the operation to burn the element. The alloy used for this purpose is nichrome. The lamp filament will not oxidise because it is in an atmosphere of inert gas. Tungsten is used because it will withstand very high temperature without melting; the inert gas is present to reduce the tendency of the very hot tungsten to evaporate.

*3.12 A physical interpretation of Ohm's law and the effect of temperature change

The connection between the current flowing in a conductor and the average drift speed v of the conduction electrons was derived in equation 2.11:

$$I = NAve, \tag{3.13}$$

where N is the number of conduction electrons in unit volume each having charge e, and A is the area of cross-section of the conductor.

To arrive at some sort of physical basis for Ohm's law, what is now required is a corresponding expression relating this drift speed v to the p.d. V which has been applied across the conductor. The effect of this p.d., in the simple case of a straight length l of the conductor, is to produce an electric field E in the conductor, where E is equal to the potential gradient V/l. (This relation will be derived in section 6.10.)

$$E = \frac{V}{l}. \tag{3.14}$$

A conduction electron experiences the effect of this field as a force, $F = Ee$, pulling the electron in the direction of the potential gradient towards the positive end of the conductor,

i.e.

$$F = \frac{V}{l}e. \tag{3.15}$$

If the electron were in a vacuum, the effect of this force would be a continuous acceleration along the whole of the electron's path, and it would gain a very considerable speed in consequence, as the electrons do in the vacuum of a cathode ray tube or an X-ray tube. But the conduction electrons in a metal are differently situated. Each such electron is moving about in the metal, in a random manner,

analogous to the motion of the molecules in a gas, and its drift in the electric field is superimposed on this random motion, rather like the drift of gas molecules blown along in a breeze. When the electrons collide with each other, or with atoms of the metal, they do not lose their random motion, because they collide elastically. Nor would they lose their directional drift either, were the atoms all perfectly arranged in their crystal lattices. What does destroy this directional motion, however, is any sort of irregularity in the crystal, due, perhaps, to a dislocated atom, the presence of a foreign impurity atom, or merely because the thermal vibrations of some atoms are considerably higher than the average in amplitude. The result is that an electron does not gather directional speed from the electric field for long before coming up against some such irregularity, whereupon it loses its speed again, and has to build it up a second time. Consequently, the average drift speed v is only small, being effectively the same throughout the length of the conductor. If the applied p.d. V is increased, thus increasing the strength of the accelerating field, the average drift speed is larger, and so the current increases.

To treat the problem more mathematically, we can write

$$v = \tfrac{1}{2}at, \tag{3.16}$$

where t is the average time interval in which the electron can gather speed in the field, a is the acceleration, and the speed acquired in time t is $2v$, being twice the average value.

But $F = m_e a$, where m_e is the mass of the electron,

and $F = \dfrac{V}{l} \cdot e$ (equation 3.15).

Hence we can substitute in equation (3.16) the value

$$a = \frac{V}{l} \cdot \frac{e}{m_e} \tag{3.17}$$

to give

$$v = \frac{1}{2}\frac{Ve}{lm_e}t. \tag{3.18}$$

Singling out the p.d. V in this equation, we have

$$V = \frac{2lm_e}{et}v, \tag{3.19}$$

which is the relation between V and v which we aimed to derive. Combining equations 3.13 and 3.19, we have

$$R = \frac{V}{I} = \frac{2m_e}{Ne^2t}\frac{l}{A} \tag{3.20}$$

i.e. provided that all the factors inside the bracket can be taken to be constant, the resistance R of the conductor is constant and is proportional to the factor l/A.

The only quantities inside the bracket which might be variables are N and t. N is the concentration of free charge carrying electrons in the conductor, and this is effectively constant in a metal, even if the temperature changes. The time interval t depends, not on the directional drift speed v, but on the speed of the random motion of the electrons, and we can write

$$t = \frac{x}{\bar{c}}, \tag{3.21}$$

where x is the mean distance travelled between collisions with some lattice irregularity, and $\bar{c}$ is a measure of the average random speed. Thus, t is likely to be

a constant quantity also, at any rate if the temperature is constant. Hence we would expect a metal at constant temperature to have a constant resistance.

The above theory has included the suggestion that the intrusion of foreign atoms into the crystal lattices of the pure metal would increase the irregularities which destroy the directional drift of the electrons, i.e. they would reduce the mean free distance x. For this reason, the resistivity of an alloy is greater than that of the pure element.

The effect of temperature rise

To account for the effect of temperature rise, we need to know a little more about the effect of the thermal vibrations of the atoms in the crystal lattice.

We can assume the equilibrium separation of these atoms to be represented by a diagram of the kind shown in Fig. 3.6; if the atoms move too far apart they experience a mutual attraction, and if brought too close together they experience

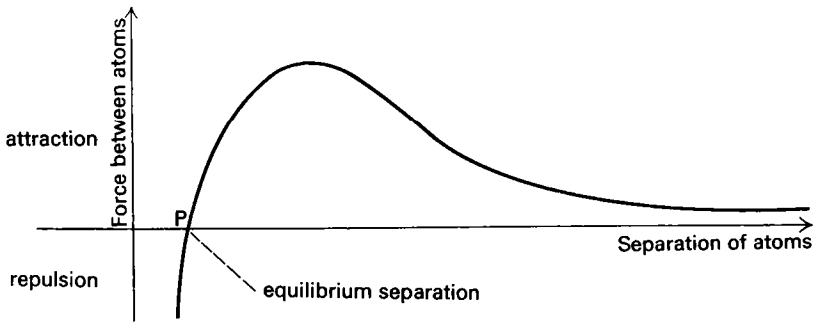


Fig. 3.6 The force between neighbouring atoms in a crystal

repulsion, so that the equilibrium separation is represented by the point P on the graph, and their thermal vibration might be expected to take the form of oscillation about this point. The forces between neighbouring atoms, however, do not vary linearly with their separation, and the graph where it passes through P is not a straight line; the atoms find it easier to come further apart than to approach closer together. The result of this is that atoms which vibrate with large amplitude have a mean separation which is larger than that of their less energetic neighbours. Since not all the atoms have the same energy of vibration, this causes irregularities in the lattice.

If the temperature of the metal is raised, thus increasing the vibrations of the atoms, this effect is enhanced, putting more irregularities in the crystal. With reference to equations 3.21 and 3.20, this represents a reduction in the free path x and hence a reduction in the value of t; the electrons have less time in which to acquire their drift velocity, and the current is reduced, i.e. the resistance is increased.

The same temperature effect does not occur with the irregularities introduced by foreign atoms into an alloy, because these are all irregularities in the first place, and temperature rise cannot further increase their number, and so the resistance of the alloy increases proportionally less with temperature rise.

In many cases of a semiconductor, the above effect is swamped by the extra effect of temperature rise in producing a great increase in the concentration of charge carriers N, with the net result that the resistance goes down.

3.13 Varying the range of a meter

The remainder of this chapter will be taken up with practical aspects of electrical resistance, and these will be dealt with by considering specific examples.

We start with the problem of varying the range over which an ammeter or a voltmeter can be used, which is done by putting suitable resistors in parallel or in series with the instrument. The instrument most commonly used to measure current and p.d. is basically the same in both cases. This is a sensitive current-measuring instrument, i.e. a milliammeter, or a microammeter, which has been adapted to serve the particular special purpose.

3.14 The ammeter

A milliammeter is converted to an ammeter by connecting a low resistance shunt, i.e. a resistor in parallel with it. This allows a proportion of the current to bypass the meter, thus lowering its sensitivity, and incidentally has the desirable effect of lowering the resistance of the meter.

Example

A milliammeter reads 0–15 mA, *i.e. full scale deflection (f.s.d.) is obtained with a current of* 15 mA. *The resistance of the meter is* 5.00 Ω. *How can this be converted for use as an ammeter, having f.s.d. with a current of* 1.5 A, *and what is the resistance of the converted meter?*

When a current of 1.5 A = 1 500 mA enters the instrument, it is required that

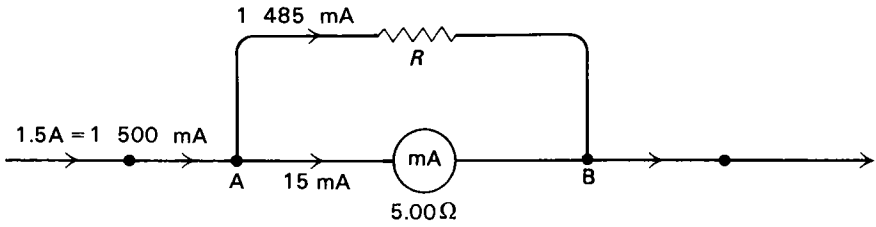


Fig. 3.7 Conversion of milliammeter for use as an ammeter

the meter shall register full scale deflection, i.e. 15 mA must flow through the meter. Hence, the shunt must take the remaining 1 485 mA.

Applying Ohm's law:

$$V_{AB} = 15 \times 5 = 75 \text{ mV}$$
$$V_{AB} = 1\,485 \times R,$$

where R is the resistance of the shunt. Hence, the required value R for the shunt is given by

$$R = \frac{15 \times 5}{1\,485}\,\Omega$$
$$= \frac{5}{99}\,\Omega$$
$$= 0.0505\ \Omega.$$

The resistance of the complete meter, including both milliammeter and shunt, is

$$R' = \frac{V}{I}$$

$$= \frac{75}{1500}\ \Omega$$

$$= 0.0500\ \Omega.$$

The ideal ammeter would have zero resistance, because it is connected in a circuit in series; if it has any resistance at all, this must to some extent disturb the circuit.

The importance of shunt connections

An ammeter which is designed to work over a single range is provided with a shunt having soldered connections. A multirange instrument, however, requires a

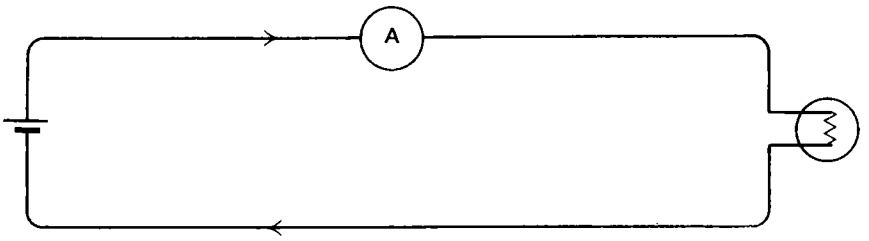


Fig. 3.8 Connection of an ammeter in a circuit. The resistance of the ideal ammeter is zero

choice between interchangeable shunts; to alter the range, the selected shunt is changed, either by means of a switch, or by the use of terminals. It will be realised that, as in the example above, the resistance of the shunt is very low. A bad connection at switch or terminals would, therefore, result in a large error. The switch must be carefully designed; terminals require a careful experimenter, to see that they are clean, and tightly connected.

A swamping resistor

Variation of resistance with temperature also operates against the accuracy of a shunt. The shunt, because it carries most of the current, is likely to rise in temperature, and it is therefore made of manganin, so that its resistance shall remain constant. But it is not practicable to make the winding of the milliammeter of

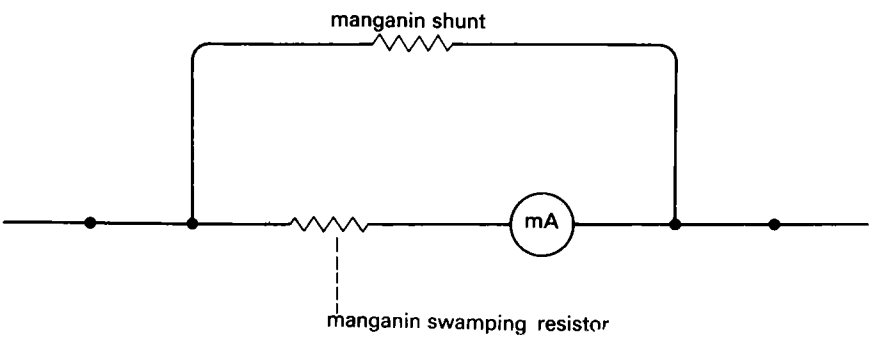


Fig. 3.9 A swamping resistor

manganin also, and this consists of copper. Should the temperature of the whole instrument rise, this would alter the resistance of the milliammeter coil, without changing the resistance of the manganin shunt, and an error would result. Some ammeters have a manganin coil connected in series with the copper coil, to reduce its effective temperature coefficient. This increases the resistance of the meter, of course, which is a disadvantage, but its accuracy is less affected by temperature changes. This extra coil is called a swamping resistor.

3.15 The voltmeter

The type of voltmeter most commonly used is a milliammeter, or a microammeter, with a high value resistor connected in series.

Example 1

A milliammeter reads 0–15 mA. *Its resistance is* 5.00 Ω. *How can this be converted for use as a voltmeter, having full scale deflection with a p.d. equal to* 15 V, *and what is the resistance of the converted meter?*

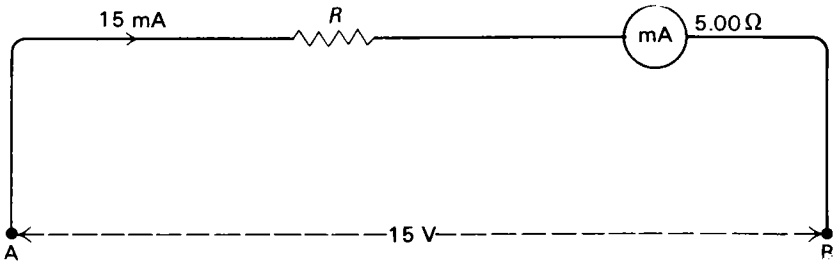


Fig. 3.10 Conversion of milliammeter for use as a voltmeter

When a p.d. of 15 V is applied across the terminals AB, it is required that the meter shall register full scale deflection, i.e. 15 mA must flow through the meter. Hence, the resistance of the converted instrument must be equal to R', where

$$\frac{15}{R'} = 15 \times 10^{-3}\ \Omega^{-1},$$

$$\therefore\quad R' = 1000\ \Omega.$$

Since the resistance of the milliammeter is 5.0 Ω, a further resistance R must be connected in series, where

$$R = 995\ \Omega.$$

The ideal voltmeter would have infinite resistance, because it is connected in parallel with the component across which the p.d. is to be measured, and this should not provide an alternative path for the current.

Example 2

A p.d. of 15 V *is applied across two resistors in series, having resistances of* 500 Ω *and* 1000 Ω *respectively. Deduce the value of the p.d. across each separate resistor. A voltmeter, having resistance* 1000 Ω, *is now connected across each of the two resistors in turn. What reading will the voltmeter register, in each case?*

(*a*) *In the absence of the meter*

The current I is common to both resistors,

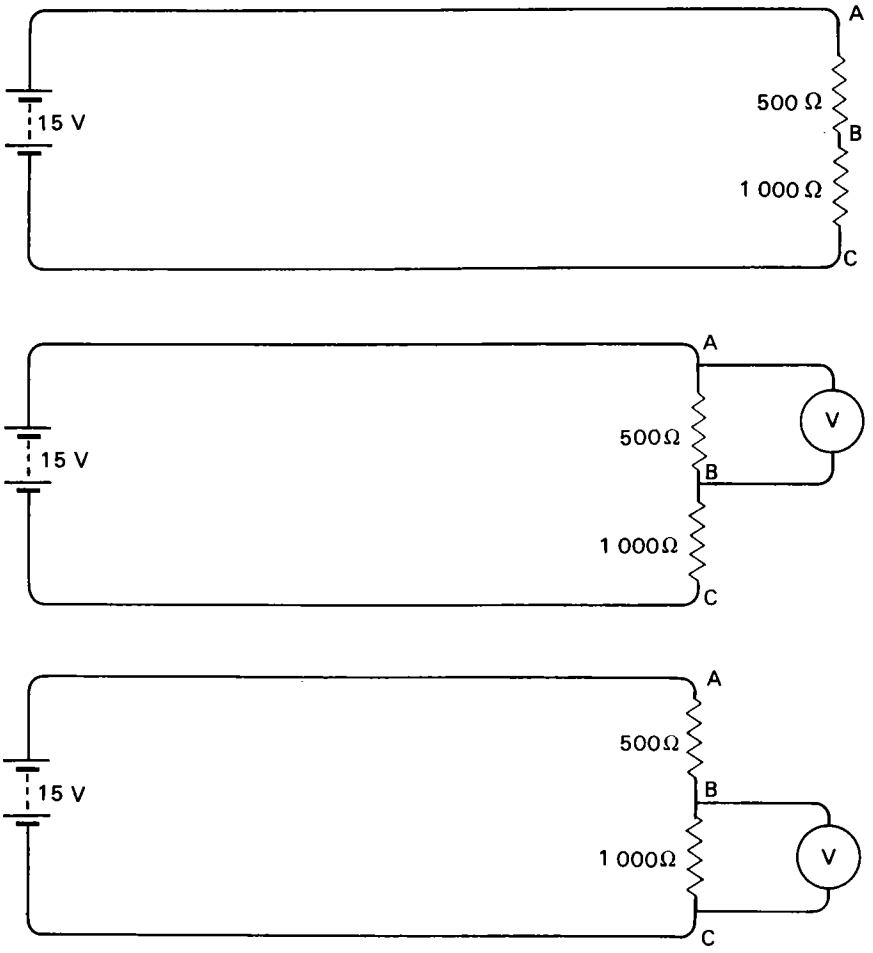


Fig. 3.11 (a) (b) (c)

i.e.

$$V_{AB} = I \times 500,$$

$$V_{BC} = I \times 1000,$$

and

$$V_{AB} + V_{BC} = V_{AC} = 15 \text{ V}.$$

$$\therefore \quad I = \frac{15}{1500} = \frac{1}{100}\text{A}$$

$$\therefore \quad V_{AB} = 5 \text{ V}$$

and

$$V_{BC} = 10 \text{ V}.$$

(*b*) *When the voltmeter is connected across AB*

Because the meter is not perfect, it puts its resistance, 1 000 Ω, in parallel with the 500 Ω resistor. The effective resistance between A and B is given by

$$\frac{1}{R_{AB}} = \frac{1}{500} + \frac{1}{1000}\,\Omega^{-1},$$

$$\therefore \quad R_{AB} = \frac{1000}{3}\,\Omega$$

$$\text{and } R_{BC} = 1000\,\Omega.$$

As in (*a*), the applied p.d. is divided up in the proportion of the two resistances,

i.e.
$$\frac{V_{AB}}{V_{BC}} = \frac{R_{AB}}{R_{BC}} = \frac{1}{3},$$

where
$$V_{AB} + V_{AC} = 15\text{ V},$$

$$V_{AB} = \frac{15}{4} = 3.75\text{ V},$$

$$V_{BC} = \frac{15.3}{4} = 11.25\text{ V},$$

and the voltmeter, when connected across AB, reads 3.75 V, instead of the value, 5 V, which existed before it was connected.

(*c*) *When the voltmeter is connected across BC*

This puts the voltmeter resistance, 1000 Ω, in parallel with the 1000 Ω resistor, which reduces the effective resistance between B and C to

$$R_{BC} = 500\,\Omega.$$

The circuit is now equivalent to two equal resistors in series,

$$R_{AB} = 500\,\Omega,$$
$$R_{BC} = 500\,\Omega,$$

and the p.d. is divided equally between them.

$$\therefore \quad V_{BC} = 7.5\text{ V}$$

and this is the value registered by the voltmeter, instead of the value, 10 V, which existed before it was connected.

This large discrepancy, between the two p.d.s originally present, and the values registered on the voltmeter, arises because the resistances in the circuit are large ones, commensurate with the resistance of the voltmeter. The discrepancy would not have been so startling if the resistors AB and BC had been smaller, e.g. 5 Ω and 10 Ω.

The example worked out above illustrates the limitation of the commonest type of voltmeter, which is basically a milliammeter or a microammeter, and stresses the point that such an instrument must have as high a resistance as possible. Reconsideration of the calculation in example 1, which opened this discussion of the voltmeter, would show that the final resistance of the instrument depends on the sensitivity of the basic meter from which it is developed. If this were a microammeter, for example, reading 0–150 μA, instead of the milliammeter taken in the example, the resistance of a voltmeter reading 0–15 V would be 100000 Ω, specified as being 6700 Ω/*V*, which is a better instrument.

3.16 Inclusion of ammeter and voltmeter in a circuit

When both ammeter and voltmeter are included in a circuit, there is a problem about the best way in which to connect them, arising from the fact that both instruments fall short of the ideal. The two possibilities are shown in Fig. 3.12.

In the first case, the voltmeter correctly reads the p.d. across the component X, but the ammeter reading includes the small current taken by the voltmeter, in addition to the current which it is intended to read, i.e. that which flows in the component X.

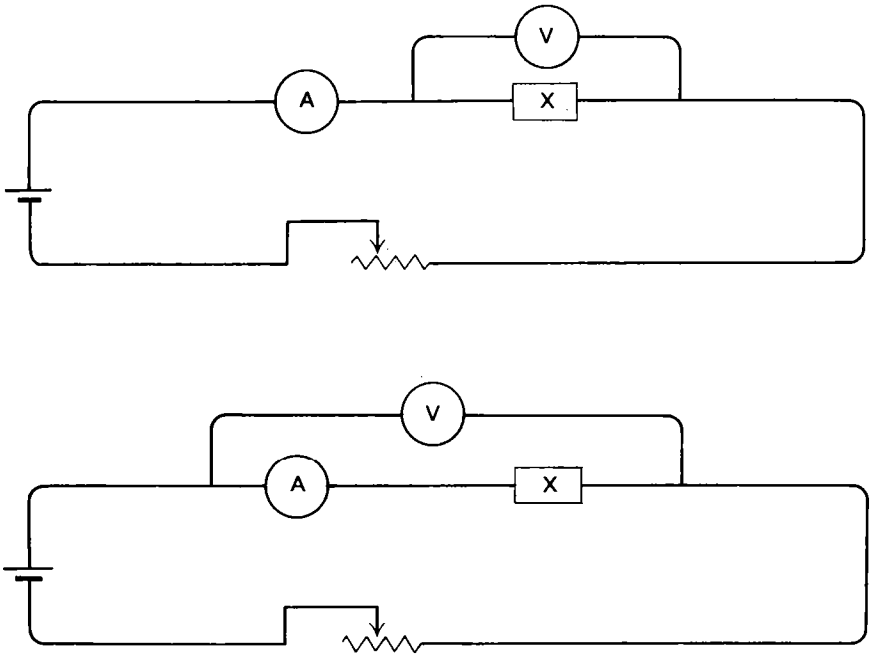


Fig. 3.12 Alternative ways to connect ammeter and voltmeter (a) Preferable when the resistance of X is low (b) Preferable when the resistance of X is high

In the second case, the ammeter correctly reads only the current taken by the component, but the voltmeter reading includes the small p.d. across the ammeter, in addition to the p.d. across X.

The first arrangement is preferable if the resistance of the voltmeter is much greater than the resistance of X, whereas, if the voltmeter resistance is by comparison low, the second circuit is better. But it is impossible, with this type of instrument, to be absolutely right, because the practical ammeter has some resistance, although slight, and the most common types of voltmeter draw a small current, whereas ideally they would take none.

3.17 An ohm-meter

The milliammeter can also be adapted to the approximate measurement of resistance. This is the principle of the ohm-meter, the circuit of which is shown in Fig. 3.13. The resistance to be measured is connected across the terminals X, Y, and its value is judged by the current which flows; clearly, if its value is zero, the current is maximum, and, if its value is infinitely large, the current is zero. The meter may be graduated to read ohms directly, but its scale is far from linear, as the calibration curve indicates, being very closed up at the high resistance end of the scale.

It is clear that the deflection depends, not only on the resistance R under test,

but also on the e.m.f. of the cell which provides the current. Variations in this e.m.f. may be approximately compensated by adjustment of the variable resistor R_S, connected in parallel with the meter. Altogether, however, the ohm-meter is not an accurate instrument, and has little merit other than convenience.

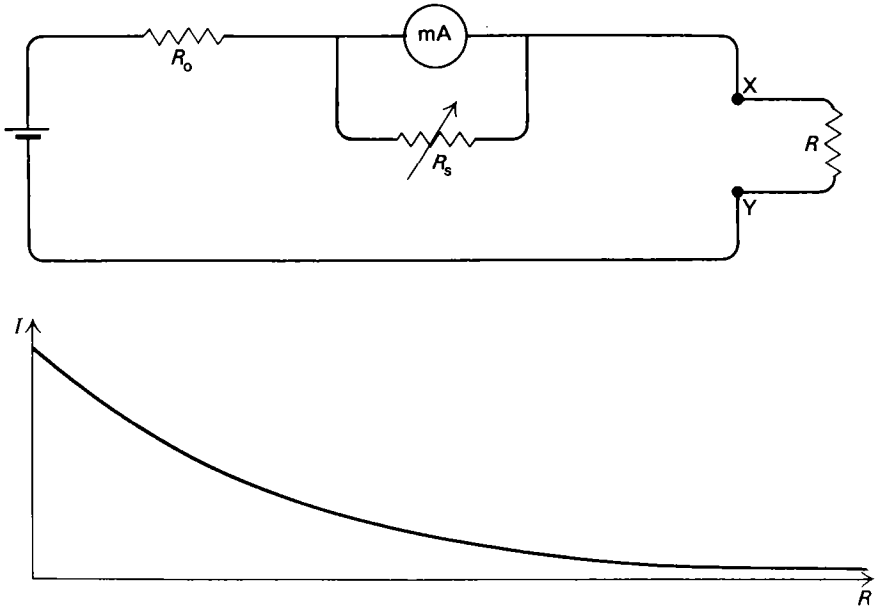


Fig. 3.13 An ohm-meter (a) The ohm-meter circuit (b) Calibration curve

3.18 Power loss in cables

Example

A consumer accepts electrical power from a distant source at 100 kW, *the resistance of the supply cables being equal to* 0.2 Ω.

(*a*) *If the p.d. at the consumer's end is equal to* 10 kV, *calculate the total potential drop in the cables and the percentage power loss in them.*

(*b*) *Repeat the calculation, assuming that the consumer accepts the same power, but at a lower p.d., equal to* 1 kV.

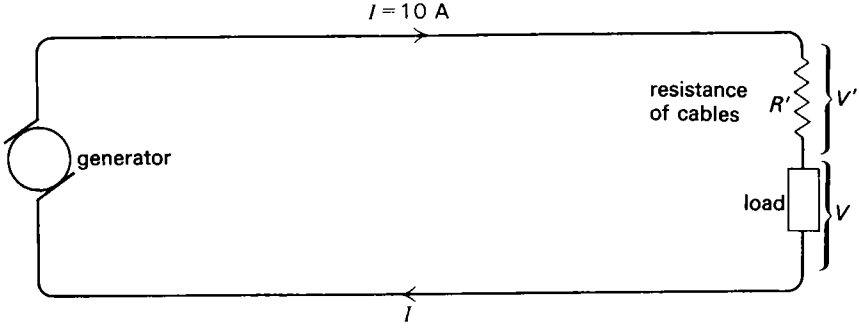


Fig. 3.14 (a) Power delivered at 10^4V

(*a*) The current I is given by

$$I = \frac{P}{V}$$
$$= \frac{10^5}{10^4}\,\text{A}$$
$$= 10\ \text{A}.$$

Therefore the total potential drop in the cables is

$$V' = IR',$$

where R' is the resistance of the cables.

i.e.
$$V' = 10 \times 0.2\ \text{V}$$
$$= 2\ \text{V}$$

and the power lost in the cables is

$$P' = V'I$$
$$= I^2R'$$
$$= 100 \times 0.2\ \text{W}$$
$$= 20\ \text{W},$$

i.e. the power loss is 0.02%.

(*b*) The same power is now being delivered at a lower potential difference, and so the current must be proportionally greater.

We have

$$I = \frac{P}{V}$$
$$= \frac{10^5}{10^3}\,\text{A}$$
$$= 100\ \text{A}.$$

The total potential drop in the cables is therefore

$$V' = IR'$$
$$= 100 \times 0.2\ \text{V}$$
$$= 20\ \text{V}$$

and the power lost in the cables is

$$P' = V'I$$
$$= I^2R'$$
$$= 10^4 \times 0.2\ \text{W}$$
$$= 2000\ \text{W},$$

i.e. the power loss is now 2%.

The conclusion to be drawn from this calculation is evident: when power is to be transmitted on a large scale, this must be done at high potential, in order to reduce the loss in power due to the current in the transmission lines. The reason for this may be summarised as follows.

The power P' which is lost in the cables is

$$P' = I^2R'.$$

If this is to be kept relatively low, the current must not be too large, since there is a limit to what can be done in reducing the resistance R' of the cables.

The power P delivered to the load, however, is

$$P = VI$$

where V is the p.d. at which the power is delivered. Since I has to be small, V must be large in compensation.

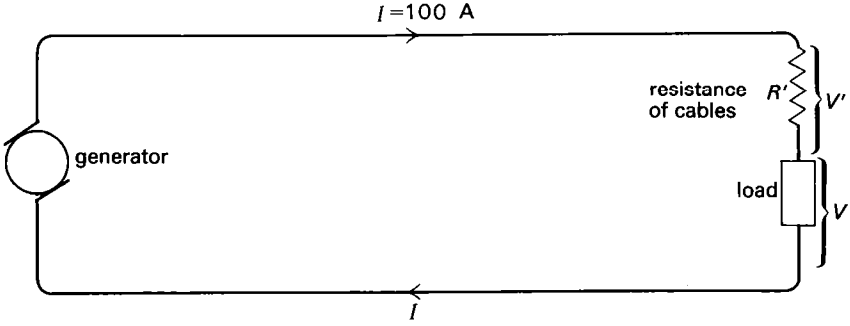


Fig. 3.14 (b) Power delivered at 10^3V

It might, perhaps, be thought that it is impossible in practice to adjust the value of the p.d. V in this way; if the resistance of the load has the fixed value R, the p.d. across it should be given by the equation

$$V = IR$$

which suggests that the current I cannot be reduced without the p.d. also being reduced, thus greatly reducing the power delivered.

This difficulty can be overcome by using a transformer. A transformer, inserted between the cables and the load, has the effect of changing the apparent resistance. A suitable transformer, for example, can deliver a current of 400 A to a load at a p.d. of 250 V, while accepting the corresponding power from the cables as a current of 10 A at a p.d. of 10^4 V. This necessitates using alternating current, because a transformer is an alternating current device.

The system used to carry electric power across country, therefore, is to generate alternating current at the most convenient p.d., and to deliver it to the consumer at a low p.d., but to have a very high p.d. between the transmission lines, a transformer being inserted at either end.

The upper limit is reached for the potential which can be used in a transmission

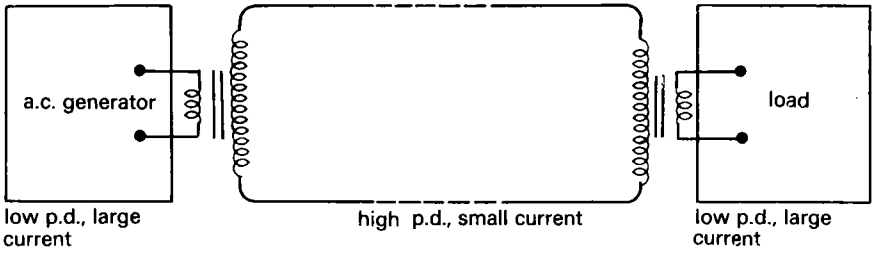


Fig. 3.15 The transmission of electric power

line when the insulation of the air begins to break down, and the power loss consequently becomes serious after all. At such high potentials, the electric field strength close to the cable becomes very high, and intermittent corona discharges through the air take place, especially at points where there are small irregularities in the surface of the cable. Corona discharge is a more serious problem with alternating current, because in this case, to maintain an alternating potential difference of high average value, the peak value reached by the potential difference in the course of each alternation must be even higher. At the present time, some countries are developing power transmission systems which use potentials higher than 400 kV, and the cables then carry direct current, in order to reduce the tendency to corona discharge; the power is still generated and utilised in alternating form in order that the advantage of transformers can still be available, but it is converted into direct current form and back again at the two ends by using mercury vapour rectifiers and inverters (or by using a solid state equivalent, called a thyristor).

There is a further case when the power must be carried as direct current, and this is when the cables are to be laid under water, as with the submarine cables which provide a power link between France and Britain. The reason in this case is the electrical capacitance which exists between the cable and the surrounding water, which causes heavy power loss if alternating current is used. (The special significance of capacitance with respect to alternating current is discussed in section 9.8.) Consequently, in such circumstances as this, too, the current is changed from alternating to direct current to carry it under the sea, and then converted back again.

3.19 Potential drop in house wiring

The objection to power loss in the wiring system of a house is not the cost of the wasted power, but the fact that this power turns into heat, which may damage the insulation, or even cause arcing at switches and sockets which were not constructed to carry the large current. The p.d. 240 V adopted for domestic use in Britain represents a reasonable compromise between economy and safety. A 2000 W electric heater constructed for use at 400 V would take a current of 5 A, which is only half the current required by a 2000 W, 200 V heater. The wiring for 400 V installation could therefore have thinner conductors, but the potential would be much more dangerous.

Low voltage installation in a laboratory

It would be a convenient arrangement, and it is sometimes adopted, to have a low potential source, e.g. 24 V, which can supply currents up to 10 or 20 A to various parts of a laboratory. But if the resistance of the leads is, say, 0.10 Ω, the potential drop in them due to a current of 10 A would be 1 V, leaving only 23 V at the terminals; consequently the terminal p.d. would vary with the current supplied, a serious disadvantage when accuracy is required.

3.20 Potential drop in a cell

Example 1

In the circuit shown in Fig. 3.16, *the voltmeter and ammeter may be assumed to behave as ideal instruments, having no effect upon the current and p.d.s in the circuit, and the cell has zero resistance.*

When the switch S is open, the voltmeter reading is 6.0 V. *The switch is now closed,*

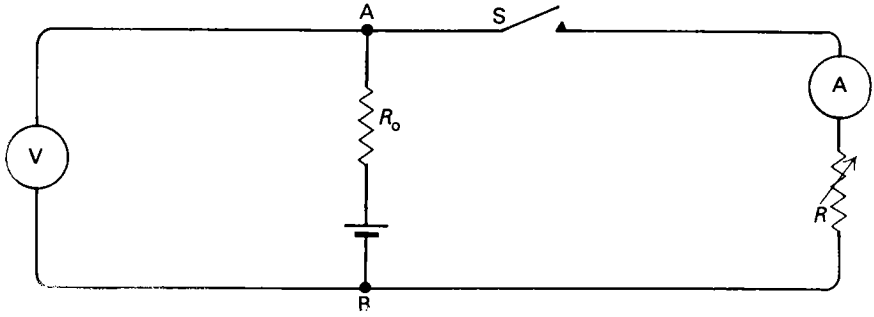


Fig. 3.16

and the value of the resistor R is adjusted so that the ammeter reading is 2.0 A. *If* $R_0 = 2.5\ \Omega$, *what is the new reading of the voltmeter?*

What is the maximum current which can be made to flow through the ammeter, and what will the voltmeter read in this case?

When the switch is open, the current is zero, and so no current flows in resistor R_0, and the p.d. across it is zero. The p.d. across terminals AB, which is 6.0 V, is therefore equal to the e.m.f. E of the battery.

When the switch is closed, the current flowing in resistor R_0 is 2.0 A, therefore the p.d. across this resistor is V_0 where

$$\begin{aligned} V_0 &= IR_0 \\ &= 5.0\ \text{V}. \end{aligned}$$

Hence the p.d. V, across the terminals AB, which is measured by the voltmeter, is

$$\begin{aligned} V &= E - V_0 \\ &= 6.0 - 5.0\ \text{V} \\ &= 1.0\ \text{V} \qquad \text{(i)} \end{aligned}$$

The maximum current available flows when the value of R has been reduced to zero, the only resistance remaining in the circuit being that of resistor R_0, and we have

$$\begin{aligned} I_{\max} &= \frac{E}{R_0} \\ &= \frac{6.0}{2.5}\ \text{A} \\ &= 2.4\ \text{A}. \qquad \text{(ii)} \end{aligned}$$

In this case, the p.d. across the terminals AB is

$$\begin{aligned} V &= IR \\ &= 0 \\ \text{or}\quad V &= E - IR_0 \\ &= 6 - 2.4 \times 2.5\ \text{V} \\ &= 0. \end{aligned}$$

Example 2

In the circuit shown in Fig. 3.17, *the voltmeter and ammeter may be assumed to behave as ideal instruments.*

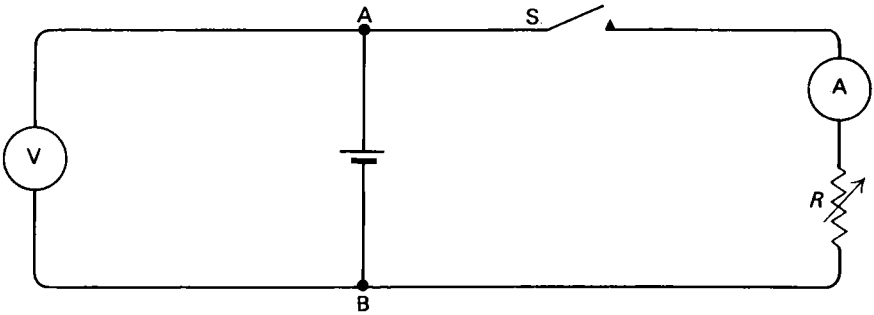


Fig. 3.17

When the switch S *is open, the voltmeter reading is* 6.0 V. *The switch is now closed and the value of the resistor R is adjusted so that the ammeter reading is* 2.0 A. *The reading of the voltmeter is now* 1.0 V. *Account for this fall in p.d.*

The values given in this problem, and the readings of the voltmeter, are identical with those in the first problem, yet the resistor R_0 is missing from the circuit. This can be explained on the basis that this resistance, $R_0 = 2.5\ \Omega$, although not belonging to an extra visible component, is still present in the circuit, being the internal resistance of the battery. It is reasonable to suppose that the ions which carry the current through the battery will experience some resistance to their passage, and in this particular case this must be equal to 2.5 Ω.

This is a general result, although the resistance in the chosen example is an artificially high one, and it applies to any form of generator, whether a cell or a dynamo; i.e. as soon as current is taken from the generator, the p.d. across its output terminals falls to a value which is lower than its e.m.f., due to the loss of some p.d. in driving current through the generator. We can write

$$V = E - V_0, \tag{3.22}$$

where the lost p.d. is

$$V_0 = IR_0,$$

R_0 being the internal resistance,

$$\therefore \quad V = E - IR_0. \tag{3.23}$$

If the internal resistance is constant (which is not always the case in practice) this equation is represented by the graph in Fig. 3.18.

3.21 The definition of e.m.f.

Referring to section 2.6, a more satisfactory definition of e.m.f. can now be given.

The e.m.f. of a generator represents the energy converted into or from electrical form for every unit of charge which flows through it, this energy conversion being reversible. The practical unit is the volt. The e.m.f. of the generator is equal in value to the p.d. across its terminals, provided that no current is flowing.

Since, as we have seen, the types of voltmeter most commonly used are, in reality, high resistance milliammeters, they must draw a small current – the better the voltmeter the smaller the current. It follows that voltmeters of this type, when connected across a cell, register a p.d. which is, at best, slightly less than the value of the e.m.f.

The lead accumulator

The internal resistance of a lead accumulator is very small. This means that the effect of a potential loss in the cell is very slight, and often passes unnoticed, which

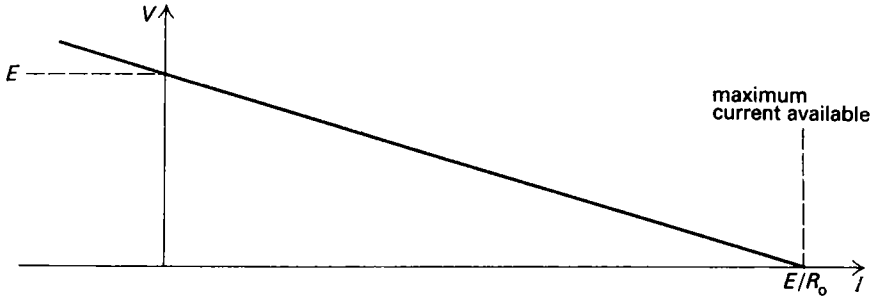


Fig. 3.18 Internal resistance. Variation in the p.d. across a generator with the value of the current, assuming that the internal resistance R_0 is constant

can be very useful in elementary experiments. It also means that a short circuit does great damage to the accumulator, because the short circuit current is very high.

3.22 The measurement of internal resistance

When a resistor R is connected across a cell which has e.m.f. E and internal resistance R_0, we can write three equations giving the value of the current, viz.:

$$I = \frac{E}{R+R_0}, \tag{3.24}$$

$$I = \frac{V}{R}, \tag{3.25}$$

where V is the p.d. across the resistor R, which is the p.d. across the cell,

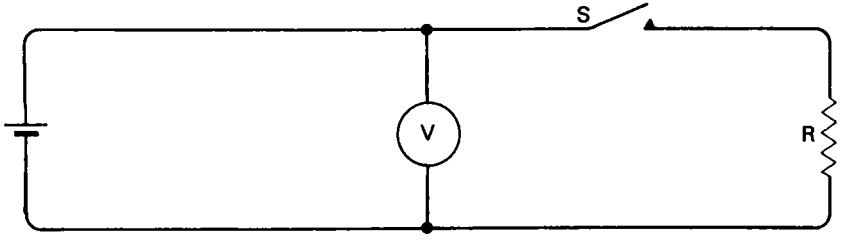


Fig. 3.19 (a)

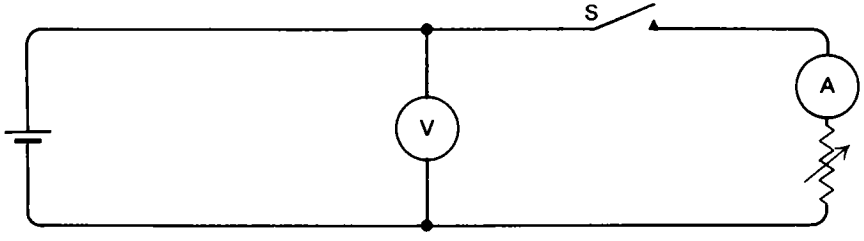


Fig. 3.19 (b)

Fig. 3.19 The measurement of internal resistance (a) Using voltmeter and standard resistor R. (b) Using voltmeter and ammeter

and
$$I = \frac{E - V}{R_0}, \tag{3.26}$$

where $(E - V)$ is the p.d. lost in the cell.

To measure the internal resistance, two of the three variables I, V, R, must be determined. For example, using the circuit in Fig. 3.19 (*a*), where R is a standard resistor, we have, from equations 3.25 and 3.26

$$\frac{R_0}{R} = \frac{E - V}{V}, \tag{3.27}$$

where E is equal to the reading of the voltmeter (assumed to draw negligible current) when switch S is open.

Alternatively, using the circuit in Fig. 3.19 (*b*),

$$R_0 = \frac{E - V}{I}. \tag{3.28}$$

3.23 A matched load

Example

A source having e.m.f. 220 V *and internal resistance* 400 Ω *supplies current to a load which has resistance R.*

Calculate the p.d. across the load resistor, and the current and power delivered to it, in each of the three cases, $R = 20\ \Omega$, 400 Ω, 8000 Ω. *Comment on the results.*

The required equations are:

$$I = \frac{E}{R + R_0}, \tag{3.24}$$

$$V = IR, \tag{3.25}$$

and
$$P = VI = I^2R = \frac{E^2R}{(R + R_0)^2}. \tag{3.29}$$

Putting in each of the values given above, in turn, we have:

(*a*) $R = 20\ \Omega$.

$$I = \frac{200}{420} = 0.476\ \text{A},$$
$$V = 0.476 \times 20 = 9.52\ \text{V},$$
$$P = 9.52 \times 0.476 = 4.53\ \text{W}.$$

(*b*) $R = 400\ \Omega$.

$$I = \frac{200}{800} = 0.25\ \text{A},$$
$$V = 0.25 \times 400 = 100\ \text{V},$$
$$P = 100 \times 0.25 = 25\ \text{W}.$$

(*c*) $R = 8000\ \Omega$.

$$I = \frac{200}{8400} = 0.0238\ \text{A},$$
$$V = 0.0238 \times 8000 = 190\ \text{V},$$
$$P = 190 \times 0.0238 = 4.52\ \text{W}.$$

Comments

The significance to be attached to these results depends on the purpose for which the source is connected across the load.

If the object is to apply the maximum possible p.d. across the load, as might be the case, for example, if the load is the signal input of an oscilloscope, then the resistance of the load must be as high as possible.

If the object is to supply the maximum possible power to the load, then the best value for its resistance, out of the three possible values quoted in the question, is 400 Ω. If equation (3.29) were used to plot values of the power delivered to the load against corresponding values of its resistance R, a curve would be obtained having the form shown in Fig. 3.20. The power delivered passes through a maximum when the resistance of the load and that of the source are equal, in which case the two are said to be matched. This is the requirement when supplying a loudspeaker from an amplifier.

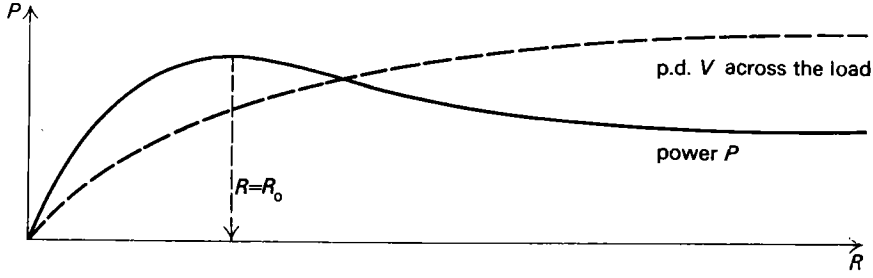


Fig. 3.20 Load matching. The power P delivered to a load resistor, plotted against values of the load resistance R.

3.24 The potential distribution in circuits

When tracing the values of the p.d. in various parts of a circuit, it is useful to bear in mind the analogous case of liquid flow. Moving along a stream in the direction of the water flow, we drop in level, i.e. in gravitational potential, and moving against the flow, we rise. Similarly, moving with the current through a resistor means a fall in potential, and moving against it means a rise. The e.m.f. of a cell included in the circuit raises or lowers the potential, dependent on the polarity of the cell.

Finally, if we follow a complete closed path, returning to the starting point, the net change in potential is zero. These considerations are really applications of conservation of energy.

Some examples follow:

Example 1

If the cells and resistors have the values marked in the circuit diagram, deduce the potential of the point P.

First, we deduce the value of the current I and label its direction, as shown in the diagram.

$$\Sigma I = \frac{E}{R}$$

$$= \frac{(3+2)\ \text{V}}{(6+1+4+5+2+2)\ \Omega}$$

$$= \frac{5}{20}\text{ A.}$$

$$= \frac{1}{4}\text{ A.}$$

Proceeding clockwise from E to P involves moving in the direction of the current, and so the potential falls as we go through each resistor.

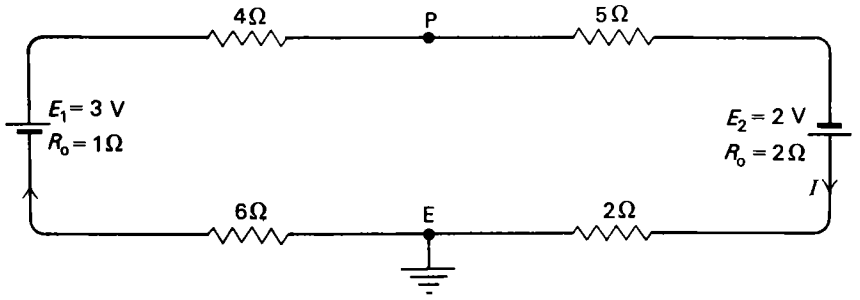


Fig. 3.21

The total potential drop in the resistors between E and P is equal to

$$\frac{1}{4} \times (6 + 1 + 4)\text{ V} = \frac{11}{4}\text{ V.}$$

The cell E_1 raises the potential by the amount

$$E_1 = 3\text{ V.}$$

Therefore the potential of P is

$$V_P = \left(3 - \frac{11}{4}\right)\text{ V}$$

$$= +0.25\text{ V}$$

As a check, suppose we proceed anticlockwise from E to P. This involves moving against the current in the resistors, and so with respect to p.d. across the resistors, the potential of P is higher than that of E, by the amount

$$\frac{1}{4} \times (2 + 2 + 5)\text{ V} = \frac{9}{4}\text{ V}$$

The cell E_2 on the other hand, is so connected that it lowers the potential, by the amount

$$E_2 = -2\text{ V.}$$

Therefore, the potential of P is

$$V_P = \left(-2 + \frac{9}{4}\right)\text{ V}$$

$$= +0.25\text{ V}$$

Example 2

A p.d. equal to 100 V *is applied between the two points* A,B *of the network, which consists of five resistors, each having resistance R equal to* 1 000 Ω.

Find the value of the potential difference V_{CF}, between the points C,F, *and the potential difference V_{DF}, between the points* D *and* F.

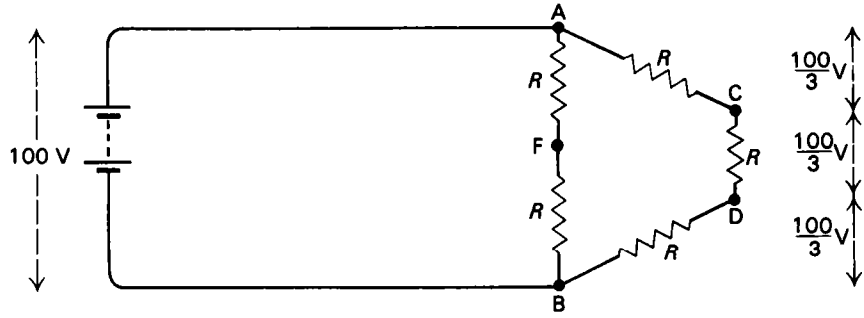


Fig. 3.22

The total potential drop in the three resistors AC, CD, DB is 100 V. Since these resistors are connected in series, all take the same current; since their resistances are equal, the p.d. across each of them is the same.

Hence $$V_{AC} = V_{CD} = V_{DB} = \frac{100}{3}\text{ V},$$

and $$V_{CB} = \frac{200}{3}\text{ V},$$

$$V_{DB} = \frac{100}{3}\text{ V}.$$

Similarly, V_{FB} is equal to half the p.d. between A and B,

i.e. $$V_{FB} = 50\text{ V},$$

Therefore, we have

$$\begin{aligned} V_{CF} &= V_{CB} - V_{FB} \\ &= \frac{200}{3} - 50\text{ V} \\ &= \frac{50}{3}\text{ V} \\ &= 16.7\text{ V}, \end{aligned}$$

and

$$\begin{aligned} V_{DF} &= V_{DB} - V_{FB} \\ &= \frac{100}{3} - 50\text{ V} \\ &= \frac{-50}{3}\text{ V} \\ &= -16.7\text{ V}. \end{aligned}$$

Example 3

In the circuit shown below, deduce the value of the p.d. across the 5.0 Ω *resistor.*

We first label the current in each of the three branches, putting an arrow to

represent the direction which we are assuming each current to follow. Remembering that, in the matter of current, what comes in goes out, it is only necessary to have two unknown quantities, for if the currents through the cells are respectively I_1, I_2, the current in the resistor must be equal to $(I_1 + I_2)$.

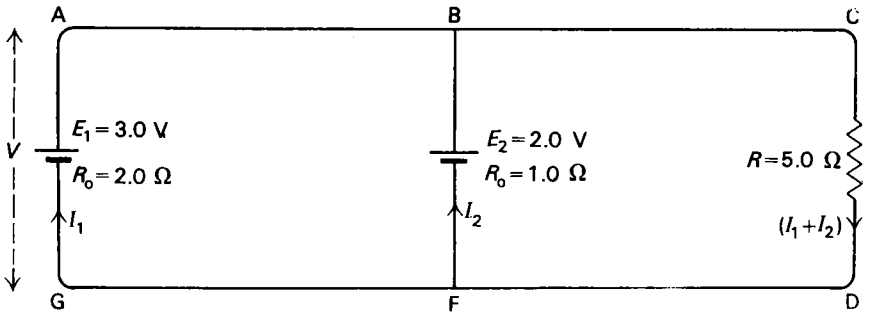


Fig. 3.23

We now have

$$V_{AG} = V_{BF} = V_{CD} = V,$$
$$\therefore \quad V = V_{AG} = 3.0 - 2.0I_1,$$
$$V = V_{BF} = 2.0 - 1.0I_2,$$
$$V = V_{CD} = 5.0(I_1 + I_2).$$

Eliminating I_1, I_2 between these equations gives

$$V = \frac{35}{17}\text{ V}$$
$$= 2.06\text{ V}.$$

3.25 Kirchhoff's laws

The last problem was solved by using certain physical principles, which are summarised, more formally, as Kirchhoff's laws. These laws apply to networks consisting of linear conductors and sources of e.m.f., through which steady currents are flowing.

Kirchhoff's laws are:

1 The algebraic sum of the currents which flow into any point in a network is zero.

$$\Sigma I = 0.$$

2 In any closed path taken round the network, the algebraic sum of all products of current I and resistance R is equal to the algebraic sum of the e.m.f.s included in the chosen path.

$$\Sigma IR = \Sigma E.$$

The first law is an expression of the principle of conservation of electric charge, i.e. charge cannot disappear, nor can it appear from nowhere.

The second law is an expression of conservation of energy. Any unit of charge taken round a closed path must have gained from the sources of e.m.f. the same amount of energy as it loses in resistors.

We can now solve the last problem again, using these two laws. This approach differs only a little from the method used previously, but it is more readily applied to solving problems of greater difficulty.

Applying Kirchhoff's laws to the circuit of Fig. 3.23, we have:

Law 1 $\Sigma I = 0$.

i.e.
$$I_1 + I_2 - (I_1 + I_2) = 0.$$

Nothing is gained here, because the first law has been used already, in the process of labelling currents in the circuit diagram. This is the usual way to proceed, since the number of unknowns is reduced in consequence.

Law 2 $\Sigma IR = \Sigma E$ in any closed path.

(*a*) Following the path CDFBC:
$$\Sigma IR = (I_1 + I_2) \times 5 + I_2 \times 1,$$
$$\Sigma E = 2.0,$$
$$\therefore \quad 2.0 = 5.0\, I_1 + 6.0\, I_2.$$

(*b*) Following the path CDGAC:
$$\Sigma IR = (I_1 + I_2) \times 5 + I_1 \times 2,$$
$$\Sigma E = 3.0,$$
$$\therefore \quad 3.0 = 7.0\, I_1 + 5.0\, I_2.$$

(*c*) Following the path ABFGA:
$$\Sigma IR = -I_2 \times 1 + I_1 \times 2,$$
$$\Sigma E = -2.0 + 3.0,$$
$$\therefore \quad 1.0 = 2.0\, I_1 - I_2.$$

Any pair of these three equations may be used to determine the values of I_1, I_2, whence we can deduce the value
$$V = 5.0(I_1 + I_2)$$
$$= 2.06 \text{ V}.$$

*3.26 Symmetrical networks

Kirchhoff's laws can be used to provide the solution for quite complicated networks, although in practice the algebra may become quite involved. If there is symmetry, however, this may be used to reduce the number of unknown quantities; provided that these can be reduced to four or five or less at the start, the problem is not too difficult.

Example

A cubical wire framework is constructed, the resistance of each wire edge of the cube being R. Deduce the resistance of the network, measured between two corners of the cube which are diagonally opposite to each other.

The problem is to determine the current I_1 which enters and leaves the network, and the p.d. V between the input and output terminals, both in terms of the current I flowing in one branch of the network. The ratio, $R_1 = V/I_1$, is the resistance required by the question. If the unknown current I is cancelled out in taking this ratio, the problem has been solved. Symmetry has been used, in addition to Kirchhoff's first law, in labelling currents on the diagram. Thus, the three wires

AB, AF, AD are symmetrically placed. If the current in AB is equal to I, it follows that the currents in AF and AD are also equal to I. The wire JH, bringing current to the output terminal H, corresponds in every way to the wire AB, taking current from the input terminal A. Hence the current in JH is equal to I, and flows towards the point H. Similarly, for the currents in CH, GH.

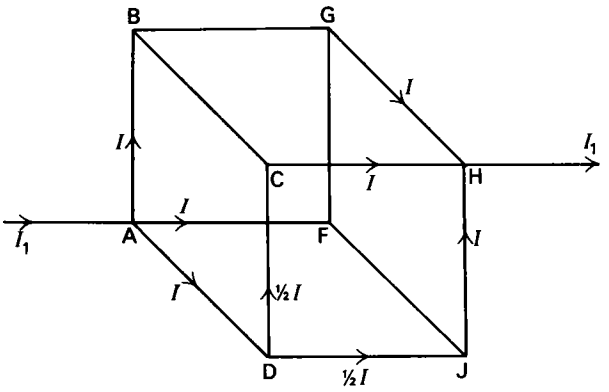


Fig. 3.24

Now we consider the divisions of the current at the point D. The two paths DC, DJ are symmetrically placed, so the current reaching the point D from A must divide equally, and the current in each of the wires DC, DJ is equal to $\frac{1}{2}I$.

We can now solve the problem. We have

$$I_1 = 3I,$$

$$V = V_{AD} + V_{DJ} + V_{JH}$$

$$= IR + \tfrac{1}{2}IR + IR$$

$$= \frac{5}{2}IR.$$

Therefore the resistance of the network is R_1 where

$$R_1 = \frac{V}{I}$$

$$= \frac{5}{2}IR \div 3I,$$

i.e.

$$R_1 = \frac{5}{6}R.$$

QUESTIONS

3.1 (*a*) An electric hot plate for use at 240 V has two heating elements in it, each of resistance 80 Ω. A switch is arranged to provide low, medium, and

high positions by connecting the two resistors in series, or in parallel, or by putting only one resistor in circuit, as the case may be. Calculate the power available in the hot plate in each of the three cases.

(*b*) A 2.0 V lead accumulator is connected across two resistors, R_1 and R_2, in series. The resistors are then connected in parallel instead. Compare the rate at which heat is produced in one resistor with the rate at which heat is produced in the other in each arrangement.

$$R_1 = 4\ \Omega;\ R_2 = 8\ \Omega.$$

(*c*) An electric heater is connected to a 13 A socket by copper leads. When the current is switched on, the heater coil gets hot, but the socket and the leads remain cold. Explain this.

(*d*) What is the resistance, when the lamp is lit up, of

(i) a 240 V, 60 W lamp,
(ii) a 12 V, 60 W lamp?

Why would you choose the second type if a point source of light is required, and why would a good connection for the leads be more important in the second case?

3.2 The values of the resistors in this network are $R_1 = 1.0\ \Omega$, $R_2 = 4.0\ \Omega$. If the resistance measured between A and B is 4.0 Ω, deduce the value of R_3.

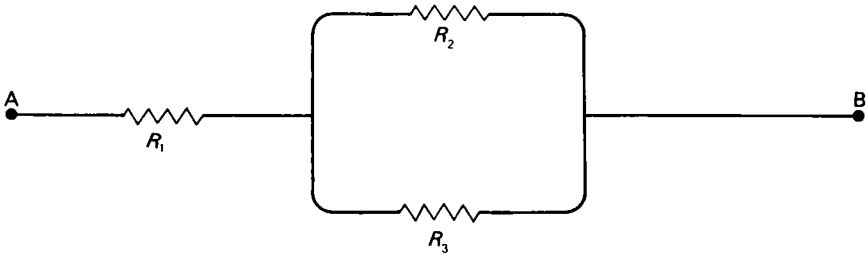


Fig. e 5

3.3 The current flowing in the resistor R has the same value when switches S_1 and S_2 are both closed as it has when they are both open. What is the value of R?

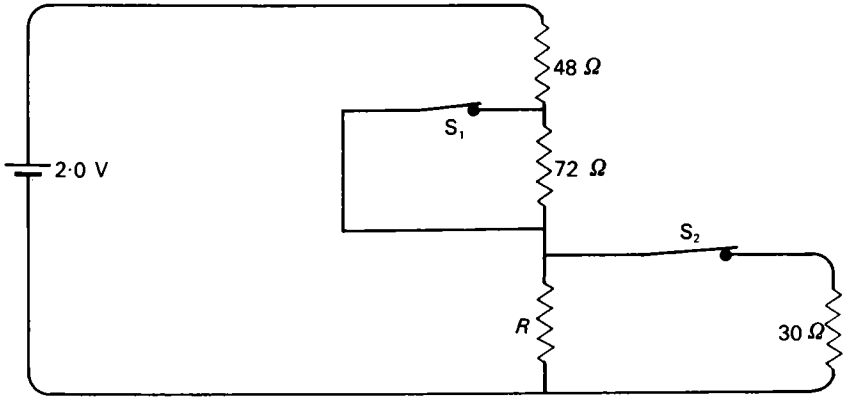


Fig. e 6

3.4 The values of the variable resistors R_1 and R_2 are initially adjusted to be 2.0 Ω and 20.0 Ω respectively. If R_1 is now increased in value by 1.0 Ω, calculate the amount by which R_2 must be altered in compensation, if the current flowing in resistor R is to remain unchanged. Explain the application of this principle in providing a fine control of current in a circuit.

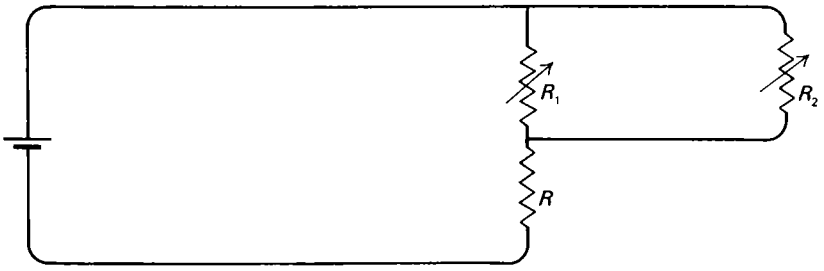


Fig. e 7

3.5 If the resistances of the two cells in the circuit may be neglected, deduce the value of the potential at the point P (*a*) when switch S is open, (*b*) when the switch is closed.

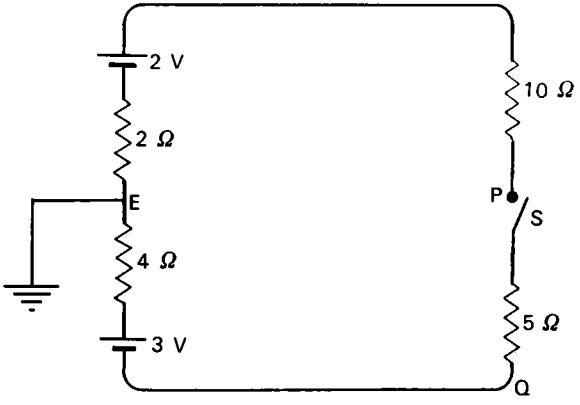


Fig. e 8

What would these two values become (*c*) if a further good earth connection were made at the point Q, (*d*) if the earth connection at E were then removed, leaving the one at Q?

3.6 The diagram e 9a represents a simple transistor circuit. The potential at A is

$$V_A = +9.0 \text{ V}$$

and the currents flowing in the resistors are, respectively,

$$I_C = 1.5 \text{ mA}$$

$$\text{and} \quad I_B = 8.0\ \mu\text{A}.$$

Deduce the value of resistor R_1 if the potential at B is $V_B = 0.7$ V, and deduce the value of the potential at C if $R_2 = 3000$ Ω.

3.7 In the transistor circuit shown in the diagram the potential at A is

$$V_A = +9.0 \text{ V}$$

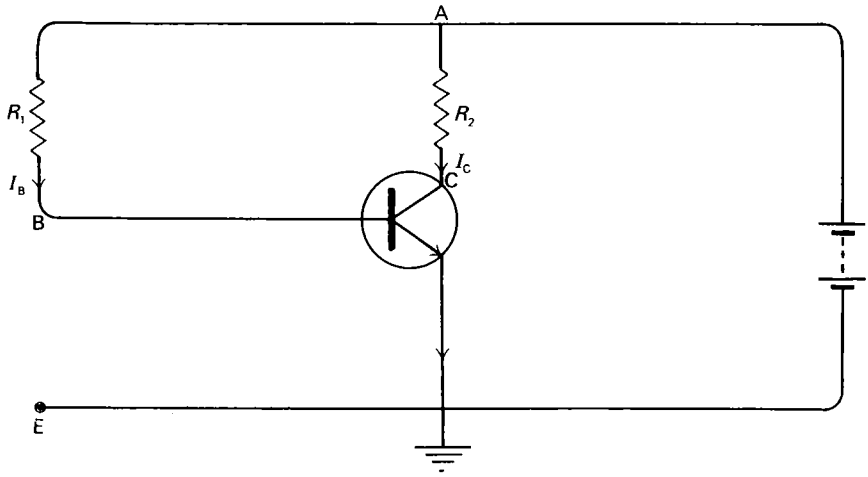


Fig. e 9a

and the currents are

$$I_c = I_e = 0.5 \text{ mA},$$
$$I_1 = 50\ \mu\text{A}$$

(*a*) Given that $R_3 = 8000\ \Omega$, and $R_4 = 2000\ \Omega$, deduce the values of the potential at c and at e.

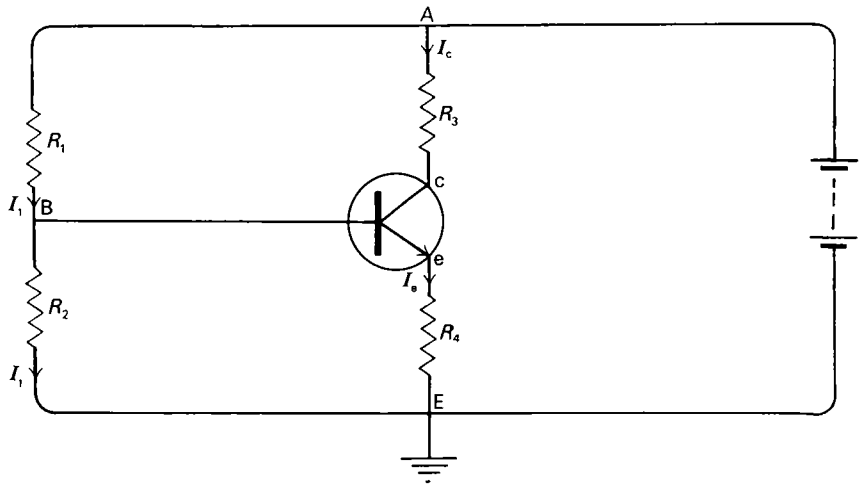


Fig. e 9b

(*b*) Given that the p.d. between B and e is $V_{Be} = 0.70$ V, deduce the values of R_1 and R_2.

3.8 In the triode circuit shown in the diagram,

$$I_A = I_C = 5.0 \text{ mA},$$

and the potential at B is

$$V_B = 200 \text{ V}.$$

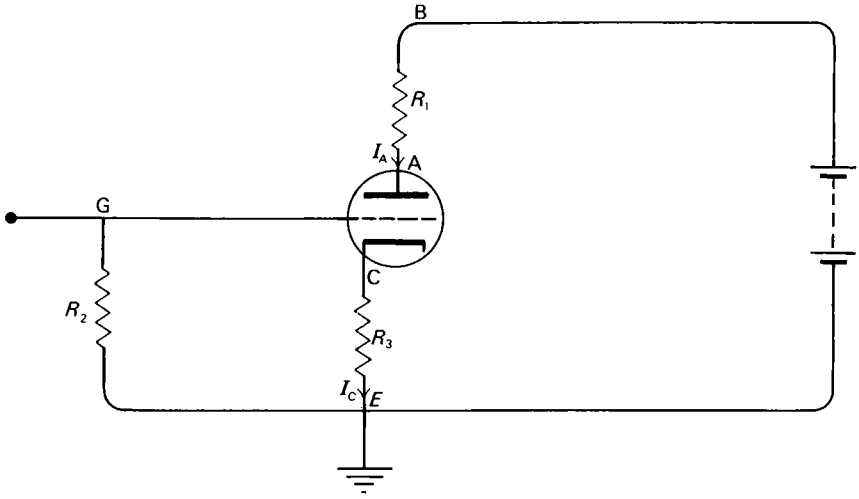


Fig. e 9c

(*a*) If the value of R_1 is $R_1 = 15000\ \Omega$, deduce the value of the potential at A.

(*b*) If the potential at C is to be

$$V_C = +1.5\ \text{V},$$

what is the value of the resistor R_3?

(*c*) If no current flows in resistor R_2, what is the potential at G relative to the potential at C?

3.9 AB, BC, CD, DA are, respectively, resistances of 10 Ω, 10 Ω, 100 Ω, and 101 Ω, connected in a network ABCDA. A p.d. of 2.01 V is maintained between points A and C, with the positive pole at A. Calculate the p.d. between B and D. (O & C)

3.10 A distant telephone, the resistance of which is 300 Ω, is connected to the exchange, 10 km away, by a pair of telephone wires whose resistance, singly, is 5 Ω/km. A twig falls across the wires at a certain point, producing the effect of a resistance R connected between the wires at that point. Measurements made at the exchange show a resistance between the terminals of the line of 130 Ω, which rises to 160 Ω when the distant telephone is disconnected. How far from the exchange is the fault to be found, and what is the value of R? (O & C)

3.11 The wire used in making up a heating element has resistivity $5.0 \times 10^{-7}\ \Omega\cdot\text{m}$, and cross-sectional area 0.8 mm^2. Calculate the length of wire required to make up a 240 V, 1200 W element.

If there are 8×10^{28} conduction electrons in every m^3 of wire, deduce the mean drift speed of the electrons in the wire when the 240 V p.d. is applied (assuming this to be d.c.).

3.12 Copper contains 10^{29} 'free' electrons per m^3, and its resistivity is 1.72×10^{-8} Ω·m. A p.d. of 10 mV is set up between two points 10 cm apart in a uniform copper wire. Calculate the average drift velocity with which electrons will move through the wire. (O & C).

3.13 A long metal rod, area of cross-section 1.50 cm^2, carries a current of 10.0 A. Leads are clipped to points on the rod 20 cm apart, and connected to a galvanometer with a 960 Ω resistor in series. The resistance of the galvanometer is 40 Ω, and it is found to register a current of 2 μA.

Calculate the resistivity of the metal. (C)

3.14 (*a*) A length of constantan wire, radius 0.6 mm, is reduced to one third

of its original resistance by coating it with an even layer of copper. Calculate the thickness of the copper. The resistivity of constantan is $49 \times 10^{-8}\ \Omega\cdot\text{m}$, and that of copper is $1.7 \times 10^{-8}\ \Omega\cdot\text{m}$.

(*b*) If a copper and an aluminium wire have equal lengths and equal resistances, what is the ratio of their masses?

(The resistivity of aluminium $= 3.2 \times 10^{-8}\ \Omega\cdot\text{m}$; of copper $= 1.8 \times 10^{-8}\ \Omega\cdot\text{m}$. The density of aluminium $= 2.70\ \text{gm/cm}^3$; of copper $= 8.9\ \text{gm/cm}^3$).

3.15 A cable consists of six straight steel wires, each of diameter 2.0 mm, surrounding a central core of aluminium wire, diameter 5.0 mm. When the aluminium is carrying a current of 8.0 A, calculate the current flowing in the steel.

(The resistivities of aluminium and steel are respectively $3.2 \times 10^{-8}\ \Omega\cdot\text{m}$ and $12.2 \times 10^{-8}\ \Omega\cdot\text{m}$.)

3.16 A resistance thermometer has resistance 10.54 Ω when placed in melting ice, and 14.69 Ω when placed in condensing steam at normal atmospheric pressure.

Calculate (*a*) the temperature coefficient of resistance of the metal of the coil, (*b*) the temperature of a liquid in which the thermometer has a resistance of 11.30 Ω, (*c*) the minimum detectable change of temperature if the least change of resistance that can be detected is 0.01 Ω. (C)

(Assume that the resistance varies linearly with temperature.)

3.17 A coil of wire is found to have a resistance of 5.00 Ω in melting ice. When the coil is heated to a temperature of 100 °C and a 100 Ω resistor is connected in parallel with it, the total resistance is 5.00 Ω. Calculate the temperature coefficient of resistance of the coil. (C)

3.18 The resistance of an electric fire is 32 Ω measured at room temperature, 15 °C. When connected to the 230 V mains, the fire dissipates 1250 W. If the element is made from nichrome wire whose temperature coefficient of resistance is $4.0 \times 10^{-4}\ ^\circ\text{C}^{-1}$, what is the working temperature of the wire? (O & C)

3.19

The resistance of this network between A and B is 220 Ω.

(*a*) Find the value of *R*.

(*b*) If the 100 Ω and 200 Ω resistors have zero temperature coefficients, yet

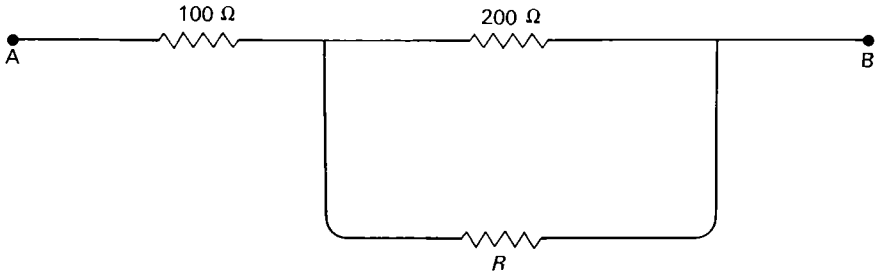


Fig e 10

the resistance of the network, while being 220 Ω at 20 °C, is 233 Ω at 100 °C, find the temperature coefficient of resistance of *R* at 0 °C.

(Assume the variation of resistance with temperature to be effectively linear.)

3.20 A milliammeter reads 0–100 μA, and its coil has resistance 120 Ω. Explain how you would convert it (*a*) to make a voltmeter, reading 0–5 V, (*b*) to make an ammeter, reading 0–1.0 A.

State the resistance of each of the converted instruments.

3.21

The battery puts a p.d. of 6.0 V across AB. Calculate (i) the p.d. across R_2

when switch S is open, (ii) the reading of the voltmeter when the switch is closed, for each of the two cases:

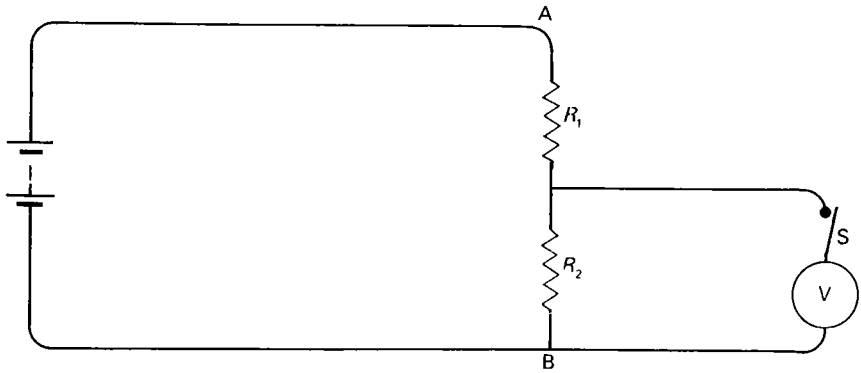


Fig e 11

(*a*) $R_1 = 2.0\ \Omega$, $R_2 = 4.0\ \Omega$;
(*b*) $R_1 = 2000\ \Omega$, $R_2 = 4000\ \Omega$,
the resistance of the voltmeter being $1000\ \Omega$.

3.22

The voltmeter has resistance $500\ \Omega$, and it is reading 1.0 V. If the p.d. V across AB is effectively constant, calculate (*a*) the value of this p.d. V, (*b*) the p.d. across the $1600\ \Omega$ resistor when the voltmeter is disconnected.

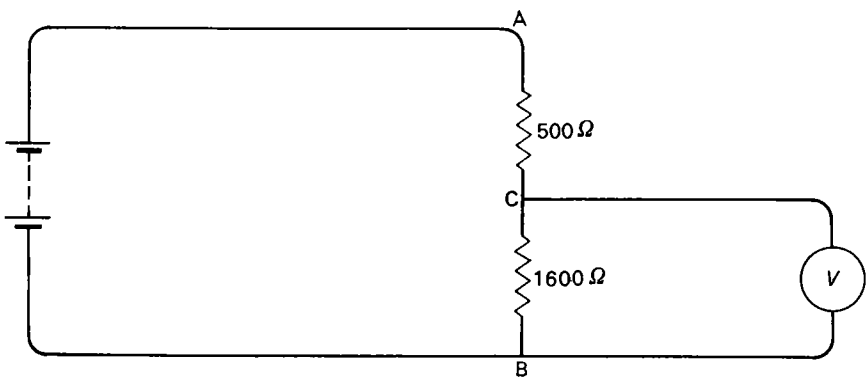


Fig. e 12

3.23

The p.d. across AB is constant, and is equal to 100 V (Fig. e 13). The resistance of the voltmeter is $1000\ \Omega$. If the reading of the voltmeter is 40 V, what is the p.d. across resistor R when the meter has been disconnected?

3.24

The circuit shown in the diagram is used to measure the current through the diode D and the p.d. across it. The voltmeter has a resistance of $10^7\ \Omega$. When A is positive with respect to B, R is made equal to $1000\ \Omega$. When A is negative with respect to B, R is made $10^6\ \Omega$. The value of the p.d. V_{AB} across AB is measured separately, using the same voltmeter. Obtain the value

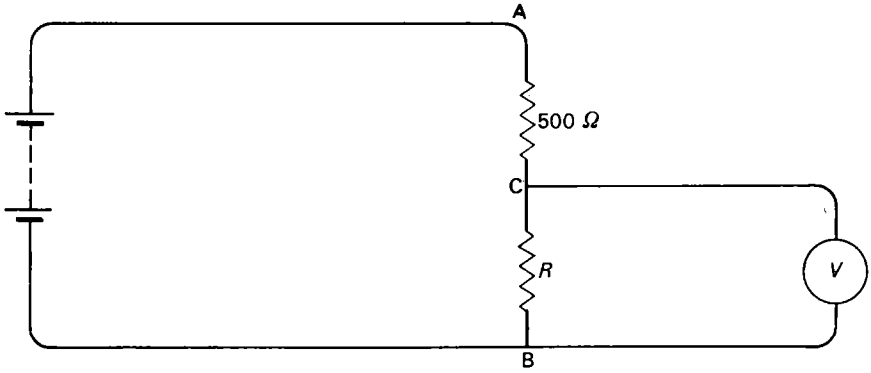


Fig. e 13

of the p.d. across the diode and the current through the diode for each of the following cases:

(*a*) $R = 10^3\ \Omega$, $V_{AB} = +3.30$ V,
$V = +0.75$ V,

(*b*) $R = 10^6\ \Omega$, $V_{AB} = -3.33$ V,
$V = -1.30$ V

V being the reading of the voltmeter when it is connected across the diode. (CAM. SCH. ENG.)

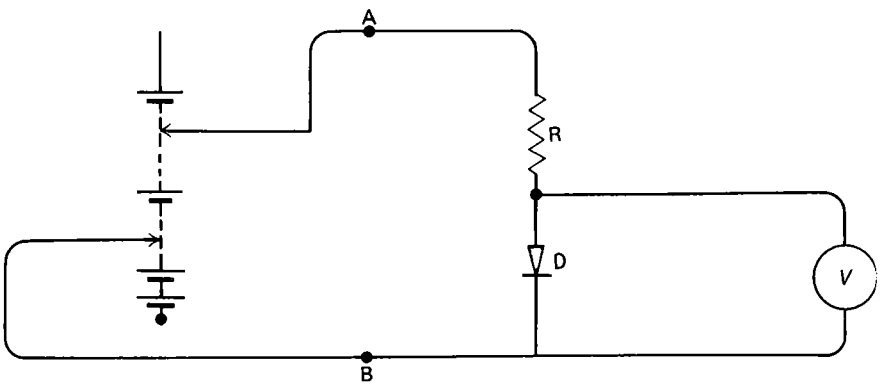


Fig. e 14

3.25 An ammeter reads 15% too low over the whole scale. The resistance of the coil is 5.00 Ω, and the resistance of the external shunt is $2.50 \times 10^{-3}\ \Omega$. How may its readings be restored to their correct values? (C)

3.26 A milliammeter has resistance 5.0 Ω at 15 °C, and reads full scale deflection for a current of 15 mA. Calculate the value of the shunt required if full scale reading is to represent 15 A.

The shunt is made from a resistance wire with zero temperature coefficient, but that of the copper coil of the meter is $4.2 \times 10^{-3}\ °C^{-1}$. If the shunted meter is at a temperature of 20 °C, what will be the true value of the current when the meter reads 10.0 A?

3.27 An inaccurate milliammeter, of resistance 10 Ω and having a linear scale, reads 1.15 mA when a current of 1.0 mA is passed through it. How might additional resistances be used to make the instrument read correctly if the ammeter and correcting resistors together are to present a resistance of 10 Ω to an external circuit? (OXF. SCH.)

3.28 An unknown resistance R is to be measured using a particular ammeter A and a particular voltmeter V, and neither the potential drop in A nor the current taken by V can be neglected.

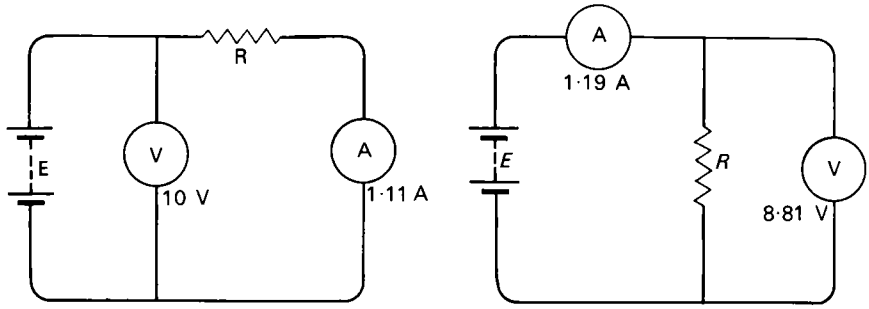


Fig. e 15

With the circuit connected as in figure (*a*), the readings of A and V are respectively 1.11 A and 10 V. With the connections changed to those shown in figure (*b*), the readings are 1.19 A and 8.81 V. The output p.d. of the source E is the same in both cases. Determine the value of R. (CAM. SCH. ENG.)

3.29 Explain the following observations:

(*a*) A bicycle lighting system consists of front lamp and rear lamp connected in parallel across the output terminals of a small dynamo, the return path for the current being through the bicycle frame. If the bulb of the front lamp burns out, the result is that the rear lamp glows more brightly than it did before.

(*b*) The plug of an electric kettle does not normally get hot, although it is in series with the heating element. Should the plug get hot, this is probably because the connections inside it require tightening up.

(*c*) A very small current through the human body can be fatal. Yet a cable carrying 1 000 A is not necessarily dangerous.

(*d*) It would be safer to supply electricity for domestic use at a p.d. of 100 V than it is at 240 V, but the installation would be more costly.

3.30

$$R_1 = R_2 = 2.0\ \Omega$$

The power taken by the variable resistor R is 10 W when the p.d. across it is 5 V (Fig. e 16). Calculate (*a*) the potential drop in each of the two resistors R_1, R_2, (*b*) the total power dissipated in these resistors, (*c*) the p.d. across AB.

If the variable resistor R were adjusted to take power of 10 W with a p.d. across it of 20 V, what would the answers to the above questions be then?

3.31 Electric power is supplied to a distant consumer through a power line which has total resistance 5 Ω. Explain why an input p.d. of 100 kV would be more satisfactory than one of 10 kV, and calculate in each case (*a*) the output p.d., (*b*) the output power when the power input is 20 000 kW. (O & C)

3.32 A town with an electricity supply at 200 V takes a load of 4×10^4 A. The supply

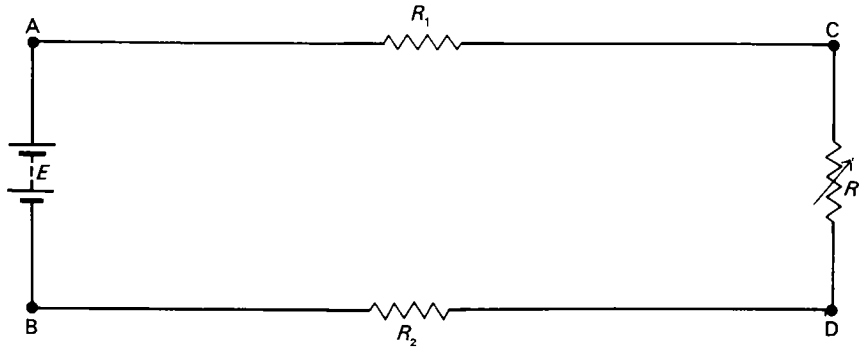


Fig. e 16

is provided by a loss free transformer, which is fed from a high voltage source through a twin cable 100 km in length. Each conductor in the cable has resistance 10^{-2} Ω/km. What must the voltage of the source be if the power loss in the cable is not to exceed 1% of the town's requirements? (CAM. SCH.)

3.33 (*a*) The p.d. across AB = 6 V.

$$R_1 = 2.0\ \Omega,\ R_2 = 4.0\ \Omega.$$

What is the p.d. between P and B, (i) when switch S is open, (ii) when S is closed?

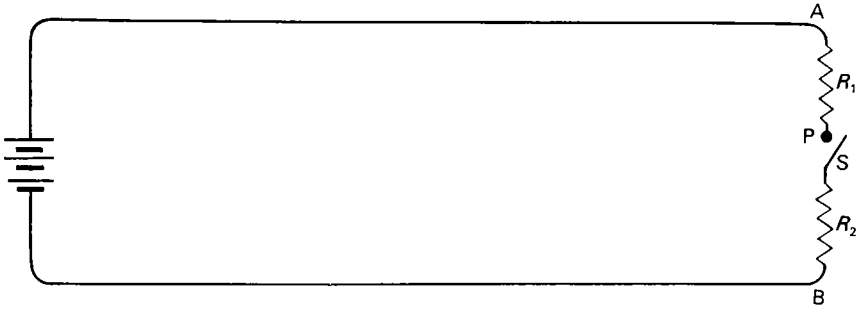


Fig. e 17

(*b*) A battery is connected across a 4.0 Ω resistor in series with a switch, and a high resistance voltmeter is connected across the battery. The reading on the voltmeter is 6.0 V when the switch is open, and 4.0 V when the switch is closed. Deduce what you can from these readings.

3.34 A storage cell has a constant e.m.f. of 2.0 V and an internal resistance of 1.0 Ω. What is the p.d. across its terminals (*a*) when they are connected by a wire of resistance 0.1 Ω and (*b*) when the cell is being charged from a 12.0 V d.c. supply of internal resistance 4.0 Ω? (O & C)

3.35 (*a*)
When switch S (Fig. e 18) is open the voltmeter reads 1.60 V; when S is closed the reading is 1.05 V. Assuming the voltmeter resistance to be very high, deduce the e.m.f. and the internal resistance of the cell.

(*b*) A cell, with e.m.f. E and constant internal resistance R_0, is connected in

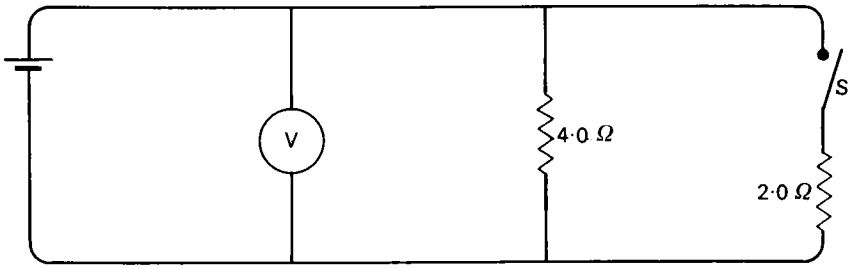


Fig. e 18

series with a variable resistor and an ammeter. Write an equation showing how the p.d. V across the terminals of the cell varies with the value of the current I drawn from it, and draw a graph to illustrate the variation. How can the value of R_0 be deduced from this graph?

3.36

(*a*) Calculate the potentials at the points A, B, C.

(*b*) Calculate the electrical power wasted in each of the cells.

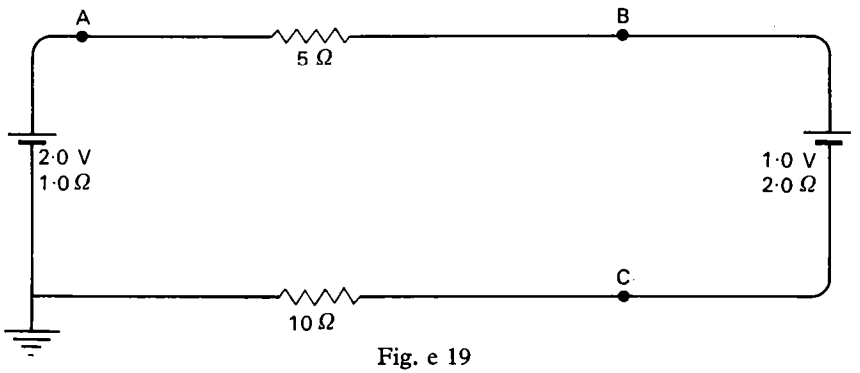


Fig. e 19

3.37 A battery of e.m.f. 8.0 V and internal resistance 30 Ω is used to drive current through a resistance of 50 Ω. A voltmeter with resistance 250 Ω is connected across the 50 Ω resistor. What does the voltmeter read? (O & C)

3.38

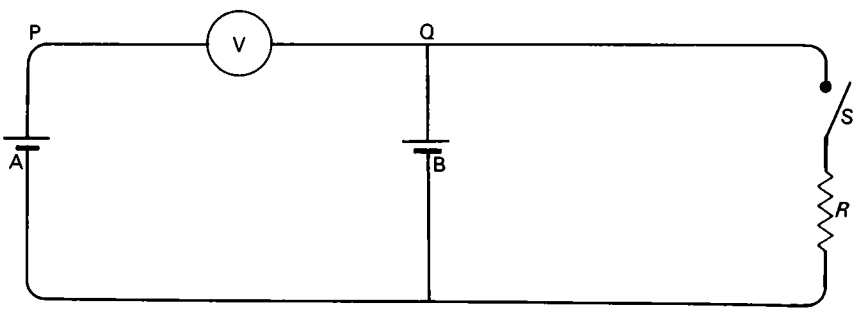


Fig. e 20

A, B are two lead accumulators. The resistance of the millivoltmeter V is high enough for the current flowing in it to be neglected.

When switch S is open circuited, the millivoltmeter reads 0.020 V, with terminal P having the higher positive potential. When S is closed, the reading rises to 0.032 V. If the resistor R has resistance 5.0 Ω, and the e.m.f. of accumulator B is 1.98 V, calculate its internal resistance.

3.39

The value of each resistor R in this circuit is $R = 1000\ \Omega$, and the battery has e.m.f. 100 V and negligible resistance.

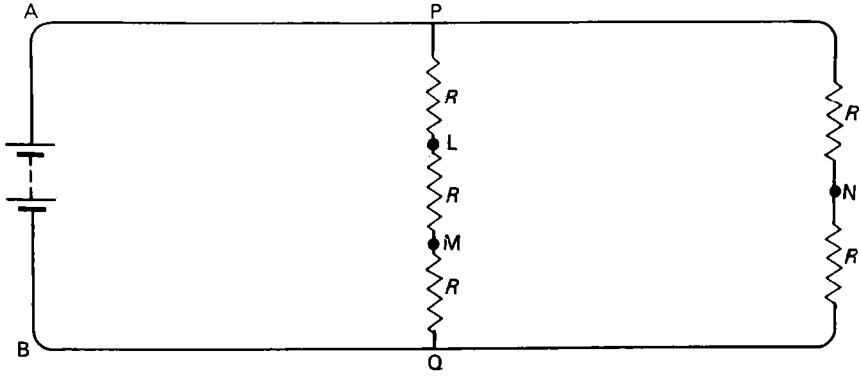


Fig. e 21

(*a*) Deduce the values of the p.d. (i) between L and Q, (ii) between M and Q, (iii) between L and N, (iv) between M and N.

(*b*) If a voltmeter, resistance 1000 Ω, is now connected between the points M and Q, what will it read?

(*c*) A resistor, value 1000 Ω, is then inserted in the circuit, in series with the lead connecting A to P. If the voltmeter has been disconnected, what is the p.d. between points L and N now?

3.40

The diagram represents a circuit arranged to charge six lead accumulators in series from a 24 V d.c. source. If each accumulator has e.m.f. 2.1 V and internal resistance 0.05 Ω, and the charging current is 3.0 A, calculate

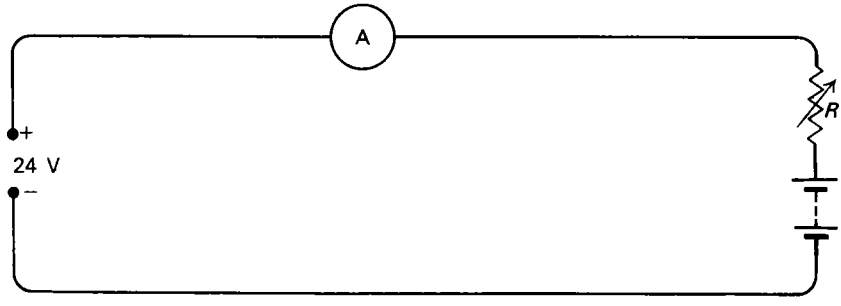


Fig. e 22

(*a*) the value of the series resistance R,

(*b*) the p.d. across the six accumulators,

(*c*) the rate at which heat is evolved in the accumulators,

(*d*) the power taken up usefully in charging the accumulators,

(*e*) the proportion of the energy taken from the source which is wasted as heat in the circuit.

3.41 A cell has an e.m.f. of 1.50 V and a resistance of 0.8 Ω. It is connected in series with a rheostat *R* and a milliammeter of resistance 3.0 Ω. A 2.0 Ω resistor is then joined in parallel with the meter. What value of *R* will cause the meter to read 20 mA? What is the p.d. between the terminals of the cell in these circumstances?

3.42 A battery A having e.m.f. 12.0 V and internal resistance 0.50 Ω is connected in parallel with a second battery B of e.m.f. 12.4 V and internal resistance 0.30 Ω. What resistance must be connected across the batteries so that B may supply twice the current supplied by A?

If this resistance is replaced by one of 2.0 Ω, what is the p.d. across the batteries then?

3.43

Under certain conditions, the electrical system of a motor car can be represented by the circuit shown. Find the power dissipated in the load

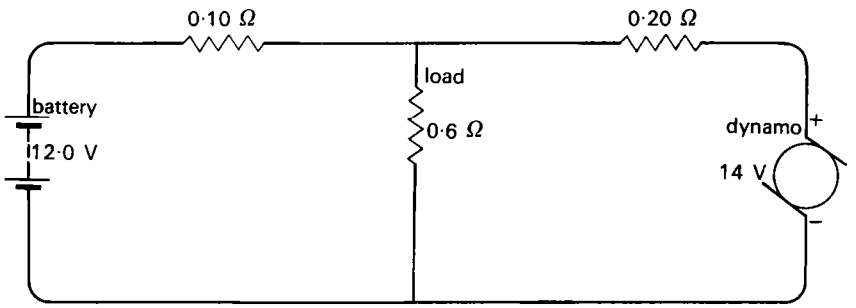


Fig. e 23

resistor, and the proportion of the power contributed by the battery and the dynamo respectively. Find also the efficiency of the system. At what dynamo voltage would the battery just begin to be charged? (OXF. SCH.)

3.44 State Kirchhoff's laws.

A d.c. dynamo of e.m.f. 50 V and internal resistance 1.0 Ω is connected in parallel with a battery of accumulators of e.m.f. 48 V and internal resistance 0.1 Ω. The combination is used to supply current to an external circuit. Calculate the currents flowing in the dynamo and battery, the p.d. across each, and the power supplied by (or taken in by) the battery when the current in the external circuit is (*a*) 1.0 A, (*b*) 2.0 A, (*c*) 3.0 A. (S)

3.45 Deduce the p.d. between the points A, B in this arrangement.

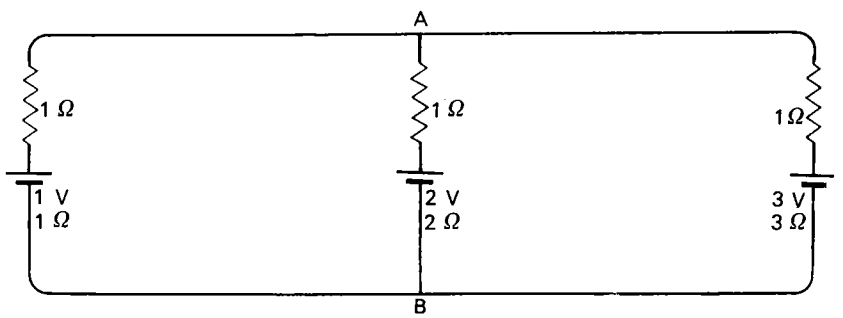


Fig. e 24

*3.46 Two cells, one of e.m.f. E_1 and internal resistance R_1, the other of e.m.f. E_2 and internal resistance R_2, are connected in parallel, with their positive poles joined together and their negative poles joined together by wires of negligible resistance. Show that the combination is equivalent to a battery of e.m.f. E and internal resistance R, where

$$E(R_1 + R_2) = E_1 R_2 + E_2 R_1$$

and

$$R(R_1 + R_2) = R_1 R_2.$$

If $E_1 > E_2$, and a resistance X is connected between the positive and negative terminals, find the condition for which the current through E_2 is zero. (O & C)

*3.47

In the circuit shown, each of the batteries has internal resistance 2 Ω, and battery B has an e.m.f. of 100 V. The voltmeter V, which draws negligible

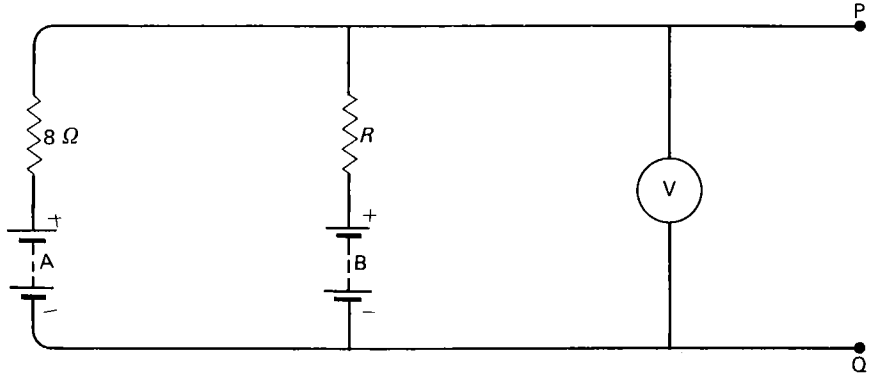


Fig. e 25

current, reads 96 V when the terminals PQ are open circuited as shown. When a 40 Ω resistor is connected across the terminals PQ, the voltmeter reading falls to 80 V.

Find the e.m.f. of the battery A and the value of the resistance R. What is the power delivered by battery B when the 40 Ω resistor is connected across terminals PQ? (CAM. SCH. ENG.)

*3.48 Five equal resistances R form the sides of a regular pentagon, and each vertex is joined to the centre of the pentagon by a further resistance R. Find the resistance between any two adjacent vertices and that between any two non-adjacent vertices. (OXF. SCH.)

*3.49 A rectangle ABCD has the following dimensions:

$$AB = CD = 10 \text{ cm},$$
$$AD = BC = 20 \text{ cm}.$$

It is formed of wire of resistance 10 Ω/m, and the two opposite corners AC are joined by a straight piece of the same wire. Calculate the resistance which would be measured between the other two corners B and D. (CAM. SCH.)

*3.50 A copper wire of circular cross-section is 1 m long and its diameter increases uniformly from 0.2 mm at one end to 0.4 mm at the other. Calculate the electrical resistance of the wire.

(Resistivity of copper $= 1.72 \times 10^{-8}$ Ω·m.) (CAM. SCH.)

*3.51 A wire of length l and resistance R is extended by an amount ϵ such that

$\epsilon \ll l$. If the volume of the wire and its resistivity remain unchanged, what is the fractional change in R?

If $R = 100\ \Omega$ and $\epsilon/l = 10^{-3}$, suggest a method of measuring the change in R. (O & C)

*3.52 A metal disc, radius r_2 and thickness d, has a central axle, of radius r_1, perpendicular to its plane. Electrical contacts are made to the axle and to a conducting band which makes contact with all points on the outer circumference of the disc. If the resistivity of the metal is ρ, what is the resistance R of the disc, measured between axle and circumference?

*3.53 A metal hemisphere of radius 10 cm is immersed near the centre of a large conducting tank containing a liquid having resistivity 60 $\Omega\cdot$m. The plane surface of the hemisphere is level with the surface of the liquid.

(*a*) Derive an expression for the resistance between two hemispherical shells within the liquid, concentric with the centre of the metal hemisphere, having radii r and $(r + \delta r)$ respectively, where δr is very small.

(*b*) Hence calculate the resistance between the hemisphere and the tank.

(*c*) When a p.d. of 10^3 V is applied between the hemisphere and the tank, where will the initial rate of rise of temperature be greatest? Estimate its value if 1.70 J are required to raise the temperature of 1 cm^3 of the liquid by 1 °C. (C)

*3.54 A variable resistor R is connected across a source which has e.m.f. E and internal resistance R_0. Obtain an expression for the power taken by the external resistance R, and show that the condition for this to be maximum is

$$R = R_0.$$

*3.55 A wire of length l, radius r, and resistivity ρ is heated in vacuo by a current I, which raises it to a steady Kelvin temperature T. Assuming that ρ is proportional to T, that the rate of radiation per unit surface area of the wire is proportional to T^4, and that heat losses by conduction through the support are negligible, find an expression for T in terms of the other quantities involved.

Compare (*a*) the radii, (*b*) the lengths of the filaments of a 100 V 60 W tungsten lamp and a 230 V 100 W tungsten lamp, assuming that the operating temperatures of the two filaments are the same.

Explain briefly why electric lamp filaments are run at the highest possible temperature consistent with a reasonable life. (O & C)

Chapter 4

The potentiometer and the bridge

4.1 Standards available in the laboratory

Direct-reading voltmeters and ammeters are convenient instruments, but limited in their accuracy, and even so their calibration should occasionally be checked. This requires reliable standards for comparison, and these are provided by having standard resistors and a standard cell available in the laboratory, the latter being a special type of cell chosen for having an accurately consistent e.m.f. (see section 5.15).

Knowing accurately the values of the standards, accurate measurement of p.d., current, and resistance can be carried out, and this is the concern of this chapter. How the standards themselves are determined is a different question, and a more fundamental one, which will be discussed in a later chapter (chapter 17). We shall assume here, for example, that the e.m.f. of a Weston standard cell is known to be 1.0186 V at 20 °C, but an experiment must have been carried out at some time to determine this value in terms of the definition of the volt. Such an experiment is termed an absolute determination, because it takes us back to first principles.

4.2 The potentiometer

The potentiometer is a device which measures p.d. by direct comparison with the e.m.f. of a standard cell, and has much greater accuracy than a pointer-reading voltmeter. We will describe it for the most part in its simplest form, which is an elementary piece of equipment; even so, its accuracy can be considerable.

The simplest practicable arrangement is to have a length, usually 1 m, of uniform resistance wire, AB (Fig. 4.1), stretched out straight between two copper contact strips fixed on a board alongside a scale. A cell D, termed the driver cell, is connected across the ends of this wire, to produce a potential gradient in it, and this potential gradient is then calibrated by reference to the e.m.f. E_S of a standard cell.

For the purpose of this discussion we suppose that the p.d. to be measured is the p.d. V across the terminals of some other cell (which is equal to its e.m.f.) and the procedure is illustrated in Fig. 4.1. If we can determine the lengths AP_S and AP of potentiometer wire which have p.d.s across them equal to E_S and V respectively, then we can write

$$\frac{V}{E_S} = \frac{AP}{AP_S} \tag{4.1}$$

from which the value V can be calculated.

The procedure is as follows: suppose that, in Fig. 4.1 (*a*), the sliding contact P′ does not yet make connection with the resistance wire AB. The p.d. between the points A and P_S on the wire is due to the potential gradient put in it by the driver

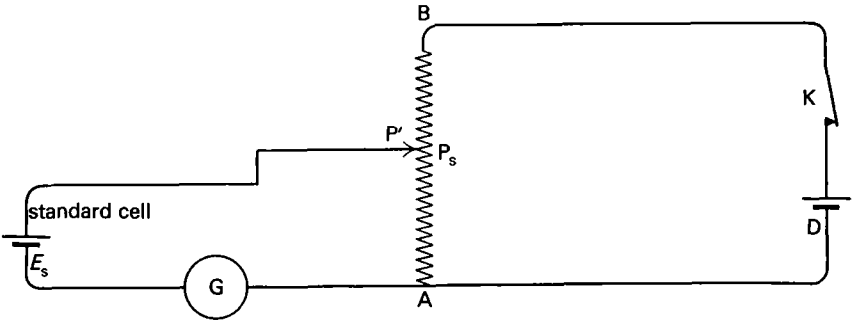


Fig. 4.1 (a)

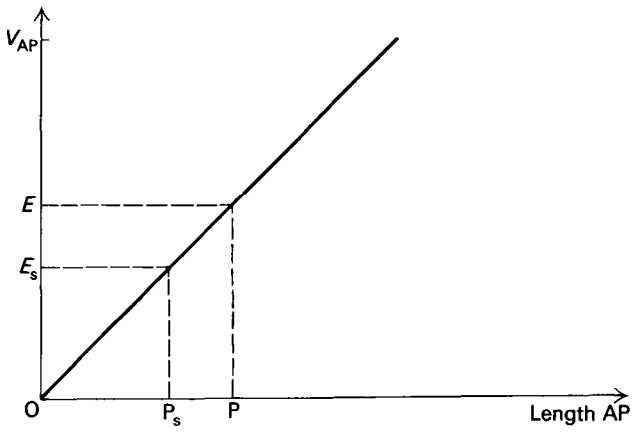


Fig. 4.1 (b)

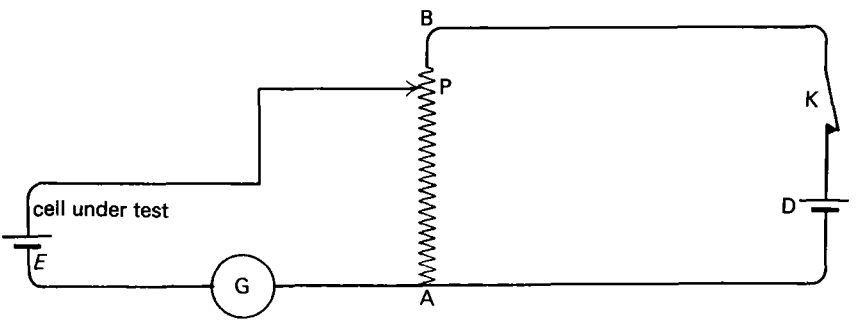


Fig. 4.1 (c)

Fig. 4.1 The potentiometer (AB may be a straight length of uniform resistance wire) (a) Calibrating the potentiometer by reference to a standard cell (b) The potential gradient set up in the wire AB (c) Using the calibrated potentiometer to measure the e.m.f. of another cell

cell D. The p.d. across AP′, on the other hand, is equal to the e.m.f. E_s of the standard cell. If, now, the point P_s is so chosen that these two p.d.s are equal, meaning that P′ and P_s are at the same potential, then P′ and P_s may be contacted together, and no current will flow through the standard cell, or through the

galvanometer connected in series with it. P_S is the balance point. Similarly the balance point P for the other cell is found.

4.3 The advantages of the potentiometer

It will be realised that the potentiometer is a form of voltmeter. By comparison with other types its greatest advantage is the length of its scale, which, even in the simple arrangement described above, is the full length of the wire AB. The precision with which the balance point can be located on the wire depends, of course, on the sensitivity of the galvanometer, and this is usually very discriminating, giving the potentiometer considerable accuracy.

A further contribution to the accuracy is the fact that the actual measurement made is one of length, the galvanometer being required merely to detect whether or not any current is flowing – the most precise thing that such a meter can do. Additionally, the potentiometer has the advantage that it acts as a voltmeter of extremely high resistance, and therefore does not interfere with the p.d. which it is being used to measure; if it draws any current at all, this must be too small to register on the galvanometer.

These advantages depend, of course, on the potential gradient in the wire AB being uniform, since it is upon this assumption that equation 4.1 depends. This means that the resistance wire must have uniform cross-section.

4.4 A direct-reading potentiometer

The circuit drawn out in Fig. 4.2 represents an arrangement by which a simple straight wire potentiometer may be calibrated to read the p.d. directly in volts.

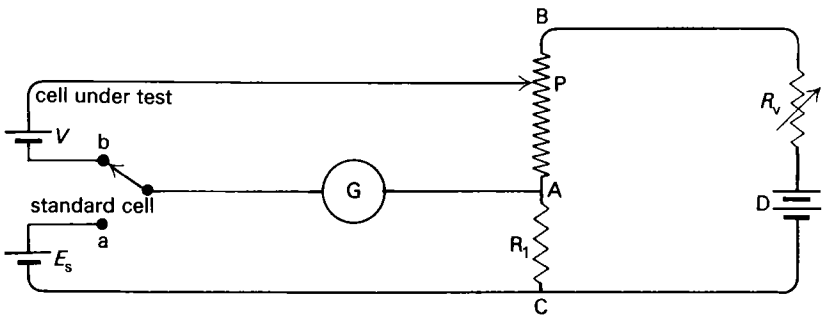


Fig. 4.2 A direct reading potentiometer

We suppose that the potentiometer wire AB is 1.00 m long, having resistance R measured in ohms. A resistor R_1 is connected in series with it, selected to have the value

$$R_1 = 1.018\,R,$$

where the e.m.f. of the standard cell is taken to be 1.018 V.

When the change-over switch S is connected to contact (a), to put the standard cell in circuit, the variable resistor R_V is used to vary the potentiometer current until balance is obtained. The p.d. across the resistor R_1 must now be equal to the e.m.f. E_S of the standard cell, and if the current in the potentiometer wire is I, then

$$I = \frac{E_S}{R_1}$$
$$= \frac{1.018}{1.018R}$$
$$= \frac{1.00}{R}.$$

Hence, the p.d. across the whole length AB of the potentiometer wire is

$$V_{AB} = IR$$
$$= 1.00 \text{ V},$$

and the potential gradient in the wire is 1.00×10^{-3} V/mm; each mm on the scale represents 1.00 mV.

The change-over switch can now be connected to contact (*b*), and the p.d. under test (shown in the diagram as being the e.m.f. of another cell) can be measured directly.

This is the principle upon which precision types of potentiometer are made.

4.5 The measurement of current by potentiometer

Although the potentiometer is basically a voltmeter, this does not preclude its application to the measurement of current. The principle in this case is to include a standard resistance R_S in the circuit, so that the current must flow through it, and to measure the potential difference V which is set up across this resistance, whence

$$I = \frac{V}{R_S}. \tag{4.2}$$

This principle is illustrated by the circuit shown in Fig. 4.3, where it is supposed that both the p.d. across a torch bulb and the current flowing through it are to be measured. Because the potentiometer, for all its more complicated appearance, is

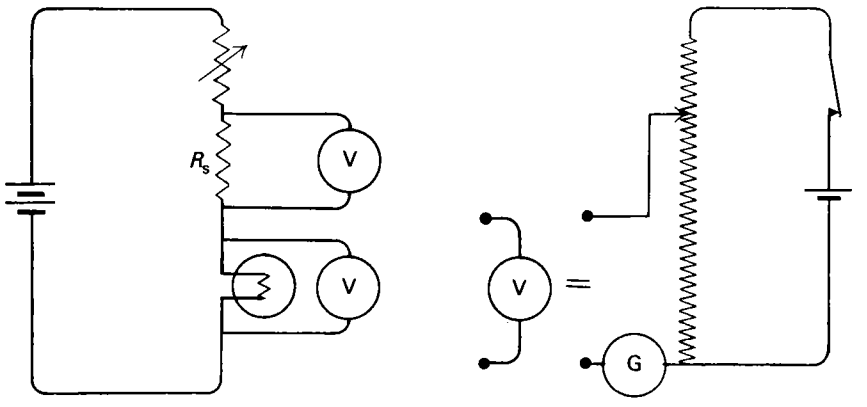


Fig. 4.3 The measurement of both p.d. and current by voltmeter or potentiometer (The potentiometer is calibrated by standard cell)

but a form of voltmeter, there is no need (unless to satisfy an examiner) to draw out the potentiometer circuit in full, and this economy has been implied in the diagram.

4.6 The calibration of a voltmeter and an ammeter

The reading of a voltmeter cannot be successfully checked by connecting a standard cell directly across it because the internal resistance of this type of cell is high, and the current drawn by the voltmeter, although small, is sufficient to cause an appreciable potential drop in the cell. Consequently, the p.d. applied across the voltmeter is something less than the standard e.m.f. of the cell.

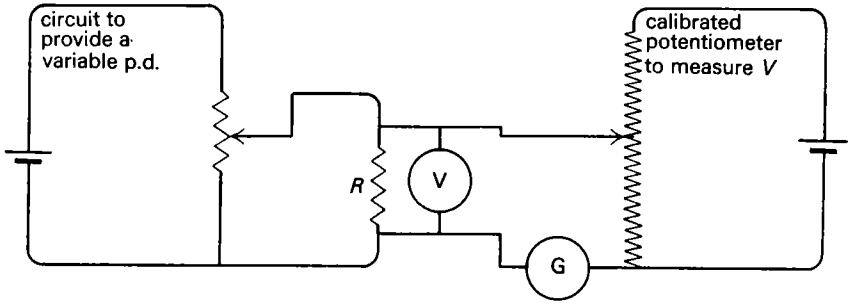


Fig. 4.4 (a)

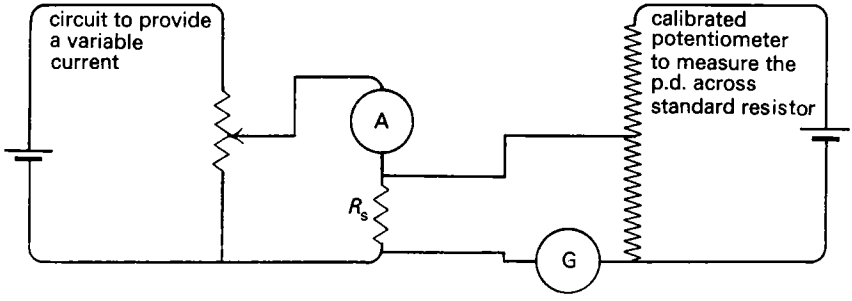


Fig. 4.4 (b)

Fig. 4.4 Circuits to calibrate voltmeter and ammeter. The potentiometer is supposed to have been already calibrated by reference to a standard cell (a) Calibrating a voltmeter (b) Calibrating an ammeter $I = \frac{V}{R_s}$

The correct way to calibrate the voltmeter, therefore, is first to calibrate a potentiometer by comparison with the standard cell, and afterwards to connect potentiometer and voltmeter in parallel across the same load, so that the readings of the two instruments may be compared.

Similarly, an ammeter is calibrated by connecting it in a circuit in series with a standard resistance, and measuring the p.d. across the resistance by means of a calibrated potentiometer (Fig. 4.4).

4.7 The standard resistor

When using a standard resistor, particularly one of low value, some attention must be paid to the connections, because a bad connection may add to the effective value of the resistance. When a potentiometer is being used in conjunction with it, this source of error may be avoided by providing the resistor with four terminals.

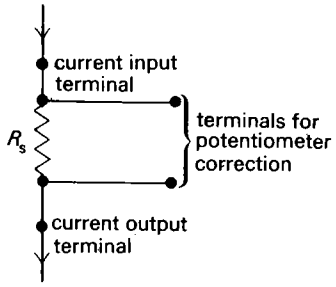


Fig. 4.5 A standard resistor, having four terminals

The current is fed in and out of the resistor at two of these terminals, while the potentiometer which measures the p.d. is connected to the other pair. The connections to the potentiometer can introduce no error because no current flows in them, and at the same time any potential drop which may be set up at the current terminals is not included in the potentiometer reading.

4.8 Measuring a high potential difference

The range of p.d.s covered by the potentiometer is normally restricted to values which are lower than the p.d. applied across it by the driver cell D, and this is usually of the order of one or two volts. To increase the range to a higher limit, a

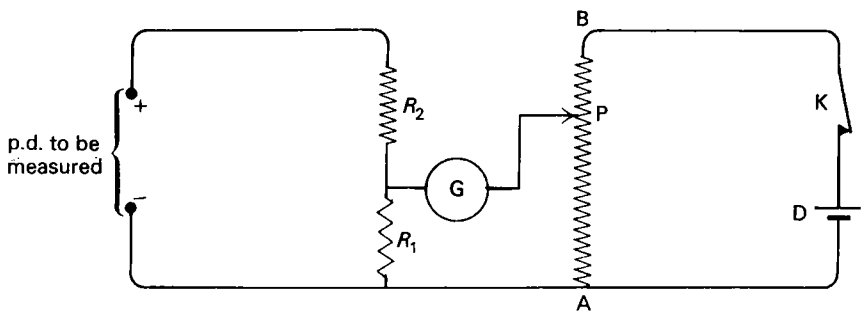


Fig. 4.6 Measurement of a high p.d.

resistance chain may be used as a potential divider, so that a known fraction of the p.d. under test is measured.

As an example, suppose that the p.d. to be measured is about 100 V. The potential divider in this case might consist of two resistors R_1, R_2 connected in series, where $R_1 = 10^3\ \Omega$ and $(R_1 + R_2) = 10^5\ \Omega$. The p.d. is applied across the two in series, and the part of it which appears across R_1 is tapped off for measurement by the potentiometer. We then have

$$V = \frac{(R_1 + R_2)}{R_1} V_1, \tag{4.3}$$

where V_1 is the measured p.d. across R_1, and V is the full p.d. which is to be determined.

Inserting the values postulated above for R_1 and R_2, this would give

$$V = 10^2 V_1.$$

A disadvantage of this circuit is that one of the features of the potentiometer has been partially lost; its input resistance is not quite so high as it would have been, because the potential divider draws some current. For this reason, the values chosen for R_1 and R_2 are as high as the sensitivity of the galvanometer permits.

4.9 Measuring a very small potential difference

To measure a p.d. of, say, 1 mV, the scale of the potentiometer requires to be extended, so that a small p.d. of this order shall provide a balance length of appreciable size. This means that the potential gradient must be reduced to the order of 1 mV/m, which can be done by inserting a high value resistor in series with the driver cell, to take up the greater part of its p.d.

Example 1

The e.m.f. of a thermocouple (which is a few millivolts) is measured by setting up the circuit of Fig. 4.7. AB is a length, 1.00 *m, of uniform potentiometer wire, resistance* 5.5 Ω, *and the value of the resistor R, which is connected in series with it, is* 2000 Ω. *The p.d. applied by the driver cell across the potentiometer wire and the resistor R in series, as measured by a voltmeter, is* 2.0 V.

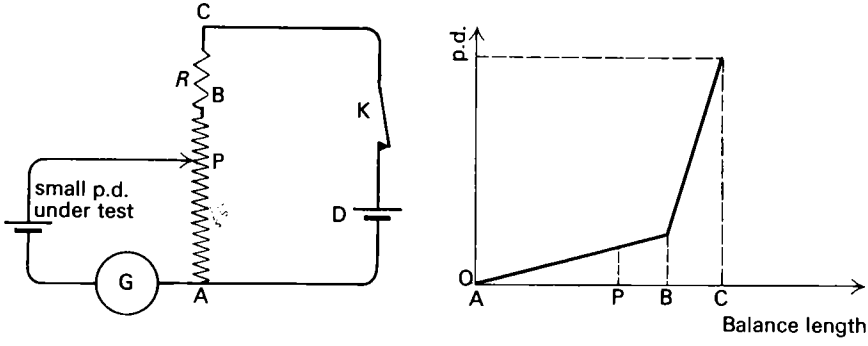


Fig. 4.7 Measurement of a small p.d.

When the thermocouple is connected across AP, the balance length is found to be 65.2 *cm. Deduce the e.m.f. E of the thermocouple.*

The p.d. across the full length AB of the potentiometer wire is V_{AB}, where

$$\frac{V_{AB}}{V_{AC}} = \frac{R_{AB}}{R_{AC}},$$

i.e.

$$\frac{V_{AB}}{2.0} = \frac{5.5}{2005.5},$$

$$\therefore \quad V_{AB} = 5.48(5) \times 10^{-3}\ \text{V}.$$

The e.m.f. of the thermocouple, therefore, is V, where

$$V = \frac{\text{AP}}{\text{AB}} \times V_{AB}$$

$$= 0.652 \times 5.48(5) \times 10^{-3}\ \text{V}$$

$$= 3.6 \times 10^{-3}\ \text{V}.$$

There is no purpose in stating this result to more than two significant figures, because of the low accuracy of the original calibration upon which the calculation is based.

A more accurate determination would refer to the e.m.f. of a standard cell, as in the next example.

Example 2

The e.m.f. of a thermocouple is measured by setting up the circuit shown in Fig. 4.8, *where* R_1, R_2 *are resistance boxes. AB is a length* 1.00 *m of uniform resistance wire, resistance* 5.45 Ω. *The value of* R_2 *is selected to be* 1000 Ω, *and the value of* R_1 *is adjusted so that a balance point,* P_S, *can be obtained with the standard cell.*

When R_1 = 1040 Ω, *the balance length,* AP_S, *is found to be* 21.1 *cm. When the thermocouple is connected across* AP, *the length required to balance the e.m.f. of the thermocouple is* AP = 68.5 *cm. Taking the e.m.f. of the standard cell to be* E_S = 1.018 V, *deduce the e.m.f. E of the thermocouple.*

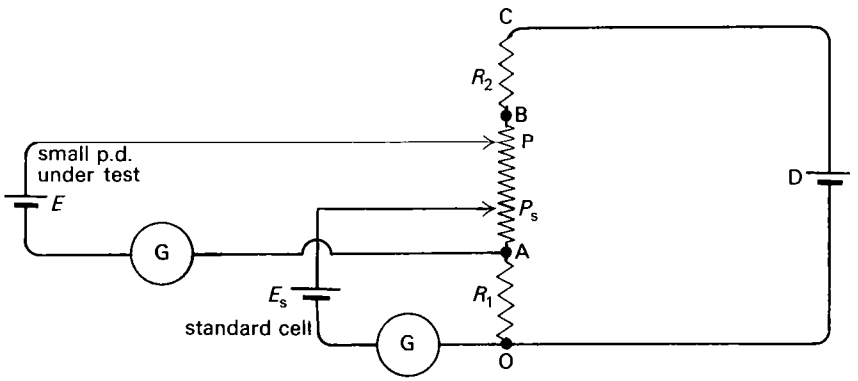


Fig. 4.8 Measurement of a small p.d. Calibration by standard cell

In this case, it will be convenient to treat the resistor R_1 as representing a considerable extension in length of the potentiometer wire AB, viz:

$$R_{AB} = 5.45\ \Omega,$$
$$\text{length AB} = 1.00\ \text{m},$$
$$R_1 = 1040\ \Omega.$$

Therefore equivalent length of potentiometer wire = l_1

where
$$l_1 = \frac{1040}{5.45} \times 1.00\ \text{m}$$
$$= 190.83\ \text{m}.$$

The length of potentiometer wire which balances the e.m.f. E_S, of the standard cell is

$$l_S = l_1 + 0.211\ \text{m}$$
$$= 191.04\ \text{m}.$$

But the length which balances the e.m.f. E, of the thermocouple is

$$l = 68.5\ \text{cm}$$
$$= 0.685\ \text{m},$$

$$\therefore \quad E = \frac{l}{l_s} \times E_s$$

$$= \frac{0.685}{191.04} \times 1.018 \text{ V}$$

$$= 3.65 \times 10^{-3} \text{ V}.$$

4.10 The comparison of resistances by potentiometer

The value of a resistor R may be compared with that of a standard resistor R_s by passing current through the two resistors, connected in series, and comparing the p.d.s set up across them. If a potentiometer is used, this provides an accurate comparison. We have

$$V = IR$$

$$V_s = IR_s,$$

whence

$$\frac{R}{R_s} = \frac{V}{V_s} \tag{4.4}$$

In the circuit diagram of Fig. 4.9, a variable resistor is included in series with the driver cell D. This enables the sensitivity of the potentiometer to be adjusted, so that suitable balance lengths may be obtained.

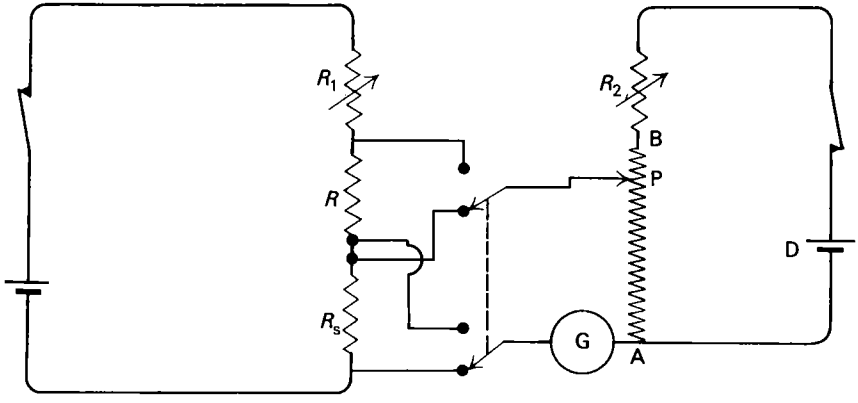


Fig. 4.9 The comparison of two resistances by potentiometer

Although this method is applicable to resistors over a wide range, it is particularly useful for comparing those of very low value, say 10^{-1} to 10^{-4} Ω. In this range the Wheatstone's bridge, next to be described, loses its accuracy, owing to loss of sensitivity and the increased relative importance of the leads and connections. In the case of the potentiometer, the resistance of leads and connections introduces no error because no current flows in them, which gives the potentiometer circuit the advantage.

4.11 The Wheatstone's bridge

The more familiar circuit for the accurate comparison of two resistances is the Wheatstone's bridge.

This circuit is said to be balanced when there is zero current in the galvanometer,

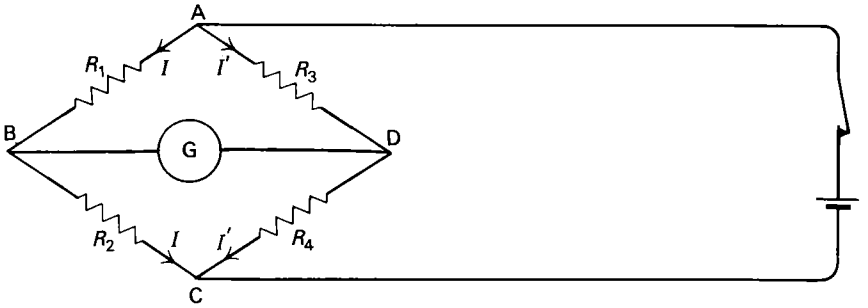


Fig. 4.10 The Wheatstone's bridge circuit

in which case the current flow is as it is assumed to be in the diagram, resistors R_1, R_2 taking a current I, and resistors R_3, R_4 taking a current I'. Also, since the p.d. across the galvanometer is zero, the points B, D must have the same potential.

For the balanced bridge circuit, the following relations therefore hold:

$$V_{BD} = 0,$$
$$\therefore \quad V_{AB} = V_{AD},$$
$$\therefore \quad IR_1 = I'R_3.$$

Also

$$V_{BC} = V_{DC},$$
$$\therefore \quad IR_2 = I'R_4.$$

Eliminating I, I' between these equations gives the balanced bridge relation

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}, \tag{4.5}$$

whence the ratio of resistors R_1, R_2 can be determined if the ratio R_3/R_4 is known.

As with the potentiometer, precision types of Wheatstone bridge arrangements are widely used, having accurately wound resistance bobbins connected to contact studs of rotary switches.

The simplest form, however, has a length of uniform resistance wire stretched on a board, which is divided into two segments by a contact key, making two

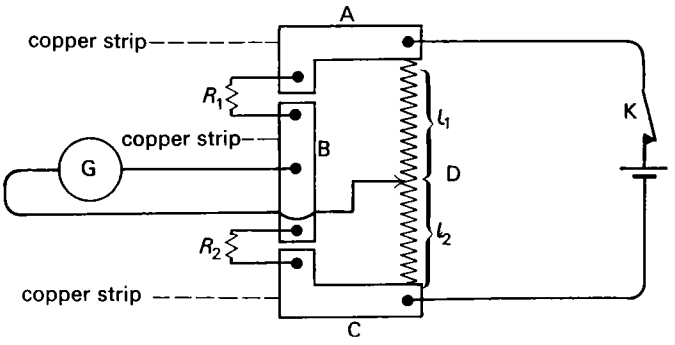


Fig. 4.11 The metre bridge

branches, R_3, R_4 of the bridge circuit. This is called the metre or the half-metre bridge, depending on the length of bridge wire used.

For the metre bridge, we have

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}, \tag{4.5}$$

$$\therefore \quad \frac{R_1}{R_2} = \frac{l_1}{l_2}, \tag{4.6}$$

where l_1, l_2 are the lengths into which the contact D divides the resistance wire when the circuit is balanced.

The thick copper strips shown in the diagram are included in the apparatus because the values to be inserted for R_1, R_2 in equation 4.6 include the resistances of the leads, and these copper strips reduce the length of lead which is necessary.

*4.12 Examples of Kirchhoff's laws, applied to unbalanced potentiometer and bridge circuits

Example 1

A potentiometer consists of a uniform resistance wire, length 1.00 *m, with resistance* 5.45 Ω, *across the ends of which a p.d. equal to* 2.00 V *is applied from a driver cell which has negligible resistance. What length of potentiometer wire will balance the e.m.f. of another cell, if this has the value* E = 1.50 V?

If the contact key is now moved to a point which is 1.0 *cm short of the balance point, deduce the value of the current which flows through the galvanometer in series with this cell. The cell and galvanometer have resistances* 1.10 *and* 15.50 Ω *respectively.*

The potentiometer circuit is shown in Fig. 4.12. It is almost obvious that the balance length is 75.0 cm.

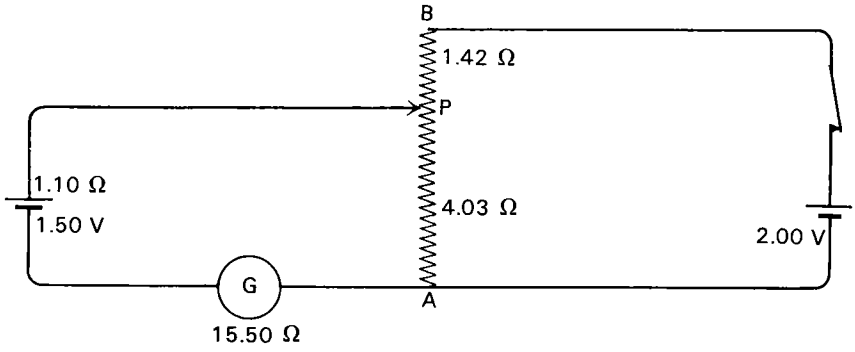


Fig. 4.12 (a)

When the contact is made 1.0 cm short of the balance point, the potentiometer wire has been divided into two segments having lengths 74.0 cm and 26.0 cm. These have resistances equal, respectively, to

$$R_1 = 5.45 \times \frac{74.0}{100}\ \Omega$$

$$= 4.03\ \Omega,$$

and

$$R_2 = 5.45 \times \frac{26.0}{100}\ \Omega$$

$$= 1.42\ \Omega.$$

This circuit is drawn out again in Fig. 4.12(*b*), to have an appearance which is more obviously a network.

The currents are first labelled in the diagram, using Kirchhoff's first law to reduce

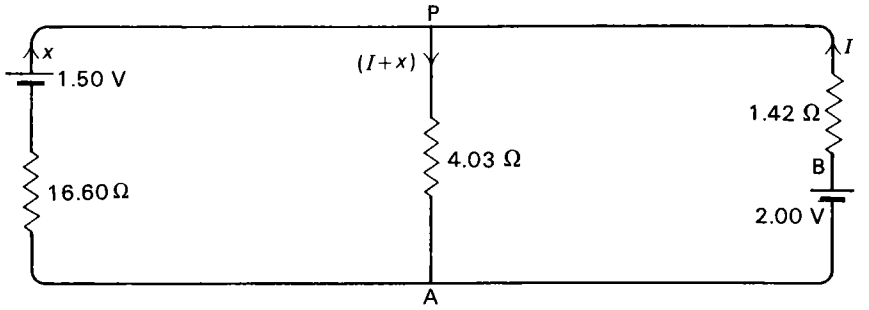


Fig. 4.12 (b)

the number of different symbols required to represent currents. Since only the current in the galvanometer is to be found, this has been labelled x, to distinguish it. Since the potentiometer is not far off balance, the value of x is likely to be small.

Applying Kirchhoff's second law, we have

$$1.50 = 16.60x + 4.03\,(I + x)$$
$$= 20.63x + 4.03I, \qquad \text{(i)}$$

and

$$2.00 = 1.42I + 4.03\,(I + x)$$
$$= 4.03x + 5.45I. \qquad \text{(ii)}$$

Eliminating I gives

$$x = 1.20 \times 10^{-3}\ \text{A}$$

Example 2

A balanced bridge circuit has four arms of equal resistance, each being 10.0 Ω. *The p.d. across the circuit derives from a cell having e.m.f. equal to* 2.0 V *and negligible internal resistance, and the galvanometer resistance is* 100.0 Ω.

Calculate the current which flows in the galvanometer if the resistance of one arm of the bridge is increased by 0.10 Ω, *and the resistance of an adjacent arm is reduced by* 0.10 Ω.

Again, Kirchhoff's first law has been used in labelling currents in the diagram,

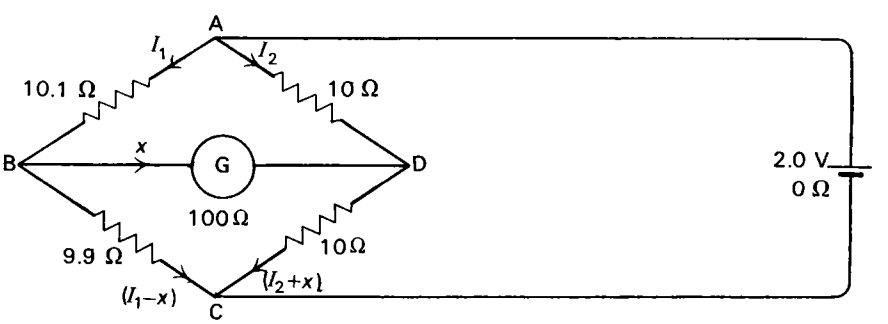


Fig. 4.13

and the symbol x has been allotted to the galvanometer current, which has to be determined.

In applying Kirchoff's second law, alternative closed paths are available. Two of the three equations required for a solution are chosen to include the branch having the cell, because its zero resistance causes these to be relatively simple equations. We have:

Loop ABCEA : $2.0 = 10.1\ I_1 + 9.9\ (I_1 - x),$

$$\therefore\ 2.0 = 20.0\ I_1 - 9.9x, \qquad \text{(i)}$$

and Loop ABCEA : $2.0 = 10.0\ I_2 + 10.0\ (I_2 + x),$

$$\therefore\ 2.0 = 20.0\ I_2 + 10.0x. \qquad \text{(ii)}$$

Taking ABDA as the third closed path,

$$10.1\ I_1 + 100x - 10.0\ I_2 = 0. \qquad \text{(iii)}$$

Solving:

from equation (i) $\qquad I_1 = \dfrac{(2.0 + 9.9x)}{20.0},$

from equation (ii) $\qquad I_2 = \dfrac{(2.0 - 10.0x)}{20.0}.$

Substituting for I_1, I_2 in (iii)

$$10.1\ \frac{(2.0 + 9.9x)}{20} + 100x - 10.0\frac{(2.0 - 10.0x)}{20} = 0,$$

whence

$$2000x = 0.2$$

$$x = 10^{-4}\ \text{A}.$$

QUESTIONS

4.1 A potentiometer arrangement consists of a 1.00 m length of uniform resistance wire AB, in which a potential gradient is set up by connecting a lead accumulator with e.m.f. approximately 2 V across it. A standard cell, with e.m.f. 1.018 V, is found to balance across 51.2 cm of the wire, and a dry cell X balances across 74.8 cm.

Draw a circuit diagram, and deduce (*a*) the p.d. across AB, (*b*) the e.m.f. of the dry cell X.

A resistance of 0.50 Ω is now inserted in the circuit, in series with the accumulator and the potentiometer wire AB. In this new circumstance, the standard cell balances across 57.9 cm of the wire. Assuming the p.d. across the accumulator terminals to have the same value as before, deduce the resistance of the potentiometer wire.

4.2 Suggest a possible reason for each of the following experiences when using a simple potentiometer circuit.

(*a*) When comparing the e.m.f.s of a dry cell and of a standard cell, it is found that the selected galvanometer clearly indicates the balance point on the wire to ±1 mm when using the dry cell, but only to about ±5 mm when using the standard cell.

(*b*) A lead accumulator is connected across the potentiometer wire, in series with a tapping key, to provide the potential gradient and the balance point is found for a dry cell. The balance length is initially equal to 75.2 cm, but to

maintain the balance it is found necessary slowly to move the contact point, as the balance length slowly increases.

(*c*) When the e.m.f.s of a standard cell and of a dry cell are being compared by potentiometer, the balance lengths are, respectively, 51.2 cm and 74.8 cm. The potentiometer is then disconnected, and a moving coil voltmeter, of resistance 500 Ω, is used to compare the cells instead; the reading is 0.55 V across the standard cell, and 1.45 V across the dry cell.

4.3 The diagram represents a circuit which is set up to measure the specific heat of copper. R_s is a standard resistor, the value of which is 0.20 Ω, and R is a heating resistor imbedded in the copper block B. When current is flowing

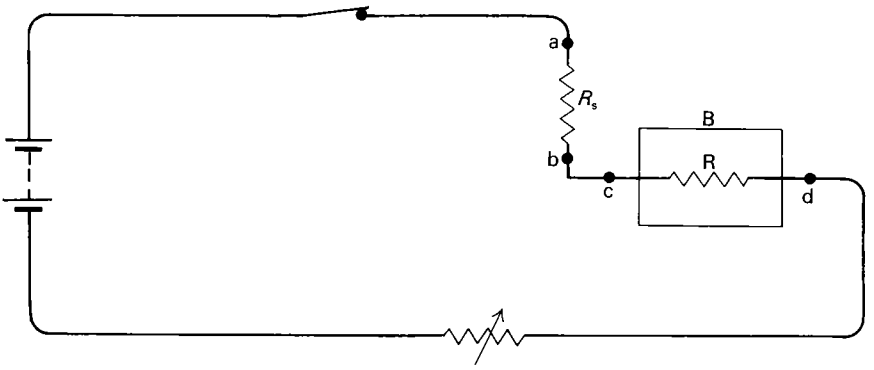


Fig. e 26

in the circuit, the p.d.s across the resistors R_s and R are measured by connecting a potentiometer circuit first across ab and then across cd. Draw out the circuit again, including the potentiometer in its two positions.

Readings are obtained as follows:

balance length for a 1.018 V standard cell = 51.2 cm,
balance length across ab = 60.3 cm,
balance length across cd = 90.5 cm.

Calculate the rate at which heat is being released in the resistor R.

4.4 (*a*) A lead accumulator, a 1.0 Ω resistance coil, and a 1 m length of potentiometer wire are connected in series. A cell is balanced across 72.0 cm of the wire. If the cell is now connected instead across the 1.0 Ω resistor and a segment of the wire, the balance point is 37.3 cm along the wire. Deduce the resistance of the potentiometer wire.

(*b*) A dry cell is balanced by a length 74.8 cm of a potentiometer wire. When a 3.0 Ω resistor is connected directly across the terminals of the dry cell, the balance length becomes 69.7 cm. Calculate the internal resistance of the dry cell.

4.5 (*a*) The cell shown in Fig. e 27 has e.m.f. E and internal resistance R_0, which is constant and does not vary with the current. The variable resistor R is a resistance box capable of carrying the current without overheating.

Sketch the graph which you would expect to obtain if you were to plot values of the reciprocal, $1/I$, of the current against corresponding values of R and explain how you could deduce R_0 from the graph.

(*b*) A resistance box R is connected across the terminals of a cell having e.m.f. E and constant internal resistance R_0. The p.d. across the resistance box is balanced against a potentiometer wire, the balance length being l. Show that

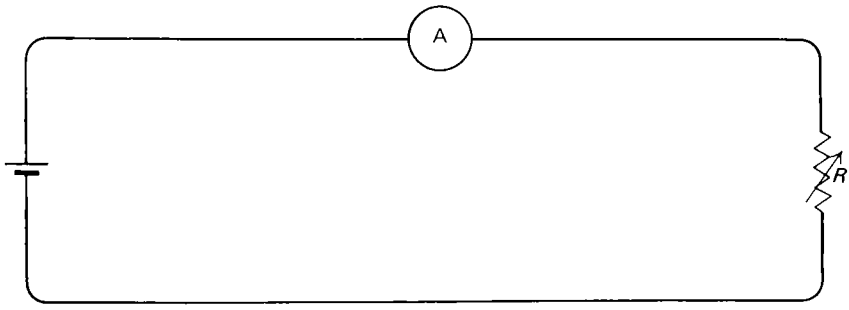


Fig. c 27

a graph plotting $y = R/l$ against values of R is a straight line, and show how you could deduce the value of R_0 from the graph.

4.6 In the circuit shown, AB is a uniform potentiometer wire 100 cm long. E_1 is an accumulator of e.m.f. 2.0 V and negligible internal resistance.

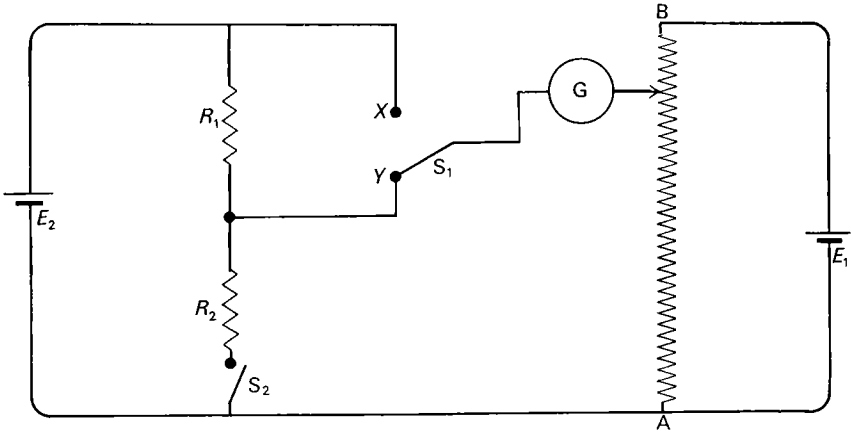


Fig. e 28

E_2 is a cell of e.m.f. 1.1 V and resistance 1.0 Ω. The values of the two resistors R_1 and R_2 are 1.0 Ω and 2.0 Ω respectively. Switch S_1 enables the galvanometer G to be connected to either X or Y. Calculate the balance lengths obtainable for each of the two positions of S_1, when S_2 is (*a*) open, (*b*) closed. (C)

4.7 A simple potentiometer apparatus consists of a 1 m length of uniform resistance wire AB stretched out between a copper strip at A and a copper strip at B. Owing to a defect in the manner in which the contact at A is made between wire and copper, there is a small zero error, and an observed balance length l represents an actual length $(l + x)$.

In order to determine the zero error, a current is sent through two resistors R_1 and R_2 connected in series, and the p.d.s across R_1 and across R_2 are balanced, in turn, against a length of the potentiometer wire. Then the balance length corresponding to the combined p.d. across the two resistors is similarly determined. If the three balance lengths are, respectively, 21.2 cm, 43.7 cm, and 65.2 cm, calculate the value of the zero error x, and the ratio of the resistances R_1 and R_2.

4.8 In an experiment to determine the resistivity of copper, a length of copper wire of resistance R_1 is connected in series with a comparable standard resistance R_2 of value 0.010 Ω, an accumulator of just over 2 V e.m.f., and a resistance R of about 20 Ω. The p.d.s across R_1 and R_2 are compared by a potentiometer arrangement consisting of a uniform wire of 30 Ω resistance, a resistor S in series, a 2 V accumulator, and the usual accessories. Draw the circuit diagram.

Explain the purpose in the circuit of the resistance R, and estimate a suitable value for S. (N)

4.9 An accumulator, a resistance box, a 400 Ω resistor, and a potentiometer wire 10 m long with resistance 40 Ω are connected in series. It is found that for a certain value of the resistance in the box, 685 cm of potentiometer wire balance the small e.m.f. of a concentration cell. For the same value of the box resistance, the p.d. across the 400 Ω resistor plus 320 cm of the potentiometer wire balances the e.m.f., 1.018 V, of a standard cell.

Calculate the e.m.f. of the concentration cell. (O & C)

4.10 A constant current was passed through a potentiometer consisting of a coil of resistance 1000 Ω connected in series with a uniform slide wire of resistance 4.00 Ω/m. A balance was obtained when a Weston cell of e.m.f. 1.018 V was connected across the 1000 Ω coil together with 450 cm of the slide wire. A balance was also obtained when a thermocouple was connected across 120.5 cm of the slide wire.

Find (*a*) the current through the potentiometer, (*b*) the e.m.f. of the thermocouple. (O & C)

4.11 A steady current is passed through a manganin potentiometer wire AB of length 10 m and diameter 5.6×10^{-1} mm, connected in series with a resistance box BC of resistance 1001 Ω and a variable resistor. A Weston cell of e.m.f. 1.018 V, with a sensitive galvanometer in series with it, is connected in parallel between the points A and C. The current in the potentiometer wire is adjusted by means of the variable resistor so that the galvanometer registers zero deflection. The Weston cell is now removed, and is replaced by a thermocouple. Zero galvanometer reading is now obtained when the thermocouple is connected between A and the point P on the potentiometer wire, where AP = 524 cm.

Draw the two circuits and calculate the e.m.f. of the thermocouple. (Resistivity of manganin $= 4.19 \times 10^{-7}$ Ω·m.) (L)

4.12 A potentiometer circuit is set up by connecting a length of uniform resistance wire AB in series with a resistance box R, a variable resistor R_1, and a 2 V accumulator. The wire AB is 1.00 m long and its resistance is 4.20 Ω. A standard cell, e.m.f. 1.018 V, is connected in parallel with the resistance box, a galvanometer being included in series with the cell. The resistance R of the box is set at 509 Ω, and the variable resistor is then adjusted so that the p.d. across the box balances the e.m.f. of the standard cell. A source of a small e.m.f. E is now connected across a segment AP of the potentiometer wire, and the e.m.f. is found to be balanced by the potential gradient in the wire when AP = 62.3 cm.

Draw a circuit diagram, and calculate (*a*) the current flowing in the potentiometer wire, (*b*) the e.m.f. E.

The circuit is now slightly altered, in the following way: the value of the resistance box is set at 427 Ω and the standard cell, instead of being connected directly across the box, is connected across the box plus a length AP_0 of the potentiometer wire, where $AP_0 = 16.3$ cm. The variable resistor is again adjusted so that the e.m.f. of the standard cell is balanced.

Calculate the p.d. across the whole wire AB in this case, and suggest a practical advantage resulting from making this alteration to the circuit.

4.13 A metre bridge is balanced when a length of iron wire is placed in the left hand gap, and a standard 10 Ω resistor in the right hand gap. When the temperature of the iron wire is 20 °C, the balance point is 42.6 cm from the

left hand end of the bridge wire. The temperature of the iron wire is now raised to 100 °C. How may the bridge balance be restored (*a*) by adjusting the position of the slide contact, (*b*) by connecting a resistor in parallel with the iron wire? (Temperature coefficient of resistance of iron wire at $0°\ C = 6.0 \times 10^{-4}\ °C^{-1}$.)

4.14 A Wheatstone's bridge contains a standard 5 Ω resistance, the unknown resistance, and a metre length of wire of resistance 20 Ω. The balance point is found to be 35.0 cm along the wire, measured from the end to which the standard resistance is attached. What would be the value calculated for the unknown resistance?

It is subsequently found that, owing to faulty soldering, there is a contact resistance of 0.5 Ω between the metre wire and the terminal to which the unknown resistance is attached. What is the true value of the unknown resistance? How might the fault be detected? (OXF. SCH.)

4.15 A telegraph line ABC, which is 50 km long, has a fault due to earthing at an unknown point B. The end C of the line is joined to the end A through resistances R_1, R_2, R_3 in series, in that order. A battery is connected from the end A to the junction of the resistors R_1 and R_2, and a galvanometer is connected from the junction of R_2 and R_3 to earth. The resistance of R_1 is equal to that of 12.5 km of the telegraph line. When $R_2 = 1425\ \Omega$ and $R_3 = 1500\ \Omega$ the galvanometer is not deflected. What is the distance from A to the fault B?

4.16 Four 10 Ω coils AB, BC, CD, DA have a cell of e.m.f. 2.0 V and negligible internal resistance connected across AC.

(*a*) Calculate the change in the p.d. between B and D that occurs if an addition of 0.1 Ω is made to CD.

(*b*) The p.d. between B and D can be restored to the old value by connecting a resistor in parallel with one of AB, BC, or DA. Calculate the resistance of this resistor, and state the position in which it must be connected. (C)

4.17

The modified form of Wheatstone bridge circuit shown in the diagram is set up in order to measure the relatively small variations in the resistance R when its temperature is changed. AB is a uniform resistance wire 1.0 m long, with resistance 4.2 Ω.

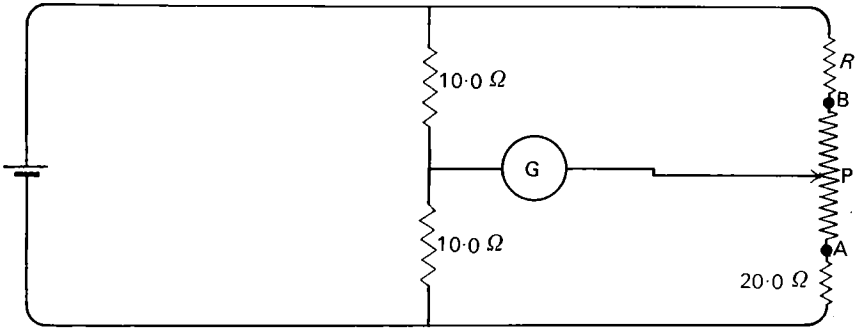


Fig. e 29

Obtain an expression showing the relation between the resistance R and the balance length x, assuming x to be measured in cm. What is the value of R when $x = 0$, 50 cm, 100 cm respectively? (AP = x)

4.18 A standard 11.0 Ω resistor is connected in one arm of a metre bridge circuit, and a standard 1.0 Ω resistor in the other arm. The balance point is found to be at 92.0 cm from one end of the wire. When the two resistors are interchanged, the new balance point is at 7.9 cm from the same end of the wire.

The discrepancy between these values and the true ratio of the two resistors is due to a fault in the soldered connection at each of the two ends of the bridge wire, the effective length of which is $(100 + x + y)$ cm, x cm being the zero error at one end, and y cm the zero error at the other.

Calculate x and y.

4.19 A Wheatstone bridge circuit ABCD contains equal resistances R in the arms AB and CD, and equal resistances R' in the arms AD and BC. An accumulator of e.m.f. E and negligible resistance is connected across AC. Calculate the p.d. between B and D.

If $R' = (1 + \lambda)R$, show that this p.d. is independent of the value of R and also is approximately proportional to λ when λ is small compared with unity. How would this p.d. be affected if the accumulator had an internal resistance $\frac{1}{2}R$? Would the use of a moving coil galvanometer to detect the p.d. between B and D affect its magnitude? (O & C)

*4.20 The slide wire of a potentiometer is 100 cm long and has resistance 7.0 Ω. It is connected by thick leads to an accumulator of e.m.f. 2.0 V and negligible internal resistance. A balance point is obtained for a Leclanché cell of e.m.f. 1.50 V and internal resistance 20 Ω. The galvanometer used has a resistance 30 Ω and a current sensitivity of 1 division per μA. What is the galvanometer deflection when the potentiometer slider is moved 0.10 cm from the balance point? (CAM. SCH.)

*4.21 State Kirchhoff's laws for currents flowing in a network of conductors.

'The first law is simply the law of conservation of charge; the second law is the law of conservation of energy.' Discuss the statement briefly.

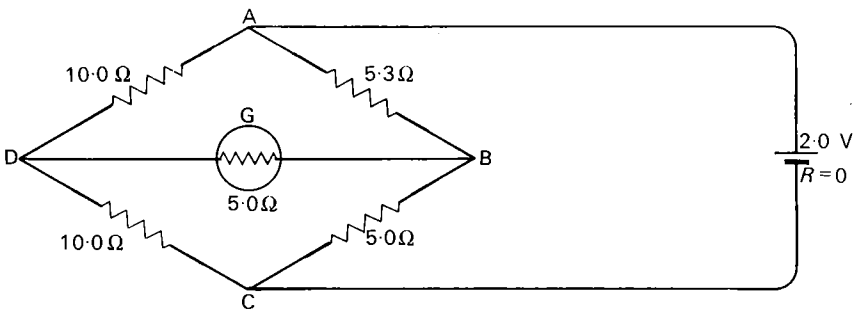


Fig. e 30

The diagram shows an unbalanced Wheatstone's bridge connected across a cell of e.m.f. 2.0 V and negligible internal resistance. Calculate (*a*) the p.d. between the points A and B, (*b*) the current flowing through the galvanometer G. (OXF. SCH.)

*4.22 A Wheatstone's bridge is made from four resistors each 100 Ω and a galvanometer, resistance 100 Ω, and a 2.0 V cell of negligible resistance.

If one of the resistors has temperature coefficient of resistance 4×10^{-3} °C^{-1} and is used as a thermometer, find the least change in temperature detectable if the galvanometer is sensitive to a current of 10^{-6} A. (CAM. SCH.)

*4.23 A Wheatstone's bridge ABCD has resistance R_1 between A and B and between B and C, R_2 between A and D, and R_3 between D and C. An ammeter of negligible resistance is connected across BD and a cell of negligible internal resistance across AC. The current in the ammeter changes from $+I$ to $-5I$ as R_3 is changed from a very large value to a very small value. What is the ratio of R_1 to R_2? Sketch a rough graph of the variation of ammeter reading against the values of R_3, and point out how the sensitivity of the bridge near balance may be obtained from such a graph. (OXF. SCH.)

*4.24 A Wheatstone's bridge consists of four resistors P, Q, R, and S, with a driving cell of e.m.f. E_1. The galvanometer is replaced by a second cell of e.m.f. E_2, connected as shown in the diagram.

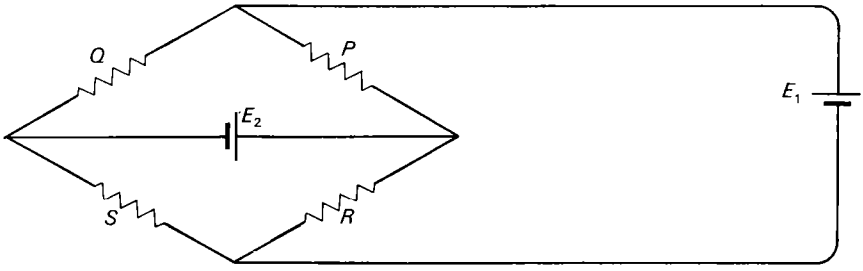


Fig. e 31

Find under what condition the current through P will be zero, assuming that the internal resistances of the cells are negligibly small. Show further that if the condition is to be achieved by adjustment of the resistor S only, E_1/E_2 must lie in the range 1 to R/Q. (O & C)

*4.25 Four resistances A, B, C, D are used to form a Wheatstone bridge. The join of A and D is connected to the positive terminal of a cell of 2.04 V e.m.f. and of negligible resistance, and the negative terminal of the cell is connected to the join of B and C. Resistance A is 620 Ω; resistance B produces a large thermal e.m.f. but may be considered to be a resistance of 380 Ω connected in series with a cell that generates an e.m.f. $E = \alpha I^2$, connected so that its positive pole is joined to the negative terminal of the 2.04 V cell. For small currents, $\alpha = 4 \times 10^3$ V/A^2. If balance is obtained with resistance $D = 530$ Ω, find the value of resistance C. (CAM. SCH.)

*4.26 In a certain metal rectifier the current is found to be proportional to the square of the applied p.d. (i.e. it cannot be described by Ohm's law). The rectifier is connected in one arm of a Wheatstone bridge in which each of the other arms is a resistance of 10 Ω. A 2.0 V battery is applied to the bridge, which is found to be balanced. The voltage of the battery is now increased to 6.0 V and the resistance in the arm opposite to the rectifier is changed until a balance is again obtained. What is the new value of this resistance? (CAM. SCH.)

*4.27 A Wheatstone bridge consists of four resistors AB, BD, AC, and CD. A 2.0 V cell is connected between A and D, and a galvanometer of resistance 20 Ω between B and C. Initially all the arms of the bridge are equal to 10 Ω. If the arms AB and CD are both increased by 1.0×10^{-4} Ω, how much current will flow through the galvanometer? (CAM. SCH.)

*4.28 The resistance per unit length of a potentiometer wire is found to vary lineally from 9 Ω/m at one end to 10 Ω/m at the other. The p.d. applied across the ends is 2.0 V. What is the p.d. between the low resistance end and the mid point of the wire? (OXF. SCH.)

Chapter 5

Electrolysis and cells

5.1 Electrovalency and an electrolyte

When we state that an electric current in a metal is due to a drift of electrons, we are making a claim with little prospect of establishing it by practical demonstration. With current through an electrolyte, the case is different, because the charge carriers are ions, and these make their passage obvious by their liberation at the electrodes. It was the study of this electrolysis which gave the first indication of the particle aspect of electric charge, some years before the electron was identified. Such considerations make an account of electrolysis an essential part of an account of electrical phenomena, although a detailed treatment is nowadays regarded as belonging to chemistry.

An electrolyte is essentially a compound, of which sodium chloride, NaCl, is a typical example, in which the atoms, or groups of atoms, are held together in the crystal lattice of its solid form by electrovalent bonds. The sodium and chlorine atoms in the salt crystal are not neutral atoms, but ions, Na^+ and Cl^-, formed because each chlorine atom has accepted one electron in excess of its neutrality at the expense of a sodium atom. It is the electric attraction between the ions of opposite sign which binds them together, and this is what is termed electrovalency.

The neutral sodium atom, with atomic number $Z = 11$, has one electron in its M shell, as pictured in Fig. 1.1, whereas the neutral chlorine atom has seven, its atomic number being 17; ten electrons in each case are taken up in filling the K and L shells. Thus, by the exchange of a single electron between them, both the sodium and the chlorine acquire the stable electron configuration of an inert gas with eight electrons in the outermost shell; the sodium has lost its lone M electron, and the chlorine has made its M shell up to eight. In attaining this stability, however, the sodium atom has become the positively charged sodium ion, Na^+, and the chlorine atom has become the negatively charged chlorine ion, Cl^-.

The crystal lattice of sodium chloride, therefore, consists of alternate Na^+ and Cl^- ions, vibrating about their equilibrium positions. Since the valency electron of the sodium atom has been taken up by the chlorine, there are no electrons free to carry charge through the crystal, and sodium chloride at room temperature is an insulator, becoming a conductor only when the ions can break free of their neighbours. This occurs when the crystal is melted, or goes into solution, or is vaporized, and the charge carriers are then the ions, Na^+ and Cl^-, not electrons.

When an electric current flows through molten sodium chloride, or through sodium chloride solution, the sodium ions, Na^+, carry positive charge to one electrode, and the chlorine ions, Cl^-, carry negative charge to the other. The result is the liberation of chlorine at one electrode, and of sodium, or of hydrogen displaced by sodium from the water, at the other electrode. This is electrolysis, and electrovalent compounds, which can be electrolysed in this way, are alternatively called electrolytes.

Some ionic compounds, such as glass, are put to use as insulators. If glass is heated, however, it becomes a conductor a little before reaching its melting point, and molten glass is a good conductor. The insulation is broken down, also, if the material is subjected to a very powerful electric field, which amounts to saying that there is a limit to the voltage which it can withstand.

5.2 Faraday's laws of electrolysis

Figure 5.1 shows a circuit in which three electrolysis vessels, or voltameters, are connected in series. The three electrolytes have ions represented by the equations

$$NaCl = Na^+ + Cl^-,$$
$$H_2SO_4 = 2H^+ + (SO_4)^{2-},$$
$$CuSO_4 = Cu^{2+} + (SO_4)^{2-}.$$

These equations involve the assumptions that electrons can be taken over by a group of atoms, e.g. the sulphate radicle, $(SO_4)^{2-}$, and that valency in an electrovalent compound is a matter of the number of electrons which the atom or radicle has gained or lost in becoming an ion. The monovalent sodium loses one electron from its neutral atom; copper, to form the cupric ion, Cu^{2+}, loses two.

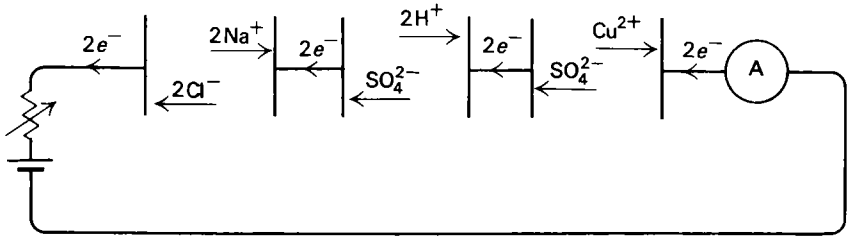


Fig. 5.1 Electrolysis

Two electrons are represented in the diagram as arriving at the cathode of each voltameter. To complete the circuit, two monovalent positive ions (cations) or one divalent positive ion must in the same time arrive at each cathode from the electrolyte, and two monovalent negative ions (anions) or one divalent negative ion must arrive at each anode.

This result is the substance of Faraday's laws of electrolysis, which are:

1 The mass of substance liberated at an electrode by electrolysis is proportional to the product of current and time, i.e. to the amount of charge which has been carried round the circuit.

2 When the same current is passed for the same period through different electrolytes, the mass of substance liberated at any electrode is proportional to the quotient relative atomic mass/valency.

The close relation between these laws and the ionic theory, including the theory that there is a basic unit of electric charge, is self evident.

5.3 Some definitions

Relative atomic mass A_r (atomic weight)

In comparing the masses of different atoms or ions, the standard adopted for comparison is the mass of the atom of the carbon isotope, $^{12}_{6}C$, which is the most abundant isotope of carbon, and has 6 protons and 6 neutrons in the nucleus.

A quantity of mass m_u, called the unified atomic mass unit, is defined as 1/12 of the mass of one atom of ^{12}C, i.e., in the case of ^{12}C, we can write

$$A_r = \frac{m_a}{m_u} = 12 \text{ exactly},$$

where m_a is the mass of the carbon atom, and A_r is its relative atomic mass, referred, to, until very recently, as its atomic weight. The traditional name is misleading however, because it is not a weight at all, but a ratio.

The relative atomic mass of any other element, likewise, is the ratio of the mass of its atom to that of the atomic mass unit, and we can write, for the general case,

$$A_r = \frac{m_a}{m_u}, \tag{5.1}$$

or

$$A_r = 12\frac{m_a}{m_a(^{12}C)}.$$

The ratio in mass, for example, of the atoms of hydrogen and ^{12}C is $\frac{1.008}{12}$, whence the relative atomic mass of hydrogen is

$$A_r = 1.008.$$

The mole

The mole is a unit used to measure equal numbers of atoms, molecules, ions, etc. of different substances.

One mole of the carbon-12 isotope is, by definition, the amount of carbon which has a mass of exactly 0.012 kg.

One mole of any other substance is the amount of substance which contains as many elementary units (i.e. atoms, molecules, ions, etc., as the case may be) as there are atoms in 0.012 kg of carbon-12.

e.g. 1 mole of hydrogen atoms has mass equal to 1.008×10^{-3} kg.

The Avogadro constant N_A

The number N_A of elementary units contained in 1 mole, e.g. the number of atoms in 0.012 kg of ^{12}C, is called the Avogadro constant. Its value is

$$N_A = 6.023 \times 10^{23} \text{ mol}^{-1}.$$

5.4 The Faraday constant F

Applying the above definitions to the case of electrolysis, we find that Faraday's second law leads to the equation

$$F = \frac{Q}{n}, \tag{5.2}$$

where Q is the quantity of charge which flows through the electrolyte, and n is the number of moles of monovalent ions which are liberated at one electrode in consequence.

The quantity F is a constant, which is called the Faraday constant. Its value is

$$F = 9.649 \times 10^4 \text{ C/mol}^{-1}.$$

Assuming the ionic explanation of electrolysis to be correct, that the charge carried by a single monovalent ion is identical with the electronic charge e, it follows that the value of F is given by the relation

$$F = N_A e. \tag{5.3}$$

This relation gives the basis for one accurate determination of the electronic charge e, the Avogadro constant N_A being determined by measurements involving the diffraction of X-rays by a crystal. We have

$$e = \frac{F}{N_A}$$

$$= \frac{9.649 \times 10^4}{6.023 \times 10^{23}} \text{ C}$$

$$= 1.60 \times 10^{-19} \text{ C}.$$

Example

Calculate the mass of copper which would be deposited on the cathode by electrolysis of copper sulphate solution with the passage of 1 coulomb of electric charge.

The number of moles of a monovalent ion which would be liberated is given by equation (5.2)

$$n = \frac{Q}{F}$$

$$= \frac{1}{9.65 \times 10^4} \text{ mol.}$$

But the cupric ion Cu^{2+}, is divalent, i.e. each ion carries charge $+2e$. Hence the number of moles of copper atoms deposited on the cathode is

$$\frac{n}{2} = \frac{1}{2 \times 9.65 \times 10^4} \text{ mol.}$$

The relative atomic mass of naturally-occurring copper (which is a mixture of isotopes) is

$$A_r = 63.6,$$

whence, the mass of 1 mole of copper atoms is

$$M = 63.6 \times 10^{-3} \text{ kg.}$$

The total mass of copper liberated by the passage of 1 coulomb is

$$m = \frac{n}{2} \times M$$

$$= \frac{63.6 \times 10^{-3}}{2 \times 9.65 \times 10^{-4}} \text{ kg}$$

$$= 3.29 \times 10^{-7} \text{ kg.}$$

5.5 The migration of ions

According to the kinetic theory, the ions in solution, or in the molten electrolyte, are in a condition of continuous random motion. When an electric field is set up between the two electrodes, the current is due to a drift of the ions in the field, superimposed on this random motion, positive ions drifting towards the cathode with average speed v_1, and negative ions towards the anode with average speed v_2. Taking the simplest case of monovalent ions, the current is given by the equation

$$I = NA(v_1 + v_2)e, \tag{5.4}$$

where N = the number of ion pairs in unit volume,
A = the conducting area perpendicular to the current flow,
e = the charge carried by each monovalent ion,
v_1, v_2 = the mean drift speeds of positive and negative ions.

The number of charge-carrying ions in unit volume is very large, and the speeds v_1, v_2 which correspond to quite a considerable current are only very small. A typical value is about 3×10^{-6} m/s. The slowness of this movement can be explained on the basis that the positive ions, moving in one direction, experience a drag due to the negative ions moving in the opposite direction, and vice-versa. There is also the attraction between ions and the molecules of solvent, i.e. the effect of viscosity.

5.6 Preferential discharge

The ions which carry the current through a solution of electrolyte are not always identical with the ions which are liberated at the electrodes. One example of this is the electrolysis of sodium chloride solution. In this case, the current is carried by Na^+ and Cl^- ions, and the chlorine comes off at the anode, but hydrogen, and not sodium, is liberated at the cathode. This is accounted for by the slight ionization of the water, which provides some hydrogen ions (or, more accurately, oxonium ions, $(OH_3)^+$) and these take preference over the sodium ions in accepting an electron at the cathode, a process known as preferential discharge. More water then ionizes, to replace the liberated hydrogen ions, the net effect being the replacement of chlorine ions, Cl^-, by hydroxyl ions, $(OH)^-$, in the solution, which is the conversion of sodium chloride to sodium hydroxide

$$2Na^+ + 2Cl^- + 2H_2O \rightarrow 2Na^+ + 2(OH)^- + Cl_2 + H_2.$$

Thus, preferential discharge depends on the relative ease with which different positive ions retrieve their missing electron at the cathode, and with which negative ions part with their surplus electron at the anode. The hydrogen ion, for example, regains its electron at the cathode more readily than the sodium ion; the chlorine ion, Cl^-, loses its electron at the anode more readily than the hydroxyl ion.

5.7 Polarization of a voltameter

An instructive comparison can be made between a voltameter, in which electrolysis is carried out, and an electric cell, which acts as a source of current. Both consist of two electrodes inserted in a solution of electrolyte.

When a voltameter is first set up, the arrangement is symmetrical, both electrodes being of the same material. If a chemical change results from the electrolysis, the energy conversion which accompanies it is from electrical to chemical form. An example of this is the electrolysis of dilute sulphuric acid, using electrodes of platinum or of carbon, the result of which is to break water up into its component hydrogen and oxygen; the gain in chemical energy is obvious from the fact that the evolved gases form an explosive mixture in their readiness to turn back into water.

An electric cell, on the other hand, is essentially unsymmetrical, the two electrodes being made of different materials placed in a solution of an electrolyte. Resulting from this lack of symmetry there is usually a p.d. of the order of 1 V between these electrodes when on open circuit, and the cell is capable of maintaining an electric current. The resulting chemical reaction in the cell, as compared with that in a voltameter, converts energy from chemical to electrical form, and this is the source of the e.m.f.

The energy conversion in the cell is a reversible one, which is typical of the energy change associated with an e.m.f.; if the direction of the current is reversed by applying a suitable p.d. from an external source, the chemical reaction is reversed, the energy required for this being provided electrically because the current is flowing against the direction of the e.m.f. of the cell. In order to account for the electrical energy taken up in a voltameter in the process of electrolysis, there must be a similar back e.m.f. set up in the voltameter, and this appears as a result of the behaviour of ions released by the electrolysis. Taking the example of the electrolysis of dilute sulphuric acid, not all the evolved gases escape; a thin film of hydrogen coats the cathode, with the result that the previously symmetrical voltameter has been converted into an unsymmetrical cell, with an e.m.f. which opposes the flow of current. The voltameter is said to have become polarized, and the energy changes which accompany the electrolysis are represented by the equation

$$VI = EI + I^2R, \tag{5.5}$$

where V is the applied p.d. across the voltameter, E the back e.m.f. due to polarization, and R the resistance of the electrolyte. The quantity EI is the power converted into chemical energy, the remainder, I^2R, being converted to heat.

Because of this polarization e.m.f., the current through the sulphuric acid voltameter, plotted against values of the applied p.d., gives the graph shown in figure 5.2, the equation being

$$I = \frac{(V-E)}{R}, \tag{5.6}$$

where, in this particular case, E is about 1.5 V. This behaviour is typical of electrolysis, except in the comparatively rare cases where there is no overall chemical change.

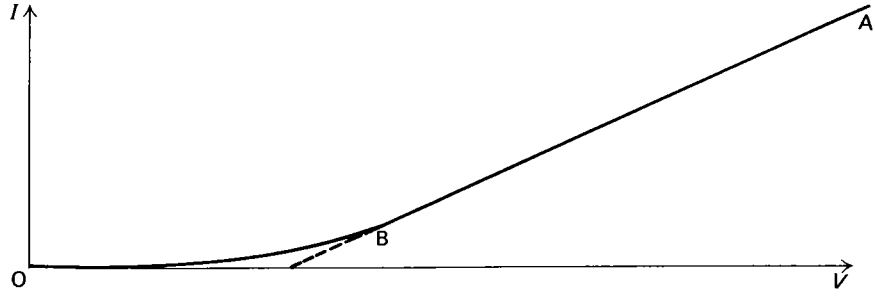


Fig. 5.2 Electrolysis of dilute sulphuric acid, using platinum electrodes. The variation of current with p.d. (The shape of the curve between O and B may be explained on the basis that the polarization is rendered partially ineffective by diffusion of polarizing hydrogen layer away from the electrode

Example

The combustion of hydrogen and oxygen releases 1.4×10^8 J *of energy for every kilogram of hydrogen. Estimate the value of the back e.m.f. resulting from polarization when dilute sulphuric acid is electrolysed to produce hydrogen and oxygen from water.*

We have, from equation (5.2)

$$Q = Fn,$$

where Q is the quantity of electric charge which passes, F, the Faraday constant, is 9.65×10^4 C/mol, and n is the number of moles of hydrogen atoms released by electrolysis.

But 1 mole of hydrogen atoms $= 1.008 \times 10^{-3}$ kg,

therefore 1 kg of hydrogen atoms = 10^3 mol, approximately,

$$\text{and } \frac{Q}{m} = 9.65 \times 10^7 \text{ C/kg.}$$

The gain in chemical energy resulting from the electrolysis is assumed to be equal to the heat released by combustion when the reaction proceeds the other way,

i.e. $$\frac{W}{m} = 1.4 \times 10^8 \text{ J/kg.}$$

The back e.m.f. must therefore be

$$E = \frac{W}{Q}$$

$$= \frac{1.4 \times 10^8}{9.65 \times 10^7} \text{ V}$$

$$= 1.5 \text{ V.}$$

5.8 Measurement of the resistance of an electrolyte in solution

In measuring the resistance of an electrolyte, steps must, in general, be taken to allow for the possible existence of a polarizing e.m.f., and this can best be done by preventing its appearance. A suitable method is to set up a Wheatstone bridge circuit, with the voltameter in one branch, and using an alternating source of audio-frequency to energize the bridge, say 1 000 or 2 000 Hz. No ions are liberated at this frequency, so there is no polarization. The galvanometer used to detect the balance condition must respond to the alternating current and a telephone receiver is often used for this purpose.

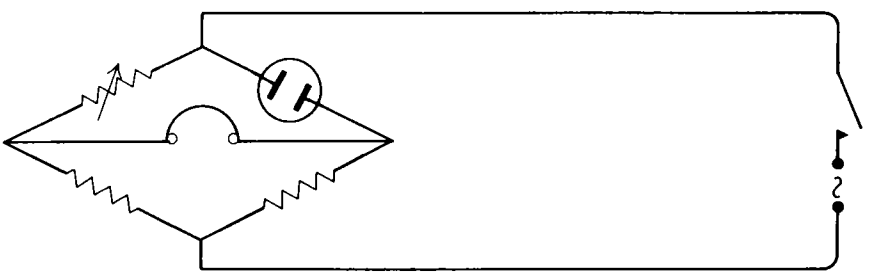


Fig. 5.3 Arrangement to measure the resistance of an electrolyte

5.9 The variation in resistance of an electrolyte with temperature

If the temperature of an electrolyte in solution is raised, the effect is to decrease the resistance, because the viscosity of the liquid has been reduced and this allows the mobility of the ions to increase. This is the reverse of what happens in a metal, where temperature rise lowers the average drift speed of the conduction electrons, causing the resistance to go up.

5.10 Electrode potential

We have stressed the point that a symmetrical voltameter has no e.m.f., and produces one only if polarized, thus losing its symmetry and acting as a cell. This

does not explain, however, how the e.m.f. arises, and an attempt to do this leads to the postulate of an electrode potential. This is a p.d. which is supposed to be set up between an electrode and an electrolyte due to an exchange of ions between the two. Thus, if atoms of the electrode were to enter the electrolyte, they would do so in the form of positive ions, and this would leave the electrode negative; negative ions leaving the electrolyte would likewise make the electrode negative, and positive ions leaving the electrolyte would make the electrode positive. The electrode potential represents the equilibrium situation.

In the symmetrical voltameter, the two electrode potentials would be equal and opposite, cancelling each other out, but in a polarized voltameter or a cell they would be unequal, giving rise to an e.m.f.

5.11 Polarization of a cell

If electrolysis can change the effective electrodes of a voltameter owing to a surface layer of liberated ions, the same effect can occur in a cell, and when this happens the result is to lower the e.m.f. of the cell and to increase its resistance. Unlike the case of the voltameter, however, polarization of a cell is not an essential part of the process by which it works, and proper design of a cell either prevents polarization or removes it as soon as it occurs. There is not room here to describe any but a small selection of practical cells; in each of these the prevention of polarization is a noticeable feature.

5.12 The Leclanché cell

The Leclanché cell has the advantage that its materials are relatively non-corrosive and this makes it suitable for use in a 'dry' form, which is the familiar dry cell. The positive and negative electrodes are respectively carbon and zinc, giving it an e.m.f. of 1.5 V, and the electrolyte is ammonium chloride, NH_4Cl.

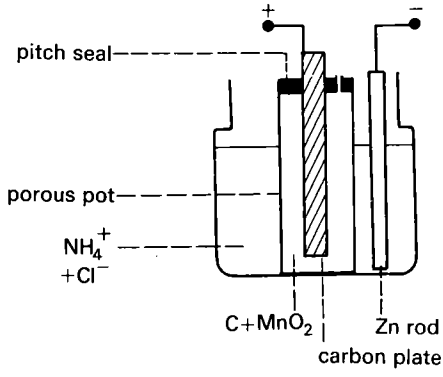


Fig. 5.4 The Leclanche cell (a) wet cell

When the cell is supplying current, zinc ions, Zn^{2+}, go into solution from the zinc electrode, and ammonium ions, NH_4^+, are discharged at the carbon, forming ammonia and hydrogen, thus:

at the zinc electrode: $Zn + 2Cl^- \rightarrow Zn^{2+} + 2Cl^- + 2e^-$,

at the carbon electrode: $2NH_4^+ + 2e^- \rightarrow 2NH_3 + H_2$.

The carbon electrode tends to become polarized by a film of hydrogen, and the

depolarizing agent used to remove this is manganese dioxide, MnO_2. Because this is an insoluble powder, very little of it would be sufficiently close to a solid carbon rod to have effect, so the manganese dioxide is mixed with powdered carbon and packed round the carbon rod, with the result that the whole of the mixture is effectively the positive electrode.

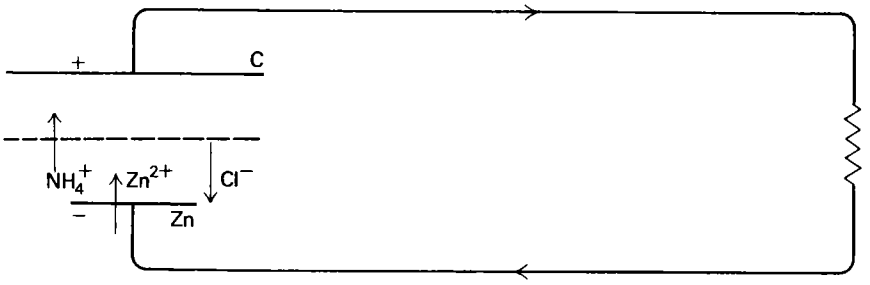


Fig. 5.4 (b)

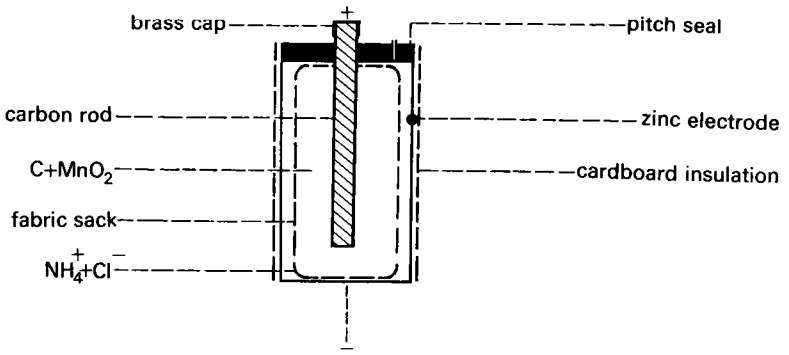


Fig. 5.4 (c)

Fig. 5.4 The Leclanché cell (b) The migration of ions (c) One type of dry cell

In the dry cell, the negative zinc electrode is used also as the cylindrical container, and this is lined with a quite thin layer of the ammonium chloride electrolyte made into a paste with zinc chloride, which is deliquescent and keeps it moist. The geometry reduces the resistance of the cell because in crossing through the electrolyte the current merely has to flow between the large area of zinc and the almost equally large surface area of the carbon and manganese dioxide mixture.

5.13 The Weston cadmium cell

This cell is designed with the single purpose of producing a standard, the e.m.f. being accurately consistent provided that the cell is made up according to specification. Also, its e.m.f. varies only slightly with temperature. The standard cell is virtually useless as a source of current or energy, and any attempt to use it for such a purpose would quickly destroy it as a standard; in any case, its resistance is very high, being several hundred ohms.

The positive and negative electrodes are respectively mercury and cadmium, and the electrolyte is cadmium sulphate in saturated solution. Mercurous sulphate

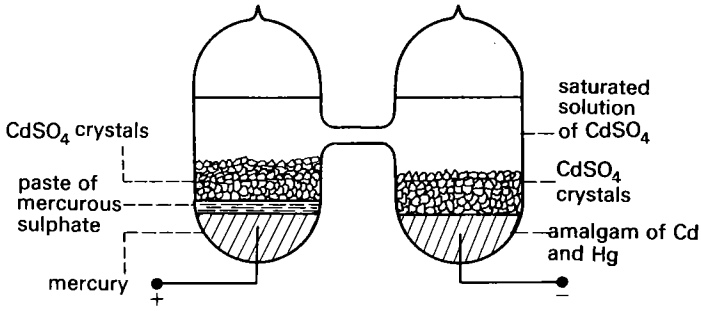


Fig. 5.5 The Weston cadmium cell

paste, which covers the mercury, acts as the depolarizing agent by preventing Cd^{2+} ions from reaching the mercury, since they displace mercury from the mercurous sulphate instead.

5.14 The lead accumulator

In theory, the direction of the energy conversion can be reversed in any cell by using a p.d. from an external source to reverse the direction of the current. In practice, however, in most cases the products of the forward reaction are obtained in the wrong places for them to be available for the reversed process. For this reason the dry cell is available only as a primary cell, meaning one which exhausts its chemicals once and for all. This is an expensive, even if convenient, way to obtain electricity.

The lead accumulator is an example of a secondary cell, being so designed that the energy conversion can be entirely reversed, so that it can repeatedly be charged and discharged. The charging process is, of course, an example of polarization resulting from electrolysis, the polarization e.m.f. being then utilised as the e.m.f. of the charged accumulator. The charged cell has positive and negative electrodes respectively of lead dioxide, PbO_2, and lead, both obtained in a spongy form held on special gridded plates of lead, so that the electrolyte can penetrate them, and the electrolyte is a fairly concentrated solution of sulphuric acid.

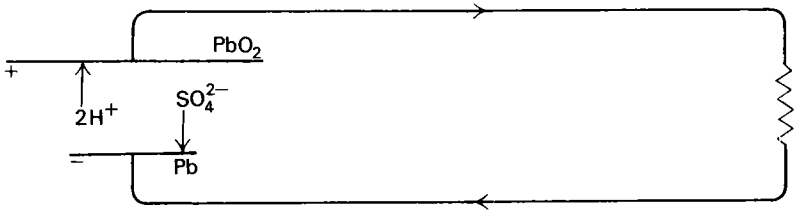


Fig. 5.6 The discharging lead accumulator

The reactions which takes place in the discharging cell are:

at the negative electrode: $Pb + SO_4^{2-} \rightarrow PbSO_4 + 2e^-$,

at the positive electrode: $2e^- + PbO_2 + 4H^+ + SO_4^{2-} \rightarrow PbSO_4 + 2H_2O$,

the overall chemical change being

$$Pb + PbO_2 + 2H_2SO_4 \rightarrow 2PbSO_4 + 2H_2O.$$

There is no polarization because the lead peroxide is its own depolarizing agent. Also, the lead sulphate forms as spongy material on the gridded plates, with the result that the discharged cell is in a condition suitable for charging it up again by passing a reverse current from an external source.

The e.m.f. of the cell remains almost constant, at about 2 V, throughout its discharge, and begins to fall only when a recharge is becoming due. This is an advantage of this type of secondary cell, but it means that measurement of the e.m.f. is no reliable guide to its state of charge; this is better tested by measuring the relative density of the acid, because this becomes progressively more dilute as discharge proceeds, owing to sulphate ions leaving the electrolyte for the electrode.

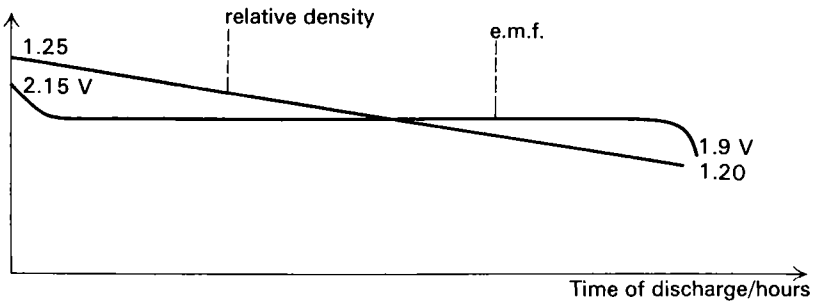


Fig. 5.7 The lead accumluator. The variation of e.m.f. and acid density with discharge time

A matter of practical importance when recharging the accumulator is the effect of carrying the process too far, or too rapidly for the chemistry of the electrodes. In either case, the effect is to release hydrogen and oxygen at the electrodes in explosive proportions, which accounts for the regulations to which a battery-charging room is subject.

5.15 The nickel-iron (NIFE) cell

The nickel-iron accumulator has advantages over the lead type in that it can withstand much rougher treatment, both mechanical and electrical, and it can be left for long periods without attention, a neglect which would destroy a lead accumulator. It has disadvantages too, however, being more expensive and having the lower e.m.f. of about 1.2 V, which falls steadily in value as discharge proceeds.

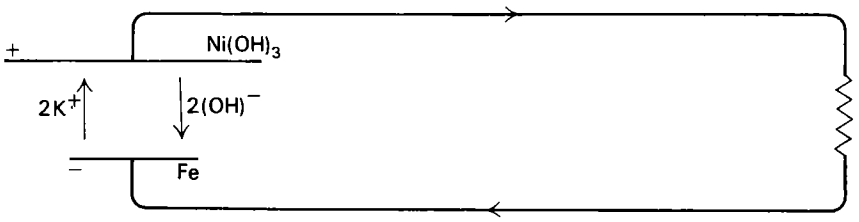


Fig. 5.8 The discharging Ni–Fe cell

The charged electrodes of this cell are nickelic hydroxide, $Ni(OH)_3$, (positive), and finely powdered iron (negative), both held in structures of steel, and the electrolyte is potassium hydroxide, KOH. The reactions at discharge are:

at the positive electrode: $2e^- + 2Ni(OH)_3 + 2K^+ \rightarrow 2Ni(OH)_2 + 2K^+ + 2(OH)^-$
at the negative electrode: $Fe + 2(OH)^- \rightarrow Fe(OH)_2 + 2e^-$.

It will be noticed that the electrolyte concentration remains constant throughout. The nickel cadmium cell is rather similar to the Ni–Fe, using cadmium in place of iron.

5.16 Fuel cells

By converting chemical energy directly to electrical energy, the primary cell has the possibility of being an exceptionally useful source of electric power. In practice, however, it can be used on only a small scale, because in all practical types the electrodes, which are the energy source, are expensive, and because a large power cell would be prohibitively large and heavy. Similarly, a secondary cell is a means of storing electrical energy, and there is no known means of doing this on a large scale, but the capacity of the secondary cell is limited by bulk and weight. Even an electrically propelled motor vehicle remains, so far, an unrealised proposition for the motor industry, except in specialized circumstances. Yet it would have many important advantages.

A fuel cell is a type of cell conceived to go beyond such limitations, the principle being to bring a conventional fuel into reaction with air or oxygen, not by burning it, but by using the fuel as one electrode of a cell, and air or oxygen as the other. This would be done by continuously feeding fuel and oxygen into contact with the electrolyte at plates which act, not as consumable electrodes taking part in the chemistry, but merely as essential mechanical parts of the construction, and perhaps as catalysts. The use of common fuels such as hydrocarbons reacting with air has not yet been achieved as a commercial proposition, but there has been more success using hydrogen and oxygen. The Bacon fuel cell, working with these two gases, is used as the electrical source for operating controls in Apollo spacecraft; the water produced as a result of the reaction is kept separate from the electrolyte, and used to provide drinking water in the spacecraft.

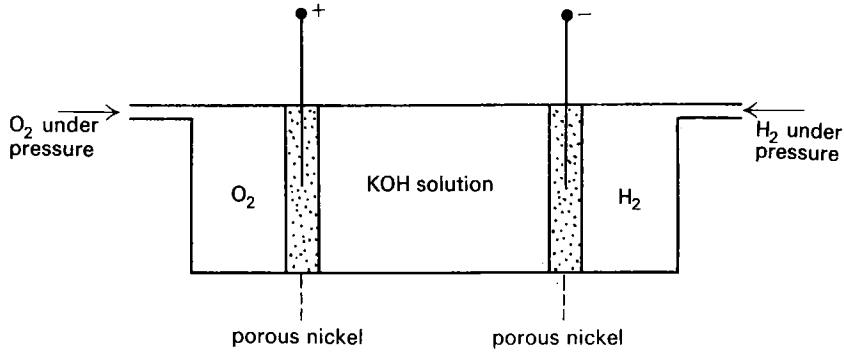


Fig. 5.9 The principle of the Bacon fuel cell. Means of extracting the water are not shown in this simple diagram

The Bacon fuel cell uses potassium hydroxide in solution as the electrolyte, and the electrodes are porous nickel plates, in which the hydrogen and oxygen, under a very considerable pressure, come in contact with the electrolyte. The principle is illustrated in Fig. 5.9.

QUESTIONS

5.1 The cell shown in the circuit has e.m.f. 2.0 V and internal resistance 1.0 Ω. Assuming the electrode potential at B to be −0.5 V, and that at C to be + 1.5 V, sketch a graph to show the potential at various points along the line ABCDE (*a*) when the switch S is open, (*b*) when the switch is closed.

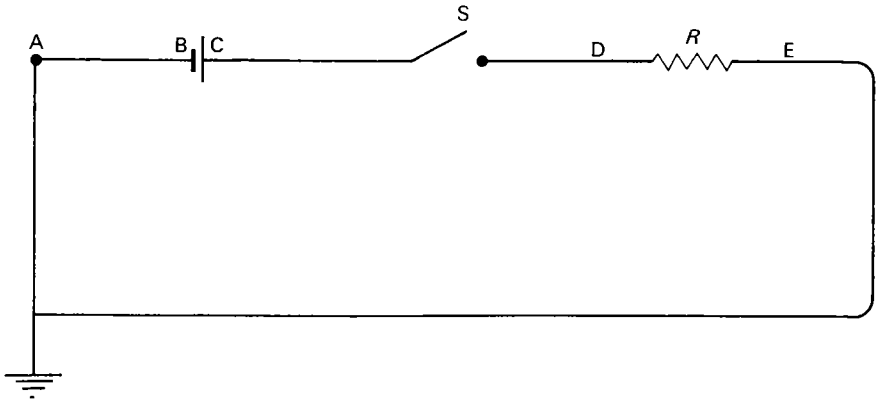


Fig. e 32

5.2 A voltameter with large platinum plates contains dilute sulphuric acid in which 10^{-2} g of metallic copper has been dissolved. A constant current of 10^{-2} A is then passed through the solution for 25 min. Calculate (*a*) the mass of copper deposited on the cathode, (*b*) the additional time for which the same current must be passed before hydrogen is liberated at the cathode.

(Faraday constant = 9.6×10^4 C/mol
1 mole of copper atoms = 63.6 g.
The cupric ion, Cu^{2+}, is divalent.) (O & C)

5.3 A water voltameter and a copper voltameter were connected in series with an ammeter and a variable number of accumulators, each of e.m.f. 2.0 V and negligible resistance. The following readings of current I and the number of accumulators N were taken.

N	4	3	2	1
I/mA	120	80	40	5

When the water voltameter was removed, and the experiment was repeated with the copper voltameter alone, the corresponding readings were:

N	3	2	1
I/mA	220	150	75

Plot the appropriate graphs and explain their general appearance. Calculate values for the resistances of the copper voltameter and the water voltameter. (C)

Chapter 6

Basic electrostatics

6.1 The process of charging a body electrically

The point was made in chapter 1 that when electric charge is given to a body the process consists of separating positive charge from negative charge; one body cannot be charged up negatively without an equal positive charge appearing somewhere else as a result. The claim was made that this is so without any evidence being quoted to support it, but it is the fundamental process of electrostatics and several experiments will be described in this chapter which illustrate it. One of these, which will serve to take us across the line from current electricity to electrostatics, uses the arrangement shown in Fig. 6.1, where a battery is connected across a system of two parallel metal plates (a capacitor).

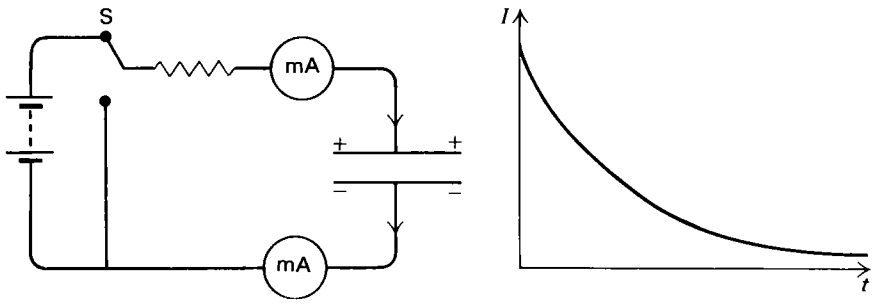


Fig. 6.1 Transfer of electric charge

When the switch is first made, what happens is that current flows on to one plate and off the other, charging one plate positively and the other negatively by equal amounts. The current dies away in strength as the plates progressively charge up, finally ceasing entirely when the p.d. between the two plates is equal to that across the battery. Apart from this fact that it dies away, the current obeys the same rules as it would in any other circuit; at any instant the value of the current is the same at all parts of the circuit (except, of course, in the space which separates the plates) and both meters always agree in their readings. Similarly, if the switch is now moved across to discharge the metal plates, the charge flows back whence it came through the two meters, until the discharging process is complete.

Another feature of this experiment, which is typical of electrostatics, is the fact that the total amount of charge which is transferred from one plate to the other is only very small. This is because the force with which unlike charges attract each other is so great that the p.d. produced between them when they are separated

soon builds up to a high value – a fact which accounts for the electric sparks which often accompany experiments in electrostatics.

6.2 Electric field and lines of force

The electric field which surrounds a charged body is defined in terms of the force experienced by a second charged body if it is placed in the field, and this is expressed by the equation

$$E = \frac{F}{Q}, \tag{6.1}$$

i.e. the electric field strength E is defined as being the force which acts per unit charge, and its SI unit is newton per coulomb.

If electric lines of force are drawn to mark the direction of the field, meaning the direction of the force which would act on positive charge, they can also be used to indicate its strength, by having them concentrated together at places where the field is strong. This is arranged in the following way.

Suppose that, where there is a uniform field of strength E, we consider an area A which is perpendicular to the field and make N lines of force cross this area. Then the number N is so chosen that we can write

$$E = \frac{N}{A}, \tag{6.2}$$

i.e. the field strength E is represented by the number of lines which cross unit area.

Where the field is not uniform, the field strength E at any point is represented by taking a small area δA perpendicular to the field and containing the point, and having a small number δN of lines crossing this area, such that

$$E = \frac{\delta N}{\delta A}. \tag{6.3}$$

A typical diagram of lines of force is shown in Fig. 6.2, and in this we endow the lines with certain properties, while having to postpone any attempt to justify ourselves until a later chapter. These properties, which are in fact consistent with the inverse square law of force between point charges (chapter 1.3) are as follows:

(*a*) Electric lines of force are continuous lines which start from a positive charge and end on a negative charge.

(*b*) The number of lines of force which start from a positive charge $+Q$, or

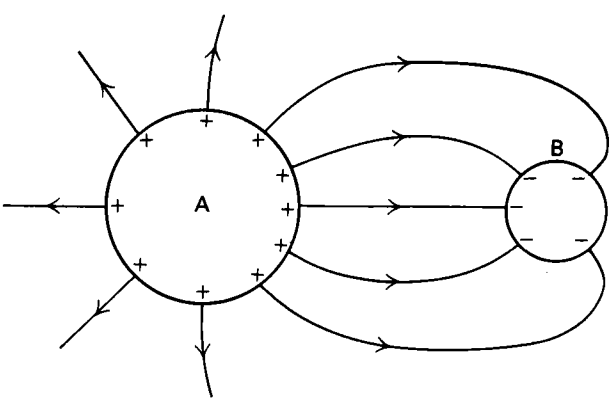


Fig. 6.2 A diagram showing electric lines of force

end on negative charge $-Q$, is proportional to the magnitude of Q. Thus, in the diagram, 10 lines are shown leaving the positively charged body A, while 5 lines finish up at the negatively charged body B. It would be correct to deduce that the total charge on A is twice as great as that on B and opposite in sign.

6.3 Induced charge

If a charged body is brought up close to a conductor, it causes a separating out of charge in the conductor, by attracting unlike charge to the near side and repelling like charge to the far side, a process which is termed charging by induction.

If the conductor is momentarily connected to earth while the charged body is still present, the charge of like sign is repelled to earth, and only induced charge of opposite sign is left on the conductor (Fig. 6.3(*a*)), and remains there if the charged body is removed. Alternatively, if the conductor is made in two halves, these may be separated from each other while still in the influence of the charged body, thus preventing the two charges from flowing back to neutralize each other when the charged body is removed (Fig. 6.3(*b*)).

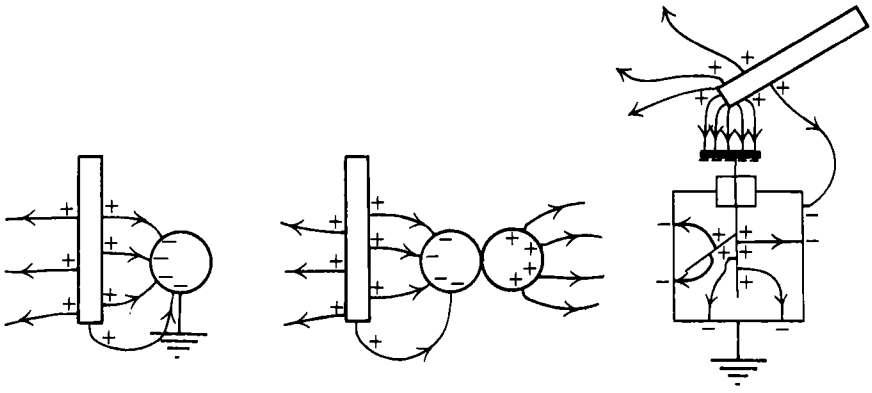


Fig. 6.3 Examples of charging by induction (a) (b) (c)

It is by this process of electrostatic induction that a gold leaf or similar electroscope registers the effect of a charged body held near by, and this is an aspect of induction which is of importance to demonstrations in electrostatics. As Fig. 6.3(*c*) suggests, the extent to which the gold leaf will rise depends on how many of the lines of force from the charged body end up on the electroscope, and a true indication of the size of the charge will be given only if all the lines of force from the body end up in this way. This was a conclusion which was reached by Faraday in the famous series of experiments to be described in the next section, which are known as Faraday's ice-pail experiments.

6.4 Faraday's ice-pail experiments

If the electroscope is to collect all the lines of force from a charged body, it must evidently be connected to a conductor which completely surrounds the charge, and the nearest to this ideal that Faraday achieved was to lower the charged body well down inside a deep metal vessel. This happened to be a metal ice pail, from which the experiments take their name.

The first sequence of operations is shown in Fig. 6.4. While the charged body is near the surface of the ice pail, many of its lines of force escape from the top and induce charge elsewhere, with the result that the rise of the gold leaf is only partial. Once the body is well down inside the pail, however, the number of lines which escape is negligible, and the rise of the leaf gives a true indication of the size of the charge, and is quite independent of the exact position taken up by the body inside the ice pail. If the charged body is now taken out of the ice pail again, without ever having made contact with it, the leaf, of course, falls to zero again, since there were equal and opposite charges induced on the inside and the outside of the pail (or on the outside plus the electroscope).

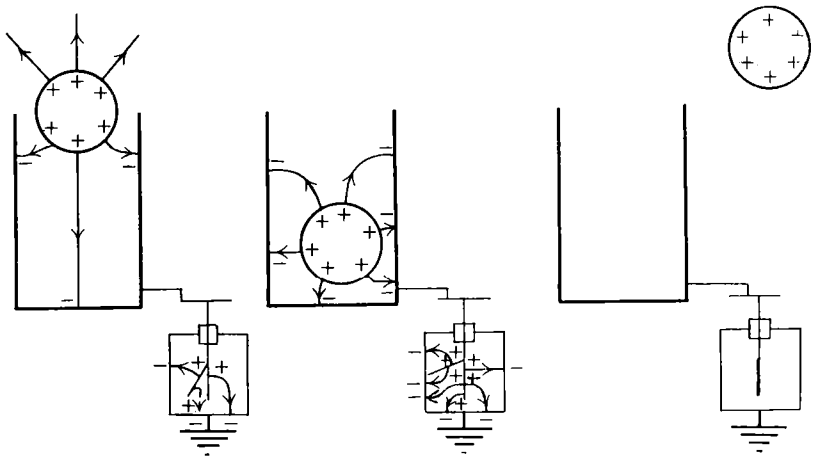


Fig. 6.4 The first of Faraday's ice-pail experiments

This, the first of Faraday's experiments, has a very useful application in electrostatics experiments when charges are to be compared in magnitude and in sign. Thus, if charges Q, Q' are lowered in succession into the ice pail, the rise of the leaf indicates the relative magnitudes of Q and Q'; if they are lowered in together, the leaf indicates the resultant charge to be $(Q + Q')$ or $(Q - Q')$, dependent on their relative signs. Nor does it matter whether the charged bodies are conductors or insulators, since they do not part with their charges.

The second of Faraday's ice-pail experiments is illustrated in Fig. 6.5, and in this case the charged body must be a conductor. It is lowered well inside the ice pail, and this time it is brought in contact with the ice pail. There is no abrupt rise or fall of the leaf to betray the instant at which contact was made. If the body is now taken out of the ice pail, however, the leaf stays up, which it did not do in the first experiment. If the ice pail is now discharged, and the body lowered into it a second time, the leaf fails to rise at all, showing that the body lost all its charge to the ice pail when it was first contacted to it.

The interpretation given to this result is that, when the charge $+Q$ was lowered well down inside the ice pail, the induced charges were equal to $-Q$ on the inside of the pail (because it collected all the lines of force) and $+Q$ on the outside of the pail and the electroscope. Thus, when the charged body made contact with the ice pail, its own charge was exactly neutralised by the induced charge $-Q$ on the inside of the pail, leaving the whole charge $+Q$ on the outside of the pail and on the electroscope.

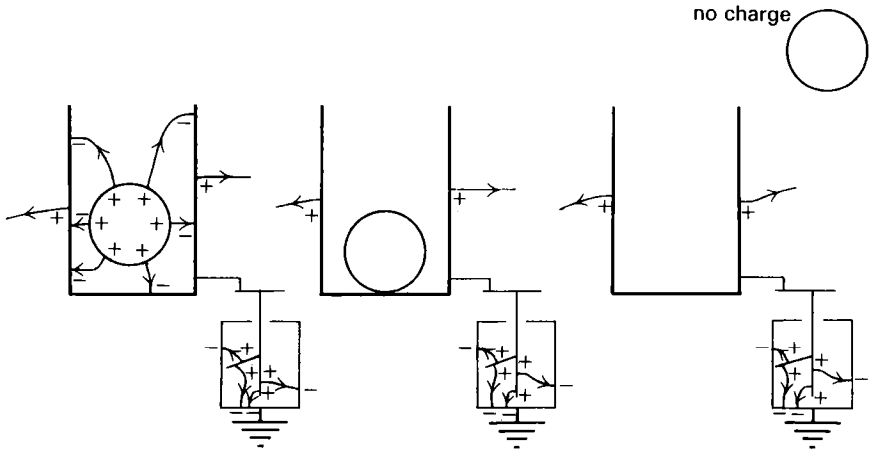


Fig. 6.5 The second of Faraday's ice-pail experiments (a) (b) (c)

A useful application of this result in practical electrostatics is in providing a method by which a charged conductor may be made to surrender all its charge to a second conductor, provided that this is hollow, instead of merely sharing it.

A further experiment, which forms a natural extension of the two carried out by Faraday, consists of momentarily earthing the ice pail while a charged conductor is well down inside the pail, but not making contact with it.

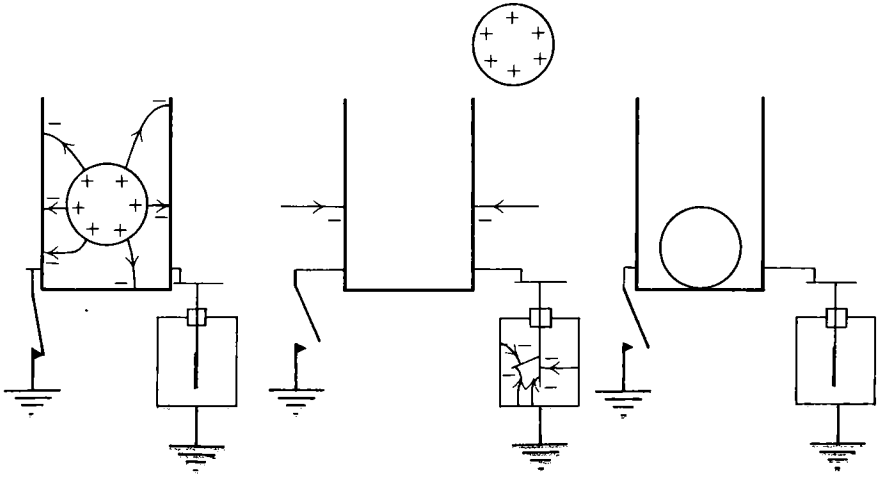


Fig. 6.6 An extension of Faraday's ice-pail experiments (a) (b) (c)

This causes the charge induced on the outside of the pail and on the electroscope to be neutralized by charge from the earth, so that the leaf goes down, and the only induced charge left is the charge $-Q$ on the inside of the ice pail. If the earth connection is now broken, and the charged body removed from the pail, the electroscope and pail are left in possession of this charge $-Q$, and the leaf goes up. Finally, if the charged body is lowered again into the ice pail and the two are this

time contacted together, the leaf falls, and all signs of any charge vanish, showing that inducing and induced charges were equal and opposite.

A summary of the results which are demonstrated by these ice-pail experiments is as follows:

The total charge which is induced by a charge $+Q$ on the conductors which surround it is equal to $-Q$. Thus, when lines of force originate from a charge $+Q$, these lines, in air or in a vacuum, are continuous lines until they terminate on a charge $-Q$.

Where lines of force cross a boundary between air or vacuum and some other insulating material, we shall see presently that there is a discontinuity, but that it remains valid to say that they start and end on equal and opposite charges.

6.5 Charging by friction

If ebonite is stroked with polythene sheet, it charges up positively, and it does so because it loses some electrons to the polythene. So we have said (section 1.1); but we have suggested no experiment to support the statement. This is an example where an ice pail may be used, enabling us to demonstrate that, if the ebonite goes positive, the polythene goes negative by an exactly equal amount, showing that it is indeed an exchange of charge which has taken place. The procedure can be as follows.

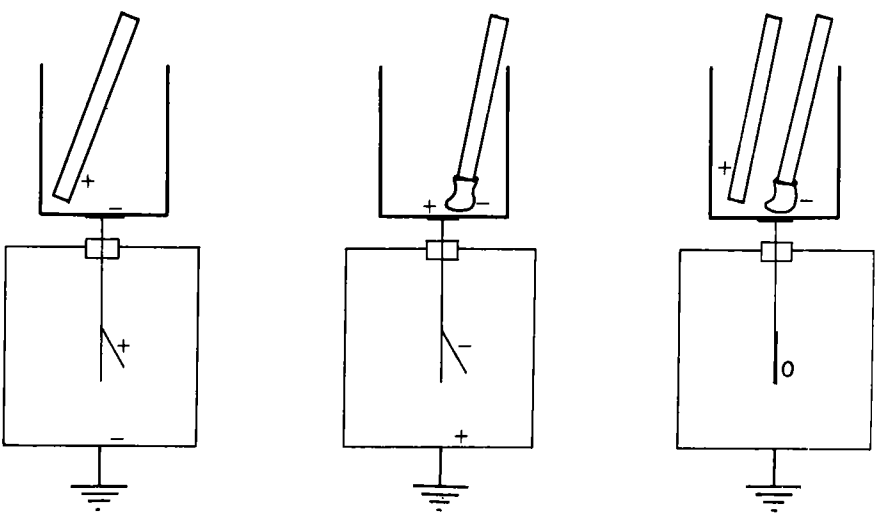


Fig. 6.7 Transferring charge by stroking one material on another

A small sheet of polythene is fixed round one end of an insulating rod, and used to charge up one end of an ebonite rod. None of the charge so produced, or transferred, can escape, because both materials are insulators. Before charging them, however, each is held for a brief moment above a low flame, which effectively removes any charge which might have been present at the outset. One end of the ebonite is now stroked across the polythene sheet, to charge the ebonite positively. If ebonite and polythene are then lowered, in turn, into the ice pail, the electroscope leaf rises; it rises by the same amount in each case. But if the two are placed in the ice pail together, the leaf stays down, only going up if one or other of them

is removed. The charges are evidently equal in size, but opposite in sign, which is what we set out to show.

6.6 Potential at a point in space – equipotential surfaces

When a body is charged up, there must be a p.d. between the body and whatever holds the charge of opposite sign from which its own charge has been separated, and an example of this is the system of two parallel plates which was shown in Fig. 6.1. A similar example is shown in Fig. 6.8, where a conducting sphere is represented as being charged up by connecting it to earth through a battery. The inclusion of the battery in this case makes the presence of the potential difference obvious, but there would in fact be a p.d. just the same, no matter what process was used to charge the sphere; this is so because the separation of unlike charges which the charging process involved was a process which took up energy.

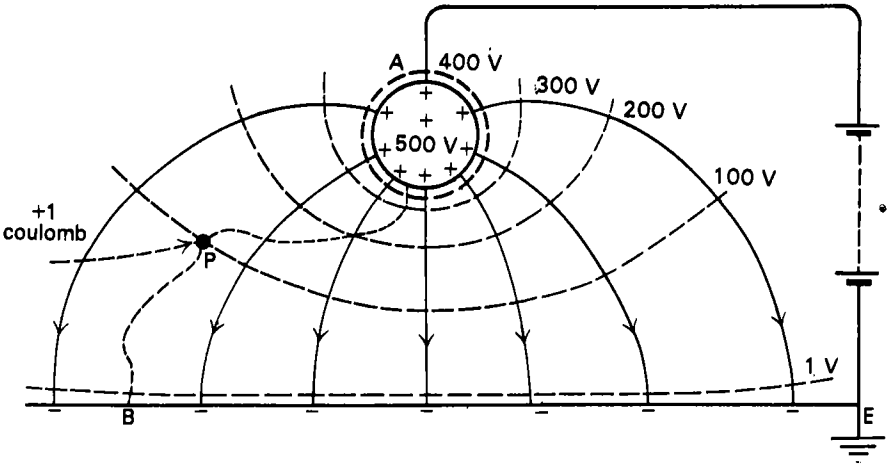


Fig. 6.8 Equipotential surfaces and electric lines of force

The potential V of the sphere is defined, consistent with the definition given in chapter 2, by the equation

$$V = \frac{W}{Q}, \tag{6.4}$$

where W is the amount of work which would have to be done against the electric force of repulsion, if charge Q were to be transferred from a point such as B on the earthed plane BE to the surface of the sphere.

Putting the equation in this simple form, we have to assume that the transfer of the charge Q would leave the potential V unaltered, which is justified for the arrangement in the diagram, where a battery is connected. It would not be true in the general case, however, and a more accurate definition would refer to transfer of a charge δQ which is too small to alter the potential appreciably, and the equation would then be

$$V = \frac{\delta W}{\delta Q}. \tag{6.5}$$

Now let us suppose that a small body carrying positive charge of 1 C (an impossibly large amount of charge, of course, in practice) is moved from B up to the sphere A. Since the potential of A is postulated in the diagram to be +500 V, the work required to carry out this operation would be 500 J, and this would be so, no matter what particular path from B to A was chosen. By the time that the charge has been carried as far as the point P, shown in the diagram, we suppose that 100 J out of these 500 have already been expended. Following the definition of electric potential, this must mean that the point P, a point in space, is at a potential of +100 V, and so is any other point on the line shown dotted on the diagram and passing through P. Bearing in mind that this diagram represents an arrangement in three dimensions, the dotted line is really part of a surface, and this is an equipotential surface, analogous to a contour on a map.

We conclude that electric potential is not restricted to such things as conductors and the terminals of batteries, and there is always a p.d. between any two points on the same line of force. This must be so, because work is required if a charge is to move against the field along a line of force.

Lines of force and equipotential surfaces are perpendicular to one another. This follows from the fact that, by definition, electric charge can be moved about on an equipotential surface without the expenditure of energy; on the other hand, to move the charge along a line of force, work must be done against or by the force.

As a special case of this, it follows that lines of force are perpendicular to the surface of a conductor which holds static charge. This is because, since there is no current, all parts of the conductor must be at the same potential, and its surface is therefore an equipotential surface.

The above principles have been observed in sketching the lines of force and equipotential surfaces in Fig. 6.8.

6.7 The electroscope as a voltmeter

The leaf of an electroscope rises whenever it is in an electric field, i.e. whenever there are lines of force from the leaf to the conducting case of the electroscope. Since there is always a p.d. between any two points on a line of force, the above statement means that the electroscope leaf will rise whenever there is a p.d. between leaf and case, and the extent of the rise may be used to measure the size of this p.d., i.e. the electroscope may be used as an electrostatic voltmeter. The same applies to any electroscope which operates on similar principles. The Braun electroscope, for example, is a very similar instrument and one application of it in this role was described in section 2.1.

The voltmeter aspect of the electroscope may be illustrated by an experiment represented in Fig. 6.9. The usual way to use the electroscope is to earth its case

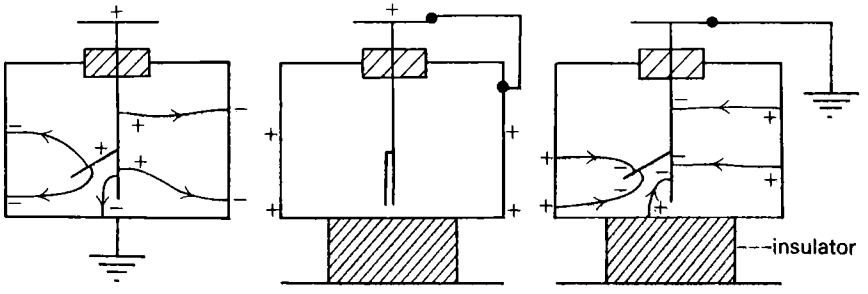


Fig. 6.9 The electroscope registers p.d. (a) Earthed case (b) Case and leaf are connected together (c) Earthed leaf

by placing it on the bench, which, in terms of electrostatics experiments, is a good enough earthed conductor. The leaf rises, therefore, whenever anything is done to cause it to have potential.

But suppose the electroscope is placed on an insulating stand, as in Fig. 6.9(*b*). If leaf and case are connected together, the whole apparatus may be charged up, yet the leaf does not rise. This is because there is no p.d. between leaf and case, and there can, therefore, be no lines of force joining leaf to case. (We will presently see that, when a single conductor is charged up, all the charge resides on the outside, so there is no charge on the leaf, in this instance, either.) Now suppose the conducting connection between leaf and case is removed, and the platform of the electroscope is touched with the finger, to earth the leaf (Fig. 6.9(*c*)). There is now a p.d. between leaf and case, and the leaf goes up, although it is earthed.

6.8 Potential is a scalar

Electric field strength E is a vector quantity. This is illustrated by its defining equation

$$E = \frac{F}{Q}. \tag{6.1}$$

Field strength E is the force per unit charge, and force is a vector.

Electric potential V, however, is a scalar quantity, and this is illustrated by the defining equation 6.4 or 6.5. It is a ratio of work or energy to charge, and energy is a scalar.

Suppose, for example, that we consider a point P which is subject to the combined influence of two neighbouring charges. To evaluate the resultant field strength E at the point, the directions of the two contributory fields, E_1 and E_2, must be taken into account as well as their magnitude, i.e. the fields must be added by vector methods. The resultant potential V at the point, on the other hand, is simply the arithmetical sum of the two contributory potentials, $V_1 + V_2$.

Example

Two charged bodies, A *and* B, *are placed near a point* P. *If body* B *is removed, the potential at* P *is* +500 V; *if body* A *is removed instead, the potential at* P *is* +700 V. *What is the potential at P in the presence of both charged bodies,* A *and* B?

The resultant potential is simply

$$\begin{aligned} V &= 500 + 700 \text{ V} \\ &= 1200 \text{ V}. \end{aligned}$$

Arguing from first principles: to bring positive charge up to P from earth would require work to be done against the repelling forces of both charges, and this work is additive.

6.9 Charging a conductor by induction – the potential aspect

The process of charging a conductor by induction can now be discussed in more detail, from the aspect of potential.

The charged sphere A shown in Fig. 6.10 is supposed to be at a potential of +500 V, and it is surrounded by equipotential surfaces as drawn in the first diagram. If a body BC is now placed in the field, there is a p.d., say 100 V, between B and C, B being the more positive. Consequently, if BC is a conductor, electrons must flow from C to B, with the result that the end B of the conductor

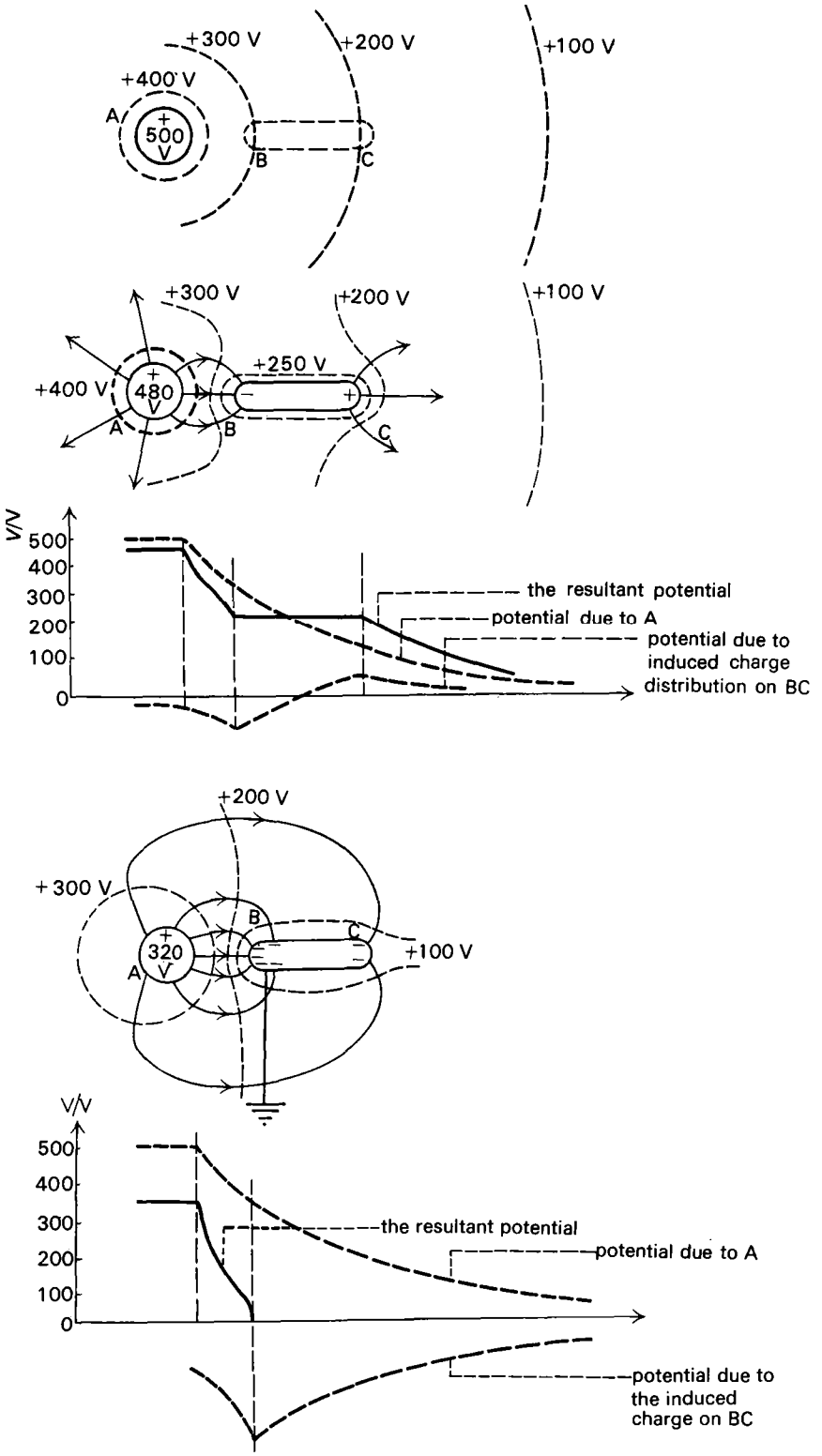


Fig. 6.10 Charging a conductor by induction (a) The equipotential surfaces surrounding a charged sphere A (b) Induced charge in an isolated conductor BC – the resultant equipotential surfaces (c) Induced charge on an earthed conductor BC – the resultant equipotential surfaces

becomes charged negatively at the expense of the end C, which becomes positive. Equilibrium is reached when all parts of BC are at the same potential, say 250 V. The induced charge distribution on BC has distorted the original equipotential surfaces, making the surface of BC into an equipotential surface.

In the third diagram, the conductor BC is supposed to be connected to earth. In this case, electrons must flow into the conductor from earth, making the whole conductor negatively charged, so that its resultant potential shall be zero throughout. The positive potential due to the charged sphere A has been neutralized at every point on BC by the effect of its own negative charge. This necessitates a greater concentration of negative charge at B than there is at the more remote end C. The earth connection is shown in the diagram as being made to the end B, but in fact it is quite immaterial whereabouts this contact is made; the effect would be the same in any case.

The effect of negative charge induced on the earthed conductor BC in the third diagram reacts back, of course, on the charged sphere A, bringing down the potential of A from the original value, +500 V, to something smaller, suggested on the diagram as being +320 V. This is a result of importance to the principle of the capacitor, and this aspect will be discussed later in the chapter.

6.10 The electrophorus

Given a charged body, we now have two methods by which it may be used to put charge on a conductor. This may be done by placing the charged body and the conductor in contact, so that the charge is shared between them (a process which works best if both bodies are conductors). The alternative is to hold the conductor near the charged body, so that it is in the influence of its electric field, and then to earth the conductor momentarily, in which case it acquires charge by induction.

There are two essential practical differences between the results of these two processes. When a conductor acquires charge by contact, it does so only at the expense of the originally charged body, which is partially discharged in consequence, and clearly both bodies finish up with charge of the same sign. By the induction method, on the other hand, the conductor acquires charge of the opposite sign, and it does so without affecting in any way the charge held by the originally charged body.

The electrophorus is a simple device for obtaining an unlimited amount of charge by induction. A slab of insulating material, such as polythene, is charged initially, by rubbing it on flannel or some similar material. A metal disc with an

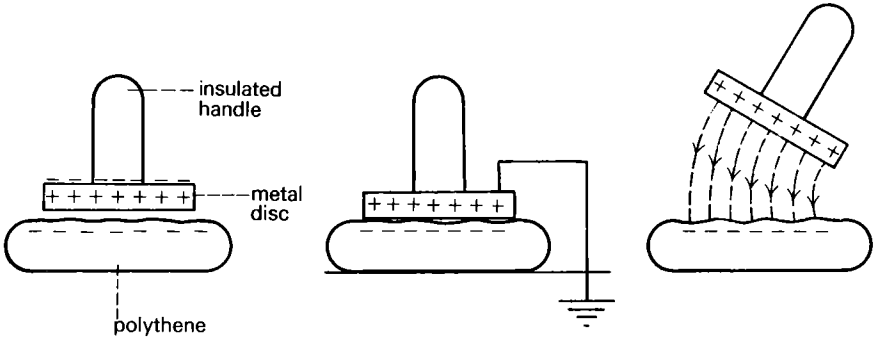


Fig. 6.11 The electrophorus – the sequence of operations (a) (b) (c)

insulated handle is then placed on the polythene slab, and momentarily earthed. In this way, charge of opposite sign is induced on the metal disc, which can then be removed, and allowed to transfer its charge elsewhere. This process does not discharge the polythene (except at the very small number of places where the disc actually makes contact with the polythene) and so the process may be repeated any number of times. The energy required to build up charge in this way comes from the operator, who has to remove a positively charged disc from the attraction of negatively charged polythene, and has to do work in the process.

The Wimshurst machine is another form of induction machine which works on the same principle as the electrophorus. Although very successful in practice, the details of its working make rather tedious reading, which would not be justified here. The basic process is that the sequence of operations is carried out automatically as a repetitive cycle, as two discs, facing each other, are rotated in opposite directions about a central axle, perpendicular to their planes.

6.11 Potential gradient

Equipotential surfaces resemble contours; where they are closely packed, i.e. where the potential gradient is large, the electric field is strong. We will prove that potential gradient and electric field strength are, in fact, identical.

Suppose that between two points, A and B, on a line of force, distance x apart, the field strength E is constant and is directed from A towards B. If we place a positive charge $+Q$ anywhere on the line AB, a force will act on it, directed away from A and towards B, given by

$$F = EQ. \tag{6.1}$$

Consequently, if the charge were first to be at B, and were then transferred to A, work W would have to be done against this force, and is given by

$$W = Fx. \tag{6.6}$$

But the p.d. between A and B, $(V_A - V_B)$, is, by definition, equal to the work done per unit charge,

i.e.
$$V_A - V_B = \frac{W}{Q}. \tag{6.4}$$

Whence, we have

$$V_A - V_B = \frac{Fx}{Q}$$

$$= Ex,$$

$$\therefore \quad E = \frac{(V_A - V_B)}{x}, \tag{6.7}$$

i.e. field strength = potential gradient.

A more general form for this equation is

$$E_x = -\frac{dV}{dx}, \tag{6.8}$$

where E_x is the component of the field strength along the direction of x. In this form, the equation is appropriate to the measurement of field strength at a point. The negative sign expresses the fact that the direction of the field is the direction of decreasing potential.

The unit of electric field strength

We now have a choice between alternative names for the same SI unit of electric field strength, and both these names are in common use, viz:

$$E = \frac{F}{Q}, \tag{6.1}$$

and the unit is the newton per coulomb;

$$E = \frac{-\mathrm{d}V}{\mathrm{d}x}, \tag{6.8}$$

and the unit is volt per metre.

It is stressed that these are alternative names for the same thing. We may say, for example, that a particular field strength E is 5000 N/C, or that it is 5000 V/m.

There is a third name, too, which might have been used, but is not, and this follows from the quantitative use of electric lines of force:

$$E = \frac{\mathrm{d}N}{\mathrm{d}A}, \tag{6.3}$$

and the unit is lines of force per square metre.

6.12 Some examples

Example 1

Two plane metal plates are placed horizontally, one above the other, separated by a distance of 4.0 *mm. A p.d. of* 220 V *is applied between the plates, and a spray of very small, charged oil drops is introduced into the field between the plates. Assuming the top plate to be positive, deduce the upward force which acts on a drop, due to the field, if the drop carries the negative charge of* 6 *electrons. The electronic charge* $e = -1.6 \times 10^{-19}$ C.

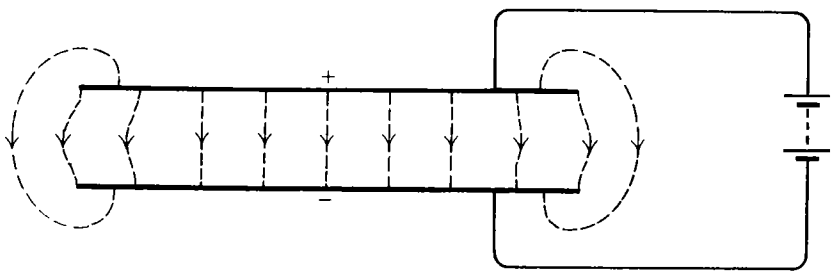


Fig. 6.12

Neglecting the edge effects which are indicated in the diagram, and taking the field between the plates as being uniform, we have

$$E = \frac{V_A - V_B}{x}$$

$$= \frac{200}{4 \times 10^{-3}} \text{ V/m}$$

$$= 5 \times 10^4 \text{ V/m}.$$

The charge on the oil drop $= Q$,

where
$$Q = -6 \times 1.6 \times 10^{-19}\ \text{C}$$
$$= -9.6 \times 10^{-19}\ \text{C}.$$

The upward force on the drop is therefore
$$F = EQ$$
$$= 5 \times 10^4 \times 9.6 \times 10^{-19}\ \text{N}$$
$$= 4.8 \times 10^{-14}\ \text{N}.$$

Example 2

A, B *in Fig.* 6.13 *are two small positively charged spheres,* A *having a larger charge than* B. *The potential due to these charges varies along the line* AB *in the manner shown in the diagram.*

(*a*) *Mark in on the diagram the position of any neutral point, i.e. of any place where the resultant field strength is zero.*

(*b*) *Show graphically how the field strength E varies along the line through* A *and* B.

(*c*) *Sketch the lines of force and equipotential surfaces due to the two charges.*

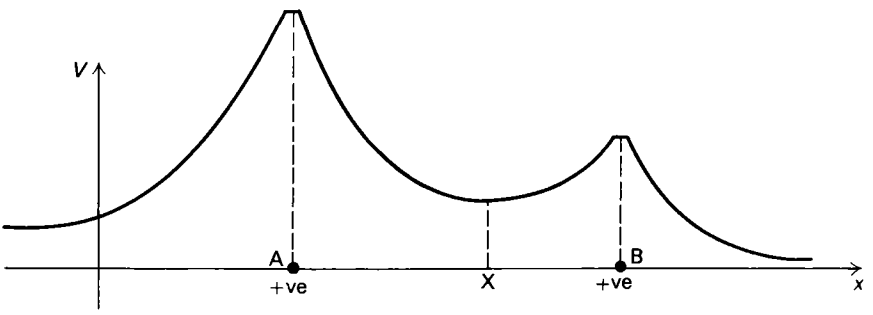


Fig. 6.13

(*a*) Where the field strength is zero, this means that the potential gradient is zero, which is so at the point X in the diagram, and X marks the position of the only neutral point.

(*b*) To show how the field strength E varies along the line of AB, we merely have to trace how the gradient of the potential graph varies, since this gradient, reversed in sign, is identical with the field strength. The result of doing this is shown in Fig. 6.14.

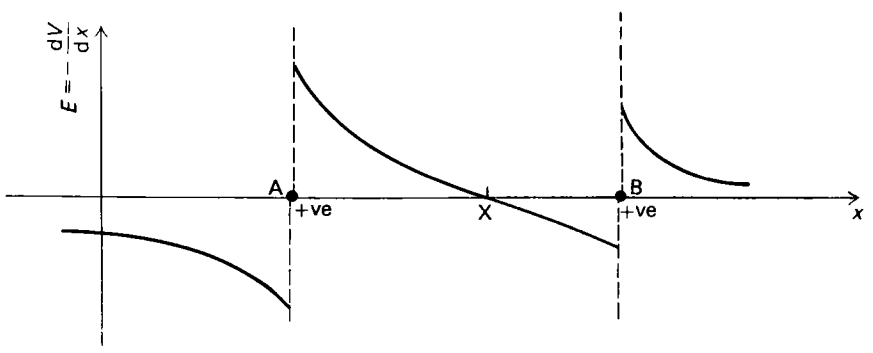


Fig. 6.14

It will be noted that it is possible for the value of the field strength to change abruptly at a point; in particular, since it is a vector, there is a reversal in sign on crossing over from one side to the other of an electric charge. A similar abrupt change in value is not possible in the case of potential, because this would involve having an infinite potential gradient, i.e. infinite field strength, at such a point.

(*c*) In plotting the lines of force and equipotential surfaces shown below (Fig. 6.15), the lines of force were drawn in first, using the position of the neutral point X as a guide. The form of the equipotential surfaces could then be estimated, bearing in mind that lines of force and equipotential surfaces are perpendicular to one another.

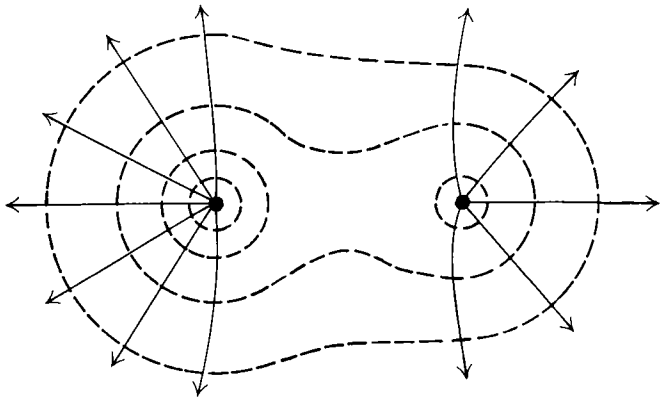


Fig. 6.15 Lines of force and equipotential surfaces

6.13 The electrostatic circumstances of a charged conductor

We now have to investigate the conditions in which a conducting body exists when it holds static electric charge, i.e. the conditions which apply to such things as its potential, charge distribution, field strength, etc. In each of these cases the ruling factor can usually be easily deduced; it can also be demonstrated.

(a) Potential

The potential must be the same at all points of the charged conductor, whether inside it or on its surface. This is because current will flow until the charge distribution is such that this condition is realized. The statement applies only to a conductor, of course (although it need not be a good one), and it applies only when all current has ceased.

This may be demonstrated by using an electroscope as a voltmeter (Fig. 6.16). The case of the electroscope is earthed, and a wander lead is used to connect the leaf directly to various points on the charged conductor. Whether this contact is made outside the conductor or inside it, the result is the same, and the rise of the leaf is constant, showing that the potential is constant.

(b) Field strength

The lines of force just outside a charged conductor are perpendicular to its surface. This is because the surface of the conductor is an equipotential surface.

Inside a hollow conductor, the field is zero, i.e. there are no lines of force at all – provided, that is, that there is not a second charged body inside it which makes no conducting connection with it.

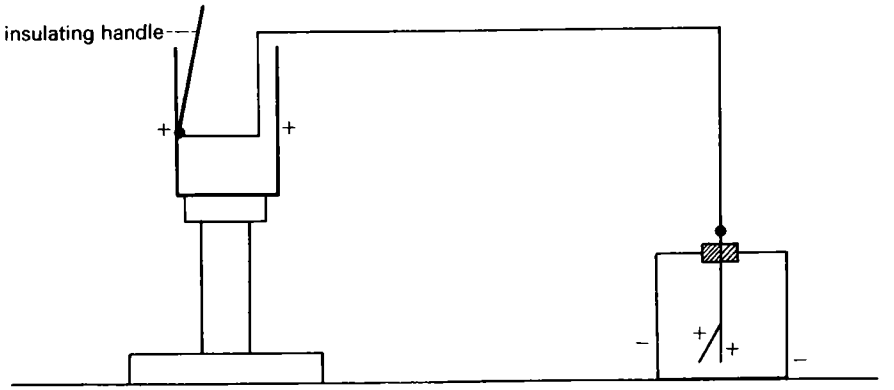


Fig. 6.16 Apparatus to test the potential of a charged conductor

We can argue that, if there were any such line of force, and if there were no charged body inside on which the line could terminate, it would necessarily link two points on the same conductor. But this would necessitate the existence of a p.d. between the two points, and we have just demonstrated that there is no such p.d. In other words, an electric field inside the hollow conductor would be a potential gradient, and this must be zero.

The absence of field inside a conducting shell, due to charges on the shell or outside it, is the basis of electrostatic screening.

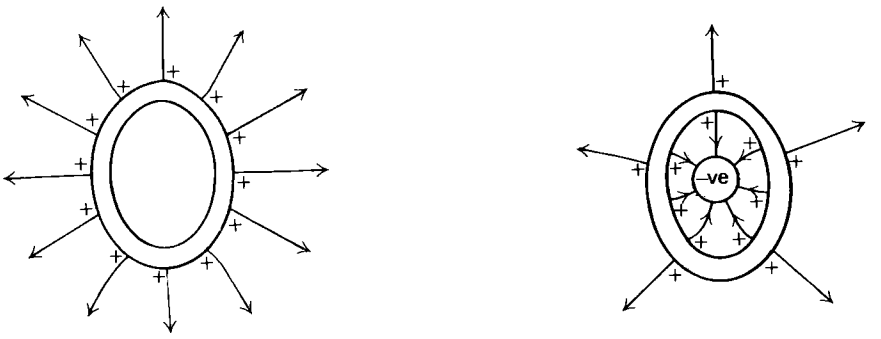


Fig. 6.17 There is no field inside a charged conductor unless there is a separate charged body inside (a) (b)

(c) Charge

The charge on the conductor is all on its outside surface, and there is no charge inside.

This, again, assumes that there is not a separate charged body inside, as there is in Fig. 6.17(*b*). If there were any such charge, lines of force must originate from it, i.e. there would be a field inside the conductor, which there is not.

The absence of charge inside a charged conductor is demonstrated as part of Faraday's ice-pail experiments (Fig. 6.5); if a charged conductor makes contact with the inside of the ice pail, it loses all its charge to the ice pail, i.e. all the charge goes to the outside.

There are various well known alternative demonstrations, one of which uses Biot's hemispheres.

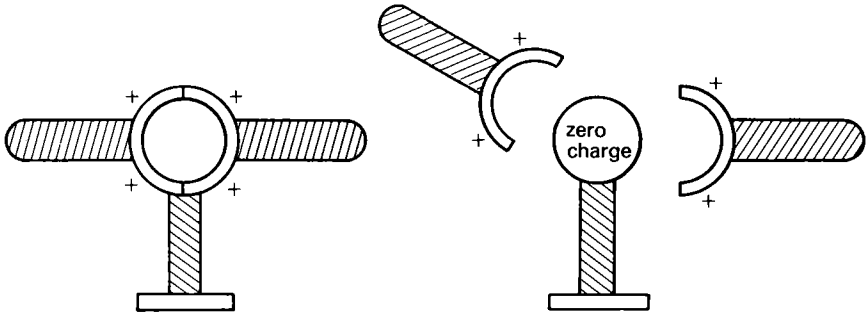


Fig. 6.18 Biot's hemispheres

These are two hollow metal hemispheres, with insulated handles, which may be placed to surround a charged and insulated metal sphere so that they make contact with it. If the whole arrangement is first charged up, and the two hemispheres are then removed simultaneously, subsequent test with an electroscope reveals that all the charge was removed by the surrounding hemispheres, leaving none at all to charge the sphere.

(d) The charge density

The charge has to distribute itself over the surface of the conductor in such a way that the potential shall be the same everywhere on the conductor, and this results in the charge density, i.e. the charge per unit area

$$\sigma = \frac{\delta Q}{\delta A}$$

being greatest where the curvature is sharpest. This may be demonstrated by using a proof plane, which is a small metal disc with an insulated handle.

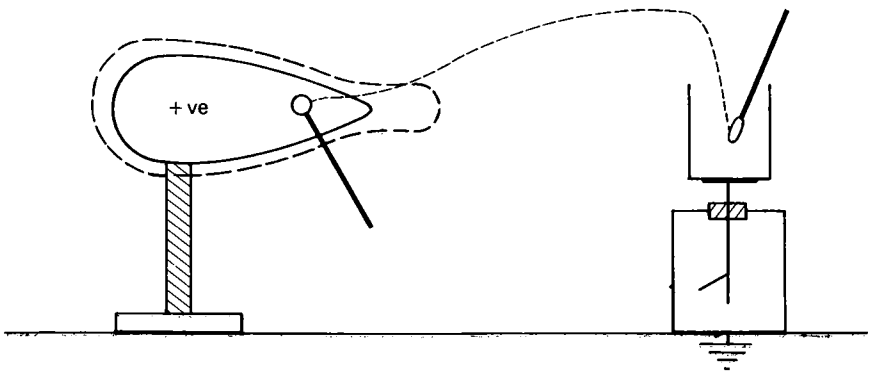


Fig. 6.19 Charge density. Exploring charge distribution with a proof plane. The distance of the surrounding dotted line from the charged conductor is supposed to represent the charge density, which is greatest at the pointed end

The proof plane is placed flat on the surface of the conductor at different parts of the conductor in succession, so that when it is lifted off it shall take charge away with it in proportion to the charge density where it was placed. This charge is then tested by holding the proof plane inside an ice pail connected to an electroscope, without allowing the plane to contact the ice pail. In this way its charge is tested

without losing it, and so the conductor remains charged up throughout the proceedings.

(e) The field strength at the surface of a charged conductor

The field strength at the surface of the conductor is greatest where the curvature is sharpest. This must be so because the concentration of lines of force is proportional to the charge density. In particular, the lines are very concentrated indeed where a conductor comes to a point (Fig. 6.20(*a*)).

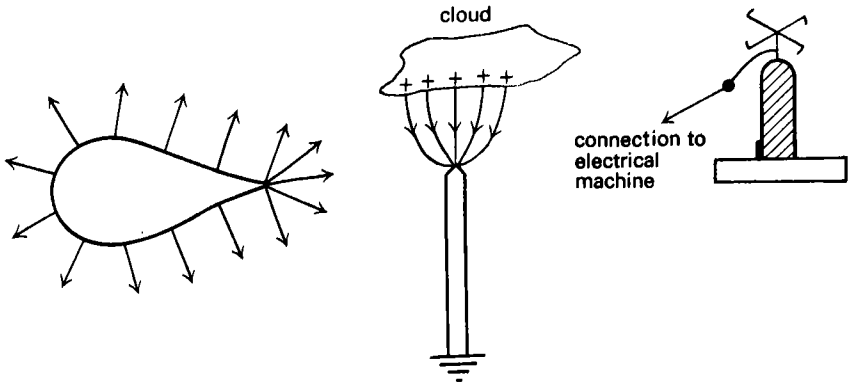


Fig. 6.20 Examples of field concentrated at a point (a) (b) A lightning conductor (c) An electric windmill

This powerful localized field near a pointed conductor has very considerable practical importance, because it easily reaches such a high value that the air in the vicinity of the pointed conductor becomes ionized, and carries charge away, discharging the conductor. When this occurs, it is called corona, or brush discharge, because the air near the point becomes luminous, with some resemblance, if viewed in the dark, to a luminous brush.

The discharging action is accompanied by the movement of ions, and molecules from the air, attached to ions, causing an 'electric wind', which can be demonstrated, on the turbine principle, by connecting 'an electric windmill' to an electrostatic machine.

Point discharge is the principle of the lightning conductor, which serves to discharge a thundercloud before the potential has reached a high enough value to discharge the cloud to other objects such as buildings. It is also the principle of charging and discharging combs, which can exchange charge with a body moving past them without actual contact being made.

6.14 The Van de Graaff generator

The Van de Graaff is an electrostatic generator which is very simply based on some of the principles discussed in the last section. The electrophorus and the Wimshurst machine produce high potential simply because the energy required to separate unlike charges is large. The total amount of charge involved, however, is too small for these two instruments to be much more than toys, although large Wimshursts were at one time used to energize early types of X-ray tube. The Van de Graaff is in a different class; a large generator can produce a potential of several million volts, and it was one of the important forerunners of the modern high-energy particle accelerator.

6.14 The Van de Graaff generator

The principle involved is the simple one that we first encountered in Faraday's ice-pail experiments, that charge always goes to the outside of a conductor; consequently, if a charged conductor B is contacted to the inside of a hollow conductor A, it loses all its charge to A no matter how high the potential may be to which A has already been raised. This principle means that, in theory, the potential of A may be raised to an unlimited extent, merely by continually feeding charge to its inside surface. The practical limit to this is set by the rapidity with which A is losing its charge again. A constant limiting potential is reached when the conductor A is receiving and losing charge at the same rate. It is therefore an advantage to feed charge into it as rapidly as possible, and to reduce the loss of charge as much as possible.

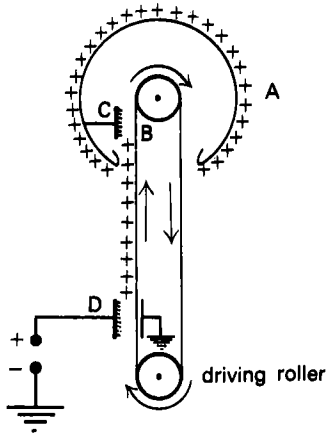


Fig. 6.21 (a) The Van de Graaff generator

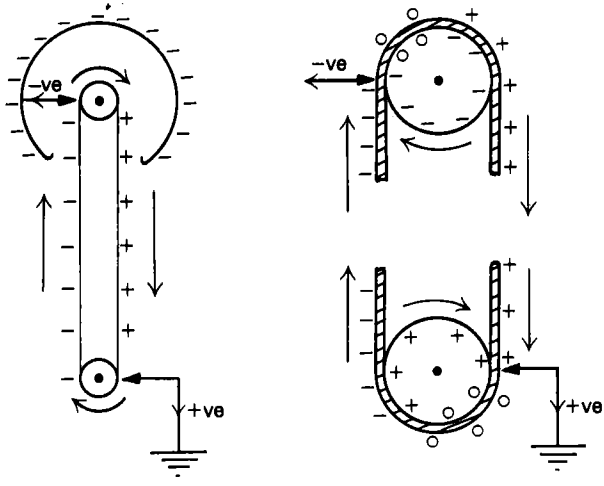


Fig. 62.1 (b) A self-exciting Van de Graaf generator. The discharging comb at the top collects negative charge off both belt and roller, while the comb at the bottom removes the positive charge from both.

The basic arrangement is a rotating belt of insulating material, such as rubber, which receives charge at the bottom and carries it up to a collecting comb C inside a large spherical conductor A, where it loses its charge to the sphere by the process of brush discharge at the comb. With large generators, the initial charge is sprayed continuously on to the bottom of the belt from a conventional high potential transformer and rectifier, so that the charge may be carried up to the sphere at a considerable rate; this initial charge must be at high potential, so that brush discharge shall take place across to the belt from the discharging comb D. The small type of generator used in a school laboratory, however, obtains its charge by friction. The lower roller, of metal or perspex, charges the rubber belt negatively by friction, while the upper roller, of polythene, charges it positively. Consequently, the belt carries negative charge up to the sphere and brings positive charge away from it.

If the potential is to reach a really high value, it is essential, as we have already pointed out, to keep the discharge rate down; as soon as the sphere is losing charge as fast as it is receiving it, the potential will go no higher. The loss is due to three causes. The first of these in unavoidable, being the current taken by the load which is put across the generator: the potential may be used, for example, to bring protons up to high speed in an accelerator tube, in which case the charge carried by the protons represents a discharge current. The second cause of loss is leakage of charge through the insulating supports of the metal sphere and through the belt itself; this is a matter partly of design and partly of cleanliness. The third cause of loss is leakage of charge into the surrounding air; if the surface were, in fact, a smooth isolated sphere, the electric field strength at its surface would be given by the equation (which will be derived later, in section 13.7)

$$E = \frac{V}{r}, \tag{6.9}$$

where r is the radius of the sphere. The field would not be stronger than this anywhere else – this is illustrated in Fig. 6.22 – and so this equation may be taken to indicate the maximum potential to which the sphere can be charged. If the field strength at its surface were to break down the insulation of the surrounding air by ionizing it, brush discharge would result. The equation shows that the limiting value for the potential is directly proportional to the radius of the sphere. If there are projections of any kind, however, making flaws in the smooth surface of the sphere, the lines of force are concentrated at these projections, causing brush discharge to occur at a lower potential. In addition to this, if the generator is housed in a laboratory, induced charge on the walls would disturb the symmetry of the lines of force, and this, too, would increase the field strength locally.

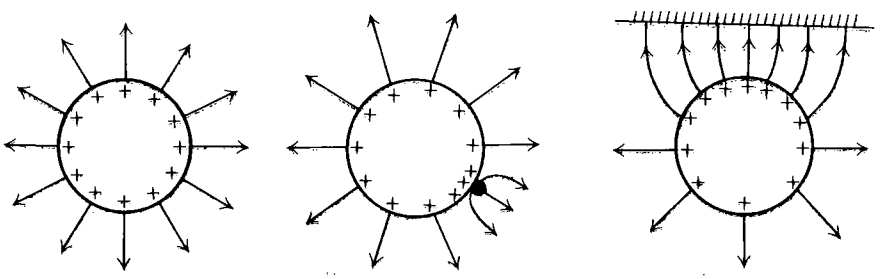


Fig. 6.22 The charged sphere – local field effects caused by lack of symmetry (a) The symmetrical field of an isolated conducting sphere. The field strength is greatest at its surface (b) The localized field at an irregularity on the smooth surface (c) The effect of a neighbouring wall

For this reason, the first powerful Van de Graaff generators were housed in very large sheds. The present method, however, is to encase the generator in a large container, filled with gas at a pressure of about 20 atmospheres. This is because the limiting field strength which a gas can withstand without ionizing is in direct proportion to the pressure, so that this device increases the insulating properties of the gas about twenty times. A suitable gas to use in the container is 'Freon 12' (CF_2Cl_2), chosen because its dielectric strength is more than twice that of air, which means that it does not conduct so readily.

6.15 A capacitor

When a conductor has been designed with the object of storing electric charge, it is called a capacitor, and the ratio of its charge Q to its potential V is termed its capacitance C,

i.e.
$$C = \frac{Q}{V}. \tag{6.10}$$

More precisely, V is the p.d. between the conductor which holds the charge Q and whatever conductor has the equal and opposite charge $-Q$ from which the first charge has been separated. A practical capacitor consists of two parallel metal sheets which store these equal and opposite charges, and the lines of force of the charges cross the space between the sheets.

The SI unit of capacitance, coulomb per volt, is called the farad, but this is too large for most practical purposes, and the units more commonly quoted in practice are the microfarad (μF) and the picofarad (pF) where

$$1\ \mu\text{F} = 10^{-6}\ \text{F},$$
$$1\ \text{pF} = 10^{-12}\ \text{F}.$$

Because capacitance is never very big, the capacitor is not an important component where direct current is involved; its application is largely concerned with alternating current, particularly at high frequency, where it is repeatedly being charged up, first in one direction, then in the other. In this field it is a very important device.

In nearly all cases, the charge Q stored in a capacitor is directly proportional to the p.d. between its plates, so that its capacitance C is a constant.

6.16 The factors which control capacitance

If the amount of charge stored in a parallel plate capacitor is kept constant by isolating it after charging it, the effect of making adjustments which alter its capacitance can be studied by observing the changes in the p.d. between the plates. This is because we have

$$C = \frac{Q}{V}.$$

Hence, if Q is constant, any decrease in the p.d. V indicates that the capacitance has been increased. This principle is the basis of the sequence of experiments illustrated in Fig. 6.23, where a charged metal plate is connected to the leaf of an electroscope, the case of which is connected to earth. Thus the extent to which the leaf rises indicates the potential of the metal plate.

The gold leaf is observed to fall to a small deflection if the area of the charged metal is increased, or if a parallel, earthed metal plate is brought up close to it; thus both these operations has increased the capacitance.

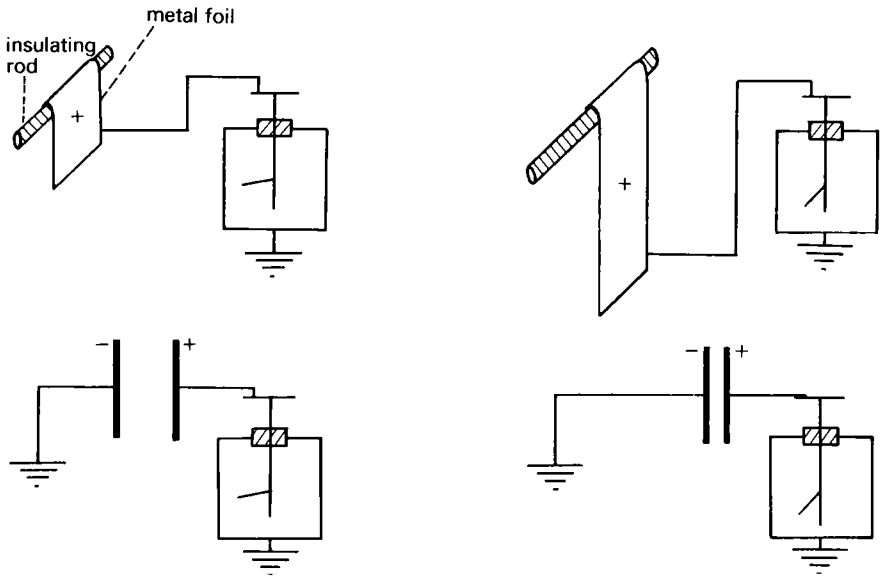


Fig. 6.23 Experiments to demonstrate factors which affect capacitance (a) Area. As the sheet is unfurled, the potential falls (b) Separation of the metal plates. When the plates are brought close together, the potential falls

These two results are summarized in the equation, which will be derived in a later chapter (section 13.9), giving the capacitance of two parallel plates separated by a vacuum (or for all practical purposes, by air), viz

$$C = \epsilon_0 \frac{A}{d}. \tag{6.11}$$

Here, A is the area of each plate, d is the distance which separates the two, and the constant factor, ϵ_0, is termed 'the permittivity of a vacuum'. Its SI value has been found, by practical measurement, to be

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ F/m, approximately.} \tag{6.12}$$

6.17 A dielectric

If a sheet of insulating material, such as glass, is inserted between the two parallel plates used in the last experiment, the gold leaf falls still further, showing that the capacitance has been still further increased (Fig. 6.24). Used in this way, the insulator is termed a dielectric.

In most practical capacitors the metal plates are separated by a solid dielectric, which increases the capacitance in this way by its presence; but its more effective function is to serve as a means of separating the plates (or thin metal sheets) at a very small distance apart, and so to increase the capacitance on this account. When the space between the two metal sheets is completely filled with dielectric, which is usually the case, we write

$$C = \epsilon_r C_0, \tag{6.13}$$

where C_0 is the capacitance of the arrangement in the absence of the dielectric, and C is the value when the dielectric has been inserted.

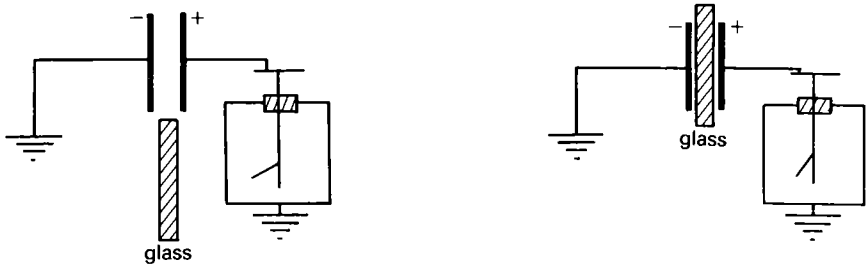


Fig. 6.24 The effect of a glass dielectric in increasing capacitance

The factor ϵ_r is termed the relative permittivity of the material, and the reason for this can be seen by reference to equations 6.11 and 6.13. Combining these equations, the capacitance of a dielectric spaced capacitor becomes

$$C = \epsilon_r \epsilon_0 \frac{A}{d}, \tag{6.14}$$

which is sometimes simplified in form by writing

$$C = \epsilon \frac{A}{d}. \tag{6.15}$$

Whereas ϵ_0 is the permittivity of a vacuum, the factor ϵ may now be termed the permittivity of the dielectric, whence we have that the relative permittivity of the dielectric is

$$\epsilon_r = \frac{\epsilon}{\epsilon_0}. \tag{6.16}$$

6.18 Choosing a dielectric

The approximate values of the relative permittivity of a number of typical dielectrics is listed in Table 1. In choosing the dielectric to use in a capacitor, however, its relative permittivity is but one of several factors which must be taken into account. Some relevant questions to be answered are listed below; their relative importance depends on the application for which the particular capacitor is being designed.

(*a*) Can the dielectric be provided in the form of a very thin insulating layer, thus producing the high capacitance which results from small separation of the plates?

(*b*) Is the dielectric a very good insulator? If it is not, the capacitor may be regarded as having capacitance with a high leakage resistance in parallel with it. This is acceptable for some applications.

(*c*) Will the dielectric withstand the effect of a large potential gradient? If it will not, there will be a comparatively low value limiting p.d., above which the capacitor cannot be used.

(*d*) How far may the temperature be allowed to rise?

(*e*) Is the dielectric flexible? Capacitors using waxed paper as the dielectric, for example, or using a plastic, can be made small in physical size by rolling them into cylindrical form. This is not possible in the case of mica.

Table 1. *Approximate values of the relative permittivity of some typical dielectrics, measured using a direct potential difference at* 20 °*C.*

Dielectric	Relative permittivity
Mica	7
Glass	5–8
*Water	80
Paraffin wax	2–2·5
Terylene	2·5
Special ceramics (containing titanates)	100–1 000

* Water cannot, of course, be used as the dielectric of a capacitor, because it is not a good enough insulator.

6.19 Some practical forms of capacitor

(*a*) *A variable capacitor* consists of a number of metal vanes on a central shaft, which can be rotated so that they lie partially within the spaces which separate a series of fixed metal vanes. The dielectric is air, and a typical value for the capacitance is 3×10^{-4} μF.

(*b*) *A waxed paper capacitor.* This is cylindrical in form. Two long sheets of thin metal foil, and two long sheets of paper impregnated with paraffin wax, are placed alternately and rolled into a cylinder.

(*c*) *A plastic dielectric capacitor.* Various polymers are used for the dielectric in such capacitors, including polystyrene, terylene, and polythene. These have the advantage of very high insulation resistance. They are made in much the same cylindrical form as the waxed paper capacitor.

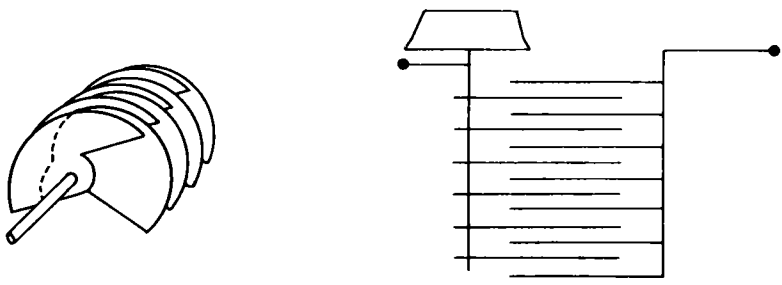


Fig. 6.25 A variable air-spaced capacitor (a) (b)

(*d*) *A mica capacitor.* This type must be flat, because the mica sheets cannot be rolled up. The mica layers which form the dielectric are coated with a film of silver to make the conducting plates of the capacitor, and stacked in layers, on the same principle as that of the variable capacitor shown in Fig. 6.25(*b*). Mica has the advantage of being an excellent insulator, and it can withstand the effect of a large potential gradient, but it cannot be used to provide a capacitance much in excess of a few microfarads without being physically large.

(*e*) *A ceramic capacitor.* The characteristics of this type of capacitor can be varied considerably, because various types of dielectric can be chosen. One class of dielectric, based on titania, TiO_2, has exceptionally high relative permittivity, of the order of 10^3, and can be used to make capacitors of very small physical size.

(*f*) *An electrolytic capacitor.* An exceptionally thin dielectric can be made by

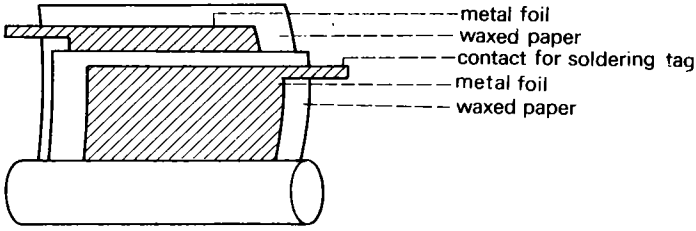


Fig. 6.26 A waxed paper capacitor

depositing it electrolytically as a layer on a metal foil, which is used as the anode. In this way, high capacitance can be produced without the finished component being unduly large. A capacitance of about $10^5\ \mu F$, for example, might be the value for a cylindrical capacitor 10 cm high and 5 cm in diameter.

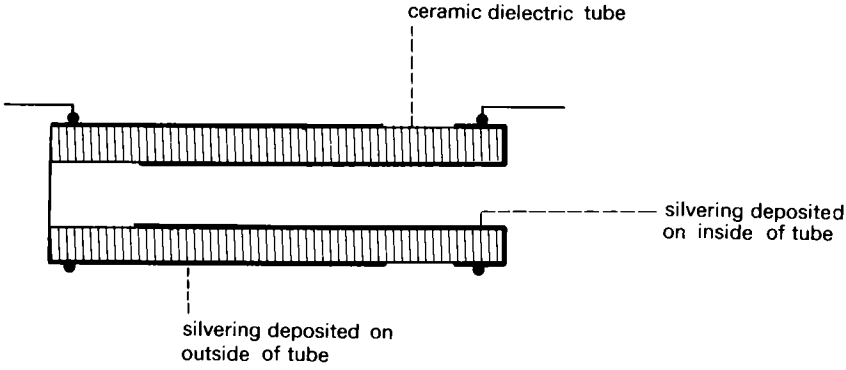


Fig. 6.27 A ceramic dielectric capacitor

The conventional type has an anode of aluminium foil, on which the electrolysis deposits a thin, insulating layer of aluminium oxide to act as the dielectric. The other conducting plate of the finished capacitor consists of the thin paper which was impregnated with the electrolyte and put in intimate contact with the metal foil, and a second foil is placed on the other side of this paper to make the second connection.

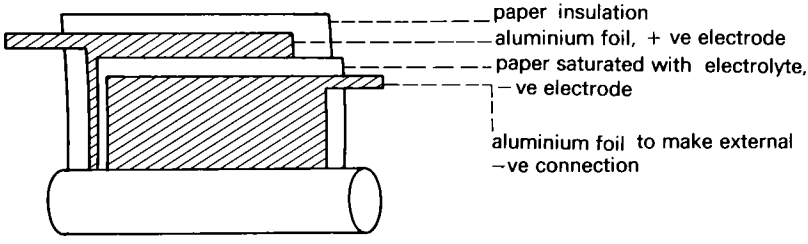


Fig. 6.28 An electrolytic capacitor

More recent types of electrolytic capacitor, which can be made even smaller in size, have tantalum anodes, the dielectric being a deposit of tantalum pentoxide.

These types of capacitor must be used in a circuit with the metal anode as the positive plate, otherwise reversal of the electrolysis is liable to destroy them. The leakage current tends to be comparatively high, but there is some advantage in this,

since it serves to maintain the dielectric layer by slowly continuing the process of electrolysis.

6.20 The theory of the dielectric

The construction of all these capacitors is based on the principle that two conducting sheets placed close together have greater capacitance than a single sheet, and that dielectric put between the two increases the capacitance still further. We have yet to consider why this is so.

The effect of placing an earthed sheet close to a charged one (Fig. 6.23(*b*)) may be regarded as an aspect of electrostatic induction. If the first sheet holds, say, a positive charge, then the induced charge on the earthed sheet is negative, and this lowers the potential of the positive sheet by its proximity, and so increases the capacitance (section 6.9).

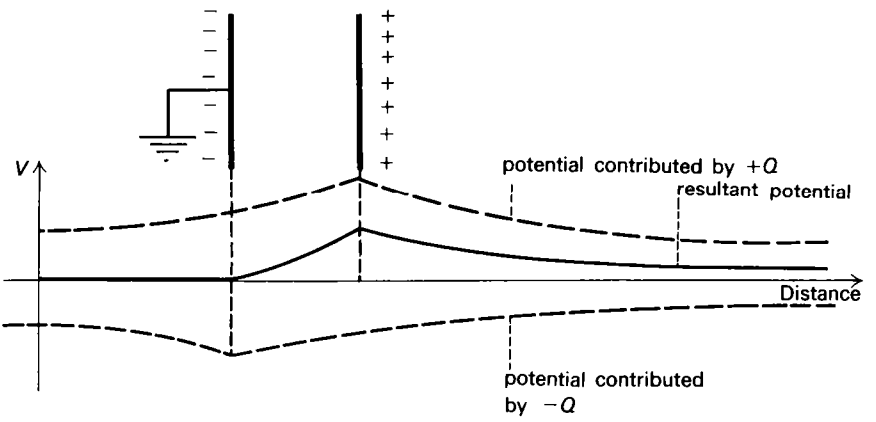


Fig. 6.29 The effect of an earthed conducting plate in increasing capacitance

The effect of a dielectric inserted between the sheets (Fig. 6.24) in increasing capacitance may be accounted for in a very similar way. The dielectric becomes polarized by the field, i.e. the two surfaces acquire equal and opposite charges, not by conduction, but by rearrangement of the charge distribution in the molecules or atoms. In this way the dielectric acquires a polarity which is of opposite sign to that of the two charged sheets, with the result that its presence lowers the p.d. between the sheets and the capacitance is correspondingly increased.

There are two main classes of dielectric, which differ in the manner of their polarizing. The molecule of one class is an electric dipole, having positive charge

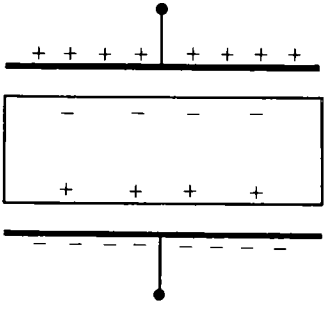


Fig. 6.30 A polarized dielectric

at one end of the molecule and negative charge at the other; in a strong electric field the molecules align themselves to lie along the field. The molecule of the other class is not naturally a polar molecule; when subjected to the electric field, however, the mean centre of the negative charge representing the electrons becomes displaced relative to the mean centre of positive charge representing the atomic nucleii, with the result that the molecule becomes an orientated dipole when in the field.

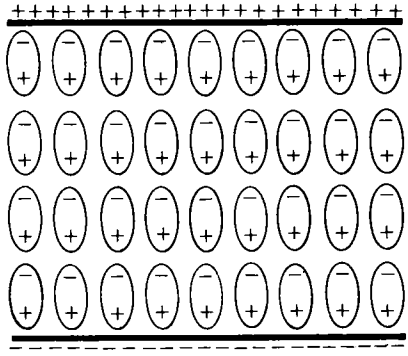


Fig. 6.31 Molecules in a polarized dielectric

6.21 A discontinuity in lines of force

If the plates of a charged capacitor are isolated, so that the charge on them cannot change, the effect of inserting a dielectric into the space which separates them is to lower the p.d. between them, and this can be interpreted as meaning that the field strength in the dielectric is less than it was in the original air space. This conclusion follows from the equation for this field strength

$$E = \frac{V_1 - V_2}{x}. \tag{6.7}$$

Because the concentration of lines of force is defined to represent field strength, the conclusion is that there are fewer lines of force inside the dielectric than there are in the space that remains between the dielectric and the plates (Fig. 6.32). If we consider this in terms of the polarized molecules, the field strength in the dielectric no doubt has local variations, but, taking the macroscopic view, we can argue in terms of the polarization charges at the surfaces of the dielectric. It would then seem reasonable to suggest that lines of force do extend continuously from positive charge to equal negative charge, but that the charge on which they terminate need not be charge on a conductor; it may equally well be the charge on the surface of the polarized dielectric. This gives a picture of the lines of force as shown in Fig. 6.32.

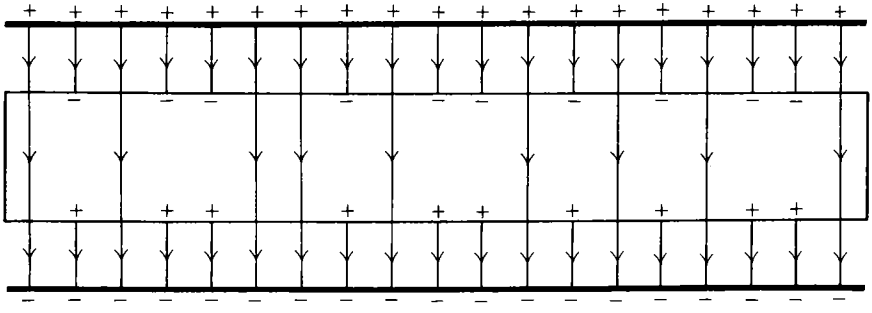


Fig. 6.32 The effect of dielectric on lines of force

QUESTIONS

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ F/m}$$

6.1 Account for the following effects:

(*a*) With the object of demonstrating that like charges repel one another, a charged polythene rod is brought up close to a pith pellet suspended from a thread. As expected, the pith is attracted towards the rod until it makes contact with it, and it is then repelled. Unexpectedly, however, the pellet, instead of continuing to be repelled, is soon attracted again to the charged rod, and then repelled a second time, and the whole cycle of operations repeats itself a number of times.

(*b*) When a charged polythene rod is held close to a pith pellet suspended from a thread, the pith is initially attracted to the polythene, although the pith holds no charge. The pellet then has a tendency to cling to the polythene, instead of being repelled by it. When a charged *brass* rod is used, however, instead of the polythene one, the pellet is repelled immediately after making contact.

(*c*) If a charged polythene rod is brought close to the metal plate at the top of a gold leaf electroscope, the leaf rises, and falls again when the rod is removed. It makes no difference if the rod actually touches the electroscope plate at one stage. If the procedure is repeated, however, using a charged metal rod instead of a polythene one, the behaviour of the electroscope is altered if the metal is allowed to contact the plate.

6.2 Given two small charged bodies, explain how you would use an electroscope in order to ascertain which of the bodies carries the larger amount of charge, and whether both charges have the same sign.

Does the success of your experiment depend upon whether or not the bodies are conductors? Explain.

6.3 In what conditions can an induced electric charge be equal to the inducing charge in magnitude? Describe an experiment to demonstrate the truth of your statement.

You are provided with two insulated metal cans, A and B, the first can A being small enough to be placed, when required, inside can B. Given also a small metal sphere suspended from insulating thread and carrying charge $+Q$, explain how you would proceed in order to place on the can B (*a*) charge $-4Q$, (*b*) charge $+2Q$.

6.4 The diagram represents a charged conducting sphere A on an insulating

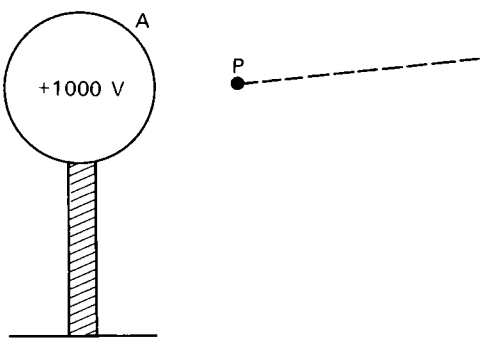


Fig. e 33

stand, near which a small metal sphere P is held by means of a long insulating rod to which it is attached.

(*a*) The sphere A is charged to a potential of + 1 000 V. The small sphere P, owing to the proximity of the charge on A, is at a potential of 500 V. Explain, in terms of the definition of potential, what each of these statements means.

(*b*) If a moving coil voltmeter is now connected between P and earth, the voltmeter does not register 500 V; on the contrary, its reading is zero. If, however, P is connected to the leaf of a gold leaf electroscope, the case of which is connected to earth, the leaf rises, showing correctly that P has potential. Account for the behaviour of each instrument, and illustrate your answer by sketching the equipotential surfaces which surround P both before and after each instrument is connected.

6.5 There is a vertical electric field at the earth's surface, which may be as large as 300 V/m. When a man is standing up, therefore, there might be a p.d. between his head and his feet of about 550 V, which should give him an unpleasant shock. Explain the fallacy in this reasoning.

6.6

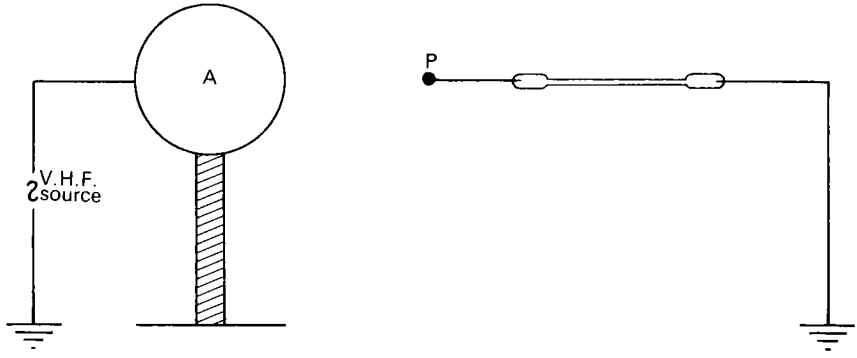


Fig. e 34

The insulated metal sphere A is charged up by connecting it to a source of very high frequency alternating potential, the peak value of which is about 1 000 V. If a gas discharge tube is now held in the vicinity, with one electrode at P, and the other connected to earth, the gas appears to glow continuously. Explain this.

(You should assume that no current flows through the air from A to P.)

6.7 In explaining the significance of p.d. and potential in chapter 2, the analogy was taken of water raised to a height h. P.d. was treated as being analogous to difference in height h; i.e. energy per unit charge was compared with energy per unit weight.

Discuss the suggestion that it would have made a closer analogy to compare p.d. V with the quantity gh, where g is the gravitational acceleration.

6.8 Illustrate, by a suitably labelled sketch in each case, how it is possible to set up electrostatic systems in which (*a*) a conductor at earth potential carries a net positive charge, (*b*) a conductor at earth potential has regions of both negative and positive charge, (*c*) a conductor with no net charge is at a positive potential with respect to earth.

How would you verify by experiment that both negative and positive charges are present in case (*b*)? (O & C)

6.9

A gold leaf electroscope is so constructed that for a few degrees deflection the gold leaf makes connection with the earthed case by the contact P. Describe and explain the behaviour of the leaf when (*a*) a positively charged,

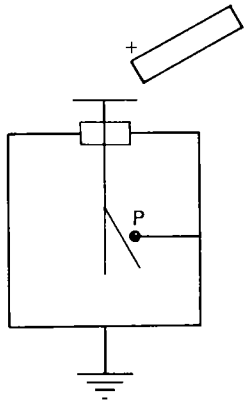


Fig. e 35

insulated body is brought towards the plate at the top of the electroscope until the leaf touches the contact P; (*b*) the charged body is then moved slowly still closer towards the plate; (*c*) the charged body is fixed in position close to the plate and the air between them is feebly ionized to render it slightly conducting. (O & C)

6.10 The section B of the composite rod shown in the diagram is made of insulating material, which holds positive charge at one end, and negative charge at the other. Sections A and C are made of metal, and are uncharged.

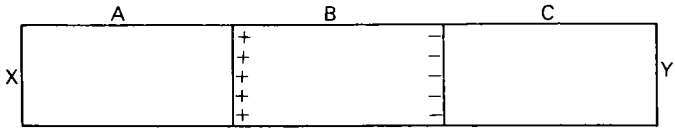


Fig. e 36

Sketch a graph showing how the potential of the rod varies along its length from X to Y, and discuss the charge distribution in each of the sections A and C.

Supposing that the sections A and C are now connected to each other by a conducting wire from X to Y, sketch the new potential variation, and discuss the charge distribution in A and C now.

6.11 A positively charged glass sphere is placed on the axis of an insulated metal rod some distance from one of its ends. Account for the resulting charge distribution over the metal rod. How do you reconcile the charge distribution on the metal rod with the fact that the only free charges in a metal are electrons?

Sketch freehand graphs showing the form of the variation of (*a*) the electric potential, and (*b*) the electric field strength along the axis of the metal rod from the surface of the glass sphere through the rod to a distant point.

How and why is the electric field between the glass sphere and the metal rod changed if the metal rod is earthed? (O & C)

6.12 The three bodies A, B, C are all conductors: A is charged to a potential of +1000 V; B is uncharged and isolated; C is earthed. Sketch the lines of force and equipotential surfaces between A and B, and between B and C. Sketch a graph showing how (*a*) potential, (*b*) field strength would vary along the line shown dotted in the diagram (Fig. e 37).

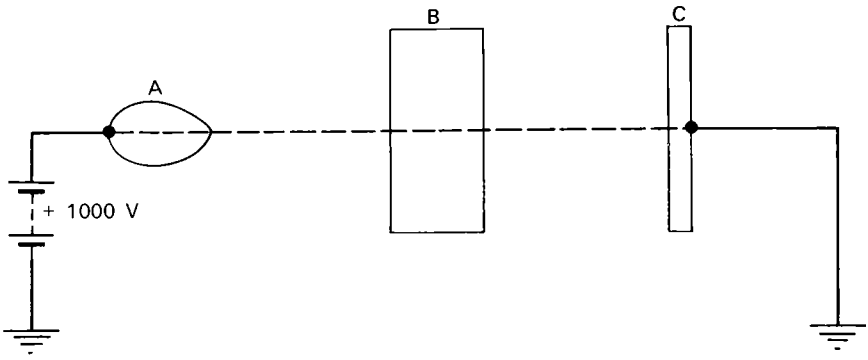


Fig. e 37

6.13 Two small conducting spheres, both positively charged, are placed 10 cm apart. If sphere A holds + 10 units of charge, and sphere B holds + 5 units of charge, sketch the lines of force and equipotential surfaces in the space between the spheres.

Show graphically how (*a*) the potential, (*b*) the electric field strength varies along the line which passes through the centres of the two spheres, and mark on each graph any point where the field strength is zero.

6.14 Answer the same questions as in 6.13, for the case when the charge on A is + 10 units and the charge on B is −5 units. Is (*a*) the potential, (*b*) the field strength zero at any point other than one lying on the direct line of the two spheres?

6.15 Two flat metal sheets, A and B, are held parallel to each other a distance of 0.5 cm apart, and a p.d. of 250 V is applied between them. A small oil drop in the field between the plates experiences a force, due to the field, of 32×10^{-15} N. What is the charge on the oil drop?

6.16

The two metal plates X, Y, each fixed to an insulating handle, are designed to demonstrate the existence of an electric field, such as there is between the

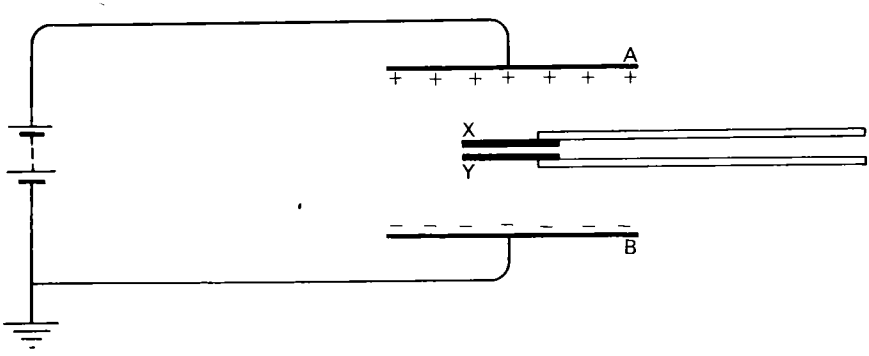


Fig. e 38

two plates A and B. The process is to place them as shown, but to lay them together in contact with each other; they are then separated, and each is tested by means of an ice pail and an electroscope. Explain the principle of this procedure.

Should the experiment show null effect, owing, you suspect, to the field

being too weak for the sensitivity of the electroscope, suggest how, by proceeding a little further in your use of the ice pail, you could obtain some result after all.

6.17 Two flat metal plates are arranged horizontally, one above the other, and separated by a distance of 3 mm. When the p.d. between the plates is adjusted to be 450 V, with the top plate positive, a small oil drop is observed to be held stationary between the plates. If the mass of the drop is 1.2×10^{-14} kg, what is the magnitude of the charge held by the drop ? If by some process, the drop were to gain a further charge of $+3.2 \times 10^{-19}$ C, what would the resultant force on the drop be now ?

(Take the value of g to be 10 m/s^2.)

6.18

The two points P, Q, situated on the line OPQ which passes through the centre of the charged metal sphere A, are 0.5 cm apart. Owing to the charge

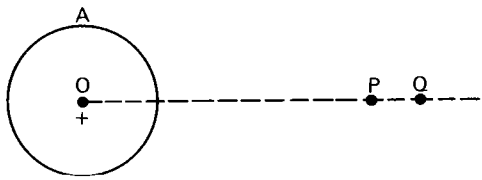


Fig. e 39

on the sphere, the potentials at P and Q respectively are 300 V and 216 V. If a small body, carrying charge $+9 \times 10^{-9}$ C, is placed between P and Q, estimate the size of the force which acts on it.

6.19

A small charged metal sphere P held on an insulating handle, is contacted to A, thus charging P to a potential of +300 V. The potential of the hollow sphere B is already +1000 V.

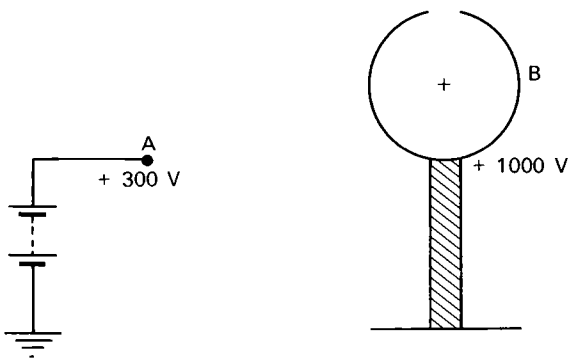


Fig. e 40

If the small sphere P is now removed from A and contacted to the inside of B, positive charge flows from P to B, although B was initially at the higher potential. In view of this result, what do you estimate to be the approximate potential of P when it is inside B, but not yet making contact with B ? Explain how P obtained this potential.

6.20

The diagram shows a 12 V battery connected between leaf and case of an electroscope. Such a p.d. is not sufficiently high to produce any effect on the

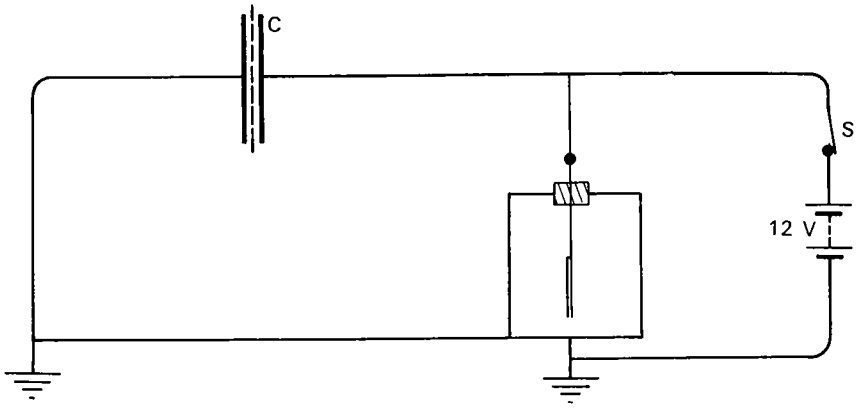


Fig. e 41

leaf, which remains undeflected. The electroscope is also connected, however, to arrangement C, which consists of two very flat metal plates held very close together, but separated by a single sheet of discharged polythene; one plate is connected to the electroscope leaf, and the other to the case.

After first connecting the battery, switch S is opened, thus leaving the electroscope connected only to C. The earthed plate and the polythene sheet are now removed; the leaf is found to rise.

(*a*) Explain this behaviour.

(*b*) How would you proceed to remove any initial charge from the polythene sheet before starting the experiment?

(*c*) How would you test the polythene for charge?

(*d*) Describe how you would adapt this experiment to demonstrate that the battery terminal marked positive has opposite polarity to the charge on a rubbed polythene rod?

6.21 It is required to construct a capacitor of value 0.20 μF by using flat metal sheets separated by mica (relative permittivity $\epsilon_r = 6.0$) of thickness 10^{-3} cm. Calculate the area required for each metal sheet. What charge would be stored at a p.d. of 300 V?

6.22 Define electrostatic potential. How would you show experimentally (*a*) that the potential is the same at all points on the surface of an insulated charged conductor, and (*b*) that the potential of such a conductor is proportional to the charge on it?

Describe the Van de Graaff generator, and explain how this machine produces high voltages by making use of simple electrostatic principles.

The sphere of a model Van de Graaff apparatus has a capacitance of 5×10^{-11} F; the leakage resistance to earth is 10^{13} Ω, but when the potential reaches 3×10^5 V the insulation of the air breaks down and the sphere discharges to earth with a spark. Find the minimum charging current which will enable a potential of 3×10^5 V to be reached; also find the rate at which the potential of the sphere begins to rise immediately after spark over if the charging current is three times this minimum value. (O)

Chapter 7

Basic magnetism

7.1 Magnetic force

There are two aspects of a magnetic force: it can be exerted by a magnet, or it can be exerted by a current, i.e. by a moving charge. Likewise, it acts on a magnet, or on a moving charge. The interaction between currents is the more fundamental and important aspect, but we will find it convenient to begin a description of elementary phenomena with the behaviour of a permanent magnet. This is where, historically, the original study began. The name magnetism derives from the locality Magnesia in Asia Minor, where an iron ore already naturally magnetized was first discovered. This is the black oxide of iron, Fe_3O_4. The main properties of the permanent magnet have been known for a long time. They were studied, for example, by Humphrey Gilbert in the reign of Queen Elizabeth I. The magnetic field of an electric current, on the other hand, was first detected by Oersted in 1819.

7.2 The permanent magnet

A magnet can be made only from one or other of the small group of metals and their alloys which are said to be ferromagnetic, of which iron is the most important member, followed by nickel and cobalt. Also, a magnet exerts its very considerable attracting force only on materials which belong to this class.

The magnet's power of attraction is centred on two patches which are usually, but not essentially, near the ends of the magnet. These areas are called the poles of the magnet. Ideally the poles are sometimes treated as being concentrated at two points, but the theory which accounts for the magnetism suggests that this cannot be the case in practice. The conception of a magnetic field arises for the same reason as applies to the electric field, namely the fact that magnets exert their influence at a distance, even across empty space.

Because the earth has a magnetic field, a pivoted magnet acts as a compass, settling itself in the earth's field in such a way that its magnetic axis, i.e. the line joining its poles, lies approximately along the geographical meridian. For this reason, one pole of the magnet is called its north pole, meaning its north-seeking pole, while the other is called its south pole.

The like poles of two magnets are found to repel each other, and unlike poles to attract. The magnet also attracts a piece of iron. It does so, because it first magnetizes the iron, with an unlike pole at the near end, and then attracts the nearer pole more strongly than it repels the more distant one. This explains why a magnet attracts only ferromagnetic materials; it does so only if the material can be magnetized. It can also be seen why a horseshoe magnet exerts its attraction with special efficiency (Fig. 7.1).

It sometimes happens that, when the south pole of a weak magnet is brought up very close to the south pole of a powerful one, the force between the two turns out to be attraction, and not the expected repulsion between like poles. The explana-

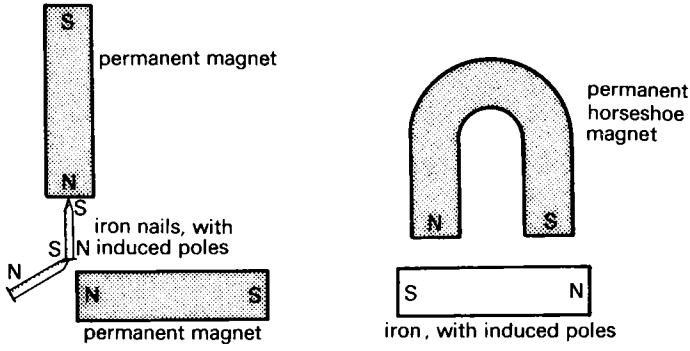


Fig. 7.1 Induced poles. A magnet attracting iron (a) A magnet attracts iron nails by first magnetizing them. A second magnet may be used to test their induced polarity (b) A horseshoe magnet attracts iron with special efficiency because both the induced poles in the iron are attracted

tion is that the stronger magnet has temporarily induced magnetism in the weaker one with reverse polarity, and thus attracts it. This simple experiment has some significance in quantitative magnetism, because it illustrates that we cannot always rely upon a magnet's poles having unchangeable strength, in the same way as electric charge is unchangeable.

This explanation of how a magnet exerts its attraction is faintly analogous to the reason that a charged polythene rod attracts a piece of pith, by first causing a redistribution of charge in the pith. The electrical case is called charging by induction, and the magnetic case is called magnetizing by induction. The analogy is not very close, however, because electric charges flow in the conducting pith, and there is no flow of 'magnetic charges' in the iron. The analogy is closer between the magnetizing of the iron and the polarizing of a dielectric when placed in an electric field.

7.3 The magnetic field. Lines of magnetic flux

Because magnetic forces, like electric forces, act at a distance, it is convenient to postulate the existence of a magnetic field in the space which surrounds a magnet, and this can be represented by drawing lines of magnetic flux, otherwise called lines of magnetic induction. The choice of these alternative names is made, rather than the more obvious one, lines of magnetic force, because in the historical development of the subject, this name has been allotted to a slightly different type of field line, which will not be used in this book.

The direction of a line of magnetic flux may be defined in several ways, all giving the same result. The simplest definition to choose is the one which is most often applied in practice – the direction of a line of magnetic flux at any point marks the direction in which a compass would point, if it were placed there. The arrows placed on flux lines show the direction of the force on a north-seeking pole. (It may appear that, if the poles of a magnet are, in practice, not points but patches, there is no definite axis to the compass which can give precise direction to a line of magnetic flux. This difficulty can be surmounted, however, by turning the compass upside down on its pivot. There must be one line in it which points in the same direction as it did before, and this is the magnetic axis, and the direction of the flux. This is illustrated in Fig. 7.2, for the case when the compass card is a disc.)

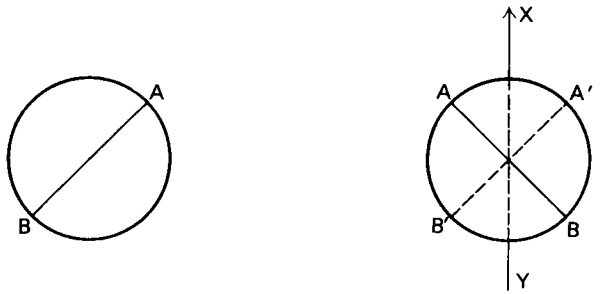


Fig. 7.2 The magnetic axis of a compass disc (a) Plan view of a pivoted compass card. AB is an arbitrary line marked along a diameter (b) The compass card has been turned over on its pivot. AB now sets in the new direction shown, A′B′ being the previous direction The direction of the flux, and the magnetic axis of the compass disc, is YX, where XY bisects the angle between AB and A′B′

7.4 The earth's magnetic field

The earth's magnetic field at any place consists effectively, over a small area, of parallel lines of magnetic flux, and a vertical plane containing such a line is called the plane of the magnetic meridian. The direction of the magnetic meridian at any point makes a small angle with the geographical meridian, and this angle is called the magnetic declination, or the magnetic variation, at the point.

Taking a view of the earth as a whole, the overall effect is as if there were a small, but very powerful, magnet imbedded at the centre of the earth, with its axis approximately along the earth's axis. Generally speaking, the earth's flux lines dip downwards in the northern hemisphere, and upwards in the southern hemisphere, with the result that the part of the earth's field, its horizontal component, which affects the compass, is strongest in the vicinity of the equator, and very weak near the magnetic north and south poles of the earth. The angle which the earth's flux makes with the horizontal at any place is called the angle of dip.

There are considerable local variations of the overall picture, and, in the laboratory, what we call 'the earth's field' is a field much affected by such materials as steel girders, magnets concealed in cupboards, and so on.

It will be noticed that a tendency would be expected for a compass to dip downwards in the northern hemisphere, and that it becomes reluctant to work at

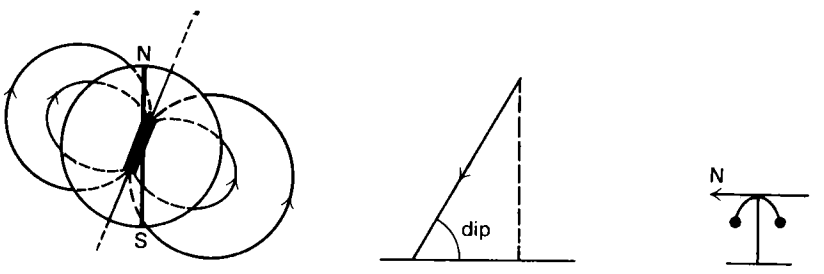


Fig. 7.3 The earth's magnetic field (a) The earth behaves as if there were a short magnet at its centre (b) Dip in the northern hemisphere (c) The weighted cup of a compass

all anywhere near the earth's magnetic poles, where the force is nearly vertical. The dip of the compass is counteracted, in practice, by gravity; the compass has a

weighted cup at its pivot, which keeps it horizontal, by putting its centre of gravity below the pivot.

7.5 The magnetic flux of a bar magnet

Three examples are shown in Fig. 7.4 of the magnetic field of a bar magnet, as modified by the effect of the earth's field. Where the resultant field is zero, owing to the earth and magnet having equal and opposite effects, the point is called a neutral point.

Two particular features of the flux lines in these diagrams are worth noting. First, the arrows on a line of magnetic flux point away from the north pole of a magnet, but towards the earth's magnetic north. This source of possible confusion

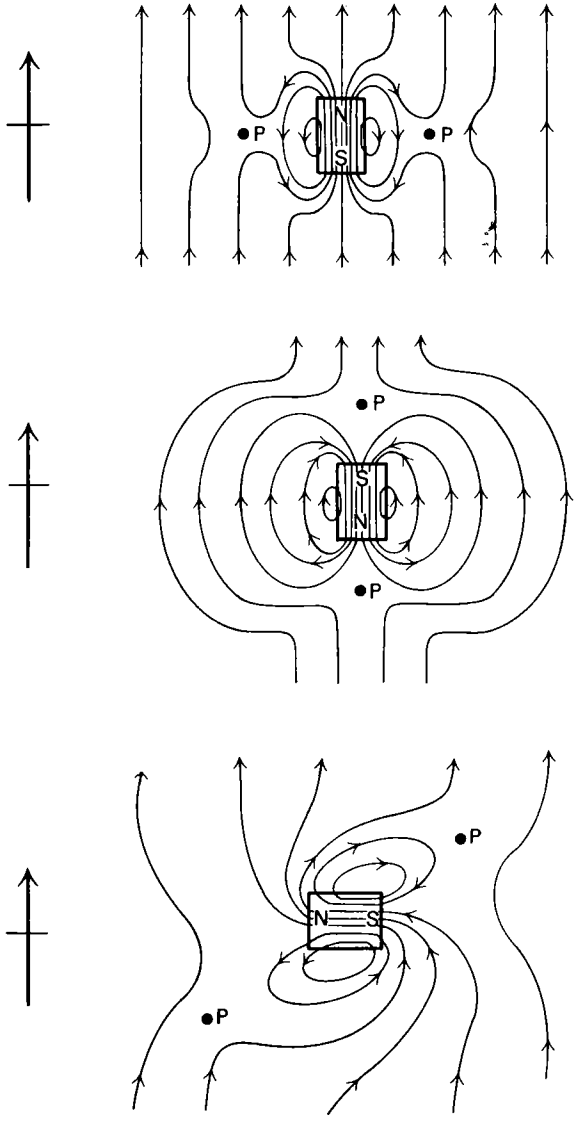


Fig. 7.4 A bar magnet in the earth's field. P = neutral point (a) North pole pointing north (b) North pole pointing south (c) North pole pointing west

arises from the nomenclature: the north pole of a magnet is a north-seeking pole, and repels the similar north pole of a compass; but the earth's magnetic north attracts the north pole of the compass, by definition, being the north which the compass pole seeks. The other point is that the flux diagram may suggest the possibility of certain effects which actually are found to occur. Diagram 7.4 (*c*) for example, has the appearance of flux lines pulling the magnet round, to act as a compass, and to set in the magnetic meridian, and this tendency on the part of the magnet does, in fact, exist.

7.6 Magnetism in other materials

In addition to ferromagnetic materials, there is a group of substances, known as ferrites, which are not so strongly magnetic as iron, but are effective enough in their magnetism to be useful in certain applications, having the specialized advantage of being effectively electric insulators. These are composite oxides, containing both iron and some other metal, and they are described as being ferrimagnetic.

All other substances display only a very slight reaction to magnetic influences, but the theoretical importance of their behaviour is, of course, considerable, since a theory which accounts for the special behaviour of iron must also account for the more ordinary behaviour of other materials. This very small magnetic effect can be detected by suspending small elongated specimens of various substances in the field of a powerful magnet. The substances are then found to belong to one or other of two distinct groups. The first group resembles a ferromagnetic material in its behaviour, but in a very slight way; the specimen finally settles down to lie lengthwise along the direction of the field, which is what a small piece of iron would do. Such substances are called paramagnetic. All other materials belong to the other group, known as diamagnetic; these settle down to lie transversely to the field instead of lengthwise. A diamagnetic material, placed in a magnetic field, weakens the flux inside it; a paramagnetic or a ferromagnetic material strengthens it.

7.7 The magnetic field due to current in a conductor

When current flows in a conductor, it produces the same magnetic forces as a magnet does, although not so powerfully. In particular, the magnetic field due to current in a solenoid resembles that of a bar magnet, and the behaviour of a solenoid is similar to that of the magnet, except that, being hollow, it can attract iron right into its centre. The solenoid exhibits polarity, much as the magnet does, and it can be suspended or pivoted, and used as a compass. An explanation of the magnetism of iron, and of the paramagnetic and diamagnetic properties of other materials, is to be sought in the motion of the electrons which surround the nucleii of the atoms, and we may regard the magnetic field of a moving charge as being the basis of the magnet's field also. This aspect will be discussed in a later chapter (chapter 16).

The field of a solenoid may be broken down to give the field due to current in a single straight conductor, and Fig. 7.5 shows the lines of magnetic flux in both these cases. A contrast between lines of magnetic flux and lines of electric force is evident here. Lines of electric force start from positive electric charge and terminate on equal negative charge. Lines of magnetic flux, on the other hand, form closed loops, which surround a conductor carrying current. It is assumed that, in the alternative case when the flux is caused by a magnet, these closed loops

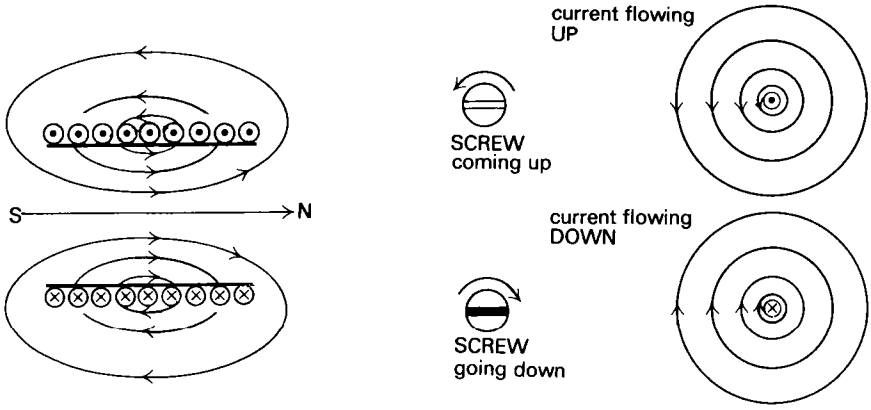


Fig. 7.5 The magnetic flux caused by an electric current (a) The magnetic flux of a solenoid (b) The magnetic flux due to current flowing in a straight conductor

thread the length of the magnet, entering the magnet at its south pole and leaving at its north pole, as they do in a solenoid. The field diagrams of Fig. 7.4 were drawn on this assumption.

A simple memory aid in deciding upon the direction of the arrow on these flux lines is Maxwell's screw rule. If a right handed screw is imagined to move in the direction taken by the current, the direction of the flux is the same as that in which the screw must be turned.

Two further properties of a bar magnet add to its resemblance to a solenoid, and are consistent with the view that the flux lines run through the length of the magnet. If a magnet is broken into two parts, this does not isolate the separate poles, but the new ends have polarity, and two separate magnets result, each with two poles. Accepting that the poles are merely the places where the flux leaves and enters the magnet, this result is to be expected. The other property is the fact that the two poles are equal and opposite in strength. This may be demonstrated by floating a magnet on a cork, on water, and placing it in a uniform field, such as that due to the earth; the magnet swings round, acting as a compass, to lie in the direction of the field, but it does not move as a whole along the line of the field, because its poles are pulled equally in opposite directions. This is true also, and only to be expected, in the case of the solenoid.

7.8 The magnetism of different ferromagnetic alloys

When an iron core is placed in a solenoid, the magnetic flux is greatly increased, owing to the fact that the iron becomes strongly magnetized, and its contribution greatly exceeds that of the current in the solenoid (unless the current is exceptionally large). Ferromagnetic materials are divided into two classes in this respect. Soft iron and soft alloys of iron magnetize readily, making powerful electromagnets. Their magnetization is easily reversed when the magnetizing field is reversed, which makes them suitable for use as the cores of such devices as the transformer, which are to be used with alternating current. Hard alloys are magnetized with more difficulty, but they do not easily lose their magnetism, and they make good permanent magnets. The powerful permanent magnets used, for example, in loud speakers, or for use with magnetron valves, are hard steels containing metals added to the iron, such as aluminium, nickel, cobalt, and copper.

7.9 A comparison between lines of electric force in a dielectric and lines of magnetic flux in a magnetic material

When a magnetic core is added to a solenoid, while keeping the current unchanged, the flux due to the magnetised material adds to the flux already there, resulting in a strengthened magnet. When dielectric is placed between the plates of a capacitor, isolated to hold constant charge, the field outside the dielectric is unaffected, and the field inside is reduced. This contrast between the two cases is consistent with the essential difference between lines of magnetic flux and lines of electric force already referred to – that lines of magnetic flux are continuous, forming closed loops, whereas lines of electric force arising directly from electric charge start and terminate on equal and opposite charges, and suffer an abrupt change in number in consequence, at the boundary of a polarized dielectric.

It is due to differences of this kind that we do not call the magnetic lines, lines of force; in this aspect of their behaviour they more closely resemble the behaviour of a different kind of field line used in electrostatics, but not used in this book – lines of electric displacement, sometimes called lines of electric induction.

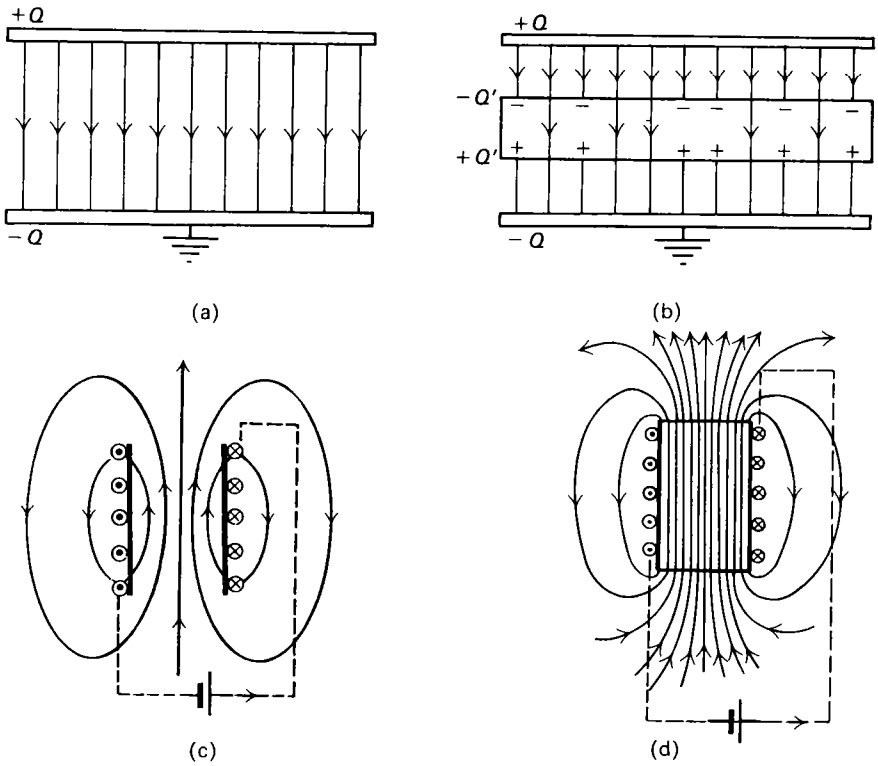


Fig. 7.6 A comparison between lines of electric force and lines of magnetic flux. (a) Lines of electric force E (b) Lines of electric force through polarized dielectric (c) Lines of magnetic flux B (d) Lines of magnetic flux through a magnetic core

7.10 The magnetic force on a current carrying conductor – the motor effect

If current in a straight conductor or in a solenoid produces a magnetic field, it must likewise experience a force if suitably placed in a magnetic field. This amounts to

saying that, if a conductor carrying current exerts force F on a magnet, then the magnet must exert equal and opposite force F on the conductor; this is an application of Newton's third law of motion. A similar conclusion applies to interaction between two conductors carrying current. The appearance of this force experienced by the conductor when placed in a magnetic field is often called the motor effect, because the electric motor is a practical application of it.

Two examples of the motor effect are shown in Fig. 7.7. The first three diagrams, illustrating the interaction between current in a straight conductor and a horseshoe magnet, demonstrate the connection with Newton's third law. The first diagram shows the field of the conductor, which exerts a downward force F on the magnet, made up from a force $\frac{1}{2}F$ on each pole. The second diagram shows the field of the magnet, which exerts an upward force F on the conductor. The lines of magnetic flux in the third diagram represent the resultant field of magnet and conductor,

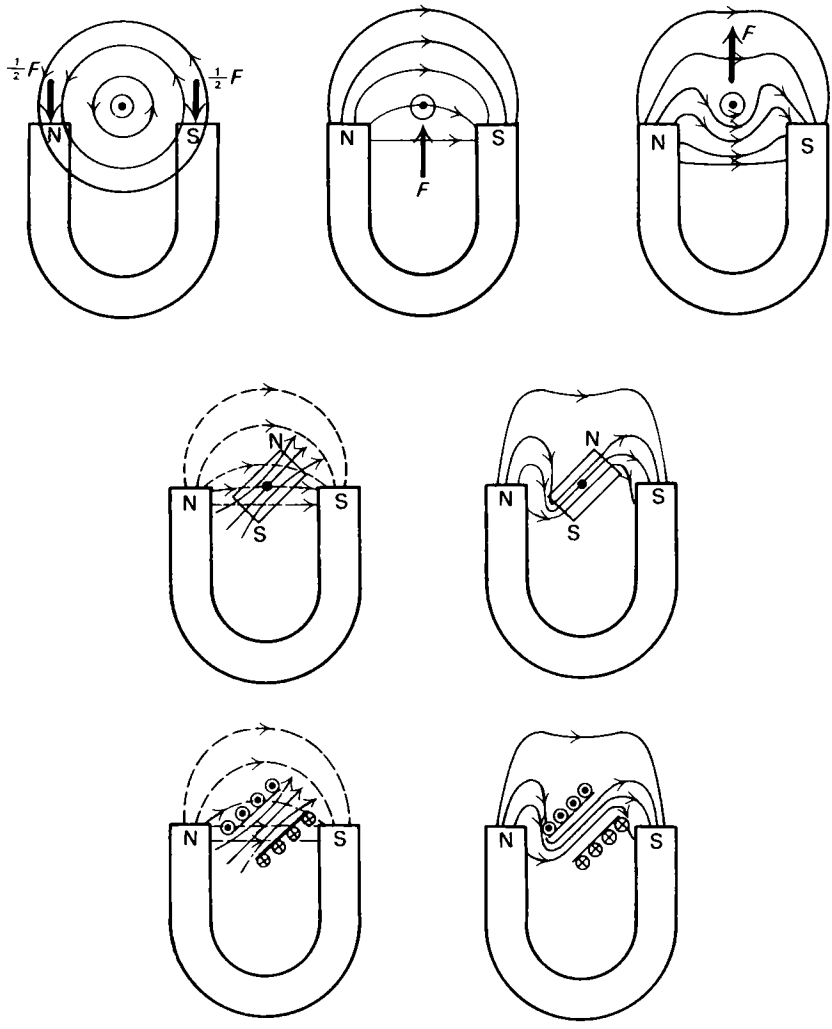


Fig. 7.7 Two examples of the motor effect (a) The field due to the current (b) The field due to the magnet (c) The combined field (d) The separate fluxes of horseshoe and pivoted magnet (e) The combined flux of horseshoe and pivoted magnet (f) The separate fluxes of horseshoe and pivoted solenoid (g) The combined flux of horseshoe and pivoted solenoid

and their shape clearly suggests the existence of an upward force on the conductor.

The last four diagrams illustrate the motor effect as being an example of the similarity between a solenoid and a magnet. A magnet, pivoted in a field, acts as a compass, turning to bring its own flux parallel to the external flux. The solenoid does the same thing.

Figure 7.8 is included merely to stress the difference in direction between the force exerted by a magnetic field on the pole of a magnet and the force exerted on a conductor which is carrying current. The force which acts on the north pole of a magnet is in the direction of the flux, which is not surprising, since this is how we

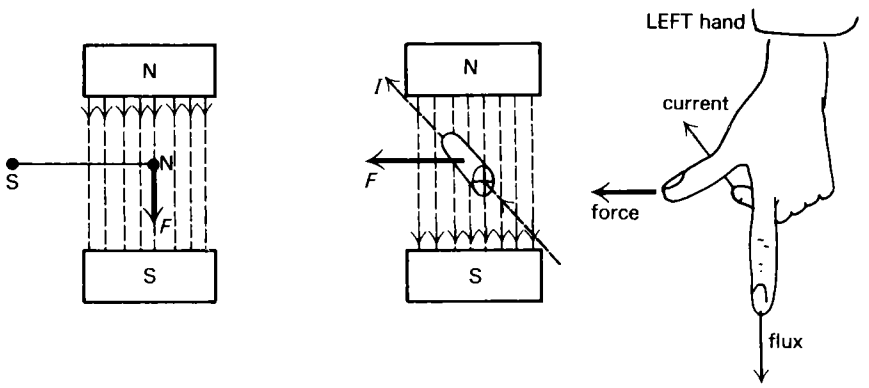


Fig. 7.8 Showing the direction of the force produced by magnetic flux (a) The north pole of a magnet (b) A conductor carrying current (c) The left-hand rule

chose to define the direction of the flux. The straight conductor, on the other hand, experiences a force, only to the extent to which the conductor is perpendicular to the flux; the force is perpendicular to both these directions. The left hand rule is sometimes used as a guide, and this is illustrated in the diagram.

7.11 The interaction between magnetic flux and moving charge

The electrons in a cathode ray oscilloscope, or charged particles crossing a cloud chamber, are deflected by magnetic flux in a manner which is consistent with the effect on current in a conductor. If the charge is moving in a direction perpendicular to the flux, the deflecting force is in the third perpendicular direction. This is how the deflections are achieved by the scanning coils of a television tube.

If the charge is moving along the lines of magnetic flux, there is no deflecting force, which suggests a definition of the direction of a magnetic field, alternative to defining it as being the direction in which a compass points, i.e. the direction of a magnetic field may, alternatively, be defined as the direction along which a stream of charged particles must travel in the field, if they are not to be deflected by it.

7.12 The flux density B

The strength of a magnetic field may be measured in terms of the force experienced by a moving charge, or of the force experienced by a conductor carrying current.

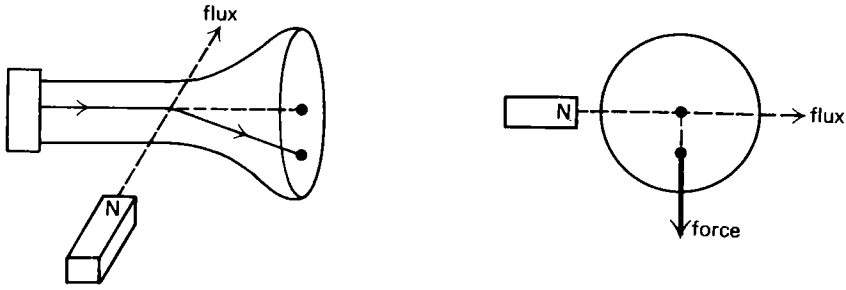


Fig. 7.9 The deflection of electrons in a cathode-ray tube

The alternative defining equations, which will later be shown to be consistent with each other (chapter 14), are

$$B = \frac{F}{Qv} \tag{7.1}$$

and

$$B = \frac{F}{Il}, \tag{7.2}$$

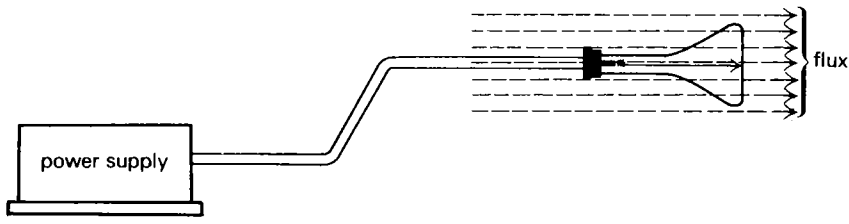


Fig. 7.10 The use of a small oscilloscope to establish the direction of a magnetic field. The tube is held to obtain zero deflection

where, in the first equation, F is the force which acts on charge Q if the charge is moving with velocity v in a direction which is perpendicular to the flux. In the second equation, F is the force which acts on length l of a straight conductor carrying current I if the conductor is held perpendicular to the flux.

The vector B, which is a measure of the strength of the magnetic field, is sometimes called the magnetic induction. More often, however, it is referred to as the magnetic flux density, a name which arises from the convention of representing it by drawing lines of magnetic flux. At a place where the flux is uniform, we imagine that a unit area placed perpendicular to the field would be crossed by an amount of flux (or a number of lines), equal to B, and hence that the total flux which crosses an area A is equal to Φ, where

$$\Phi = BA. \tag{7.3}$$

A comparison with electric field strength E

Equation (7.1) may usefully be compared with equation (6.2), which referred to electric field strength E. We write the two equations in the forms

$$F = EQ, \tag{7.4}$$

$$F = BQv. \tag{7.1}$$

These summarize the distinction between the electrostatic field, which exerts force on a stationary charge Q, and magnetic flux density, which exerts force on charge Q only when this charge is moving relative to the magnetic flux in the appropriate direction.

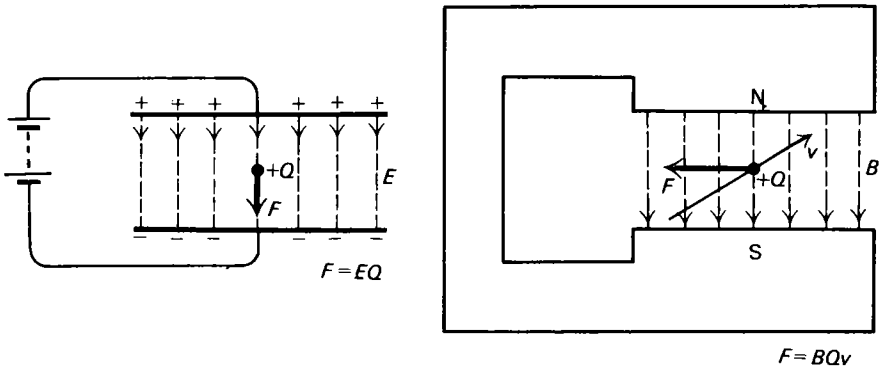


Fig. 7.11 Comparing the electric force on stationary charge and the magnetic force on moving charge

7.13 Electromagnetic induction

When the motor force, $F = BIl$, is applied to cause motion of the conductor which carries the current, which is the principle of the electric motor, the result is that energy is converted from electrical to mechanical form. This energy conversion is a reversible process, just as the energy conversion in electrolysis is reversible; if mechanical energy is put into a suitably designed motor instead of taking it out, which can be done by turning it by mechanical means, the motor then acts as a dynamo, and becomes a source of current, which means that it has an e.m.f. in it. This is termed an induced e.m.f., and the process is known as electromagnetic induction.

The simplest case of electromagnetic induction is that of a single conductor being moved across magnetic flux. Referring to Fig. 7.12, if a current were made to flow through the conductor AB from an external source in the direction from A to B, the resulting motor force would act upwards on the conductor, as applying the left hand rule would suggest. If, on the other hand, the conductor is moved upwards by exerting a force from outside, as indicated in the diagram, the conductor becomes a source of current, driving current through itself from B to A, and consequently from A to B in the external circuit. In other words, the motion across the flux has caused an e.m.f. to be induced in the conductor, directed from B towards A. This e.m.f. must also be present when the motion has been caused by the motor effect, in which case it is a back e.m.f., directed against the current flow, and this accounts for the acceptance of energy from the external source of current, for conversion to mechanical energy.

It will be observed that when an induced e.m.f. is to result, the conductor must be moved in such a way that it cuts the magnetic flux. If, for example, it were moved from left to right in the diagram, or pushed in and out, perpendicular to the plane of the diagram, there would be no e.m.f., and no flux would be cut.

Because the direction of the motor effect can be found by applying the left hand rule, it follows that the direction of an induced e.m.f., relative to the motion of

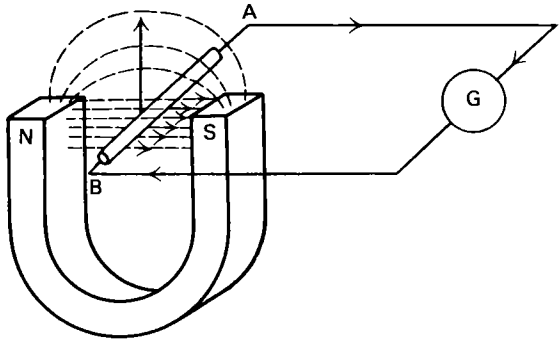


Fig. 7.12 Electromagnetic induction – a conductor cutting flux

the conductor which is cutting flux, can be found instead by using the right hand rule.

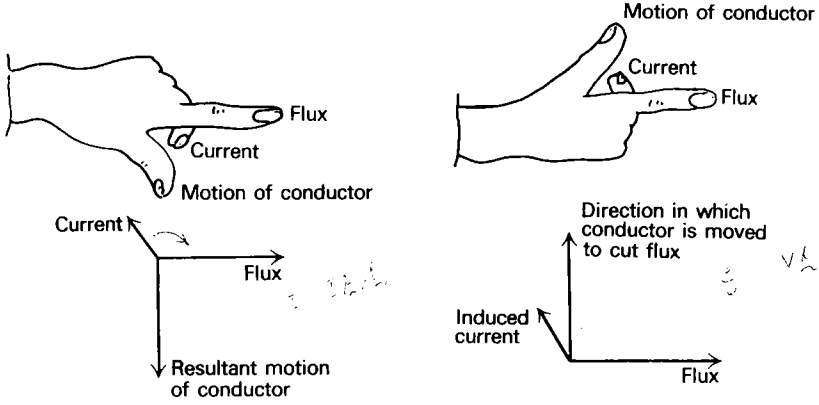


Fig. 7.13 The left-hand and right-hand rules (a) The motor effect. Left-hand rule (b) Electromagnetic induction. Right-hand rule

The important consideration, however, is conservation of energy: if induced current flows, taking energy out, its direction is such that it necessitates energy being put in. This means that the current acts against the process which induces it. This is fundamental in the example we have been discussing; if induced current flows in the conductor as a consequence of moving it upwards, the direction of the current is such that there is a motor force acting on it downwards.

A simple experiment which demonstrates this principle is illustrated in Fig. 7.14. If a bar magnet is brought up towards a suspended coil, some of the flux from the magnet cuts the coil while the magnet is moving, and an e.m.f. is induced. If the coil is short circuited, a current flows in it, so that the coil temporarily acquires its own magnetism. The direction of the current is such that the coil repels the magnet, and the magnet repels the coil, and this effect causes the coil to swing away as the magnet approaches. If the magnet is now pulled away from the coil, the induced current flows in the reverse direction, and the coil swings in pursuit of the magnet.

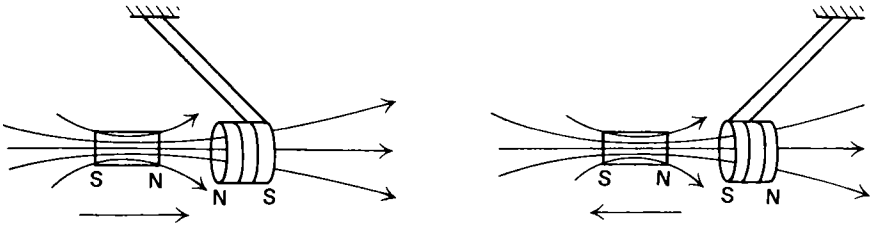


Fig. 7.14 The induced current in a short-circuited coil causes it to follow the motion of the inducing magnet

7.14 Electromagnetic induction by changing the linkage of flux

Electromagnetic induction was discovered by Michael Faraday in 1834, and at the same time, independently, in America, by Joseph Henry. Neither of these experimenters discovered it in the first place, however, by making a conductor move across flux. The apparatus used by Faraday resembled that shown in Fig. 7.15.

Two coils, A and B, quite separate from each other, are wound round the same iron toroid, and a battery and switch are connected across coil A, a galvanometer being in series with coil B. When the current in coil A is switched on or off, and at no other time, the galvanometer in the circuit of B is deflected, showing that an e.m.f. has temporarily been induced in coil B. This e.m.f. acts in one direction when the current in coil A is switched on, and in the opposite direction when the current is switched off. This form of electromagnetic induction is called mutual induction. It is the process which occurs in a transformer.

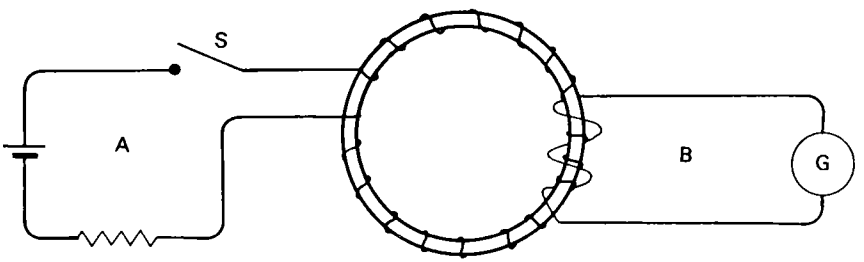


Fig. 7.15 Mutual induction. Electromagnetic induction by changing the flux linkage with a coil

The e.m.f. in this case cannot accurately be attributed to a conductor cutting flux, because nothing except the switch contact has been moved. This is a separate effect from that of flux cutting, and the e.m.f. induced in coil B is identified as being the result of changing the number of lines of magnetic flux which are passing through the coil. Since an e.m.f. is induced in each turn of the coil, and these are in series, the resultant e.m.f. is proportional to the number of turns in the coil; it is also proportional to the rate at which the flux threading the coil changes, and we can write

$$E \propto N \frac{d\Phi}{dt}. \tag{7.5}$$

The product $N\Phi$ is sometimes referred to as being the total flux linkage, where Φ represents the total flux, or the total number of lines of flux, which thread the coil.

In this case, we can say that the induced e.m.f. E, is proportional to the rate at which the flux linkage is changed.

7.15 The laws of electromagnetic induction

The laws of electromagnetic induction are framed to include both ways by which an e.m.f. can be induced, viz:

Faraday's law. Whenever a conductor cuts magnetic flux, an e.m.f. is induced in it, which is proportional to the rate at which the flux is cut. Also, whenever the quantity of magnetic flux which links a circuit is changed, an e.m.f. is induced in the circuit, which is proportional to the rate at which the flux linkages are changed.

Lenz's law. The direction of an induced e.m.f. is such that, if induced current flows, its effect will hinder the process by which it is induced.

7.16 Faraday's disc

The evidence which justifies the quantitative aspect of Faraday's law will be discussed in a later chapter (chapter 14). One part of this, however, is relatively easily

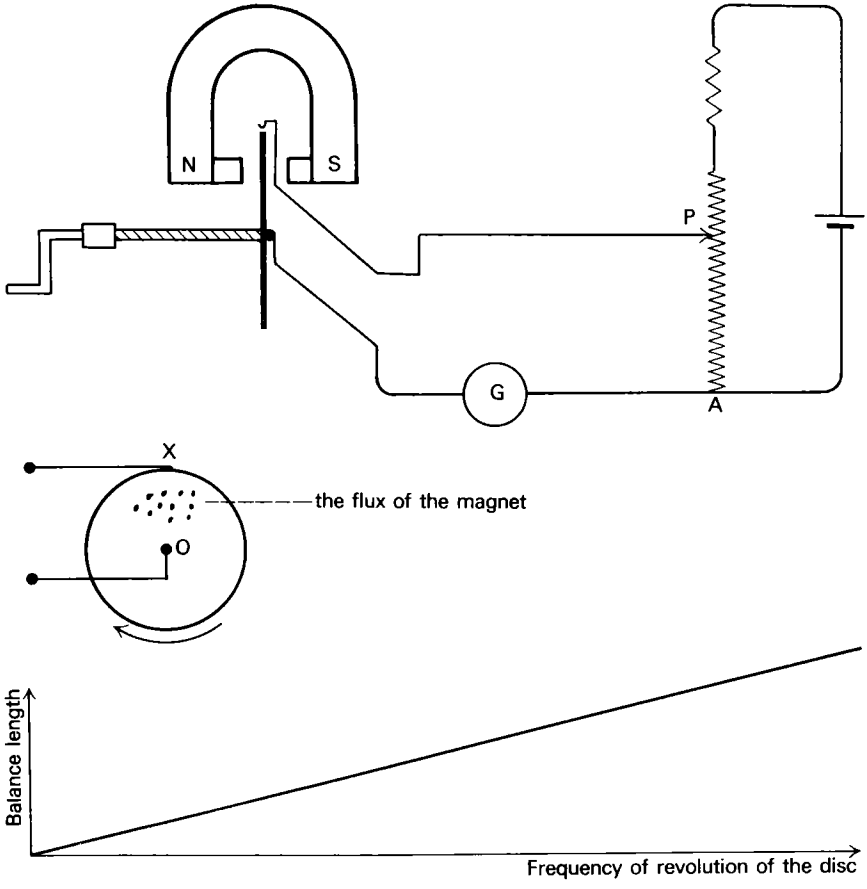


Fig. 7.16 Faraday's disc (a) (b) (c)

tested, and this is the effect of changing the speed with which a conductor is moved to cut flux. Doubling the speed of rotation of a dynamo armature, for example, would be expected to result in doubling the induced e.m.f. Faraday's disc is a simple form of dynamo, in which this basic principle is made very clear. A metal disc is mounted to rotate about a horizontal central axis, perpendicular to its own plane, in such a way that a radius OX (Fig. 7.16) is cutting the flux of a powerful horseshoe magnet. The p.d. between the axle and the point X on the rim of the disc is balanced on a potentiometer. The disc is revolved by hand, through a step-up gearing, so that it may be turned quite fast. The balance length AP, may be plotted against rate of revolution of the disc, to demonstrate that the induced e.m.f. is in proportion to the speed at which the disc is turned, and consequently to the rate at which the radius OX is cutting flux.

Because Faraday's disc is a short-circuited dynamo, and current is flowing in it, it can also be used to demonstrate Lenz's law. The disc is noticeably easier to turn when the magnet is removed, because this removes the motor action between the induced current and the magnet, which acts against the motion. This current flows because the flux of the magnet is localized in the region of the radius OX. In other radii, such as OY in Fig. 7.17, there is no e.m.f., and current flows in the direction shown in this diagram.

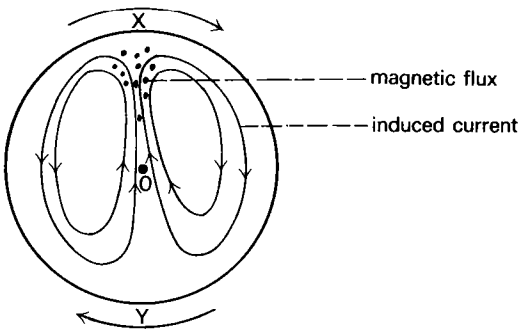


Fig. 7.17 The induced currents in Faraday's disc

7.17 Eddy currents

Currents induced in a mass of metal, such as this current which flows in Faraday's disc, are known as eddy currents. A further example is illustrated in Fig. 7.18. A brass cylinder is suspended about its central axis to hang between the poles of a horseshoe magnet. If the cylinder is made to perform torsional oscillations about this axis, the motion is very sluggish, and the cylinder soon comes to rest, as if it were being made to move through treacle. This is because the eddy currents induced in it by its motion in the flux rapidly dissipate its energy. If a stack of metal discs, insulated from each other, copper coins, for example, is substituted for the cylinder, this oscillates quite freely, because the eddy currents cannot flow through the high resistances which separate the discs.

Eddy currents are particularly apt to occur where there is alternating current. If, for example, a coil carrying alternating current is wound round a metal block, the changing flux in the centre of the metal induces an e.m.f. in the outer part, just as the motion in the flux does in the example just described. Consequently an eddy current flows in the metal, dissipating energy by heating the metal. The effect is

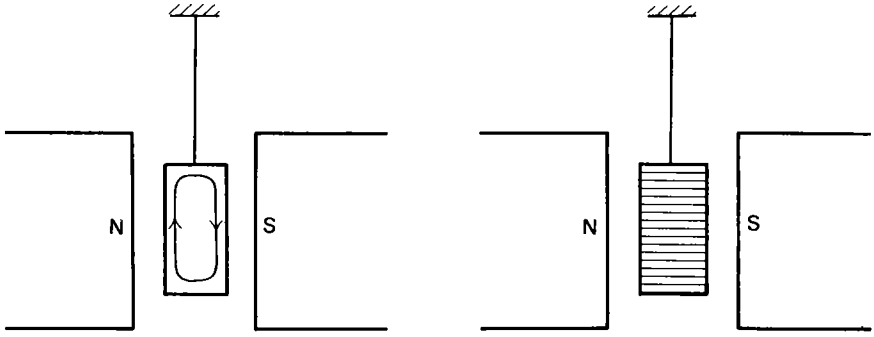


Fig. 7.18 Eddy currents (a) The torsional oscillation of a solid metal cylinder is heavily damped by eddy currents (b) A stack of insulated metal discs oscillates freely

particularly pronounced at high frequency, and this is the principle of the eddy current furnace.

More often, eddy currents in the cores of alternating current machinery are an objectionable cause of energy loss, to be avoided if possible. The iron cores of such electrical equipment as transformers are consequently made up from layers of thin stampings, insulated from each other, on the principle of the demonstration just described. At higher frequencies, the less effective magnetic properties of ferrimagnetic materials is accepted, on account of their high resistance, and this is the reason for using a ferrite aerial rod in a radio receiver.

7.18 Mutual inductance and self inductance

The electromagnetic induction which occurs in an arrangement of two coils, such as that shown in Fig. 7.16, is called mutual induction. If the rate of change of current in one circuit, the primary circuit, is dI_1/dt and the resulting e.m.f. induced in the other, the secondary circuit, is E_2, the ratio of this e.m.f. to the rate of change of current is called the mutual inductance M,

i.e. $$E_2 = M\frac{dI_1}{dt}. \tag{7.6}$$

The SI unit of mutual inductance is called the henry (H),

i.e. $$1\ \mathrm{H} = 1\ \mathrm{V.A^{-1}.s}.$$

An arrangement of two coils which has been specifically designed in order to have mutual inductance is a mutual inductor, and two special forms of it are the transformer and the induction coil. The symbols used for a mutual inductor in a circuit diagram are shown in Fig. 7.19.

Fig. 7.19 Symbols for a mutual inductor (a) Air-cored (b) Iron-cored

For an air-cored mutual inductor, the mutual inductance M is a constant. This is because the flux is proportional to the primary current. This is not so, however, if there is an iron core, because the additional flux caused by the iron is not proportional to the current, and the mutual inductance of an iron-cored inductor is not a precise quantity.

Self induction

Electromagnetic induction also occurs in a single coil if the current is changed, and, indeed, to a small degree, it occurs in any circuit. This is because a coil, and, to some extent, a single conductor, is inevitably linked by its own magnetic flux. Hence, if the current is changed, a momentary e.m.f. is induced. This is self induction. If the coil has many turns and, in particular, if it has an iron core, the self induction can be large. In d.c. circuits, self induced e.m.f.s appear only when the current is switched on, or when it is switched off, and not when it has a steady value.

The ratio of the self induced e.m.f. E to the rate of change of current $\mathrm{d}I/\mathrm{d}t$ is called the self inductance, or simply the inductance L,

i.e.
$$E = -\mathrm{L}\frac{\mathrm{d}I}{\mathrm{d}t}. \tag{7.7}$$

The SI unit is, again, the henry.

The negative sign has been included in this equation to conform with Lenz's Law. The self induced e.m.f. always acts in opposition to the change in current, which means that it is directed against a growing current, but it acts with a dying current, having a tendency to maintain it.

As with mutual inductance, the coefficient L is a constant, unless there is an iron core, in which case its value depends very much on circumstances. A device designed to have inductance is an inductor, and one special form of it, used to smooth out variations in current, is called a choke. The symbols used for inductance in a circuit diagram are shown in Fig. 7.20.

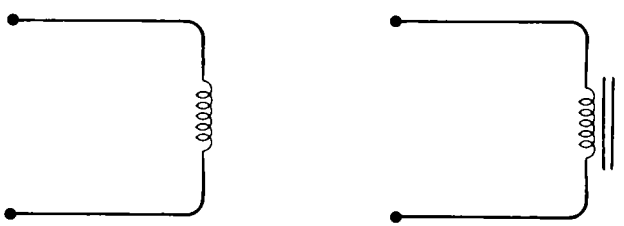


Fig. 7.20 Symbols for inductance (a) Air-cored (b) Iron-cored

7.19 Some practical forms of inductor

Small inductors are simply coils of wire, with or without a ferromagnetic core; for use at high frequency, such cores are either 'iron dust' or ferrites, as already mentioned, to keep down eddy currents. A large value inductor or choke is a coil wound on a former, with an iron core made from insulated stampings so joined together that the magnetic flux may be confined to the iron, with no air gaps.

Standard resistors

Self induction has a nuisance value in bridge and potentiometer circuits, because it may cause a momentary galvanometer deflection on making or breaking a contact,

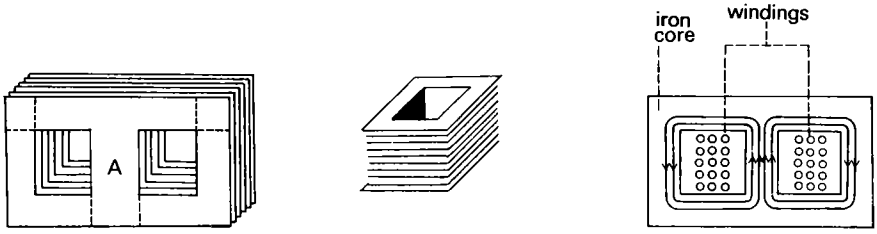


Fig. 7.21 A choke (a) The soft iron stampings (b) The coil (This is slipped over the central limb, A, of the iron core) (c) The flux

which makes the adjustment for balance more difficult to detect. For this reason and because the components may be used in a.c. circuits, where inductance is particularly important, standard resistors are non-inductively wound. The wire is folded back on itself before winding on its bobbin, so that there shall be no magnetic flux due to the current.

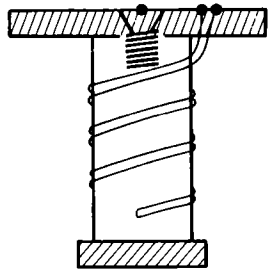


Fig. 7.22 A non-inductive resistor

7.20 The effect of a steady p.d. applied across an inductor

In practice, an inductor possesses some resistance in addition to its inductance, and when current flows the p.d. V across the component is given by

$$\begin{aligned} V &= IR - E \\ &= IR + L\frac{dI}{dt}, \end{aligned} \tag{7.8}$$

where E is the self induced e.m.f.

Taking the ideal case of an inductor having zero resistance,

$$V = -E = L\frac{dI}{dt}. \tag{7.9}$$

This equation implies that, if a steady p.d. were applied across the ideal inductor, the current would continually grow at a constant rate dI/dt, such that the induced e.m.f. E would be equal and opposite to the applied p.d. In the practical case, the current initially increases at this rate, but the effect of resistance is to limit the final value of the current to $I = V/R$, with the result shown in Fig. 7.23. This will be compared with the behaviour of a capacitor, at the start of the next chapter.

7.21 A transformer

A transformer is constructed on the same principle as the choke shown in Fig. 7.21, except that two coils are wound, instead of one, on the former which surrounds the

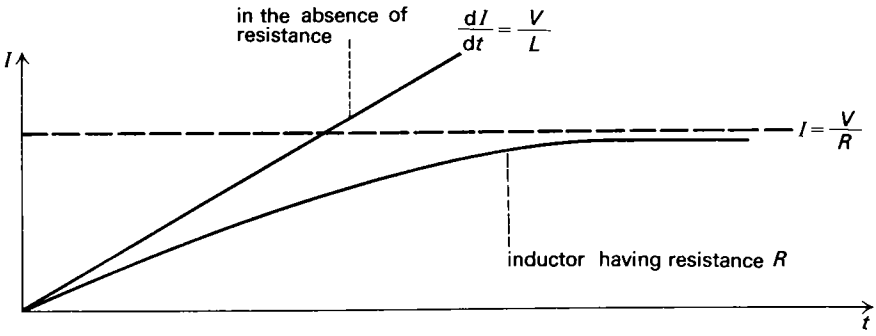


Fig. 7.23 Growth of current in an inductor

central core; these are referred to as the primary and secondary windings, respectively.

The transformer is used to change the p.d. of an alternating current by the effect of mutual induction. If a p.d. equal to V_1 is applied across the primary winding from an alternating source, the secondary winding acts as an alternating source with e.m.f. equal to E_2, where

$$\frac{E_2}{V_1} = \frac{N_2}{N_1}, \tag{7.10}$$

N_2, N_1 being the number of turns of the primary and secondary winding, respectively. The mutual induction occurs because the alternating current causes alternating flux in the core of the transformer, and this induces an e.m.f. in both primary and secondary, in proportion to the number of turns.

The transformer will be discussed in more detail in chapter 19.

QUESTIONS

7.1 Explain the meaning of (*a*) an electric field, (*b*) a magnetic field, and how, in each case, you would define the direction of the field.

A stationary electron, charge -1.6×10^{-19} C, experiences a force in the direction Ox of magnitude 4.8×10^{-15} N. When the electron is moving in the direction Ox with a velocity of 10^7 m/s, it experiences an additional force in the direction Oy equal to 3.2×10^{-15} N, whereas if it moves in the direction Oz it experiences no additional force.

Explain these effects in terms of electric and magnetic fields, their direction and magnitude.

7.2 When an electric current flows in a straight metal conductor there is magnetic flux in the space round the conductor. There is, however, no electric field; explain why this is so, and discuss the circumstances in which a force would act, or not act, on an electric charge in the vicinity.

7.3 The SI unit of electric field strength E is 1 V/m or 1 N/C; the SI unit of magnetic flux density B is 1 W/m^2 or 1T. Discuss whether, in practical terms each of the units is inconveniently large, or inconveniently small, and illustrate your answer by considering a simple practical arrangement in each case.

7.4 The diagram represents two poles of a large sized horseshoe magnet, which produce a uniform magnetic field of flux density 3×10^{-2} T in the space between them.

(*a*) Calculate the force which acts in each of the following cases:

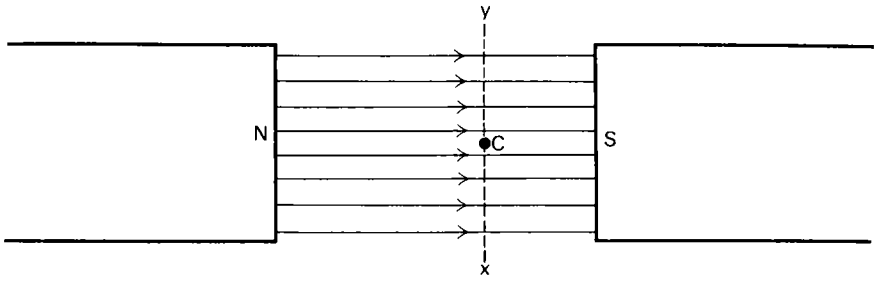


Fig. e 42

(i) A straight wire is held perpendicular to the plane of the diagram so that it passes through the point C. The wire carries a current of 2.0 A and an effective length of 5 cm lies in the field.

(ii) The same wire is placed in the field but in the plane of the diagram so that it makes an angle of 60° with the lines of magnetic flux.

(iii) A small body, carrying charge 5×10^{-12} C, is placed at C.

(iv) The same charged body is at C, but is moving in the direction *xy* at 500 cm/s.

(*b*) A circular loop of wire, radius 4.0 cm, is placed in the field with its plane perpendicular to the flux. Calculate the total flux, in webers, which passes through the loop.

(*c*) If, by some device, the flux threading the loop were to disappear at a uniform rate, taking 6.0 s to vanish, deduce the value of the e.m.f. induced in the loop during this time.

7.5 The diagram shows an iron ring, which has magnetic flux in it, due to the current flowing in circuit A.

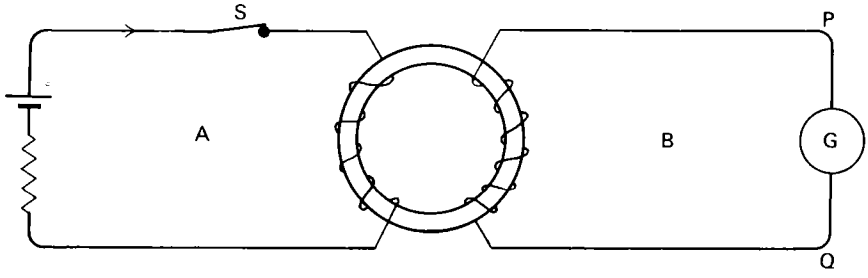


Fig. e 43

(*a*) Is the flux in the ring clockwise, or anticlockwise? Explain the basis on which you answered this question.

(*b*) When the switch S is broken, an induced current momentarily flows in circuit B. Is the direction of the current in the galvanometer from P to Q, or from Q to P? Explain your reasoning.

7.6 A dynamo has an e.m.f. of 100 V. When a current of 20 A is taken out of the dynamo, the armature becomes harder to turn than was the case when no current was being taken out. Explain why it becomes harder to turn, and identify the source of the 2000 W of power which are being taken out.

7.7 When the magnet shown in the diagram is being moved towards the coil, an induced current flows in the galvanometer. Why? This current momentarily causes the coil to behave as a weak electromagnet; which end of the coil

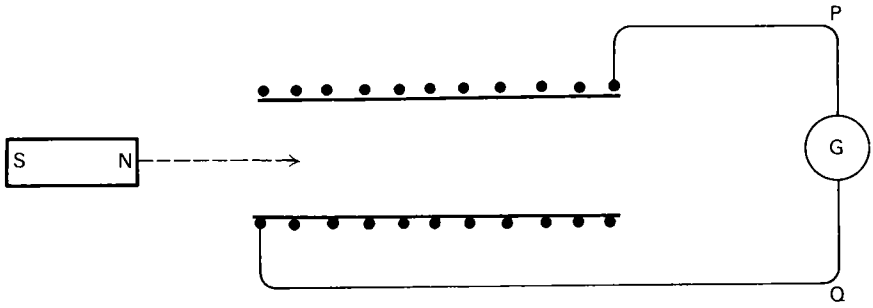


Fig. e 44

is its N pole? On what basis did you decide? In which direction does the induced current flow through the galvanometer – from P to Q, or from Q to P? Explain your reasoning.

7.8 Two coils, A and B, have mutual inductance $M = 3 \times 10^{-2}$ H. The current in A is increasing at a uniform rate, changing from zero to 10 A in 20 s. What is the e.m.f. induced in coil B during this time?

When coil B is open circuited, coil A acts as an inductor, its self inductance L being 2×10^{-2} H. If the coil is connected to a 12 V d.c. source in series with a 13 A fuse wire, estimate how long will elapse before the fuse blows after switching on the current.

(Assume the resistance of the circuit to be negligible.)

7.9 In an induction coil, the current in a primary coil produces a magnetic flux of 8.0×10^{-2} Wb, which threads a secondary coil of 5000 turns. When the magnetizing current is switched off, the flux vanishes in 10^{-1} s. What is the mean value of the induced e.m.f. in the secondary coil during this short period of time?

7.10

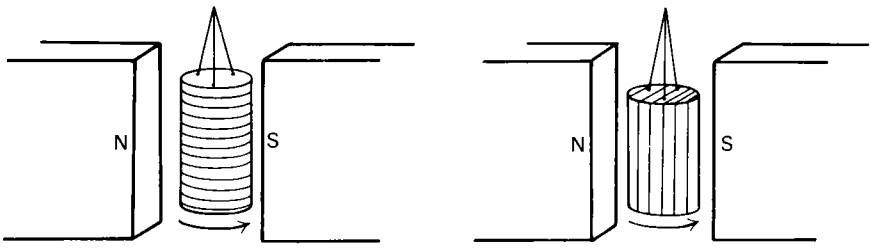


Fig. e 45

The diagram in each case represents a metal cylinder (not iron) suspended between the poles of a powerful horseshoe magnet and set revolving. In (*a*) the cylinder consists of a stack of dirty coins, whereas in (*b*) it is built up from a stack of vertical, flat metal sheets. It is found that, whereas cylinder (*a*) revolves freely, cylinder (*b*) is very sluggish and soon comes to rest. Also, a solid metal cylinder resembles (*b*) in its behaviour. Explain these effects, and state where, in the damped cases, the energy goes.

7.11 A coil of inductance 10 H is joined to a battery with e.m.f. 20 V, the total circuit resistance being 5 Ω.

Deduce (*a*) the rate of growth of current at the instant when the circuit is closed, (*b*) the final value of the current, (*c*) the mean e.m.f. induced when the circuit is broken, if the current falls to zero in 0.10 s.

Chapter 8

Capacitors and inductors in a circuit

8.1 Comparing capacitance, resistance, and inductance

When it comes to their functions in an electric circuit, the capacitor and the inductor can usefully be considered together, since they are in many ways complementary to each other in their behaviour. A comparison with the resistor may be summarized by the three defining equations

$$V = \frac{Q}{C}, \tag{8.1}$$

$$V = RI, \tag{8.2}$$

$$V = L\frac{dI}{dt}, \tag{8.3}$$

where V is the p.d. applied across the component, in each case, and in equation 8.1 we could substitute

$$Q = \int I\,dt.$$

The results of applying a steady p.d. may be briefly compared as follows. Whereas the resistor draws a steady current continuously, the capacitor accepts only a transient current when the p.d. is first applied, and this dies away to zero by the time that the capacitor has charged up to have this p.d. across its plates. The inductor, on the other hand, while it delays the initial growth of current, due to a back e.m.f. induced in it by the growing magnetic flux, can offer no opposition to a steady current, and it is only the resistance included in its circuit (perhaps only the resistance of its windings) which finally keeps the current down to a steady value at all. Thus the capacitor and inductor are complementary in their behaviour, a fact which can be further illustrated by writing equation (8.1) in different form. Differentiating this equation, and transposing, gives

$$\frac{dV}{dt} = \frac{1}{C}\cdot\frac{dQ}{dt},$$

where the current

$$I = \frac{dQ}{dt},$$

whence

$$I = C\frac{dV}{dt}, \tag{8.4}$$

as compared with equation (8.3) for the inductor

$$V = L\frac{dI}{dt}. \tag{8.3}$$

Thus, the p.d. across a capacitor cannot change without a current flowing to alter the amount of charge stored in it. It is the current, on the other hand, which cannot change in an inductor, without there being a p.d. across it.

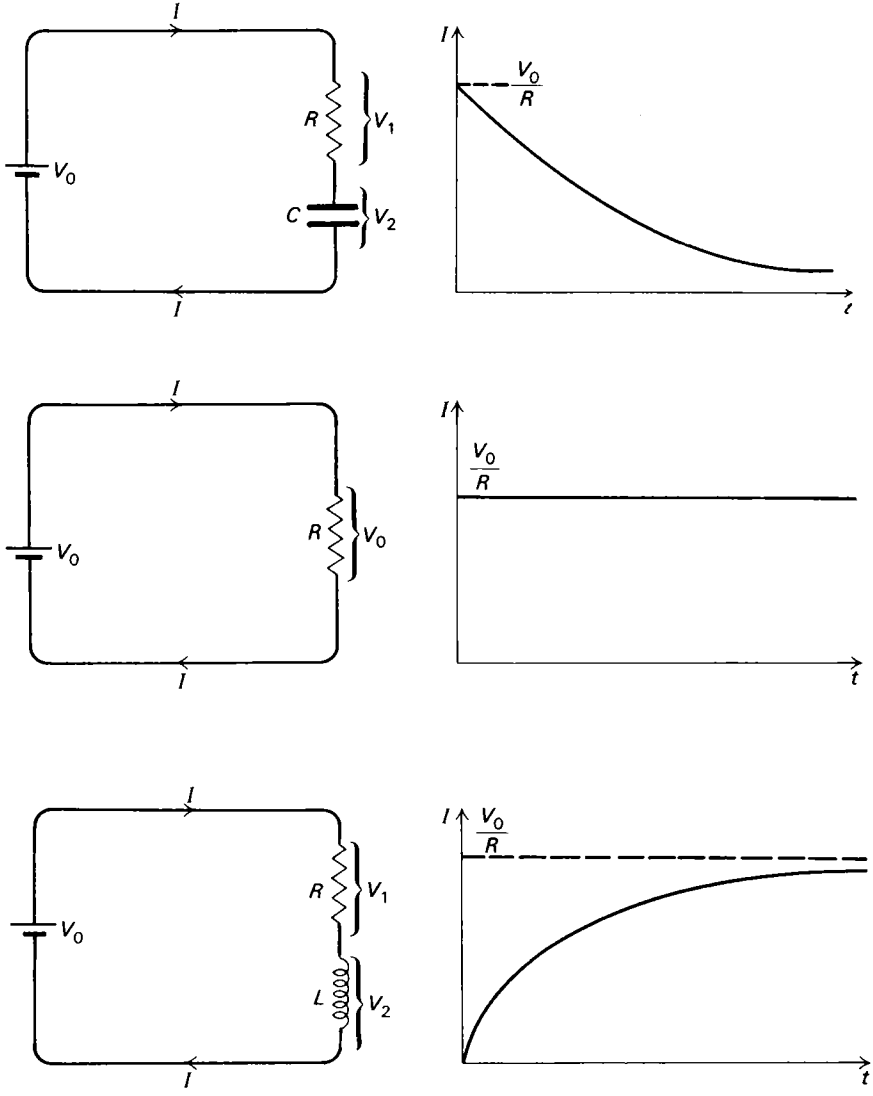


Fig. 8.1 A comparison between capacitor, resistor, and inductor, placed in a circuit (a) The capacitor (b) The resistor (c) The inductor

Applying Kirchhoff's second law to each of the circuits drawn in Fig. 8.1, we have

$$V_0 = V_1 + V_2$$

where:

For the capacitor,

$$V_1 = RI,$$

$$V_2 = \frac{1}{C}Q.$$

Initially, there is no charge on the capacitor, and so $V_2 = 0$. The p.d. is entirely across the resistor, and the initial value of the current is

$$I = \frac{V_0}{R}.$$

Finally, the whole of the p.d. is across the charged capacitor,

i.e. $$Q = CV_0,$$

and the current has fallen to zero.

For the inductor,

$$V_1 = RI,$$

$$V_2 = L\frac{dI}{dt}.$$

Initially, the current is zero, and $V_1 = 0$. The p.d. is entirely across the inductance, and the initial rate of growth of current is

$$\frac{dI}{dt} = \frac{V_0}{L}.$$

Finally, the current has ceased to grow, there is no p.d. across the inductance, and it is entirely across the resistance. The current has now reached a steady value, given by

$$I = \frac{V_0}{R}.$$

[The mathematical equations of the curves drawn in Fig. 8.1 will be derived at the end of this chapter.]

The above contrast in behaviour can be demonstrated quite strikingly by connecting a torch bulb in series with each of the three components, giving the arrangement shown in Fig. 8.2.

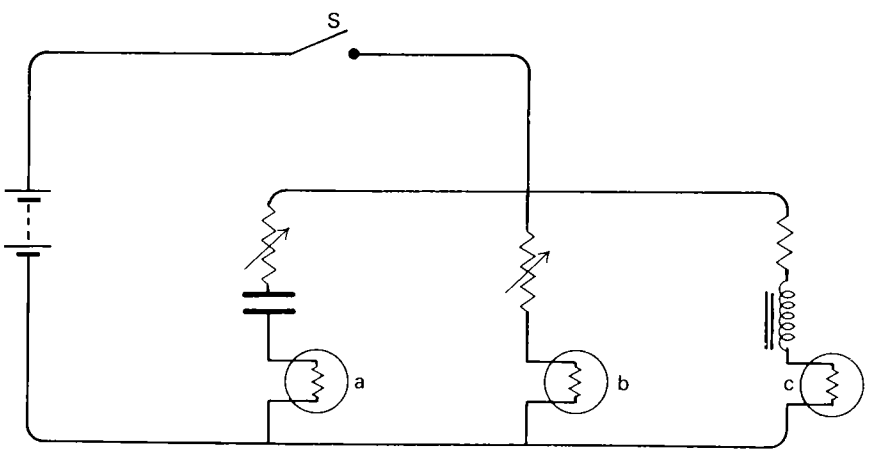


Fig. 8.2 A circuit to contrast the behaviour of capacitor, resistor, and inductor

When the switch S is made, the effect of the three current-time graphs appears in luminous form: bulb (*a*) initially flashes, as the capacitor charges up, and then goes out; bulb (*b*) lights up at once, and remains alight; bulb (*c*) lights up after

some delay, and then remains alight. This demonstration requires a very high value capacitance, e.g. a 5×10^4 μF electrolytic capacitor, and a large iron-cored choke, the resistance shown in series with this being in fact the resistance of the choke itself.

8.2 The energy stored in a charged capacitor

The charge which flows into a capacitor carries energy with it. This energy is not dissipated as heat, as it would be in a resistor; it is stored in the charged capacitor, and we may regard it as being stored in the electric field between the plates, and in the polarized dielectric.

Suppose the capacitor in Fig. 8.3. to be initially discharged. It finishes with charge Q and p.d. V, where

$$Q = CV.$$

The first increment of charge reaches the capacitor when the p.d. across it is zero, and so brings no energy with it, whereas the last increment arrives at the full p.d. V.

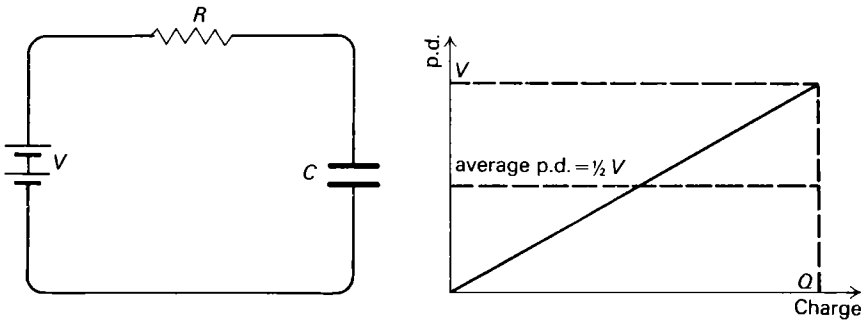


Fig. 8.3 (a) (b) Variation of p.d. with charge, assuming C to be constant

Assuming that the capacitance C is constant, which is true of most capacitors, the average p.d. during the charging process is equal to $\frac{1}{2}V$ and so the energy stored in the charged capacitor is

$$W = \tfrac{1}{2}QV. \tag{8.5}$$

Measuring Q in coulombs and V in volts, this energy W is in joules. Combining equation (8.5) with equation (8.1) gives the alternative equations

$$W = \tfrac{1}{2}CV^2, \tag{8.6}$$

$$W = \tfrac{1}{2}\frac{Q^2}{C}. \tag{8.7}$$

The charge was supplied by the battery, however, at the constant p.d. V, with the result that the energy W_0 taken from the battery was

$$W_0 = QV, \tag{8.8}$$

i.e. the capacitor has stored only half the energy taken from the battery. The explanation of this is that the charging current flowed through the resistor R, across which the average p.d. was also $\frac{1}{2}V$, and the remaining energy, $\frac{1}{2}QV$, was

dissipated in the resistor. This result is quite independent of the value of R, even in the extreme case when $R = 0$.

More strictly, equations (8.5), (8.6), (8.7) can be derived by using calculus, viz:

$$W = \int V \mathrm{d}Q$$

$$= \frac{1}{C}\int Q \mathrm{d}Q$$

$$= \tfrac{1}{2}\frac{Q^2}{C},$$

from which the other two equations follow.

The problem of storing electrical energy

It might be noted, again, that the difficulty about storing electrical energy is that you cannot store much of it. The storage capacity of accumulators, although limited, is useful, but a capacitor offers no competitition at all, even to this. $10^5\ \mu$F is a large capacitance for a single capacitor. If this were charged to a p.d. of 100 V, the energy stored in it would be

$$W = \tfrac{1}{2}CV^2$$

$$= \tfrac{1}{2} \times 10^5 \times 10^{-6} \times 10^4 \text{ J}$$

$$= 500 \text{ J},$$

which is not a large amount.

8.3 A mechanical analogy to the capacitor

The capacitor is analogous to a spring in mechanics. Provided that the spring obeys Hooke's law, the tension T is proportional to the extension x,

i.e. $$T = \lambda x, \tag{8.9}$$

where λ represents the stiffness of the spring, an elastic constant for the spring.

The energy W stored in the stretched spring is given by

$$W = \tfrac{1}{2}Tx, \tag{8.10}$$

or $$W = \tfrac{1}{2}\lambda x^2. \tag{8.11}$$

These equations compare with equations (8.5) and (8.7), where the reciprocal of capacitance, $1/C$, is analogous to the stiffness λ, and charge Q is analogous to extension x. Thus a powerful spring, which takes a large force to stretch it appreciably, corresponds to a small value capacitor, which takes a large p.d. to put appreciable charge into it.

In all such comparisons as this, it is a little unfortunate that capacitance has been defined as the ratio Q/V, and not as the inverse ratio, V/Q. This is noticeable in comparing equations (8.1), (8.2), and (8.3) (but not 8.3 with 8.4).

8.4 The energy associated with current in an inductor

When current is flowing in an inductor, some energy is stored in it, which we may regard as being stored in the magnetic field, and in the magnetized iron core, if there is one.

8.4 The energy associated with current in an inductor

Consider the circuit of Fig. 8.4. At any instant, when the growing current has the value I, the potential difference V across the inductance is

$$V = L\frac{\mathrm{d}I}{\mathrm{d}t}.$$

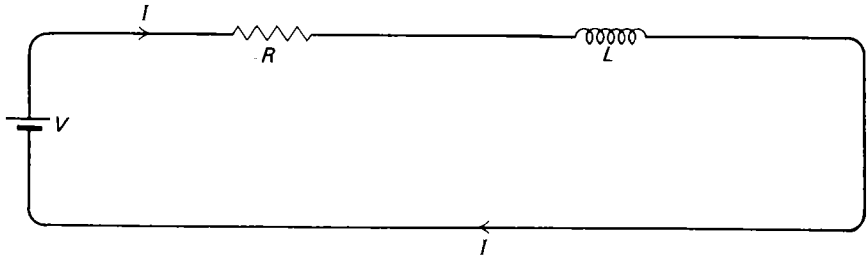


Fig. 8.4

Energy is being fed into the inductance, at this instant, at the rate

$$VI = L\frac{\mathrm{d}I}{\mathrm{d}t}\cdot I$$

whence, the total energy stored in the inductance, when the current has reached the value I, is given by

$$W = \int VI\mathrm{d}t,$$

$$= \int_0^{\mathrm{I}} LI\mathrm{d}I. \qquad (8.12)$$

Provided that there is no iron core (but only very approximately if there is one), we can take L to be constant, in which case the equation integrates to give

$$W = \tfrac{1}{2}LI^2. \qquad (8.13)$$

Comparing this with equation (8.6), we can say that energy stored in inductance is associated with the magnetic field of a current, whereas energy stored in capacitance is associated with the electric field of a p.d. (or of electric charge).

8.5 A mechanical analogy to the inductor

Making charge Q analogous to displacement x, and p.d. V analogous to force or tension T, as was done in the last section, current $I = \mathrm{d}Q/\mathrm{d}t$ becomes analogous to velocity $v = \mathrm{d}x/\mathrm{d}t$.

The mechanics equations which correspond, on this analogy, to equations (8.3) and (8.13) would appear to be

$$F = m\frac{\mathrm{d}v}{\mathrm{d}t}, \qquad (8.14)$$

and

$$E_{\mathrm{k}} = \tfrac{1}{2}mv^2, \qquad (8.15)$$

making inductance L analogous to mass m; i.e. whereas mass produces an inertial force opposing any effort to change its velocity, inductance produces a back e.m.f. opposing any tendency to change the current.

We can cite several examples of this in practice.

8.6 The choke as an electrical hammer

When a falling hammer is stopped in its course very rapidly on striking a nail, its impulsive force on the nail is very large, greatly in excess of the weight of the hammer head. Similarly, if the current in a choke is abruptly switched off, the induced e.m.f. is short-lived, but very large, and can be greatly in excess of the p.d. which was used to send the current. This effect may be demonstrated by using the circuit of Fig. 8.5.

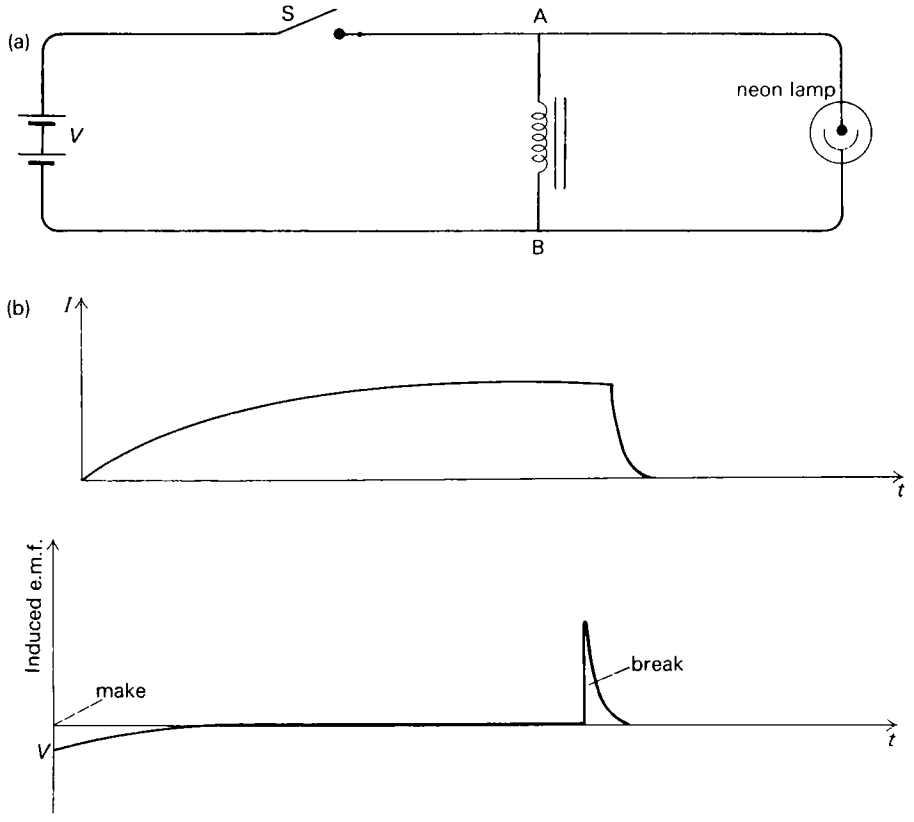


Fig. 8.5 Demonstration of the large induced e.m.f. in a choke when the current is switched off (a), (b)

The potential difference V provided by the battery is insufficient to light the neon lamp, and the lamp does not light when the current is switched on. If time is allowed for the current to grow to maximum in the choke, however, and it is then abruptly switched off, a large induced e.m.f. causes the lamp to flash. Alternatively, the experimenter may connect himself across AB, in place of the lamp. The battery e.m.f. may be only 6 or 12 V and he feels nothing when the current is switched on, but the shock can surprise him when the current is switched off. The difference between switching on and switching off is illustrated in Fig. 8.5 (*b*). When we switch on, there is a complete circuit, and the induced e.m.f. at no time exceeds the p.d. of the battery, V. The initial rate of growth of current adjusts itself so that the induced e.m.f. is initially equal and opposite to V. The current obeys the equation

$$V = RI + L\frac{dI}{dt}, \tag{8.16}$$

where the induced e.m.f. $= -L\dfrac{dI}{dt}$.

But when the current is switched off, it must stop very rapidly, since the circuit is not complete. In practice, there is likely to be arcing across the switch, which slows things down a little, but the rate of change of current will still be high.

This experiment works rather more efficiently if a capacitor is connected across the switch. This is because arcing does not then occur, and a very rapid oscillation of current in and out of the capacitor takes its place. This will be more readily understood after reading the next section.

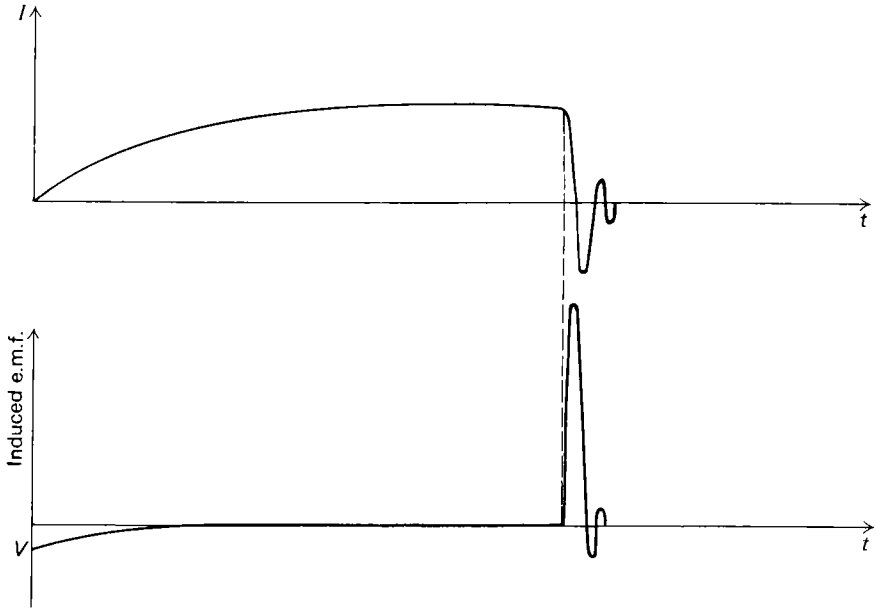


Fig. 8.6 The effect of putting a capacitor across the switch S in the circuit of figure 8.5

Self induction can be a nuisance with switchgear when large currents are involved. There may be dangerous arcing at the switch when the current is cut off, and the switches must be designed to allow for this.

Another mechanical analogy to the effect just described is involved in the design of the household water tap. The tap is made to turn the water flow off gradually, lest the impulsive force involved in stopping the flow too suddenly were to burst the pipe.

8.7 The oscillating circuit

If a mass is attached to the end of a stretched spring, and then released, an oscillation results. This is because, in contracting, the spring sets the mass in motion, and the momentum of the mass carries the spring past its equilibrium

position. The energy is alternately potential energy, $\frac{1}{2}\lambda x^2$, of the strained spring, and kinetic energy, $\frac{1}{2}mv^2$, of the moving mass.

The analogous electrical arrangement is an inductor connected across a charged capacitor. The capacitor, in discharging, sets a current flowing in the inductor, and self induction causes this flow to continue after the capacitor has discharged, recharging it with reversed polarity. An oscillating current results. The energy is alternately 'potential energy', $\frac{1}{2}Q^2/C = \frac{1}{2}CV^2$, in the charged capacitor, and 'kinetic energy', $\frac{1}{2}LI^2$, in the inductor.

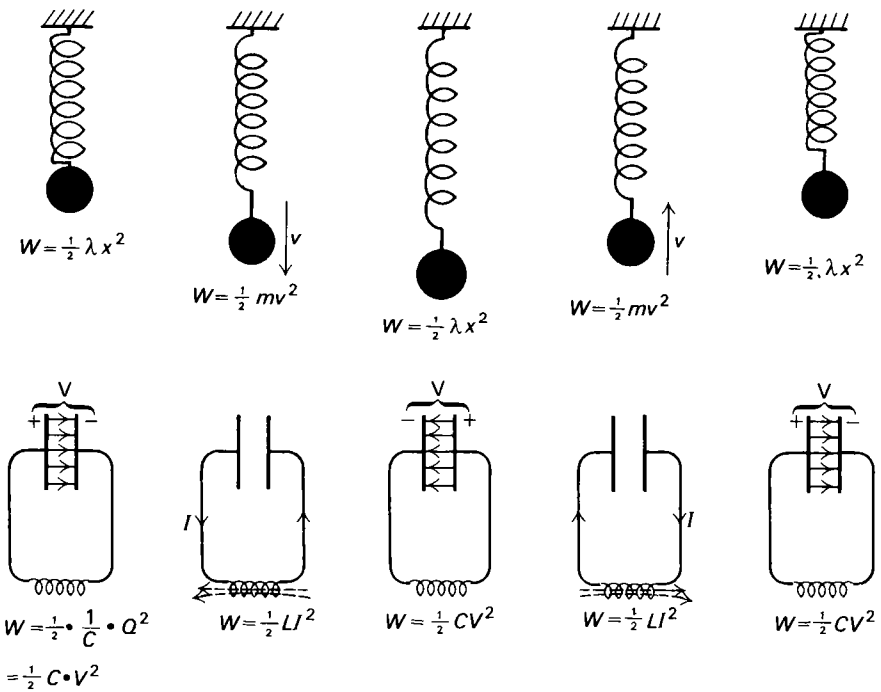


Fig. 8.7 The oscillating spring, and the oscillating circuit

Inductance L in the oscillating circuit corresponds to mass m, and the reciprocal, $1/C$, of capacitance to stiffness λ. A large mass and a weak spring both contribute towards making the mechanical oscillation slow, and the period of oscillation T is given by

$$T = 2\pi\sqrt{\frac{m}{\lambda}} \tag{8.17}$$

(assuming the mass of the spring itself to be negligible). Similarly, a large inductance L and a large capacitance C (i.e. small value for $1/C$) both contribute towards making the electrical oscillation slow, and the period of oscillation T is given by

$$T = 2\pi\sqrt{(LC)}. \tag{8.18}$$

We can have two circuits, both with the same period of oscillation. One has large C and small L, and the other has small C and large L. The first circuit

involves large swings in current, and the second involves large swings in the p.d. This is because the energy in the oscillation is given by

$$W = \tfrac{1}{2}CV_0^2 = \tfrac{1}{2}LI_0^2, \tag{8.19}$$

where V_0 is the maximum p.d. put across the capacitor, and I_0 is the maximum current which flows in the inductor.

[The mathematical derivation of equation 8.18 will be given at the end of this chapter.]

8.8 Rectification of alternating current – smoothing

Apart from their function as the components of the oscillating circuit, the practical importance of capacitors and inductors lies in their behaviour towards alternating current, and this will be the subject of the next chapter. One particular aspect which can conveniently be discussed at once is the problem of converting a.c. to d.c. The simplest means to achieve this is simply to insert a rectifying device, a thermionic or solid state diode, in series with the alternating source. Because the diode allows easy passage of current in one direction but offers very high resistance to it when the direction is reversed, the output current in this simple circuit varies in the manner shown in Fig. 8.8, and the process is described as unsmoothed, half-wave rectification.

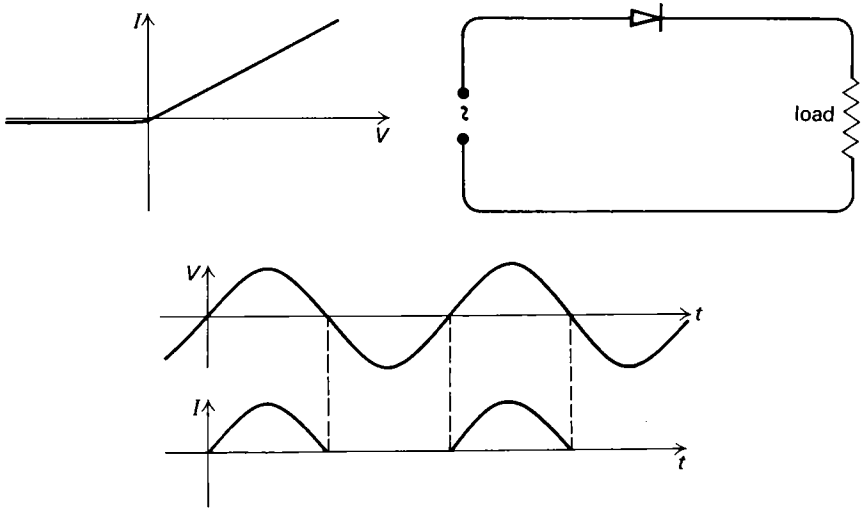


Fig. 8.8 Half-wave rectification (a) The characteristic of an ideal diode (b) A simple rectifying circuit (c) The applied alternating p.d. and its relation to current

This current is unidirectional, but it is by no means steady, and for most purposes it needs to be smoothed. This is where a capacitor comes in, and, more rarely, an inductor as well. The capacitor smooths the output p.d., and is analogous in its function to a spring, such as a car spring, which smooths out fluctuating forces. The inductor smooths the waveform of a rippling current, and is analogous to a flywheel, which prevents fluctuations in speed.

Smoothing by capacitance

A smoothing capacitor is connected across the output of the rectifying circuit, i.e. in parallel with the load, with the result shown in Fig. 8.9. The capacitor charges

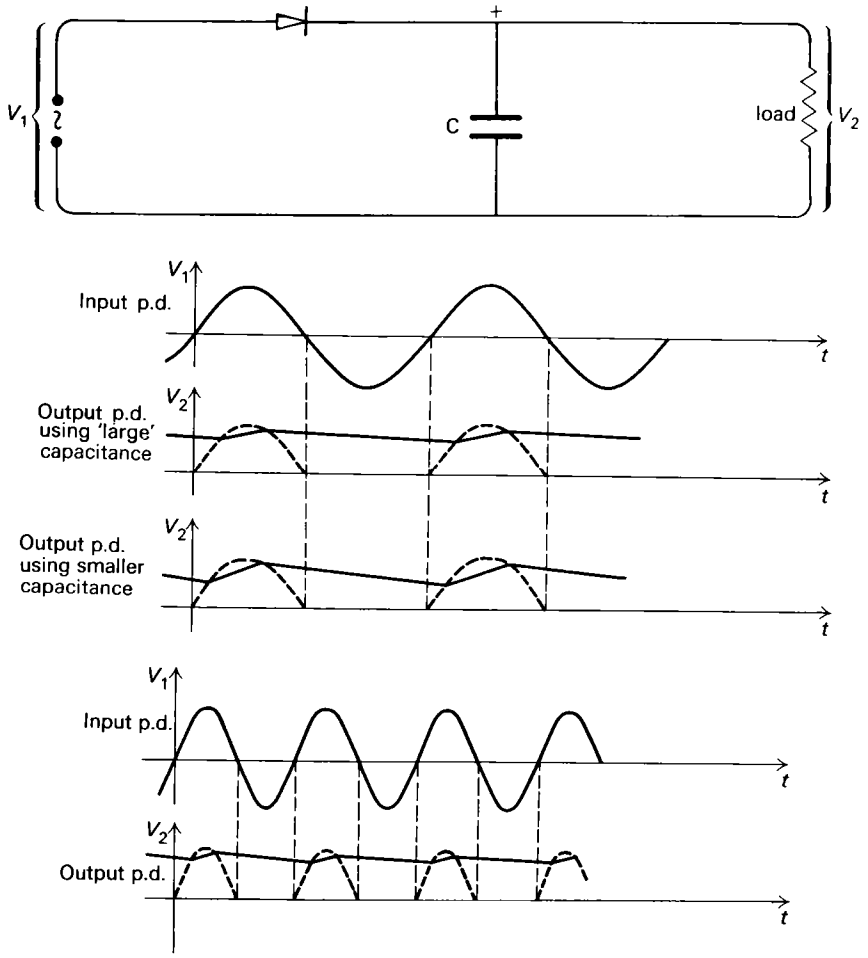


Fig. 8.9 The smoothing capacitor (a) The smoothing capacitor in circuit (b) Partially smoothed output curves, using frequency f (c) Smoothing the output, using frequency $2f$

up to very nearly the peak value of the alternating p.d. during the half cycle when the diode is conducting, and maintains a flow of current in the load during the other half cycle when the diode is cut off. The capacitor is discharging through the load during this time, with the result that the p.d. falls to some extent. The resulting variation in the output p.d., and consequently in the current, is shown in Fig. 8.9 (*c*). Because a small capacitance holds less charge, and therefore loses its p.d. more easily, the smoothing is more effective if a large value capacitor is used. It is also more effective if the a.c. frequency is high, because this allows less time for the capacitor to discharge before its charge is replenished through the rectifier.

An electrolytic capacitor is very suitable for smoothing, because it can easily be made to have large capacitance. The requirement to connect it with the correct polarity is satisfied, because the p.d. has been rectified, and its leakage resistance is unimportant, being equivalent to a slight alteration in the load.

The p.d. which must be tolerated by the diode in the smoothed rectifying circuit is equal to twice the peak value of the applied p.d. This is because the capacitor effectively maintains the potential of the output side at the steady value V_0. Meanwhile, the potential at the input side is alternating between the values $+V_0$ and $-V_0$, and the p.d. across the diode varies, in consequence, between zero and $2V_0$.

Smoothing by inductance

A smoothing inductor is connected in series with the load, and it reduces fluctuations in the current. For this reason, the inductor becomes increasingly effective as the current taken by the load is increased, which is complementary to the behaviour of a capacitor. Consequently the inductor is used to supplement the effect of a capacitor, when a heavy current is required. It is a heavy and bulky component, however, to include in a circuit, especially when designed to operate at frequencies as low as 50 Hz. This is because to be effective at low frequency its inductance must be high, and to carry large current its resistance must be low. It is now obsolescent, owing to the development of circuits involving special solid state diodes, which perform its function, on a different principle, and lack its inconvenience, and also because smoothing capacitors of very large capacitance are now available.

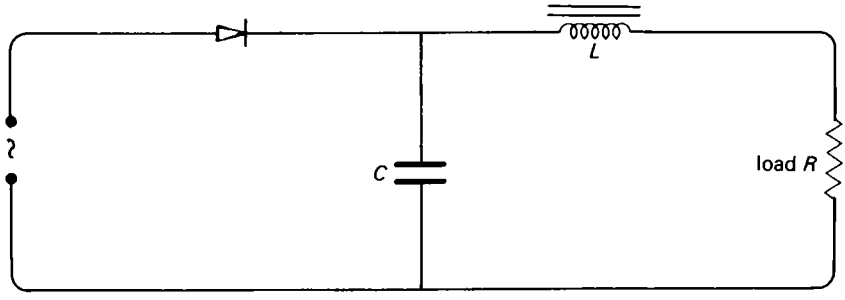


Fig. 8.10 A smoothing circuit, including capacitor and inductor

Neglecting the effect of resistance, there is no potential drop in the smoothing inductor due to a steady current, but any current fluctuations set up an induced e.m.f. in it. This means that a large proportion of any fluctuations in the applied p.d. is taken up by the inductor, and denied to the load connected in series with it, and the output current is smoothed in consequence.

Full-wave rectification

The task of a smoothing circuit is made easier if preceded by full-wave rectification, because the initial variation in p.d. is then not so pronounced. This involves using

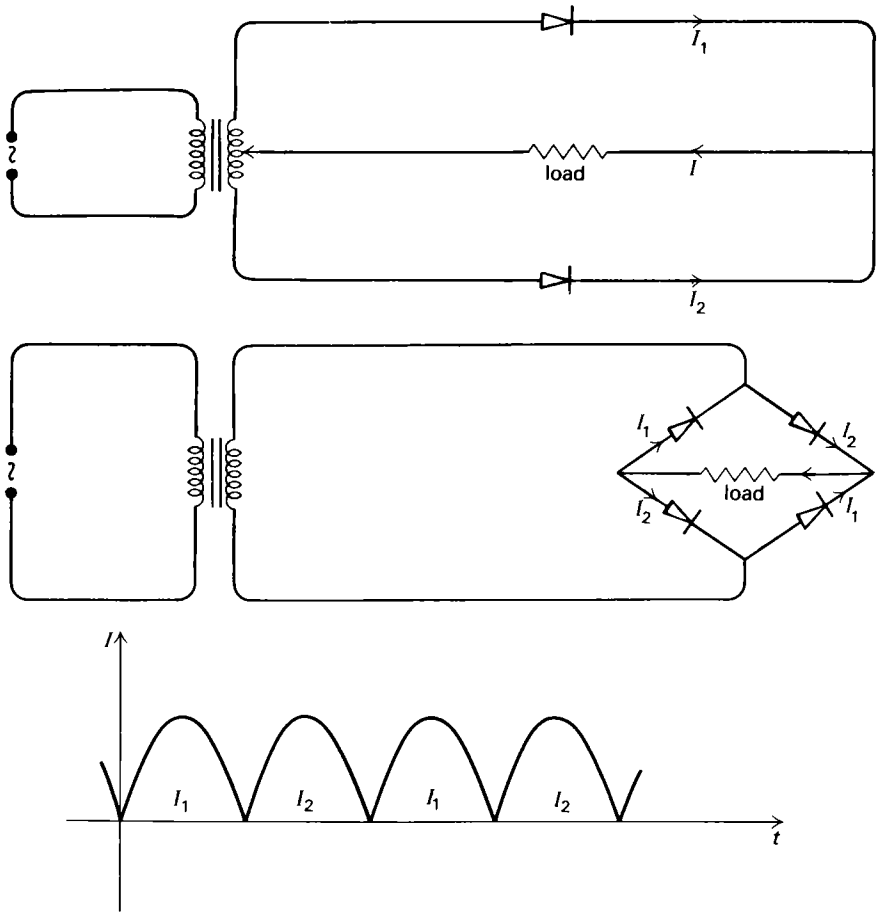


Fig. 8.11 Full wave rectification (a) A full wave rectifying circuit (b) A bridge rectifying circuit (c) The output current in the load. This graph refers to both circuits (a) and (b)

two diodes on a principle analogous to that of a double action pump; while one diode is cut off, current is flowing through the other. Alternatively, four diodes are used, in a bridge rectifying circuit.

Both these circuits are illustrated in Fig. 8.11, which is self explanatory.

Example

The input p.d. applied across a half-wave rectifying circuit has peak value 300 V *and frequency* 50 Hz. *The output p.d. is smoothed by a single capacitance of value* 50 μF. *If the current taken by the load is* 5 mA, *estimate the percentage fluctuation in the output p.d.*

As a fair approximation, the capacitor has almost the time of one cycle, 0.02 s, in which to discharge, before it starts to charge up a second time. If we guess, in view of the small value of the current, that the smoothing is fairly effective, we can

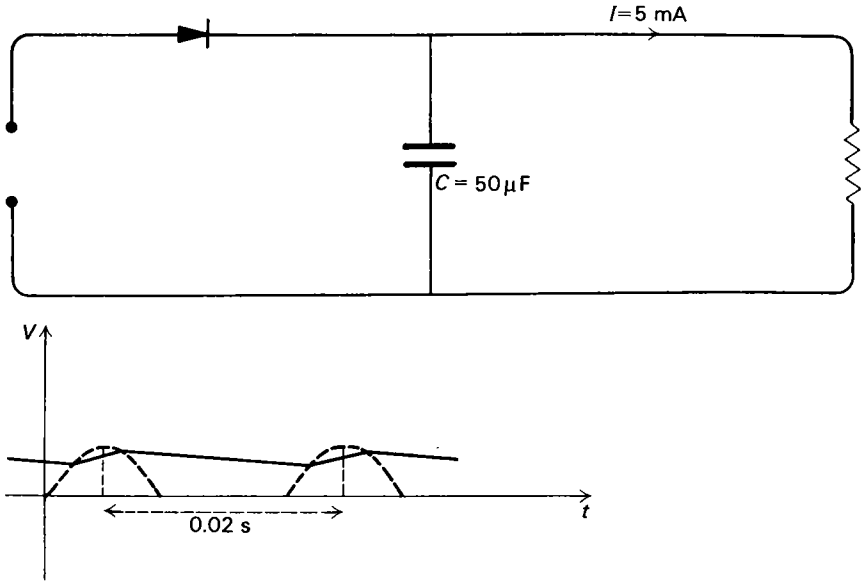


Fig. 8.12

take the discharging current to be reasonably constant, at 5 mA. On these assumptions, the charge taken from the capacitor in 0.02 s is

$$\begin{aligned}\Delta Q &= It \\ &= 5.10^{-3} \times 0.02 \text{ C} \\ &= 10^{-4} \text{ C.}\end{aligned}$$

The resulting fall in potential is ΔV, where

$$\begin{aligned}\Delta V &= \frac{\Delta Q}{C} \\ &= \frac{10^6}{50} \times 10^{-4} \text{ V} \\ &= 2 \text{ V.}\end{aligned}$$

Rather overestimating the average output p.d. V to be equal to the peak value of the input p.d., i.e. 300 V, the percentage fluctuation is approximately

$$100 \frac{\Delta V}{V} = \text{about } 0.7\%.$$

This example illustrates the circumstances in which the use of a smoothing capacitor is likely to be effective. A low voltage power supply, for example, required to provide a current of, say, 5 A at a p.d. of 30 V, would require a very much larger capacitance: to reduce the fluctuation in potential difference, in this case, to ± 2 V

would require the capacitance to be $50 \times 10^3\ \mu$F, and this would still represent a fluctuation of $\pm 7\%$.

8.9 Capacitors in series and in parallel

Capacitors in series

The capacitors shown in Fig. 8.13 are supposed to have been discharged before they were connected in the circuit. In this circumstance, there are certain parts

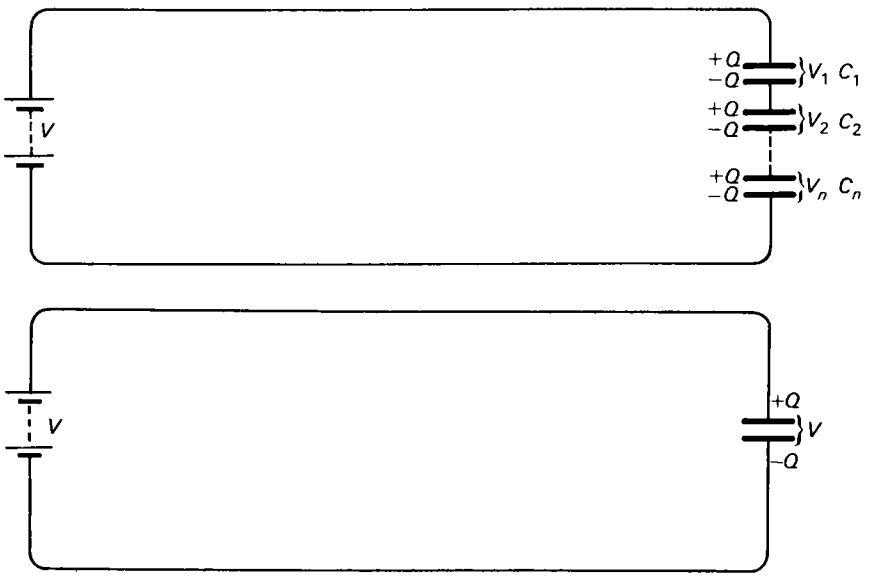


Fig. 8.13 (a) Capacitors in series (b) The equivalent single capacitor

of the circuit which must, overall, have zero charge, because there is no connection by which they can have gained any. On this principle, it follows that all the capacitors hold equal charge; if charge $+Q$ flows on to the top plate of C_1, charge $+Q$ must flow off the bottom plate of C_1 to the top plate of C_2, and so on.

If the p.d.s across $C_1, C_2, \ldots C_n$ are, respectively, $V_1, V_2, \ldots V_n$, we have:

$$V = V_1 + V_2 + \ldots V_n$$

$$= Q\left(\frac{1}{C_1} + \frac{1}{C_2} + \ldots \frac{1}{C_n}\right),$$

whereas, for the single equivalent capacitance C

$$V = \frac{Q}{C},$$

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \ldots \frac{1}{C_n}. \tag{8.20}$$

Capacitors in parallel

When capacitors are connected in parallel, the p.d. V is common to all, whereas the charge is shared between them, and we have

$$Q = Q_1 + Q_2 + \ldots Q_n$$
$$= V(C_1 + C_2 + \ldots C_n),$$

whereas, for the single equivalent capacitance C,

$$Q = VC,$$
$$C = C_1 + C_2 + \ldots C_n. \tag{8.21}$$

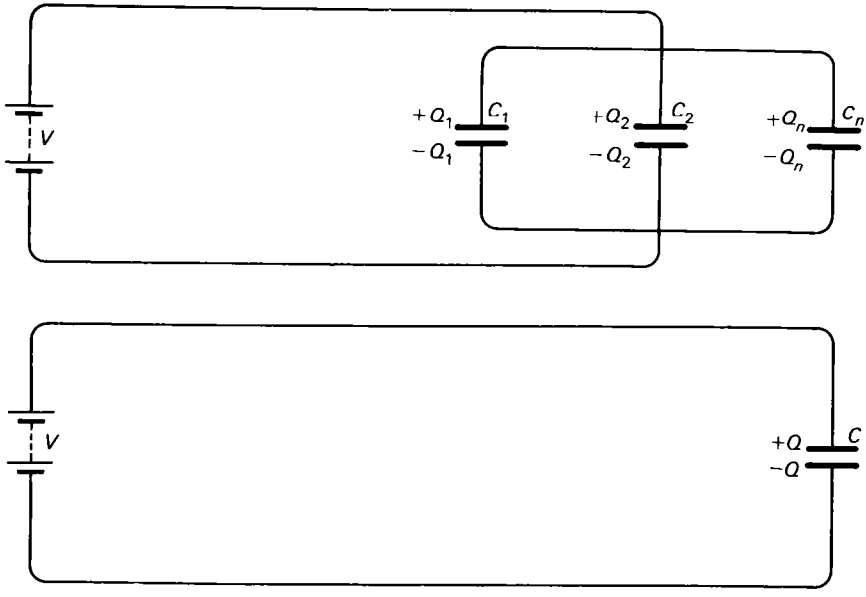


Fig. 8.14 (a) Capacitors in parallel (b) The equivalent single capacitor

A comparison with the similar arrangements of resistors is an obvious one to make here, and the equation for resistors in series corresponds to that for capacitors in parallel, and vice versa. This is because the definition of capacitance, $C = Q/V$, makes capacitance correspond to conductance, $G = I/V$, rather than to resistance $R = V/I$.

One example where this difference in form between resistance and capacitance equations shows up in practice is in the connection of resistors and capacitors in standard boxes. Whereas the resistors in a resistance box are made additive by connecting them in series, the capacitors in a capacitance box are connected in parallel.

An example which can easily trap the unwary involves calculating the distribution of p.d. applied across a chain of capacitors. In a resistance chain, the current is common to all, and the p.d. across a single resistor in the chain is therefore proportional to its resistance,

i.e.
$$V_1 = V \cdot \frac{R_1}{R_1 + R_2 \ldots R_n}, \tag{8.22}$$

and similarly for each resistor. In a capacitance chain, on the other hand, the charge Q is the same for all (assuming that the capacitors were all discharged initially) but the p.d. across a single capacitor in the chain is proportional, not to its capacitance, but to $1/C$.

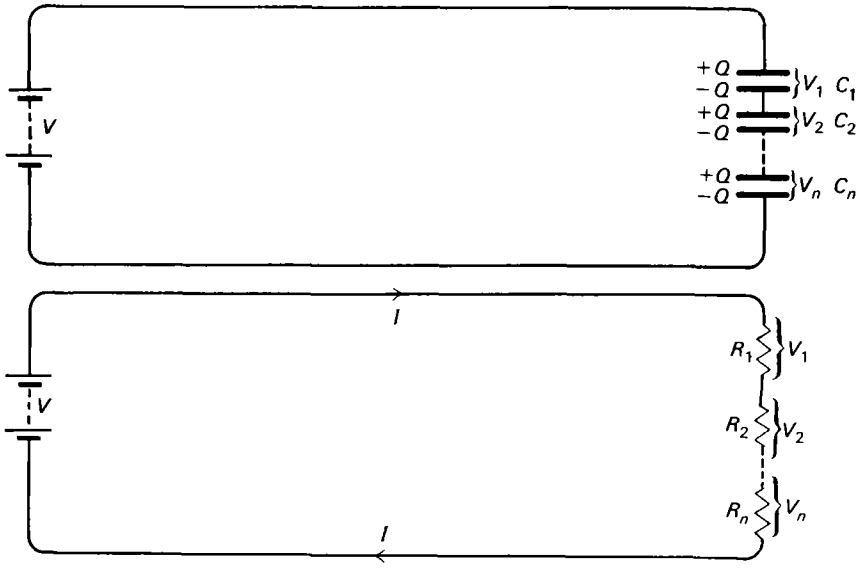


Fig. 8.15 (a) A capacitor chain $V \propto \frac{1}{C}$ (b) A resistor chain $V \propto R$

$$V_1 = V \cdot \frac{\frac{1}{C_1}}{\left(\frac{1}{C_1} + \frac{1}{C_2} + \cdots \frac{1}{C_n}\right)}, \tag{8.23}$$

and similarly for each capacitor; hence the largest capacitance takes the smallest share of the p.d.

Following the pattern set in this chapter, it would be logical to consider next the results of connecting inductors in series and parallel. The problem in this case, however, is not so simple as it is with capacitors, partly because inductors, in practice, have resistance as well as inductance, and we will omit it. What remains to be done in this chapter concerns the capacitor.

8.10 Examples

Example 1

The values of the components in the circuit of Fig. 8.16 *are:*

$$C_1 = 2\,\mu\text{F}$$
$$C_2 = 3\,\mu\text{F}$$

and the applied p.d. is $V = 100$ V.

When switch S *is initially open, deduce the values of the p.d.s across the separate capacitors* C_1, C_2, *and the charge stored in each.*

If switch S *is now closed, deduce the quantity of charge which flows in each of the resistors* R_1, R_2, *and trace the energy changes which take place.*

Initially, both capacitors (assuming that they held no charge before being placed in the circuit) now have equal charge Q because they are in series. When charge $+Q$ flows to the top plate of C_1, charge $+Q$ flows from its bottom plate to the top plate of C_2, and charge $+Q$ flows from the bottom plate of C_2.

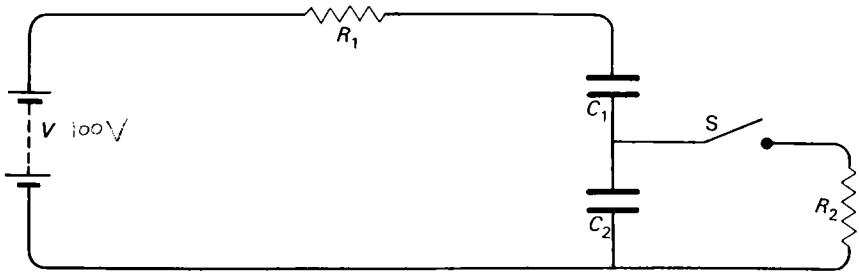


Fig. 8.16

Hence $$V_1 = \frac{Q}{C_1},\ V_2 = \frac{Q}{C_2},$$

and $$V = V_1 + V_2 = 100 \text{ V}.$$

Therefore $$V_1 = V\left(\frac{1/C_1}{1/C_1 + 1/C_2}\right) = 100 \times \frac{\frac{1}{2}}{(\frac{1}{2}+\frac{1}{3})} V = 60 \text{ V},$$

and $$V_2 = V - V_1 = 40 \text{ V}.$$

Whence $$Q = C_1 V_1 = 2 \times 60\ \mu\text{C} = 120\ \mu\text{C}.$$

When the switch S *is closed:*
Current flows in R_2 discharging the capacitor C_2, and the full p.d., 100 V, appears across C_1. The new charge Q' in C_1 is given by

$$Q' = 100\ C_1$$
$$= 200\ \mu\text{C}.$$

The effective circuit has, in fact, been changed from that drawn in Fig. 8.17(*a*) to that shown in Fig. 8.17(*b*).

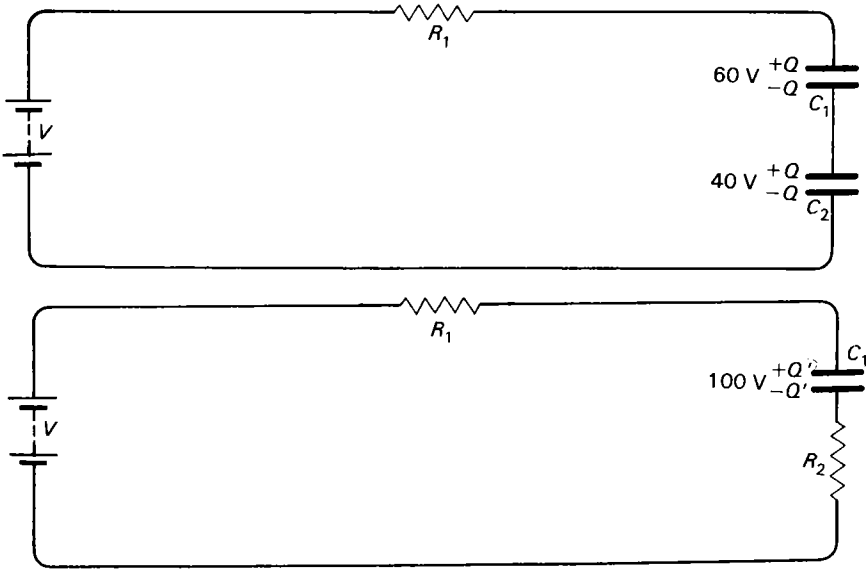


Fig. 8.17 (a) Switch open (b) Switch closed

Thus, the charge which has flowed through the battery from the bottom plate of C_2 to the top plate of C_1 to increase the p.d. across C_1 is

$$Q' - Q = 80\,\mu\text{C}.$$

At the same time, a larger amount of charge flowed through resistor R_2, for the bottom plate of C_1 and the top plate of C_2 initially held zero charge between them, whereas the bottom plate of C_1 finally holds charge $-Q'$, while C_2 holds none.

Therefore, the charge which flows in $R_2 = Q' = 200\,\mu\text{C}$.

With regard to the energy:

Initially, the energy stored in capacitors C_1, C_2 was W_1, W_2, where

$$\begin{aligned} W_1 &= \tfrac{1}{2}C_1V_1^2 \\ &= \tfrac{1}{2}\times 2\times 10^{-6}\times 3600\ \text{J} \\ &= 3.6\times 10^{-3}\ \text{J}, \end{aligned}$$

$$\begin{aligned} W_2 &= \tfrac{1}{2}C_2V_2^2 \\ &= \tfrac{1}{2}\times 3\times 10^{-6}\times 1600\ \text{J} \\ &= 2.4\times 10^{-3}\ \text{J}. \end{aligned}$$

Finally, there is no energy in C_2, and the energy stored in C_1 is W_1', where

$$\begin{aligned} W_1' &= \tfrac{1}{2}\times 2\times 10^{-6}\times 10^4\ \text{J} \\ &= 10\times 10^{-3}\ \text{J}. \end{aligned}$$

Now, in the process of charge redistribution, we have worked out that charge $(Q' - Q) = 80\,\mu\text{C}$ flowed through the battery. This took energy W_B from the battery, where

$$\begin{aligned} W_B &= (Q' - Q)\times V \\ &= 80\times 10^{-6}\times 100\ \text{J} \\ &= 8\times 10^{-3}\ \text{J}. \end{aligned}$$

Thus, the net gain of energy by the capacitors was

$$W_1' - (W_1 + W_2) = 4\times 10^{-3}\ \text{J},$$

which is less than the energy taken out of the battery. This is because the balance of the energy was dissipated as heat in the two resistors when the charge flowed in them.

Example 2

The three capacitors C_1, C_2, C_3 shown in Fig. 8.18 *have capacitance respectively* 1 μF, 2 μF, *and* 3 μF. *All three are initially charged to a potential difference of* 100 V, *the plates a, c, e being positive. The capacitors are now disconnected from the source and from each other, taking care that they do not lose their charge, and then reconnected in the manner shown in Fig.* 8.18(*b*), *with plates b and c, d and e, f and a, connected together in pairs.*

Deduce the p.d. across each capacitor.

This problem basically concerns conservation of charge. Initially, the charge held by each capacitor is given by:

$$\begin{aligned} Q_1 &= C_1V, \\ Q_2 &= C_2V, \\ Q_3 &= C_3V, \end{aligned}$$

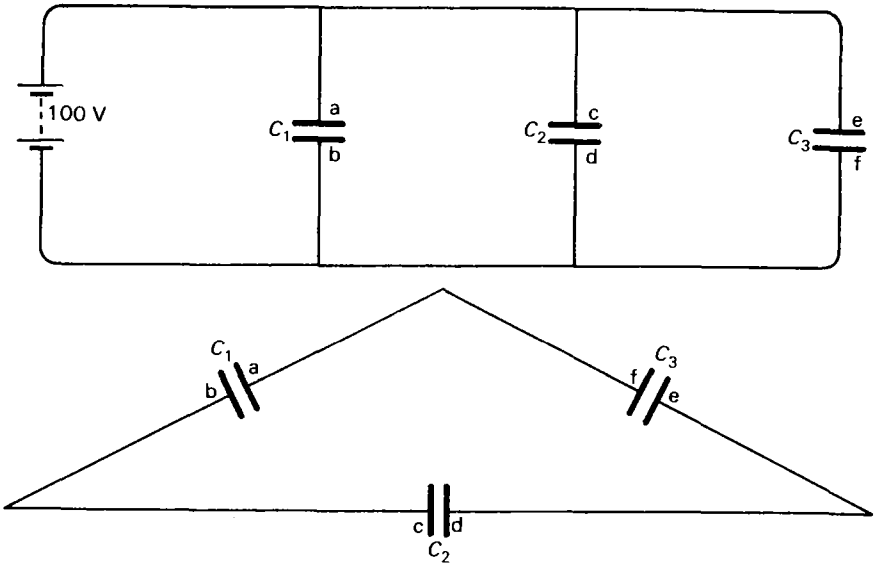


Fig. 8.18 (a) (b)

where $V = 100$ V. When the three capacitors are then joined together in the manner specified in the question, the result can only be that an amount of charge, Q, flows in each connection. Assuming that this flows from plate b of capacitor C_1 to plate c of capacitor C_2, from plate d of C_2 to plate e of C_3, and so on, the new values of the charge on each capacitor become:

$$Q_1' = Q_1 + Q,$$
$$Q_2' = Q_2 + Q,$$
$$Q_3' = Q_3 + Q,$$

whence the corresponding p.d.s are:

$$V_1' = \frac{Q_1'}{C_1},$$

$$V_2' = \frac{Q_2'}{C_2},$$

$$V_3' = \frac{Q_3'}{C_3}.$$

Now, applying Kirchhoff's second law to the circuit,

$$V_1' + V_2' + V_3' = 0,$$

whence

$$\frac{Q_1 + Q}{C_1} + \frac{Q_2 + Q}{C_2} + \frac{Q_3 + Q}{C_3} = 0,$$

$$\therefore \quad Q\left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}\right) = -\left(\frac{Q_1}{C_1} + \frac{Q_2}{C_2} + \frac{Q_3}{C_3}\right)$$

$$= -3 \text{ V}.$$

Putting in the values specified in the question, this gives

$$Q\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{3}\right) = -300$$

where Q is in microcoulombs,

$$Q = -\frac{1800}{11}\,\mu\text{C},$$

and from this, the three p.d.s can be deduced, viz:

$$V_1' = -\frac{700}{11}\text{ V},$$

$$V_2' = +\frac{200}{11}\text{ V},$$

$$V_3' = +\frac{500}{11}\text{ V}.$$

8.11 The measurement of capacitance by ballistic galvanometer

The direct way to measure capacitance would be to charge the capacitor to a known p.d., and then discharge it through a form of microammeter which is capable of measuring electric charge. The capacitance could then be calculated directly, using the equation

$$C = \frac{Q}{V}. \tag{8.1}$$

An ordinary microammeter can be used for this purpose, although it would, for preference, have a certain special feature, as explained below.
The current which flows through the meter is only transient, and this produces a deflecting impulse, not a steady deflecting torque. In this circumstance, the

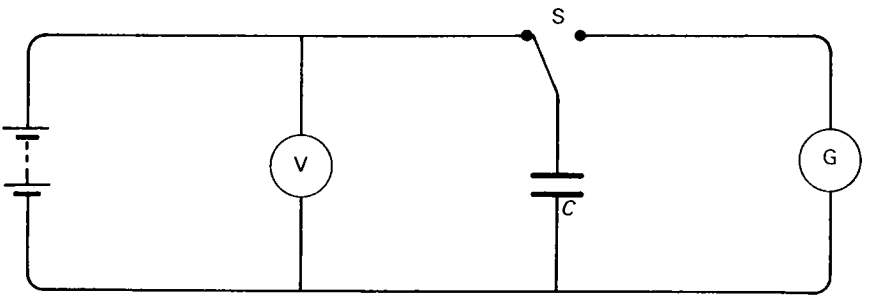


Fig. 8.19 The measurement of capacitance by ballistic galvanometer

mechanical response is similar to that of a ballistic pendulum, and a momentary deflection, or throw, θ, is produced, in place of the steady deflection which would result from a continuous current.

If a mechanical system produces a deflection which is proportional to a steady deflecting force F, then the throw which results from using it ballistically is proportional to the deflecting impulse $\int F dt$.

Similarly, if the steady deflection of a microammeter is proportional to the current I, then the throw, when it is used ballistically, is proportional to the charge Q discharged through it,

i.e.
$$\theta \propto \int I dt,$$
$$\therefore \quad \theta \propto Q. \tag{8.24}$$

This equation is valid, however, only if the discharge is effectively completed before the meter has had time to respond. An instrument which is specifically designed for this application is therefore made to be unusually slow in its movement. It is usually referred to as a ballistic galvanometer. A more detailed description is given in chapter 18.3.

This method is better suited to the comparison of two capacitances, the capacitors being inserted in turn into the circuit. In this case, the voltmeter may be dispensed with and the ballistic meter need not be calibrated, provided that it is known to have a linear scale, for we have

$$\frac{\theta_2}{\theta_1} = \frac{Q_2}{Q_1},$$

where
$$Q_2 = C_2 V, \; Q_1 = C_1 V,$$
$$\frac{\theta_2}{\theta_1} = \frac{C_2}{C_1}. \tag{8.25}$$

There is no need, either, to apply a correction to allow for damping of the meter movement, because in this comparison equation such a damping correction would cancel out.

This method is available only if the capacitors store sufficient charge to provide a reasonably large deflection. Unless the meter sensitivity is increased by using a d.c. amplifier, the capacitances must be greater than, say, $10^{-2}\,\mu$F.

8.12 Measurement of capacitance by Maxwell's method

A better method of measuring capacitance directly was devised by Maxwell, and this can be used even when the value of the capacitance is very small.

The switch S is a vibrating contact, e.g. a reed switch, or a rotating switch, which completes its two way cycle with a constant frequency f. The capacitor C is repeatedly charged up and then discharged through a microammeter, and the mean current I which flows through the meter is given by

$$I = fQ$$
$$= fCV. \tag{8.26}$$

An alternative circuit is given in the second diagram. The equation is the same, the only difference being that the meter measures the mean current which charges the capacitor, instead of the discharge. The resistor R_1 serves as a protection for the meter, and its value has no relevance to the calculation; it must not be too large, however, since its presence slows down the discharge time of the capacitor, and this must be sufficiently short to allow the discharge to be complete.

With reference to equation 8.26 and the circuit of Fig. 8.20(*b*), it can be seen that the combination of capacitor and discharging circuit is equivalent to a resistor of value R, where

$$R = \frac{V}{I},$$
$$\therefore \quad R = \frac{1}{fC}. \tag{8.27}$$

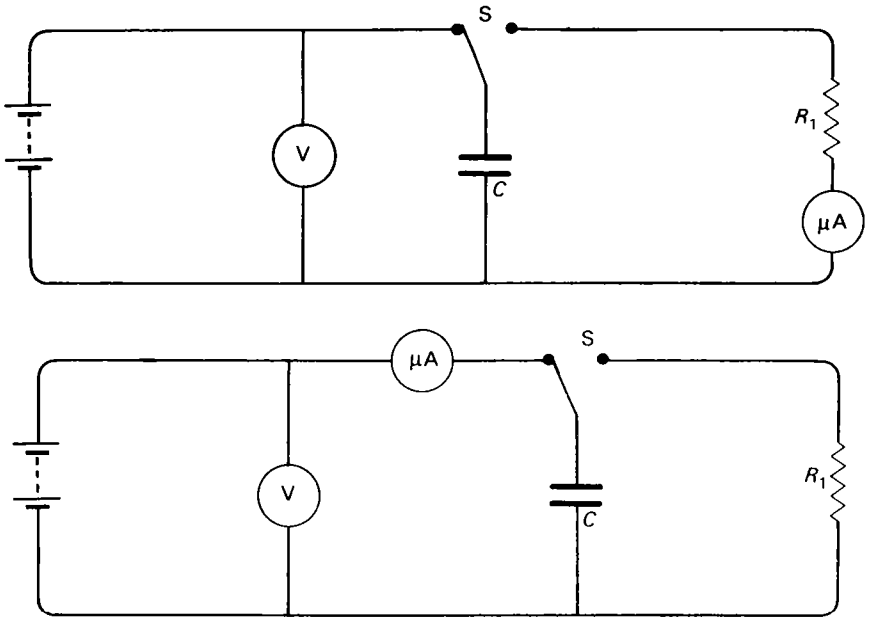


Fig. 8.20 The measurement of capacitance by Maxwell's method (a) (b)

It follows that the value of the capacitance may be determined by treating this part of the circuit as a resistance. A high value resistance box, for example, may be put in series with the meter instead of the capacitor and its discharging circuit, and adjusted to draw the same current as the discharging capacitor did. This gives the value, R, of the equivalent resistance.

Alternatively, a bridge circuit may be used, which was devised by Maxwell. Referring to Fig. 8.21, when the bridge circuit is balanced, we have

$$\frac{1/fC}{R_0} = \frac{R_1}{R_2}. \tag{8.28}$$

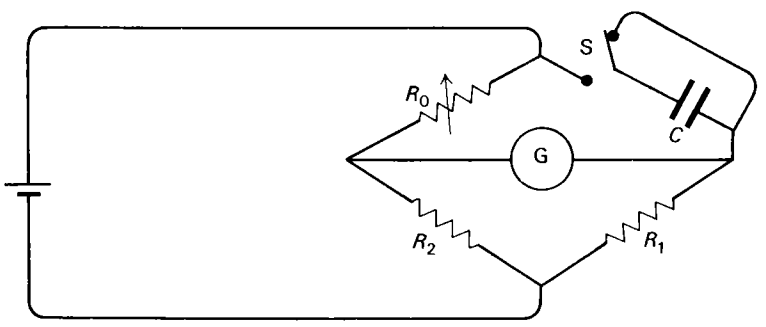


Fig. 8.21 The measurement of capacitance. Maxwell's bridge circuit

8.13 Comparison of capacitances by electrometer

This method is especially suitable for small value capacitors. Whereas by the ballistic galvanometer method the two capacitors are charged in turn to the same

p.d. and their charges are compared, the basic principle this time is to give the capacitors equal amounts of charge and compare the p.d.s across them. In practice, the value of one capacitor is compared with its altered value when connected in parallel with the other one. It is essential to use an electrostatic voltmeter, i.e. an electrometer, to measure the p.d., because any other type of instrument would simply discharge the capacitor.

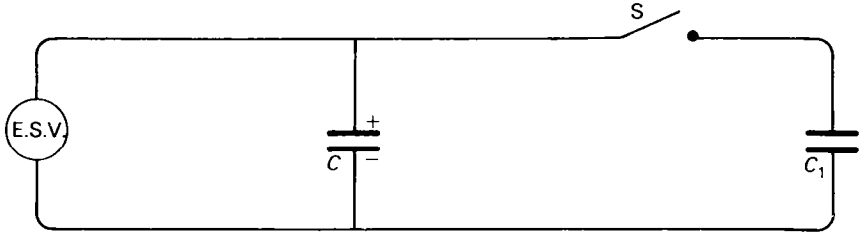


Fig. 8.22 The comparison of capacitances by sharing charge

Suppose capacitor C is first charged to a p.d. V, and is then connected in parallel with capacitor C_1, after being disconnected from the source. The p.d. will now fall to a lower value, V_1, where

$$Q = CV$$
$$= (C + C_1)V_1,$$

whence

$$\frac{C_1}{C} = \frac{(V - V_1)}{V_1}. \tag{8.29}$$

The value of capacitance C in this equation includes the capacitance of the electrometer. This may be eliminated by placing the capacitors under test in the position of C_1, in turn, and performing the experiment twice. In this case, we would have

$$\frac{C_2}{C_1} = \left(\frac{V' - V_2}{V_2}\right) \Big/ \left(\frac{V - V_1}{V_1}\right). \tag{8.30}$$

8.14 A capacitance bridge

The most versatile method for comparing capacitances, and the easiest to use in practice, is to set up a bridge circuit, using the two capacitors as two arms of the bridge and resistors as the other two arms.

Suppose a d.c. source is connected across such a bridge arrangement, as in Fig. 8.23. The two capacitors, being in series, hold equal charges (assuming them to have been discharged in the first place), and the two resistors carry the same current. Hence we have:

$$V_{AB} = \frac{Q}{C_1},$$
$$V_{BD} = \frac{Q}{C_2},$$
$$V_{AE} = R_1 I,$$
$$V_{ED} = R_2 I.$$

If the bridge is balanced, i.e. if the p.d. across BE is zero, it follows that

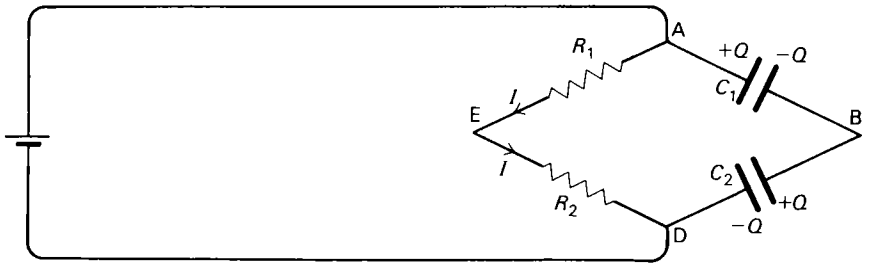


Fig. 8. 23 A Capacitance Bridge

$$V_{AB} = V_{AE}$$

and

$$V_{BD} = V_{ED},$$

whence

$$\frac{C_2}{C_1} = \frac{R_1}{R_2}. \tag{8.31}$$

The inverse ratio C_2/C_1 compares with the ratio R_1/R_2, which is typical of the functions of capacitance and resistance in a chain.

This result cannot easily be applied in practice using a d.c. source, because a galvanometer connected between the points B and E to detect the balance would disturb the charge distribution on the capacitors. The simplest practical application is to use an audio-frequency alternating source instead, having frequency of, say, 2000 Hz, and to put an a.c. detector across BE to detect the balance. A headphone is often used, in which case the resistance ratio is adjusted to give minimum sound.

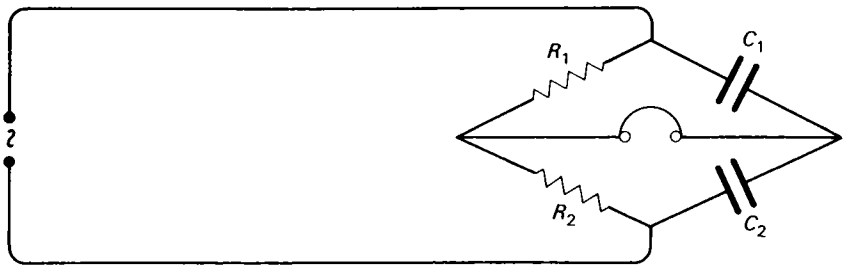


Fig. 8.24 A practical form of capacitance bridge

The a.c. bridge will be further discussed in section 10.20, using a.c. theory to derive equation 8.31.

This method is available over a wide range of capacitance values, although it becomes more difficult if these are very small. Complications also arise if the capacitors have leakage resistance.

Comparison of inductances

It would be following a logical sequence to discuss next the measurement of inductance. This is a problem, however, which has special difficulties about it, due to the fact that, in practice, an inductor inevitably has resistance as well. For

this reason, inductance measurement is outside the scope of this book, beyond observing that bridge circuits can be designed for the purpose.

8.15 The measurement of relative permittivity ϵ_r

Any of the above methods of comparing capacitances may be used to measure the relative permittivity ϵ_r of a dielectric. This was defined in section 6.19 by the equation

$$\epsilon_r = \frac{C}{C_0}, \tag{8.32}$$

where C_0 is the capacitance of any vacuum-spaced capacitor (and in practice air makes no appreciable difference) and C is its capacitance provided that the space taken up by the electric field is filled entirely with the dielectric.

In the capacitance bridge described in the last paragraph, for example, the capacitor C_2 might be an air-spaced capacitor, which is compared with C_1. The dielectric is then inserted to fill the space between the plates of C_2, and its capacitance is compared with that of C_1 a second time.

If the dielectric is a liquid, or can be melted for the purpose of insertion, a capacitor made from two concentric spheres would be very suitable, for the reason that the entire space can be filled, with no possibility of any lines of force escaping the dielectric (provided that it is used with the outer sphere earthed), This arrangement was used for this purpose by Faraday.

*8.16 The rate of charge and discharge of a capacitor – mathematical treatment

This chapter opened with a qualitative discussion of the manner in which the current grows in a circuit which includes a capacitor or an inductor. We will now consider these cases mathematically.

A discharging capacitor

Suppose that the switch connects the resistor R across the charged capacitor C at the instant $t = 0$. As the capacitor discharges, the p.d. which it puts across the resistor diminishes, and the rate of discharge becomes progressively slower.

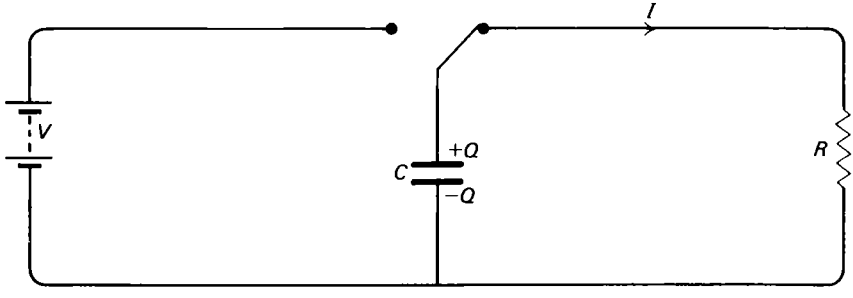


Fig. 8.25

Suppose that, at time t, the p.d. across the capacitor is V. We have

$$V = \frac{Q}{C}$$

and

$$V = RI,$$

where

$$I = \frac{dQ}{dt},$$

the minus sign signifying discharge.

$$-R\frac{dQ}{dt} = \frac{Q}{C}.$$

Integrating, this gives:

$$\left.\begin{aligned} Q &= Q_0\exp\left(\frac{-t}{RC}\right), \\ V &= V_0\exp\left(\frac{-t}{RC}\right), \\ I &= I_0\exp\left(\frac{-t}{RC}\right). \end{aligned}\right\} \tag{8.33}$$

where Q, V, I are the initial values.

The charging capacitor

In the case of the charging capacitor, the battery is connected in series throughout and the current I at the instant t is

$$I = \frac{dQ}{dt}.$$

Applying Kirchhoff's second law to the circuit, we have

$$V_0 = V_1 + V_2,$$

where $$V_1 = RI$$

and $$V_2 = \frac{Q}{C},$$

whence $$V = \frac{RdQ}{dt} + \frac{Q}{C}.$$

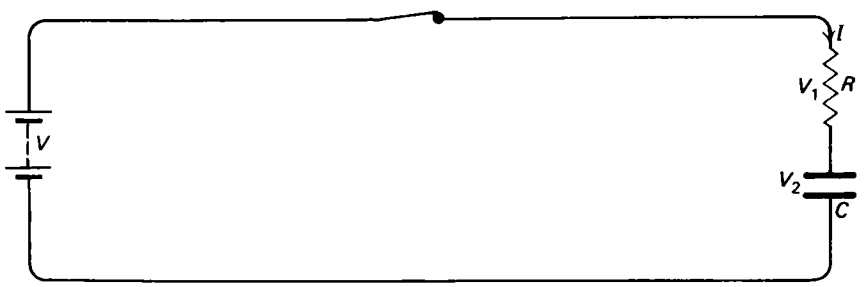


Fig. 8.26 (a)

This equation solves to give

$$Q = CV_0\left[1 - \exp\left(\frac{-t}{RC}\right)\right],$$

i.e.

$$\left.\begin{aligned} V &= V_0\left[1 - \exp\left(\frac{-t}{RC}\right)\right], \\ Q &= Q_0\left[1 - \exp\left(\frac{-t}{RC}\right)\right], \end{aligned}\right\} \tag{8.34}$$

and
$$I = I_0 \exp\left(\frac{-t}{RC}\right),$$

where V_0, Q_0 are the p.d. and charge of the fully charged capacitor, and $I_0 \quad V_0/R$ is the initial charging current.

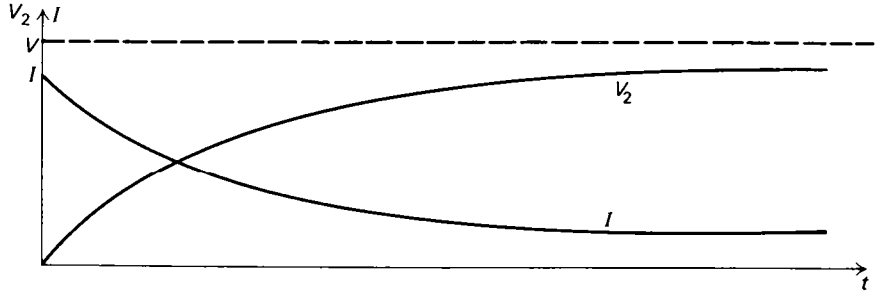


Fig. 8.26 (b)

The time constant

Putting $t = RC$ in equation 8.33 or 8.34 gives

$$I = I_0 e^{-1},$$

where
$$e^{-1} = 0.37.$$

This value of t is called the time constant of the circuit. It is the time which elapses before the discharging or charging current has fallen to 0.37 times its initial value. In the same time, the p.d. across the discharging capacitor falls to 0.37 times its initial value, and that across the charging capacitor has risen from zero to 0.63 times its final value.

*8.17 Differentiating and integrating circuits

Referring to the circuit of the charging capacitor (Fig. 8.26), we note that, at any instant during the charging process, the p.d. V_1, across the resistor, and the p.d. V_2, across the capacitor, are given by

$$V_1 = RI,$$

$$V_2 = \frac{Q}{C},$$

where
$$I = \frac{dQ}{dt}.$$

Hence, we have

$$V_1 = RC\frac{dV_2}{dt} \tag{8.35}$$

and
$$V_2 = \frac{1}{RC}\int V_1 dt. \tag{8.36}$$

Thus, if it were possible to have a situation in which the current in the resistor and the p.d. V_1 across it were the direct consequence of applying the p.d. V_2 across the capacitor, we would have a differentiating circuit; a varying p.d. V_2 applied across

the capacitor would produce, as an output signal, a p.d. across the resistor, and this would be proportional to dV_2/dt.

Similarly, if it were possible to have a situation in which the p.d. V_2 appeared across the capacitor as the direct consequence of applying the p.d. V_1 across the resistor, we would have an integrating circuit. The input signal would be the varying p.d. V_1 applied across the resistor, and the output, taken from across the capacitor, would be proportional to $\int V_1 dt$.

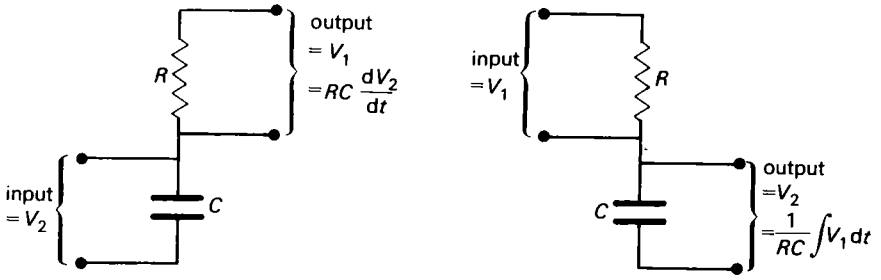


Fig. 8.27 The basis of differentiating and integrating circuits (a) Differentiating (b) Integrating

These effects cannot, of course, be obtained (except approximately) by having merely the resistor and capacitor in series; it would require the input p.d. V to be applied across the pair of components to cause a p.d. to appear across either one of them. They can, however, be produced with the aid of a high gain amplifying circuit, the technique of which is beyond the scope of this book. This is what is done in an analogue computor.

*8.18 The growth of current in an inductor – mathematical treatment

The treatment of the inductor is similar to that of the charging capacitor. In the circuit of Fig. 8.28, the resistor R may be taken to include the resistance of the inductor, the p.d. V_2 being the p.d. taken up in overcoming the back e.m.f. in the inductance.

We have $$V_0 = V_1 + V_2$$

where $$V_1 = RI$$

$$V_2 = L\frac{dI}{dt}$$

whence $$V_0 = RI + L\frac{dI}{dt}$$

which can be integrated to give

$$I = I_0\left[1 - \exp\left(-\frac{R}{L}t\right)\right], \qquad (8.37)$$

where $I_0 = V_0/R$ is the final steady value reached by the current.

The time constant

The time constant of this circuit is $t = L/R$, and this has an interpretation similar to that of the time constant $t = RC$ of the capacitor circuit.

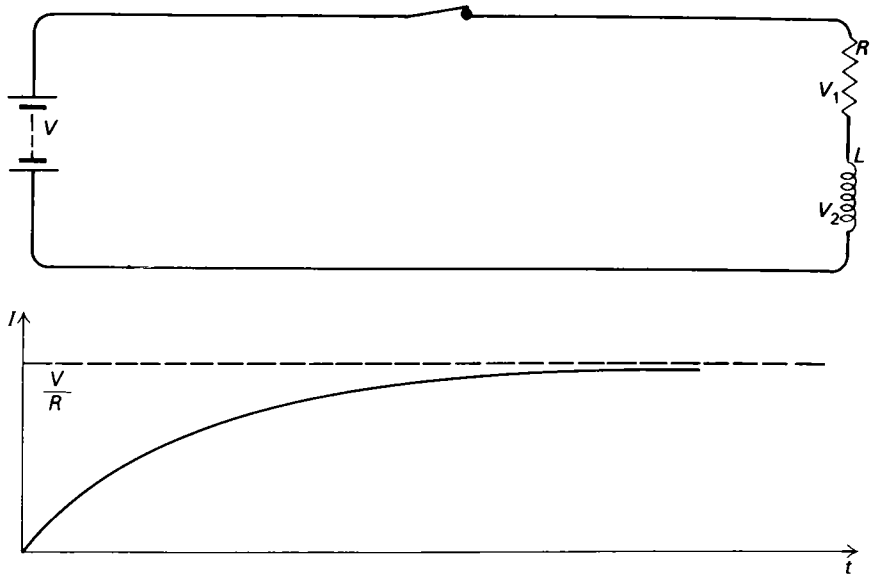


Fig. 8.28

It will be noticed that the current in an inductor reaches its steady maximum value more slowly if R is small. This is because in this case the maximum value which it has to reach is high, whereas the initial rate of growth depends only on the value of L.

*8.19 The oscillating circuit – mathematical treatment

In the ideal case, the inductor connected across the charged capacitor in this circuit has zero resistance. Taking the current to be positive when its direction is that shown in the diagram, we have

$$L\frac{dI}{dt}+\frac{Q}{C}=0$$

where

$$I=\frac{dQ}{dt}$$

$$\therefore \quad \frac{Ld^2Q}{dt^2}+\frac{Q}{C}=0, \tag{8.38}$$

which is the differential equation of an undamped sinusoidal oscillation

$$Q=Q_0\sin 2\pi ft. \tag{8.39}$$

The frequency of the oscillation is f, where

$$f=\frac{1}{2\pi\sqrt{(LC)}}, \tag{8.40}$$

and the period is

$$T=2\pi\sqrt{(LC)}. \tag{8.41}$$

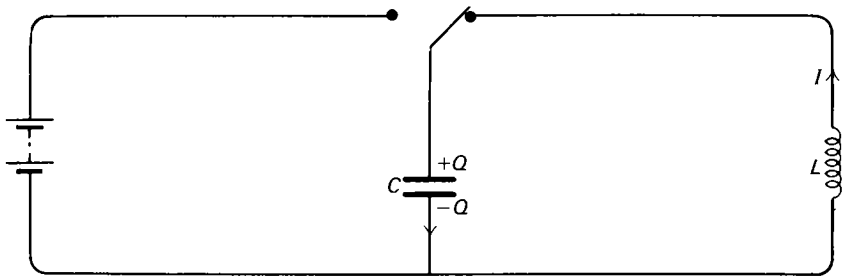


Fig. 8.29

In the practical case, of course, the inductor would have some resistance, and the oscillations are damped.

8.20 The discharging and charging capacitor – an application to the measurement of a very high resistance

It is a very stringent test of an insulator to see whether it is capable of discharging a charged capacitor, the reason being that the amount of charge held is so small that a very slight current is sufficient to remove it. An example of this is the elementary demonstration that such substances as wood and cotton are sufficiently good conductors to discharge a charged electroscope.

A calibrated electroscope, or any electrostatic voltmeter, may be used on this principle to measure very high resistances, of the order, say, of $10^{12}\ \Omega$. A suitable circuit is shown in Fig. 8.30, where a capacitor is charged up from a source put in series with the resistor.

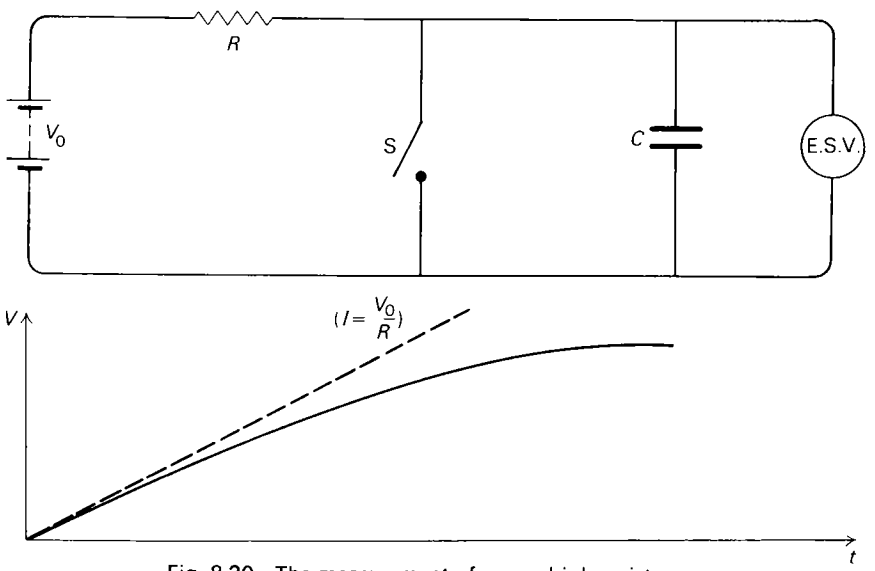


Fig. 8.30 The measurement of a very high resistance

Suppose the short circuiting switch S is initially closed, thus short circuiting the capacitor. As soon as the switch is opened, the capacitor starts to charge up

through the resistor. If the value C selected for the capacitance is appropriate, the p.d. V across the voltmeter will rise slowly as the capacitor charges up, and its growth can be timed.

Initially, we have

$$I = \frac{V_0}{R},$$

where

$$I = \frac{dQ}{dt}$$

$$= C\frac{dV}{dt},$$

whence

$$R = \frac{V_0}{C\frac{dV}{dt}},$$

V_0 being the p.d. applied from the source. The value inserted for the capacitance C must include the capacitance of the electrostatic voltmeter.

Thus, the experiment consists of measuring the initial rate at which the p.d. increases.

This method is available for a wide range of high resistances because the value of the capacitance may be selected to suit the circumstances. It is best suited for the comparison of two resistances, in which case the values of V_0 and C need not be determined.

8.21 The pulse electrometer

The pulse electrometer is a form of leaf electroscope adapted to make it particularly suitable for experiments of the type just described, i.e. for the comparison of very small currents. The principle is to add an electrode to the electroscope, called the counter electrode, which can be adjusted in position to make connection with the leaf at a pre-selected potential. This might be done, for example, with a gold leaf electroscope, as shown in Fig. 8.31(*a*).

One practical form of pulse electrometer is shown in the second diagram, in which the gold leaf is replaced by aluminium foil, and gravity control of the leaf is replaced by using the restoring force of a very fine quartz fibre, attached to the lower end of the foil.

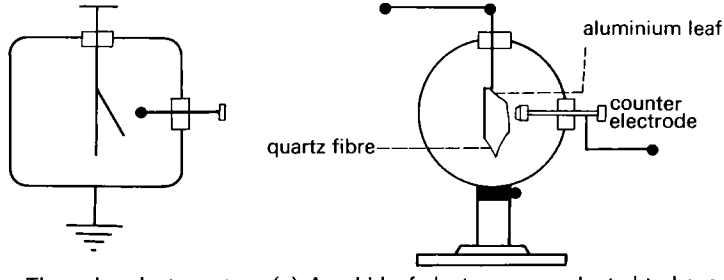


Fig. 8.31 The pulse electrometer (a) A gold leaf electroscope, adapted to have pulsing property (b) A pulse electrometer

If the pulse electroscope is connected in series with an E.H.T. source and a very high value resistor, its leaf will start 'pulsing'. The leaf repeatedly rises, as the current through the resistor starts to charge it up, until it makes contact with the

counter electrode; the leaf then returns to the zero position, and the cycle of operations starts a second time. The pulse rate is proportional to the current, and so the electrometer may be used to compare currents of very small size, down to, say, 10^{-10}A. There is a disadvantage, however, as it requires the P.D. from an E.H.T. unit to deflect the leaf at all. The transistor has now made the instrument obsolescent in its previous home, the school laboratory; it is easier to use a D.C.amplifier, which amplifies the small current before measuring it on a conventional meter (section) 18.6.

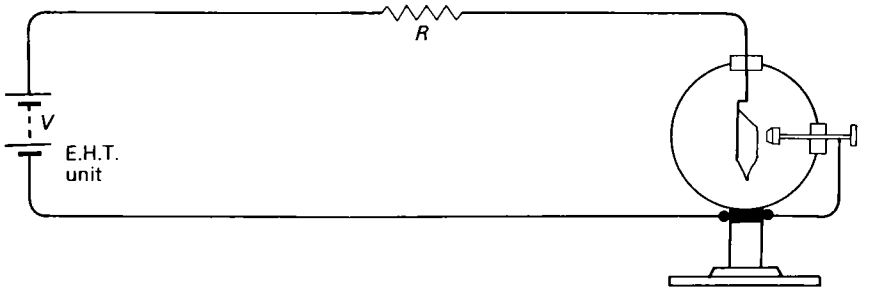


Fig. 8.32 The pulse electrometer; the comparison of two high value resistors

8.22 The flashing neon lamp

A neon lamp with a capacitor connected across it may be used to perform a function which is similar to that of the pulse electrometer, except that it draws a larger current at a lower p.d., and the range of resistance which can be used is therefore smaller in value. The relevant property of the neon lamp is that, whereas it will conduct only when the p.d. across its electrodes rises above the striking potential V_1, it then continues to conduct until the p.d. falls below a particular value V_2 which is considerably lower than V_1, and in this behaviour it is consistent.

If the circuit of Fig. 8.33 is set up, the capacitor repeatedly charges up from the p.d. V_2 to the p.d. V_1 through the resistor R, and then discharges until the p.d.

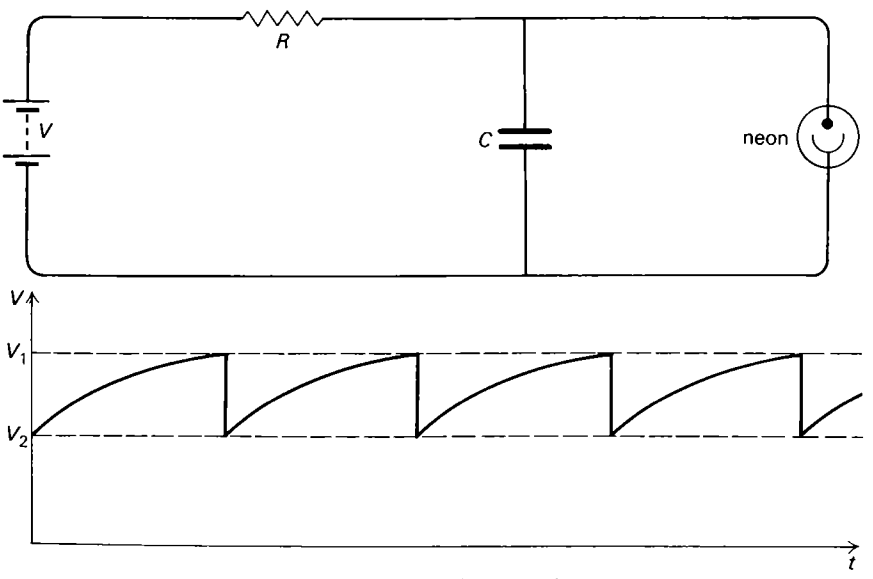


Fig. 8.33 The flashing neon lamp

has fallen to V_2 again, the discharging current flowing through the neon. The flashing frequency is proportional to the reciprocal product $1/RC$, because the discharge time is small enough to be neglected.

We can write

$$\frac{f_2}{f_1} = \frac{R_1}{R_2}, \tag{8.42}$$

where f_1, f_2 are the two flashing frequencies corresponding to resistances R_1, R_2.

Alternatively, both pulse electrometer and neon lamp may be used to compare two capacitances by connecting them in turn across the instrument and keeping the resistance R unchanged. In this case

$$\frac{f_2}{f_1} = \frac{C_1}{C_2}. \tag{8.43}$$

A further application of the neon lamp circuit is to use it as a simple stroboscopic source, since the lamp flashes with frequency f.

QUESTIONS

$$\epsilon_0 = 8.85 \times 10^{-12}\ \text{F/m}$$

capacitance of a parallel plate

$$\text{capacitor } C = \epsilon_r\, \epsilon_0 \frac{A}{d}$$

8.1 (*a*) A capacitor, with capacitance $5 \times 10^4\ \mu$F, consists of a dielectric, 10^{-1} mm thick, separating two parallel metal surfaces. If the relative permittivity ϵ_r of the dielectric is 6.0, calculate the area of each metal surface.

(*b*) If the capacitor is charged to a p.d. of 200 V, calculate the amount of charge and the amount of energy stored in it.

(*c*) How long would it take a 12 V battery of negligible resistance to provide the same amount of energy, when delivering a current of 5 A?

(*d*) If the charged capacitor were discharged in a suitable manner through a solution of copper sulphate, what mass of copper would be liberated at the cathode?

(Faraday constant $= 9.65 \times 10^4$ C/mol
(1 mole of Cu^{2+} ions $= 63.6 \times 10^{-3}$ kg.)

(*e*) If the charged capacitor were to deliver its energy in the space of 10^{-2} s in the form of an electric spark, what would be the mean power in the spark during this time?

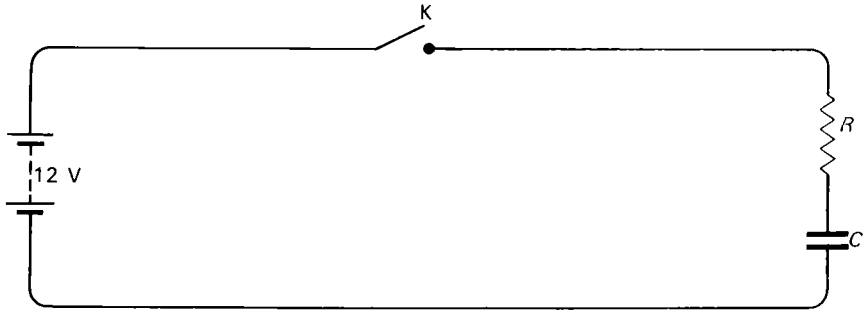


Fig. e 46

8.2

In Fig. e 46, $C = 4\ \mu F$

$R = 10^4\ \Omega$.

Calculate (*a*) the initial value of the charging current when the key is first closed, (*b*) the initial rate at which the p.d. across the capacitor is increasing, (*c*) the charging current at the instant when the p.d. across the capacitor is 8 V, (*d*) the amount of energy finally stored in the capacitor, (*e*) the amount of energy converted to heat in the resistor during the process of charging the capacitor.

8.3 An inductor of inductance 5 H and resistance 4 Ω is connected in series with a 20 V battery of negligible resistance.

Calculate (*a*) the final steady value of the current, (*b*) the initial rate at which the current grows at the instant when the circuit is made, (*c*) the final steady rate at which energy is converted to heat in the resistance, (*d*) the energy which must be dissipated when the current is switched off after reaching its final steady value.

8.4 A diode is used to provide rectified current through a very high resistance R from a sinusoidal a.c. source; the diode and resistor are connected in series across the source. Sketch labelled graphs showing how the p.d. across (*a*) the source, (*b*) the resistor, (*c*) the diode vary with time.

What would be the effect on the p.d. across (*b*) and (*c*) of connecting a capacitor of large capacitance C across R, such that $fRC \gg 1$, where f is the frequency of the source? (O & C)

8.5

Initially, capacitor A in the diagram is charged to a p.d. of 2000 V, and capacitor B is uncharged. When the key K is closed, the p.d. across A falls to 1600 V. If the capacitance of A is $4.0 \times 10^{-3}\ \mu F$, deduce the capacitance of B.

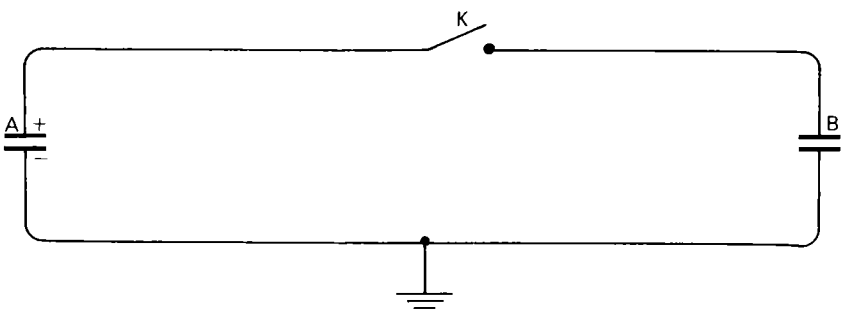


Fig. e 48

8.6

The components in the diagram have the following values:

$$R_1 = 8000\ \Omega, \quad R_2 = 4000\ \Omega, \quad C_1 = 2\ \mu F.$$

(*a*) If the p.d. between A and B is zero, what is the value of C_2?

(*b*) A resistor with resistance 4000 Ω is now connected in parallel with R_1, and a capacitor C is connected in parallel with C_1.

If the p.d. between A and B is again equal to zero, what is the value of C?

(*c*) While leaving the 4000 Ω resistor in parallel with R_1, capacitor C is removed, and a capacitor C' is connected, instead, in series with C_2, below the point B in the diagram. Again, the p.d. across AB is zero. What is the value of C'?

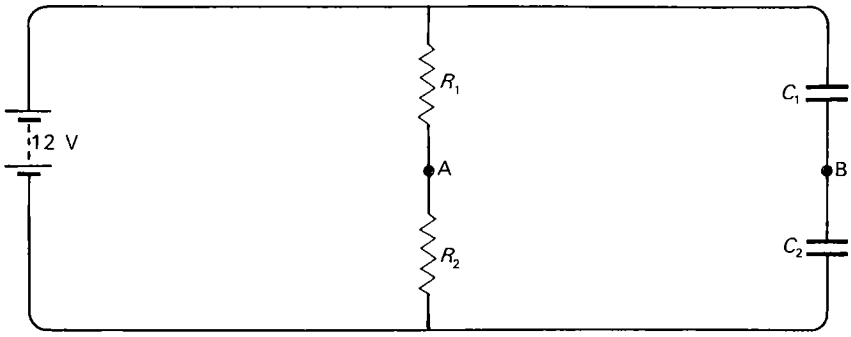


Fig. e 49

8.7

The values in this circuit are:

$$V = 70 \text{ V}, C_1 = 4\ \mu\text{F}, C_2 = 4\ \mu\text{F}, C_3 = 2\ \mu\text{F}, C_4 = 1\ \mu\text{F}.$$

(*a*) Initially, K_2 is open and K_1 is closed. Calculate the charge held by each capacitor and the p.d. across it.

(*b*) The key K_1 is now opened and K_2 is closed. Calculate the total amount of electrical energy dissipated.

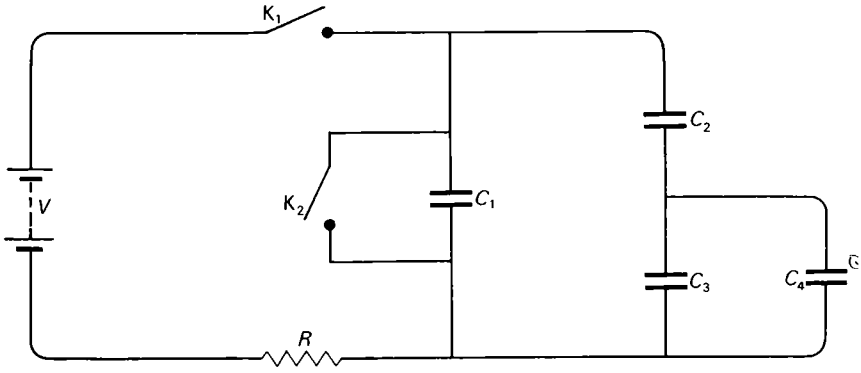


Fig. c 50

(*c*) Resistor R is a protecting device. What component in the circuit is protected, and against what accidental procedure?

8.8

Two capacitors of capacitance 1 μF and 2 μF respectively are charged in series from a 100 V d.c. supply. If the capacitors are now isolated and reconnected in parallel with each other (with their positive plates connected together), what is the p.d. across the capacitors?

How much electrical energy has been lost? Suggest the process by which the energy has been lost, and what has become of it.

8.9

Two capacitors and three resistors are connected, as shown in the diagram, to a battery of e.m.f. 6.0 V and of negligible internal resistance. The values of the components are:

$$C_1 = 10.0\ \mu\text{F}, C_2 = 4.0\ \mu\text{F}, R_1 = R_3 = 100\ \Omega, R_2 = 50\ \Omega.$$

Initially, all the switches are open and the capacitors are discharged.

Calculate the charges on the capacitors when (*a*) switch S_1 alone is closed, (*b*) switches S_1 and S_2 are closed, (*c*) all three switches are closed. (C)

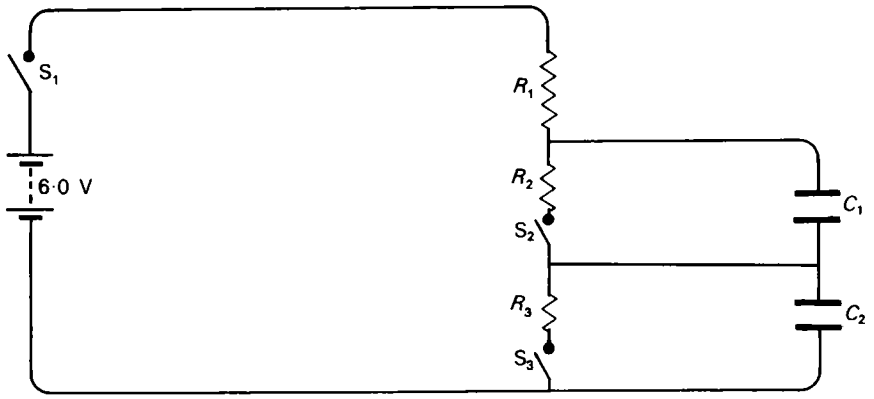


Fig. e 51

8.10

In the diagram,

$$R_1 = 3000\ \Omega,\ R_2 = 1000\ \Omega,\ C = 2\ \mu\text{F}.$$

(*a*) The key K is closed. Calculate the charge and energy stored in the capacitor *C*.

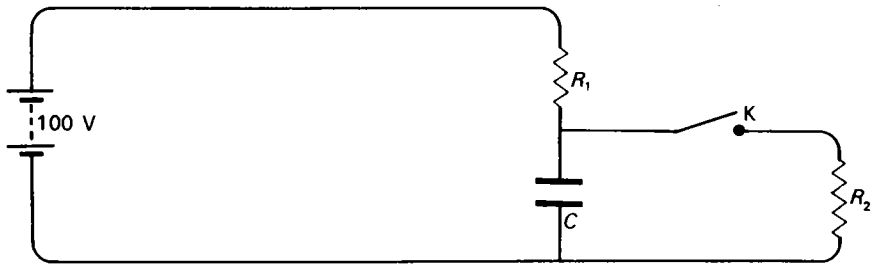


Fig. e 52

(*b*) Deduce the gain in charge and energy by the capacitor when the key is opened.

(*c*) How much charge flows through the battery after the key is opened, and how much energy is taken from the battery?

8.11

In Fig. e 53,

$$C_1 = 2\ \mu\text{F},\ C_2 = 4\ \mu\text{F},\ C_3 = 5\ \mu\text{F},\ V = 38\ \text{V}.$$

All the capacitors were discharged before they were connected in the circuit.

(*a*) Calculate the p.d. across each capacitor, and the total energy stored.

(*b*) Capacitor C_2 is now short circuited. Deduce the quantity of charge which flows through the battery in consequence, and the energy taken from the battery.

8.12

A capacitor *C* is placed in a circuit (Fig. e 54) so that it may be connected to the battery then rapidly connected across the milliammeter, and the process repeated at a constant frequency, the cycle being completed with a frequency of 100 Hz. If the battery maintains a p.d. of 150 V and the milliammeter reads 5.0 mA, deduce the capacitance of capacitor *C*.

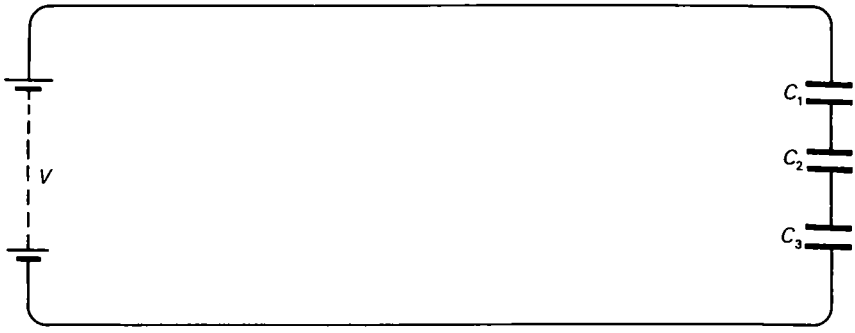


Fig. e 53

How would you modify the circuit, adding only resistors, so as to apply only 50 V to the capacitor? The battery has negligible internal resistance. (C)

8.13 A sensitive moving coil meter with a suspended system that can swing freely is calibrated both as a ballistic galvanometer and as a galvanometer to record steady currents. Its sensitivity is 8×10^{-8} C per division ballistically, and 2.5×10^{-7} A per division for steady currents.

When a capacitor is charged from a 24 V battery and discharged through the meter, the ballistic throw is 30 divisions. When a 2 V battery is used, and

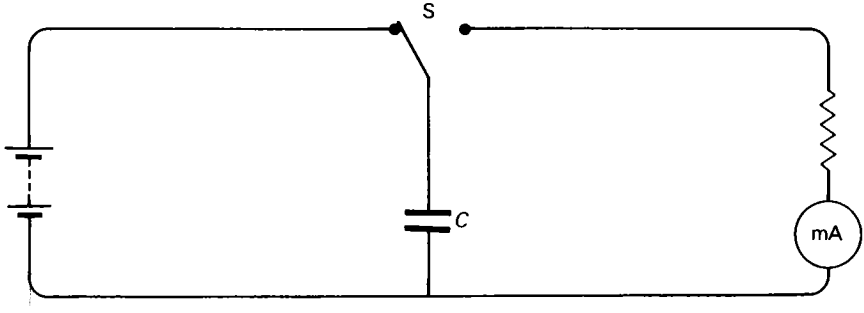


Fig. e 54

the capacitor is charged and discharged through the meter many times per second by means of a vibrating reed switch, the steady deflection is 80 divisions.

Find (*a*) the value of the capacitance, (*b*) the number of times the capacitor is discharged per second. (O)

8.14 Define 'electric potential' and 'capacitance of an insulated conductor'. Explain carefully, with reference to your definitions, why the presence of a similar earthed conductor near and parallel to an isolated sheet of metal considerably increases the capacitance of the sheet.

The figure shows a capacitor in a circuit in which a vibrating switch first charges it up to 200 V, and then discharges it through a microammeter, repeating the sequence 50 times per second. The mean current registered by the meter is 11.0 μA. Use this information to find the capacitance of the capacitor in farads.

The capacitor in fact consists of two flat square metal sheets, 50 cm by 50 cm, placed parallel, one over the other, and 2.0 mm apart. Calculate the magnitude of ϵ_0, the permittivity of a vacuum (or air), and state the units in which it is measured.

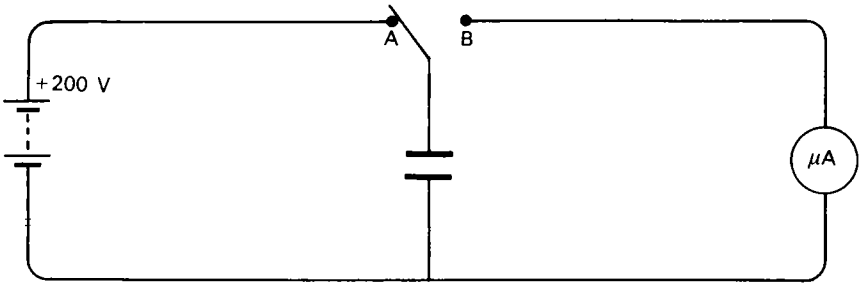


Fig. e 55

If a capacitor is to be accurately measured with such a circuit, it is essential that the part of the circuit between B and the galvanometer shall be well insulated from the part between A and the battery. What must be the minimum resistance of any possible leakage path between these parts if the error introduced is not to exceed 1 per cent in measuring the capacitance? (S)

8.15 A voltmeter of very high resistance is connected across the terminals of a 10 μF capacitor which is initially charged to 410 V. When the charging source is removed the voltmeter reading falls, and reaches 390 V after 10.0 s. Calculate the resistance of the voltmeter.

(You may use an approximate method of calculation, but must explain why it is approximate.) (O & C)

8.16 In the circuit, the neon lamp N strikes when the p.d. across it rises to 110 V, and extinguishes when it falls to 85 V. When it is extinguished its resistance may be considered infinite, and when lit very small.

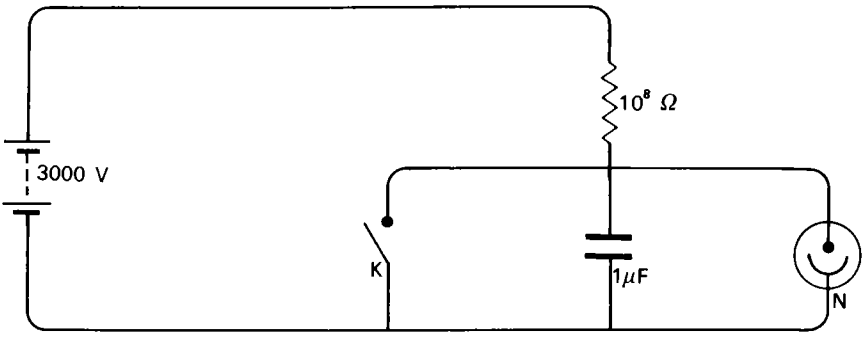


Fig. e 56

If the key K is opened at time $t = 0$, calculate approximately the time when the neon lamp first strikes, and the frequency of subsequent strikings. (CAM. SCH.)

8.17 The diagram represents a circuit intended to rectify current of frequency 50 Hz from a transformer, in order to send direct current of a steady value through the load resistor R.

Supposing that the smoothing capacitor C is not fully effective, sketch a graph showing how the p.d. across AB would vary with time, and indicate what the effect would be if the current sent through R were increased.

If the p.d. across the capacitor is brought up to 24 V once in each cycle of

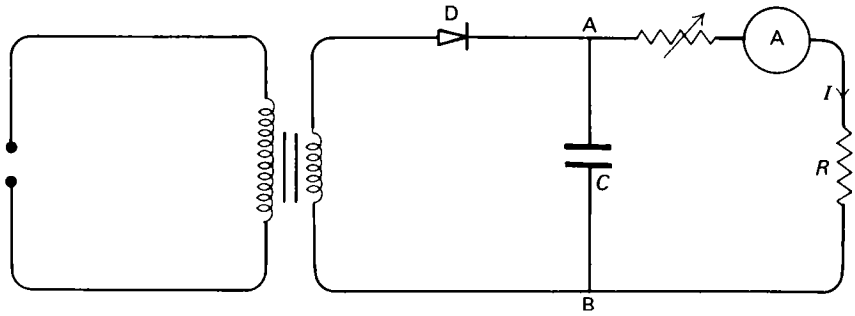


Fig. e 57

the alternating current when the diode D conducts, and if the p.d. is not to fall by more than 0.5 V during the remainder of the cycle, obtain an approximate value for the capacitance required for C in each of the following cases: (*a*) $I = 5$ A, (*b*) $I = 1$ A, (*c*) $I = 5$ A, but the frequency of the supply is 3000 Hz.

(As good approximations for the circumstances, the current may be treated as effectively constant, and the capacitor may be taken to be losing its charge over the period of a complete cycle between successive recharges through the diode.)

Do you consider this circuit to be practicable in all three sets of circumstances suggested in the question?

8.18 A current of 4 A is flowing in an inductance of 5 H. How much energy is stored in the magnetic field of the inductor? In order to prevent arcing across the switch when the current is switched off, a capacitor C is connected across the switch, to take up the energy from the inductance.

Explain qualitatively what occurs when the switch is broken. If the p.d. across the switch is not to exceed 1000 V, calculate the minimum value adequate for the capacitance of C.

8.19 An oscillating circuit is made up by connecting an inductor, with inductance 0.01 H, across a capacitor, with capacitance 2.0 μF. Neglecting any effect which resistance might have, calculate the frequency of the electrical oscillations which result when the capacitor is discharged through the inductor.

If the capacitor is initially charged to a p.d. of 100 V, calculate the energy stored in it, and the maximum value reached by the current in the inductor.

If the capacitance is now reduced to 0.02 μF, calculate the value of inductance which would be required to provide oscillations with the same frequency as before. To what p.d. must the capacitor initially be charged for the oscillations to have the same energy as before, and what is the maximum value of the current in the inductor?

*8.20 Describe briefly how you would attempt to test experimentally whether the capacitance C of a small radio capacitor labelled 0.1 μF was independent of the p.d. across it. (OXF. SCH.)

*8.21 See Fig. e 58.

$$C_1 = 1\ \mu\text{F},\ C_2 = 2\ \mu\text{F},\ C_3 = 3\ \mu\text{F};$$
$$R_1 = 1\ \Omega,\ R_2 = 3\ \Omega.$$

The p.d. across PQ = 4 V. Find the value of the p.d. across AB. (CAM. SCH.)

*8.22

$$C_1 = 3\ \mu\text{F},\ C_2 = 4\ \mu\text{F},\ C_3 = 5\ \mu\text{F}.$$

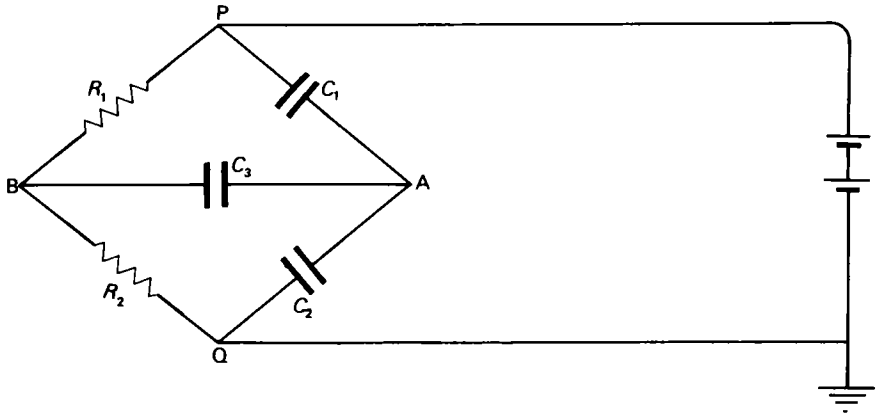


Fig. e 58

Initially, C_1, C_2, C_3 are separately charged up, and have p.d.s of 300 V, 200 V, and 100 V respectively across them, the plates A, D, F being positive.

Without discharging them, the three capacitors are now connected together to form a continuous circuit, plate A being connected to plate E, D to G, and F to B. Deduce the quantity of charge which flows from one

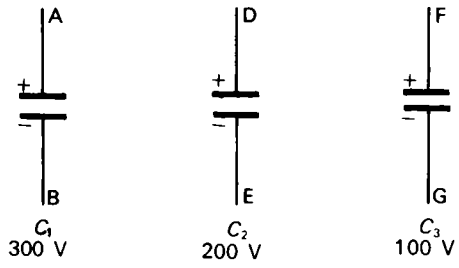


Fig. e 59

capacitor to another round the circuit, and the resulting p.d. across each capacitor.

*8.23 See Fig. e 60.

$$C_1 = 3\ \mu\text{F},\ C_2 = 2\ \mu\text{F},\ C_3 = 4\ \mu\text{F}.$$

(*a*) Initially, S_1 is closed and S_2 is open. Find the charge stored in each capacitor and the p.d. across each.

(*b*) S_1 is now opened, and S_2 is closed. What is the total charge which flows through resistor R, and how much energy is dissipated in it?

(*c*) Switch S_2 is kept closed, and switch S_1 is now closed also. How much charge flows through the battery, and how much energy is taken from the battery?

*8.24 Three plane parallel conducting plates A, B, C are spaced with AB = 2 mm and BC = 4 mm. The dimensions of the plates are much larger than the spacings, so that end corrections may be ignored. A and C are joined together, and B is raised to a potential of 100 V with respect to A and C. The plates are then isolated, and C is raised to a potential of 50 V above A.

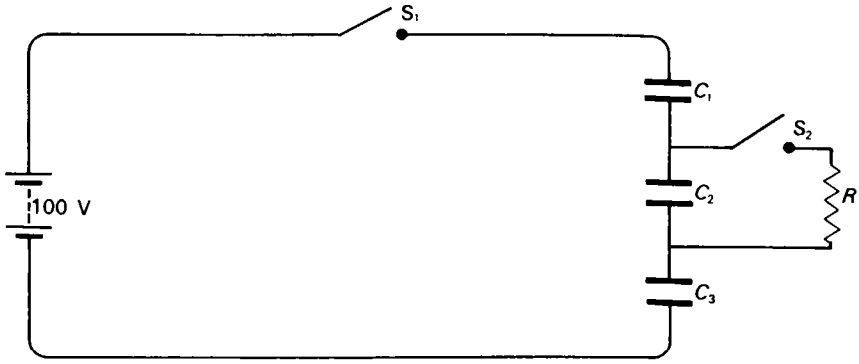


Fig. e 60

What is the p.d. between B and A? (CAM. SCH. ENG.)

(*Hint:* treat this as a problem concerning two capacitors, one with plates A and B, the other with plates B and C; also, for parallel plates, $C \propto d^{-1}$ where C = capacitance and d = separation.)

*8.25 In the circuit shown, both capacitors are initially uncharged. The switch S vibrates so as to make contact alternately with A and with B. If V_n denotes

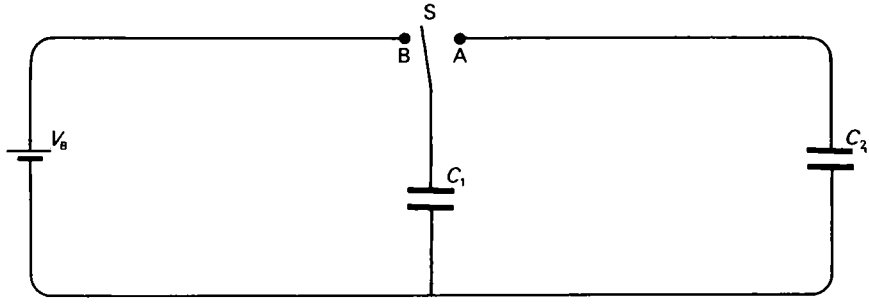


Fig. e 61

the p.d. across both capacitors after n cycles, show by considering one further cycle that

$$V_B - V_{n+1} = \frac{C_2}{C_1 + C_2}(V_B - V_n).$$

Hence find the minimum number of cycles needed to charge C_2 to more than 99 per cent of V_B, if $C_2 = 2C_1$. (OXF. SCH.)

*8.26 The current I through a certain non-linear circuit element is given by the relationship

$$5I = V^2,$$

where I is measured in μA and V, the applied p.d., is in V. The element is initially charged to a p.d. of 100 V. How long will it take for this p.d. to drop to 50 V? (OXF. SCH.)

*8.27 When a certain capacitor is disconnected from a battery, the p.d. between its plates gradually decreases. Show that if the capacitor is discharging through the dielectric material only (and not through the supports, for example) the time taken for the p.d. to fall to one half of its original value is independent of the size of the capacitor.

Calculate this time when the relative permittivity of the dielectric material is 3.14 and its resistivity is $3.6 \times 10^{10}\ \Omega\cdot\text{m}$.
(Take $\log_e 2 = 0.70$.) (CAM. SCH.)

*8.28

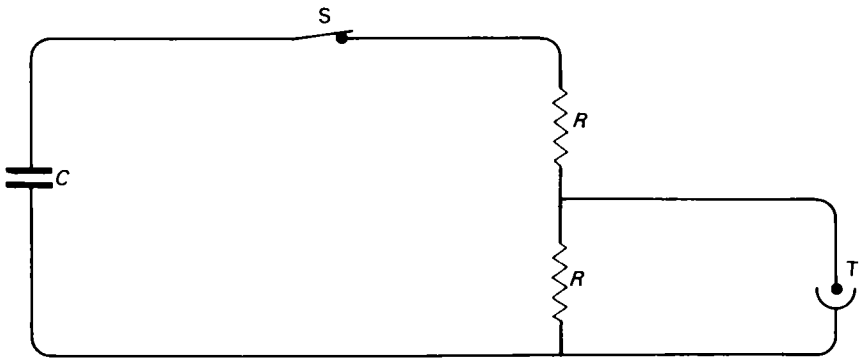


Fig. e 62

C is initially charged to p.d. V_0. T is a discharge tube. So long as T is conducting, its characteristic is such that the p.d. across it stays constant at $\frac{1}{3}V_0$. If T conducts as soon as the switch is closed, how long will elapse before T becomes non-conducting? (OXF. SCH.)

*8.29

In the circuit shown, the capacitor charges through the resistor $4R$ until the voltage at X is +300 V with respect to earth. The switch then connects X to contact 2. When X has reached earth potential, the switch reconnects it to

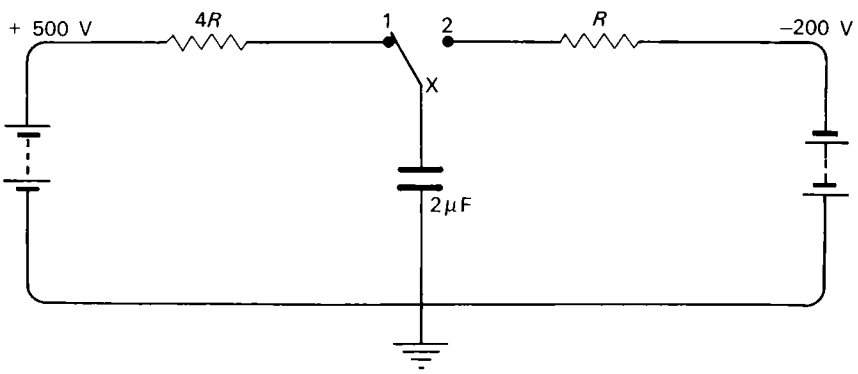


Fig. e 63

contact 1, and the process repeats itself continuously. For a certain value of R, the time for a complete cycle is 5×10^{-3} s. At what average rate is electrical energy converted into heat by the resistors?

By considering approximate values of the currents charging and discharging the capacitor, or otherwise, estimate the value of R needed to give this cycle time.

(Illustrate the process by sketching a graph of V_X against t.) (CAM. SCH. ENG.)

*8.30 In the two circuits in the diagram,

$$C = 2.0\ \mu\text{F},\ R = 10^6\ \Omega,\ V = 10\ \text{V}.$$

The switch is first connected to contact 2, and then moved across to contact 1. Derive an expression giving the value of the p.d. across the capacitor at

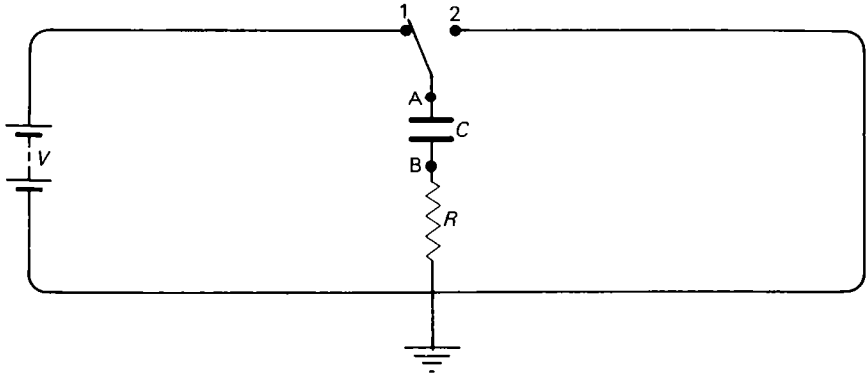


Fig. e 64 (a)

time t after the switch has made contact to 1, where t is measured in seconds.

If the switch is moved from 2 to 1 and back to 2, with a time interval of 10 s between each change, sketch the variation of potential with time of each of the points A, B, C, D in the two circuits. (Assume that the capacitor is effectively charged to 10 V in 10 s.)

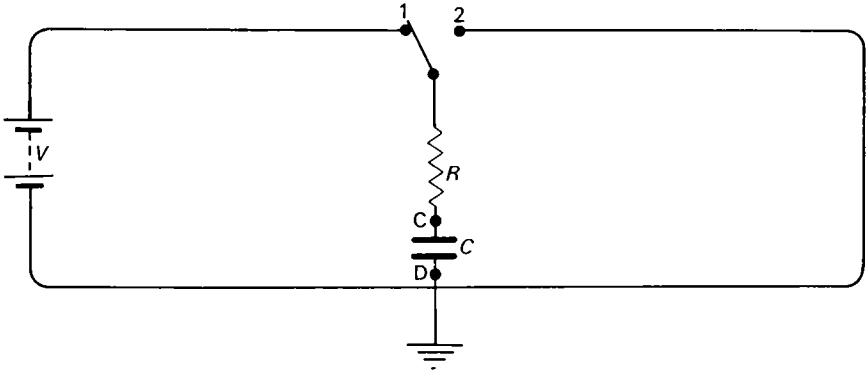


Fig. e 64 (b)

*8.31 A battery of voltage V_0 and internal resistance R_0 is suddenly connected to a circuit consisting of inductance L in parallel with a resistance R.

Find an equation relating the current I through the coil to its rate of change, and show that the value of I at the instant t is given by

$$I = \frac{V_0}{R_0}\left\{1 - \exp\left[\frac{R_0 R t}{L(R + R_0)}\right]\right\}. \qquad \text{(OXF. SCH.)}$$

*8.32 A coil of resistance R and inductance L is connected to a source of e.m.f. E. Derive an expression for the growth of the current, and calculate the time taken for it to reach 63 per cent of the final value, when $R = 1\ \Omega$ and $L = 1$ mH.

What is the work done in establishing the current at this time if $E = 1$ V? How much of this work is dissipated as heat? (CAM. SCH. ENG.)

($e^{-1} = 0.37$)

*8.33 A relay coil has an inductance of 0.5 H and a resistance of 20 Ω. The coil is suddenly connected to a source having a constant p.d. of 10 V. After 0.02 s the source is suddenly disconnected and the terminals of the coil are simultaneously short circuited. If the relay contacts operated by the coil close when the coil current is 0.2 A and increasing, and open when the coil current is 0.1 A and decreasing, find the total time during which the contacts remain closed. (CAM. SCH. ENG.)

8.34 Describe in detail how you would attempt to compare (*a*) two resistances each of the order of 0.1 Ω, (*b*) two resistances each of the order of 20 MΩ. State the accuracy you would expect to obtain in each case. What objections are there to the Wheatstone bridge method in each case? (S)

Chapter 9

Alternating current

9.1 Waveform

Because alternating current is a continually varying current, and particularly because it repeatedly reverses in direction, it does not always resemble a steady direct current in its effect, and a measurement of it must represent some kind of average value being taken. These are the reasons that alternating current is treated in a chapter of its own.

A current is properly described as alternating if it varies periodically, that is to say if the current repeatedly passes through the same cycle of values, with a constant frequency. Additionally, it must do so in such a way that the average value of the current over a period of time is zero. If the variation is periodic, but the average is not zero, this is because the current has a d.c. component in addition to an alternating one.

An alternating e.m.f. is readily obtained by electromagnetic induction: all that is required is to rotate a magnet at constant speed in such a way that its flux threads a coil in a periodic manner as the magnet's poles repeatedly sweep past, thus inducing an alternating e.m.f. in the coil. A bicycle dynamo is a simple example of this, and large generators work on much the same principle. An alternative technique, particularly useful at higher frequencies, is to use an oscillating circuit of some kind, in which the oscillations are maintained by the agency of a thermionic valve or a transistor.

A graph which traces the variation of current or e.m.f. with time shows the waveform, and both the currents represented in Fig. 9.2 have waveforms which classify them as alternating current.

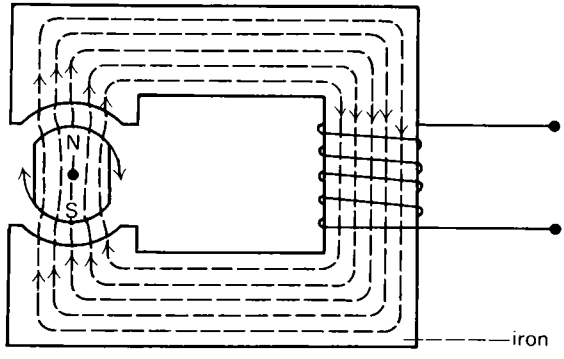


Fig. 9.1 Showing the principle of the bicycle dynamo. As the magnet rotates, with frequency f, the variation in flux threading the coil induces an alternating e.m.f. in it, having frequency f

The other distinctive features of an alternating current besides its waveform are its peak value I_0, which is the maximum value reached by the current in the course of its oscillation, and the frequency f, measured in hertz, i.e. cycles per second.

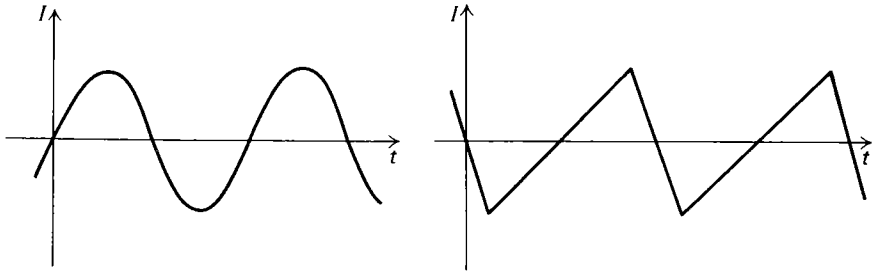


Fig. 9.2 Alternating current. Two examples of an alternating waveform (a) Sinusoidal (b) Sawtooth

9.2 Root mean square current

The ordinary a.c. meter does not follow the variations in current as an oscilloscope does; its pointer does not oscillate in sympathy with the alternating current or p.d., but registers a steady deflection. This raises a question as to what particular form of average it should read. The peak value, which an oscilloscope would measure, is clearly too big, whereas the mean value, i.e. the ordinary arithmetical mean, is zero. It is actually calibrated to read the root mean square current or p.d., because this gives a reasonable estimate for comparison with d.c., as explained below.

In some respects, of course, the effects of d.c. and a.c. are quite different, and no equivalence between the two exists. Alternating current, for example, will not charge an accumulator, whereas direct current will not operate a transformer. The common ground for comparison must be sought in an irreversible effect, that is, one which does not reverse with reversal in direction of the current, e.g. the Joule heating effect in a resistor.

The rate P at which heat is produced in a resistor is given by

$$P = RI^2 \tag{9.1}$$

i.e. if the behaviour of the resistor can be described by Ohm's law, making R constant, P is proportional to the square of the current. If the current varies, as is the case with alternating current, the mean value of P is proportional to the mean value of I^2. An a.c. meter is therefore graduated to register the square root of this mean square value, i.e. the root mean square. The alternating current, measured in this way to have an r.m.s. value I, will then produce the same heating effect in a resistor R as a direct current does which has the steady value I, and equation (9.1) applies in both cases. Similarly with potential difference.

The meaning of the r.m.s. value, $I_{\text{r.m.s.}}$, is illustrated graphically in Fig. 9.3. If we were to measure the current I at a series of equal intervals of time over a complete cycle, square all these values, and then average the squares, this would give the mean square current, or something very close to it. (To be accurate, a very large number of values of I must be taken, at very short time intervals.)

$$\text{The mean square current} = \frac{I_1^2 + I_2^2 + \ldots I_n^2}{n}$$

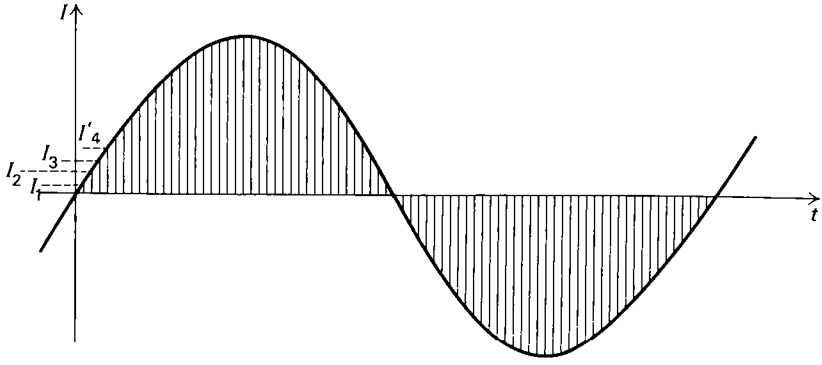


Fig. 9.3 Illustrating the meaning of r.m.s. current. $I_{\text{rms}} = \left\{\frac{(I_1^2 + I_2^2 + \ldots I_n^2)}{n}\right\}^{\frac{1}{2}}$

and the r.m.s. current is the square root of this average,

i.e. $$I_{\text{r.m.s.}} = \left[\frac{I_1^2 + I_2^2 + \ldots I_n^2}{n}\right]^{\frac{1}{2}}. \tag{9.2}$$

The accurate definition of r.m.s.

More accurately, the defining equation is

$$I_{\text{r.m.s.}}{}^2 = \frac{1}{T}\int_0^T I^2 \, dt, \tag{9.3}$$

where I is the value of the current at the instant t, and T is the periodic time of one complete cycle, $T = f^{-1}$.

This equation is interpreted in Fig. 9.4, for the case of a current having a sawtooth waveform. If the values of I^2 are plotted against t, the expression $\int_0^T I^2 \, dt$ gives the area below the graph for the time T of one complete cycle. The mean square

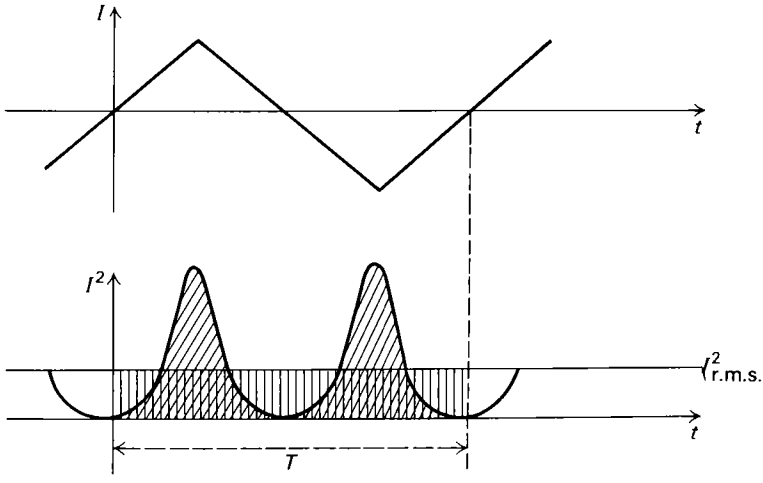


Fig. 9.4 r.m.s. current (a) A sawtooth waveform (b) I^2 against t for the sawtooth waveform above

value, $I_{\text{r.m.s.}}^2$, is also plotted, the area below this being equal to $I_{\text{r.m.s.}}^2 T$. By the definition given in equation 9.3, these two areas are equal.

9.3 The equation of sinusoidal current

Mathematically, sinusoidal current, i.e. current having the waveform illustrated in Fig. 9.2(*a*), is represented by the equation

$$I = I_0 \sin 2\pi ft, \tag{9.4}$$

or

$$I = I_0 \sin \omega t, \tag{9.5}$$

where f = the frequency with which the value of the current is repeated, and $\omega = 2\pi f$ is called the angular frequency. This is, for example, the angular velocity of the rotating magnet in Fig. 9.1.

I = the value of the current at the instant t

I_0 = the peak value of the current.

The choice of the instant taken as zero time is arbitrary, and the equation has been written here in the form which applies if $t = 0$ when $I = 0$. The time could, for example, have been taken as zero at the instant when $I = I_0$, in which case equation 9.4 would be replaced by the form

$$I = I_0 \cos 2\pi ft$$

The equations which will be used and derived in the remainder of this chapter apply to sinusoidal waveform, and cannot be taken to have direct application in other cases. The fact is, however, that the sinusoidal variation is the basis of all periodic variations, as summed up by Fourier's theorem, which is stated below. The sinusoidal oscillation is the only one which consists of a single frequency, rather than being a mixture of several frequencies, and for this reason it is said to have pure waveform; this is the type of waveform which practical a.c. generators are designed to produce.

Fourier's theorem

Fourier's theorem states, in mathematical terms, that any periodic variation, such as the sawtooth of Fig. 9.4(*a*), may be analysed into, or built up from, a series of sinusoidal components. If the basic frequency is f, these components have frequencies f, $2f$, $3f$, $4f$ etc.

One practical example of this in another branch of physics is the complicated waveform of a note from a musical instrument, which results from the combination of the fundamental note and its overtones.

It follows from Fourier's theorem that an analysis of the way in which a device or a circuit behaves towards pure waveform a.c. provides the basis, also, of its behaviour towards more complicated waveforms.

9.4 The r.m.s. value of sinusoidal a.c.

When the waveform is sinusoidal, the variation of I^2 with time has a special symmetry and the value of I^2 oscillates between the two extremes, zero and I_0^2. Owing to this symmetry, the mean value is $\frac{1}{2}I_0^2$, and the r.m.s. value of the current is

$$I_{\text{r.m.s.}} = \frac{1}{\sqrt{2}} I_0. \tag{9.6}$$

Similarly, the r.m.s. value of a sinusoidal p.d. is

$$V_{\text{r.m.s.}} = \frac{1}{\sqrt{2}} V_0. \tag{9.7}$$

It is stressed that these equations apply only to sinusoidal waveform, and are not true in the general case. They do not apply, for example, to the case illustrated in Fig. 9.4.

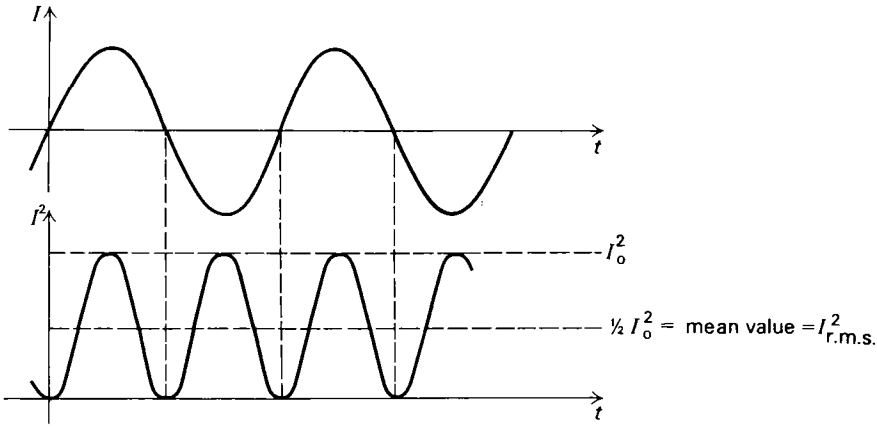


Fig. 9.5 The r.m.s. value of sinusoidal a.c. (a) (b)

The mathematical justification for equation 9.6 is as follows.
We have:

$$I = I_0 \sin 2\pi ft,$$

$$I = I_0 \sin \omega t, \tag{9.5}$$

$$I^2 = I_0^2 \sin^2 \omega t$$

$$= \tfrac{1}{2} I_0^2 (1 - \cos 2\omega t). \tag{9.8}$$

This is the equation of the curve drawn in Fig. 9.5(*b*) and it represents a cosine curve with a repetition rate equal to $2f$, which is symmetrical about the mean value $I_{\text{r.m.s.}}{}^2$, where

$$I_{\text{r.m.s.}}{}^2 = \tfrac{1}{2} I_0^2,$$

i.e. the mean value of I^2 is $\frac{1}{2} I_0^2$ because the second term inside the bracket in equation 9.8 averages, over a complete cycle, to zero.

*A more general derivation

The use of calculus has deliberately been avoided in the above derivation, but this could not have been done so easily for a more complicated waveform. An alternative treatment, which has more general application, is as follows.

We have

$$I_{\text{r.m.s.}}{}^2 = \frac{1}{T} \int_0^T I^2 \mathrm{d}t, \tag{9.3}$$

where $$I = I_0 \sin \omega t.$$

$$I_{\text{r.m.s.}}{}^2 = \frac{1}{T} I_0^2 \int_0^T \sin^2 \omega t \, \mathrm{d}t$$

$$= \frac{1}{T} \cdot \tfrac{1}{2} I_0^2 \int_0^T (1 - \cos 2\omega t) \mathrm{d}t$$

$$= \tfrac{1}{2} I_0^2.$$

This is because $\int_0^T \mathrm{d}t = T$ and $\int_0^T \cos 2\omega t \, \mathrm{d}t = 0$.

The relations between r.m.s. and peak values of current and p.d. are of basic importance to what follows. This is because, whereas it is the r.m.s. value which is used in practice, it is the peak value which directly concerns the theory.

9.5 Resistance, capacitance, and inductance in an a.c. circuit

Because the flow of alternating current involves continuous change of current, including repeated reversals in its direction, capacitance and inductance have important effects which they lack in the d.c. circuit. The capacitor allows current to flow in the process of repeatedly charging it and discharging it, and the inductor provides opposition to the changes in current by the process of self induction. In describing in turn the effects of resistance, capacitance, and inductance, we will first consider the basic physics, and a more precise mathematical analysis will be given at the end of the chapter.

9.6 Resistance

When alternating current flows in resistance, the p.d. and current are in phase with each other; at the instant when the p.d. reaches maximum, the current reaches maximum also.

The equations which apply to direct current are valid in terms of r.m.s. values,

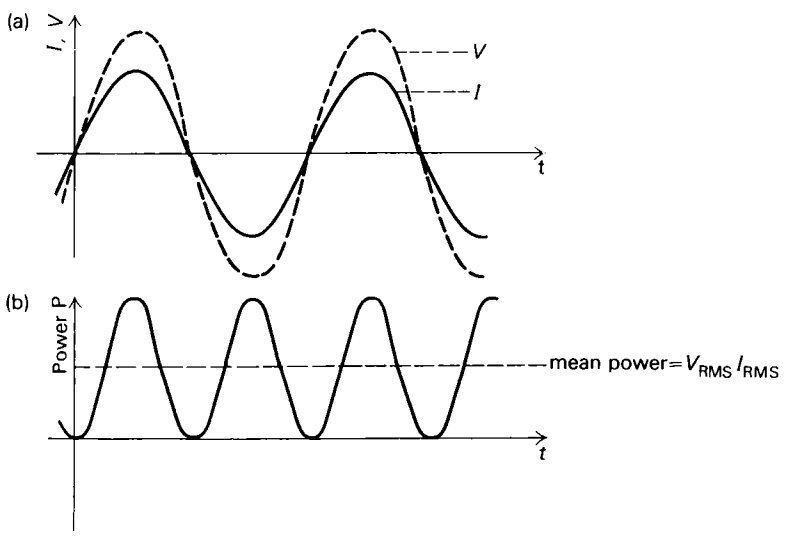


Fig. 9.6. Alternating current in a resistor (a) (b)

which is the reason that the measuring instruments are calibrated to register r.m.s. values. We have

$$V_{\text{r.m.s.}} = RI_{\text{r.m.s.}}, \tag{9.9}$$

and

$$P_{\text{av}} = V_{\text{r.m.s.}}\, I_{\text{r.m.s.}}$$
$$= I^2_{\text{r.m.s.}}\, R, \tag{9.10}$$

where P_{av} is the average rate at which energy is converted to heat.

9.7 Capacitance

When an alternating source is connected across a capacitor, current flows to and fro continuously, as the capacitor repeatedly charges up, discharges, and charges up again with reversed polarity. The successive stages in this behaviour, in the course of one complete cycle of the alternating current, are illustrated in Fig. 9.7. This is very different from the effect of connecting a d.c. source, where current flows only for sufficient time at the outset to charge the capacitor.

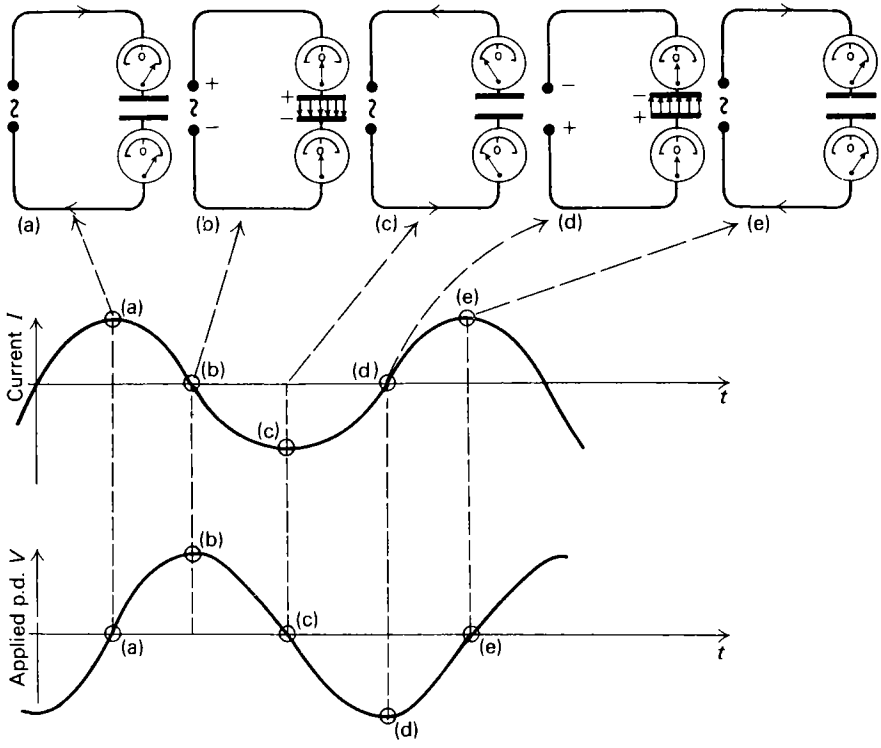


Fig. 9.7 Alternating current in a capacitor

The successive instants of time represented by points (*a*), (*b*), (*c*), (*d*), (*e*) in the diagram are separated from each other by intervals corresponding to one quarter cycle of the alternating current. The current is first shown flowing on to the top plate of the capacitor, and off the bottom plate. The p.d. across the capacitor grows as it charges up until, at instant (*b*), when the current has fallen to zero, the p.d. is maximum. Because this p.d. must, at all times, be equal to the p.d. applied

from the source, the applied p.d. is also maximum at the same instant. During the next half cycle of the current, from instant (*b*) until instant (*d*), current is flowing in the reverse direction, the capacitor reaching its maximum reverse p.d. at the instant shown at (*d*). So it goes on, and the current flows continuously, first in one direction, then in the other.

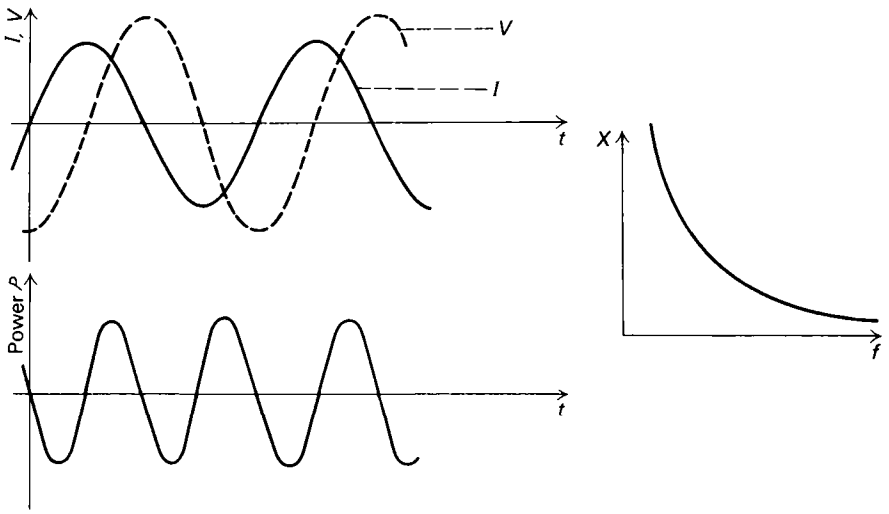


Fig. 9.8 Alternating current in capacitance

Because the p.d. across the capacitor results from current which has already flowed to charge it, it follows that the p.d. lags behind the current in its oscillations, the lag being a quarter of a cycle. So long as the current is flowing in one direction, whether growing in size or dying away, it is still charging the capacitor, and so the p.d. across the capacitor is still growing, to reach maximum at the instant when the current has fallen to zero and is about to flow in the opposite direction. Because of this difference in phase between p.d. and current, the mean power taken by the capacitor, i.e. the mean value of the product VI, is zero; the capacitor does not emit heat, but it repeatedly stores some energy, and then surrenders it again.

The maximum amount of charge stored in the capacitor in each half cycle is equal to the product of its capacitance C and the peak value V_0 of the applied p.d. It acquires this charge, first with one polarity, then with the other, with frequency f. Consequently, an average value for the current which flows may be taken to be proportional to these three factors and we can write

$$I \propto fCV_0$$

It will presently be shown mathematically that the exact relation between r.m.s. values is

$$I_{\text{r.m.s.}} = \omega C V_{\text{r.m.s.}}, \tag{9.11}$$

where $\omega = 2\pi f$.

This clearly resembles the familiar equation 9.10 of the resistor, and we write

$$X_C = \frac{V_{\text{r.m.s.}}}{I_{\text{r.m.s.}}}, \tag{9.12}$$

where $$X_C = \frac{1}{\omega C}, \tag{9.13}$$

and the SI unit of X_C, being V/A, is the ohm.

This property X_C of the capacitor is called its reactance. It is given a distinctive name, and represented by a distinctive symbol, because it differs in its effect from resistance in the following important respects.

The p.d. and current in the reactance are not in phase with each other, as they are in resistance. On the contrary, the current leads the p.d. by a quarter of a cycle. Consequently, the mean power conversion in a reactance is zero. Whereas, for resistance, we have the equations

$$P_{av} = V_{r.m.s.} I_{r.m.s.}$$

and $$P_{av} = I^2_{r.m.s.} R,$$

the corresponding equations

$$P_{av} = V_{r.m.s.} I_{r.m.s.}$$

and $$P_{av} = I^2_{r.m.s.} X_C$$

are not valid in the case of reactance, being replaced simply by

$$P_{av} = 0.$$

Whereas resistance is in most cases independent of frequency, the reactance of a capacitance is in inverse proportion to frequency, as shown by equation 9.13 (assuming that the capacitance C is constant). It is this discrimination between currents and p.d.s of different frequencies that gives capacitance its special practical importance. The capacitor accepts current at high frequency more readily than at low frequency, and will not pass d.c. at all, except for a short period after the p.d. has first been applied.

9.8 Inductance

The functions of capacitance and inductance in an a.c. circuit are complementary. The capacitor favours high frequency current, and the inductor favours low frequency.

When alternating current flows in an inductor, no p.d. is required at the instants when the current is not changing. (This assumes, of course, that the resistance is zero.) When the current is growing, however, a forward p.d. is necessary, to overcome the self induced back e.m.f., and a reverse p.d. is required when it is dying away again, or when it is growing in the reverse direction. This means that the p.d. leads the current in phase by a quarter cycle, as compared with the case of the capacitor, where the current leads the p.d.

The p.d. at any instant is proportional to the inductance L and to the rate at which the current is changing, and this evidently depends on both the size of the peak value I_0 which the current is going to reach and the frequency with which it does so. Consequently, an average value for the p.d. may be taken to be proportional to these three factors, and we can write

$$V_0 \propto f L I_0.$$

It will presently be shown mathematically that the exact relation between r.m.s. values is

$$V_{r.m.s.} = \omega L I_{r.m.s.}, \tag{9.14}$$

where $$\omega = 2\pi f.$$

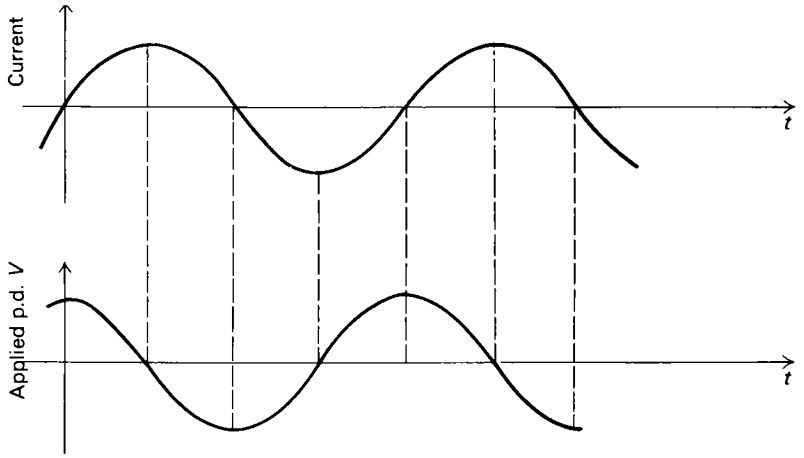


Fig. 9.9 Alternating current in an inductor. Current and applied p.d.

This resembles the case of the capacitor, and we write

$$X_L = \frac{V_{\text{r.m.s.}}}{I_{\text{r.m.s.}}}, \tag{9.15}$$

where

$$X_L = \omega L, \tag{9.16}$$

X_L being the reactance of the inductor.

As with the capacitor, this reactance does not continuously take power from the source, and we cannot write $P_{\text{av}} = V_{\text{r.m.s.}} I_{\text{r.m.s.}}$, or $P_{\text{av}} = I^2_{\text{r.m.s.}} X_L$. Energy is repeatedly becoming stored in the magnetic flux of the inductor and surrendered again, and the mean power is zero.

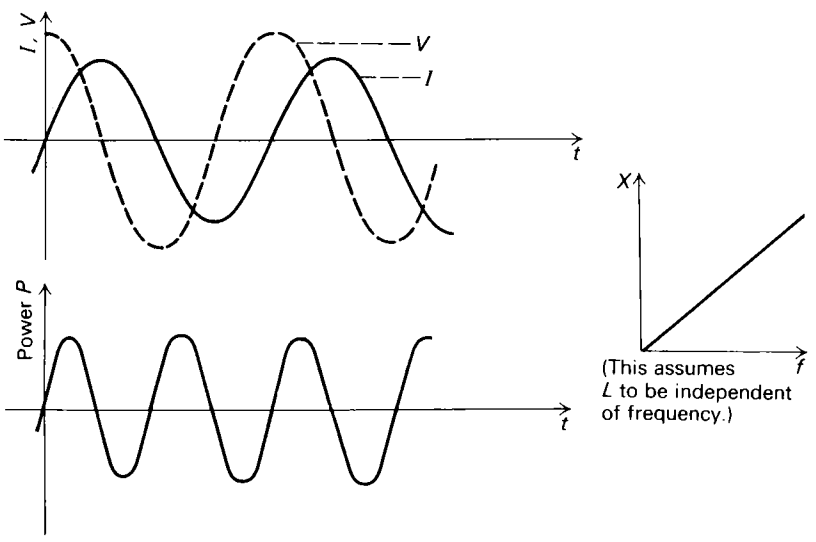


Fig. 9.10 Alternating current in an inductance

A comparison of equations 9.13 and 9.16 illustrates the difference between capacitance and inductance in their frequency discrimination. One or two practical reservations must be made. Any practical form of inductor necessarily includes resistance in its windings, and a pure inductance, unless a superconducting one, is unattainable. Also, if the inductor has a ferromagnetic core, its inductance L is not constant, being dependent on the value of the current, and possibly on the frequency too. In this case, therefore, the reactance is not constant.

9.9 Examples

Example 1

A sinusoidal alternating p.d. with r.m.s. value 240 V *and frequency* 50 Hz *is applied across a* 0.8 μF *capacitor. Deduce the r.m.s. value of the current, and the maximum amount of energy which becomes stored, at any time, in the capacitor.*

Recalculate these two values for the new circumstances when the frequency has been raised to 3000 Hz, *the p.d. remaining unaltered.*

When f = 50 Hz:

$$X_C = \frac{1}{\omega C} = \frac{1}{2\pi \times 50 \times 0.8 \times 10^{-6}}\ \Omega = 3980\ \Omega,$$

$$\therefore \quad I_{\text{r.m.s.}} = \frac{V_{\text{r.m.s.}}}{X_C} = \frac{240}{3980}\ \text{A} = 0.060\ \text{A}.$$

The maximum p.d. across the capacitor in the course of a cycle is the peak value V_0 where

$$V_0 = V_{\text{r.m.s.}}\sqrt{2} = 240 \times 1.414\ \text{V} = 340\ \text{V}.$$

The maximum amount of energy stored in the capacitor is therefore

$$W = \tfrac{1}{2}CV_0^2 = \tfrac{1}{2} \times 0.8 \times 10^{-6} \times 340^2\ \text{J} = 0.046\ \text{J},$$

or

$$W = \tfrac{1}{2}CV_0^2 = CV^2_{\text{r.m.s.}} = 0.8 \times 10^{-6} \times 240^2\ \text{J} = 0.046\ \text{J}.$$

When f = 3000 Hz:

The reactance of a capacitor is in inverse proportion to frequency.

Therefore

$$I_{\text{r.m.s.}} = \frac{V_{\text{r.m.s.}}}{X_C},$$

where, in this case, only frequency has been changed,

i.e.

$$I_{\text{r.m.s.}} \propto f.$$

The new value, therefore, is given by

$$I_{\text{r.m.s.}} = 0.060 \times \frac{3000}{50}\ \text{A} = 3.6\ \text{A}.$$

The energy stored in each cycle depends only on capacitance and applied p.d., which have not changed,

$$\therefore \quad W = 0.046\ \text{J}.$$

Example 2

A sinusoidal alternating p.d. with r.m.s. value 240 V *and frequency* 50 Hz *is applied across an inductor, of which the effective resistance is zero and the inductance is* 20 H.

Deduce the r.m.s. value of the current, and the maximum amount of energy stored in the flux of the inductor.

Recalculate these two values for the new circumstance when the frequency has been raised to 3000 Hz, *the p.d. remaining unaltered.*

When $f = 50$ Hz:

$$X_L = \omega L = 2\pi \times 50 \times 20\ \Omega = 6280\ \Omega,$$

$$\therefore \quad I_{\text{r.m.s.}} = \frac{V_{\text{r.m.s.}}}{X_L} = \frac{240}{6280}\ \text{A} = 0.038\ \text{A}.$$

The maximum amount of energy stored in the inductor is

$$W = \tfrac{1}{2}LI_0^2 = LI_{\text{r.m.s.}}{}^2 = 20 \times (0.038)^2\ \text{J} = 0.029\ \text{J}.$$

When $f = 3000$ Hz:

$$X_L \propto f$$

and

$$I_{\text{r.m.s.}} = \frac{V_{\text{r.m.s.}}}{X_L}.$$

The value of $V_{\text{r.m.s.}}$ has not been changed,

$$\therefore \quad I_{\text{r.m.s.}} \propto \frac{1}{f}.$$

The new value is therefore given by

$$I_{\text{r.m.s.}} = 0.038 \times \frac{50}{3000}\ \text{A} = 0.00063\ \text{A}.$$

The energy stored $= LI_{\text{r.m.s.}}{}^2$,

$$\therefore \quad W \propto I_{\text{r.m.s.}}{}^2 \propto \frac{1}{f^2},$$

$$\therefore \quad W = 0.029 \times \left(\frac{50}{3000}\right)^2\ \text{J} = 8.1 \times 10^{-6}\ \text{J}.$$

These results assume inductance L to be constant.

9.10 Root mean square current and potential difference as vectors

At any one instant, the relations between currents and p.d.s in a network must be consistent with Kirchhoff's laws, whether the current is direct or alternating. Thus, with respect to values taken at any one selected instant of time, we have:
where conductors meet at a point (Fig. 9.11(*a*))

$$\Sigma I = 0,$$

and where components are connected in series (Fig. 9.11(*b*))

$$V = \Sigma V_n,$$

V_n being the p.d. across a single component, and V the p.d. across them all.

These simple rules, however, do not apply in the general case to the values of current or p.d. registered by meters inserted in an a.c. circuit; nor should we expect them to do so. This is because the instruments are measuring root mean square values, not instantaneous ones, and these do not necessarily combine in the

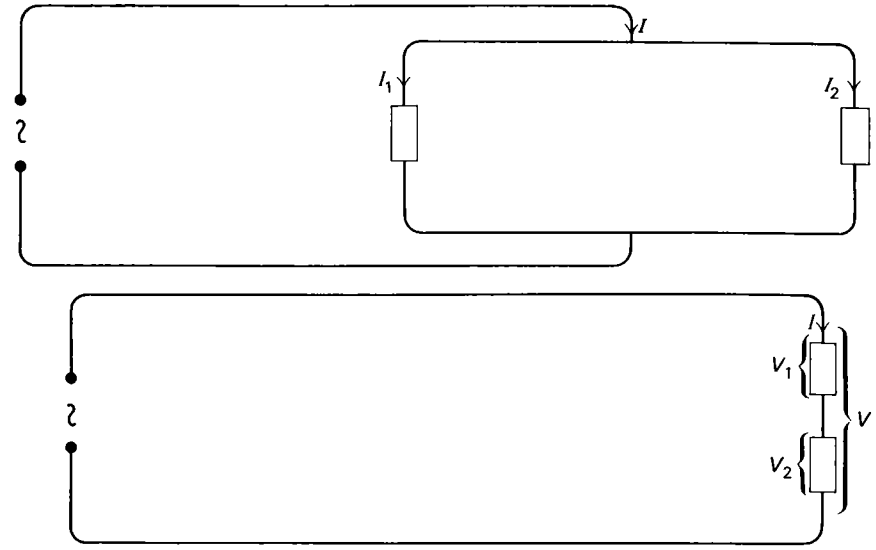


Fig. 9.11 Components in series and parallel – instantaneous values of *I* and *V* (a) (b)

same way as steady currents and p.d.s do. This is illustrated in Fig. 9.12. Suppose the waveforms of the two potential differences V_1, V_2 in the circuit of Fig. 9.11(*b*) were as shown on the graph, being only partially in phase with each other. The waveform of the resultant potential difference, V, may be found by adding together the two values V_1, V_2 at successive intervals of time, and this has been done in the diagram. The peak value, V_0 is evidently smaller than the sum of the separate peak values, $(V_1)_0$ and $(V_2)_0$, and this is because, not being in phase, these two peak values do not appear at the same instant of time. Because the a.c. voltmeter or ammeter is calibrated to register r.m.s. values, and these are proportional to the peak values, the same qualification applies to the readings of the meters.

Fortunately, the treatment of r.m.s. values is not rendered impossibly difficult

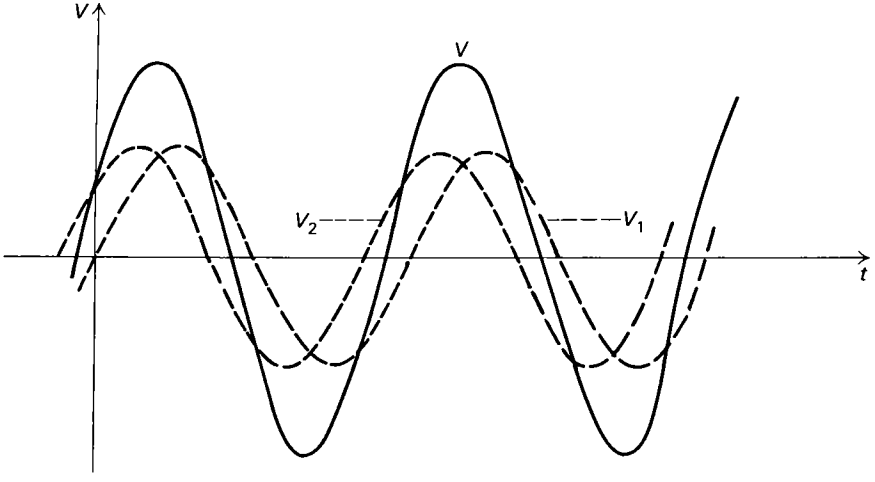


Fig. 9.12 The sum of two p.d.s which are only partially in phase

by the above complication, because they can be treated as vectors, a process in which phase angle plays a prominent part.

Phase angle ϕ

We can take as an example the relation between p.d. and current for a device through which alternating current is flowing. It may well be that the two are not completely in phase with each other, and, for the general case, we write

$$I = I_0 \sin \omega t,$$
$$V = V_0 \sin(\omega t + \phi), \tag{9.17}$$

where ϕ is the phase difference between the two. If ϕ is positive, the p.d. is ahead of the current in phase, as it reaches maximum before the current does. If $\phi = \pi/2$ rad, the p.d. is maximum at the instant when the current is zero, and this applies to the p.d. across an inductive reactance. In the case of capacitative reactance, where the p.d. lags behind the current in phase, then $\phi = -\pi/2$ rad.

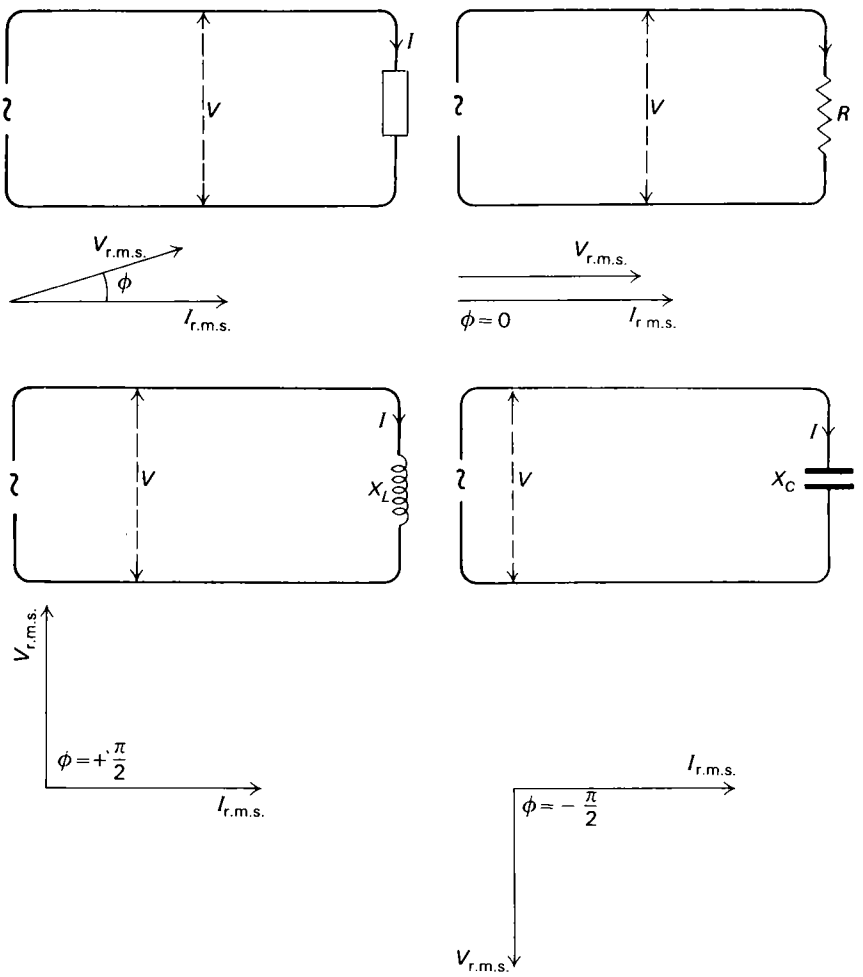


Fig. 9.13 (a) p.d. and current vectors where V leads I by ϕ (b) p.d. across a resistance $\phi = 0$ (c) p.d. across an inductive reactance $\phi = \frac{\pi}{2}$ (d) p.d. across a capacitive reactance $\phi = -\frac{\pi}{2}$

Vector representation

There is more than one method available for treating vectors, and the one which we will choose is the graphical one, identical with that applied in mechanics when drawing a parallelogram or a triangle of forces. Supposing, for example, that a p.d. $V_{\text{r.m.s.}}$ exists across a component in a circuit when a current $I_{\text{r.m.s.}}$ is flowing in it, the p.d. and the current differing in phase by an angle ϕ. These may be represented in magnitude and phase by drawing vectors, as shown in Fig. 9.13(*a*). The other examples shown in this diagram represent the three special cases, where current flows in resistance, inductance, and capacitance respectively.

9.11 Components connected in series

When two components are connected in series in an a.c. circuit, with the result that the current is common to both, this is a statement which applies to the r.m.s. value of the current also, but the two separate r.m.s. potential differences must be added by treating them as vectors. If the two p.d.s are in phase with each other, of course, this process reduces to treating them arithmetically, and this would be the case, for example, if both components were resistors. If, on the other hand, one is a resistance while the other is a reactance, so that the phase difference between the two p.d.s is $\pi/2$ rad, then the vector diagram is one or other of the two drawn in Fig. 9.14.

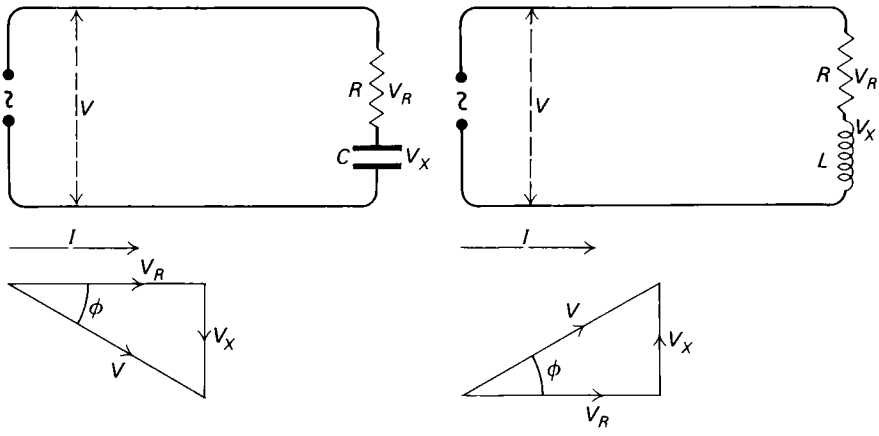


Fig. 9.14 P.d. vectors for resistance and reactance in series (a) Capacitance (b) Inductance

In both cases, we have

$$V^2 = V_R^2 + V_X^2, \tag{9.18}$$

where V, V_R, V_X are r.m.s. values, V being the resultant p.d., and

$$\tan \phi = \frac{V_X}{V_R}, \tag{9.19}$$

where ϕ is the phase angle between V and I.

9.12 Impedance Z

Because the p.d. and current in the circuit referred to above are neither in phase nor are they out of phase by as much as $\pi/2$, the ratio of p.d. V to current I cannot correctly be described either as resistance R or as reactance X, and we write

$$\frac{V_{\text{r.m.s.}}}{I_{\text{r.m.s.}}} = Z, \tag{9.20}$$

where Z is called the impedance of the circuit, or of the relevant component.

Impedance is the general term for the ratio $V_{\text{r.m.s.}}/I_{\text{r.m.s.}}$, resistance and reactance being special cases of it. If the p.d. leads the current in phase, whether $\phi = \pi/2$ or not, the impedance is inductive, and the component may be treated as resistance and inductance in series. Similarly, if the p.d. lags behind the current, the impedance is capacitative.

9.13 Vector treatment of impedance

The three p.d.s in equation 9.18 are related to the r.m.s. current by the equations

$$V = ZI,$$
$$V_R = RI,$$
$$V_X = XI.$$

Fig. 9.15 Resistance and reactance vectors for components connected in series (a) Capacitance (b) Inductive

Thus, cancelling I in equations 9.18 and 9.19, we obtain

$$Z^2 = R^2 + X^2, \tag{9.21}$$

and

$$\tan\phi = \frac{X}{R}. \tag{9.22}$$

Similarly, the current factor may be omitted in the vector diagram, giving a diagram relating resistance, reactance, and resultant impedance instead (Fig. 9.15).

9.14 The power consumption in an impedance

We have already seen that no energy is converted to heat in a reactance, a result which is sometimes expressed by the statement that the current in a reactance is wattless.

For reactance and resistance in series, the mean power P_{av}, which is used in the resistance only, is

$$\left.\begin{aligned} P_{av} &= V_R I \\ &= I^2 R \\ &= VI\frac{R}{Z} \end{aligned}\right\}, \tag{9.23}$$

because $$I = \frac{V}{Z},$$

$$P_{av} = VI\cos\phi, \tag{9.24}$$

because $$\cos\phi = \frac{R}{Z},$$

where the p.d.s and current are r.m.s. values. The ratio $\cos\phi = R/Z$ is called the power factor. The significance of this name is that the product of the r.m.s. values of p.d. and current, VI, does not, in general, give the power P as it does in the case of direct current and we have, instead, equation 9.24.

One equation, however, which is familiar in d.c. theory is still applicable without modification, viz.

$$P = I^2 R.$$

Example

A sinusoidal alternating p.d. with r.m.s. value 240 V *and frequency* 50 Hz *is applied across an inductor which has inductance* 2.0 H *and resistance* 500 Ω. *Calculate the r.m.s. value of the current, the power converted to heat in the inductor, the phase angle between current and p.d., and the power factor.*

The inductor may be treated as a resistance and inductance in series, and we have

$$Z^2 = R^2 + X^2,$$

where $$R = 500\ \Omega,$$

$$X = \omega L = 2\pi \times 50 \times 2.0\ \Omega = 628\ \Omega,$$

$$\therefore\quad Z = 800\ \Omega.$$

$$I_{r.m.s.} = \frac{V_{r.m.s.}}{Z} = \frac{240}{800}\ \text{A} = 0.30\ \text{A},$$

$$P = I^2 R = 0.09 \times 500\ \text{W} = 45\ \text{W},$$

$$\tan\phi = \frac{X}{R} = \frac{628}{500} = 1.256,$$

$$\therefore\quad \phi = 51.5^\circ = 0.286\pi\ \text{rad}.$$

$$\text{Power factor} = \frac{R}{Z} = \frac{500}{800} = 0.625.$$

9.15 Capacitance and inductance in series

When capacitance and inductance are connected in series, the two p.d.s are π rad out of phase with each other, the result being that they act in opposite directions. This is because the p.d. across the capacitance is $\pi/2$ rad behind the current in phase, whereas the p.d. across the inductance is $\pi/2$ rad ahead. In practice, there is always some resistance in the circuit also, because a practical inductor has resistance, and Fig. 9.16 is drawn for a circuit which includes all three components.

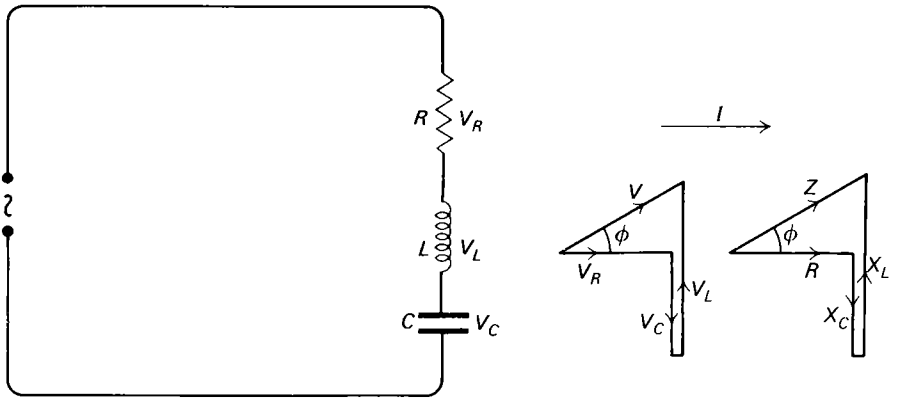


Fig. 9.16 Vector diagrams for R, L, C in series

From the diagrams, we have:

$$V^2 = V_R{}^2 + (V_L \sim V_C)^2,$$
$$Z^2 = R^2 + (X_L \sim X_C)^2, \tag{9.25}$$
$$\tan\phi = \frac{(X_L \approx X_C)}{R}.$$

Example

In the circuit of Fig. 9.16, *the p.d. V of the a.c. source has the r.m.s. value* 20 V, *the frequency being* 50 Hz. *The values of* R, L, C *are respectively* 30 Ω, 3.5 H, *and* 3.0 μF, *where R includes the resistance of the inductor winding.*

Deduce the peak value of the p.d. across the capacitor. Find also the peak value of the p.d. across the inductor, assuming its resistance to be 10 Ω.

We have:

$$X_L = \omega L = 350\pi\ \Omega = 1\,100\ \Omega,$$

$$X_C = \frac{1}{\omega C} = \frac{10^6}{300\pi}\ \Omega = 1\,060\ \Omega,$$

$$X_L - X_C = 40\ \Omega.$$

$$\therefore\quad Z^2 = R^2 + (X_L - X_C)^2 = 900 + 1\,600\ \Omega^2 = 2\,500\ \Omega^2,$$

$$\therefore\quad Z = 50\ \Omega,$$

$$\therefore\quad I = \frac{V}{Z} = \frac{20}{50}\ \text{A} = 0.40\ \text{A},$$

where this is the r.m.s. value.

The r.m.s. value of the p.d. V_C across the capacitor is

$$V_C = X_C I = 1\,060 \times 0.40\ \text{V} = 424\ \text{V},$$

and the peak value is $(V_C)_0$

where $(V_C)_0 = 424 \times 2^{\frac{1}{2}}\ \text{V} = 600\ \text{V}.$

Also, the impedance of the inductor is Z_L where

$$Z_L{}^2 = R_L{}^2 + X_L{}^2 = (100 + 121) \times 10^4\ \Omega = 1\,100\ \Omega, \text{ very nearly.}$$

The r.m.s. value of the p.d. V_Z across the inductor is therefore

$$V_Z = Z_L I = 1\,100 \times 0.40 \text{ V} = 440 \text{ V},$$

and the peak value is $(V_Z)_0$

where $(V_Z)_0 = 440 \times 2^{\frac{1}{2}} \text{ V} = 622 \text{ V}$.

These results may well be thought remarkable. The peak value of the p.d. applied across the circuit is only 28 V, being $20 \times \sqrt{2}$, yet the p.d. across the capacitor and that across the inductor reach much higher values than this. The explanation is that these two p.d.s are always in opposition to each other, since they differ in phase by the angle $\phi = \pi$, very nearly. Thus only the difference between V_L and V_C contributes to the p.d. across the complete circuit at any instant.

A mechanical analogy to this might be a massive ball attached to a vertical spring. If the top end of the spring is held in the hand, and vibrated up and down, quite a slight movement of the hand can cause large amplitude oscillations of the ball, and consequently give rise to large forces between spring and ball. The ball, in this case, corresponds to the inductor, and the spring to the capacitor. To obtain this effect, however, the frequency must be appropriately chosen.

9.16 Series resonance – the acceptor circuit

What we have been discussing in the last example is an effect of electrical resonance. The inductor and capacitor in series make up a circuit which is capable of oscillation on its own account (sections 8.7 and 8.21) and the frequency of free oscillation is

$$f_0 = \frac{1}{2\pi\sqrt{(LC)}}. \tag{9.26}$$

If an alternating p.d. having frequency f is applied across this circuit, it causes a flow of current which also has this frequency f and not the frequency f_0 which is natural to the circuit; that is, a forced electrical oscillation occurs. The values chosen for the components in the last example were such that the natural frequency was very nearly equal to 50 Hz, viz:

$$L = 3.5 \text{ H},$$
$$C = 3.0\,\mu\text{F},$$
$$\therefore \quad f_0 = \frac{1}{2\pi\sqrt{(10.5 \times 10^{-6})}} \text{ Hz}$$
$$= 49 \text{ Hz}.$$

In this circumstance there was resonance, and the oscillation in p.d. across capacitor and inductor built up to a very large size. This response can be deduced from equation 9.25. Thus, the impedance Z of the circuit is given by

$$Z^2 = R^2 + (X_L - X_C)^2 \tag{9.25}$$

and this is minimum, i.e. resonance occurs, when

$$X_L - X_C = 0,$$

i.e.

$$2\pi f L = \frac{1}{2\pi f C},$$

whence

$$f = \frac{1}{2\pi\sqrt{(LC)}},$$

Plotting impedance against frequency gives a graph with the form shown in Fig. 9.17.

Because this series circuit accepts current at its natural frequency more readily than at any other, it is termed an acceptor circuit for the frequency f_0.

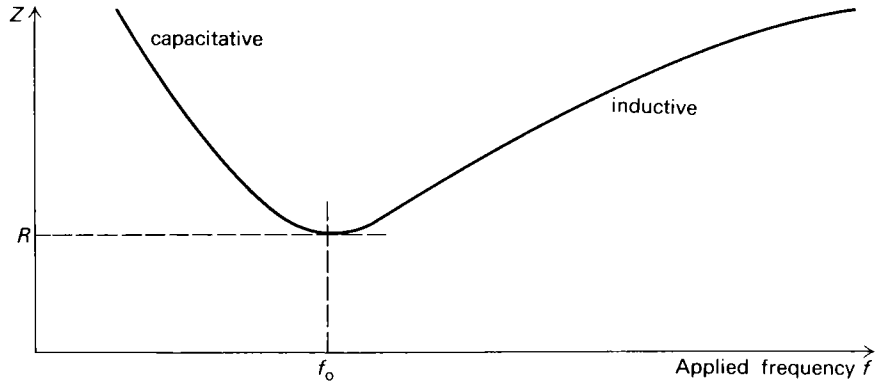


Fig. 9.17 Capacitor and inductor in series – the variation of impedance with frequency

9.17 Parallel resonance – the rejector circuit

If an alternating p.d. is applied across an inductor and a capacitor which are connected in parallel, the arrangement behaves as a rejector circuit for the frequency f_0, i.e. the impedance is maximum at this frequency, and not minimum, as it is with the series circuit. In this case, because the components are in parallel, it is the p.d. and not the current which they have in common, and current in the inductor repeatedly charges the capacitor, first with one polarity, then with the other, requiring only a small contribution from the source of the p.d.

The parallel resonance circuit is what is used as the tuned input circuit of a radio receiver.

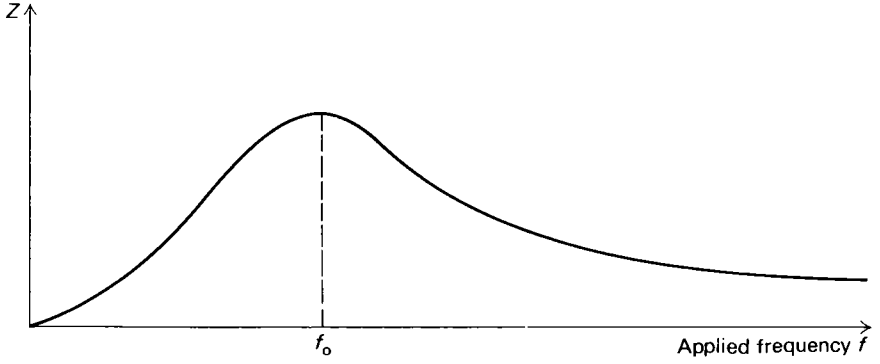


Fig. 9.18 Capacitor and inductor in parallel – the variation of impedance with frequency

9.18 Coupling and decoupling capacitors

Much of the importance of the capacitor in the circuits of electronics depends on the fact that its reactance is frequency dependent. It is frequently the case in these

circuits that a d.c. potential becomes superimposed on an alternating one, or a low frequency potential is mixed with a high frequency one, and it is necessary to separate the two out again. This is the function of a coupling and of a decoupling capacitor. Several examples of these two applications will appear in chapter 12. In the two circuits of Fig. 9.19 d.c. and a.c. are represented as being superimposed upon each other by connecting a transformer in series with a battery.

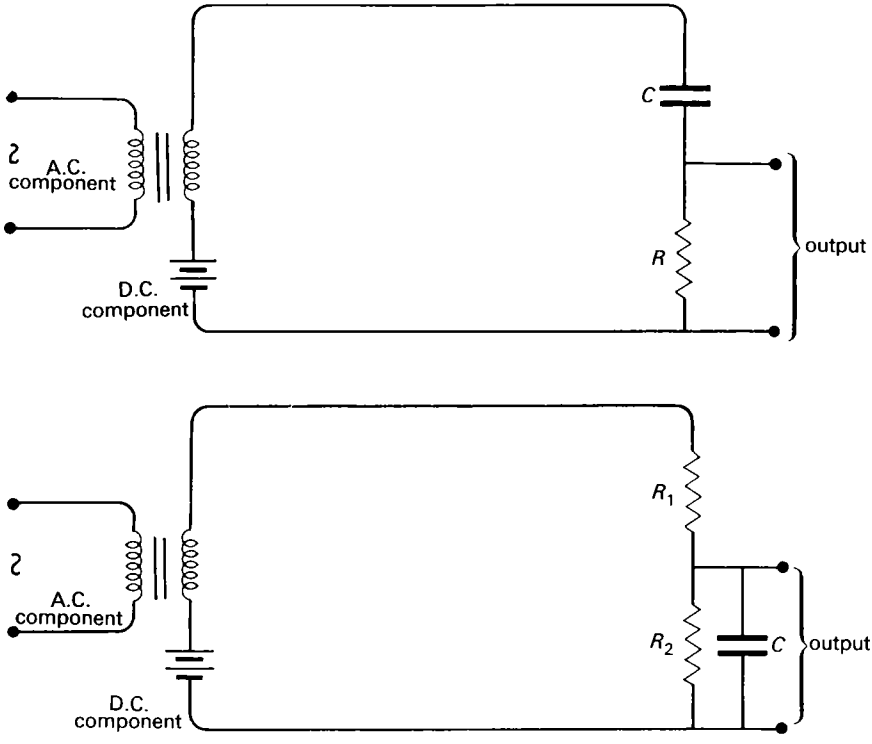


Fig. 9.19 Coupling and decoupling capacitors (a) A coupling capacitor (b) A decoupling capacitor

In the first circuit, the capacitor will charge up until the whole of the d.c. component exists as a p.d. across it, after which no further direct current flows, and the direct p.d. across the resistor is zero. Part of the alternating p.d. is across R, however, and can be applied across the input of a further circuit without any d.c. component remaining. In this way the capacitor is being used to pass on the alternating potential, and is termed a coupling capacitor. It will clearly perform this function more effectively if its reactance is low compared with the value of the resistance R, so that the greater part of the alternating p.d. shall appear across the resistor,

i.e. $$\frac{1}{2\pi fC} \ll R.$$

The second circuit in the diagram illustrates the function of a decoupling capacitor, the purpose being to apply a proportion of the direct p.d. from the

source across the resistor R_2, while the alternating component is removed by effectively short circuiting it with the comparatively low reactance capacitor C. Hence the required condition in this case, also, is

$$\frac{1}{2\pi fC} \ll R_2.$$

Provided that this condition holds, nearly the whole of the alternating p.d. is taken up by the resistor R_1, leaving a proportion of the direct p.d. only to appear across R_2. The capacitor is said to have decoupled the alternating component.

9.19 The capacitance bridge

When the equation for a capacitance bridge was derived in section 8.16, a method was adopted which did not require a.c. theory. A more satisfactory treatment can now be given in terms of reactance.

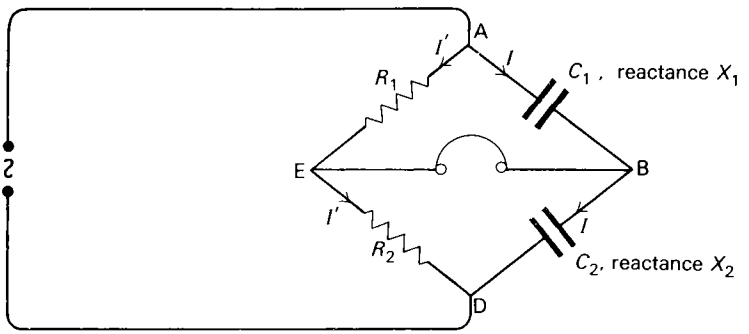


Fig. 9.20 The capacitance bridge

Assuming the bridge to be balanced, i.e. that no current flows in the telephone, the r.m.s. currents are shown in the diagram. We have:

$$V_{AB} = V_{AE}$$

and $$V_{BD} = V_{ED},$$

because $$V_{BE} = 0.$$

Therefore $$X_1 I = R_1 I'$$

and $$X_2 I = R_2 I'.$$

$$\frac{X_1}{X_2} = \frac{R_1}{R_2},$$

where $$X_1 = \frac{1}{\omega C_1}$$

and $$X_2 = \frac{1}{\omega C_2}.$$

Therefore $$\frac{C_2}{C_1} = \frac{R_1}{R_2}, \qquad (9.27)$$

which is the equation which was derived in section 8.16.

Phasing condition

In general, it is not sufficient to balance the r.m.s. values of the p.d.s. The p.d. across BE will be zero only if the p.d.s across AB and AE are not only equal in magnitude, but also in phase with each other. In the simple case of the capacitance bridge, where two branches are purely resistive and the other two are purely reactive, the second condition is automatically fulfilled in satisfying the first, but in the general case a second adjustment would be required to deal with the phase. An inductance bridge, for example, is more complicated on this account, owing to the resistance component included in every practical form of inductor.

9.20 A mathematical treatment of impedance

A formal treatment of elementary a.c. theory was neglected earlier in this chapter in favour of discussing the basic physics. A mathematical approach is not difficult, however, and resistance, inductance, and capacitance will now be discussed in turn in this way. The starting point in each case is to set out two equations, one of which is the equation either for the sinusoidal current or for the sinusoidal p.d., and the other is the basic equation which relates instantaneous values of current and p.d. for the component under discussion; e.g. if this is a resistor, the basic equation is $V = RI$.

Resistance

The starting equations are:

$$I = I_0 \sin \omega t, \tag{9.29}$$

where $\omega = 2\pi f$,

and

$$V = RI. \tag{9.30}$$

whence, eliminating I,

$$V = RI_0 \sin \omega t,$$

i.e.

$$V = V_0 \sin \omega t,$$

where

$$V_0 = RI_0.$$

Interpreting this result, we have:

(*a*) V and I are in phase

(*b*)

$$Z = \frac{V_{\text{r.m.s.}}}{I_{\text{r.m.s.}}} = \frac{V_0}{I_0} = R.$$

(*c*) The value of the impedance Z is therefore independent of frequency, unless the resistance R varies with frequency.

(*d*) The mean power $= P_{\text{av}}$, where

$$\begin{aligned} P &= VI \\ &= V_0 I_0 \sin^2 \omega t \\ &= \tfrac{1}{2} V_0 I_0 (1 - \cos 2\omega t) \\ \therefore \quad P_{\text{av}} &= \tfrac{1}{2} V_0 I_0 \\ &= V_{\text{r.m.s.}} I_{\text{r.m.s.}} \\ &= I_{\text{r.m.s.}}{}^2 R. \end{aligned}$$

Inductance

The starting equations are:

$$I = I_0 \sin \omega t \tag{9.29}$$

and

$$V = L\frac{dI}{dt}, \tag{9.31}$$

whence $V = \omega L I_0 \cos \omega t,$

i.e. $V = V_0 \sin(\omega t + \phi),$

where $V_0 = \omega L I_0$

and $\phi = \frac{\pi}{2}$ rad.

Interpreting this result, we have:

(*a*) V and I differ in phase by $\pi/2$ rad, where V is ahead of I.

(*b*)

$$Z = \frac{V_{\text{r.m.s.}}}{I_{\text{r.m.s.}}} = \frac{V_0}{I_0} = \omega L.$$

(*c*) The value of the impedance Z is therefore directly proportional to the frequency f.

(*d*) The mean power = P_{av}, where

$$P = VI = V_0 I_0 \cos \omega t \sin \omega t = \tfrac{1}{2} V_0 I_0 \sin 2\omega t,$$

$$\therefore \quad P_{\text{av}} = 0,$$

because the average value of $\sin 2\omega t$ over a period of time involving complete alternating cycles is zero.

For this reason, the impedance of the inductance is termed its reactance X, being a special form of impedance which converts no power to heat.

Capacitance

Applying a similar treatment to capacitance, the starting equations would be:

$$I = I_0 \sin \omega t \tag{9.29}$$

and

$$V = \frac{Q}{C}. \tag{9.32}$$

If we were now to attempt to substitute for Q in the second of these equations, however, this would involve an integration because $Q = \int I dt$. In these circumstances it is easier to start, instead, by assuming the sinusoidal equation for p.d. rather than that for current, viz.

$$V = V_0 \sin \omega t. \tag{9.33}$$

Also, differentiating equation 9.32, and transposing, we have

$$I = C\frac{dV}{dt}. \tag{9.34}$$

These equations have exactly the same form as the starting equations 9.29 and 9.31 for dealing with inductance, except that:

I and V are interchanged

and C replaces L.

The conclusions reached must therefore be the same, subject to these substitutions, viz.:

$$I = I_0 \cos \omega t$$
$$= I_0 \sin(\omega t + \phi),$$

where $I_0 = \omega C V_0$

and $\phi = \frac{\pi}{2}$ rad.

Interpreting this result, we have:

(*a*) V and I differ in phase by $\pi/2$ rad, where I is ahead of V.

(*b*)
$$Z = \frac{V_{\text{r.m.s.}}}{I_{\text{r.m.s.}}}$$
$$= \frac{V_0}{I_0}$$
$$= \frac{1}{\omega C}.$$

(*c*) The value of the impedance Z is in inverse proportion to frequency f.

(*d*) The mean power $= P_{\text{av}}$ and $P_{\text{av}} = 0$.

The impedance Z of a capacitance is therefore termed its reactance X.

*9.21 The vector treatment of r.m.s. values

The addition of r.m.s. values of current or p.d. by treating them as vectors may be justified in the following way.

Supposing that two components in a circuit are connected in series, and the p.d.s across them are respectively given by

$$V_1 = (V_1)_0 \sin \omega t$$

and $$V_2 = (V_2)_0 \sin(\omega t + \phi),$$

so that the instantaneous value V of the p.d. across the two at the instant t is given by

$$V = V_1 + V_2$$
$$= (V_1)_0 \sin \omega t + (V_2)_0 \sin(\omega t + \phi).$$

This expression can, of course, be simplified by trigonometrical methods, and the vector diagrams drawn in Fig. 9.21 are little more than diagrams illustrating this process.

Thus, the line OA_1 in Fig. 9.21 (*a*), is a rotating vector representing the peak p.d. value $(V_1)_0$. It is supposed to be rotating about the point O with angular velocity ω, the zero of time being taken as an instant when OA_1 is directed along the x axis.

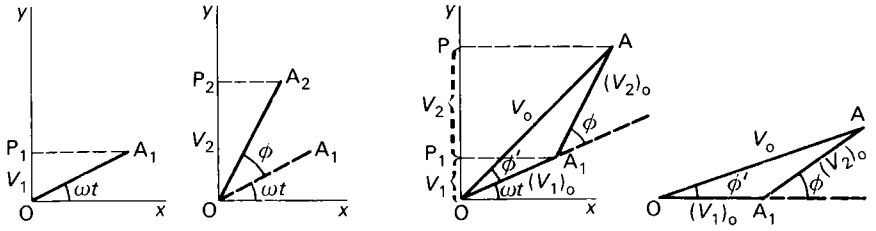


Fig. 9.21 Vector addition of two p.d.s (a) $V_1 = (V_1)_0 \sin \omega t$ (b) $V_2 = (V_2)_0 \sin (\omega t + \phi)$ (c) $V = V_1 + V_2$ (d) Vector addition of peak values

At a time t, therefore, OA_1 makes an angle ωt with the x axis, as shown in the diagram, and its projection OP_1 on the y axis is given by

$$OP_1 = OA_1 \sin \omega t.$$

Hence, since the length OA_1 is drawn to represent the peak value,

i.e. $$OA_1 = (V_1)_0,$$

it follows that the instantaneous value V_1 at time t is represented by the length OP_1.

Similarly, the line OA_2 in Fig. 9.21(*b*) is a rotating vector drawn to represent the peak value $(V_2)_0$. At the instant when time t has been taken to be zero, OA_2 makes an angle ϕ with the x axis. The projection OP_2 on the y axis at time t represents the instantaneous value V_2, because

$$\begin{aligned} OP_2 &= OA_2 \sin(\omega t + \phi) \\ &= (V_2)_0 \sin(\omega t + \phi) \\ &= V_2. \end{aligned}$$

Because the frequency is the same for both p.d.s the two vectors OA_1, OA_2, at all times make angle ϕ with each other.

The two vector diagrams of Fig. 9.21(*a*) and (*b*) are shown superimposed in the third diagram, and it is evident that the projection OP on the y axis now gives the instantaneous value of the resultant potential difference V at the instant t, for we have

$$\begin{aligned} OP &= (V_1)_0 \sin \omega t + (V_2)_0 \sin(\omega t + \phi) \\ &= V_1 + V_2 \\ &= V. \end{aligned}$$

Also, the rotating vector OA gives the peak value $(V)_0$ of the resultant p.d., and the angle ϕ' gives the phase difference between V and V_1.

If all we want to know is the peak value and the phase, much of this diagram becomes irrelevant to our purpose, and all that is required is the vector triangle drawn in Fig. 9.21(*d*).

To add r.m.s. values, a diagram may be drawn which is similar to this one, putting in r.m.s. values in place of peak values. This is justified by the fact that peak and r.m.s. values are proportional to each other.

QUESTIONS

9.1 An alternating p.d. of frequency 50 Hz has sinusoidal waveform, and peak value 28.3 V.

(*a*) Write down an expression giving the value of the p.d. at any instant t, assuming that $V = 0$ when $t = 0$.

(*b*) What is the r.m.s. value of the p.d.?

(*c*) If the p.d. is applied across a 5 Ω resistor, what is the mean rate of converting energy into heat in the resistor?

(*d*) Find the r.m.s. values of the current which would flow when the p.d. is applied across (i) a 2 μF capacitor (ii) a 50 μF capacitor, (iii) a 0.5 H inductor with negligible resistance.

9.2 The amount of magnetic flux Φ which threads a coil of wire varies sinusoidally with frequency 500 Hz. If the peak value of the flux is 6×10^{-2} Wb, and the coil has 400 turns write down an expression giving the amount of flux Φ which threads the coil at time t, assuming that its value is zero when $t = 0$.

Using the law of electromagnetic induction, stated in the form $V = N d\Phi/dt$, obtain an expression for the e.m.f. induced in the coil, and deduce its r.m.s. value.

If the coil has resistance 10 Ω, and its ends are joined by a resistor of resistance 30 Ω, calculate the mean rate at which energy is converted to heat in this resistor.

9.3 Explain what is meant by peak value and root mean square value of an alternating current. Establish the relation between these quantities for a sinusoidal waveform.

What is the r.m.s. value of the alternating current which must pass through a resistor immersed in oil in a calorimeter so that the initial rate of rise of temperature of the oil is three times that produced when a direct current of 2.0 A passes through the resistor under the same conditions? (N)

9.4 A 'saw tooth' voltage rises uniformly with time from zero to a maximum value V in time t. It then drops instantaneously to zero and then rises again as before, the process being continued indefinitely. What voltage would a moving coil voltmeter read? What quantity of electricity would such a voltage send through a pure resistance R in unit time? (S)

9.5 A 2.0 V accumulator of negligible internal resistance is connected in series with a switch and a 5.0 Ω resistance wire. The switch is closed for 0.10 s, opened for 0.30 s, and the process steadily repeated.

(*a*) Calculate the r.m.s. value of the current.

(*b*) How many joules of heat would be produced in the wire in 12 s? (C)

9.6 Two capacitors, one of capacitance 2.00 μF and the other of capacitance 4.00 μF, are charged to p.d.s of 100 V and 50 V respectively. The positive plates of the capacitors are now connected together by a wire of negligible resistance, and the negative plates are connected through a resistor of 50 Ω.

If the flow of charge may be considered to have ceased after 0.50 s, calculate the r.m.s. value of the current which has passed (taking it over this period). (L)

9.7 The following readings of the current through a certain device, and the p.d. applied across it, were taken.

I/A	1.25	0.75	0.25	0	0	0	0
V/V	50	30	10	0	−10	−30	−50

The device is now connected in series with an 80 Ω resistor and a 30 V accumulator. By means of a commutator, the connections to the accumulator are regularly reversed 6.0 times per second. Draw a graph showing the variation of the current in the circuit with time over a period of one second, and calculate the average value of the rate at which heat is developed in the resistor. (C)

9.8 An electric fan-heater is marked 2 kW, 50 Hz, 240 V, a.c. only. The current I, taken by the device varies with time according to the equation

$$I = I_0 \sin \omega t.$$

Find the values of: I_0, ω, and the maximum instantaneous power consumed by the device. (Assume that the fan behaves as a pure resistor.)

Comment on the following statements made concerning the heater:

(*a*) 'It has a fan as well, so it must give much less heat than a 2 kW radiator.'

(*b*) 'It says 240 V, so a cable which will stand 250 V is good enough.'

(*c*) 'It is just a heater, so it must work on d.c. as well as a.c.'

(*d*) 'I never bother to connect the third wire.' (C)

9.9 A flat coil is mounted so that it can be rotated about an axis in its own plane, and is placed in a magnetic field so that its axis of rotation is perpendicular to the magnetic flux. The coil is now rotated at the constant speed of 5 rev/s. If the area enclosed by the coil, which has 500 turns, is 40 cm^2, and the flux density of the magnetic field is 0.06 T, deduce:

(*a*) the total flux Φ which threads the coil at the instant when the plane of the coil makes an angle θ with the field;

(*b*) the e.m.f. V induced in the coil at this instant, where, by Faraday's law of electromagnetic induction, $V = N\mathrm{d}\Phi/\mathrm{d}t$;

(*c*) the r.m.s. value of the e.m.f. induced in the coil;

(*d*) the r.m.s. value of the current which flows in a 10 Ω resistor connected across the coil (take the resistance of the coil to be 3.3 Ω, and its inductance to be negligible);

(*e*) the mean power converted to heat in the 10 Ω resistor.

9.10 An alternating p.d., with r.m.s. value 0.12 V, is applied across a capacitance of 6.3 μF. Calculate the r.m.s. current which flows in the capacitor when the frequency is (*a*) 50 Hz, (*b*) 3000 Hz, (*c*) 20 kHz.

What is the maximum amount of energy stored in the capacitor in the course of each cycle?

9.11 An alternating current, r.m.s. value 0.30 A, flows in a 0.05 H inductor, the resistance of which may be neglected. Calculate the r.m.s. value of the p.d. which has been applied across the inductor when the frequency is (*a*) 50 Hz, (*b*) 3000 Hz.

What is the maximum amount of energy stored in the inductor in the course of each cycle?

9.12 Explain how a sinusoidal alternating voltage may be generated by a rotating coil, and obtain an expression for the voltage developed at time t.

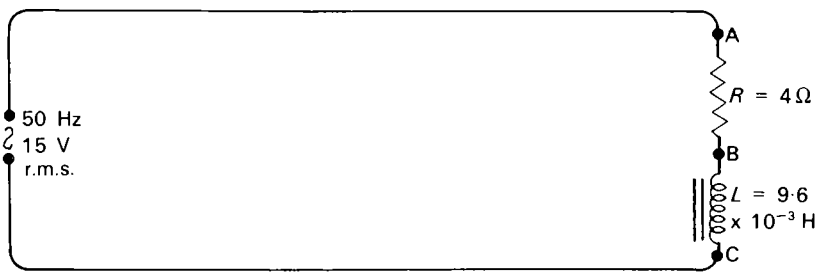


Fig. e 65

For the circuit of the figure, carrying alternating current of 50 Hz, calculate, correct to two significant figures:

(*a*) the reactance of the inductor L,

(*b*) the impedance of the circuit between A and C,

(*c*) the r.m.s. value of the circulating current,

(*d*) the r.m.s. p.d. between A and B and between B and C,

(*e*) the power dissipated in the circuit,

(*f*) the phase difference between the current and the supply voltage. (O)

9.13 A capacitor with capacitance 2.0×10^{-4} μF is connected in series with a resistor and a source of alternating e.m.f. whose frequency is $10^6/2\pi$ Hz. The p.d.s across resistor and capacitor respectively are found to be equal. Calculate the value of the resistor.

If the r.m.s. current through the circuit is 1/7 mA, what are (*a*) the r.m.s value of the applied p.d., (*b*) the p.d. across the resistor? (O & C)

9.14 To produce a powerful alternating magnetic field, a large coil is connected to the a.c. supply mains, of 240 V r.m.s. at 50 Hz. An ammeter in series with the coil registers 10 A and after 8 h of continuous operation the electric supply meter indicates a consumption of 4.8 kWh of energy. Assuming that all the energy is accounted for by resistance losses in the coil, calculate its resistance.

Calculate the reactance of the coil, and hence its inductance. (O & C)

9.15 A resistance, 1000 Ω, and a capacitance, 2×10^{-6} F, are connected in series across the 200 V, 50 Hz supply. Calculate the impedance of this combination, and find the p.d. across (*a*) the resistor, (*b*) the capacitor, (*c*) the supply, at the instant when the current has its maximum value. (O & C)

9.16 An alternating p.d. V is applied across a capacitor C and a resistor $R = 5000\ \Omega$ in series. If the p.d. across the resistor is 0.8 V when the frequency is 10 kHz, what is the capacitance of C? $V = 1{\cdot}0$V

What would be the p.d. across the resistor if the frequency were reduced to 1 kHz, the applied p.d. V retaining the same value?

9.17 A p.d. is applied across a resistor R and a capacitor C which are connected in series. The applied p.d. consists of a 50 Hz alternating ripple superimposed on a d.c. component of 200 V, the peak value of the ripple being 10 V.

Sketch p.d. – time graphs with reference to (*a*) the applied p.d., (*b*) the p.d. across the capacitor, (*c*) the p.d. across the resistor, given that $R = 1000\ \Omega$, $C = 20\ \mu$F.

Label approximate values of V on your graphs.

9.18 A 50 Hz sinusoidal alternating p.d. has peak value 340 V. Taking the p.d. to be zero at the instant when time $t = 0$, calculate the value of the p.d. at each of the following instants:

$$t = 5 \times 10^{-3}\ \text{s}, \quad 3.33 \times 10^{-3}\ \text{s}, \quad 2.5 \times 10^{-3}\ \text{s}, \quad 1.66 \times 10^{-3}\ \text{s}.$$

9.19 A neon lamp lights when the p.d. across it exceeds 170 V and goes out when the p.d. falls to 140 V. If the lamp is connected to 240 V r.m.s., 50 Hz mains, for what fraction of time is it lit up?

9.20 The circuit is connected to an a.c. source of frequency 50 Hz, and a current of 2.0 A flows. R is a resistor of resistance 100 Ω, L is an inductor of reactance 80 Ω and resistance 60 Ω, and C is a capacitor of reactance 150 Ω (all at 50 Hz).

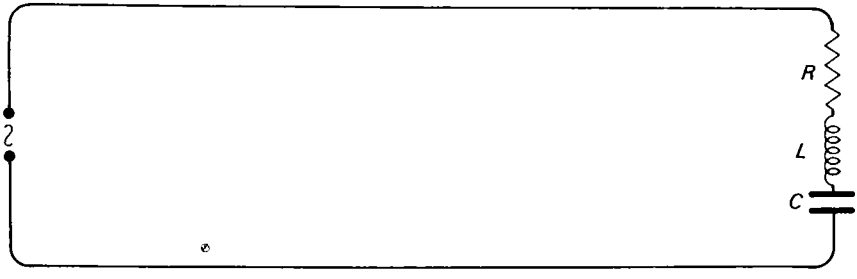


Fig. e 66

Calculate:

(*a*) the impedance of L,

(*b*) the p.d. across each component,

(*c*) the supply voltage,

(*d*) the power factor of the complete circuit,

(*e*) the inductance of L,

(*f*) the capacitance of C.

(*g*) the phase relationships between the supply voltage and that across each of the three components, and represent them in a vector diagram. (L)

9.21

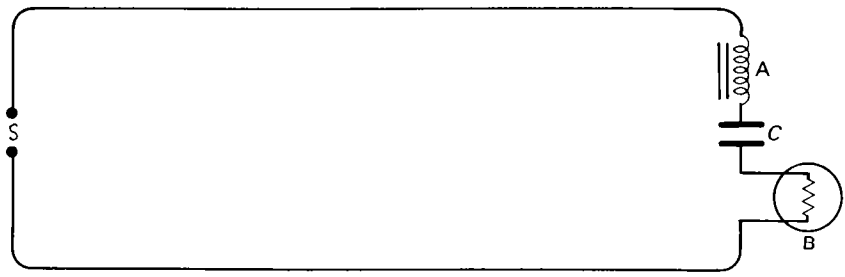


Fig. e 67

In the circuit shown, A is a coil of self inductance 2.0 H and resistance 500 Ω, C is a capacitor of capacitance 0.25 μF. B is a filament lamp, and S is a constant voltage a.c. supply of negligible impedance and continuously variable frequency. Explain why, as the frequency is continuously increased from a low value, the brightness of the lamp increases, reaches a maximum, and then decreases.

The resistance of the lamp is 50 Ω when it is glowing with maximum brightness and the r.m.s. current is then 0.30 A. Calculate for these conditions (*a*) the frequency of the supply, (*b*) the supply voltage, (*c*) the phase angle between the p.d.s across A and C. (L)

9.22 A non-inductive resistor of resistance 20 Ω and a coil are connected in series across a source of p.d. 8.0 V r.m.s. alternating at 50 Hz. The r.m.s. p.d.s. across the resistor and coil are 4.0 V and 6.0 V respectively. Find the current in the circuit and, by calculation or by drawing a vector diagram on graph paper, determine the resistance and inductance of the coil. (N)

9.23 A choke, of self inductance L and resistance R_1, is connected in series with a resistance R_2, and connected across the 230 V, 50 Hz mains. A very high resistance a.c. voltmeter shows a p.d. of 125 V across R_2, and 150 V across the choke. Comment on these results.

Given that $R_2 = 250\ \Omega$, deduce the values of L and R_1. (O & C)

*9.24 A 50 Hz alternating current of magnitude 10 A r.m.s. passes along a copper wire of diameter 1 mm. If copper contains 9×10^{28} free electrons per m^3, what is the amplitude of the displacement of the electrons in the copper associated with the current? (OXF. SCH.)

*9.25 A coil of self inductance L is connected to a resistance R. A magnetic flux Φ threads the coil and varies sinusoidally with angular frequency ω (i.e. $\omega = 2\pi f$). Show that, if ω is large enough, the mean power dissipated in R is independent of ω, and find this mean value.

A bicycle dynamo consists of a permanent magnet which rotates relative to a fixed coil, thereby providing an alternating flux. Indicate graphically how you would expect the brightness of the lamp to vary with the bicycle's speed. (OXF. SCH.)

*9.26 Define root mean square value of a current, and explain why r.m.s. values are used on the instruments of electricity supply authorities.

Show that if a current increases linearly with time from 0 to 20 mA in 10^{-2} s, then falls instantly to 0, and repeats this 100 times a second, the r.m.s. value is $20/\sqrt{3}$ mA.

What would the r.m.s. value be if, instead of rising from 0 to 20, the rise was −10 to +10, with a sudden fall to −10? What would be the r.m.s. value if the current increased linearly from −10 to +10 in 10^{-2} s, fell linearly from 10 to −10 in the next 10^{-2} s, and so on?

What current reading would be given for the current originally referred to (0 to 20 mA in 10^{-2} s) by (*a*) a moving coil galvanometer, (*b*) a current balance in which the force between one coil and other coils, all carrying the current, was balanced by weights? Explain how you arrive at these results.

Give three different causes for non-linearity in the scale of an accurately calibrated moving iron ammeter of the repulsion type. (S)

*9.27 A periodic current has a rectangular waveform which comprises a constant value of 1 A for 10^{-3} s followed immediately by the constant value −1 A for the next 10^{-3} s, and so on. What is its r.m.s. value? If this current flows in a resistance of 1 Ω, find the mean and maximum values of the power dissipated. Draw the waveform of the voltage required to make the same current flow in a pure capacitor of 1 μF, and show how the stored energy varies. (OXF. SCH.)

*9.28 If the p.d. V between the plates of a capacitor is changing, show that current flowing into it is

$$I = C\frac{\mathrm{d}V}{\mathrm{d}t}.$$

A periodic p.d. is applied across a capacitor and has a sawtooth waveform which rises from $-V_0$ to $+V_0$ at a constant rate in a time t_1, and falls again to $-V_0$, also at a constant rate, in a time t_2.

Find (*a*) the waveform of the current through the capacitor, (*b*) the r.m.s. values of the p.d. and of the current. (OXF. SCH.)

*9.29 A parallel plate capacitor, with plate separation 2 mm, is charged to a p.d. of 100 V and isolated. One plate is then made to execute S.H.M. of amplitude 0.05 mm about its original position so that it remains parallel to the fixed plate but moves in a direction perpendicular to its plane. Show that a sinusoidal alternating p.d. appears superimposed on the d.c. p.d. across the capacitor, and find its r.m.s. value.

The capacitor, instead of being charged and then isolated, is permanently connected to the 100 V supply through a very high resistance. Explain qualitatively what happens in this case when one plate vibrates. (O & C)

*9.30 An a.c. generator of negligible internal impedance whose output p.d. is 100 V r.m.s. is connected in series with a resistance of 100 Ω and a pure inductance of 2 H. If an a.c. voltmeter gives the same reading whether connected across the resistance or the inductance, what does it read? What is the frequency of the generator?

Calculate the power dissipated in the resistor at the instant when the current is half its maximum value. Calculate also the power supplied to the inductance at the same instant. Show that the sum of these two quantities is equal to the power delivered by the generator at this instant. (O & C)

*9.31 A variable capacitor C is connected in series with a resistor R across an alternating voltage supply of frequency f and peak value V_0. C is adjusted until the peak energy stored by the capacitor is a maximum.

What are then the peak values of the p.d. across (*a*) the capacitor, (*b*) the resistor, and what is the phase difference between the current and the supply p.d.?

If $R = 1000\ \Omega$, $V_0 = 100$ V, $f = 50$ Hz, and C is adjusted as above, how many electrons flow on to one plate of the capacitor as its potential changes from its most positive value to its most negative? (O & C)

Chapter 10

AC demonstrations—the cathode ray oscilloscope

10.1 The principle of the cathode ray oscilloscope

Many of the results deduced in the last chapter can be demonstrated or measured by a cathode ray oscilloscope because this is, in effect, a voltmeter having very low inertia, particularly suited to following periodic variations in p.d.

The cathode ray tube used in an oscilloscope is an evacuated glass tube, having the familiar conical shape shown in Fig. 10.1, in which a narrow beam of high

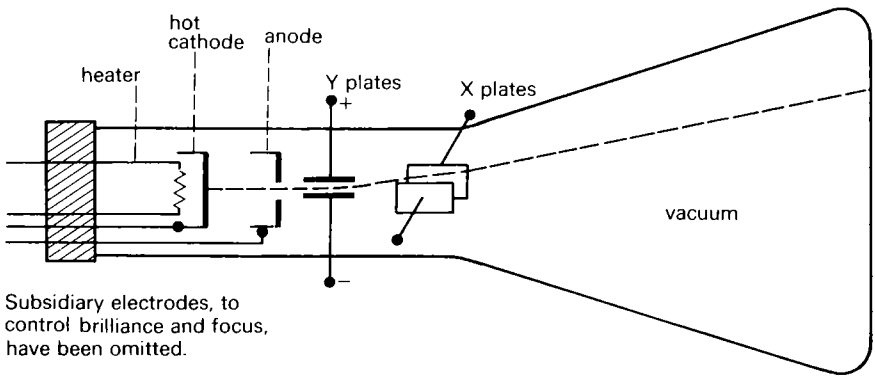


Fig. 10.1 A cathode-ray tube

speed electrons is launched from a system of electrodes, situated in the neck of the tube and known as the electron gun. These electrons finally strike the wide face of the tube, which is coated with fluorescent material, producing a bright spot of light to mark the point at which they arrive.

Just after emerging from the electron gun, the electron beam passes between two pairs of metal plates in succession, which are set at right angles to each other. These are the deflector plates. If a p.d. is applied across a pair of deflector plates, thus setting up an electric field in the space which separates them, the electron beam is deflected by the field towards the positively charged plate, and the spot on the tube face is displaced by an amount which is proportional to the applied p.d. Thus the deflection of the spot can be used to serve the same purpose as the deflection of a voltmeter needle – to measure the p.d. which has been applied. The deflected system, being a beam of electrons, has very little inertia, with the result that the movement of the spot will faithfully follow oscillations of the applied p.d., even up to radio frequencies.

If a second p.d. is now put across the other pair of deflector plates, the spot is

again deflected, in a direction perpendicular to the previous one, and the final position of the spot registers the resultant effect of the two p.d.s in the manner of a point plotted on a graph, having two perpendicular displacements from the origin. The pairs of plates are usually set to give deflections respectively in the horizontal and the vertical, and are called, in consequence, the X plates and the Y plates.

All the deflector plates are connected to earth through high value resistors to provide a leakage path for the electrons, should any of these chance to accumulate on the plates. A useful consequence of this is that a p.d. may be produced between a pair of plates by applying the p.d. between one plate and earth, instead of directly across the two plates, and this is quite a common procedure.

A variable d.c. potential, applied to one or both Y plates at the earthy side of the leakage resistor, allows the otherwise undeflected spot to be moved along the Y axis. This is called *Y shift*. *X shift* can similarly be provided.

For p.d.s applied directly across the deflector plates, the sensitivity is not high, being of the order of 20 V for 1 cm deflection. Most instruments include an amplifier, to increase the sensitivity when required.

10.2 The electron gun

The main electrodes of the electron gun are the heated cathode, which emits the electrons, and the anode, which provides the potential gradient to accelerate them.

The cathode is an indirectly heated metal cylinder, with an oxide coated face, from which the electrons are emitted by thermionic emission. The anode has a central hole through which the accelerated electrons can emerge. After this, they must maintain their speed until they strike the fluorescent tube face, which means that there must be conducting connection between the anode and the end of tube, to put the two at the same potential. For practical reasons, this is made earth potential, and the accelerating p.d. is applied by making the cathode about 500–1500 V negative relative to earth.

If these were the only electrodes in the gun, the electrons would be successfully speeded up and some of them would emerge from the gun, but the majority would travel no further than the anode and the remainder would not be focused at a single spot on the tube face. Two additional electrodes are included to make the electrons follow the path shown in Fig. 10.2(*a*). These are the grid, and the focusing electrode.

The grid is a cylinder, very close to the cathode, and made a variable amount negative relative to the cathode by a potential dividing resistor called *the brilliance control*. Grid and cathode together make up an electron lens system, which focuses the electron beam to pass through the hole in the anode.

The focusing electrode is a subsidiary anode, and the two anodes together make up a similar electron lens system, used to focus the beam to the spot on the tube face. The potential divider used, in this case, to vary the relative potentials of the two anodes, is *the focus control*.

The electron lens action is suggested by the contours of equipotential surfaces shown in Fig. 10.2(*b*). Electrons are attracted towards the positive end of a potential gradient in much the same way as a ball moves down a hill. Taking positive potential, for this purpose, to be lower than a negative one, electrons moving across a potential gradient tend to move towards any potential channel, and this is what the curved equipotential surfaces of the electron lens provides.

The slanting neck of the tube is coated internally with graphite, which is connected to the anode, and earthed. To provide screening against the effects of external electric and magnetic fields, the tube is surrounded by earthed μmetal, an alloy which becomes magnetized and demagnetized very readily.

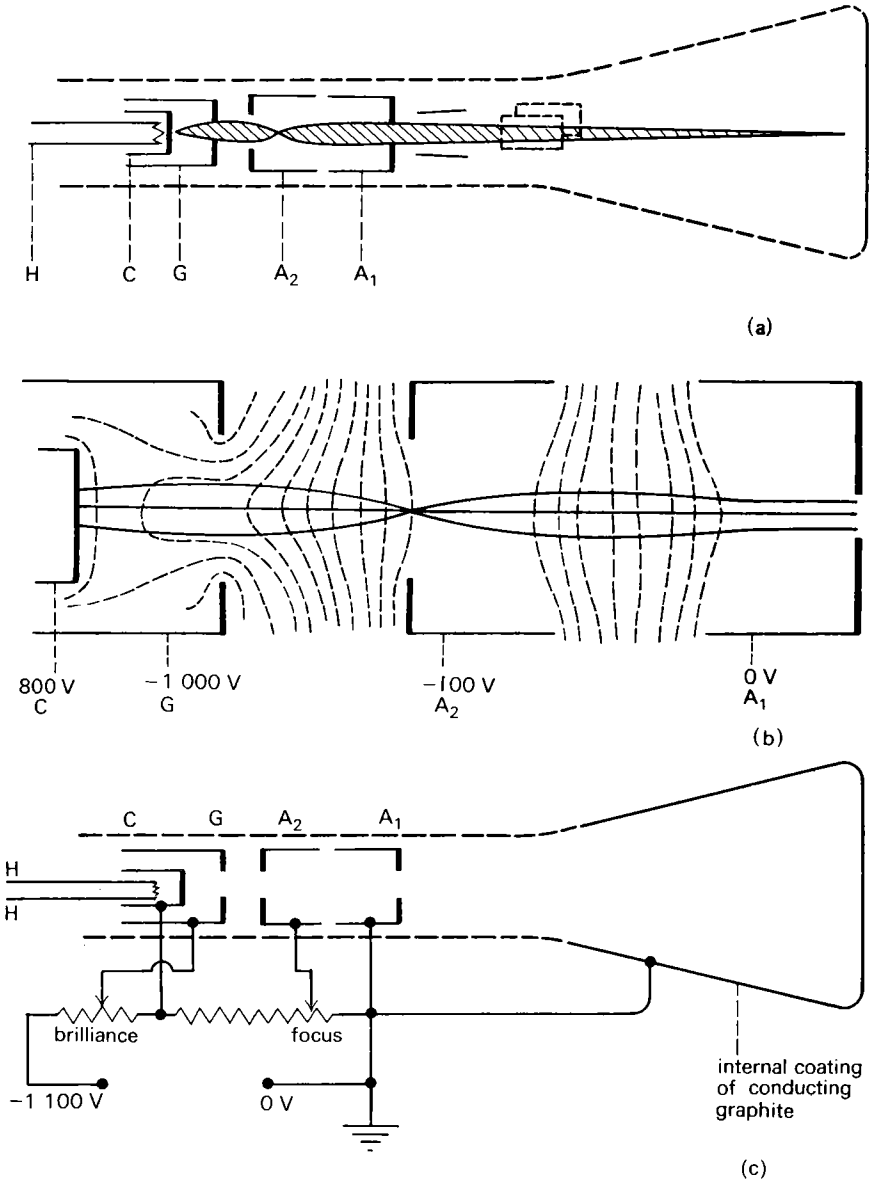


Fig. 10.2 The electron gun (a) The electrodes, and the electron beam (b) Equipotential surfaces. The dotted lines are contours of equipotential, being a down gradient, for electrons, from left to right (c) The resistance chain connected across the power supply

10.3 Magnetic deflection and focusing – the television tube

In some oscilloscopes, means is also provided to deflect the electron beam by the magnetic field of an electric current. This is the method of deflection which is always used with the type of cathode ray tube used in television. In this case, the beam must focus to a specially bright and sharp spot which requires higher energy electrons, provided by a higher p.d. than that used in oscilloscope tubes. These

high energy electrons would require an inconveniently strong electric field to deflect them, so this is achieved magnetically instead. The magnetic fields are produced by electric currents flowing in deflecting coils, placed outside the narrow neck of the tube. Focusing, too, is achieved magnetically, by means of a non-uniform magnetic field directed along the axis of the tube. A discussion of magnetic focusing is outside the scope of this book.

10.4 The measurement of peak a.c. voltage

If an alternating p.d. is applied across the Y plates of the cathode ray tube, the spot continuously oscillates up and down, tracing a vertical line, from positive to negative peak of the p.d. The length of this line is proportional to twice the peak voltage, i.e. $2V_0$. That is, the oscilloscope is a peak voltmeter.

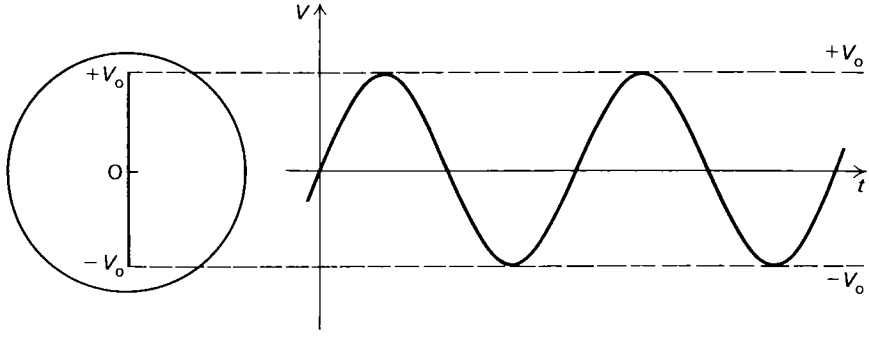


Fig. 10.3 The C.R.O. as a peak voltmeter

10.5 Measurement of a d.c. voltage

If a d.c. potential were applied to a deflector plate which is already connected to a source of X or Y shift potential, the two would interfere with each other. Because of this and other similar reasons, the X and Y input terminals are usually connected to the deflector plates through capacitors. A capacitor in series is no great obstacle to a.c. signals, but a d.c. potential will not permanently register. To obtain a deflection of the spot in response to a d.c. potential, at least one deflector plate must be connected direct to its terminal. The p.d. to be measured is applied between this terminal and earth, and a steady deflection of the spot results, which is proportional to this p.d.

10.6 Measurement of current

Using magnetic deflection, a current could be measured by sending it through the deflector coils, because the deflection is proportional to this current. More usually, however, electrostatic deflection would be used, and in this case the current is passed through a resistor, and the p.d. across this resistor is applied across the deflector plates.

10.7 Oscillograms

Because the spot can be made to respond to two p.d.s simultaneously, it can be made to trace out a graph showing how one p.d. V_2, applied across the Y plates, varies with another potential difference V_1, applied across the X plates. But the

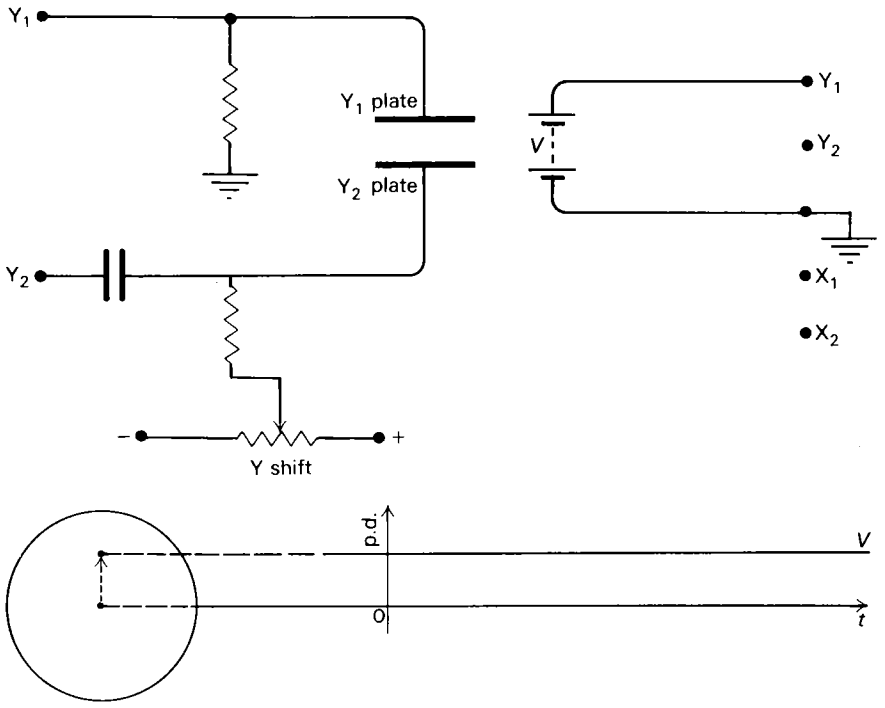


Fig. 10.4 Measurement of a d.c. p.d. (a) Connections to Y plates (b) External connection of d.c. p.d.

spot does not resemble a pen in leaving a permanent trace behind it. For the graph to be continuously visible on the tube face the spot must repeatedly trace out the same pattern. This means that the potential differences V_1 and V_2 must be repetitive, i.e. they must have periodic waveforms. Also, these must be synchronized with each other, in order that the successive traces shall be identical and coincident.

10.8 Frequency measurement

A very simple example of an oscillogram is one which is very useful for comparing the frequencies of two alternating p.d.s. It may be used, for example, to measure the frequency of the output from an audio-frequency oscillator, by comparing this with the frequency of the a.c. mains supply, which may be assumed, in Britain, to be 50 Hz.

The mains frequency is applied, at a suitable p.d., across the X plates; most oscilloscopes have internal provision for this. The output from the audio-oscillator is applied across the Y plates. Two typical patterns are shown in Fig. 10.5, for

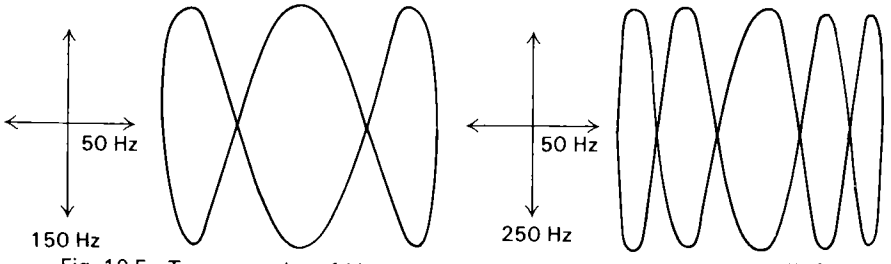


Fig. 10.5 Two examples of Lissajous figures (a) Ratio 3 : 1 (b) Ratio 5 : 1

which frequencies of 150 Hz and 250 Hz have been applied, in turn, across the Y plates, having 50 Hz across the X plates. In the second case, the spot oscillates vertically 5 times as frequently as it does so horizontally. If a pencil point is made to follow the pattern in the diagram, it will be seen to move up and down five times, to once to and fro horizontally. Such a pattern is called a Lissajous figure.

This method of comparing frequencies is practicable only if the two frequencies have a simple ratio to each other. Otherwise, the two voltage variations are not synchronized, and the pattern is continually changing.

10.9 Plotting waveform

The oscillogram which is most commonly used plots waveform, i.e. p.d. against time. To do this, the oscillating p.d. is applied across the Y plates, while the horizontal trace must represent time. The spot is made to travel repeatedly from left to right across the tube face, marking equal distances in equal times, returning rapidly to its starting point to retrace its path continuously at a constant repetition rate. The varying p.d. applied across the X plates to have this effect requires to have a sawtooth waveform, as shown in Fig. 10.6(*b*). The electronic circuit which produces it is called *a linear timebase circuit.*

Here, again, for the resultant trace to be steady, the spot must repeatedly trace out the same path each time that it crosses the tube face, which means that the signal frequency must be an exact multiple of the time base repetition rate, i.e. the two must be synchronized.

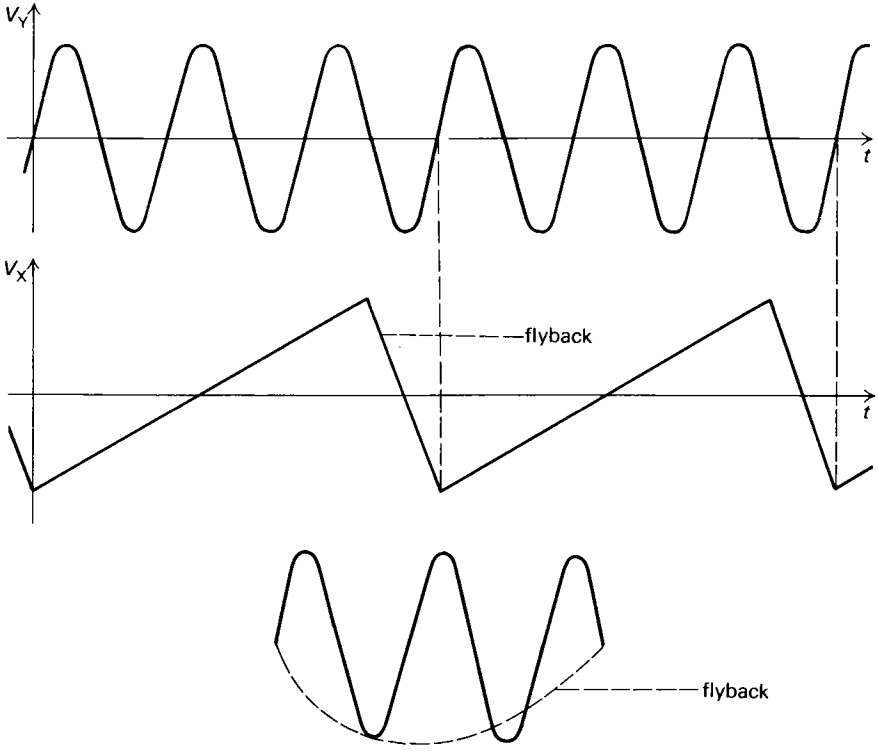


Fig. 10.6 The linear time base (a) The signal applied to a Y plate (b) The linear time base applied across the X plates (c) The trace on the face of the tube. The flyback is usually suppressed, i.e. made invisible

10.10 The equivalence of a.c. and d.c.

Some typical uses for an oscilloscope in a school laboratory will now be described, with reference to the a.c. theory given in the last chapter.

We start with a comparison of the effects of d.c. and a.c. in a resistor. It was explained in section 9.4 that an alternating p.d. with r.m.s. value V should have the same heating effect when applied across a resistor as a steady d.c. potential difference of value V. Further, provided that the waveform is sinusoidal we have

$$V_{\text{r.m.s.}} = \frac{1}{\sqrt{2}} V_0.$$

A suitable circuit to test these conclusions is shown in Fig. 10.7, where the oscilloscope is used to measure both a d.c. potential difference and the peak value

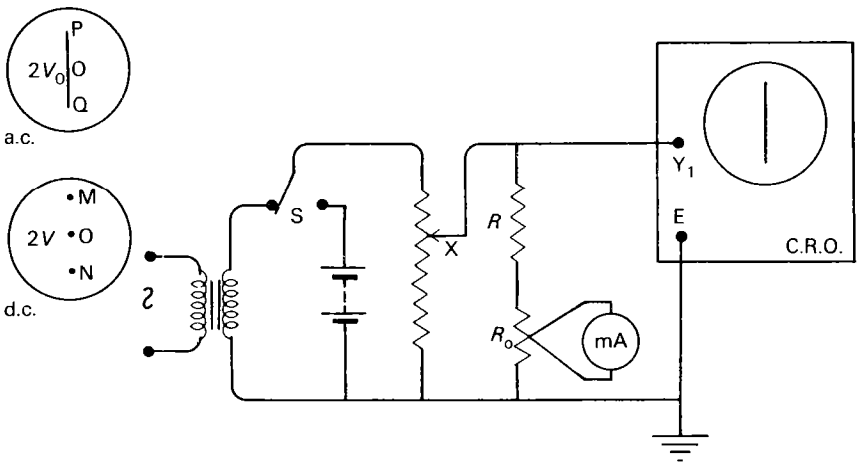


Fig. 10.7 A circuit to demonstrate the relation between the effective value and the peak value of an alternating p.d.

for a.c. The meter shown in the diagram is a thermo-milliammeter in which a thermocouple in series with a moving coil meter indicates the degree to which the thin wire resistor R_0 is heated by the current flowing in it. This is connected in series with a resistor R, to convert the arrangement into a voltmeter.

An alternating p.d. is applied between a Y plate of the C.R.O. (direct connection) and earth, and at the same time it causes current to flow through the thermo-milliammeter. The length of the vertical trace, PQ, on the tube face is proportional to the peak to peak value of the alternating p.d., i.e.

$$\text{PQ} \propto 2V_0.$$

The change-over switch S is now moved across to connect the direct p.d. from an H.T. supply across the C.R.O. instead, and the potential divider contact X is moved to produce the same reading on the thermo-milliammeter as was previously obtained with the alternating p.d. This steady p.d. shifts the spot on the tube face upwards to M, where the deflection OM is proportional to the steady p.d. V, which is equivalent to the alternating p.d. having the peak value V_0. The polarity of the H.T. supply is now reversed, giving an equal downward deflection ON. The sum of the two deflections, i.e. the distance MN, is proportional to $2V$,

$$\text{MN} \propto 2V.$$

The significance of the two identical readings of the meter is that an alternating p.d. with peak value V_0 and the steady direct p.d. of value V are equivalent, both producing heat at the same rate in the resistor. The ratio is

$$\frac{V_0}{V} = \frac{PQ}{MN}.$$

A series of corresponding values of PQ and MN may be obtained by adjustments of the potential divider. According to theory, the ratio should be equal to $\sqrt{2}$. This is so, however, only for sinusoidal waveform, and the oscilloscope can be used to check this waveform if the linear time base is switched on to the X plates.

10.11 The variation of impedance with frequency

In this series of experiments, the oscilloscope is used to measure frequency by comparison with the 50 Hz mains frequency, using Lissajous figures. A variable audio-frequency oscillator is used as an a.c. source to send current through a selected impedance connected in series with an a.c. milliammeter, and the value of the impedance is measured for a sequence of measured frequencies, in multiples

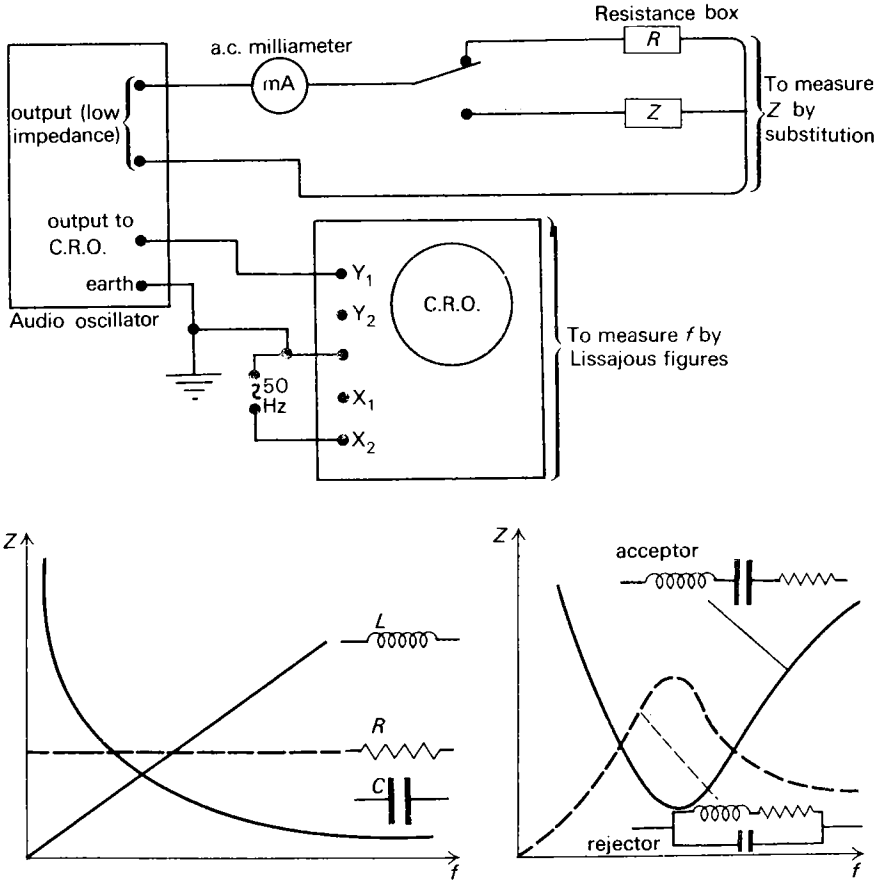


Fig. 10.8 The variation of impedance (or reactance) with frequency (a) An arrangement to measure impedance and frequency (b) Inductor, resistor, and capacitor. (In practice, the inductor would have some resistance, and relation between Z and f would not be strictly linear) (c) Acceptor and rejector circuits

of 50 Hz. The measurement of impedance can be carried out by the method of substitution, i.e. after measuring the current, a resistance box is substituted for the impedance in the circuit. The value of the resistance is selected to give a meter reading identical with that obtained using the impedance, and the value of this impedance is taken to be numerically equal to that of the resistance. A series of corresponding values of frequency f and impedance Z may be obtained in this way, and a graph plotted to show the relation between them. In this way, the results obtained in the last chapter may be tested, using arrangements of capacitor, inductor, capacitor and inductor in series, and so on, as summarized by the graphs drawn in Fig. 10.8(*b*) and (*c*).

[This experiment is subject to some systematic error, particularly if the resistance of the audio-oscillator or of the milliammeter is at all high. This is because a complicating factor has been neglected, involving phase. For this reason, the audio-oscillator chosen must have a low impedance output.]

*10.12 Measurement of phase difference by oscilloscope

The phase angle between the p.d. and the current in an impedance can be deduced by taking measurements from an oscilloscope trace. If the p.d. across the impedance is applied across the Y plates, and a p.d. proportional to current is applied across the X plates, this trace is an ellipse.

Alternating current is sent through the impedance and a resistance R_0 in series. The p.d. across the resistor, being equal to the product R_0I, is proportional to the

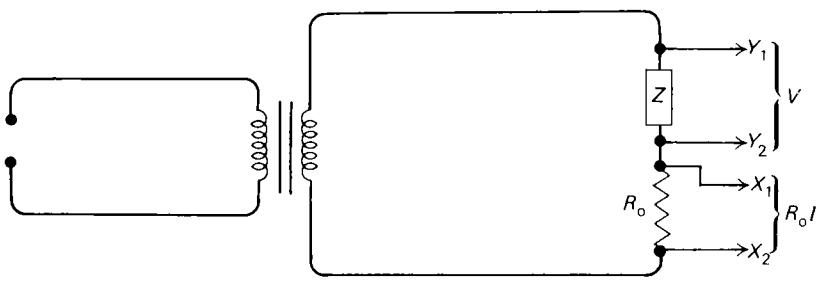


Fig. 10.9 Measurement of phase angle The equations for the two p.d.s are

$$R_0I = R_0I_0 \sin \omega t$$

and

$$V = V_0 \sin(\omega t + \phi)$$

current, and this is applied across the X plates; the potential difference V across the impedance is applied across the Y plates.

The equations for the two p.d.s are

$$R_0I = R_0I_0 \sin \omega t$$

and

$$V = V_0 \sin(\omega t + \phi).$$

If the connections to the Y plates were removed, leaving the alternating p.d. R_0I across the X plates, the trace would be a horizontal line, proportional in length to twice the peak value, $2R_0I_0$. Similarly, if the X plates were disconnected instead, the trace would become a vertical line, proportional in length to $2V_0$. When both

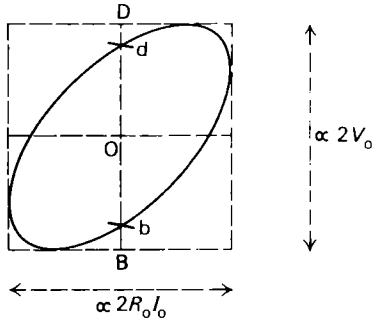


Fig. 10.10

sets of connections are made, the spot traces out an ellipse bounded by a rectangle having sides equal to these two lengths.

Now, when $t = 0$, we have

$$V = V_0 \sin \phi$$

and

$$R_0 I = 0.$$

Hence, at this instant, the vertical deflection Od in Fig. 10.10 is proportional to $V_0 \sin \phi$. But OD is proportional to the peak value V_0.

Hence,

$$\sin \phi = \frac{V}{V_0}$$

$$= \frac{\text{Od}}{\text{OD}}$$

$$= \frac{\text{bd}}{\text{BD}}.$$

All that is required, therefore, is to measure the length bd from the ellipse and then to disconnect the X plates and measure BD, which is the length of the vertical trace which remains. The ratio of these two distances is equal to $\sin \phi$.

The trace for a reactance

If the impedance Z is a pure reactance, the phase angle ϕ is 90°, which means that the maximum vertical deflection, due to V_0, occurs at the instant when the horizontal deflection, due to $R_0 I$, is zero. The ellipse now has vertical and horizontal axes.

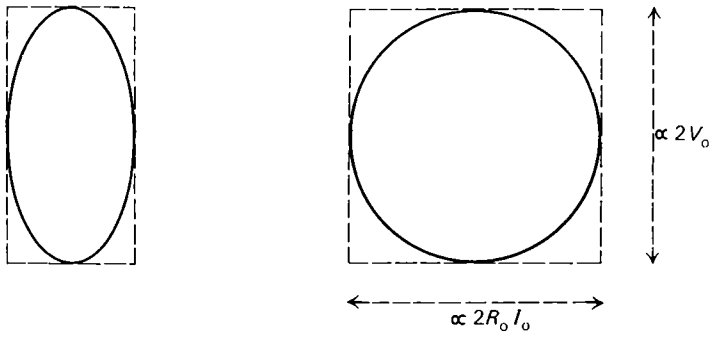


Fig. 10.11 Figures obtained for a reactance

If the value of R_0 is chosen to make the horizontal and vertical traces have equal lengths, the ellipse becomes a circle. Further, if both sets of deflector plates have the same sensitivity (or have been corrected to have this effect by using an amplifier (or a potential divider), we can say that, for the circular trace

$$R_0 I_0 = V_0,$$

$$\therefore \quad X = \frac{V_0}{I_0}$$

$$= R_0.$$

QUESTIONS

10.1 Draw a good labelled diagram of a cathode ray oscilloscope; indicate one way in which suitable voltages could be applied to the various electrodes, and show how they are applied. (Details of the time base circuit are not required.) How would you calibrate an oscilloscope for use as a voltmeter?

Explain briefly (*a*) why the tube should be highly evacuated, (*b*) why the screen and the region around it are made electrically conducting, (*c*) why a very high accelerating voltage is necessary if the oscilloscope is to be used to measure alternating voltages of high frequency. (O)

10.2 Draw a sketch to show the essential parts of a cathode ray oscillograph having electrostatic deflection. With the help of your sketch explain how in a cathode ray oscillograph (*a*) the electrons are produced, (*b*) the electrons are focused, (*c*) the spot is made visible, (*d*) the brightness of the spot is controlled.

What is meant by stating that a cathode ray oscillograph is fitted with a linear time base of variable frequency? (N)

10.3 What is meant by the statement that the sensitivity of a certain cathode ray tube for vertical deflection is 2 V/mm? With the help of a diagram, explain the factors in the design of the tube which govern this sensitivity.

How would you use a cathode ray oscilloscope to do the following experiments?

(*a*) Observe the waveform of the a.c. supply mains voltage (approximately 240 V r.m.s. at 50 Hz).

(*b*) Check the calibration of an audio-oscillator of variable frequency at a scale reading of about 1000 Hz, the frequency of the supply mains being assumed to be exactly 50 Hz. (O & C)

10.4 The Y plates of an oscilloscope are connected to a source which gives a sinusoidal alternating p.d. of frequency 250 Hz. Sketch the trace which you would expect to observe on the face of the tube in each of the following circumstances:

(*a*) No signal is applied to the X plates, and no time base voltage is switched on.

(*b*) A linear sawtooth p.d. is applied across the X plates, which changes linearly from $-V$ to $+V$ with a repetition rate of 50 Hz, taking 1/60s to cover the voltage range and 1/300 s to return to the starting point.

(*c*) A sinusoidal alternating voltage of frequency 50 Hz is applied across the X plates.

(Assume in all cases that the applied p.d.s are insufficiently large to deflect the spot beyond the edge of the tube face.)

10.5 Explain what is meant by (*a*) a linear time base, (*b*) a sinusoidal time base in a cathode ray oscilloscope.

The X and Y deflection sensitivities of a cathode ray oscilloscope are each

5 V/cm. A sinusoidal p.d. at 50 Hz and of r.m.s. value 20 V is applied across the Y plates of the instrument. A p.d. of the same form and frequency, but of r.m.s. value 10 V, is simultaneously applied across the X plates. Sketch and explain the pattern seen on the oscilloscope when the p.d.s are (*a*) in phase, (*b*) 90° out of phase.

Indicate approximate dimensions in your sketches. (N)

10.6 A 50 Hz a.c. source with sinusoidal waveform and r.m.s. p.d. 30 V is connected across a diode and a 1000 Ω resistor in series. The p.d. across the resistor is applied across the Y plates of an oscilloscope. At the same time, a linear sawtooth p.d. is applied across the X plates, which changes in value from −100 V to +100 V in 0.05 s, falls back to −100 V in 0.01 s, and then repeats the cycle continuously.

If the sensitivity of both X and Y plates is 20 V/cm, and the diode has effectively zero resistance towards current in one direction and infinite resistance towards current in the reverse direction, give a labelled sketch of the trace which you would expect to observe on the face of the tube.

Suggest how the trace would be altered in appearance if a 20 μF capacitor were connected in parallel with the 1000 Ω resistor.

10.7 An oscilloscope is supposed to have a linear timebase p.d. applied across the X plates, but the time-base circuit is defective, with the result that the p.d. actually provided has the waveform shown in the diagram.

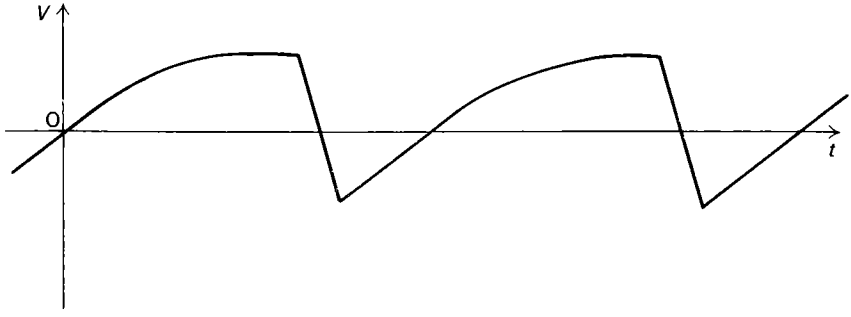


Fig. e 68

Sketch the trace which would appear on the tube face, if a sinusoidal alternating p.d. were applied across the Y plates having frequency which is 5 times the repetition rate of the time-base.

10.8 A capacitor *C* and a resistor *R* are connected in series across an alternating source which supplies a p.d. of 100 V at a frequency of 500 Hz. The X plates of an oscilloscope are connected across the resistor. If the capacitance of *C* is 0.80 μF and the resistance of the resistor *R* is 400 Ω, and the X and Y sensitivities of the oscilloscope are equal, sketch the appearance of the trace on the tube face of the oscilloscope when the Y plates are connected (*a*) across the resistor R, (*b*) across the capacitor C, (*c*) across the terminals of the alternating source.

*10.9 Two plane coils of identical construction have a common diameter, and are rotated about this diameter in a magnetic field, while the angle θ between the coils is kept constant. One coil is connected to the X plates of a C.R.O. and the other to the Y plates.

Describe and explain what may be observed on the tube face for (*a*) differing speeds of rotation, and (*b*) different values of θ. (CAM. SCH.)

*10.10 The deflector plates of an oscilloscope are 2 cm long and 0.5 cm apart. Calculate the sensitivity of the instrument (in cm/V) if the potential of the

gun is 2000 V and the distance between the centre of the deflector plates and the tube face is 25 cm.

What factors do you consider will affect the sensitivity when the p.d. across the deflector plates oscillates at high frequency?

($e = 1.6 \times 10^{-19}$ C, $e/m_e = 1.76 \times 10^{11}$ C/kg) (CAM. SCH.)

*10.11 Obtain an expression for the sensitivity, y/V_y, of a cathode ray tube in terms of the factors l_1, d, V, l_2, where the Y deflector plates consist of two parallel plates, and

l_1 = the length of the deflector plates, parallel to the direction of travel of the undeflected cathode rays,

d = the separation of the deflector plates,

V = the effective accelerating p.d. applied across the gun electrode,

l_2 = the shortest distance between the Y plates and the screen,

y = the deflection on the screen when a p.d. V_y is applied across the Y plates.

Hence, deduce the sensitivity, in mm/V, in the case when $l_1 = 2.0$ cm, $d = 0.5$ cm, $l_2 = 20$ cm, and the gun p.d. is 1200 V.

It will be noted that the expression for sensitivity does not include either the electronic charge e or charge/mass ratio e/m_e. Suppose that a beam of protons could be provided as readily as a beam of electrons, explain why a 'proton tube' would not have the same success as the cathode ray tube.

Chapter 11

Insulators and semiconductors. Thermocouples

11.1 Conductors, semiconductors and insulators

In discussing the effects of an electric current, it is natural first to look at metals and electrolytes because these are the best conductors. This is not to say, however, that all other substances are perfect insulators. Elementary electrostatics experiments readily demonstrate that they are not, when, for example, such materials as wood or cotton are shown to discharge a charged electroscope by their conduction.

In some cases of this kind, the conduction may be due to moisture. Nevertheless, a very large number of non-metallic substances are at least slightly conducting in their own right, and solids can be divided fairly clearly into three classes: metallic conductors, semiconductors, and insulators. Many insulators, however, become semiconducting if the temperature is raised sufficiently.

Table 2. *Approximate values of resistivity at room temperature*

		Substance	*Resistivity*/Ω·m
Metallic conductors	Elements	silver	1.6×10^{-8}
		copper	1.8×10^{-8}
		aluminium	2.7×10^{-8}
	Alloy	manganin	4.4×10^{-7}
semiconductors		graphite	about 1×10^{-5}
		very pure germanium	0.5
		very pure silicon	2×10^{3}
Insulators		diamond	10^{12}
		polythene	about 10^{16}

The conducting properties of a substance depend on the availability of charged particles, electrons or ions, which are free to move through the substance, and this involves not only the structure of its atoms but also the type of bonding which holds them together. The introduction of foreign atoms into a crystal lattice may also have a very marked effect. We have already seen that this increases the resistivity of a metal, but in many cases it greatly reduces that of a semiconductor, as, for example, in *n*- and *p*-type germanium and silicon.

We may take carbon as an example to illustrate the effect of crystal structure. This element appears twice in table 2, as graphite and as diamond. The arrangement of carbon atoms in the diamond crystal leaves no relatively free electrons or ions, and a diamond is effectively a good insulator. Graphite, on the other hand, having a different crystal structure, almost classifies as a good conductor.

11.2 Covalent bonds

In the solid state, substances are usually crystalline, and we can recognise four distinct ways in which the atoms may be linked together in a crystal, two of which, metallic and ionic bonding, have already been described. The third way is by covalent bonding. Before discussing this, it will be useful to expand slightly the description of the atom which was given in chapter 1.

The electrons surrounding the nucleus of an atom were shown in Fig. 1.1 as occupying a sequence of electron shells, the K, L, M, N shells etc. Each shell was represented as being a clearly defined orbit. This simple picture implies a more precise knowledge of the behaviour of the electrons than in fact we can hope to obtain. For a single particle as small as an electron we must be content with information which is less exact, and it is unrealistic to claim that we know the electron to be moving in the defined path of a specified orbit, circular or otherwise. This is because any process by which the position of the electron might be accurately located would change its motion, and leave us uncertain of its energy. Alternatively, to obtain exact knowledge of its energy would involve losing track of its position. In the event, most of our information comes from a spectrometer, which gives exact knowledge of the electron's energy, and this is at the expense of not knowing its exact position or track. Consequently, each electron in the atom is specified, not by an orbit, but by a rather ill defined space, representing its probable location, and this is known as an orbital. Many of the orbitals are not symmetrical with respect to the nucleus, which means that the electron occupies more of its time on one side of the nucleus than it does on the other. When an electron occupies a particular orbital, then, we cannot map out its exact path, but we do have accurate knowledge of its energy.

Put in more formal terms, the above expression of uncertainty about the behaviour of a very small particle is known as *Heisenberg's principle*. [This is not a principle which need worry us when dealing with larger masses, or with a large collection of small particles, such as the electrons in a beam of cathode rays; such large scale effects turn probabilities into what are almost certainties, and conventional equations apply.]

It was explained in chapter 1 that it requires eight electrons in the outer occupied shell of an atom to represent special stability, and that this is the arrangement in each of the atoms of the inert gases. No very serious modification needs to be applied to this statement. We still attach the same importance as before to these electron shells, and to the magic number of eight electrons in the outer shell. Where we will now go a little further than in chapter 1 is in remarking that these eight electrons, although they all belong to the same shell, do not all have the same orbital in the shell. No single orbital can contain more than two electrons, and the eight electrons in the outer shell of an inert gas are contained in four orbitals, not all of which represent exactly the same level of energy.

The innermost shell of the atom, the K shell, is an exception to the above rule. This has one single orbital only, and consequently it cannot hold more than two electrons. The first inert gas in the periodic table, which is helium, has atomic number $Z = 2$, i.e. it has only two electrons in its neutral atom, and these complete the complement of this single orbital in its K shell.

Ionic bonds

A slightly more accurate description of the ionic bonding between sodium and chlorine than was given in section 5.1 would refer to orbitals. Only one orbital in the M shell of the neutral sodium atom is occupied at all, and this holds the single valency electron of sodium. The seven electrons in the M shell of chlorine, on the

other hand, fill three orbitals, with two electrons each, leaving the fourth orbital holding the seventh electron and having one vacant place in it. The sodium, Na^+, and chlorine, Cl^-, ions in the sodium chloride crystal are formed because the valency electron has left its place in the M orbital of the sodium atom, and has taken up, instead, this vacant place in the M orbital of a chlorine atom. By this electron exchange, the M shell of the sodium atom has been completely emptied, and the four orbitals in the M shell of the chlorine atom have been filled, thus giving both the Na^+ and Cl^- ion the stable outer shell possessed by an inert gas.

Covalent bonds

When two chlorine atoms combine to form a single chlorine molecule, Cl_2, the bonding is of a different kind, and this is an example of covalent linkage. In this case, the orbital from one of the atoms, holding one electron, overlaps the similar

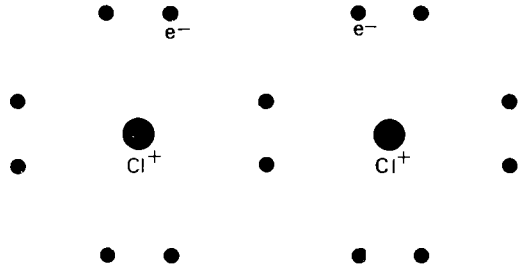


Fig. 11.1 A diagram representing the covalent bonding between two atoms in the chlorine molecule, Cl_2

orbital of the other atom, and these two orbitals share their two electrons, instead of exchanging them, giving the effect of completed shells. This is illustrated in Fig. 11.1. Only the valency electrons are shown in this and in similar diagrams, because they are the only ones concerned in the bonding.

In a covalent crystal, which may be an element or a compound, the atoms in the crystal lattice are held by covalent bonding between them and their neighbours on all sides. The diamond crystal is an example of this. Carbon, with atomic number 6, has two electrons in its K shell, and four valency electrons in its L shell, and it shares its four orbitals in this shell, one with each of four other carbon atoms in the lattice. This is compared with the structure of graphite in Fig. 11.2. The carbon atoms in graphite form successive planes with a hexagonal lattice, and only three of the four valency electrons fill overlapping orbitals, so taking part in the covalent bonding. This leaves the fourth electron, which more readily becomes a free charge carrier, and this accounts for the relatively high conductivity of graphite.

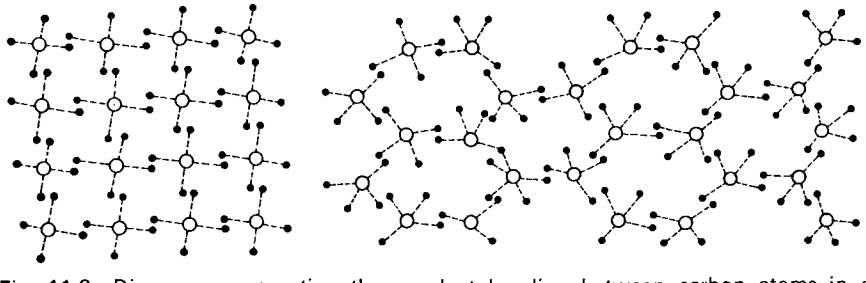


Fig. 11.2 Diagrams representing the covalent bonding between carbon atoms in a diamond and in graphite (a) Carbon atoms in a diamond crystal (b) Carbon atoms in graphite

It is evident that covalent bonding does not produce ions, and such materials are not electrolytes. Both covalent and ionic crystals may, however, be semiconductors. This aspect will be discussed presently.

11.3 Chain molecules, and Van der Waal's forces

Many of the best insulators, such as paraffin wax, bakelite, rubber, and polythene, have molecules which consist of long chains of atoms, linked together in the molecule by covalent bonds. Polythene, for example, i.e. polyethylene, has the equation $CH_2 \cdot CH_2 \cdot CH_2 \cdot CH_2 \ldots$, continuing for a very large number of carbon atoms. Substances of this kind lack the symmetry of a crystal lattice, but the molecules are arranged with a certain amount of regularity, and consequently the material might be said to be crystalline. The long atom chains may be strung out side by side, as in a synthetic fibre such as terylene or nylon; they may be partly strung out, and partly pushed up on themselves, as in polythene; or they may be partially curled up, as they are in rubber, giving rubber its peculiarly exaggerated elastic properties.

The bonding between the separate molecules is not the strong force that we find with ionic, covalent, or metallic bonds, but is the electric force between atoms and molecules which extends outside the range at which the exchange or sharing of electrons can occur. This kind of force is still effective, to a small degree, over the larger distance which separates the molecules in a gas, and it is known as a Van der Waal's force. At a few places in the molecules there may additionally be a sharing of electrons between atoms of neighbouring molecules, i.e. a covalent linkage, which strengthens the bonding, and makes the material more brittle. This is the case with vulcanized rubber.

Many of the synthetic polymers, such as nylon, terylene, and polythene, are moisture resisting, which enhances their value as insulators. These substances charge with electricity very readily, a tendency which is noticeable in nylon and terylene clothing.

11.4 An intrinsic semiconductor

A solid of ordinary chemical purity includes a small proportion of foreign atoms in its crystal lattices, and additionally there are imperfections, such as the presence of a surplus atom or of a vacancy in the lattice where an atom should be. If the substance is a semiconductor, the presence of free charge carriers is often due to one or other of these irregularities. Specialised semiconducting devices, however, such as transistors and diodes, are made from wafers of single crystal, initially prepared in a state of exceptional purity by special physical processes. From a theoretical point of view such a pure material is the simplest one to describe, and so we will start with this.

It will be convenient to choose the three elements, carbon, silicon, and germanium, both because the last two of these are commonly used as transistors and so on, and also because these three elements occupy corresponding places in the periodic table, each having four electrons in the outer shell. Their electron distributions are as follows:

	Atomic number Z	Number of electrons in each shell			
		K	L	M	N
C	6	2	4	–	–
Si	14	2	8	4	–
Ge	32	2	8	18	4

The fact that the germanium atom has eighteen electrons, not eight, in its M shell, involves a complication in atomic theory which need not concern us here.

The crystal lattices of silicon and germanium are similar to that of the diamond crystal, all four valency electrons being taken up with the covalent bonding. It might be expected, perhaps, on this account, that all three would be insulators, as there does not appear to be any electron or ion which might be free to act as a charge carrier. This does not allow, however, for the possible effect of thermal energy. It may be, particularly at high temperature, that some of the valency electrons can acquire sufficient thermal energy to free them from their covalent bonding. If this is so, these electrons resemble the conduction electrons in a metal; they are free to wander at random through the crystal, and to convey a current, but not to leave the crystal. They are unlikely, however, to be so abundant as in a metal, and the crystal would only be a semiconductor.

Perhaps it is not surprising that germanium is a better conductor than silicon at ordinary temperature, while a diamond does not conduct at all until its temperature reaches about 1 000 °C. The valency electrons of germanium are in the highest energy shell of the three, and might be expected to free themselves the more readily. This difference in behaviour is reflected in the resistivities of the three materials at room temperature, which were tabulated in table 2, in section 11.1.

11.5 Charge-carrying electrons and holes

When an electron breaks free of its place in the valency bonding, and becomes a negative charge carrier, it is leaving a place which was previously electrically neutral. Consequently, the vacancy left behind it in the crystal bonding represents a centre of positive charge, lacking, as it does, the negatively charged electron which should be there. This is illustrated in Fig. 11.3.

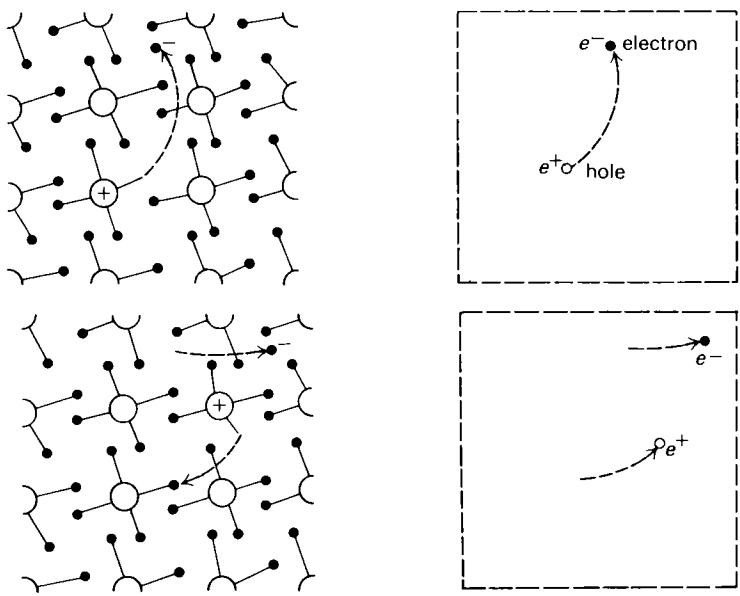


Fig. 11.3 Charge-carrying electrons and holes (a) The crystal lattice, showing one electron which has become a charge carrier leaving a charge-carrying positive hole (b) Diagram showing only the charge-carrying electron and the charge-carrying hole (c) The charge-carrying electron has moved, and the hole has been moved by an electron exchange (d) The effective movements of electron and hole

These positive vacancies in the bonding left behind by the freed electrons are known as *holes*. A hole can be filled, with comparative ease, by migration of an electron from one of the bonding orbitals near by. This is because the electron requires no extra energy to equip it for such a move; its new lodging place is exactly similar to the one which it has forsaken, and, in moving across, it neutralizes the positive charge of the hole which it moves to fill. Such an electron, however, in the process of neutralizing one hole, merely leaves a similar, positively charged vacancy behind it. This amounts to saying that, although the actual particle which has moved is an electron, in electrical terms it is the positively charged hole which has moved.

The overall effect is this. When the temperature of the crystal is sufficiently high for some valency electrons to be freed from their bonding orbitals, charge carriers are released, in effect, in pairs. For every charge-carrying negative electron, which takes its charge with it, there is also an electron which carries neutrality with it, and leaves positive charge in its wake, and this is, in effect, a charge-carrying positive hole. This is the effect which is suggested in Fig. 11.3(*b*) and (*d*), which is about the simplest way to represent charge-carrying electrons and holes in a semiconductor.

It is also possible, of course, for a hole to become filled, not by an electron from one of the bonding orbitals, but by one of the free charge-carrying electrons which created the holes in the first place. The result of this would be to reduce the number of charge carriers in the crystal, because this electron would not leave a positive hole behind it. At any specified temperature, however, there is an equilibrium concentration of free electrons and holes, which amounts to saying that, when such a pair neutralize each other, another pair is released elsewhere.

This has been a description of the exceptionally pure crystal, produced by a special modern technique. The charge carriers are negative electrons and positive holes in equal numbers, and their origin is the release of electrons from the covalent bonding; unlike the case of a less pure substance, there is nowhere else for the charge carriers to come from. This type of semiconductor is termed *intrinsic*. When an electric current flows in it, this consists of a drift of electrons in one direction, and of holes, not necessarily with the same mobility, in the opposite direction.

11.6 An extrinsic semiconductor

Most semiconducting devices, such as the transistor, are made by starting with the pure crystal of the type just described, and then injecting selected impurities into it, in controlled amounts. This has the effect of putting charge carriers into the crystal which are extra to the intrinsic ones, and indeed suppress a proportion of these, and the material is now termed an extrinsic semiconductor, the added charge carriers being extrinsic charge carriers. [The more ordinary semiconducting material, which has ordinary chemical purity, is an extrinsic semiconductor too, since its conducting properties are largely due to impurities and imperfections, and one or two devices, such as metal rectifiers, are made from materials of this kind, but these are relatively unimportant.]

Extrinsic charge carriers, unlike intrinsic ones, are not released in pairs, and may all have one sign, a fact which accounts for the versatility of semiconducting devices, since both sign and concentration of charge carriers can be fixed by design. Extrinsic crystals in which charge-carrying electrons predominate are referred to as *n*-type, and any holes which are present are termed minority charge carriers; similarly, if holes predominate, the crystal is *p*-type, and the minority carriers are electrons.

Since it is the extrinsic electrons or holes which are wanted, the ideal substance from which to make an extrinsic semiconductor would not release intrinsic charge carriers at all, at any rate at normal temperature; both germanium and silicon are suitable in this respect, particularly silicon, the resistivity of which is very high at room temperature (table 2, section 11.1). If a very small proportion of foreign atoms is injected into the pure crystal, a procedure known as doping, the effect can be very pronounced; the addition of as little as one part in ten million may increase the conductivity about a hundred times.

The foreign atoms selected for the purpose of doping must be capable of taking up places in the crystal lattice which would normally be occupied by native atoms. In the case of silicon or germanium, these are either atoms of boron, aluminium, or indium, which have three electrons in their outer shell, or atoms of phosphorus, arsenic, or antimony, which have five.

The *n*-type crystal

The effect of an antimony atom, introduced into the crystal lattice of germanium, is illustrated in Fig. 11.4. (Phosphorus would have a similar function in silicon.) The antimony atom occupies a place in the lattice which would otherwise have been taken up by germanium. But the antimony has one electron extra to those required for the covalent bonding. This electron, which is labelled e^-, has, in the diagram, rather the appearance of being quite free and even likely to contribute a surplus negative charge to the crystal as a whole. This is not really the case, however; the antimony nucleus contains the extra proton needed to neutralize the charge of the extra electron, which is attracted to it and occupies one of the orbitals of its outer shell. This extra proton has been represented in the diagram by inserting a positive sign in the antimony nucleus. It is true, however, that the extra electron is not bound to its atom as firmly as the electrons in the bonding orbitals are, and it becomes a charge carrier very readily. When it does so, the antimony ion which it leaves behind becomes a centre of positive charge, but it is not a charge carrying hole, capable of moving about, because it has no vacancy in the bonding, into which electrons from neighbouring bonding orbitals might move.

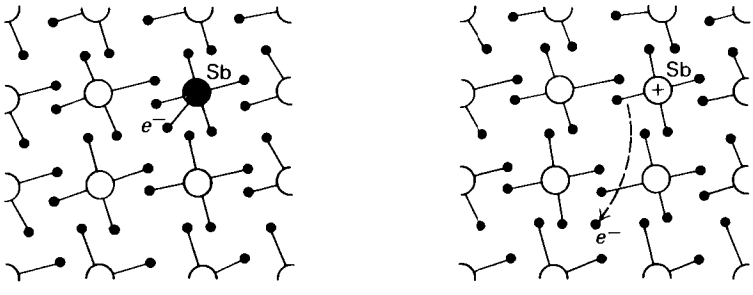


Fig. 11.4 *n*-type germanium, doped with antimony (a) The extra electron is still in an orbital of the Sb atom (b) The extra electron has become a charge carrier

This explains why the doped crystal at ordinary temperature has greatly increased conductivity. It requires only a small increment of thermal energy to release the 'surplus' electrons from their parent atoms. These extrinsic charge carriers all being negative electrons, the crystal has become an *n*-type semiconductor. The impurity, because it supplies extra electrons, is a donor impurity.

The presence of these extrinsic electrons reduces the concentration of intrinsic

holes, because these tend to trap some of the electrons, thereby becoming neutralized, and so the number of minority carriers is even smaller than might be expected.

The *p*-type crystal

The *p*-type crystal is produced in a similar way by injecting atoms of a trivalent atom, boron, aluminium, or indium. This is termed an acceptor impurity, because it leaves a vacancy in the valency bonding which in favourable circumstances can accept an electron. So long as this vacancy remains in the orbital of the acceptor

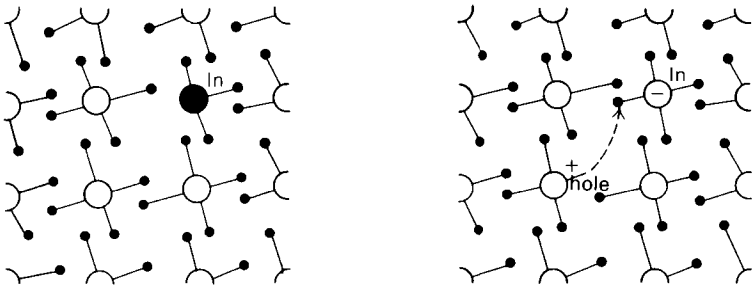


Fig. 11.5 *p*-type germanium, doped with indium (a) The vacancy is initially in an orbital of the neutral indium atom (b) The vacancy has been transferred to a germanium orbital, becoming a positive hole. The corresponding negative charge is locked in the lattice with the negative indium atom

atom, it is a vacancy, but not a charge-carrying hole. It belongs to the acceptor atom, which would be negatively charged if it were filled, and it is not, as yet, a centre of positive charge.

If an electron is to leave a bonding orbital elsewhere in the lattice to occupy the vacancy in the orbital of the acceptor atom instead, it requires a small increment of energy to enable it to do so. Once this has been provided, the vacancy left behind by the migrating electron has become a positively charged hole, which is free to be shifted about in the crystal by the valency electrons and so is a positive charge carrier. The corresponding negative charge is now locked in the crystal, with the electron which filled the original vacancy in the orbital of the acceptor atom.

11.7 Ion vacancies

Some electrovalent compounds are extrinsic semiconductors owing to imperfections in the crystal lattice caused by the absence of ions which should be there. Cadmium sulphide is an example, and it is used in one type of photoelectric cell because of certain special aspects of its action as a semiconductor. Absence of an S^{2-} ion from the lattice would mean loss of neutrality at that point, the negative charge of the ion being lacking. Two electrons are trapped in the vacancy to maintain the neutrality. These electrons readily become extrinsic charge carriers, and the crystal is an *n*-type semiconductor. One method by which the electrons can acquire the necessary energy to free them is by the absorption of light, and this is the process when the crystal is used in a photocell.

11.8 The variation of resistance with temperature

When the effect of temperature rise on the resistivity of a metal was discussed earlier, in section 3.12, the corresponding effect with a semiconductor was briefly mentioned. We can now enlarge on this, which concerns equation 3.13

$$I = NAve. \tag{3.13}$$

For a metallic conductor, the concentration N of charge carriers is very large in any case, and is not appreciably affected by temperature rise. Temperature rise does, however, cause a reduction in the drift velocity v of the electrons for reasons already given, and the resistance of the metal is consequently increased.

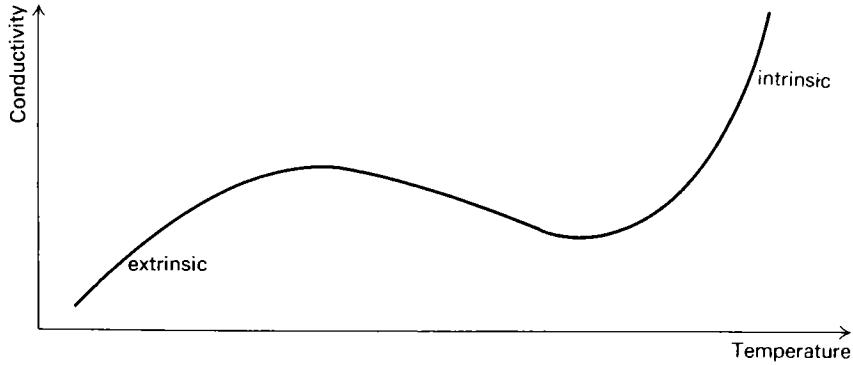


Fig. 11.6 An extrinsic semiconductor. Variation of conductivity with temperature

In the case of an intrinsic semiconductor, a much larger effect of temperature rise is the increase in the value of N, a greater number of electrons and holes being released in the crystal lattice. Consequently the resistance of an intrinsic semiconductor goes down when the temperature is raised.

With an extrinsic semiconductor, the number of potential extrinsic charge carriers is limited to the number of donor or acceptor atoms which are present in the crystal lattice. The increase in conductivity with temperature rise which results as they are released reaches its upper limit at the temperature which releases them all. The conductivity then proceeds to fall with further temperature rise. Finally, the temperature is high enough to release intrinsic electrons and holes from the valency orbitals of the parent crystal. The conductivity, consequently, proceeds to rise again, sometimes very steeply, as intrinsic conduction takes over. With many semiconducting devices, care must be taken to avoid reaching this condition, whether by direct overheating, or by passing an excessive current. It is an advantage of silicon over germanium that it will tolerate higher temperatures because its intrinsic charge carriers are harder to release.

11.9 The thermistor

The variation of resistance of a semiconductor with temperature is used, in the thermistor, to provide a thermometer of very high sensitivity, small size, and low thermal capacity. Thermistors are usually made from very pure oxides, such as manganese (II) oxide, with a small controlled amount of a selected impurity. Their response to temperature change also means that the resistance falls with increased current owing to the effect of Joule heating on the temperature.

A thermistor may be used in series with a metallic resistor to compensate for the increase in resistance of the metal with temperature.

Another application is as a voltage stabiliser. The potential dividing circuit shown in Fig. 11.8 has a thermistor put in parallel with the resistor R_2. Should any effect which increases the p.d. across the resistance chain AC occur, a large

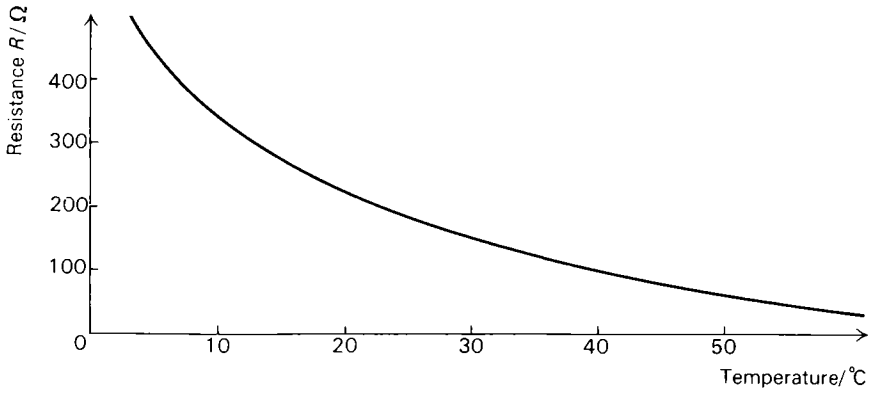


Fig. 11.7 A thermistor. Variation of resistance with temperature

current flows through the thermistor, hence lowering its resistance, with the result that the section BC of the potential divider takes up a reduced proportion of the p.d. V and the potential of B may remain constant.

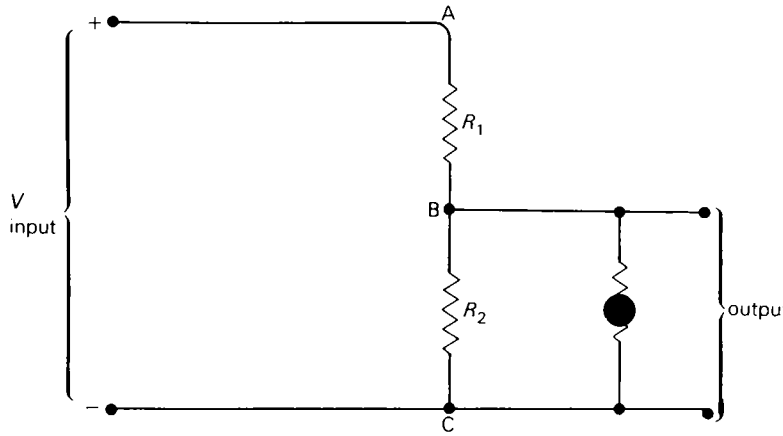


Fig. 11.8 A thermistor used as a voltage stabilizer

11.10 Thermoelectricity – the Seebeck effect

One further topic appropriate to this chapter is thermoelectricity, which concerns the direct conversion of heat to and from electrical energy. The Joule heating effect in a resistor is one case of this; but this is irreversible, and what we now have to consider are reversible effects.

If a circuit, consisting of two different metals and therefore involving two joins between the metals, is made, and if these joins are kept at different temperatures, a continuous current flows in the circuit, indicating the existence of a small e.m.f. This was discovered in 1826 by Thomas Seebeck and is known as the Seebeck effect, the whole arrangement being termed a thermocouple.

The source of the electrical energy in this case was later shown to be heat, which is absorbed in one part of the circuit, including the hot junction, and emitted, in smaller amounts, in the other part, including the cold junction. The effect is on too small a scale, however, in the case of metals, to afford a useful energy source,

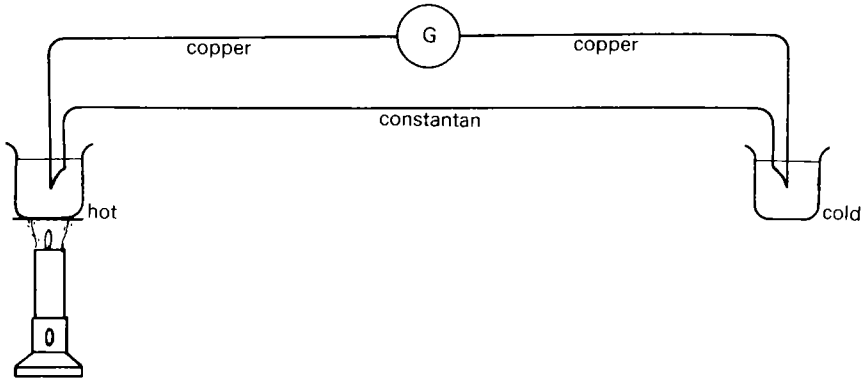


Fig. 11.9 A thermocouple – the Seebeck effect

and metal thermocouples have been utilized only in thermometry and for the detection and measurement of infra-red radiation. The e.m.f. obtainable using a pair of metals does not exceed a few millivolts, although this can, of course, be multiplied up by connecting thermocouples in series. Thermocouples of suitable *p*- and *n*-type semiconductors are found to be more effective, however, and from these, successful batteries and refrigerators to serve specialized purposes can now be made.

The Seebeck effect sometimes has a nuisance value in accurate work with potentiometers and bridges, by causing a small e.m.f. to appear in a circuit where it is not expected. This is a reason for preferring manganin to constantan as the material for a standard resistor, because the Seebeck e m.f. with manganin and copper is particularly small, whereas that with constantan and copper is comparatively large.

The connection between the two metals was described at the start of this paragraph as a 'join', and the term 'junction' was substituted later, as a concession to common practice. The initial distinction was deliberately made because this is not a true junction in the same sense as the boundary between the *p* and *n* sections of a single crystal in a diode or a transistor will be described as a junction in the next chapter. All that is required in a thermocouple is good electrical connection between the two metals, and these may be twisted together; they may even be separated from each other, and connected by a third wire, provided that the two ends of this have the same temperature. The effect is the same in all cases.

11.11 The variation of the Seebeck e.m.f. with temperature

The variation of the Seebeck e.m.f. of a thermocouple with temperature may conveniently be investigated by using a potentiometer circuit. The cold junction may be held at constant temperature in melting ice, while the temperature of the hot junction is taken up to say, 350 °C by placing it in a substance suitable for the temperature range, such as molten solder. Results of such an experiment are shown in Fig. 11.10 for two pairs of metals, copper-constantan, and copper-iron.

Both the curves shown in the diagram are approximately parabolic in shape, but it is evident that one pair of metals may be suitable for temperature measurement over a selected range, whereas another pair may not. A constantan-copper thermocouple may be used successfully between 0 °C and 300 °C. The circuit shown in the

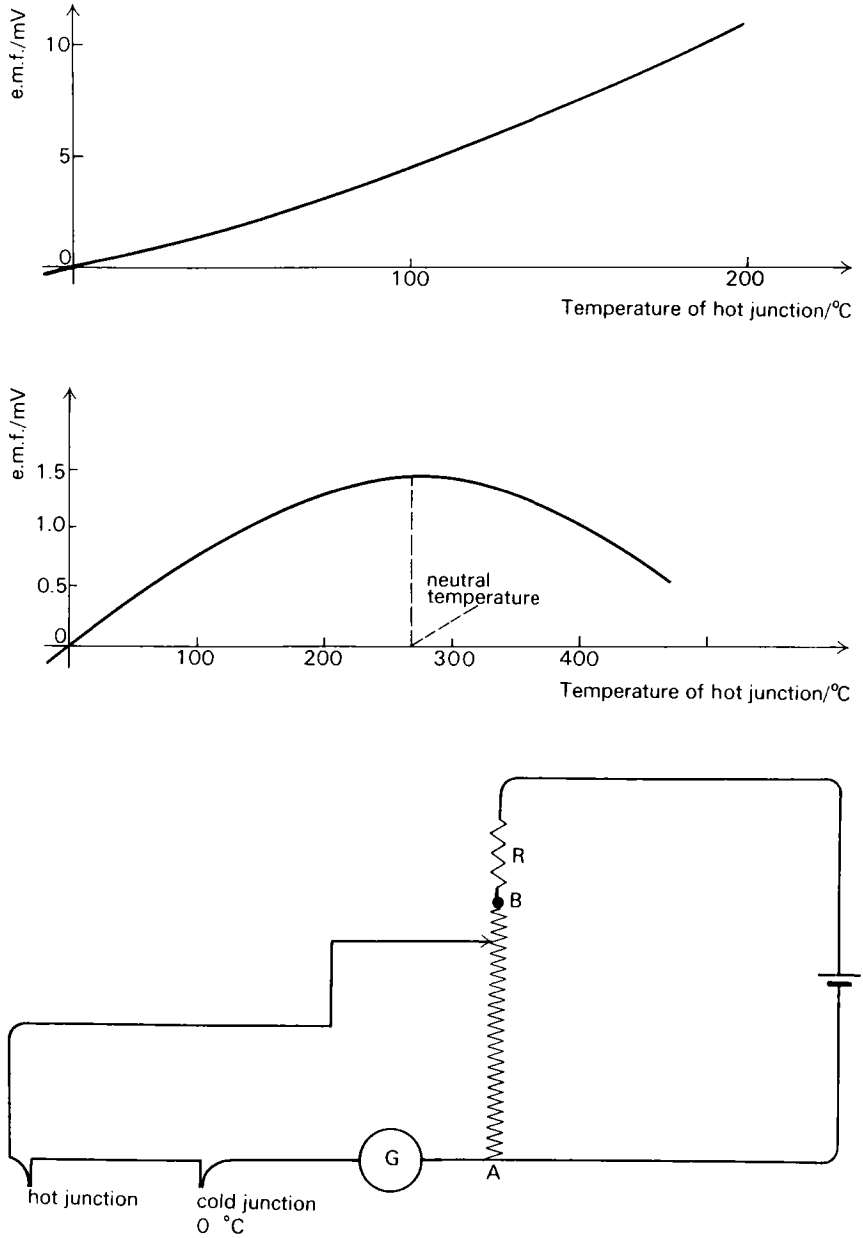


Fig. 11.10 Thermocouples. The variation of Seebeck e.m.f. with temperature of the hot junction (a) Copper-constantan, with the cold junction at 0 °C (b) Copper-iron, with the cold junction at 0 °C (c) Potentiometer circuit

diagram is suitable for thermometry, and the temperature may be deduced from the equation

$$E = a + b\theta + c\theta^2, \tag{11.1}$$

where E is the Seebeck e.m.f.; a, b, c are constants, and θ is the temperature of the hot junction measured in °C. If the cold junction is in melting ice, $a = 0$.

Among the advantages of the thermocouple as a thermometer are the large temperature range available, the very small physical size and low thermal capacity of the junction (which corresponds to the thermometer bulb), the rapid response to temperature changes, and the flexibility of the device, since the junction may be put in relatively inaccessible places.

11.12 The origin of the Seebeck e.m.f.

We have already pointed out that a thermocouple does not involve the close contact between one material and another which is a true junction. This is because the origin of the Seebeck e.m.f. is not at the joins of the two metals but is the resultant of two e.m.f.s, one in each of the two separate metals. When the junctions are at different temperatures, each metal has a temperature gradient in it, and this causes a potential gradient to appear in the metal as well. The need for having a pair of metals rather than one arises simply from the fact that it would be impossible to detect or to take current from the Seebeck e.m.f. in one metal alone, owing to the equal and opposite e.m.f. which would inevitably appear in the return half of the circuit.

The most obvious effect of a temperature gradient in a metal is thermal conduction. The average energy possessed by the conduction electrons at the hot end, in their random motion, is slightly greater than at the cold end, and this energy is continually being transferred down the temperature gradient by electrons exchanging energies over the distance of their mean free path. This is the transfer of energy which is thermal conduction, but it is not a transfer of charge. In addition to this, however, since the temperature gradient implies the existence of a slight velocity gradient, there is a tendency for more electrons to be moving away from the hot end in a given time than are moving towards it, with the result that the cold end acquires a negative charge. Thus, a potential gradient is set up in the metal, which stabilises owing to electrical conduction; i.e. the temperature gradient causes an electron drift towards the cold end, and the resulting potential gradient causes a compensating electron drift by conduction towards the hot end.

Several factors militate against metal thermocouples being very effective sources of a Seebeck e.m.f. The most obvious of these is the fact that the charge carriers are in all cases electrons, with the result that the cold end of the metal is always the negative end, and the e.m.f.s in the two metals of the thermocouple inevitably act in opposition to each other. Clearly, this is not the case with a thermocouple made up by joining together an n-type and a p-type semiconductor, and for such materials the Seebeck e.m.f. is often considerably larger.

11.13 The energy of a thermocouple

When the thermocouple is being used to supply current, it is acting as a continuous source of energy, and it derives this energy directly from heat. Thus, when current is flowing, but not otherwise, heat is taken in at some parts of the circuit and heat is ejected, in smaller amounts, at others; the balance turns into electrical energy. In view of the explanation which we have given of the origin of the Seebeck effect, attributing it to potential gradients separately set up in the two metals, we might expect the heat exchanges which provide the energy to occur separately in the two metals also. This proves to be only partly true, however, as there are two distinct forms of heat exchange involved, only one of which occurs along the lengths of the separate metals, while the other occurs at the junctions; the energy of the thermocouple comes from the combined effects of both. The first of these two to be discovered was the heat exchange at the junctions, and this is known as *the Peltier*

effect. The ejection or absorption of heat along the length of each separate metal was discovered later by Thomson (later Lord Kelvin) and it is known as *the Thomson effect*.

Both Thomson and Peltier are reversible heating effects: both are directly proportional to the current, and the direction of the energy conversion is reversed when the direction of the current is reversed; that is to say that, where current flowing in one direction causes heat to be taken in, reversal of the current causes heat to be ejected. We may compare this with the Joule heating effect in a resistor, which represents an irreversible conversion of electrical energy to heat, independently of the direction of the current, a result which is consistent with the presence of the squared term in the equation

$$P = RI^2, \tag{11.2}$$

where P is the rate at which heat is ejected in a resistance R.

11.14 The Peltier effect

The junction effect was discovered by Peltier in 1834. When current flows across the join between two metals, heat is absorbed at the join at a rate which is directly proportional to the current, and we can write

$$P = {}_1\Pi_2 I, \tag{11.3}$$

where ${}_1\Pi_2$ represents a coefficient, known as the Peltier coefficient, for the two metals 1 and 2.

To complete the circuit, the current must flow across the other join between the metals in the reverse direction, and heat is consequently ejected there. However, the Peltier effect is temperature dependent, with the result that, if the junctions are at different temperatures, there is a net conversion of energy, which contributes to the energy of the Seebeck e.m.f.

One simple way to demonstrate this effect is shown in Fig. 11.11. A current is passed from a battery through two junctions connected in series, and the connection is then switched across from the battery to a galvanometer. This registers a current for a short time, which may be attributed to the Seebeck effect, indicating that the two junctions are now at different temperatures whereas initially they were not.

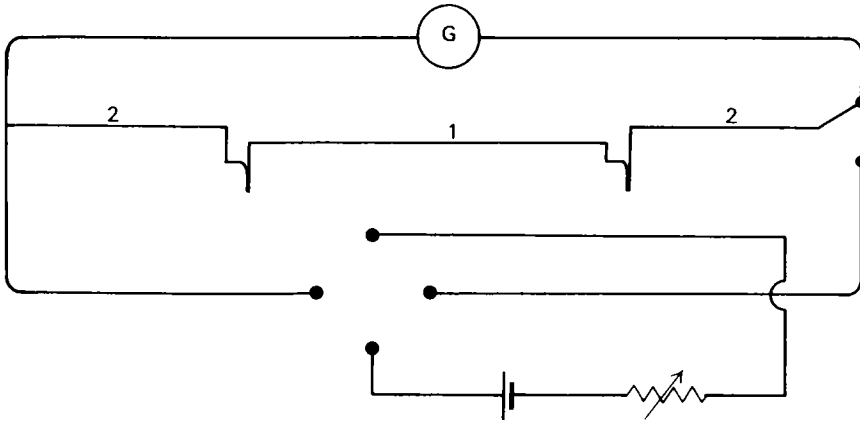


Fig. 11.11 A circuit to demonstrate the Peltier effect

We can reason that this temperature difference has been caused by the Peltier effects at the two junctions, which cooled one junction down and heated the other up while the current was flowing. This argument might be spurious, of course, the temperature difference being the result of unequally distributed Joule heating effects. However, this can be checked by repeating the experiment with the direction of the current reversed, in which case the galvanometer deflection reverses also, in accordance with the reversible nature of the Peltier effect.

The cause of the Peltier effect can be sought in the energy of the conduction electrons of a metal, relative to the ions in its crystal lattices. This is unlikely to be the same level of energy in one metal as it is in another, with the result that electrons must take in or eject some energy when an electric current transfers them from one metal to another. Because their energy depends to a small degree on temperature, the Peltier coefficient varies with temperature, being larger where this is high.

11.15 The Thomson effect

The other heat exchange, which occurs separately along the length of each metal, results from the temperature gradient which has been set up in it by putting the two junctions at different temperatures. If current flows in a metal wire when there is such a temperature gradient in it, heat is absorbed at a rate which is directly proportional to the current, and if the direction of the current is reversed, heat is ejected. When a thermocouple is in circuit, therefore, heat is absorbed in one metal and ejected from the other, but the two effects differ in magnitude between the two metals, and a net conversion of energy results which makes its contribution to the energy supplied by the Seebeck e.m.f.

The cause of the Thomson effect appears to be that, when an electric current flows towards the hot end of the metal, this is in reality a drift of conduction electrons towards the cold end. These electrons, having slightly greater thermal energy where the metal is hotter, carry this energy with them over short distances, but it is repeatedly dissipated at each successive collision in the lattice, and so released as heat. Conversely, if the direction of the current is reversed, to carry electrons up the temperature gradient, they take in the extra thermal energy at successive collisions.

11.16 Thermoelectric effects in semiconductors

The thermoelectric e.m.f. set up in a *p*-type semiconductor when there is a temperature gradient in it makes the cold end positive for the same reason that it makes the cold end of a metal or of an *n*-type semiconductor negative. Consequently, in a thermocouple which is made up from a *p*-type and an *n*-type semiconductor, the e.m.f.s in the two components act in the same direction round the circuit. For this and some less obvious reasons, the Seebeck e.m.f. in such a thermocouple is many times larger than it is for a pair of metals, being of the order of 500 μV/°C as compared with, say, 40 μV/°C for copper and constantan.

A further advantage of the semiconductor is the fact that, unlike the case of a metal, thermal conduction is due largely to exchange of energy between ions in the crystal lattices and not to behaviour of the charge carriers, with the result that the semiconductor can be a relatively good electrical conductor while being, at the same time, almost a thermal insulator, conveying very little heat directly from the hot to the cold junction. For these reasons, the Seebeck effect can be utilized, using semiconductors, to make practical thermoelectric generators, although on a rather small scale.

A further feature of thermoelectricity in semiconductors is that the Peltier effect is large, whereas the Thomson effect is not, which makes them adaptable for use as Peltier refrigerators, taking in heat at the cold junction and ejecting it at the hot. The reason for this is that the energy level of the conduction electrons in the *n*-type material, relative to the ions of the lattices, is much higher than the corresponding energy level of the electrons which are responsible for moving the positive holes about in the *p*-type material. These electrons are still involved in the covalent bonding, whereas the free electrons of the *n*-type are not, which accounts for the difference. Consequently, when a current takes electrons from *p*-type to *n*-type, a comparatively large amount of energy is taken in at the junction.

The thermoelectric generator is built on the same principle as a thermopile, being a stack of blocks of alternately *p*- and *n*-type elements connected in series by copper straps. In addition to low thermal conductivity and high thermoelectric power, the materials are required to tolerate high temperature, so that a sufficient temperature difference between hot and cold junctions may be maintained. The electrical resistance is kept low by doping the materials heavily with donor and acceptor atoms respectively, and by having blocks of large cross-section.

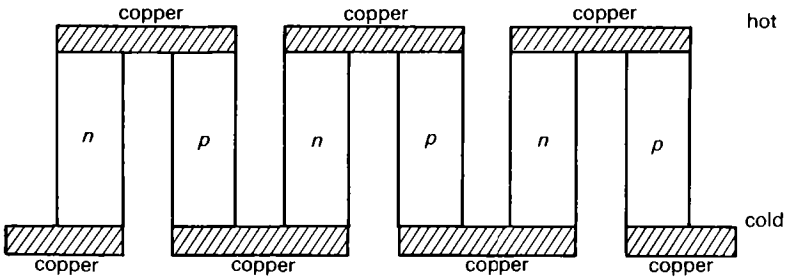


Fig. 11.12 A thermoelectric generator

The Peltier refrigerator is made on the same principle, from similar materials. In this case, a heavy current is passed, from an external source, in opposition to the direction of the Seebeck e.m.f., with the result that the arrangement acts as a heat pump, taking in heat at the cold junction and ejecting it at the hot junction. The principle of construction is similar to that of the generator, with the addition of metal fins at the junctions to aid the heat transfer.

Chapter 12

Diodes and triodes

12.1 Rectifiers

Electrons emitted into a vacuum and electrons or holes released in a semiconductor can be controlled in ways which are impossible with the vast number of conduction electrons moving about in a metal, and this makes a variety of useful devices possible. The most obvious of these is the rectifier, which will convey current through it readily in one direction but refuses to carry it, or offers a high resistance to it, if the direction is reversed. *The solid state rectifier* is a slice of a single crystal which has been doped in such a way that it is unsymmetrical, being *n*-type on one side and *p*-type on the other. It has low resistance towards current which carries the electrons and holes towards the boundary between the two regions, and very high resistance towards current which carries them away from it. *The thermionic rectifier* is a diode in which the electrons are emitted into a vacuum from a hot cathode and drawn across to the positive anode; since no electrons are emitted from the anode, no current flows if the polarity is reversed.

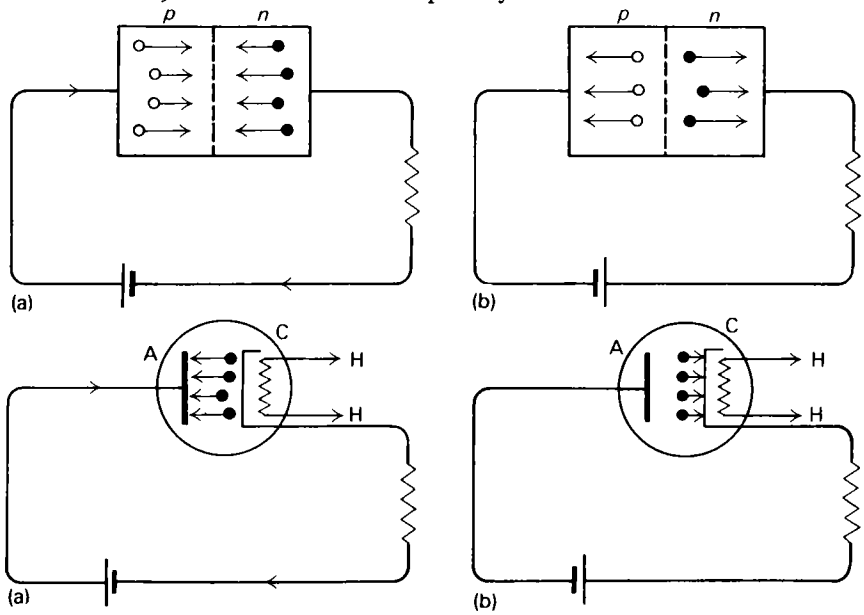


Fig. 12.1 The solid state and the thermionic diode. Rectifying action (a) Forward polarity (b) Reverse polarity

12.2 The junction diode

There are several different techniques for making an *n–p* junction diode, of which perhaps the simplest is the alloying process, which was developed first and which

is used to make small germanium diodes. This starts with a thin wafer, a few millimetres square and a fraction of a millimetre thick, taken from a germanium crystal which has already been doped to be *n*-type. A pellet of the trivalent element, indium, is then melted on one face of the crystal wafer and allowed to solidify. During this process, the indium and germanium alloy together in the vicinity of the pellet and the acceptor indium atoms penetrate the germanium lattice, occupying places in it. The resulting infusion of holes converts the crystal from *n*- to *p*-type on the side which has the indium, because most of the electrons which originally made it *n*-type are taken up in annihilating a proportion of the newly introduced holes, leaving the surplus holes to become the majority charge carriers. Connecting wires are now soldered on to the indium pellet and to the *n*-type part of the crystal slice, which is then surrounded with protecting grease and enclosed in a small capsule. This must be made light proof, because the diode is photosensitive; the action of light would be to release intrinsic charge carriers where they are not wanted.

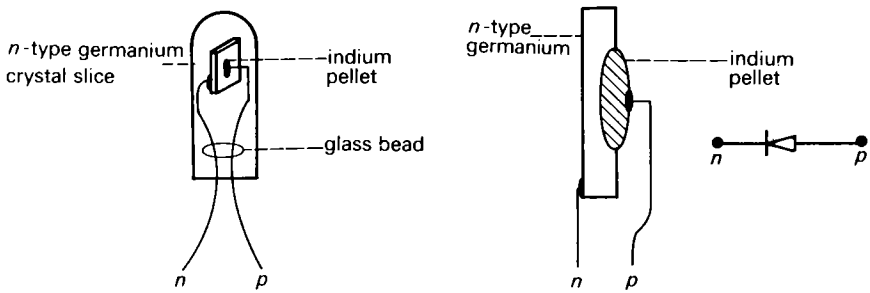


Fig. 12.2 A small germanium diode

The diode owes its low resistance in the forward direction to the fact that it has a comparatively large area of cross-section perpendicular to the current flow, and a very short path for the current to follow through the small thickness of the crystal. It owes its very high resistance in the reverse direction to the formation of a thin layer at the *n*–*p* junction which is devoid of charge carriers. This is called the depletion layer. It plays an essential part in a number of semiconducting devices, including the transistor and the photodiode.

The depletion layer

This results from the migration of charge carriers across the junction, holes reaching the *n*-type region from the *p* side and electrons reaching the *p*-type region from the *n* side. In both cases, when an electron meets a hole, mutual destruction as charge carriers results (meaning that the free electron occupies the bonding vacancy which is the hole), and so a thin layer at the junction becomes bereft of charge carriers. This is the depletion layer, and it exists at the junction before any p.d. is applied; a reverse p.d. merely serves to strengthen it, but a forward p.d. tends to destroy it.

When an electron deserts the *n* region in this way, and fills a hole in the *p* region, it leaves behind it its positively charged donor atom, locked in its place in the crystal lattice. Similarly, for every hole which leaves the *p* region, a negatively charged acceptor atom is left behind. Consequently, there is a potential gradient in the depletion layer, positive on the *n* side and negative on the *p* side, and this

represents a potential barrier, which acts against further migration of electrons and holes across the barrier. For this reason, the depletion layer is very thin.

The strength of the potential barrier depends on the material of the crystal, being in the region of 0.15 V for germanium and 0.6 V for silicon.

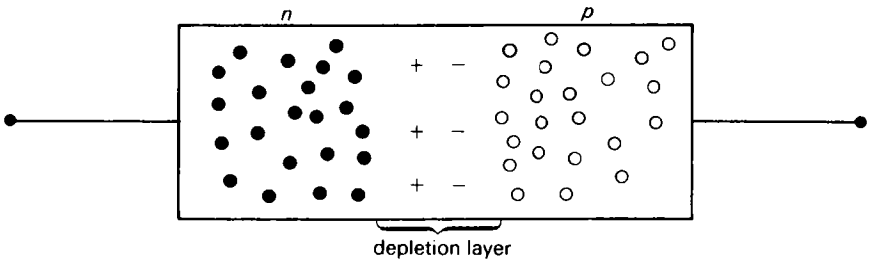


Fig. 12.3 An open-circuited diode. The depletion layer

Figure 12.4 represents an attempt to illustrate how the depletion layer and the potential barrier come about. Donor and acceptor atoms are represented here as having the four valency bonds required by the crystal lattice. Since this means that a donor atom has parted with the fifth electron of its outer shell, it is shown as having positive charge. Similarly, the acceptor atom has added an extra electron to put four into its outer shell, thus acquiring negative charge and releasing a positive hole somewhere in the lattice. The vast majority of atoms in the lattice are, of course, neutral atoms of germanium or silicon, but these are not represented in the diagram.

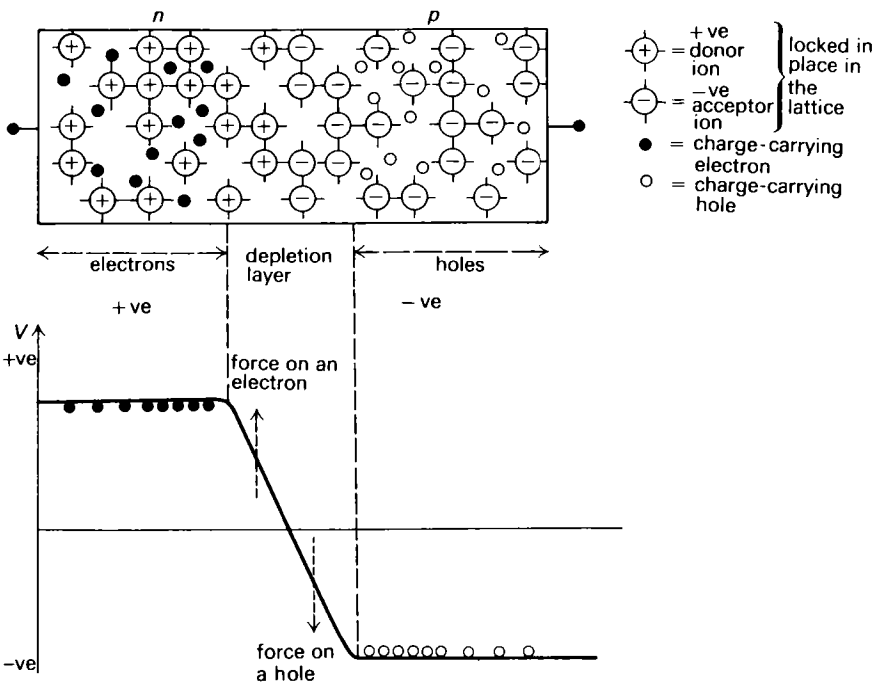


Fig. 12.4 Diagram representing the depletion layer and the potential barrier in an open-circuited diode

Perhaps it should be pointed out that the potential barrier at the junction is not a true e.m.f., as there is no source of energy there which could provide a continuous current. What we have is charge locked in the insulating depletion layer, positive at the *n* side and negative at the *p* side, with a semiconducting section of crystal on each side of this layer. This resembles a charged capacitor, rather than a source of e.m.f.

The ideal diode would have a characteristic behaviour of the kind illustrated in Fig. 12.5(*a*), having a constant and low resistance in the forward direction and infinite resistance in the reverse direction. The actual diode, however, behaves as in the second diagram; it requires a forward p.d. of appreciable size before forward conduction properly starts, and it loses its insulating property in the reverse direction if the reverse p.d. becomes too great.

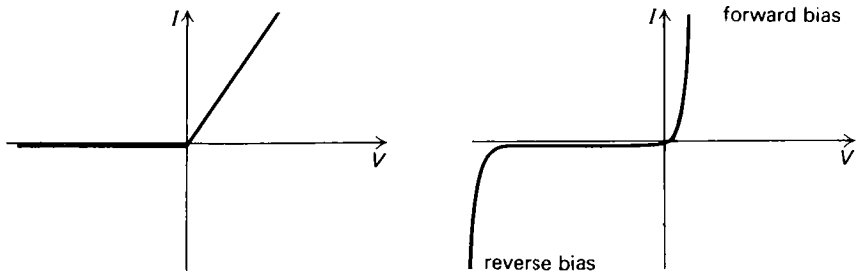


Fig. 12.5 Practical and ideal diode characteristics (a) An ideal diode (b) A germanium or silicon diode

Typical characteristics are drawn in Fig. 12.6 for two small duty diodes, one of germanium and one of silicon. (A magnified scale has been used for plotting reverse p.d. and current, because this current is so small that otherwise it would not show up at all.)

The shape of the forward characteristic is due to the potential barrier across the junction. Before forward conduction can be fully effective, the barrier p.d. must first be overcome by applying a forward p.d. of sufficient size, and this takes about 0.15 V for germanium and 0.6 V for silicon. For this reason, germanium is the more acceptable of the two for rectifying a small p.d.

The reverse resistance would be infinitely high, were it not for intrinsic conduction. The presence in the crystal slice of a small number of intrinsic charge carriers means that there are a few holes and electrons on the wrong sides of the *n–p* junction, holes on the *n* side and electrons on the *p* side, and these are able to carry a small current through the depletion layer when reverse bias is applied. These intrinsic charges are also responsible for the rapid decrease in the resistance which results if the reverse p.d. is taken up too high. This amounts to breakdown of the diode in its function as a rectifier. Because the depletion layer is very thin, and the whole of the applied p.d. appears across it, a very powerful electric field is set up in it by a reverse bias. If this field becomes too great, the minority carriers gain sufficient energy from it to release further intrinsic electrons and holes in the crystal lattice; these repeat the process and an avalanche of charge carriers is released. Consequently, intrinsic conduction takes over.

A silicon diode has a higher reverse resistance than a germanium one because its intrinsic conductivity at room temperature is much lower. The maximum reverse p.d. which a diode can tolerate without breakdown depends on the particular diode, but it is essentially greater for silicon than for germanium.

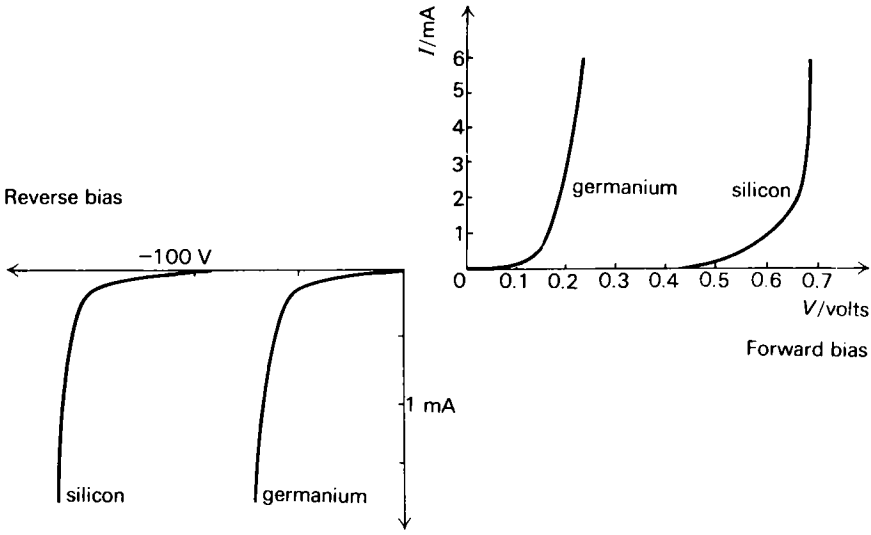


Fig. 12.6 Characteristic curves for small duty germanium and silicon junction diodes

The effect of excessive temperature rise

A junction diode must be protected against the possible effect of an excessive temperature rise, whether from external causes or due to the current itself. This is because the reverse current, being due to intrinsic conduction, is highly temperature dependent. The particular danger here is that the effect would build up, the temperature rise caused by the initial current releasing further intrinsic carriers, and so on – an effect known as thermal runaway. For this reason, heavy duty diodes

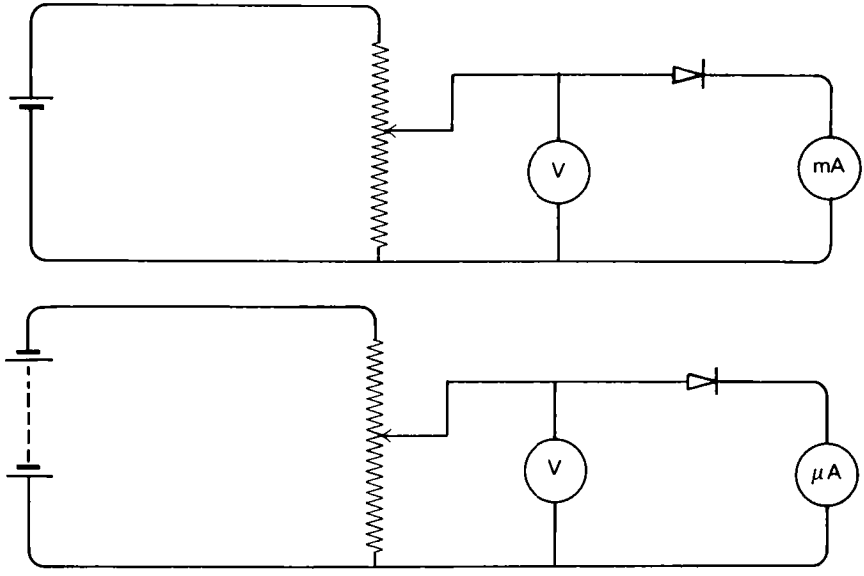


Fig. 12.7 Circuits for plotting the characteristics of a diode (The voltmeter records the p.d. across diode and microammeter in series, and may require some correction to allow for the p.d. across the meter) (a) To plot the forward characteristic (b) To plot the reverse characteristic

are thermally connected to some form of heat sink, which is a thick metal plate to conduct the heat away. Here, again, silicon is much less open to this danger than germanium, and will tolerate a higher temperature without a catastrophic release of intrinsic charge carriers. A heavy duty silicon diode can carry currents up to 100 A and more; this type of diode is doped by a diffusion process, the silicon being exposed to phosphorus vapour to make it *n*-type, or to boron vapour to make it *p*-type.

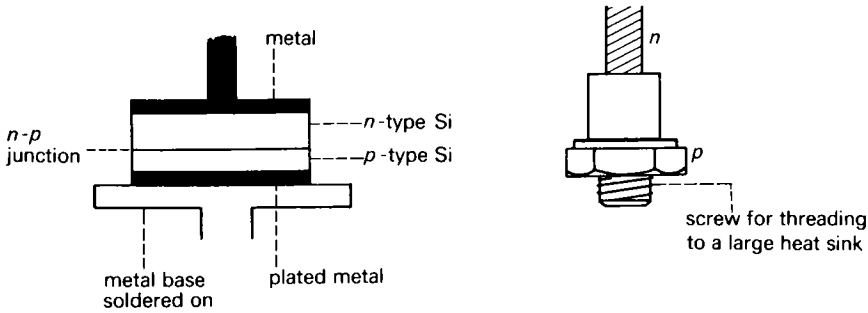


Fig. 12.8 A silicon power rectifier (a) The diode (b) The diode in its casing

12.3 The point contact diode

Because the junction diode, particularly one of silicon, requires a forward p.d. of appreciable size to be applied before it is able to conduct at all, it follows that it is useless for the rectification of very low potential oscillations, such as are met in the early stages of a radio receiver. In these circumstances, a point contact diode is used instead. This consists of a tungsten wire, which makes contact with a thin wafer of *n*-type germanium or silicon, and it works because a *p–n* junction appears in the crystal in the immediate vicinity of the contact.

The point contact diode has a respectable history behind it, being a natural development from the polycrystalline material with a metallic catwhisker contact, which was the rectifier in early crystal radio receivers. It is available for use only with small currents, and its forward resistance is not very low.

12.4 Thermionic emission – the thermionic diode

The thermionic diode has the longest history among types of rectifier. It was invented by Fleming in 1904. Electrons are injected into a vacuum from the surface of a metallic electrode by raising its temperature, and are drawn away from it towards a second electrode, made positive relative to the first. Clearly this is a unidirectional device; since electrons are emitted only from the hot electrode the reverse resistance is virtually infinite. Because the diode conducts only when the heated electrode is made negative, this is referred to as the cathode, the positive electrode being the anode.

A disadvantage of the thermionic diode, as compared with the other types of rectifier, is the necessity to heat the cathode, which makes it into a device having more than two connections, and requiring a separate heating supply. The type shown in Fig. 12.9(*a*) has an indirectly heated cathode, in the form of a nickel sleeve, and a separate tungsten heating spiral inside this sleeve and insulated from it. The nickel is coated with a mixture of barium and strontium oxides, because this emits electrons effectively at a lower temperature.

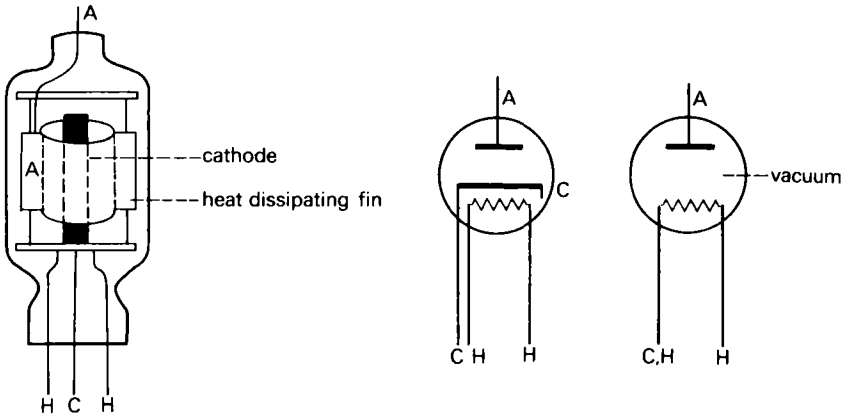


Fig. 12.9 A thermionic diode (a) An acorn diode, about twice actual size (indirectly heated) (b) Indirectly heated cathode (I.H.) (c) Directly heated cathode (D.H.)

The horizontal part CD of the characteristic curve represents saturation, the electrons being drawn across to the anode as fast as the hot cathode emits them. Naturally, if the cathode is run at a lower temperature, thus reducing the emission, saturation is reached with a smaller current. The only practical significance of this horizontal part of the curve, however, is that it marks the limit of what the diode can do; in practice it is run well below saturation, since otherwise the cathode

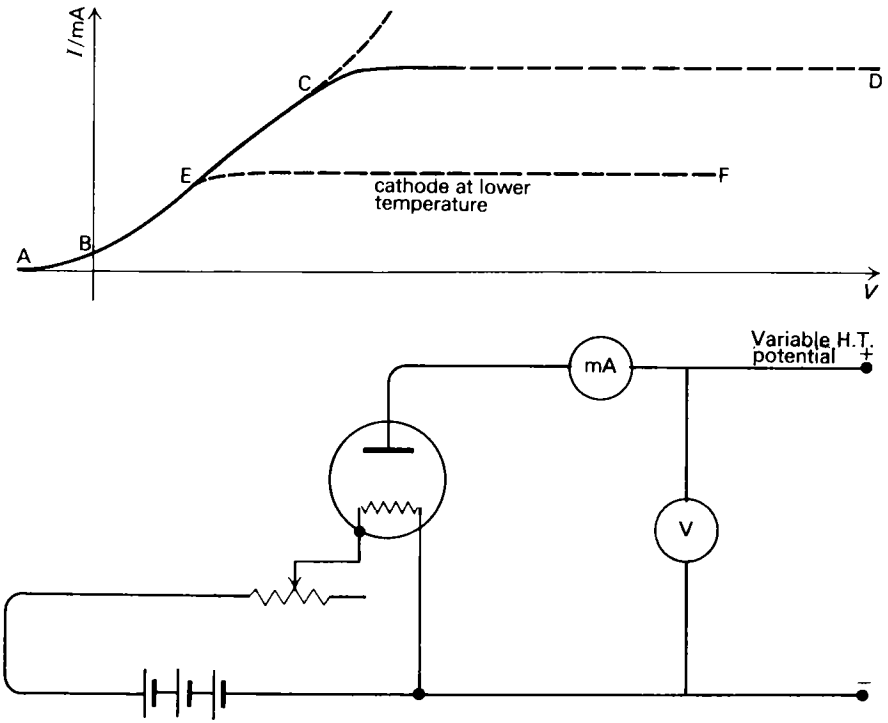


Fig. 12.10 I– V characteristics for a thermionic diode, and a suitable circuit, using a directly heated type (a) (b)

coating may suffer damage. For this reason, the most suitable type of diode to use if the whole curve is desired for instructional purposes is a directly heated one, which has a more robust tungsten cathode.

The space charge

Provided that saturation has not been reached, the value of the current in a thermionic valve, whether diode or triode or something more complicated, does not depend on the electron-emitting ability of the cathode, but on the potentials applied to the electrodes. The function of the cathode is merely to maintain a concentration of electrons in the vacuum by thermionic emission, continuously

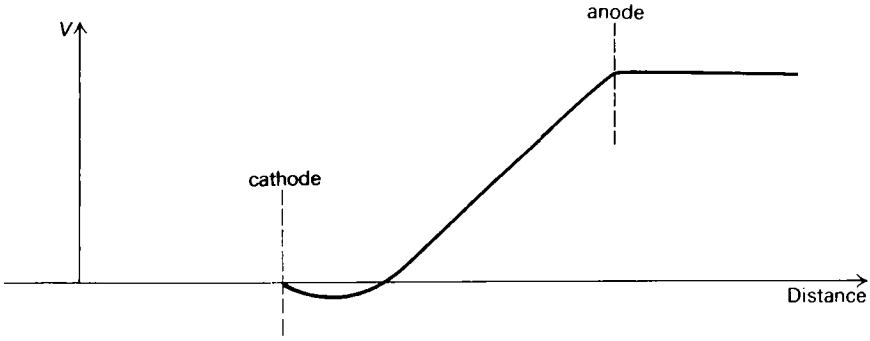


Fig. 12.11 The potential distribution in a forward biased thermionic diode

replacing those which are drawn off to the electrodes in the process of conveying the current. In performing this function, the cathode provides an electron population in the vacuum which is particularly concentrated near the cathode surface, producing a region of negative charge, which is termed a space charge. This represents a negative potential barrier near the cathode, which limits the further emission of electrons, so that an equilibrium is maintained and the tube current is not drawn directly from the cathode but from the space charge.

12.5 The thermionic triode and the transistor

The thermionic triode was invented in 1907 by de Forest, who inserted a wire mesh electrode, the control grid, in the vacuum between the cathode and anode of a diode. For forty years this was the basic device in electronics, more complicated valves, such as the tetrode and the pentode, merely being modifications of it. Its solid state counterpart, the transistor, was invented in 1945 by Bardeen and Brattain, who added a second contact whisker to a point contact diode. This led two years later to the production, by Shockley, of a junction transistor based on the construction of a junction diode, and this is the modern transistor; the point contact type, unlike the point contact diode, is now obsolete. All these were American inventions.

What gives both the vacuum triode and the solid state transistor their importance is their amplifying function, and this can be easiest illustrated in terms of the thermionic triode. When a forward p.d. is applied between cathode and anode, the current of electrons crossing the vacuum between these two electrodes derives its power from the source of the p.d. The size of the current, however, is controlled by a quite separate p.d. applied between cathode and grid; this is because the electrons must pass through this grid on their way to the anode, and the grid is placed very close to the cathode. This gives an arrangement whereby a low power signal in one circuit, which takes the form of a modulation of potential or current,

results in a high powered modulation of the current in a second circuit. This is amplification. As a comparison, a transformer can be used to increase the potential of an alternating signal, but only at the expense of the current; there is no power gain in this case, as there is in a true amplifier.

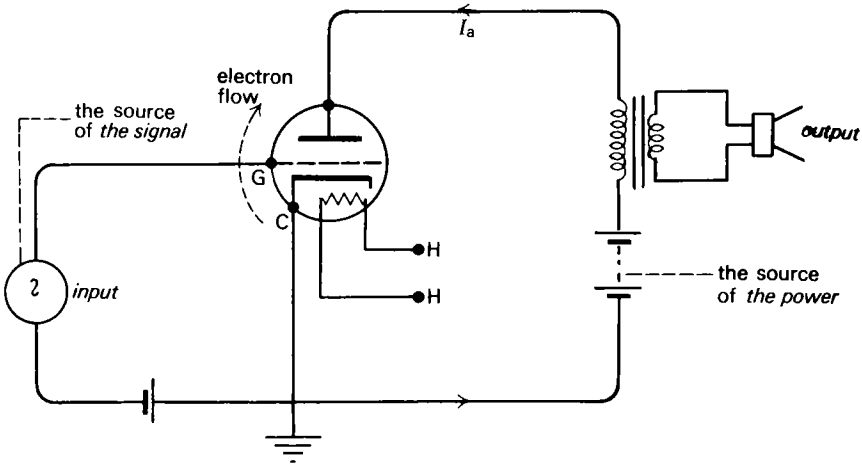


Fig. 12.12 A thermionic triode as an amplifier

Given an amplifier, there are various applications for it, one of which is distortionless (i.e. linear) amplification of an alternating signal potential, which might be obtained by the circuit shown here. Another possibility is to use the input signal merely to switch the output current on and off, in the manner of a high speed relay, which is an application of the transistor in a digital computer. The amplifying action may be made to include deliberate distortion of a waveform, which is used in pulse shaping, or in the integrating and differentiating circuits of an analogue computer. Again, a part of the output power may be fed back to the input; if the phase relations are right, this serves to build up the strength of the input signal, a process known as positive feedback, and it is used to maintain an electric oscillation.

12.6 The transistor

In a transistor, what corresponds to the vacuum of the thermionic triode is the high resistance depletion layer of a reverse biassed *n–p* junction, and the battery which provides this bias is the source of the power. Let us imagine that we start with a junction diode which is biassed in this way, so that all the extrinsic charge carriers present in the doped crystal are, having respect to their signs, at the bottom of the potential gradient which exists across the depletion layer. The electrons in the *n* region are on the positive side of the gradient, and the positive holes in the *p* region are on the negative side. In both cases this is where the potential gradient would put them, and there is therefore no tendency for current to flow (Fig. 12.13).

Suppose, now, that a circuit is set up which is capable of injecting a current of positive holes by some means into this depletion layer at the *n* side of the crystal, which is normally inhabited by electrons. These injected invaders would be in a different situation from the *n* region's electrons, being, due to their positive sign, not at the bottom of the potential gradient, as it affects them, but at the top of it; they would be attracted across the junction by the potential gradient, gaining considerably in energy in the process, and this energy would derive from the

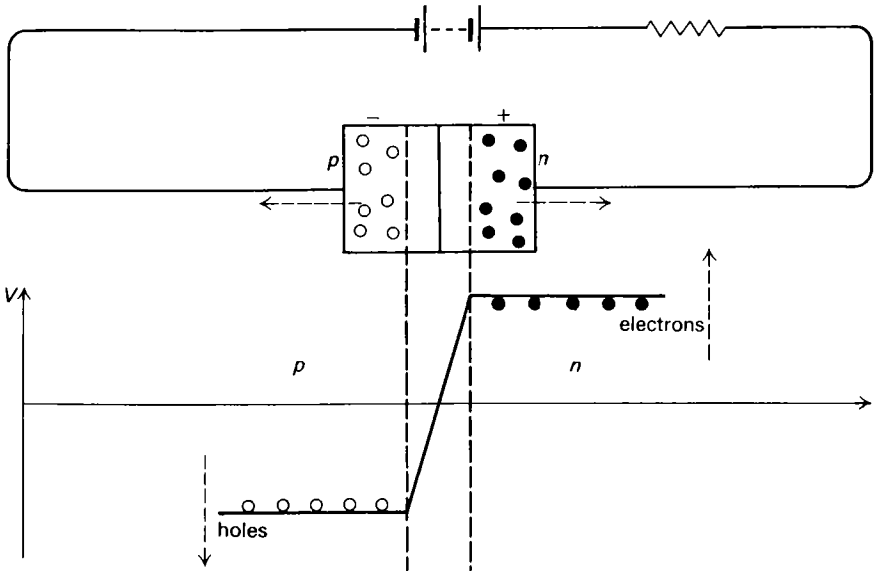


Fig. 12.13 A reverse biased junction diode

biassing battery (Fig. 12.14). If the current of holes injected from the input circuit is varied, to represent an input signal, this variation is present, with enhanced energy, in the current of holes arriving at the *p* side of the junction, and this represents the output signal. Thus the input circuit controls the current variation, while the biassing battery supplies the power. This is the basis on which the *p–n–p* transistor works.

Alternatively, instead of injecting holes at the *n* side of the depletion layer, electrons may be injected at the *p* side, which is normally inhabited by positive holes. In this case, the electrons, being negative, would be attracted up the potential gradient to the *n* side of junction, with a similar gain in energy. This is the basis of the *n–p–n* transistor.

Taking the example of a *p–n–p* transistor, the way in which the holes are made available for injection into the depletion layer of the *n–p* junction is by the addition of a second *p*-type region to the crystal very close to the first, with the result that there are now two *p*-type regions in it, separated from each other by a very thin *n*-type layer. One type of *p–n–p* transistor, for example, is made by an extension of the process already described for the alloy junction germanium diode pictured in Fig. 12.2. In this case, a second indium pellet is fused on the other face of the *n*-type crystal slice, additional to the one which makes the diode. This is an alloy junction *p–n–p* germanium transistor.

Because of its function, the added *p*-type region from which the holes are injected is called the emitter, and the original *p*-type region, which finally collects the emitted holes, is called the collector, while the *n*-type crystal wafer is the base.

12.7 An *n–p–n* silicon planar transistor

Since the successful action of the transistor depends entirely on its extrinsic charge carriers, there are obvious advantages in choosing silicon in preference to germanium, as intrinsic conduction is negligible in silicon at room temperature and a considerable temperature rise can be tolerated without intrinsic conduction

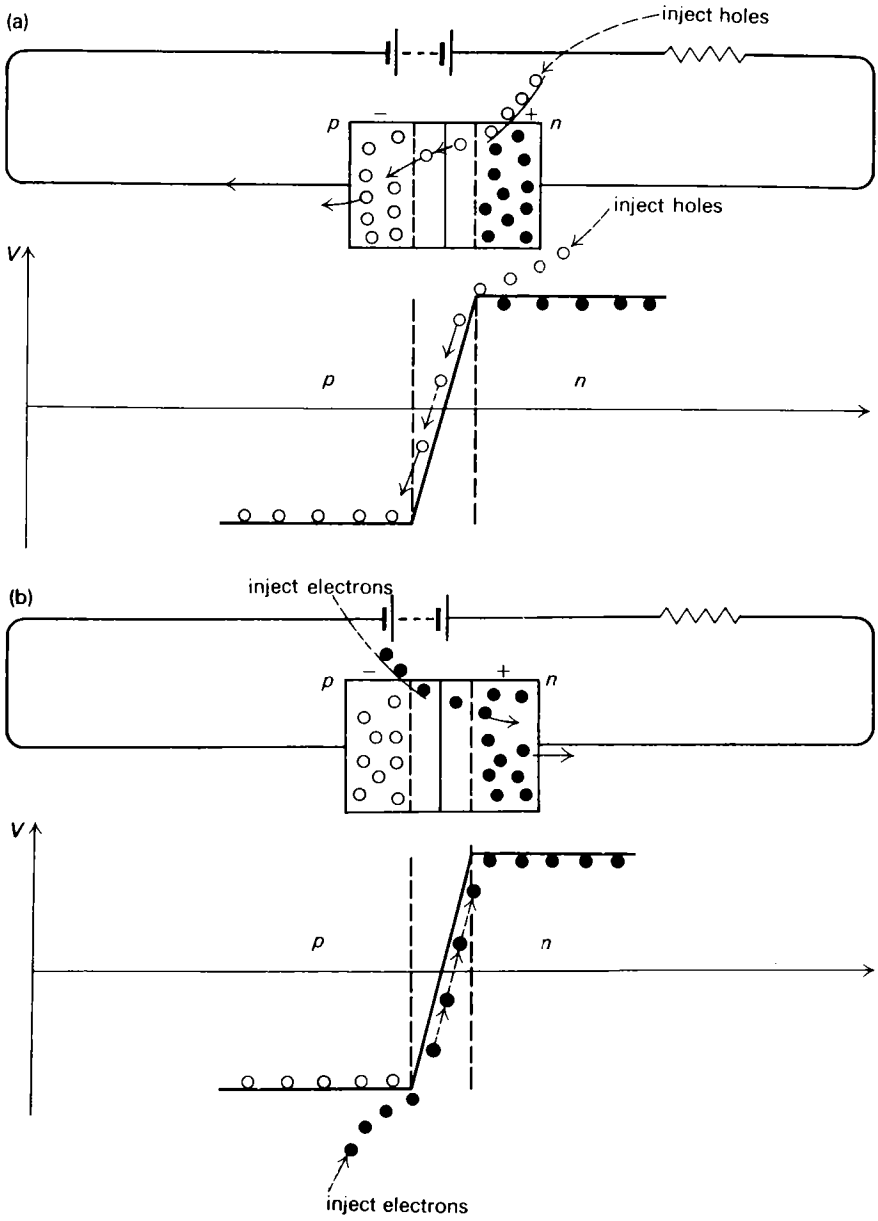


Fig. 12.14 Diagram representing the effect of injecting charge carriers into the depletion layer of a reverse biassed junction (a) *p-n-p* (b) *n-p-n*

becoming serious. Most of the earlier transistors were germanium simply because it was an easier material to work with in manufacture, having, for example, a lower melting point. The silicon planar transistor, however, which is usually of the *n–p–n* type, has now largely displaced the small germanium ones, and the remainder of this description will consequently refer to the *n–p–n* type, in which the important charge carriers from the *n*-type emitter are not holes, but electrons.

When the transistor is correctly connected in a circuit, a reverse bias of several volts applied between base and collector strengthens the depletion layer at the

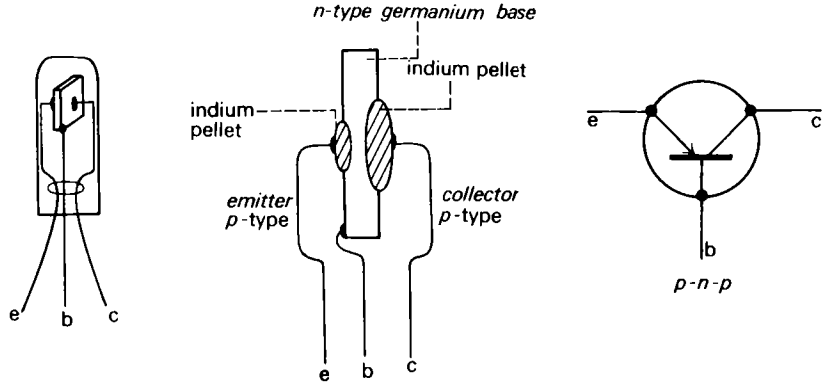


Fig. 12.15 A small alloy junction *p-n-p* germanium transistor

base-collector junction and produces a steep potential gradient in it. The electrons from the *n*-type emitter are injected into this depletion layer from the base side, which is negative with respect to the collector, and the electrons are consequently attracted by the potential gradient into the positive collector. The electrons do not leave the emitter, however, without being given some electrical encouragement to do so; the presence of the emitter region in the crystal produces in it a second *n–p* junction, and there would be a depletion layer here also, with a potential barrier across it, if this junction were left unbiassed. For the emitter to inject its electrons into the depletion layer of the base-collector junction, a small forward bias must first be applied between emitter and base, sufficient to break down the potential barrier across the junction. This biassing p.d. must be that required for a forward biassed diode, i.e. about 0.6 to 0.7 V. (It would be about 0.15 V for germanium.)

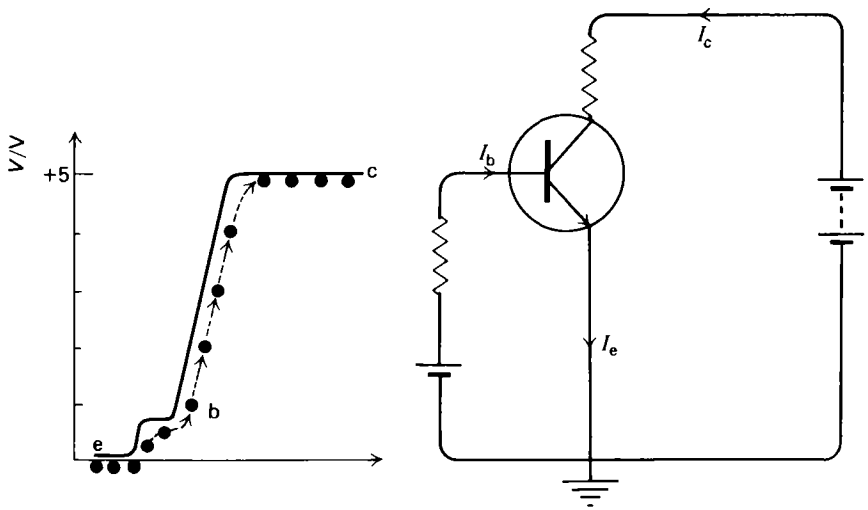


Fig. 12.16 A typically biased *n-p-n* silicon transistor

There is an important condition to be satisfied in the construction of the transistor; it is essential that the base region, where it separates the emitter from the collector, shall be made extremely thin – less, say, than 0.1 mm thick, and preferably even thinner. Otherwise, the crystal would not be a transistor at all,

but would operate, merely, as two diodes connected back to back; all the electrons reaching the base from the emitter would complete their circuit, not by travelling on into the collector, but by flowing out through the base connection, as they would in a forward biassed diode. For a good transistor, at least 99 per cent of the electrons escape doing this, and cross the thin base region instead, to reach the depletion layer of the base-collector junction; here the forward potential gradient draws them into the collector. To summarize, we can write

$$I_e = I_b + I_c, \tag{12.1}$$

where I_e, I_b, I_c are the currents flowing in the emitter, base, and collector connections, respectively.

For the planar silicon transistor, the base region is exceptionally thin, and the ratio I_c/I_b is in the range 150 to 500, so that the emitter and collector currents are very nearly identical, only a very small proportion of the emitter current escaping from the base. (For an alloy-junction germanium type a more typical value would be 50.)

It will be noticed that the only charge carriers which are concerned with the useful current in the *n–p–n* transistor are the electrons which come from the emitter. The holes in the base and the electrons in the collector are required only in order to produce the essential depletion layer at the base-collector junction. Nevertheless, part of the current which flows from emitter to base is due to holes from the base which flow into the emitter. This is not a useful current – rather the reverse – and in order to keep it small the base is only lightly doped.

12.8 The construction of an *n–p–n* silicon planar transistor

A silicon planar transistor is manufactured as one of about a thousand similar transistors, all produced in the same sequence of processes from one single slice of a silicon crystal which has been doped to be *n*-type. This slice, which is about 3 cm square, becomes the collector (not the base) of each of the finished transistors, and a single transistor cut out from it is only 0.5 mm square. The base region, separating emitter from collector, is a very small fraction of a millimetre thick.

The basic process of manufacture is to produce the *p*-type base regions of the transistors by exposing chosen areas of the surface of the *n*-type crystal slice to an atmosphere of boron vapour, while the rest of the surface is masked by a protecting coating. Some boron settles on the exposed areas, and the slice is then heated, which causes the boron to penetrate the crystal lattice to the required depth. The smaller areas in the middle of each base region which are to become the surfaces of *n*-type emitter regions are then exposed to phosphorus vapour, while the remainder is protected, and the crystal is heated again. Finally, aluminium is allowed to coat selected areas of emitter and base, in order that leads can be bonded to them. The crystal is cut into separate transistors, and the collector of each is soldered in place on the surface of a metal stem.

As with all semiconducting devices made from single pure crystals, extreme purity is very important, and the conditions observed throughout the successive processes, in order to maintain this purity, are very stringent.

12.9 The transistor in an amplifying circuit

It can now be seen how the action of a transistor can be used to produce amplification. The input signal is applied in the form of a small modulation of the bias p.d. between base and emitter. This modulates the current of charge carriers crossing from emitter to base, and so produces an output signal in the form of a modulation

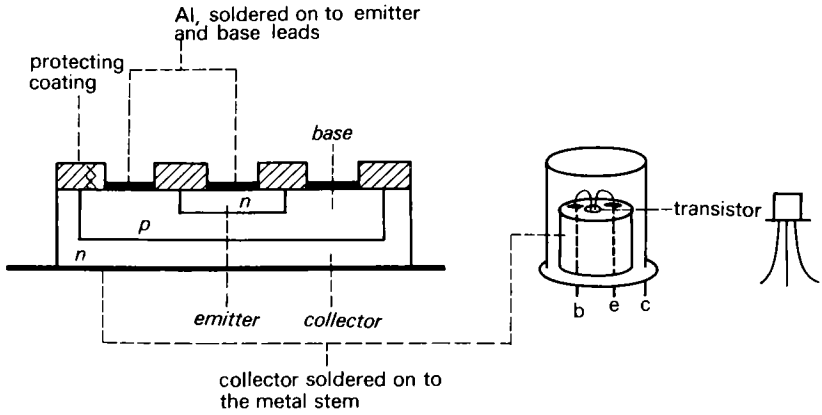


Fig. 12.17 A silicon planar *n-p-n* transistor (a) The transistor crystal slice, much magnified (b) The transistor in position (c) Actual size

of the current flowing in the collector circuit. This involves a considerable power gain, derived from the collector battery, due to the comparatively large drop in potential (typically, 5 V) when the charge carriers cross the depletion layer from base to collector.

A simple transistor amplifying circuit is drawn in Fig. 12.18. This has the emitter connected to both the input and the output circuits, and is known as a common emitter circuit.

Comparing the circuits of Figs. 12.12 and 12.18 it can be seen that the emitter, base, and collector of the transistor correspond, respectively, to the cathode, grid, and anode of the thermionic valve.

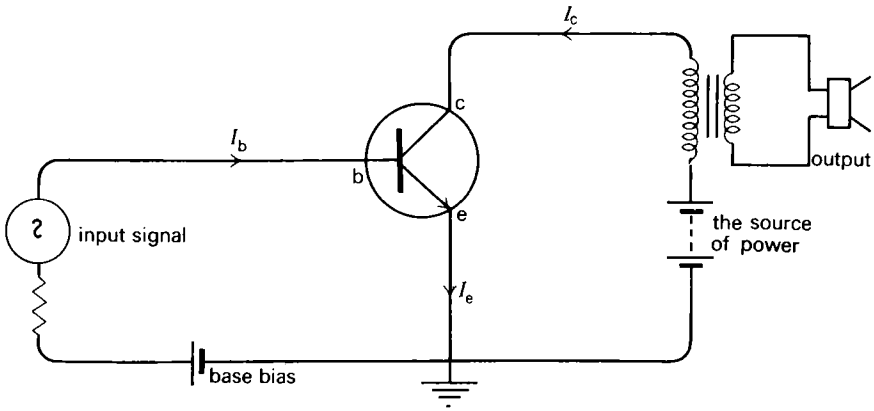


Fig. 12.18 A simple amplifying circuit, drawn for an *n-p-n* transistor

12.10 A load resistor

When, as is usually the case, several stages of amplification are required, one stage could be coupled to the next by means of a transformer, i.e. the loudspeaker in the last diagram would be replaced by the input circuit of the next transistor. A more common practice, however, is to use a load resistor and a coupling capacitor instead, and a circuit of this type is shown in Fig. 12.19, with a simple arrangement as well for providing the bias potentials.

The load resistor R is inserted in series between the collector of the transistor and the supply battery. This permits the bias potential to be applied to the collector whilst allowing the collector potential to vary in response to the variations in current which represent its output signal – which it could not have done if direct connection had been made to the battery. In the absence of a signal, the potential V_c of the collector is now given by

$$V_c = V - RI_c, \tag{12.2}$$

where V is the battery potential and RI_c is the potential drop in the resistor R, due to the collector current I_c which flows in it.

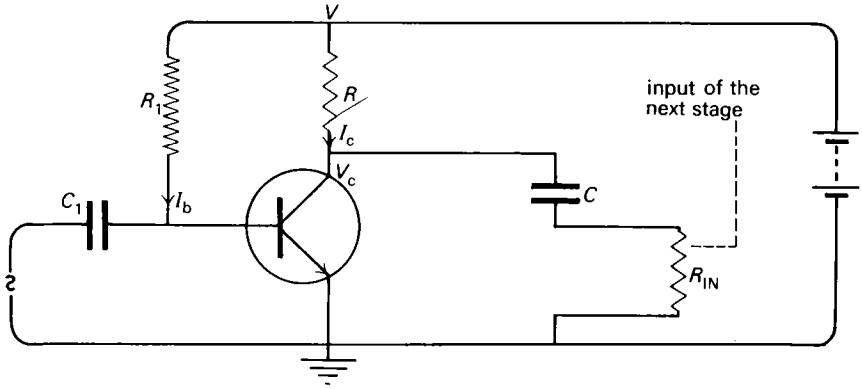


Fig. 12.19 A simple amplifying circuit, using a load resistor R and a coupling capacitor C

The coupling capacitor C

The resistor R_{IN} drawn in the circuit diagram represents the effective input resistance of the next amplifying stage. The requirement is to transfer the signal to the base of the next transistor without, at the same time, applying to it the bias potential which is intended for the collector of its predecessor, and which would be quite inappropriate to it. This discrimination is the function of the coupling capacitor C, which earns its title by its action in passing on the signal; it might, alternatively, have been termed a blocking capacitor, because it blocks the bias potential of the collector, preventing it from reaching the next transistor.

The action of a coupling capacitor has been discussed already (section 9.18). So far as the d.c. bias potential of the collector is concerned, the capacitor simply charges up from the battery, and has a potential drop equal to V_c across it. If the collector current varies, however, owing to a signal passing through the transistor, the capacitor presents only a low reactance towards this varying current, and so the variations in potential and current reach the base of the next transistor.

The overall variation in potential V_c at the collector, when a small alternating p.d. has been applied as a signal across the input terminals of the amplifying circuit, is shown in Fig. 12.20. It will be noticed that there is phase inversion, i.e. the output oscillation is 180° out of phase with the input.

Applying bias to the base

There remains the problem of applying the correct forward bias to the base, to draw the charge carriers across from the emitter, and the arrangement suggested

in the circuit diagram is the simplest method, although not the best, of achieving this without the use of a separate battery. This method is to connect the base to the supply battery through a high resistance R_1 of appropriate value. This acts as a

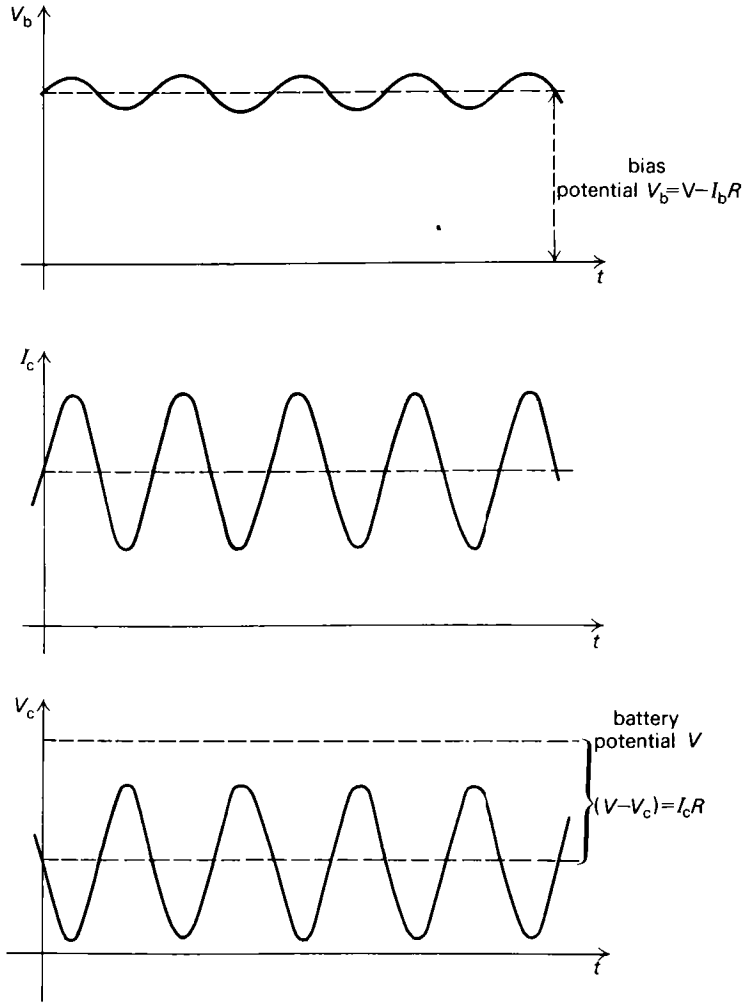


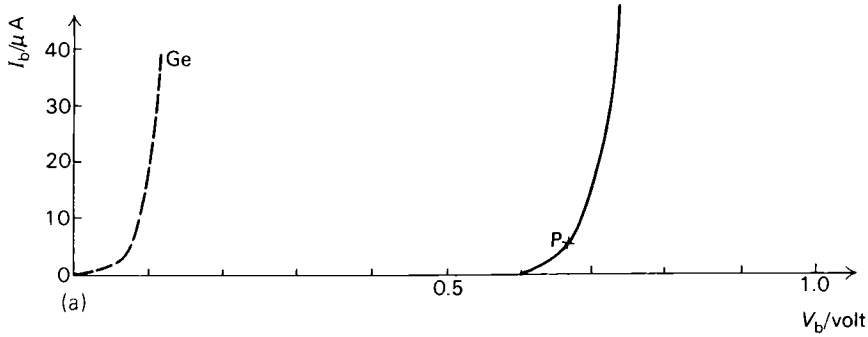
Fig. 12.20 Amplifying, using a load resistor. Waveforms corresponding to a small alternating input signal (a) The base potential (b) The collector current (c) The collector potential

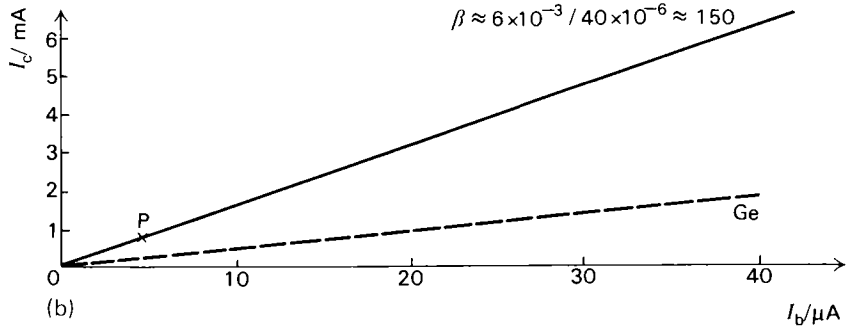
potential dropping resistance, because the base current I_b flows in it, and the d.c. bias potential V_b on the base is consequently given by the equation

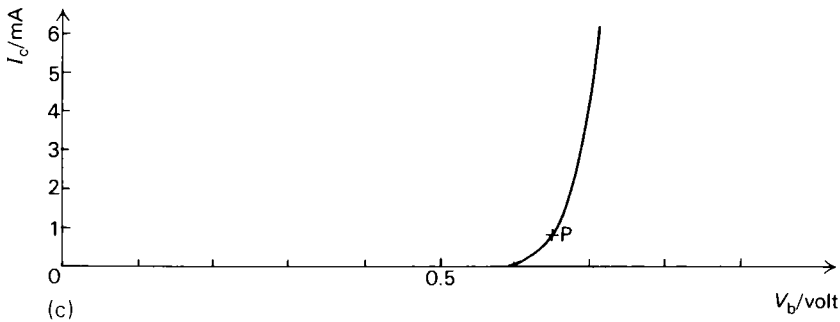
$$V_b = V - R_1 I_b. \tag{12.3}$$

12.11 Transistor characteristics

The circuit which we have been describing is termed a common emitter circuit because the emitter is connected to both the input and the output circuit; this is the most usual arrangement. Choosing the case of a typical small planar silicon *n–p–n* transistor, characteristic curves are drawn in Fig. 12.21, showing how the various currents and potentials depend on each other. These refer to the common







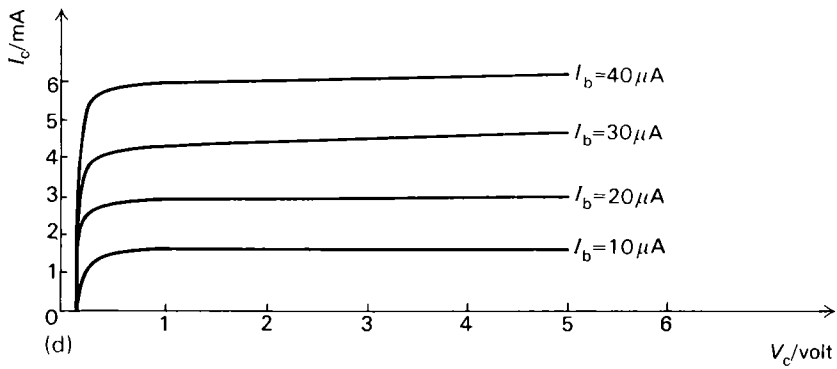


Fig. 12.21 Characteristic curves for a small planar *n-p-n* silicon transistor (a) The base characteristic, I_b against V_b (b) The transfer characteristic, I_c against I_b (c) A further mutual characteristic, I_c against V_b (d) The collector characteristic, I_c against V_c

emitter circuit, and the potential V_e of the emitter has been taken as zero. Similar curves are plotted in the diagram, for comparison, in dotted lines, which refer to a small alloy junction germanium transistor, but the description which follows primarily concerns the silicon. For all transistors, temperature is a factor which cannot safely be neglected owing to the possible effects of intrinsic conduction, and in no case must the temperature be allowed to rise so high that a very appreciable release of intrinsic charge carriers results. Silicon is, of course, less subject to such effects than germanium, and intrinsic conduction makes no appreciable contribution to the characteristic behaviour of a silicon transistor, provided that its temperature is kept down.

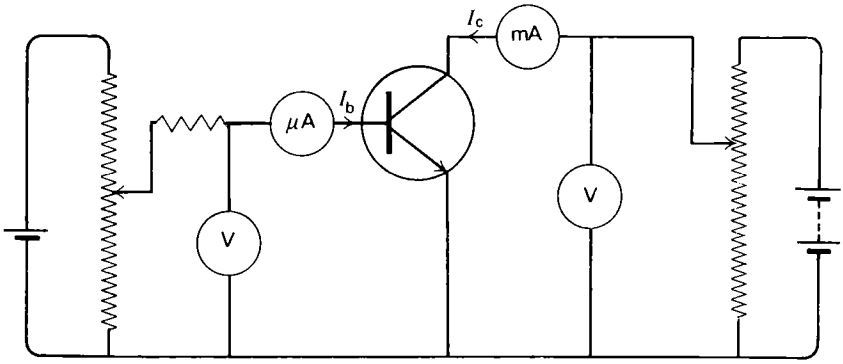


Fig. 12.22 A circuit suitable for obtaining data for the characteristic curves of figure 12.21

The transistor action is as we have already described it; a small variation in the base potential V_b causes a variation in the collector current I_c with a gain in energy; the important part of the current flows from emitter to collector, being helped on its way by the potential applied to the base. The flowing of base current makes no useful contribution to the transistor action, rather the reverse, and as the techniques of transistor action improve and the base becomes thinner and thinner, so the size of the base current is reduced. Nevertheless, in describing the transistor characteristics, it is necessary to start with the base characteristic, showing how base current I_b varies with base potential V_b. This has the same form as the corresponding curve for a forward biassed diode, the only difference being that the base current in the transistor is much smaller than that of a comparable diode, owing to the large proportion of the charge carriers crossing the junction from emitter to base which are drawn off to the collector. Nevertheless, the characteristic illustrates that the base potential is largely controlled by its current, and an input signal cannot cause this potential to vary very much. All silicon transistors operate with the base potential departing very little from the required forward bias value of 0.6–0.7 V; a stronger signal is, of course, a larger potential variation, but this seldom exceeds, say, ±0.1 V and it is the larger current which is more evident.

The transfer characteristic

This shows how collector current I_c varies with the base current I_b and is effectively a straight line, the slope of which is called the current amplification factor

$$\beta = \frac{I_c}{I_b}. \tag{12.4}$$

The value of β is, typically, 150 or more for the planar silicon transistor, and about 50 for an alloy junction germanium one. There is not a similar linear relationship

between the output current I_c and the applied potential V_b, so that, if the input signal is measured as a small change ΔV_b in potential, the output signal ΔI_c is not exactly proportional to it, and allowance sometimes has to be made for this in designing an amplifying circuit.

The collector characteristics illustrate the fact that the value of the collector current I_c depends almost entirely on what happens to the base and only to a small degree on the collector potential. This is consistent with the explanation of transistor action which has already been given, since this current consists of charge carriers attracted out of the emitter by the base potential. There is one condition, however, in which the above statement does not hold, and this is when the collector potential is reduced to about 0.3 V or less. In this case, there is no effective potential gradient in the base-collector depletion layer and transistor action fails to take place. Consequently, as the collector potential is further reduced, the collector current falls abruptly to zero. In this condition, the transistor is said to have bottomed, and this is sometimes deliberately arranged, when, for example, the transistor is used as a switch.

For normal amplifying action, the collector is put at a potential well outside the knee of the collector characteristic, where the base potential and current control its behaviour.

12.12 Calculations

We will now suppose that the transistor inserted in the circuit of Fig. 12.19 is the silicon transistor for which the characteristic curves of Fig. 12.21 have been drawn, and will proceed to calculate suitable values for the components in the circuit.

Assuming this to be an early stage of amplification, where the signal strength is small, there is little danger of causing distortion, and the transistor can be biassed so that its current is small, thus reducing the drain on the battery. A suitable working point might correspond to the values

$$I_c = 0.75 \text{ mA}.$$

$$\begin{aligned} I_b &= I_c/\beta \\ &= 0.75/150 \text{ mA} \\ &= 5\,\mu\text{A}. \end{aligned}$$

The collector potential could be about 4 V, which puts it well beyond the knee of the collector characteristic. We then have, from equation (12.2),

$$V_c = V - RI_c,$$

where V is the battery p.d., which we will take to be 9 V.

Therefore $$4 = 9 - R \times 0.75 \times 10^{-3},$$

i.e. $$R = \frac{5.10^3}{0.75}\,\Omega$$

$$= 7000\,\Omega,$$

which gives the value for the load resistor.

Similarly, for the simple method chosen for biassing the base, we would have, from equation (12.3),

$$V_b = V - R_1 I_b,$$

where the value of V_b is to be about 0.6 to 0.7 V.

Therefore
$$0.7 = 9.0 - R_1 \times 5.10^{-6},$$
$$R_1 = \frac{8.3}{5 \times 10^{-6}}\ \Omega$$
$$= 1.6\ \text{M}\Omega.$$

These calculations refer to the point marked P on each characteristic.

A second transistor, having the same pattern number as that referred to in the above calculations, would not have identical characteristics, because transistors cannot be manufactured with this degree of precision. One of the drawbacks of the simple base biassing circuit described above is that the value suitable for the resistor R_1 would have to be separately determined for each individual transistor.

The value of the coupling capacitor C

The ideal coupling capacitor would have infinite reactance towards the bias potentials, which is practicable, and zero reactance at the signal frequency, which is not. The best that can be done, in practice, is to make this reactance as low as possible by comparison with the effective input resistance R_{IN} of the next stage, which is in series with it.

The necessary condition, therefore, is
$$X_C \ll R_{IN},$$
i.e.
$$\frac{1}{2\pi f C} \ll R_{IN},$$
therefore
$$C \gg \frac{1}{2\pi f R_{IN}}.$$

This is in order that the greater part of the signal p.d. shall appear across the input of the next transistor, and not be lost, instead, as a p.d. across the capacitor. This requires a higher value capacitor at audio-frequency than it does at radio-frequency. Taking an audio-frequency $f = 50$ Hz, and assuming $R_{IN} = 8000\ \Omega$, we have
$$C \gg \frac{1}{100\pi \times 8000}\ \text{F},$$
i.e.
$$C \gg 0.4\ \mu\text{F}.$$

*12.13 A more satisfactory biassing arrangement

This biassing arrangement which has been described so far has the theoretical advantage of being the simplest in principle. A more satisfactory one in practice, however, which is more commonly adopted, is that shown in Fig. 12.23. Here the values of the components can be worked out without reference to the characteristics of one particular transistor, and protection is afforded against the effects of possible variations in the transistor's behaviour, which might result, for example, from temperature rise.

In this circuit, the base potential is applied from the potential divider, R_1, R_2, connected across the power supply, and the collector and emitter potentials are determined by the values of the load resistor R and an emitter resistor R_3 and by the currents flowing in them. The details are best explained by making a typical calculation.

We suppose that the transistor is to operate with a collector current I_C of about 1 mA, and a p.d. of about 4 V between collector and emitter. The values of the

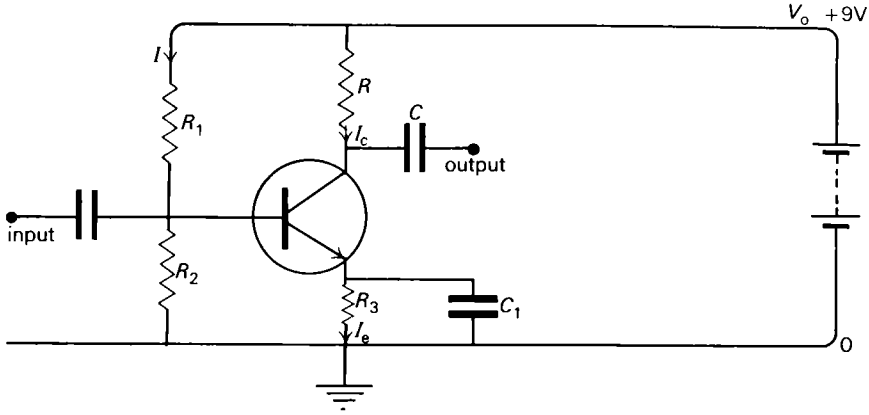


Fig. 12.23 A stabilized biasing circuit

four resistors are so chosen that the transistor potentials relative to earth are approximately:

$$V_c = 5 \text{ V},$$
$$V_e = 1 \text{ V},$$
$$V_b = 1.7 \text{ V}.$$

Thus, $V_{ce} = 4$ V, as intended,

and $V_{be} = 0.7$ V,

which is the correct forward bias for the base.

The calculations are as follows.

(a) The load resistor R

$$R = \frac{V_0 - V_c}{I_c},$$

where $V_0 = 9$ V.

Therefore $R = \dfrac{9-5}{10^{-3}}\,\Omega = 4000\,\Omega.$

(b) The potential divider R_1, R_2

The total resistance, $R_1 + R_2$, is so chosen that the current flowing in these resistors from the battery is about $I_c/10$. This is so much larger than the value of the base current, which also flows in R_1, that the contribution of the base current in fixing the value of the base potential is negligible. At the same time, the current is small enough not to represent an objectionable extra load on the battery, and the value of R_2, which connects the base to earth, is large enough for its effect in partially short circuiting the input signal to be acceptable.

We have

$$I = \frac{I_c}{10} = 100\,\mu\text{A},$$

$$R_1 + R_2 = \frac{9}{100 \times 10^{-6}}\,\Omega = 90000\,\Omega,$$

$$R_2 = \frac{V_b}{I} = \frac{1.7}{100 \times 10^{-6}}\,\Omega = 17000\,\Omega,$$

$$\therefore \quad R_1 = 73000\,\Omega.$$

(c) The emitter resistor R_3

$$R_3 = \frac{V_e}{I_e},$$

where $V_e = 1$ V

and $I_e = I_c + I_b = 10^{-3}$ A,

since I_b is small enough to be neglected.

Therefore $$R_3 = \frac{1}{10^{-3}}\,\Omega = 1000\,\Omega.$$

The function of resistor R_3 is to stabilize the behaviour of the transistor. Should there be any tendency for the collector current to increase, owing to a rise in temperature, for example, this would cause a rise in the potential of the emitter without affecting that of the base. Thus the forward bias between base and emitter would be reduced, preventing the current from increasing further.

(d) The decoupling capacitor C_1

Were it not for the presence of the capacitor C_1, the emitter resistor R_3 would be too effective in its action, and the amplifying function of the transistor would be largely suppressed. This is because there would be no discrimination between an unwanted variation in collector current and an essential one: the effect of the resistor R_3 would be to suppress them both by a process known as negative feedback; any change which an input signal might produce in the potential of the base would be largely nullified by a corresponding change in the potential of the emitter. The function of the capacitor C_1 is to provide what is effectively a short circuit from emitter to earth with respect to alternating current caused by the signal, while the steady component of the emitter current is still forced to flow in R_3: this is referred to as a decoupling action (section 9.18). The required value for the capacitance of C_1 is high, particularly if the signal is likely to include low frequencies, because the reactance must be small compared with the resistance of R_3, which is not itself very large.

Thus, $$\frac{1}{2\pi f C_1} \ll R_3,$$

where $R_3 = 1000\,\Omega$.

*12.14 A power transistor – temperature effects

As the signal grows in strength on passing through successive stages of amplification, larger collector and base currents are necessary because the stronger signal takes the form of larger swings in current. As the characteristics of a transistor indicate, the corresponding variations in potential are never very large – a respect in which the thermionic triode behaves differently. The transistor used in the output stage of an amplifier, therefore, is one designed to carry a comparatively heavy current, and its output resistance is lower than that of a less powerful type.

The heavy duty transistor is subject to the same danger as the heavy duty diode of thermal runaway, due to the effect of excessive temperature rise in releasing

intrinsic charge carriers in the wrong places. As compared with the diode, in fact, this danger is accentuated by amplification caused by a form of transistor action.

To guard against this thermal runaway, the collector of a power transistor is soldered to a metal base which can be bolted to a metal heat sink.

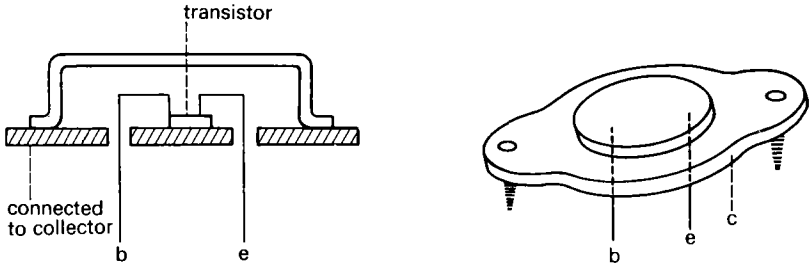


Fig. 12.24 A heavy duty silicon transistor. The metal base plate is bolted to a metal heat sink

12.15 The transistor in an oscillating circuit

It has already been explained that an amplifier can be used to maintain an oscillating signal simply by amplifying the signal and feeding back a part of the output to serve as an addition to the input. For this process to be successful, the signal which is fed back must be in phase with the input oscillation, or at any rate partially so, in order that the two oscillations shall combine with each other helpfully, and this is then termed *positive feedback.*

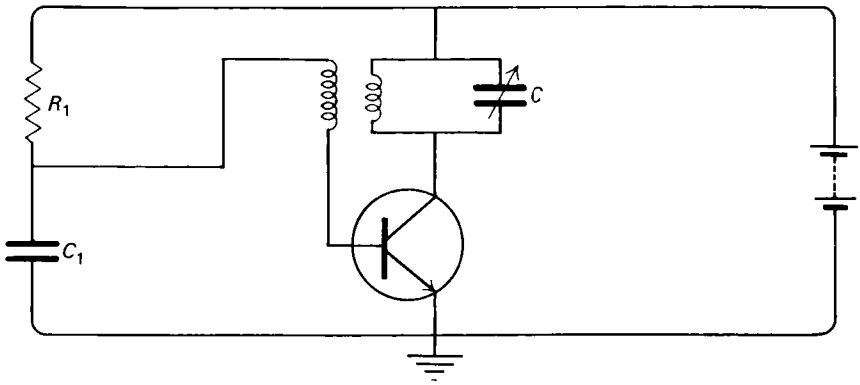


Fig. 12.25 An oscillating circuit

If the oscillation is to have sinusoidal waveform and adjustable frequency, the simplest circuit to describe, although not the most popular in practice, is the oscillating system of inductance and capacitance, and an example of this is given in Fig. 12.25. In this particular circuit, the bias for the base of the transistor is applied through the resistor R_1, and C_1 is a decoupling capacitor, which provides a low reactance connection to earth at the frequency of the signal without short circuiting the d.c. bias potential.

12.16 The thermionic triode

The thermionic triode has been displaced by the transistor for many purposes, and the advantages gained in choosing the transistor are not difficult to trace; an invitation to list them is included among the questions at the end of this chapter. The thermionic valve has its advantages too, however. One of them is its immunity from temperature effects, and another its ability to work at very high potentials. Consequently, there is small danger yet that it will become a museum piece.

The particular triode shown in Fig. 12.26 has plane electrodes, but in another type these would be coaxial cylinders, with the anode on the outside and the cathode along the central axis; in this case, the grid would be a wire spiral.

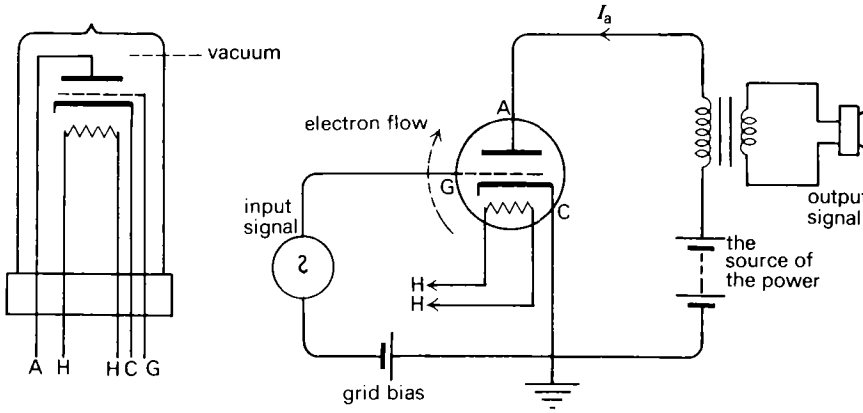


Fig. 12.26 A thermionic triode (a) The triode (b) A simple amplifying circuit

As with the vacuum diode, thermionic emission from the hot cathode serves to maintain a population of electrons in the vacuum of the valve, and most of these are concentrated near the cathode surface, forming a negative space charge. The positive potential applied to the anode, usually in the region of 100 V or more, attracts electrons from the space charge, which is then replenished by further thermionic emission from the cathode. The grid, being in the form of a wire mesh or spiral, allows the electrons to pass through it, but it is placed very close to the cathode, which makes the electric field near the cathode largely dependent on the grid potential. The result of this geometry is that the strength of the electron current which reaches the anode is largely controlled by the p.d. applied between grid and cathode, whereas the power in the current derives from the source of the anode potential. The amplifying possibility which this provides has already been discussed (section 12.5) and this is illustrated in Figs. 12.26(*b*) and 12.12.

12.17 The control grid

There is a contrast between the process by which the transistor and the thermionic triode provide the charge carriers by which they function. Whereas the emitter of the *n–p–n* transistor emits electrons only because the emitter-base junction is forward biassed, the cathode of the triode does so merely because it is hot. It is unnecessary, therefore, to apply forward bias to the grid; it would, indeed, be a disadvantage to do so, for the grid current which would flow in consequence would be large enough almost to short circuit the input signal. Consequently the triode

is operated with negative bias applied to the grid, insufficient to prevent the electrons from reaching the anode, but sufficient to prevent them reaching the grid. Thus, no grid current flows, and the valve is controlled entirely by variations in potential.

The action of the negative grid in controlling the anode current without preventing it from flowing is similar to that by which an electron lens shapes the equipotential surfaces in its neighbourhood, the result being that, whereas the potential of the actual grid wires is negative, that of the spaces between them is not, and the electrons pass through these spaces on their way to the anode. This is illustrated in Fig. 12.27.

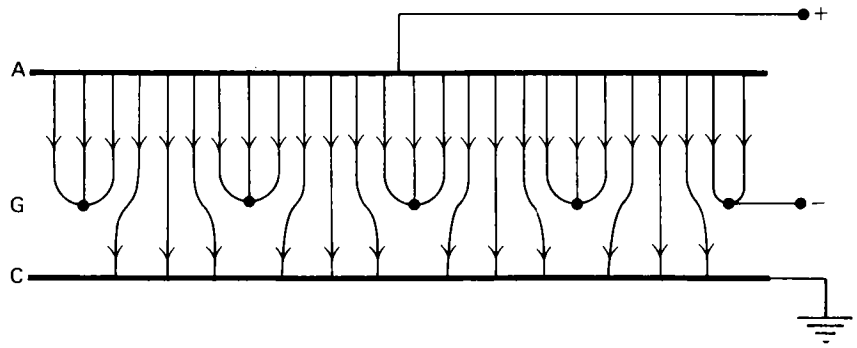


Fig. 12.27 Diagram representing the lines of force in a triode, when the anode current is not cut off. The electrons are attracted in the opposite direction to the arrows

12.18 Mutual characteristics

The action of a triode is usefully illustrated by drawing a mutual characteristic curve, showing how the current I_a which reaches the anode varies with the potential difference V_g between the grid and the cathode. Since the value of the anode current depends also on the potential of the anode, although to a lesser degree, a set of characteristic curves can be drawn, corresponding to a sequence of selected values of this anode potential V_a. An example of these, for a particular triode, is shown in Fig. 12.28.

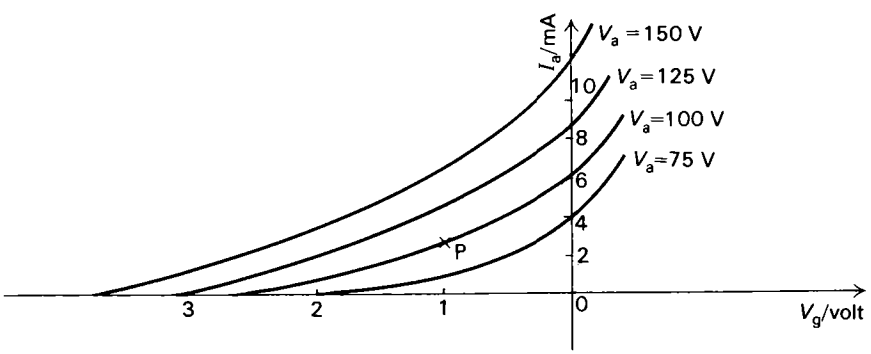


Fig. 12.28 Mutual characteristics for a thermionic triode

If the triode referred to in the diagram were operated as an amplifier with its anode at a potential of 100 V, a suitable value for the grid bias would be −1 V, corresponding to the point P marked on the curve. An alternating input signal superimposed on this bias would then cause oscillations in the grid potential without either driving the grid positive or driving it so far negative that the valve current is completely cut off; either of these eventualities would result in a distorted output from the anode, but the input signal could vary by more than ±0.5 V without having this effect.

*12.19 A typical amplifying circuit

A conventional amplifying circuit is shown in Fig. 12.29, in which the output signal is obtained in the form of a varying p.d. set up across an anode load resistor

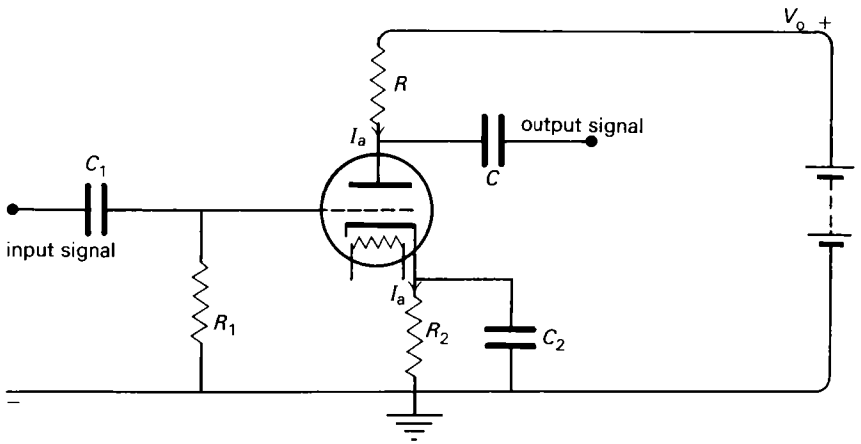


Fig. 12.29 An amplifying circuit, using a thermionic triode

R, and capacitors C_1, C are coupling capacitors. The negative grid bias is obtained by inserting a bias resistor R_2 in series with the cathode, while connecting the grid to earth through a high resistance grid leak R_1. In this way, the valve current flowing in R_2 places the cathode at a positive potential, while the potential of the grid is zero. Capacitor C_2 has a high capacitance, since its function is a decoupling one, providing a low reactance shunt to the bias resistor at the frequency of the signal, as described in section 12.13.

Assuming that the H.T. potential V_0 is 160 V, and that the triode is operated in conditions represented by the point P in Fig. 12.28, we have

$$I_a = 3 \text{ mA},$$

$$V_a = V_0 - RI_a,$$

where $$V_a = 100 \text{ V}.$$

Therefore $$R = \frac{160 - 100}{3 \times 10^{-3}} \Omega$$

$$= 20000 \Omega. \qquad \text{(a)}$$

Also, $$V_{gc} = -1 \text{ V},$$

where $V_g = 0$

and $V_c = R_2 I_a$.

Therefore
$$R_2 = \frac{1}{3 \times 10^{-3}}\,\Omega$$
$$= 330\,\Omega \qquad \text{(b)}$$

and
$$\frac{1}{2\pi f C_2} \ll 330,$$

where f is the frequency of the input signal.
Assuming that this will go as low as 50 Hz,

$$C_2 \gg \frac{1}{330 \times 2\pi \times 50}\ \text{F},$$

i.e.
$$C_2 \gg 10\mu\text{F}. \qquad \text{(c)}$$

12.20 Transit time

In the limit, the highest frequency at which a triode or transistor can operate successfully, whether as an amplifier or as an oscillator, is dependent on the effect known as transit time. The electrons must have sufficient time to permit them to cross from cathode to anode, or from emitter to collector, before the oscillations have time to reverse the polarity of the signal. This time is not available when the frequency range has reached the high value of the microwave band, 1000–10000 MHz, and valves in these frequencies work on different principles.

The thermionic triode has an advantage over the transistor with respect to transit time, because the very slow passage of the electrons or holes in the solid state crystal, as compared with the very rapid motion of electrons in the vacuum of the triode, outweighs the fact that in the transistor the distance to be travelled is much shorter.

12.21 A d.c. amplifier

None of the circuits described in this chapter would amplify a direct current or p.d., because of the blocking effects of the capacitors which have been included in the process of arranging for bias potentials. It is possible, particularly when using transistors, to design circuits which avoid the need for capacitors, and this is done when a d.c. amplifier is required. Such circuits are not, however, quite basic enough in their principle to describe here.

QUESTIONS

12.1 What is meant by (*a*) intrinsic, and (*b*) extrinsic conductivity?

Explain how and why the electrical conductivity of a crystal slice of germanium varies with: (*a*) temperature over the range 0–250 °C, (*b*) trivalent impurity atoms present, (*c*) pentavalent impurity atoms present.

Explain one method of forming a *p–n* junction in germanium. Why does this junction exhibit a rectifying action? (A.E.B.)

12.2 Explain what is meant by *p*-type and *n*-type semiconductors. Describe a *p–n* junction diode. Draw a graph which shows the variation of the current

through such a diode with the p.d. across it, and explain why the diode behaves differently when the p.d. across it is reversed. (O & C)

12.3 Give a brief description of the construction of a high vacuum diode. Draw a graph which shows the variation of the current through such a diode with the p.d. across it, and account for the main features of the curve.

Describe how the introduction of a third electrode, the grid, makes possible (*a*) control of the current which reaches the anode, and (*b*) the amplification of a voltage. (O & C)

12.4 (*a*) Whether a semiconductor is *n*-type or *p*-type, the particles which actually move through the substance are electrons. Explain the difference between an electron movement which is referred to as the movement of a negative charge-carrying electron, and an electron movement which is described in terms of a positive charge-carrying hole.

(*b*) If a p.d. is applied across a *p*-type semiconductor, the conduction is described as being due to a drift of positively charged holes down the potential gradient. Explain why this process does not result in the material becoming bereft of its holes, and so being changed into an insulator.

(*c*) An *n*-type semiconductor is so described because it contains a larger concentration of charge-carrying electrons than of positive holes. Is the material negatively charged in consequence? If not, why not?

12.5 Write explanatory notes on the following observations:

(*a*) Whereas a transistor radio operates immediately the switch is made, a television receiver produces a picture only after quite a considerable delay.

(*b*) A solid state diode, or a transistor, when designed to draw an appreciable current, is provided with a heat sink.

(*c*) A germanium diode is preferred to a silicon one for rectifying small voltages.

12.6 Describe the junction transistor. Sketch curves to show the variation of the collector current with the collector-emitter voltage for various values of the emitter-base current, and explain their form. (O & C)

12.7 Explain, with appropriate circuit diagrams, how you would investigate the static characteristics of an *n–p–n* transistor in the common emitter connection.

Sketch the curves which are obtained, and comment on their special features. (A.E.B.)

12.8 Discuss the advantages and disadvantages of semiconductor diodes and junction transistors over their respective vacuum tube equivalents. (A.E.B.)

12.9 Why is it essential that the base region of a transistor should be exceptionally thin?

The useful charge carriers in an *n–p–n* transistor are electrons from the *n*-type emitter. Discuss this, and explain why the base and collector are doped at all.

Although the useful part of the emitter current is due to electrons which

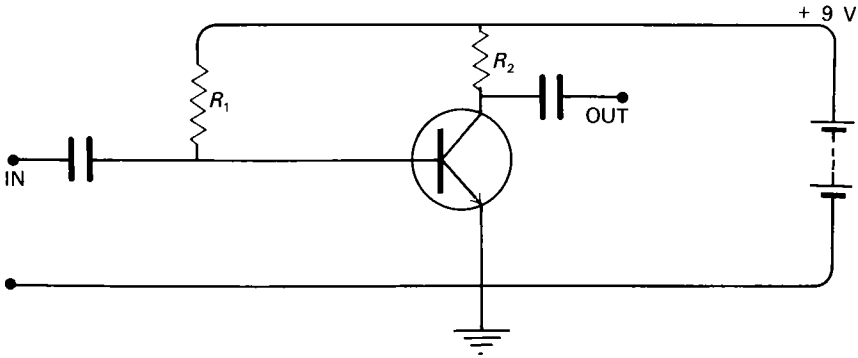


Fig. e 69

cross from emitter to base, some of the current, nevertheless, is due to holes which cross from base to emitter. Explain the difference between the two.

12.10 Fig. e 69 represents a simple biassing arrangement for an *n–p–n* transistor, to be used as a common emitter amplifying circuit.

Selecting 4 V as a suitable value for the collector potential, and assuming that you decide to work with a collector current of 2 mA, calculate suitable values for the base biassing resistor R_1 and the load resistor R_2. Take the current amplification factor of the transistor to be $\beta = 250$.

12.11 Using the circuit shown in question 12.10, a sinusoidal input signal is applied to the base which overloads the amplifier, with the result that the output waveform is distorted, both top and bottom peaks of the wave being cut off. Explain.

12.12 The simple biassing circuit shown in question 12.10 is seldom used in practice, because there is no stabilizing effect. If a rise in temperature, for example, were to alter the characteristics of the transistor, it might draw an excessive current.

A rather better arrangement is to connect the base resistor R_1, not directly to the battery, but to the collector, as shown in the diagram below. In this way, some negative feedback is provided.

By considering the effect on the base potential if the collector current were to increase by 1 mA, explain what is meant by negative feedback, and how it operates in this case.

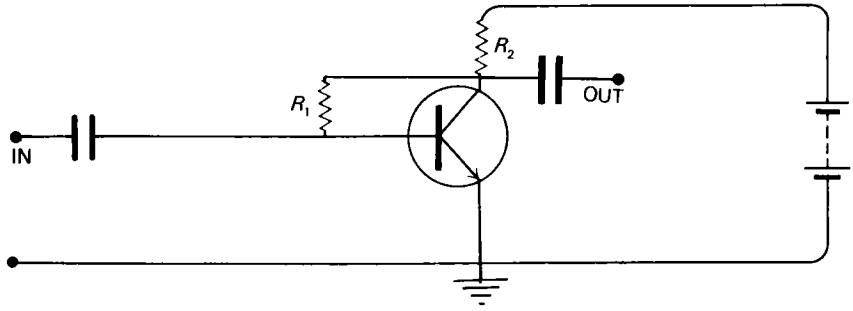


Fig. e 70

*12.13 In this circuit (e 71) a device with three terminals is connected to a 15 V battery through resistors. The device has the following properties:

(i) The current I_z leaving terminals Z is determined only by the voltage V_{xz} between terminals X and Z. The relation between I_z and V_{xz} is

$$I_z = 10^{-15}\exp(40\ V_{xz}),$$

where I_z is in amperes and V_{xz} in volts.

(ii) The current I_x entering at terminal X is a constant fraction $1/\beta$ of I_z.

If $\beta = 50$, find the value of R required to make $I_z = 10^{-3}$ A. If this value of R is used, what value would I_z have if β were 100?

An answer correct to 2 significant figures is sufficiently accurate. (CAM. SCH. ENG.)

*12.14 The device referred to in the last question is presumably a transistor. State which of the three terminals X, Y, Z is the base, which is the collector, and which is the emitter. Does the question refer to an *n–p–n* transistor, or a *p–n–p*? Is it silicon or germanium? Draw out the circuit again in the form of a conventional transistor circuit diagram.

*12.15 Draw a circuit diagram for a triode valve connected as a simple amplifier,

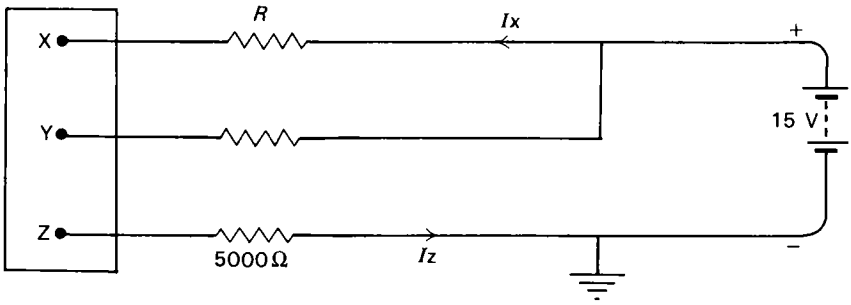


Fig. e 71

and explain how its ability to amplify an alternating voltage arises, and indicate how its gain may be calculated. What is the phase relationship between the output and input signals? Discuss qualitatively how you would expect the gain to vary with the frequency of the signal when a second valve is capacitatively coupled to the first.

Discuss the advantages of using a transistor for the same purpose, and indicate any differences between the circuit which would be used in this case and that for the triode. (OXF. SCH.)

Chapter 13

Electric field equations

13.1 The inverse square law

The descriptions of the electric and the magnetic field which were given in chapters 6 and 7 were largely confined to certain practical aspects, with the object of accounting for the parts played in a circuit by capacitance and inductance. A full treatment, and in particular the more mathematical aspect, was postponed, and the purpose of this chapter is to make up the omission with respect to the electric field, the subject of magnetic flux then being taken up in the first few chapters which follow.

The starting point for deriving electric field equations is Coulomb's inverse square law of force between point charges. If two charges Q, Q' are concentrated at points, distance r apart, in vacuo, the force of repulsion between them is given by

$$F = \alpha \frac{QQ'}{r^2}, \tag{13.1}$$

where α is a constant.

In practice, if the charges are placed in air and not in a vacuum, no appreciable change in their effect on each other can be detected.

13.2 Rationalized equations

The equation for capacitance of a vacuum (or air) spaced parallel plate capacitor was stated, without proof, in chapter 6 (equation 6.11) to be

$$C = \epsilon_0 \frac{A}{d}, \tag{13.2}$$

where A is the area of each plate, d the separation of the plates, and ϵ_0, called the permittivity of a vacuum, is a constant.

There is, in fact, a very simple relationship between the two constants α (equation 13.1) and ϵ_0 (equation 13.2), which will presently be derived, and this is

$$\left.\begin{aligned} \epsilon_0 &= \frac{1}{4\pi\alpha}, \\ \alpha &= \frac{1}{4\pi\epsilon_0}. \end{aligned}\right\} \tag{13.3}$$

Quite a large proportion of the equations in electrostatics share with equation 13.2 the feature that the constant α is included as part of the factor $1/4\pi\alpha$, and it is convenient, for this reason, to dispense with the use of α altogether, substituting the constant ϵ_0 in its place in all equations. In making this alteration, the form of equation 13.1 is altered, to become

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{QQ'}{r^2}. \tag{13.4}$$

It will be noted that this choice of ϵ_0, in preference to α, has removed the factor π from equation 13.2, which refers to plane surfaces having a uniform electric field, and has included it, instead, in equation 13.4, which refers to point charges having a radial field. This is typical, also, of other equations. It is, perhaps, more logical to associate π with radial fields and spherical surfaces, and for this reason the system of equations written to include ϵ_0, and not α, is said to be rationalized, e.g. equation 13.4 is the rationalized form of equation 13.1.

When electrical equations are presented for use with SI units, the rationalized form is universally adopted, and so this will be done here, making equation 13.4 the starting point, in preference to equation 13.1.

It is unfortunate that the adoption of c.g.s. units, which has the more venerable history, traditionally involves, in effect, the choice of α instead, i.e. the use of unrationalized equations. This is one of several factors which make it difficult for the student to consult both SI and c.g.s. texts.

13.3 The value of ϵ_0

Adopting SI units, the units of measurement for all the quantities F, Q, Q', r in equation 13.4 have been fixed independently of this equation. The unit of charge Q, for example, is the coulomb, which is the ampere second, defined by

$$Q = It. \tag{13.5}$$

This means that the SI value for the constant ϵ_0 cannot be fixed by suitable definition, or deduced from theoretical reasoning; it can be determined only by practical measurement. (This point has already been made with reference to the alternative constant α, in section 1.13.) This essential experiment consists of setting up an apparatus to suit the form of an equation such as equation 13.4 or 13.2, and measuring all the other factors involved. In practice it is the more accurate to choose a capacitance relation, such as equation 13.2, and a suitable experiment will be described later in this chapter. As stated already, the measured value is

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ F/m},$$

to an accuracy better than 1 per cent.

When writing out rationalized equations, it will be found useful to retain the factor $1/4\pi\epsilon_0$ in one piece, wherever it appears, as has been done here in equation 13.4. The reason for this is the chance fact that the value of $1/4\pi\epsilon_0$ (which is α) works out, with acceptable accuracy, to be a very convenient number, viz.

$$\alpha = \frac{1}{4\pi\epsilon_0} = 9.0 \times 10^9 \text{ in SI units.}$$

Hence, an alternative form in which to quote the value of ϵ_0 is

$$\epsilon_0 = \frac{10^{-9}}{36\pi} \text{ F/m.}$$

13.4 Field strength and potential

We will now proceed to derive expressions for the field strength E and the potential V at points in the vicinity of various charged conductors, and to do this we start

from the definitions of field strength and potential which were given in chapter 6, and from the inverse square law, equation 13.4, which rests on experimental evidence.

Summarizing these, we have:

The field strength E at a point is a vector, defined by the equation

$$E = \frac{F}{Q}, \tag{13.6}$$

where F is the force which acts on charge Q when it is placed in the field.

[More accurately, we should write

$$E = \frac{\delta F}{\delta Q}, \tag{13.6a}$$

for the reason that the placing of a charge of finite size at the point might disturb the original charge distribution which causes the field.]

The potential V at a point is a scalar, defined by the equation

$$V = \frac{W}{Q}, \tag{13.7}$$

where W is the work required to bring a charge Q up to the point from some other point which is arbitrarily taken to have zero potential. This arbitrary zero is usually taken to be earth potential.

[More accurately, we should write

$$V = \frac{\delta W}{\delta Q} \tag{13.7a}$$

for the same reason as that given in the case of field strength.]

Potential gradient. The above are two equations of definition. In addition, it follows from them (section 6.11) that the field strength E at a point is equal to the negative potential gradient $-\mathrm{d}V/\mathrm{d}x$ at that point, i.e.

$$E_x = \frac{-\mathrm{d}V}{\mathrm{d}x}, \tag{13.8}$$

where E_x is the component of E in the direction of x.

If the field strength is uniform over a finite distance x, and the potential at each of the two points distance x apart is respectively V_1 and V_2, equation 13.8 may, alternatively, be written

$$E_x = \frac{V_1 - V_2}{x}. \tag{13.8a}$$

The force between point charges. Finally, we have the experimentally established law of force between two point charges

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{QQ'}{r^2}. \tag{13.4}$$

The experimental evidence for this will be discussed at the end of this chapter.

The units. Using SI units, we have:

potential V measured in volts, which are joules per coulomb;
field strength E measured in newtons per coulomb, alternatively called volts per metre (reference equation 13.8).

13.5 The field strength and potential at a point near an isolated point charge Q

The definition of E given in equation 13.6 breaks the inverse square law, equation 13.4, into two separate parts, which may be regarded as representing, respectively, cause and effect.

A charge Q' placed at a point which is distance r from a second charge Q experiences a force F which is given by equation 13.4. Alternatively, we can say that the charge Q' experiences the force because it has been placed in a field of strength E, this field being caused by the influence of the charge Q,

i.e. $$F = Q'E, \tag{13.6}$$

where $$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2}. \tag{13.9}$$

This second equation is the required equation giving the field strength E at distance r from a point charge Q.

The potential V at a point distance r from Q is found by deducing the amount of work which must be done if a small charge Q' is to be brought up to the point from earth. We must suppose the earth to be a long way off; otherwise, the charge Q will not be isolated, as it will induce negative charge on any earthed conductors in the vicinity.

Suppose the charge Q' is moved to the point A along the straight path PA (Fig. 13.1). When the charge has reached the point P, distance r from Q, the force F in the direction AP is given by

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q \cdot Q'}{r^2}, \text{ from equation 13.4.}$$

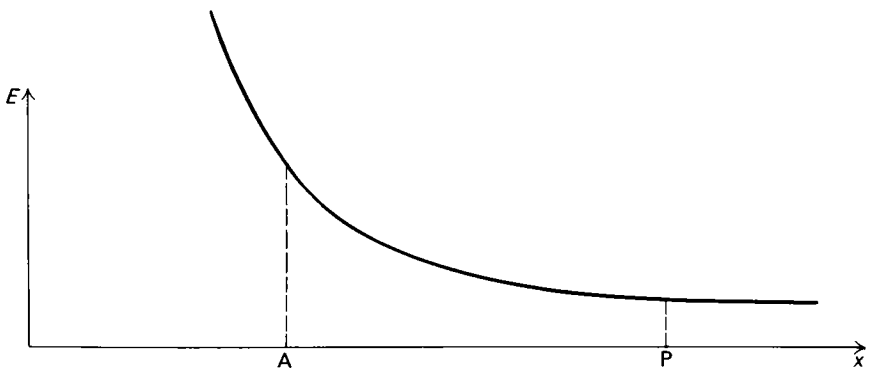


Fig. 13.1

To move the charge Q' a further small distance, $-\delta r$, towards A will require an amount of work δW to be done which is given by

$$\delta W = F(-\delta r)$$
$$= \frac{1}{4\pi\epsilon_0} \cdot \frac{-QQ'}{r^2} \delta r.$$

Hence, to move the charge all the way from the distant earth point up to the point A, the total work required is

$$W = \int_{\infty}^{r} \mathrm{d}W$$

$$= \frac{1}{4\pi\epsilon_0} \int_{\infty}^{r} \frac{-QQ'}{r^2} \cdot \mathrm{d}r$$

$$= \frac{1}{4\pi\epsilon_0} \cdot \frac{QQ'}{r}$$

Hence the potential V at A is

$$V = \frac{W}{Q'}$$

$$= \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r}. \tag{13.10}$$

Alternatively, this result comes from equation 13.8, viz.:

$$E_x = \frac{-\mathrm{d}V}{\mathrm{d}x}$$

$$V = \int_{\infty}^{r} -E_x \mathrm{d}x,$$

where

$$E_x = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{x^2}.$$

$$V = \frac{1}{4\pi\epsilon_0} \int_{\infty}^{r} \frac{-Q}{x^2} \mathrm{d}x$$

$$= \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r}.$$

The graph in Fig. 13.1 illustrates why calculus is necessary; the charge Q' cannot be moved through a finite distance towards Q without the field strength, and hence the force which acts on the charge, changing appreciably.

Summary. For an isolated point charge Q

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2}, \tag{13.9}$$

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r}. \tag{13.10}$$

13.6 Gauss' theorem

The inverse square law, equation 13.4, refers specifically to charge concentrated at a point, and, in the more practical case of charge distributed over a surface, to deduce the value of the field strength or of the potential by applying this equation would involve calculus, often in very difficult or even impossible form. It is convenient, instead, to simplify the problem by making a quantitative application of the concept of lines of force, a procedure originally adopted by Faraday.

Lines of force were defined quantitatively in section 6.2 by the equation

$$E = \frac{\delta N}{\delta A}, \tag{13.11}$$

i.e. a uniform field E is represented by having a number N of lines crossing area A, measured in a plane perpendicular to the field, such that

$$E = \frac{N}{A},$$

i.e. the number crossing unit area is made equal to E. The equation is put in the form of equation 13.11 to cover the more general case, of a point in a field which is not uniform.

This equation becomes consistent with the inverse square law equation 13.9, if the number of lines of force which radiate outwards from a point charge Q is taken to be equal to

$$N = \frac{1}{\epsilon_0} Q. \qquad (13.12),$$

To prove this, we consider a spherical surface, as in Fig. 13.2, having radius r with the charge Q situated at its centre.

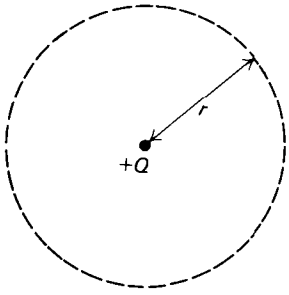


Fig. 13.2 The field strength E at all points on the dotted spherical surface is $E = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2}$

By symmetry, the field strength E is the same at all points on this surface, being given by

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2}. \qquad (13.9)$$

Also, the total area of the surface is A, where

$$A = 4\pi r^2.$$

Whence, the total number N of lines of force which must be assumed to radiate from the charge Q must be such that

$$E = \frac{N}{A},$$

i.e.

$$N = EA$$

$$= \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2} \cdot 4\pi r^2$$

$$= \frac{1}{\epsilon_0} Q. \qquad (13.12)$$

Noting that this expression does not include the distance r, we see that, as a consequence of the inverse square law, the lines may be drawn as continuous lines, radiating outwards from the charge $+Q$.

The importance of this quantitative aspect of lines of force is that it applies even when the charge Q is not concentrated at a point, but is distributed over a surface. This was proved, by Gauss, to be a consequence of the inverse square law.

Strictly, Gauss' theorem is a theorem stated in mathematical terms, and it will be given in this form later in this chapter. All that is required here, however, is a modified form of it, which may be stated as follows.

A modified form of Gauss' theorem

If we imagine that a number of lines of force equal to Q/ϵ_0 radiate from charge $+Q$, and that these lines are continuous, finally terminating on charge $-Q$, then the field strength E at any point due to the charge Q is correctly measured as being equal to the ratio $E = \delta N/\delta A$, where δN is the number of lines which cross a small area δA, perpendicular to the field, and containing the point. This is a result of the inverse square law of force between point charges, and is true, no matter how the charge Q may be distributed.

The effect of a dielectric

The effect of placing a dielectricin the field was discussed in section 6.21, and illustrated in Fig. 6.32. The number of lines of force suffers an abrupt change at the surface of the dielectric because some of them end at the charge which resides on the surface of the polarized dielectric. This makes the problem of a dielectric more difficult to deal with, and we will not, initially, consider such cases.

We can now proceed, by using Gauss' theorem, to derive equations giving the field strength and potential due to simple cases of distributed electric charge, but it will be assumed that no dielectric is present.

13.7 The isolated conducting sphere

The derivations which follow require the sphere to be isolated, in order that the charge distribution may be taken as symmetrical. The effect which an earthed conducting sheet in the vicinity would have is shown, as an example, in Fig. 13.3.

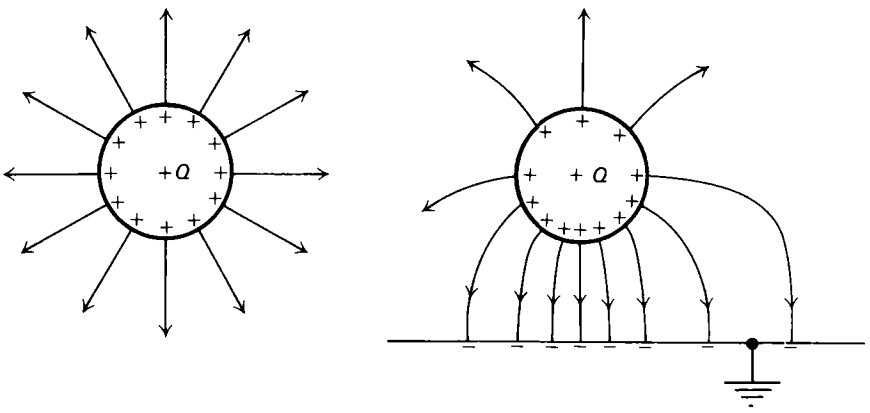


Fig. 13.3 (a) The field of an isolated conducting sphere (b) The effect of an earthed conducting sheet

***(a)* The field strength E**

(i) At a point which lies outside the sphere

The total number of lines of force leaving the sphere is N, where

$$N = \frac{1}{\epsilon_0}Q. \tag{13.12}$$

We can take an imaginary spherical surface, radius r, concentric with the charged sphere, and assume, from the symmetry, that these lines are uniformly spaced out when they cross it, i.e. the field strength E is the same at all points on this surface

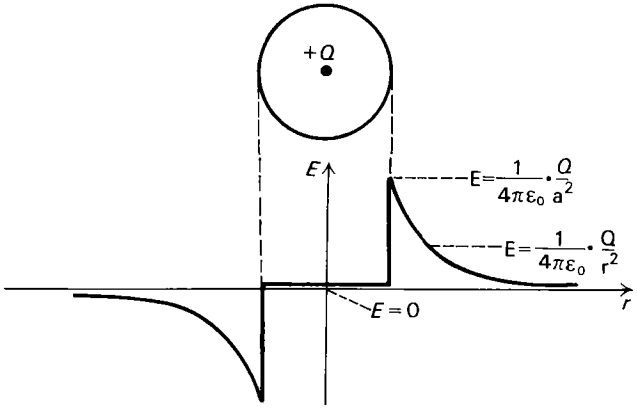


Fig. 13.4 The field of an isolated charged sphere. The field strength is here taken to be positive if it is directed towards the right

and is perpendicular to the surface. The total area which the lines have to cross at this distance r is

$$A = 4\pi r^2.$$

Hence, the field strength at any point P, distance r from the centre of the charged sphere, is

$$E = \frac{N}{A} \tag{13.11}$$

$$= \frac{1}{\epsilon_0} Q \div 4\pi r^2,$$

i.e.

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2}. \tag{13.13}$$

(ii) At a point on the surface of the sphere, radius a

Equation 13.13 still applies and we can write, for this special case,

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{a^2}. \tag{13.14}$$

(iii) At a point which lies inside the sphere

There are no lines of force inside a charged conductor,

i.e.

$$E = 0. \tag{13.15}$$

***(b)* The potential**

(i) At a point outside the surface of the sphere

$$V = \int_{\infty}^{r} -E\mathrm{d}x$$

$$= \frac{1}{4\pi\epsilon_0} \int_{\infty}^{r} -\frac{Q}{x^2} dx$$

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r}. \tag{13.16}$$

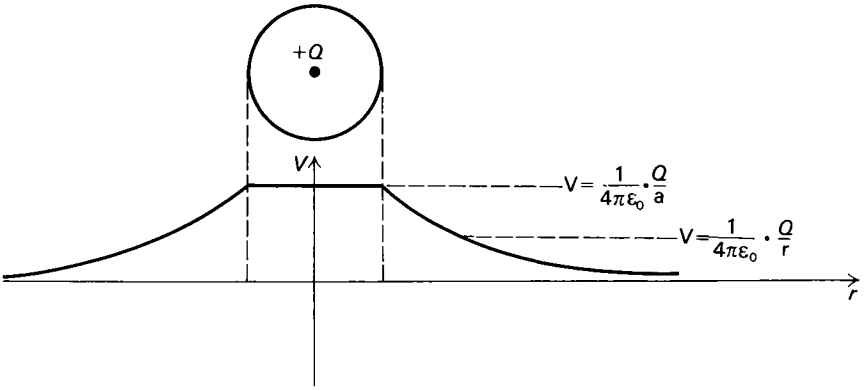


Fig. 13.5 The potential due to an isolated charged sphere

(ii) At a point on the surface of the sphere

radius a

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{a}. \tag{13.17}$$

(iii) At any point inside the sphere

The potential gradient, i.e. the field strength, inside the sphere is zero, and so the potential must be the same as the potential at its surface,

i.e.

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{a}. \tag{13.17}$$

(c) **The capacitance of the sphere**

From equation 6.10, we have

$$C = \frac{Q}{V},$$

where V is the potential of the sphere.

But

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{a},$$

whence

$$C = 4\pi\epsilon_0 a. \tag{13.18}$$

For example, the capacitance of an isolated sphere, radius 1 cm, is

$$C = \frac{1}{9 \times 10^9} \times 10^{-2}\ \text{F}$$

$$= 1.11 \times 10^{-12}\ \text{F}$$

$$= 1.11\ \text{pF}.$$

The above equations, and especially the diagrams which illustrate them, indicate the reason that high potential apparatus is terminated in rounded knobs; the smaller the radius of a charged conductor, the larger is the field strength at its surface, and discharge is the more likely to occur.

13.8 The capacitance of concentric spheres

A capacitor made from concentric spheres, having the outer sphere earthed, would make a very reliable form of capacitor, being completely shielded from outside influences. The only disadvantage, but an important one in practice, is its inconvenient shape.

The effect of the earthed shell is a good example of how the proximity of an earthed conductor, having charge of opposite sign induced on it, lowers the potential of a charged conductor, and so increases the capacitance. This is illustrated in Fig. 13.6.

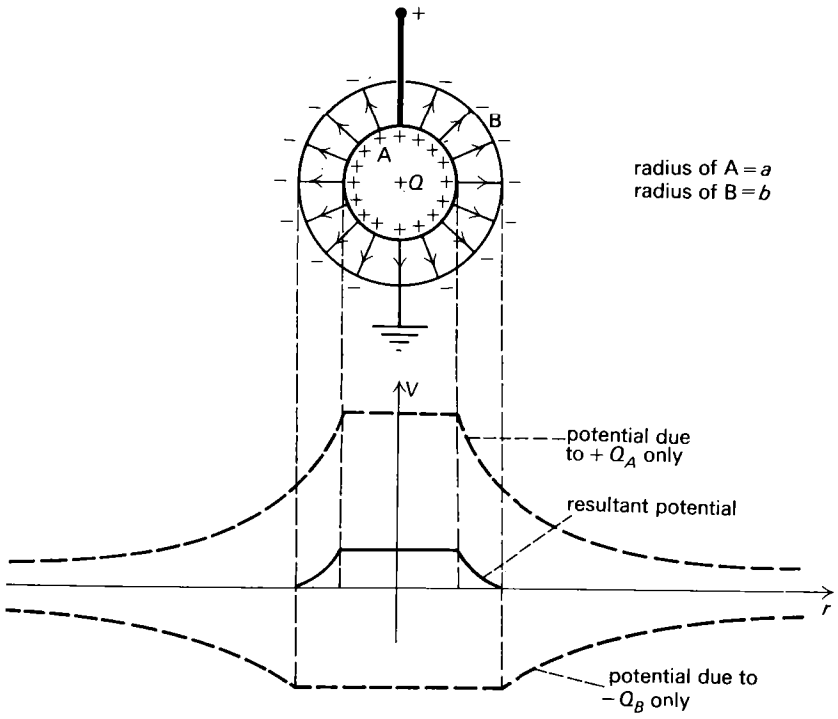


Fig. 13.6 Concentric spheres, with the outer sphere earthed The potentials due to the charge $+Q$ on the inner sphere A, and due to the charge $-Q$ on the outer sphere B, are separately plotted in the diagram. The resultant potential V at any point is the difference between the two (i.e. the algebraic sum) because that due to the charge on B is negative

If the induced charge on the outer sphere B is Q', then the potential of B due to its own charge is equal to

$$\frac{1}{4\pi\epsilon_0} \cdot \frac{Q'}{b},$$

whereas its potential due to the charge Q on the inner sphere A

is
$$\frac{1}{4\pi\epsilon_0}\cdot\frac{Q}{b},$$
whence, the resultant potential on the outer sphere is
$$V_B = \frac{1}{4\pi\epsilon_0}\left(\frac{Q'}{b}+\frac{Q}{b}\right).$$
But this is equal to zero, because the sphere is earthed,

i.e.
$$V_B = 0,$$
$$Q' = -Q.$$

The potential V_A on the inner sphere is determined in the following way. The potential due to its own charge is
$$\frac{1}{4\pi\epsilon_0}\cdot\frac{Q}{a}$$
and the potential due to the induced charge Q' on B is
$$\frac{1}{4\pi\epsilon_0}\cdot\frac{Q'}{b} = \frac{1}{4\pi\epsilon_0}\cdot\frac{-Q}{b}.$$
(This is because the potential contributed by B, at all points inside it, is the same as it is at its own surface.) Hence, the resultant potential on the inner sphere is
$$V_A = \frac{1}{4\pi\epsilon_0}Q\left(\frac{1}{a}-\frac{1}{b}\right).$$
This gives the capacitance of the arrangement
$$C = \frac{Q}{V_A - V_B},$$
$$C = 4\pi\epsilon_0\frac{ab}{(b-a)}. \tag{13.19}$$
An alternative derivation considering the field strength E is illustrated in Fig. 13.7 Because the field strength contributed by a charged conductor to points which lie inside it is zero, the charge induced on the outer sphere B does not affect the field strength at points inside it at all, and its only effect is to reduce the field at points outside it to zero.

Hence, at any point between the two spheres, the resultant field strength is E, where
$$E = \frac{1}{4\pi\epsilon_0}\cdot\frac{Q}{r^2},$$
being due to the charge $+Q$ on the inner sphere A only. Hence the p.d. between the spheres is
$$V = \frac{1}{4\pi\epsilon_0}\int_b^a \frac{-Q}{r^2}\mathrm{d}r,$$
i.e.
$$V = \frac{1}{4\pi\epsilon_0}Q\left(\frac{1}{a}-\frac{1}{b}\right)$$
as before.

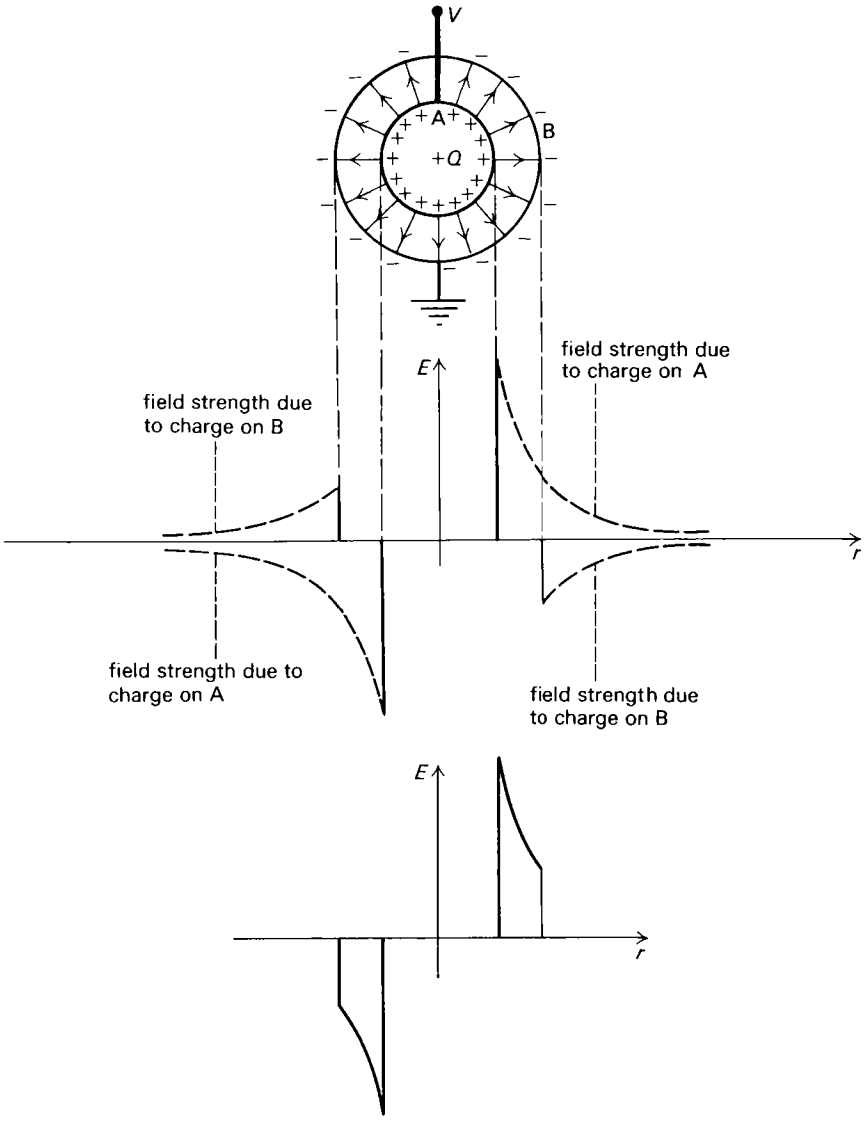


Fig. 13.7 Concentric spheres – the field strength (a) The field strengths contributed by the separate spheres (b) The resultant field strength

13.9 The field strength between two charged conducting sheets

If a p.d. is applied between two parallel plane conducting sheets, a fairly accurate idea of the distribution of lines of force may be obtained by first sketching in the equipotential surfaces. The result is something like that shown in Fig. 13.8.

If we neglect the spreading out of lines of force near the edges of the plates, where the equipotential surfaces are more widely separated, the field between the plates may be taken as uniform, i.e. the potential gradient is uniform, and we can say that

$$E = \frac{V}{d}, \tag{13.20}$$

where V is the p.d. applied across the plates, and d is their separation.

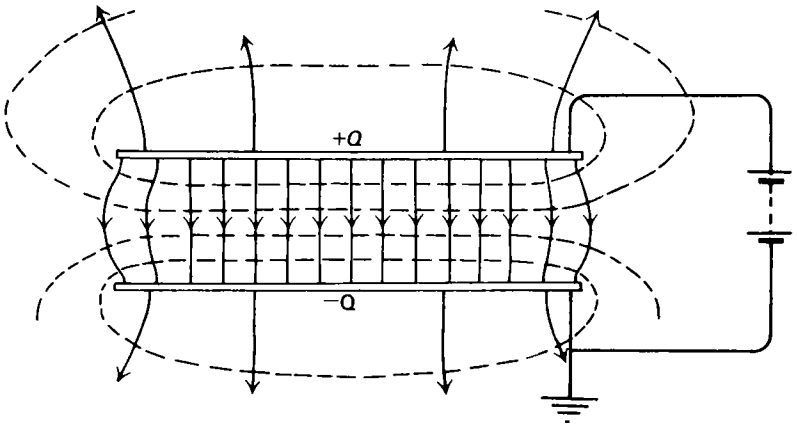


Fig. 13.8 The field and equipotential surfaces between parallel conducting sheets

To obtain an expression for the field strength in terms of the charges Q, $-Q$ on the plates, we use Gauss' theorem. The total number of lines of force is N, where

$$N = \frac{Q}{\epsilon_0}. \tag{13.12}$$

Neglecting the edge effect, and therefore assuming that these lines all run straight across from one plate to the other, we have

$$E = \frac{N}{A}, \text{ from 13.11,}$$

i.e.

$$E = \frac{Q}{\epsilon_0 A}, \tag{13.21}$$

where A is the area of each plate.

The capacitance of the capacitor made from two parallel plates is obtained by combining equations 13.20 and 13.21,

viz.

$$C = \frac{Q}{V},$$

i.e.

$$C = \frac{\epsilon_0 A}{d}, \tag{13.2}$$

and this is the equation which came under discussion in section 13.2, as being an example which is simplified by rationalizing.

A guard ring

If the separation d of the two capacitor plates, is very small, while the area A is big, as is very often the case, the edge effect represents only a very slight error in evaluating the capacitance from equation 13.2. In cases where the error would be appreciable, and accuracy is required, the edge effect is taken up by a guard ring, which has the same p.d. applied across it as the actual capacitor, but is not used to contribute to its charge – Fig. 13.9.

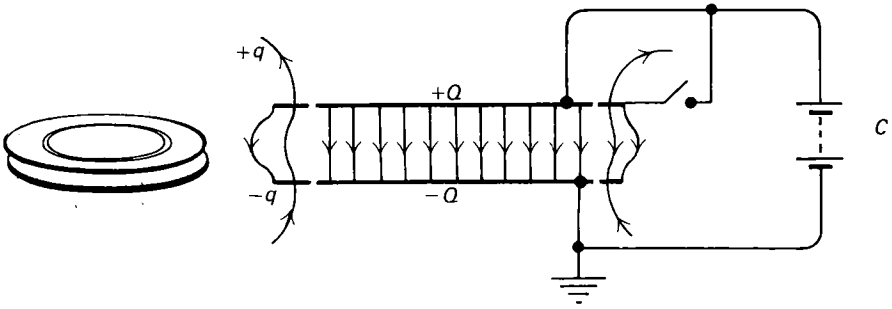


Fig. 13.9 A guard ring

The force on a charged particle placed in the field between two parallel plates

If a particle having charge $+Q'$ is placed in the field, the force F which acts on it is

$$F = EQ',$$

and

$$\left.\begin{aligned} F &= \frac{Q}{\epsilon_0 A} \cdot Q' \\ F &= \frac{V}{d} \cdot Q'. \end{aligned}\right\} \qquad (13.22)$$

13.10 The force of attraction between two charged plates

If two parallel plates hold equal and opposite charges, the force of attraction between them is given by

$$F = Q \times E',$$

where E' is the field strength at the position occupied by one plate, due to the charge on the other. This field is not as strong as the field strength E which exists in the space between the plates, because this is the effect of both charges, whereas the top plate, say, cannot experience any force due to its own field, but only that due to the field of the bottom plate.

In Fig. 13.10, the lines of force are shown which would be present due to the charge $+Q$ if the charge $-Q$ were absent from the other plate. Similarly the lines

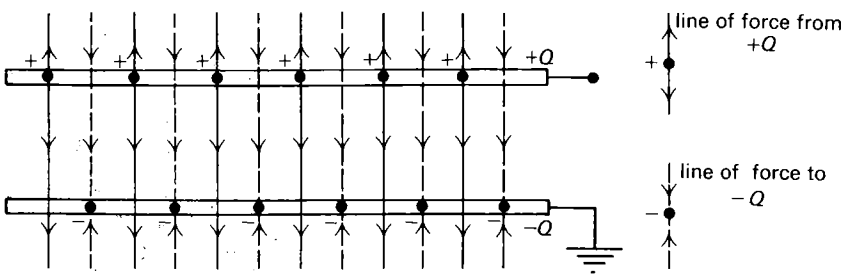


Fig. 13.10 The separate fields due to charges on two plates

of force are also shown, as dotted lines, which would be present due to charge $-Q$ alone. It can be seen that, from this point of view, each plate contributes, at any point, a field strength E', where

$$E' = \frac{E}{2},$$

with the result that the field strength at a point between the plates is

$$E = E' + E'$$

and the field strength at a point outside the plates is

$$E' - E' = 0.$$

Hence, it follows that the attracting force between the two plates is

$$F = Q \times E',$$

i.e.

$$F = \tfrac{1}{2}QE.$$

Substituting from equation 13.21, this gives

$$F = \tfrac{1}{2}\frac{Q^2}{\epsilon_0 A}, \tag{13.23}$$

whence, putting

$$Q = CV = \epsilon_0 A\frac{V}{d},$$

$$F = \tfrac{1}{2}\epsilon_0 A\frac{V^2}{d^2}. \tag{13.24}$$

Interpreting these two equations: if the plates are isolated, making the charge Q constant, the attracting force is independent of the separation (provided that d is not made so large that the edge effect becomes important); if, on the other hand, a battery is kept connected, making V constant, the attracting force is in inverse proportion to the square of the separation.

13.11 The attracted disc electrometer

All electrostatic voltmeters work on the principle of electrostatic repulsion or attraction between two charged metal plates or sheets, and the deflecting force depends on the square of the p.d., making them respond to a.c. as well as d.c. Equation 13.24 is the equation for one type of electrostatic voltmeter invented by Lord Kelvin, called the attracted disc electrometer.

The attracted disc is the metal disc A, which is attracted towards the earthed metal disc B, the attracting force being balanced against the upward contracting

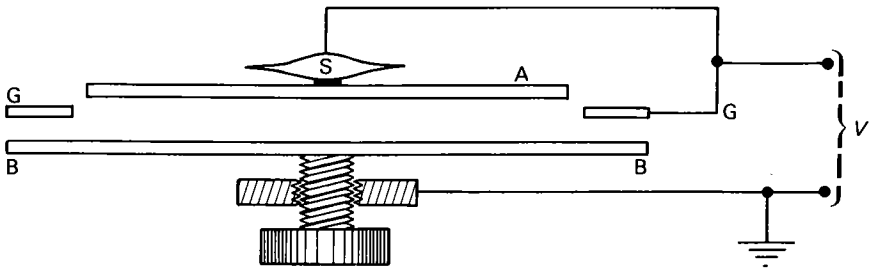


Fig. 13.11 The attracted disc electrometer

force of a leaf spring S. The disc is surrounded by a guard ring G. If a p.d. is applied between the two discs, causing disc A to be pulled downwards against the spring, the attracting force can be varied by altering the separation of the plates by means of an accurate screw movement. This control is used to bring disc A exactly in alignment with the guard ring. The force of attraction can afterwards be determined by loading plate A, when discharged, with weights, until these bring it in alignment with the guard ring again; in which case, we can write

$$mg = \tfrac{1}{2}\epsilon_0 A\frac{V^2}{d^2}, \tag{13.25}$$

d being the separation of the two plates when A lies in the same plane as the guard ring. All other quantities in the equation being measureable, the p.d. V can be calculated.

*13.12 The capacitance of coaxial cylinders

We suppose the inner cylinder to be raised to a potential, the outer cylinder being earthed. In this case, the electric field is entirely confined to the space between the two conductors, and charge $+Q$ on length l of the inner conductor induces charge $-Q$ on length l of the outer conductor.

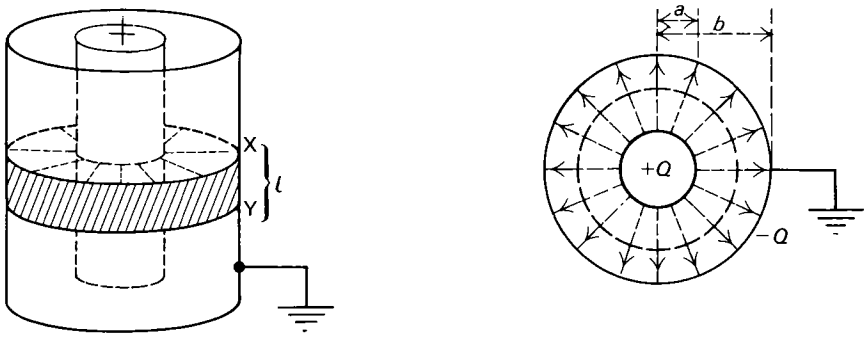


Fig. 13.12 Coaxial cylinders (a) (b)

Taking a slice XY contained between two planes perpendicular to the axis of the cylinders, distance l apart, the total number of lines of force contained between these two planes is N, where

$$N = \frac{Q}{\epsilon_0},$$

from Gauss' theorem.

Taking an imaginary cylindrical surface, radius r, coaxial with the two conductors, the total area of this surface which is crossed by these lines is A, where

$$A = l.2\pi r.$$

By symmetry, the field strength E is the same at all points on this surface, whence

$$E = \frac{N}{A}$$

$$= \frac{Q}{\epsilon_0}\cdot\frac{1}{2\pi r l},$$

and the potential difference, V, between the conductors is given by

$$V = \int_b^a -E\mathrm{d}r$$

$$= \int_b^a -\frac{Q}{\epsilon_0} \cdot \frac{1}{2\pi r l}\mathrm{d}r$$

$$= \frac{Q}{\epsilon_0 2\pi l} \int_b^a -\frac{1}{r}\mathrm{d}r,$$

i.e.

$$V = \frac{Q}{\epsilon_0 2\pi l} \log_e \frac{b}{a},$$

where a, b are the radii of the inner and outer cylinders respectively.

The capacitance of unit length is

$$\frac{C}{l} = \frac{Q}{lV}$$

$$= \frac{\epsilon_0 2\pi}{\log_e \frac{b}{a}}. \qquad (13.26)$$

One example of a pair of conductors which takes this form is the coaxial cable which is commonly used for conveying high frequency alternating current. Such cable consists of a central conductor surrounded by a coaxial earthed sheath, which serves as the return wire. The two are separated from each other by a good insulator, such as polythene. The virtue of this at high frequencies is that the alternating field is completely confined, and the tendency of high frequency current to lose energy by radiating it away is thereby eliminated. Although it is not the primary object to use the cable as a capacitor, its capacitance has some importance in affecting its performance, and capacitance features in the theory of high frequency cables.

13.13 The effect of dielectric

The equations derived so far for capacitance have assumed the electric field to be in a vacuum (or, in practice, in air). If the space between the conductors is completely filled with a dielectric, each equation is simply corrected by writing

$$C = \epsilon_r C_0, \qquad (13.27)$$

where C_0 is the capacitance with no dielectric present, ϵ_r is the relative permittivity of the dielectric, and C is the new value of the capacitance.

This was discussed in section 6.17, where it was pointed out that, although ϵ_r is very often constant in practice, it is not necessarily so.

If the dielectric only partially fills the space between the plates, its effect in simple cases can be deduced by the method suggested in the following examples.

Example 1

A slab of dielectric is inserted between two parallel conducting plates, leaving an air gap. Obtain an expression for the capacitance of the arrangement.

The actual capacitor is equivalent to two capacitors connected in series, as in Fig. 13.13(*b*), because the insertion of an uncharged metal plate on top of the dielectric (shown with its top and bottom surfaces separated), would make no difference.

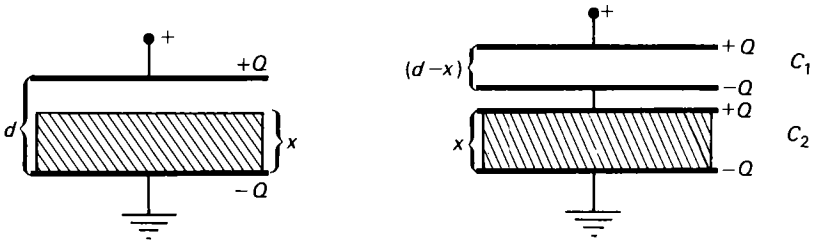


Fig. 13.13 (a) The actual capacitor
(b) An equivalent arrangement

Hence, $$\frac{1}{C}=\frac{1}{C_1}+\frac{1}{C_2}, \text{ (section 8.9)}$$

where $$C_1=\frac{\epsilon_0 A}{d-x}$$

and $$C_2=\epsilon_r\epsilon_0\frac{A}{x}$$

$$\therefore \quad \frac{1}{C}=\frac{1}{\epsilon_r\epsilon_0 A}\left[\epsilon_r(d-x)+x\right]$$

$$\therefore \quad C=\frac{\epsilon_r\epsilon_0 A}{[\epsilon_r d-(\epsilon_r-1)x]}$$

Example 2

A slab of dielectric is inserted between two parallel conducting plates to fill the whole space over an area A', leaving the remaining area $(A-A')$ to be air-spaced. Obtain an expression for the capacitance.

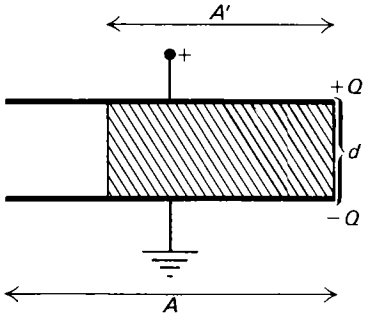


Fig. 13.14

This is, in effect, two capacitances C_1 and C_2 connected in parallel,

i.e. $$C=C_1+C_2 \quad \text{(section 8.9)}$$

where $$C_1=\epsilon_0\frac{(A-A')}{d}$$

and
$$C_2 = \epsilon_r \epsilon_0 \frac{A'}{d}.$$

$$C = \epsilon_0 \frac{[A + (\epsilon_r - 1)A']}{d}.$$

*13.14 The field strength at a point near the surface of any charged conductor

Suppose that the problem is to find the strength of the field E at a point P which lies just outside the surface of the charged conductor shown in Fig. 13.15. Suppose that the charge density (on the part of the conductor which is nearest to the point P) is taken to be

$$\frac{Q}{A} = \sigma.$$

The number of lines of force per unit area leaving this part of the conductor is

$$\frac{\delta N}{\delta A} = \frac{Q}{\epsilon_0 A} = \frac{\sigma}{\epsilon_0}.$$

But lines of force are perpendicular to a conducting surface when they leave the surface (assuming there is no current). It follows that, provided that the point P is sufficiently close to the surface so that the lines of force have not yet fanned out, the number of lines crossing unit area at P is also σ/ϵ_0.

The field strength at P is therefore

$$E = \frac{\sigma}{\epsilon_0}. \tag{13.28}$$

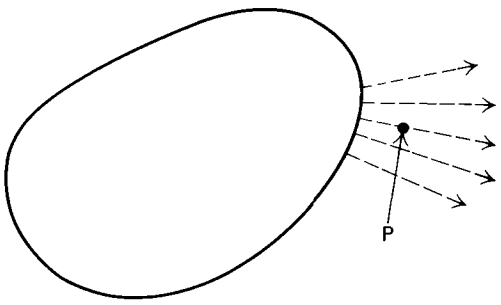


Fig. 13.15

*13.15 The pressure which tends to expand a charged sphere

Suppose that we treat the field strength E at a point P just outside a charged sphere as being the sum of two components, E_1 due to the charge on a small area δA of the surface, shown shaded in Fig. 13.16, and E_2 due to the charge on the remainder of the sphere. Then, from equation 13.28, we have

$$E = E_1 + E_2 = \frac{\sigma}{\epsilon_0}.$$

Now suppose that we cross through the surface δA, and move to the point R, which lies just inside the conductor. We know by experiment that the field strength inside is zero, a matter which will be further discussed in the next paragraph. The points P and R may be taken so close together that the value of E_2 is effectively the same for both. On the other hand, in crossing to the other side of the shaded area δA, E_1 has been reversed in direction. So we have:

at P, $$E_1 + E_2 = \frac{\sigma}{\epsilon_0},$$

at R, $$-E_1 + E_2 = 0,$$

whence $$E_1 = E_2 = \frac{\sigma}{2\epsilon_0}.$$

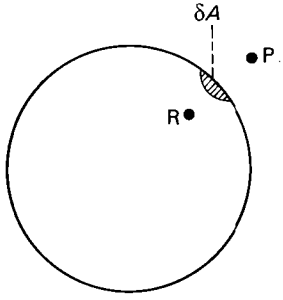


Fig. 13.16

Now the charge δQ on the area δA is equal to

$$\delta Q = \sigma \delta A$$

and this charge is subject to the field E_2, which is caused by the charge on the remainder of the conductor. Hence there is a force, δF, outwards on the area δA, given by

$$\delta F = E_2 \delta Q$$

$$= \frac{\sigma}{2\epsilon_0} \cdot \sigma \delta A.$$

The outward pressure p on the surface is therefore

$$p = \frac{\delta F}{\delta A} = \frac{\sigma^2}{2\epsilon_0}. \tag{13.29}$$

The cloud chamber

Surface tension represents a tendency of small drops to contract, resulting in a tendency for them to evaporate. This means that, owing to surface tension, a space can be made supersaturated with vapour, provided that it is dust free, because condensation would require, initially, the formation of very small drops. We have just shown that electric charge causes an outward pressure, i.e. a tendency of charged drops to expand, which is the reverse of the surface tension effect. As a consequence of this, if charged particles are produced in an air space which is supersaturated with a vapour, there is a tendency for vapour to condense on them. This is the principle of Wilson's cloud chamber, in which ionizing particles reveal

their passage by leaving a 'vapour trail', i.e. droplets of liquid condensed on the ions in their wake (section 24.2).

*13.16 The experimental evidence for the inverse square law

We now return to the equation at which this chapter began, expressing the inverse square law of force between charges concentrated at points,

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{QQ'}{r^2}. \tag{13.4}$$

It was pointed out earlier that the inverse square law rests entirely on experimental evidence. It was not convenient to discuss this further at the outset, because the most reliable experiments depend on derived equations. The omission can now be made up.

An attempt can be made to test the inverse square law by direct investigation of the repelling force between two small charged spheres, and such an experiment was carried out by Coulomb, using a torsion balance to measure the small force involved. The great difficulty about such a direct test is that the law refers to charges concentrated at points, and bodies small enough to be taken as approximating to points can hold very little charge. Equation 13.4 applies also, it is true, to charged spheres, but only when the charge on each of them is uniformly distributed. This condition cannot hold in practice, owing to induction effects between the two spheres repelling like charges to the far sides.

The real evidence for the inverse square law, therefore, is supplied by experimental verification of a sequence of equations derived from it, which are more open to test than the law itself. One particular experiment is, however, regarded as the crucial test, and this is the proof that there is no charge inside a hollow charged conducting sphere. The connection between this and the inverse square law is given below. This test was first made by Cavendish, a few years before Coulomb's torsion balance experiment. It has been repeated at intervals since, using increasingly sensitive methods of detection, and by this time the law is known to be true to a very high degree of accuracy. The evidence from atomic physics, such as Rutherford's experiments on the scattering of α particles by atomic nucleii, shows the law to hold, also, down to very small distances from the atomic nucleus.

Cavendish's proof

The apparatus used by Cavendish was, in principle, similar to Biot's hemispheres (section 6.13). The concentric spheres (Fig. 13.17(*a*)) are charged up, while the lid L is closed, in which case the inner sphere is in conducting connection with the outer sphere, and completely enclosed by it. The lid L is then lifted, thus breaking the connection between the two spheres, and the outer one is then earthed, to discharge it. A test is now made to see whether any charge remains on the inner sphere, using the most sensitive device available. All tests of this kind up to date have failed to reveal any trace of charge. Because this is so, it follows that there can have been no electric field inside the outer sphere when it was charged. It remains to demonstrate that this is a direct result of the inverse square law.

Consider the hollow charged conducting sphere drawn in Fig. 13.17(*b*). By symmetry, the charge density

$$\frac{\delta Q}{\delta A} = \sigma$$

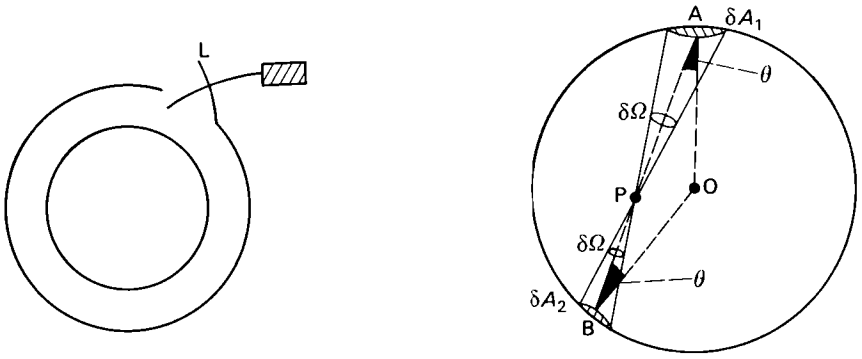


Fig. 13.17 The inverse square law (a) (b)

is the same at all parts of its surface. It is required to prove that the field strength is zero at the point P, chosen at random to be somewhere inside the sphere.

The point P is taken to be the vertex of two narrow cones, both having the same vertex angle $\delta\Omega$, drawn to cut off small areas $\delta A_1, \delta A_2$ on the surface of the sphere at A and B. The amounts of charge $\delta Q_1, \delta Q_2$ held by these two areas are

$$\delta Q_1 = \sigma \delta A_1$$

and

$$\delta Q_2 = \sigma \delta A_2.$$

Now, *on the assumption that the inverse square law holds*, the resultant field strength δE at P due to these charges is

$$\delta E = \frac{1}{4\pi\epsilon_0}\left[\frac{\delta Q_1}{\mathrm{AP}^2} - \frac{\delta Q_2}{\mathrm{BP}^2}\right]$$

$$= \frac{\sigma}{4\pi\epsilon_0}\left[\frac{\delta A_1}{\mathrm{AP}^2} - \frac{\delta A_2}{\mathrm{BP}^2}\right]$$

Now the solid angle $\delta\Omega$ is given by

$$\delta\Omega = \frac{\delta A_1 \cos\theta}{\mathrm{AP}^2} = \frac{\delta A_2 \cos\theta}{\mathrm{BP}^2},$$

where θ is the angle between the lines APB, and the normal AO, or BO, to each surface, the angle being the same for both A and B because they lie on a sphere.

Hence

$$\frac{\delta A_1}{\mathrm{AP}^2} = \frac{\delta A_2}{\mathrm{BP}^2}$$

and

$$\delta E = 0.$$

The effect of the whole sphere at P may be obtained by making up pairs of cones similar to those provided by δA_1 and δA_2, until the whole sphere has been accounted for, which leads to the conclusion that the resultant field at P is zero. This result has been obtained only by assuming the inverse square law to be true, which was an essential link in the argument.

13.17 The measurement of ϵ_0, the permittivity of a vacuum

One further aspect of Coulomb's inverse square law, equation 13.4, remains to be discussed, and this is the measurement of the constant ϵ_0, or of the constant α,

which is the same problem. It has been pointed out already (section 13.3) that this can be determined only by measuring it. Any of the capacitance expressions which have been derived in this chapter may be used as the basis of the experiment. Thus, taking as an example the parallel plate capacitor, we have

$$C = \epsilon_0 \frac{A}{d} \tag{13.2}$$

An air-spaced capacitor with this geometry can be set up (since the presence of air instead of a vacuum makes no appreciable practical difference) and the area A and separation d of the plates are measured. This leaves only the value of the capacitance C to be determined, and the method chosen for this must be one that does not depend on knowing the value of another capacitance, i.e. a comparison experiment would not be helpful. This is the special purpose for which Maxwell devised the method which was described in section 8.12, in which the capacitor is repeatedly charged up to a p.d. V and discharged through a microammeter, with a repetition frequency f. In these circumstances the capacitance becomes equivalent to a resistance R, where

$$R = \frac{1}{fC},$$

and it is in terms of R and f that it is measured. Any of the forms of the experiment which were described in section 8.12 may be used, and the measured values of A, d, and C are then inserted in equation 13.2 to give the value of ϵ_0.

This experiment can be repeated quite easily, although with only moderate accuracy, in the school laboratory, using a reed switch working from the 50 Hz mains supply to perform the continuous switching cycle. The capacitor can be made up, for example, from two large sheets of plate glass about 0.5 m square, each coated with a conducting layer of aquadag, or of metal foil, and separated by spacers of measured thickness, such as pieces of microscope slide, or thin perspex washers. To obtain greater accuracy, however, a guard ring would be needed in order to avoid the edge effects which put the capacitance equation 13.2 in error. For this reason, the parallel plate form of capacitor is not the best arrangement to adopt as the basis of a precise measurement, although it is the easiest one to set up. In accurate determinations, concentric spheres or coaxial cylinders have been used in preference, the outer conductor in each case being earthed. In these forms of capacitor, the lines of force are more effectively confined, and their capacitance equations therefore apply with greater accuracy.

*13.18 A formal statement of Gauss' theorem

Gauss's theorem was presented in section 13.6 in terms of lines of force. More formally, it is stated mathematically, as follows.

Suppose that a closed surface is drawn, which encloses a number of electric charges Q_1, Q_2, etc. Taking a small area δA of this surface, which includes a point P, suppose that the field strength at P is equal to E, and that the field makes an angle θ with the normal to the surface. Then the normal component E_n of the field strength at P is $E_n = E \cos \theta$. Gauss' theorem states that the integral, $\oint E_n \mathrm{d}A$, taken over the whole of the closed surface, is related to the total charge, ΣQ enclosed, by the equation

$$\oint \epsilon_0 E_n \mathrm{d}A = \Sigma Q, \tag{13.30}$$

this being a direct consequence of the inverse square law.

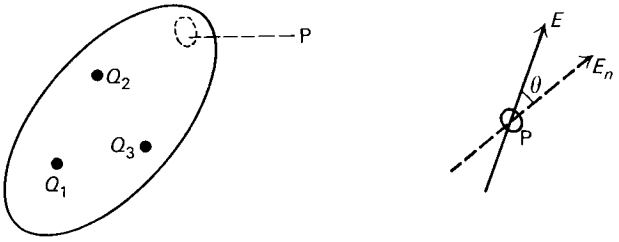


Fig. 13.18 Gauss' theorem

Expressed in terms of lines of force, $E_n \delta A$ is the number of lines which cross the small area δA, so that the integral $\oint E_n \, dA$ is equal to the total number of lines which cross the closed surface.

QUESTIONS

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ F/m}$$

$$\frac{1}{4\pi\epsilon_0} = 9.0 \times 10^9 \text{ SI units}$$

13.1 A conducting sphere, radius 10 cm, is charged to a positive potential of 3000 V. Assuming the sphere to be effectively isolated, calculate:

(*a*) the total charge held by the sphere,

(*b*) the number of electrons which have been removed from it,

(*c*) the energy stored,

(*d*) the field strength at the surface of the sphere,

(*e*) the potential at 5 cm from its centre,

(*f*) the potential at 50 cm from its centre,

(*g*) the work which would be required to move a charge of $+4 \times 10^{-9}$ C from a point 50 cm from its centre to another point at 20 cm from its centre.

13.2 Two circular flat parallel metal plates, each of radius 10 cm, are separated by an air gap 1.0 mm across, and charged to a p.d. of 250 V.

Calculate:

(*a*) the capacitance of the plates,

(*b*) the energy stored,

(*c*) the amount of charge stored on each plate,

(*d*) the number of electrons which have been transferred from one plate to the other,

(*e*) the force which would act on a small oil drop situated between the plates, if the drop holds a charge of 1.6×10^{-18} C,

(*f*) the force of attraction between the two plates.

13.3 The insulation of the air in the atmosphere breaks down owing to ionization when it is subjected to an electric field which exceeds 3×10^6 V/m in strength.

(*a*) Deduce the radius of the smallest conducting sphere which could be charged to a potential of 10^6 V before starting to discharge through the

atmosphere. Why would the leakage tend to occur at a lower potential than this in practice?

(*b*) Calculate the largest amount of charge which could be retained by a conducting sphere of radius 1 mm and the potential at which this charge would be held, assuming the sphere to be exposed to the atmosphere.

13.4 Two spheres, radii r_1 and r_2, are separated by a distance d which greatly exceeds r_1 and r_2, and they are joined by a conducting wire and charged to a potential V.

Compare (*a*) the charges on the two spheres, (*b*) the charge density on the surfaces of the two spheres, (*c*) the field strengths at the two surfaces.

Describe how you could compare the two charges, and the two charge densities, in practice, and comment on the relevance of your results to the phenomenon of electrical discharge at a pointed conductor.

13.5 (*a*) Estimate the electric field strength and the potential at the surface of a sodium ion Na^+, by assuming it to be an isolated conducting sphere of radius 10^{-10} m. (O & C)

(*b*) The Bohr model of the hydrogen atom consists of a single electron moving in a circular orbit round the hydrogen nucleus, which is a single proton. When the atom is in the ground state, the radius of this orbit is approximately 5×10^{-11} m. Calculate (i) the potential of the electron, (ii) the frequency with which it revolves in its orbit.

(*c*) An α-particle with velocity 3.5×10^6 m/s strikes a gold film, atomic number $Z = 79$. Calculate the closest possible distance of approach between the α-particle and a gold nucleus.

The mass of a helium atom is 6.64×10^{-27} kg. (C)

(*d*) Assuming the nucleus of a nitrogen atom to be effectively a sphere of radius 3×10^{-15} m, calculate the electric potential at the surface of the nucleus. What is the minimum energy in joules, which an α-particle would require to enable it to surmount this potential barrier?

(The atomic number of nitrogen is $Z = 7$. An α-particle, He^{2+}, carries charge $+2e$.)

(*e*) A stream of protons, accelerated in a vacuum through a p.d. V, is directed at a target of lithium. Treating the lithium nucleus as being a sphere of radius 2.3×10^{-15} m, use conventional equations of electrostatics to estimate the minimum value needed for the accelerating potential V, if protons are to penetrate some of the lithium nucleii. (Quantum theory predicts some exceptions to this limitation.)

(The atomic number of lithium is $Z = 3$.)

13.6 A sphere of insulating material with radius 1.0 cm is given an electrostatic charge uniformly distributed over its surface. Its potential is then 10^4 V. An uncharged metal sphere of radius 1.0 cm is placed 20 cm away, and is then connected to earth.

Draw three graphs showing qualitatively how the potential varies along a line through the centres of the spheres, (*a*) before the metal sphere is in position, (*b*) when the metal sphere is in position but uncharged, (*c*) when the metal sphere is earthed. In case (*c*), estimate the charge on the metal sphere. (CAM. SCH.)

13.7 The points A, B, and C form an equilateral triangle of side z. Point charges of equal magnitude Q are placed at A and B. Find the electric field strength and the potential at C due to these charges when (*a*) both charges are positive, and (*b*) the charge at A is positive and the charge at B is negative.

O is the midpoint of AB and XOY is the perpendicular bisector of AB; XO and OY are very large distances compared with AB. Draw rough graphs to show how the magnitude of the electric field strength and the potential vary along the line XOY in cases (*a*) and (*b*) above. (O & C)

13.8 Two small polystyrene spheres, each of mass 5.0×10^{-2} g and having metallized surfaces, are suspended from the same point by insulating threads. The

distance from the centre of each sphere to the point of suspension is 50 cm. When each sphere carries a charge $+Q$, the system is in equilibrium with the two centres separated by 10 cm. Estimate the value of Q, and explain why your answer can be only approximate.

13.9 Charges of magnitude $+1.0 \times 10^{-10}$ C, $+2.0 \times 10^{-10}$ C, $+2.0 \times 10^{-10}$ C, and $+1.0 \times 10^{-10}$ C, are placed respectively at the corners A, B, C, D of a square of side 10 cm. Find (*a*) the electric field strength at O, the point of intersection of the diagonals, (*b*) the electric potential at O, and at E, the midpoint of the side AD, (*c*) the work done if a point charge of 2.0×10^{-10} C is moved from E to O.

13.10 A spherical conductor A, of radius a, is surrounded by a concentric thin conducting spherical shell, radius b. Find an expression for the capacitance of (i) A when B is insulated and uncharged, (ii) B when A is insulated and uncharged, (iii) the system made up by connecting A to B.

13.11

A charge Q is uniformly distributed throughout a spherical volume of radius r. Obtain an expression for the electric field strength at distance x from the centre, (*a*) when $x > r$, and (*b*) when $x < r$.

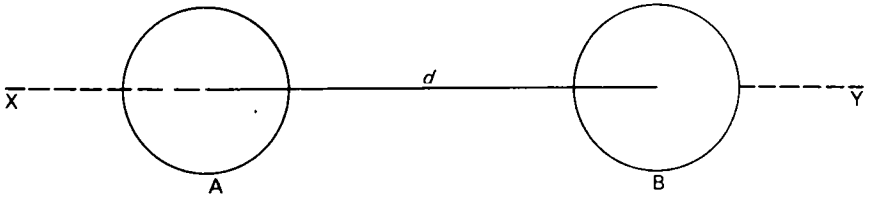


Fig. e 72

In the diagram, A and B are identical, solid, conducting spheres, each having a charge $+Q$ and situated in vacuo. Explain why the force between the spheres is not equal to $\frac{1}{4\pi\epsilon_0} \cdot \frac{Q^2}{d^2}$.

Sketch graphs showing the variation of (*a*) the electric potential and (*b*) the electric field strength along the line XY. In what way, if any, would these graphs be modified if A and B were thin-walled, hollow, conducting spheres? (O & C)

13.12 Two similar spherical particles each have mass m and carry charge Q. Initially, one is at rest on a smooth plane surface and the other is projected along the surface directly towards it from a great distance with velocity v. Describe qualitatively the subsequent motion of the particles, and find an expression for the closest distance to which they approach each other. (CAM. SCH.)

(This problem can be solved largely by applying conservation of momentum and conservation of energy.)

13.13 An isolated sphere, initially at earth potential, having radius 10 cm, is charged to a potential of 300 V. How much energy is stored?

The sphere is disconnected from the source of potential and then, without touching this sphere, a second one, made up of two thin hemispherical shells, of radius 15 cm, is arranged concentrically about it and connected to earth. What is the p.d. between the inner and outer spheres, and how much energy is now stored in the system?

Account for the change in energy. (O & C)

13.14 A capacitor consists of two concentric conducting spheres, the inner one, of radius a, being insulated and the outer one, of radius b, being earthed. How much electrostatic energy will be stored in the capacitor if $a = 10^{-1}$ m,

$b = 1.2 \times 10^{-1}$ m, and the inner sphere is charged to a potential of 900 V? What would be the effect on (*a*) the potential of the inner sphere, (*b*) the electrical energy, if the outer sphere were removed, the inner one remaining insulated? How do you account for the change in energy? (O & C)

13.15 A parallel plate capacitor consists of two square plates each of side 30 cm, separated by a distance of 2 mm in air. One of the plates is earthed, and the other plate is charged to a potential of 3000 V. If the charge stored is found to be 1.20 μC, calculate the permittivity, ϵ_0, of a vacuum. (The effect of the air between the plates is negligible.)

13.16 An oil drop of mass 3.25×10^{-12} g falls vertically, with uniform velocity, through the air between two vertical parallel plates which are 2.0 cm apart. When a p.d. of 1000 V is applied across the plates the drop moves towards the negatively charged plate, its path being inclined at 45° to the vertical.

Explain why the vertical component of its velocity remains unchanged, and calculate the charge on the drop. If the path of the drop suddenly changes to one at 26° 30′ to the vertical, and subsequently to one at 37° to the vertical, what conclusions can be drawn? (O & C)

13.17 A capacitor consisting of two parallel circular plates each of diameter 20 cm placed 0.1 cm apart in air is permanently connected to an electrostatic voltmeter of capacitance *C*. The system is charged to a p.d. of 80 V and then isolated. When the plate separation is increased to 0.5 cm the voltmeter reading rises to 300 V. Explain the change in p.d. and calculate the value of *C*. (O & C)

13.18 A capacitor, consisting of two parallel plates in contact with an insulating slab of dielectric of thickness 0.10 cm and relative permittivity 4.0, has one plate earthed and the other at a potential of + 100 V. The positive plate is then moved away from the dielectric to produce an air gap of 0.07 cm, giving a plate separation of 0.17 cm. Calculate the potential of the positive plate in this new position, assuming that it remains insulated. (O & C)

13.19 A parallel plate air-spaced capacitor is made from two discs, each of radius 8 cm, placed 1 cm apart. The upper plate is earthed and hung from a balance. The lower plate is initially at earth potential, and is then raised to a potential of 1500 V. What is the change in the apparent weight of the upper plate? Ignore edge effects. (CAM. SCH.)

13.20 An attracted disc electrometer consists of two parallel circular plates, each of radius 3.0 cm, and separated by a distance of 2.0 mm. What is the p.d. between the plates when the force of attraction between them is 2.5×10^{-5} N?

13.21 Calculate the p.d. to which the plates of a parallel plate capacitor, each of area 100 cm^2 and 1 cm apart in vacuo, must be charged if the force of attraction is to be 5.0×10^{-4} N. Neglect edge effects.

When a parallel-sided slab of insulating material, of the same area and shape as the plates and 0.5 cm thick, is placed symmetrically between them while they remain connected to the source of p.d., the force of attraction is doubled. Calculate the relative permittivity, ϵ_r, of the insulator. (O & C)

13.22 State Gauss' theorem in electrostatics. Prove the theorem for the particular case of a point charge at the centre of a spherical surface.

Assuming the general validity of the theorem, obtain the expression for the electric field strength at a point very close to a conductor on which the surface density of charge is σ. What is the mechanical force experienced by a small element δA of the surface of the conductor?

An insulated metal plate of area 0.20 m^2 is held at a distance of 3.0×10^{-3} m from a parallel earthed conducting plate; it carries a charge of 10^{-6} C. What is the mechanical force on the insulated plate? In which direction does it act? What is the p.d. between the surfaces? Find the mechanical force on the insulated plate and the p.d. between the plates when the separation is increased from 3.0×10^{-3} m to 3.00×10^{-2} m. (O)

13.23 The diagram represents two proof planes, each having area 3.0 cm^2, which are initially held in contact with each other in an electric field of strength

3×10^4 V/m, with the planes perpendicular to the field. If the two proof planes are now separated from each other while still in the field, calculate the charge stored in each of them. (Assume that the planes were initially discharged.)

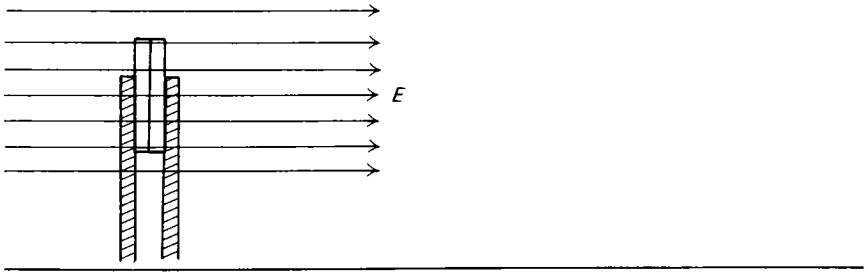


Fig. e 73

13.24 How would you show experimentally that the free charge on a conductor rests entirely on its surface?

State Gauss' theorem and use it to find the electric field strength at a point close to the surface of a conducting sphere of radius r carrying a charge Q.

A model belt-driven Van de Graaff machine has a sphere of radius 0.50 m, which is maintained at the steady potential of 30 kV. The resistance of the supporting column is 10^9 Ω. The belt, 0.05 m wide, travels at 0.80 m/s. Find (*a*) the current supplied to the sphere, (*b*) the power supplied to the sphere, (*c*) the surface density of charge on the belt and on the sphere, (*d*) the energy stored by the sphere, (*e*) the potential gradient close to the surface of the sphere. (O)

*13.25 Two concentric conducting spherical shells of radii r_1 and r_2 ($>r_1$) carry charge Q_1 and Q_2 respectively. Derive expressions for the potentials of the shells, assuming them to be of negligible thickness. Is it possible for the inner shell to have a positive charge but be at a negative potential, and if so, under precisely what conditions?

How would your answers to the preceding paragraph be modified if the space between the shells were filled with oil of relative permittivity ϵ_r? (OXF. SCH.)

*13.26 Two equal and opposite charges, +Q and −Q, are placed at points A and B, distance $2s$ apart. O is the midpoint of AB. Find the electric potential and the electric field strength at (*a*) a point P lying on AOB at a distance r from O ($r \gg s$), (*b*) a point Q lying on the perpendicular bisector of AB at a distance r from O ($r \gg s$). (O & C)

*13.27 A capacitor consists of a 10 cm radius inner sphere and a 15 cm radius outer sphere which is earthed. The inner sphere is charged to a potential of 100 V with respect to the outer sphere and then isolated. What is (*a*) the capacitance of the capacitor, (*b*) the stored electrostatic energy?

A non-conducting liquid is then poured through a small hole until the lower half of the space between the spheres is filled with liquid. Regarding this arrangement as two capacitors in parallel, what are the new capacitance and the stored electrostatic energy, if the relative permittivity of the liquid is 2.0?

How do you account for the change in energy?

*13.28 Three concentric hollow metal spheres of radii 4, 6, 8 m have charges $+8 \times 10^{-9}$, -6×10^{-9}, and $+4 \times 10^{-9}$ C respectively. Show graphically how the potential and electric field strength vary along a radial line.

What are the potentials and electric field strengths at 2, 5, 7, and 10 m from the centre if the spheres are filled with air? (OXF. SCH.)

*13.29 Three conducting concentric spherical shells, each 1.0 mm thick, have outside radii 1.0, 2.0, 3.0 cm respectively. The outer sphere is earthed, and the innermost one is charged to a potential of 10^3 V, the intermediate sphere being isolated and uncharged. Calculate the charge on the innermost sphere and the potential of the intermediate one.

If the spheres were made of elastic material, would the electric forces tend to make them expand, or contract? (O & C)

*13.30 A metal sphere of radius a is surrounded by a concentric spherical metal shell of radius b. The inner sphere is earthed, and the outer shell holds a charge $+Q$. Deduce the magnitude of the induced charge Q' on the inner sphere, and sketch graphs showing how (i) the potential due to Q, (ii) the potential due to Q', (iii) the resultant potential, vary with distance r from the centre.

By treating the system as being two capacitors in parallel, or otherwise, derive an expression for the capacitance of two concentric spheres when the charge is applied in this way.

*13.31 Derive an expression for the electric field strength at a point close to a charged conductor.

A small spherical pendulum bob of mass 10 g has a uniform electrical charge density on its surface. When a horizontal sheet with the same uniform charge density is placed below the bob, the period of the pendulum is changed by 1 per cent. Calculate the potential of the bob before the charged plate is placed in position.

(Take the value of g to be 10 m/s^2.) (CAM. SCH.)

*13.32 A parallel plate capacitor consists of two horizontal metal plates, each with area 10^{-2} m^2. The lower plate is fixed and earthed and the upper plate is suspended by a spiral spring so that, when there is no p.d. across the capacitor, the separation of the plates is 2.0 cm. If a force of 0.040 N extends the spring by 1.0 cm, obtain an expression relating the separation x of the plates to the p.d. V between them.

Show that the p.d. is a maximum when $x = 4/3$ cm, and find this maximum value. What happens if a battery of p.d. greater than this maximum value is connected across the capacitor? (O & C)

To illustrate your answer to this question sketch two graphs, using the same pair of axes, to show (*a*) how the tension of the spring varies with the distance x separating the plates, (*b*) how the force of attraction between the plates varies with their separation x. Comment on the stability of the system.

*13.33 Two isolated parallel metal plates separated by a distance d carry equal and opposite charges σ per unit area. The gas between the plates is irradiated with a short burst of X-rays in such a way that a layer of gas of negligible thickness midway between the plates is partially ionized, there being n electrons and n positive ions per unit area of the layer. The electrons drift towards the positive plate with velocity v_1 and the positive ions drift in the opposite direction with velocity v_2 ($< v_1$).

If the charge on the electron has magnitude e and edge effects can be neglected, how does the p.d. between the plates vary with time from the instant of irradiation until all the charges have reached the plates? (O & C)

*13.34 The plates of a parallel plate capacitor are each of area A, and are separated by a distance x of air. The plates are initially charged to a p.d. V from a battery, and then isolated. The separation is now increased to y, where $y > x$. Deduce the gain in energy of the capacitor, and trace the source of this energy.

Starting with the same initial conditions of separation x and p.d. V, the plate separation is again increased to y, but with the battery still connected, to keep the p.d. at the constant value V. Deduce (*a*) the amount of charge which flows through the battery, (*b*) the change in energy stored in the capacitor, (*c*) the energy supplied to or taken from the battery, (*d*) the work done in separating the plates.

*13.35 The plates of a parallel plate capacitor are each of area A and are separated

by a distance $4d$. The following operations are performed in sequence: (*a*) the plates are connected to a battery of voltage V; (*b*) with the battery connected, the plate separation is reduced to d; (*c*) the capacitor is then disconnected from the battery and the plate separation restored to $4d$.

Calculate for each operation (*a*), (*b*), (*c*) the energy supplied by the battery and the mechanical work done in changing the plate separation. (OXF. SCH.)

(These derivations involve tracing the charge and energy in the capacitor at the different stages.)

*13.36 A capacitor consists of two parallel plane metal sheets of rectangular shape, each having sides of length a, b, separated by a distance d. A slab of dielectric, relative permittivity ϵ_r, is inserted between the plates in such a way that an area $b \times x$ of the capacitor is filled with the dielectric, leaving an area $b \times (a - x)$ without dielectric. Write down an expression for the capacitance C of the arrangement.

The capacitor is now charged up to a p.d. V and its plates are isolated. Deduce the small change, δW, in the stored energy which results when the dielectric is moved a further small distance δx into the space between the plates, and hence deduce the force F which acts on the dielectric. Give an expression relating this force to (*a*) the charge Q on one plate, (*b*) the p.d., V between the plates.

*13.37 A cable consists of a wire or radius 1 mm imbedded coaxially in a dielectric cylinder of relative permittivity $\epsilon_r = 2.25$, sheathed in metal braid. If the capacitance of 1 m of the cable is 6×10^{-11} F, what is the radius of the sheath?

This dielectric has the property that any part of it which is exposed to an electric field greater than 10^7 V/m is permanently damaged and becomes a conductor. What is the maximum p.d. which can be applied between the wire and the braid without damaging the dielectric?

Discuss whether the whole of the dielectric is damaged when a p.d. slightly greater than this maximum is applied. (O & C)

Chapter 14

Magnetic flux

14.1 The magnetic interaction between currents

The basic difference between an electrostatic force and a magnetic force is this; whereas the electrostatic force acts between two static electric charges, a magnetic force is the interaction between charges which are moving relative to the observer. This statement may seem irrelevant, it is true, to the case of two permanent magnets, where nothing is apparently moving, but atomic theory attributes this magnetism, at any rate in part, to the motions of the electrons in the atoms.

In developing electrostatic theory, the force between point charges provides an obvious starting point, but there is nothing quite so simple to take this place in magnetism. Lacking an obvious starting point, therefore, we will choose one, instead, for its convenience, namely the magnetic interaction between currents flowing in two straight parallel conductors. The advantage of this is that it is the arrangement upon which the definition of the ampere is based (section 1.11). This definition specifies two parallel conductors which are infinitely long and placed in a vacuum. In practice, however, the presence of air would make no appreciable difference, because air does not significantly affect the magnetic properties of a vacuum, but any substance which might have some such effect, notably any ferromagnetic material, must not be in the vicinity.

In these conditions, we have

$$F = \beta 2 I I' \frac{l}{r}, \tag{14.1}$$

where I, I' are the values of the currents flowing in the same direction in the two conductors, r is the distance which separates the conductors, F is the force of attraction which one conductor exerts, owing to the currents, on length l of the other, β is a constant.

The definition of the ampere allots a specific value to the constant β (section 1.11). Using SI units,

$$\beta = 10^{-7}\ N/A^2. \tag{14.2}$$

The facts that the force of attraction F between the two conductors exists, and that its magnitude is in inverse proportion to their separation r, are matters which rest on experimental evidence. There is little point, however, in devising a direct experiment, because this particular arrangement of conductors, although theoretically simple, cannot be made up in practice, except approximately. The validity of the equation is established, rather, by testing other similar relations which are consistent with it, but apply to configurations of coils which can be used with greater accuracy.

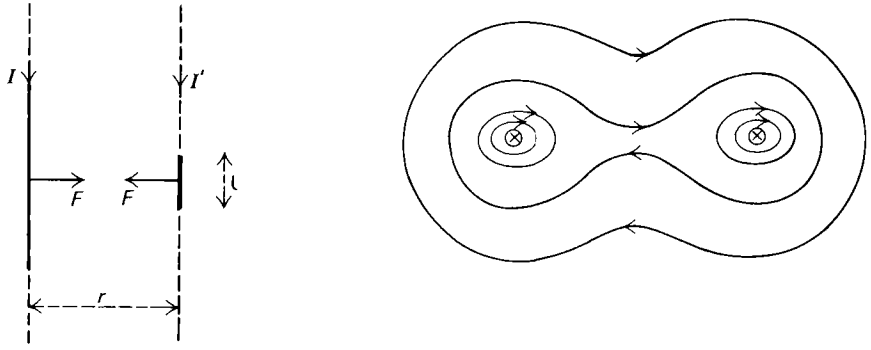


Fig. 14.1 The magnetic attraction between currents flowing in two parallel conductors (a) (b) The resultant flux of the two currents

14.2 Magnetic flux density B

We can now introduce magnetic flux density B in a very similar way to that in which electric field strength E was introduced in electrostatics, as representing the influence which a current produces in the space around it.

The conductor which carries the current I' is treated as being in the magnetic field produced by the current I which flows in the other conductor, and to experience the force F in consequence. Following the definition of flux density which has already been given in section 7.12, we have

$$F = BI'l, \tag{14.3}$$

and the relation between B and the current I in the straight conductor which causes it must therefore be

$$B = \beta\frac{2I}{r}, \tag{14.4}$$

so that, combining these two equations, we shall obtain equation 14.1 back again. In this way, equation 14.1, which gives directly the force acting between two currents, has been split up into two parts, one giving the effect of a magnetic field on a conductor placed in it which carries current, and the other giving the strength of the field in terms of the current which causes it.

This procedure is illustrated in Fig. 14.2. Whereas the resultant flux due to the two currents is shown in Fig. 14.1, a different point of view has been adopted in Fig. 14.2. This shows only the flux caused by current I in the first conductor, and

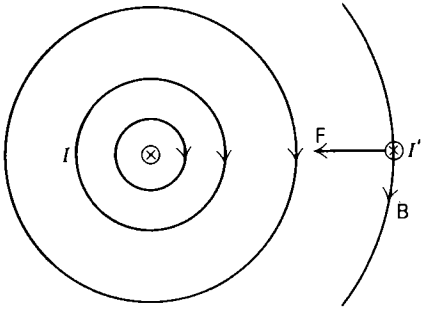


Fig. 14.2 The magnetic attraction between currents flowing in parallel conductors – one conductor is represented as being placed in the field due to the other

the conductor which carries current I' is shown as being placed in the influence of this flux and experiencing force F due to the motor effect.

The question of the direction of the magnetic flux is not as simple as that of the direction of an electric field in electrostatics, and this aspect has been discussed in chapter 7. From several possible definitions we will choose the simplest, which identifies it with the direction in which a compass would point, if placed in the field, i.e. that in which a force would act on the north pole of a magnet. The case of the force acting on moving charge or on a current is essentially more complicated, because the direction of the movement or orientation of the conductor which carries the current must also be defined, and the full force F given by equation 14.3, for example, acts only if the conductor is perpendicular to the flux; the direction of this force can be found by applying Fleming's left hand rule.

This chapter will be concerned with the effects by which magnetic flux can be detected and measured, and with relations developed from equation 14.3. Since this equation represents a definition of flux density B, we can take it to be valid in all circumstances, quite independently of whether or not there are magnetic materials in the vicinity.

Relations such as equation 14.4, which connect flux density B with the current which causes it, will be discussed in the next chapter; it will be here that the nature of the medium must be taken into account.

14.3 The units of magnetic flux density and magnetic flux

The SI unit of magnetic flux density is called the tesla, or, until recently, the weber per square metre. This is defined by inserting SI values for F, l, I, in equation 14.3,

i.e.
$$B = \frac{F}{Il}, \tag{14.3}$$

1 tesla = 1 N/A·m = 1 Wb/m^2.

[A more general form for this equation is

$$B = \frac{\delta F}{I\delta l \sin\theta}, \tag{14.5}$$

where θ = the angle between the flux and the conductor, and δF = the small force which acts on a small length, δl, of the conductor.]

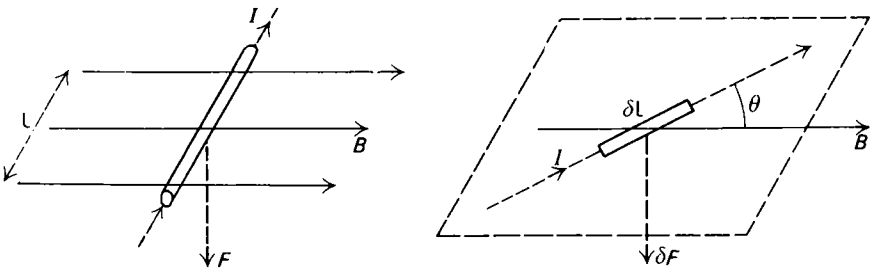


Fig. 14.3 The force on a conductor which carries current and is placed in a magnetic field (a) $F = BIl$ (b) $\delta F = BI\,\delta l \sin\theta$

The name, 'flux density', which we are adopting to refer to the magnetic field property B, arises from the device of representing the magnetic field quantitatively by lines of magnetic flux. (An alternative name is 'magnetic induction'.)

If we take an area A of an imaginary surface, perpendicular to the direction of the flux, the total quantity of flux Φ, which crosses this area is defined by the equation

$$\Phi = BA. \tag{14.6}$$

More generally, supposing a small area δA to be so placed that the normal to this area makes angle θ with the flux, the equation becomes

$$\delta\Phi = B\,\delta A \sin\theta \tag{14.7}$$

The SI unit of magnetic flux Φ is called the weber. Hence, rewriting equation 14.6 in the form

$$B = \frac{\Phi}{A}$$

the magnetic field property B becomes flux density, and its unit, the tesla, earns its alternative name, now obsolescent, weber per square metre.

We may, if we choose, think of the flux in terms of flux lines, the total flux Φ being pictured as being the total number of lines Φ, and a flux of, say, 10 Wb being a flux of 10 lines. This interpretation is similar to the convention of using electric lines of force to represent the strength E of an electric field.

For practical purposes, the SI units, both of B and of Φ, are really very large. A very effective permanent magnet can produce at most a flux density of about 1 T, and the maximum available using a steady magnetising current in a coil with a ferromagnetic core is 3 or 4 T. A typical example of a weaker field is the horizontal component, B_{hor}, of the earth's magnetic flux density in England, i.e. the part of it which affects the compass. The value of this is

$$B_{\text{hor}} = 1.8 \times 10^{-5}\ \text{T}.$$

This is in contrast with the SI unit of electric field strength E, which is inconveniently small from a practical point of view, i.e. $E = 1$ V/m represents a very weak electric field.

It is, perhaps, because the SI unit of flux density B is so large that the corresponding c.g.s. unit is at the present time in very common use, and so we will include it here. The c.g.s. unit of flux density B is the gauss, where

$$1 \text{ tesla} = 10^4 \text{ gauss}, \tag{14.8}$$

and the c.g.s. unit of magnetic flux Φ is the maxwell, where

$$1 \text{ weber} = 10^8 \text{ maxwell}. \tag{14.9}$$

The gauss is, of course, maxwell per square centimetre.

14.4 The force on a charged particle moving across magnetic flux

If a charge Q moves in a direction perpendicular to magnetic flux, a force acts on it, which is given by

$$F = BQv, \tag{14.10}$$

where B is the flux density and v is the velocity of the charge. The direction of this force, being given by the left hand rule, is perpendicular to the flux and to the direction in which the charge is moving.

The more general form of this equation is

$$F = BQv \sin\theta, \tag{14.11}$$

where θ is the angle between the velocity v and the direction of B.

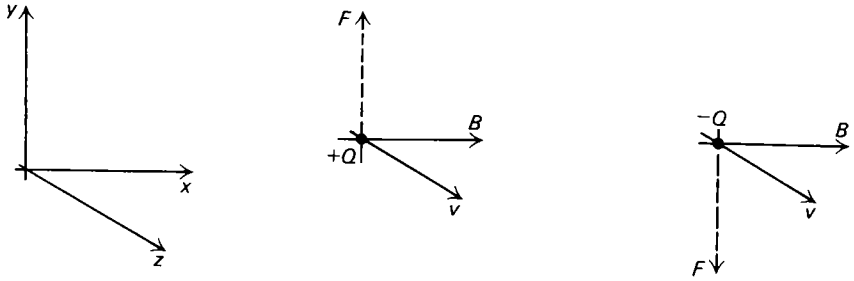


Fig. 14.4 The force on a charge which moves across magnetic flux

This very important equation was quoted in chapter 7.12, as being an alternative equation by which flux density B may be defined, and it was contrasted with the electrostatic equation

$$F = EQ, \tag{14.12}$$

as illustrating the essential difference between a magnetic and an electric field.

Accepting equation 14.3 however, as the definition of flux density, equation 14.10 will now be derived from it.

Suppose we have a continuous stream of particles, each carrying charge $+Q$, moving with velocity v perpendicular to the magnetic flux, and suppose there are n such particles in length l of the stream. The number of particles passing any point in unit time is equal to the number contained in a distance v, and this is equal to nv/l.

The electric current I represented by this stream of charged particles is equal to the amount of charge which passes any point in unit time, i.e.

$$I = Q\frac{nv}{l}.$$

The force which acts on length l of the stream is given by equation 14.3, viz

$$F_n = BIl,$$
$$F_n = BQnv.$$

But this length contains n particles. Hence, the force F which acts on a single particle of charge Q is

$$F = \frac{F_n}{n},$$

i.e.

$$F = BQv. \tag{14.10}$$

14.5 The curved path of a charged particle crossing a uniform magnetic field

Suppose that a particle of mass m, carrying charge $+Q$, is moving in a plane which is perpendicular to a magnetic field of uniform flux density B. The deflecting force due to the field, being perpendicular to the motion, does not alter the speed of the particle, but merely deflects it. Because the deflected path is also perpendicular to the flux, it follows that the magnitude of the deflecting force is constant, and its direction is always perpendicular to the motion of the particle. This is the condition for the path to be circular. The electromagnetic force, $F = BQv$, is the centripetal force which holds the particle to this circular track, of radius r

i.e.
$$BQv = \frac{mv^2}{r},$$

$$v = Br\frac{Q}{m}. \tag{14.13}$$

Equation 14.13 is very important in practice and a number of circumstances in which it is applied will be found later on in this book. Examples are the determination of the charge/mass ratio for an electron (section 21.5), the mass spectrograph (section 21.9), the cyclatron (section 26.9), and the application of a magnetic field across a cloud chamber (section 24.2) to reveal the speed or charge of ionizing particles.

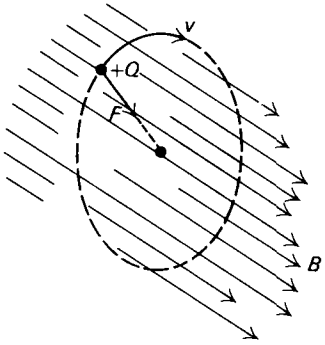


Fig. 14.5 The circular path of a charged particle moving in a plane perpendicular to uniform magnetic flux

14.6 The torque on a coil

If a coil, carrying current, is placed with its perpendicular axis (meaning the axis ns in Fig. 14.6) making an angle with a magnetic field, a torque acts on the coil,

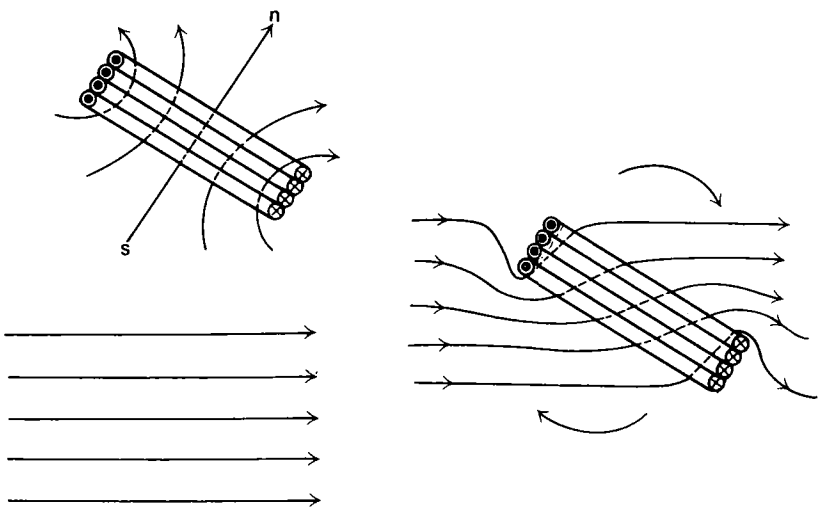


Fig. 14.6 The torque on a coil (a) The field of the coil (b) The external field (c) The resultant field

trying to turn it to bring this axis parallel with the field. This is because in this condition the field due to the coil would have the same direction as that of the external field. The effect is an important one in practice, being the basis of both the moving coil meter and the d.c. electric motor.

Suppose the coil in the diagram has rectangular cross section, with width b and depth l, and that it has N turns, carrying current I. The forces which act on each of the four sides of the coil are shown in Fig. 14.7.

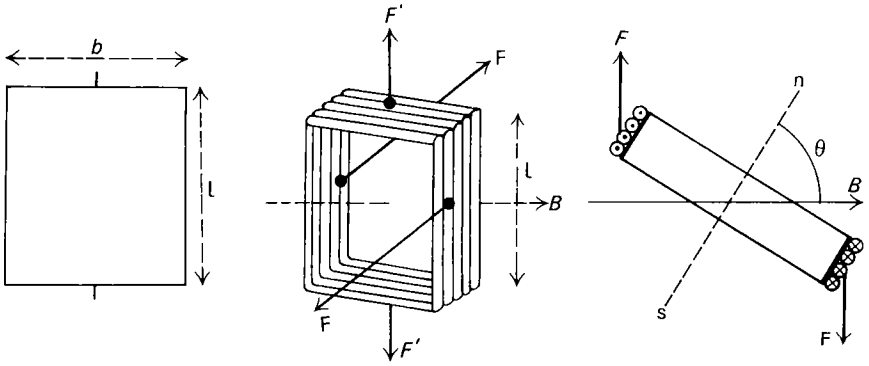


Fig. 14.7 The forces which act on the coil

The forces F' which act on the top and bottom sides of the coil are equal and opposite, and act in the same straight line. But the forces F on the two vertical sides are separated by a distance $b \sin \theta$, where θ is the angle between the central axis ns, perpendicular to the coil, and the field. There is therefore a torque, Γ on the coil, given by

$$\Gamma = F \cdot b \sin \theta,$$

where $$F = BNlI,$$

$$\therefore \quad \Gamma = BNlI \cdot b \sin \theta,$$

where $$lb = A, \text{ the area of the coil.}$$

Hence $$\Gamma = BNAI \sin \theta. \qquad (14.14)$$

Although the above derivation has assumed the coil to have rectangular shape, equation 14.14 also applies to the general case, A being the area enclosed by the coil, whatever its shape may be.

14.7 The torque on a magnet

A magnet which is held with its axis at an angle θ with a magnetic field experiences a torque in the same way as the coil, acting to bring the axis of the magnet parallel with the field, which is, of course, the basic principle of the compass. We write

$$\Gamma = Bm \sin \theta, \qquad (14.15)$$

where m represents a property of the magnet which takes the place of the product, NAI, for the coil. This is called *the magnetic moment* of the magnet, and it is defined by equation 14.15. Likewise, the product, NAI, is sometimes referred to as being the electromagnetic moment of a coil which carries current I.

To justify writing equation 14.15 for a magnet, when flux density B has been defined in terms of its effect upon a current flowing in a conductor, it is really necessary to establish the equivalent behaviour of magnet and coil by experiment. This equivalence was first established by Ampere.

[Electromagnetic theory includes some rather artificial difficulties. One of these concerns the definition of magnetic moment m. Many text books write equation 14.15 in the form

$$\Gamma = Hm \sin \theta,$$

where $H = B/\mu_0$. The factor μ_0 (which is equal to $4\pi\beta$) will appear in equations in the next chapter. The magnetic moment of a coil then becomes

$$m' = \mu_0 NAI.$$

In this way, the same name and symbol are attached, according to the point of view, to two different properties, which differ from each other by the factor μ_0.

[The version which is defined by equation 14.15 will be used consistently in this book.]

14.8 Electromagnetic induction

A mathematical relation can be attached to the laws of electromagnetic induction (section 7.13) by applying conservation of energy to the principle of the electric motor. We will take the case of the primitive form of linear motor shown in Fig. 14.8. [The argument could equally well be applied to a rotating armature, but this would merely complicate the theory.]

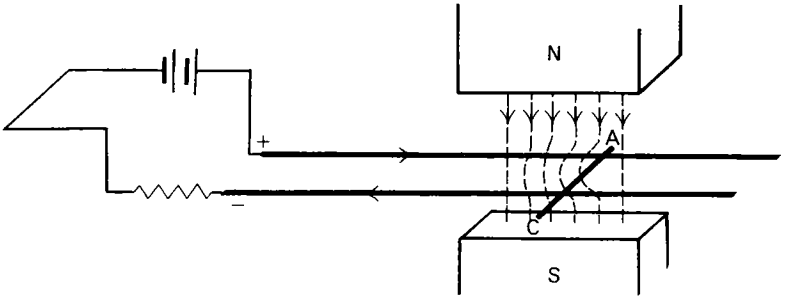


Fig. 14.8 A linear electric motor

A straight conductor AC bridges the gap between two horizontal busbars, across which a p.d. is applied, with the result that current I flows in the conductor in the direction from A to C. Being perpendicular to the flux between the poles of a horseshoe magnet, the conductor is subject to a motor force F moving it to the right in the diagram, where

$$F = BIl, \tag{14.3}$$

l being the distance which separates the busbars.

We now suppose that this force to the right is made to act against a load which exerts an equal and opposite force F, with the result that the conductor is moving to the right with the constant velocity v. Work is now being done by this simple motor at the rate P, where

$$\begin{aligned} P &= Fv \\ &= BIlv. \end{aligned} \tag{14.16}$$

The energy required to do this work must come from the source of the current, and this energy conversion comes about as the result of electromagnetic induction; the moving conductor is cutting magnetic flux, with the result that an e.m.f. V is induced in it, which acts against the flow of current. It is the energy supplied from the source in driving current against this e.m.f. which provides the motor power. The power converted in this way is

$$P = VI. \tag{14.17}$$

Whence, from equations (14.16) and (14.17),

$$VI = BIlv,$$

$$\therefore \quad V = Blv. \tag{14.18}$$

Now

$$lv = \frac{A}{t}$$

where A is the area swept out by the length AC and the moving conductor in time t,

$$V = B\frac{A}{t}.$$

But the product BA is equal to the change in the amount of flux Φ which threads the circuit in time t, due to the motion of the conductor, i.e.

$$BA = \Phi_2 - \Phi_1.$$

Hence

$$V = \frac{\Phi_2 - \Phi_1}{t}, \tag{14.19}$$

where $(\Phi_2 - \Phi_1/t$ is the rate at which flux is being cut.

Alternatively, we can write

$$V = \frac{d\Phi}{dt} \tag{14.20}$$

and this is a mathematical expression of Faraday's law of electromagnetic induction,

14.9 Considering the force on the charge carriers in a moving conductor

The above equation may also be derived by the different approach of considering the effect of the magnetic flux on the individual charge carriers in the moving conductor (e.g. the conduction electrons, in the case of a metal).

The arrangement suggested in Fig. 14.9 is the same as the last one, except that

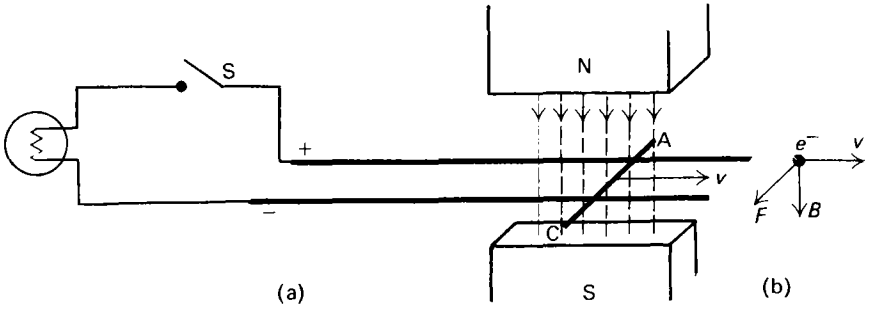


Fig. 14.9 A simple generator, on the principle of the linear motor (a) A generator (b) The force on an electron in the conductor AC

no external source is connected across the busbars, and the system is being used as an electric generator, not as a linear motor. The conductor AC is being moved to the right with velocity v, not because of a motor force, but because an external force is being used.

Since the conductor is cutting the flux by moving with velocity v, every conduction electron in it is doing the same, and consequently each electron is subject to a motor force F, which drives it towards the end C of the conductor. In this way, electrons begin to accumulate in the busbar connected to C, causing a deficiency of electrons in the busbar connected to A, and so there is now a p.d. V between the two, representing an e.m.f. V induced in the moving conductor.

Equilibrium is reached (supposing switch S to be open) because this induced e.m.f. sets up a potential gradient, i.e. an electric field, in the conductor AC, given by

$$E = \frac{V}{l},$$

and this causes a force F' to act on each electron in opposition to the motor force F, where

$$F' = Ee$$
$$= \frac{V}{l}e,$$

e being the charge of the electron.

But
$$F = Bev, \tag{14.10}$$

and so we have

$$Bev = \frac{V}{l}e,$$

whence the induced e.m.f. is

$$V = Blv, \tag{14.18}$$

which is the same equation as before.

It will be noticed that in both these derivations the direction of the induced e.m.f. is such that work must be done if induced current is taken out, which is in accordance with Lenz' law.

14.10 Electromagnetic induction caused by a changing current

Referring to Fig. 14.9, the effect of moving the conductor AC may be described in two ways: the electromagnetic induction may be associated with the fact that the moving conductor is cutting flux, or with the fact that, owing to the increasing area enclosed, the circuit is enclosing an increasing amount of flux. From the second of these two points of view, equation 14.20 would be interpreted as stating that the induced e.m.f. in the circuit is equal to the rate at which the amount of flux threading the circuit is changing. On this interpretation, the equation applies also (although we cannot give a direct theoretical proof of it) to the mutual induction between two coils, or to the self induction in a single coil, i.e. to cases where no conductors are moving, but flux variation is provided, instead, by variations in the magnetizing current. If the coil has N turns, the equation becomes

$$V = \frac{N\mathrm{d}\Phi}{\mathrm{d}t}. \tag{14.21}$$

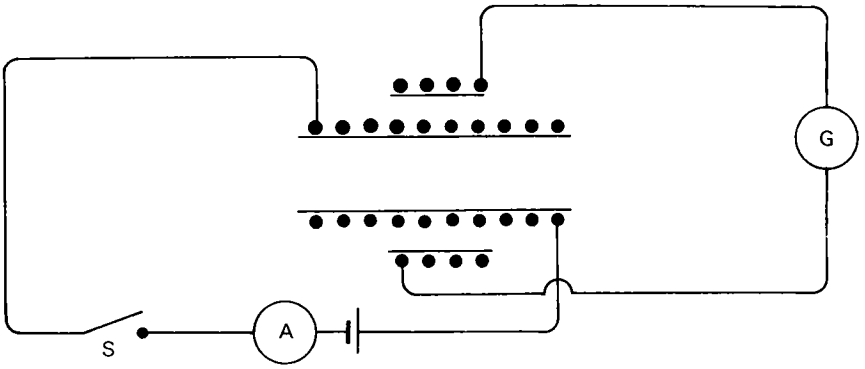


Fig. 14.10 Electromagnetic induction by changing flux – mutual induction

14.11 The charge set in motion by a change in flux

To maintain an e.m.f. continuously in a coil, the flux which threads it must be changing continuously, and a single abrupt change produces only a transient e.m.f., i.e. an electrical impulse, analogous to an impulsive force in mechanics.

Suppose the flux threading a coil having N turns is changed from Φ_1 to Φ_2 in a time t. The mean value of the induced e.m.f., which exists only for the period of time t, is given by

$$V = \frac{N(\Phi_2 - \Phi_1)}{t}.$$

If the coil is short circuited, and the total resistance of the circuit is R, the mean value of the current I, which also lasts for time t, is

$$I = \frac{V}{R}.$$

Hence, the total charge which circulates, as the result of changing the flux, is

$$Q = It$$

$$= \frac{V}{R}t,$$

$$Q = \frac{N(\Phi_2 - \Phi_1)}{R}, \qquad (14.22)$$

which is independent of time t.

[More correctly, we start from equation 14.21. At any instant,

$$V = N\frac{d\Phi}{dt},$$

$$I = \frac{N}{R} \cdot \frac{d\Phi}{dt},$$

and

$$Q = \int I dt$$

$$= \frac{N}{R}\int \frac{d\Phi}{dt} dt$$

$$= \frac{N(\Phi_2 - \Phi_1)}{R}.\Big]$$

14.12 Some examples

Example 1 Faraday's disc

A metal disc with radius 5.0 cm *is turned at constant speed about a central axle which is perpendicular to its plane. A uniform magnetic field, also perpendicular to the plane of the disc, has flux density* 0.050 T. *As a result of electromagnetic induction, a p.d. is produced between two brushes, one of which makes contact with the circumference of the disc, and the other with the surface of its axle. If the radius of the axle is* 0.5 cm, *and the disc makes* 4 *revolutions in every second, deduce the value of the p.d.*

ness of disc
Her ?

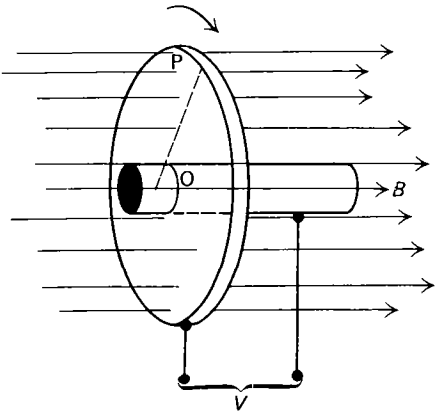


Fig. 14.11 Faraday's disc

In the course of one complete rotation of the disc, every radius in it cuts once all the flux which crosses the plane of the disc. An e.m.f. is consequently induced in every radius, so that there is a steady p.d. between axle and circumference. If the flux which threads the area outside the axle is Φ, we have:

$$\begin{aligned}\Phi &= BA\\ &= B\pi(r_2^2 - r_1^2),\end{aligned}$$

where r_2 is the radius of the disc and r_1 is the radius of the axle. If the frequency of revolution of the disc is f, then all this flux is cut f times in unit time. Hence,

$$V = fB\pi(r_2^2 - r_1^2), \tag{14.23}$$

where, in the example given,

$$\begin{aligned}f &= 4\ \mathrm{s}^{-1},\\ B &= 0.050\ \mathrm{T},\\ r_2 &= 5.0 \times 10^{-2}\ \mathrm{m},\\ r_1 &= 0.5 \times 10^{-2}\ \mathrm{m},\\ V &= 4 \times 0.050 \times \pi(25.0 - 0.25) \times 10^{-4}\ \mathrm{V}\\ &= 1.57 \times 10^{-3}\ \mathrm{V}.\end{aligned}$$

Example 2 A search coil

A small coil, which encloses area 1.0 cm^2 *and has* 20 *turns, is connected to a ballistic galvanometer. The coil is initially held in a magnetic field of flux density* 0.2 T, *with*

the face of the coil perpendicular to the flux. If the resistance of the circuit is 30 Ω, *deduce the value of the charge which flows through the galvanometer, if the coil is moved out of the field.*

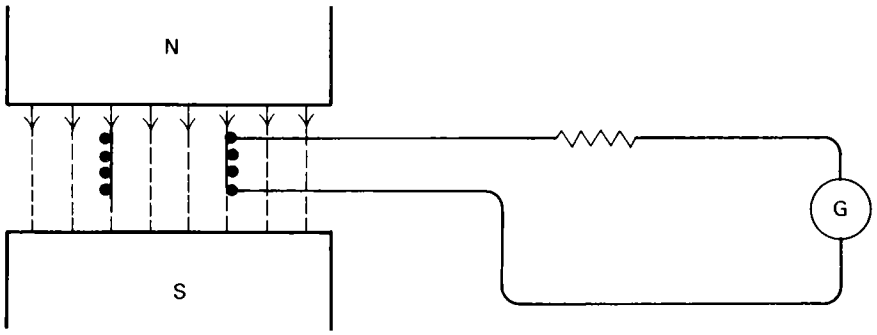


Fig. 14.12 A search coil

Initially, the flux Φ which threads the coil, is

$$\Phi_1 = BA,$$

where A is the area enclosed by the coil.

Finally, the flux Φ_2 which threads the coil is zero,

$$\Phi_2 = 0.$$

We have

$$Q = \frac{N(\Phi_2 - \Phi_1)}{R}, \tag{14.22}$$

$$\therefore \quad Q = \frac{-NBA}{R},$$

where

$$N = 20,$$
$$B = 0.2 \text{ T},$$
$$R = 30\ \Omega,$$
$$A = 1.0 \text{ cm}^2$$
$$= 10^{-4} \text{ m}^2,$$

$$Q = \frac{20 \times 0.2 \times 10^{-4}}{30} \text{ C}$$

$$= 1.33 \times 10^{-5} \text{ C}$$
$$= 13.3\ \mu\text{C}.$$

Example 3 A rotating coil—a simple a.c. generator

A circular coil with 500 *turns, enclosing area* 70 cm², *is rotated at constant speed about a diameter which is perpendicular to a uniform magnetic field. If the coil is turned at* 3000 rev/min, *and the flux density of the field is* 1.5×10^{-2} T, *obtain an expression giving the e.m.f. in the coil at the instant t, and deduce its r.m.s. value.*

At the instant shown in the diagram, the central axis of the coil, which is perpendicular to its plane, makes an angle θ with the magnetic flux. At this instant, the amount of flux threading the coil is

$$\Phi = BA \cos \theta,$$

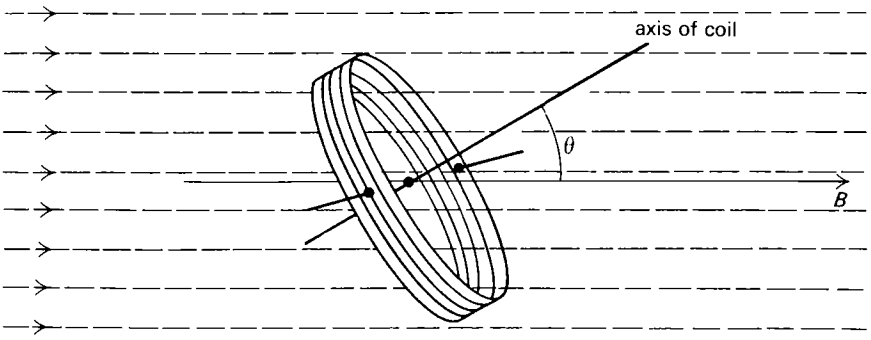


Fig. 14.13 A rotating coil

and the induced e.m.f. in the coil is

$$V = N\frac{d\Phi}{dt},$$

$$V = -NBA \sin\theta \frac{d\theta}{dt},$$

where the negative sign is of no importance in this problem.

Taking the time t to be zero at the instant when the angle θ is zero, we have

$$\theta = 2\pi f t$$

and

$$\frac{d\theta}{dt} = 2\pi f,$$

where f is the frequency of revolution of the coil. Hence,

$$V = NBA \cdot 2\pi f \sin(2\pi f t), \qquad (14.24)$$

i.e. the e.m.f. is an alternating one, with frequency f. The r.m.s. value of the e.m.f. is

$$V_{\text{r.m.s.}} = NBA\frac{2\pi f}{\sqrt{2}},$$

where

$$N = 500,$$
$$B = 1.5 \times 10^{-2}\ \text{T},$$
$$A = 70\ \text{cm}^2$$
$$= 7 \times 10^{-3}\ \text{m}^2,$$
$$f = 50\ \text{s}^{-1},$$

$$V_{\text{r.m.s.}} = \frac{500 \times 1.5 \times 10^{-2} \times 7 \times 10^{-3} \times 2\pi \times 50}{\sqrt{2}}\ \text{V}$$
$$= 11.6\ \text{V}.$$

14.13 The measurement of magnetic flux by using a search coil

An apparatus might be devised to measure magnetic flux or flux density which is based on any of the equations which have been derived in this chapter. Only a limited number of these are used in practice, however, and the conventional

method is to use a search coil, on the principle of example 2 in the last section. The actual design of the coil would depend on the particular circumstances and the strength of the field under investigation, and several examples of this will occur later in the text. A typical use would be the one given in the example, i.e. to measure the powerful flux between the poles of a horseshoe magnet. In this case the coil encloses a small area, which is precisely measured, and it is attached to a rod by which it can be handled. The leads are connected to a ballistic galvanometer. The coil is placed in the field, with its plane perpendicular to the flux. It is then rapidly removed from the field, and the throw, θ, of the galvanometer is observed. We have

$$Q = \frac{N\Phi}{R} \tag{14.22}$$

$$= k\theta,$$

where R is the resistance of the complete circuit, and k is the calibration constant, Q/θ, of the meter. The mean value of the flux density B over the area of the coil is given by

$$B = \frac{\Phi}{A}.$$

The principle of the ballistic galvanometer was outlined in section 8.11, where it was explained that a very rapid flow of the charge is essential. This is why the flux through a search coil must be changed abruptly.

Unless a calibrating device is included, this method is available only for comparing two fields; provided that the resistance of the circuit is kept unaltered, the ratio B'/B of the flux densities is equal to the ratio θ'/θ of the two throws; the effect of damping can be neglected because this cancels out.

Calibrating the ballistic meter by using a standard mutual inductance

Calibration of the meter can be effected by including a standard mutual inductance in the circuit of the search coil, as shown in Fig. 14.14. This is an arrangement of

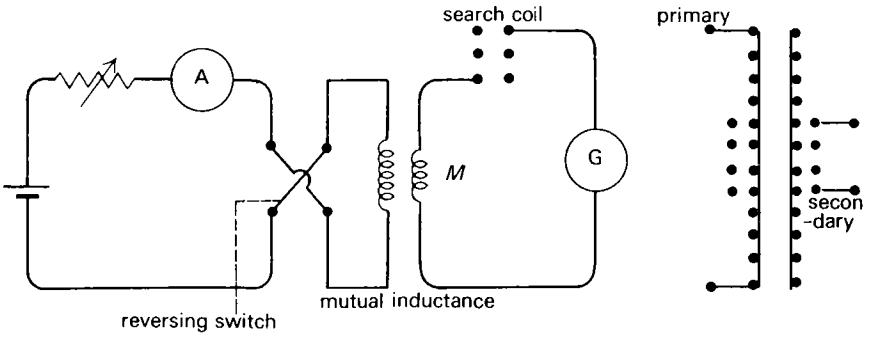


Fig. 14.14 Circuit to calibrate a ballistic galvanometer (a) (b) The mutual inductance

two coils, carefully made to have a mutual inductance M which can be accurately calculated. It may, for example, consist of one long solenoid which is the primary coil, and a shorter secondary coil wound coaxially round the primary. The mutual inductance between the two coils may be calculated by using equations of the type developed in the next chapter, provided that the core is effectively non-magnetic.

Suppose the search coil is first removed from the flux which is to be measured, and the galvanometer throw is θ, where

$$\frac{N\Phi}{R} = k\theta. \tag{14.25}$$

It is now required to measure or eliminate the constant k. To do this, a measured current I is reversed in the primary coil of the mutual inductor. Mutual induction results in a rapid circulation of charge in the circuit of the secondary coil, causing a galvanometer throw θ'. Thus the e.m.f. induced in the secondary coil is V, where

$$V = M\frac{\mathrm{d}I}{\mathrm{d}t}.$$

Hence, the induced current I' in the secondary circuit is

$$\begin{aligned} I' &= \frac{V}{R} \\ &= \frac{M}{R}\cdot\frac{\mathrm{d}I}{\mathrm{d}t} \end{aligned}$$

and the total charge which flows in consequence is

$$\begin{aligned} Q' &= \int I'\mathrm{d}t \\ &= \frac{M}{R}\int_{-I}^{+I}\frac{\mathrm{d}I}{\mathrm{d}t}\mathrm{d}t \\ &= \frac{M}{R}\cdot 2I. \end{aligned}$$

$$\frac{M}{R}\cdot 2I = k\theta'. \tag{14.26}$$

The constant k can now be eliminated from equations 14.25 and 14.26:

$$\frac{N\Phi}{M\cdot 2I} = \frac{\theta}{\theta'}. \tag{14.27}$$

It will be noted that both the search coil and the secondary coil of the mutual inductor, being connected in series, are included in the circuit for both parts of the experiment. In this way, it is arranged that the circuit resistance R shall be constant, with the result that it cancels out in deriving equation 14.27.

14.14 The vibration magnetometer

It is a familiar fact that a compass, placed where a magnetic field is strong, wastes no time in indicating its direction; the compass oscillates very rapidly, almost quivering, as it swings into line with the flux. Put the compass in a weak field, and it is sluggish, oscillating only very slowly. This variation in frequency provides a very simple means of comparing flux densities.

For small amplitude vibrations, the motion of the oscillating magnet is simple harmonic, having periodic time t, given by the equation

$$t = 2\pi\left(\frac{\mathcal{J}}{m\cdot B}\right)^{\frac{1}{2}}, \tag{14.28}$$

i.e. the frequency of vibration is

$$f = \frac{1}{2\pi}\left(\frac{m \cdot B}{\mathcal{J}}\right)^{\frac{1}{2}}, \tag{14.29}$$

where $\mathcal{J}$ is the moment of inertia of the magnet about its axis of vibration, and m is its magnetic moment.

Thus, if the magnet is moved from one field to another, the two flux densities may be compared simply by timing the oscillations:

$$\left.\begin{aligned} \frac{1}{t^2} &\propto B, \\ f^2 &\propto B \end{aligned}\right\} \tag{14.30}$$

Used in this way, the magnet is called a vibration magnetometer. The dimensions of magnet chosen, and the method of pivoting or suspending it, depend on circumstances such as the strengths of the fields under investigation. Typically, it consists of a very short permanent magnet, say 1 cm long, set in a small cylindrical metal block to give it a suitable moment of inertia, and suspended by a torsionless thread of silk or nylon in a draught excluding glass container. An example of its application might be to make a survey of the effective value of the earth's horizontal flux density component, B_{hor}, at various places in a laboratory. It would be necessary merely to time the oscillations of the magnet at these different places.

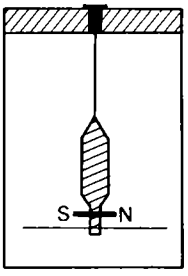


Fig. 14.15 A type of vibration magnetometer

The underlying assumption in the principle of the vibration magnetometer is that the magnetic moment of its magnet is constant, so that the frequency of oscillation varies only with the strength of the field. There can be circumstances, of course, when this assumption is unjustified, if for example, the magnetometer is put too close to a powerful magnet.

*14.15 The Hall effect

A sensitive type of magnetometer can be made from a small slice of semiconducting crystal by using what is known as the Hall effect. This effect was first detected by Hall in a metallic conductor in 1880, and we will describe it first with reference to a metal, although it is only in a semiconductor that it is sufficiently pronounced to have found a practical application.

When charged particles moving across a space such as a cloud chamber or a cathode ray tube are subjected to a transverse magnetic field, the motor force, $F = BQv$, deflects them, as we have already seen, into a curved path. When a solid conductor, on the other hand, such as a metal wire, which is carrying current, is

subjected to the field, the motor force, $F = BIl$, acts on the conductor as a whole, and there is no tendency for it to bend into a curved shape. This is because the charge carriers, in this case, are not in isolation, but are drifting through the solid, which holds equal charge of opposite sign. The motor force on a metallic conductor acts, in the first place, on the moving electrons, and these transfer the force to the metal conductor as a whole, due to the electrical attraction between them and the positive ions of the metal. Nevertheless, the electrons are deflected towards one side of the metal, causing a very small transverse p.d. in it. This is the effect which was discovered by Hall, and the p.d. is called a Hall potential difference. This is illustrated in Fig. 14.16.

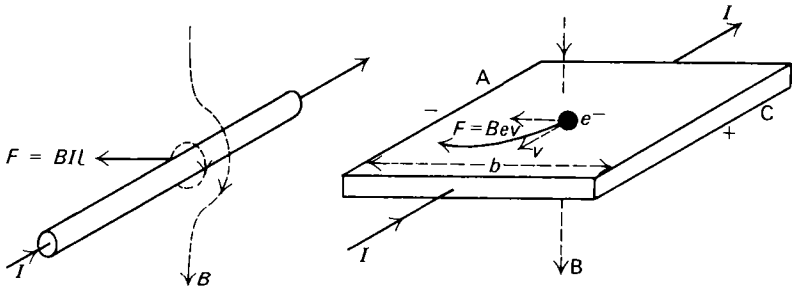


Fig. 14.16 The Hall effect in a sheet of metal (a) The motor force on a metal wire (b) The Hall effect in a metal sheet

Figure 14.16(*a*) shows the direction of the motor force, BIl, which acts on a straight conductor carrying current and placed perpendicular to a magnetic field. The electrons which carry the current are moving in the opposite direction to the conventional positive current I labelled in the diagram. In the metal sheet of the second diagram, a conduction electron is shown, which is deflected by the motor force towards the edge A of the sheet. This leaves the side C with an excess of positive charge, and the result is the Hall potential difference. An expression for the size of this p.d. can be derived as follows.

The redistribution of charge in the sheet becomes stable when the magnetic deflecting force, Bev, on an electron is balanced by the electrostatic force, Ee, which results from the Hall potential difference V. This p.d. represents a transverse potential gradient across the sheet, given by $E = V/b$ where b is the width of the sheet. Whence, we have

$$Bev = \frac{V}{b}e,$$

i.e.

$$V = Bbv. \tag{14.31}$$

It can be seen that, to obtain an appreciable p.d., we require a powerful magnetic field and a relatively high value for the velocity v of the electrons. In the case of a metal, the concentration of charge carriers is so great that only an excessive current would represent an appreciable value for the velocity v. Inserting quite optimistic values in equation (14.31), viz $B = 1$ T, $v = 10^{-4}$ m/s, gives a p.d. of 10^{-6} V if the sheet is 1 cm wide.

In the case of a semiconductor, on the other hand, the concentration of charge carriers is very much smaller, and the speed of the charge carriers can be considerably larger than it is in a metal, without the current being high. Consequently, Hall potential differences of appreciable size can be obtained in semiconductors.

For use as a magnetometer, an n- or p-type semiconductor is chosen for which the mobility of the charge carriers is high. The principle of its use is represented diagrammatically in Fig. 14.17.

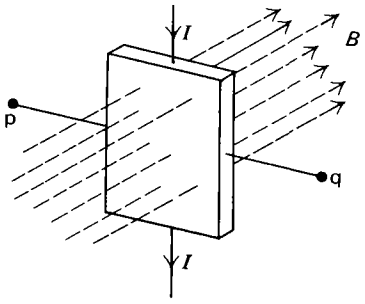


Fig. 14.17 The principle of a Hall effect magnetometer

The specimen, a thin wafer of n- or p-type crystal, is held with its broad face perpendicular to the flux to be measured, and a current I is sent through it. The Hall potential difference V appears across the terminals pq. Reference to equation 14.31 shows that, for a given value of the velocity v of the charge carriers i.e. for a fixed value of the current I, the Hall potential difference V is proportional to the flux density B. This is a sensitive device, with small physical size, and is available for use, also, with high frequency alternating fields.

A further application is in multiplication by analogue. The Hall potential difference is proportional to the current I and the flux density B. If this flux density is produced by a current I' in a coil in such a way that it is proportional to this current, the Hall potential difference becomes a measure of the product of the two currents:

$$V \propto B \times I \propto I' \times I.$$

*14.16 A distinction between p- and n-type semiconductors

An interesting aspect of the Hall effect is that its polarity is reversed, if we choose a p-type in place of an n-type semiconductor. This enables the two to be distinguished, and is a justification, if one is needed, for treating the positive hole as a charge carrier.

The polarity of the effect in an n-type crystal is illustrated in Fig. 14.16, since this refers to electrons. The reverse polarity in the case of the p-type crystal can be understood by comparing Fig. 14.18.

*14.17 The concentration of charge carriers

The concentration N of holes or electrons in the extrinsic crystal may also be deduced from measurement of the Hall p.d.

We have

$$V = Bbv, \tag{14.32}$$

where

$$I = NAev,$$

A being the cross-sectional area of the crystal wafer, perpendicular to the current flow, and e being the electronic charge.

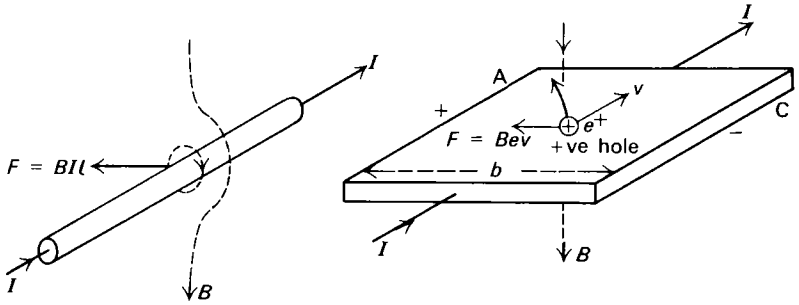


Fig. 14.18 The Hall effect in a p-type semiconductor (a) The motor force on a metal wire (b) The Hall effect in a p-type crystal

Hence,
$$V = \frac{BIb}{NeA}. \tag{14.33}$$

All quantities in this equation other than N can be directly measured, and the value of N is obtained. In fact, the factor $R_H = 1/Ne$ is the only part of this expression for V which is a property of the material from which it is made, the other factors being controllable variables. We can write

$$\frac{V}{b} = R_H \cdot B \cdot \frac{I}{A}, \tag{14.34}$$

where V/b is the potential gradient produced in the wafer when the current density is I/A and the flux density is B. The constant R_H is called the Hall coefficient of the crystal.

These equations do not apply in the case of an intrinsic semiconductor, because the holes and electrons in it have Hall effects with opposing polarity, and the resultant is likely, in any case, to be small.

14.18 Justifying the equations

The equations in this chapter have been derived by taking as the starting point equation 14.3

$$F = BIl. \tag{14.3}$$

They might, equally well, have been derived from equation (14.10)

$$F = BQv. \tag{14.10}$$

We ought, therefore, to discuss what justification there was for writing these equations in the first place. In so far as their basis is the evidence of experiment, this is the consistency with which relations derived from them measure up to practical test, rather than the results of certain specially designed experiments. We can quote a few examples of this.

The design of the moving coil milliammeter is based on the assumption that the deflecting torque which acts on a coil placed in a uniform field is proportional to the current flowing in the coil; on this depends the linear scale of the meter, and

the assumption is upheld by the successful performance of the meter. If all other factors are kept unchanged, the current sensitivity of the meter is directly proportional to the number of turns wound on the coil, that is to say that it is proportional to the length l of conductor which is placed in the field. Such results as this help to establish the relation

$$F \propto Il.$$

There is also some logical basis for it. If two conductors, each of length l, and each carrying current I, are placed alongside, to be perpendicular to magnetic flux, the resultant force, being the sum of the two separate forces F acting in the same direction on the two conductors, is equal to $2F$. But these two conductors must be equivalent, in the charge carried by them, to a single conductor with length l and carrying current $2I$,

i.e. $$F \propto I.$$

The practical evidence for equation 14.10 is the behaviour of charged particles such as α-particles when subjected to a magnetic field in a cloud chamber, or the path taken by ions in the magnetic field of a mass spectrograph. In particular, the path of the charged particle in a uniform field is circular, with radius proportional to the particle velocity.

There is no practical evidence that the force is proportional to flux density B; nor is such evidence needed, because equations 14.3 and 14.10 are alternative definitions of B.

The laws of electromagnetic induction can also be partially tested by experiment, by showing, for example, that the e.m.f. of a dynamo is proportional to its speed of rotation, and Faraday's disc was chosen in section 7.16 as a simple example of this. This type of evidence may be taken to justify the relation

$$V = \frac{d\Phi}{dt}$$

in as much as this concerns motion of a conductor across flux. But our acceptance of the laws of electromagnetic induction is also based on theoretical derivation, such as that presented in sections 14.8 and 14.9.

QUESTIONS

For the electron: $e = -1.6 \times 10^{-19}$ C

$m_e = 9.1 \times 10^{-31}$ kg

$\frac{e}{m_e} = 1.76 \times 10^{11}$ C/kg

14.1 In the space between the two poles of a large sized horseshoe magnet, the magnetic field is uniform, and the flux density is

$$B = 3.0 \times 10^{-4} \text{ T}.$$

Fill in the blank space in each of the following statements.

(*a*) A rectangular shaped coil, with 50 turns, carrying a current of 2.0 A, is so placed that one side of the coil lies in the field, making an angle of 30° with the flux. The length of the side is 5.0 cm. The force which acts on it is ....., in a direction which makes an angle ..... with the flux.

(*b*) A small cathode ray tube is so placed that the cathode ray stream, if undeflected, would be perpendicular to the flux. If the velocity of each electron is 1.7×10^7 m/s, the force which acts on it is . . . . . , perpendicular to its motion, and this curves its path into a circle, of radius $r =$ . . . . . .

(*c*) A small coil is pivoted to turn about a central axis, coplanar with its turns, and it is placed in the field, with this axis perpendicular to the flux. The coil has 60 turns, each carrying current 0.5 A, and enclosing an area 4 cm^2. When the plane of the coil makes an angle of 30° with the flux, the torque which acts on the coil is $\Gamma =$ . . . . . .

(*d*) A metal rod, length 20 cm, is held perpendicular to the flux and moved, at a velocity of 30 cm/s, in a direction which is perpendicular both to its own length and to the magnetic flux. The e.m.f. induced in the rod is . . . . . .

(*e*) A metal disc, radius 5.0 cm, is mounted on a central axle which is perpendicular to its plane, and placed in the field so that this axle is parallel to the lines of magnetic flux. The total flux which crosses the disc = ... The disc is now turned about its axle at 8 rev/s. The p.d. between centre and circumference of the disc, due to electromagnetic induction, is $V =$ . . . . . .

(*f*) A coil with 30 turns, each enclosing an area of 12 cm^2, is placed in the field so that the planes containing the turns are perpendicular to the flux, and the ends of the coil are joined to a ballistic galvanometer, the total resistance of the circuit being 50 Ω. The coil is now pulled out of the field. The total charge Q which flows through the galvanometer in consequence is $Q =$ . . . . . .

(*g*) A small magnet is allowed to oscillate in the field, about a central axis which is perpendicular both to the magnet and to the flux. When this magnet oscillates in a field of flux density 1.0×10^{-4} T, the time of 20 vibrations is 86.5 s. The time t of 20 vibrations is now $t =$ . . . . . .

14.2 An electron is moving with a velocity of 10^7 m/s in a plane which is perpendicular to a uniform magnetic field of flux density B. If the electron is curved into a path of radius 5 cm by the effect of the field, deduce the value of B. What would be the radius of the path in the case of an α-particle moving with the same velocity, if the α-particle has mass approximately 7500 m_e, and charge $+2e$, where m_e, e refer to the electron?

14.3 When low energy electrons are moving at right-angles to a uniform magnetic field of flux density 10^{-3} T, they describe circular orbits 2.82×10^7 times per second. Deduce a value for the charge/mass ratio e/m_e. (O & C)

14.4 A charged particle moves in a circle under the influence of a magnetic field of flux density 0.6 T. Show that the frequency of revolution is independent of the velocity of the particle, and find the value of this frequency for a proton. (e/m for a proton $= 9.65 \times 10^7$ C/kg.)

Describe briefly any way in which this constant frequency property has found practical application. (See 'the cyclotron', chapter 26.) (OXF. SCH.)

14.5 A horizontal straight wire 5.0 cm long weighing 1.20 g/m is placed perpendicular to a uniform horizontal magnetic field of flux density 0.60 T. If the resistance of the wire is 3.80 Ω/m calculate the p.d. that has to be applied across the ends of the wire if it is to be made just self supporting. (C)

14.6 An overhead power cable with 50 m spans running due E–W (magnetic) carries a direct current. What p.d. would have to be applied across each span in order to render the cable self supporting in the vertical plane, and in which direction?

(Density of the material of the cable $= 9.0 \times 10^3$ kg/m^3, resistivity of the material $= 1.8 \times 10^{-8}$ Ω·m, horizontal component of the flux density of the earth's magnetic field $= 1.8 \times 10^{-5}$ T.) (O & C)

14.7 An electric dipole consists of two equal and opposite electric charges $+Q$ and $-Q$ held a fixed distance d apart. The dipole is moving in a vacuum with velocity v perpendicular to a uniform magnetic field of flux density B. What force and/or couple is experienced by the dipole if its axis is pointing

(*a*) along its direction of motion, (*b*) along the direction of the magnetic field, (*c*) perpendicular to both the field and the direction of the motion?

Is there a stable orientation for the whole axis, and if so in what direction? (O & C)

14.8 A rectangular coil 1.0 cm by 0.5 cm, having 500 turns, is free to rotate about a central axis which is parallel to its longer sides. It is situated in a uniform magnetic flux of density 0.20 T which is normal to the axis of rotation. Calculate the torque which acts on the coil when a current of 5 mA is flowing in it, the plane of the coil making an angle of 60° with the direction of the magnetic flux.

14.9 A closed circular loop of wire, radius 5 cm, is placed with its plane at right angles to a uniform magnetic field in which the flux density is changing at the rate of 10^{-2} T/s. Calculate (*a*) the e.m.f. induced in the loop, (*b*) the energy dissipated in the loop in 10 s, if its resistance is 2.0 Ω.

Explain why there is a tension in the loop while the magnetic field is changing. How does it vary with time? (O & C)

14.10 (*a*) An aircraft with wing span 30 m flies horizontally at a speed of 10^3 km/s. What is the p.d. between the wing tips if the vertical component of the earth's magnetic flux density is 4.0×10^{-5} T?

(*b*) Find the p.d. between two ends of a horizontal wire 1 m long pointing E–W, 2 s after it has dropped from rest. The horizontal component of the earth's magnetic flux density is 2.0×10^{-5} T.

(*c*) An electric fan, with blades sweeping out a circle of radius 15 cm, rotates at 800 rev/min with its axis horizontal and parallel to the magnetic meridian. What is the p.d. between axis and rim if the earth's horizontal magnetic component is 2.0×10^{-5} T?

14.11 A horizontal metal disc of radius 10.0 cm is rotated about a central vertical axis in a region where the value of the earth's magnetic flux density is 5.3×10^{-5} T, and the angle of dip is 70°. A sensitive galvanometer of resistance 150 Ω is connected between the centre of the disc and a brush pressing on the rim. Assuming that the resistance of the disc is negligible, what will be the current through the galvanometer when the disc is rotated at 1500 rev/min? If the system is frictionless, calculate the power required to maintain the motion. (C)

14.12 A straight metal rod 50 cm long can slide with negligible friction on parallel conducting rails. It moves at right angles to a magnetic field of flux density 0.72 Wb/m^2. The rails are joined to a battery of e.m.f. 3.0 V and a fixed resistor of value 1.6 Ω; other resistances may be neglected.

Find (*a*) the force required to hold the rod at rest, (*b*) the maximum speed the rod attains when released. (C)

14.13 A circular coil of negligible resistance, with 100 turns of mean area 2.0 cm^2, rotates at 50 rev/s about a horizontal diameter in the vertical flux between the poles of a magnet. The coil is suitably connected through a split ring commutator and brushes to a moving coil meter of resistance 80 Ω. The brushes are set so that the current flows in one direction only through the meter. If the meter gives a steady reading of 5.0 mA, calculate the charge which passes around the circuit for one half revolution of the coil. Hence calculate the flux density of the field.

How would you use the device to find the direction of the flux at a point in a magnetic field? (O & C)

14.14 A penny farthing bicycle with conducting tyres is ridden along a wet road, moving E–W (magnetic). Discuss what currents flow, if any. (OXF. SCH.)

Give an expression for the induced e.m.f. in the circuit, if the wheels have radius r_1, r_2, the bicycle is travelling with velocity v, and the earth's horizontal component is B_{hor}.

14.15 A flat coil of 150 turns enclosing area 300 cm^2 and of total resistance 50 Ω is connected to a galvanometer of resistance 40 Ω. Starting with its plane horizontal, the coil is rotated through a half turn about a central axis in its

plane which points along the earth's magnetic meridian. If the quantity of electric charge which flows through the meter is 4.0 μC, deduce the vertical component of the earth's magnetic flux density.

14.16 A small search coil consists of 8 turns of wire enclosing a mean area 1.5 cm^2. The coil is placed between the poles of a horseshoe magnet, with its plane perpendicular to the flux, and it is connected to a ballistic galvanometer, the total resistance of the circuit being 1000 Ω. When the coil is suddenly removed from the flux, the throw of the galvanometer is 23 scale divisions. When a 1 μF capacitor, charged to a p.d. of 1 V, is discharged through the galvanometer, the throw is 25 divisions. Calculate the flux density between the poles of the magnet.

14.17 A metal framed casement window of length 1.20 m and breadth 0.50 m, lies in the east wall of a house. Regarding the window frame as a conducting circuit of resistance $2 \times 10^{-3}\ \Omega$, calculate the quantity of electricity which will circulate in the frame when it swings open through 90°.
(The horizontal component of the earth's flux density is 2.0×10^{-5} T.)

14.18 A small cathode ray tube is placed in a uniform magnetic field so that the electrons move in a plane which is perpendicular to the magnetic flux, and emerge from the gun travelling perpendicular to the flux. The effect of the magnetic flux is to curve the path of the electrons into the arc of a circle. A uniform electric field E is now applied extending over exactly the same distance as the magnetic flux density B, and in such a direction that the forces exerted upon the electrons by the two fields are in opposition. The relative strengths of the two fields are now adjusted to cancel each other out, with the result that the electrons resume their straight path.

When the magnetic flux density B is 5.0×10^{-5} T, the required value for the electric field strength E is found to be 8.5×10^3 V/m. Deduce (*a*) the velocity of the electrons, (*b*) the p.d. V through which the electrons have been accelerated by the cathode ray gun, (*c*) the radius of their path in the magnetic field when the electric field is removed, (*d*) the direction of E relative to that of B.

14.19 A short bar magnet is suspended at the centre of a circular coil so that it is free to oscillate in a horizontal plane. The axis of the coil, perpendicular to its plane, is horizontal and in the earth's magnetic meridian. The period of small angular oscillations of the magnet is 146 s when no current flows in the coil. When the current is gradually increased from zero the period of oscillation is found to decrease steadily until, with a certain current, the period is 67 s. What will be the period of oscillation if the current is now reversed? If the horizontal flux density of the earth's field is 2×10^{-5} T, what is the flux density due to this current?

14.20 An X-ray ejects an electron with a velocity 10^6 m/s from an isolated atom in a vacuum. Find how long it will take for the electron to return to its starting place if it is subjected to (*a*) a uniform magnetic field of flux density 10^{-4} T in a direction perpendicular to its motion, or (*b*) a uniform electric field of 10^2 V/m in a direction parallel to its motion. The electric field causes the electron to come to rest and then reverse its direction of motion. How far does the electron travel in each case before arriving back at its starting point? (OXF. SCH.)

14.21 A closed tube of square cross-section measures 1 cm × 1 cm × 10 cm. It is filled with mercury and placed in a magnetic field of 1 T transverse to the axis of the tube. A current of 1000 A is passed between opposite sides of the tube in the direction at right angles to the field. What is the pressure difference between the ends?

If such a device were used as a pump, what factors would limit the maximum speed of flow? (CAM. SCH.)
(Hint to last part: where does the energy come from? This is a form of electric motor.)

*14.22 A horizontal trough, 10 cm square vertical cross-section and length l, is

filled with mercury and placed in a uniform horizontal magnetic field of flux density 0.1 Wb/m^2 perpendicular to itself. A small rectangular block of glass floats in the mercury with its lower surface 0.1 cm below the mercury surface. If a current of 1000 A is passed through the mercury, along the horizontal length l of the trough, calculate the change in the height of the glass block which is immersed.
(Density of mercury $= 13.6 \times 10^3$ kg/m^3, $g = 10$ m/s^2.) (CAM. SCH.)

*14.23 A magnetic field of flux density 1 T is directed vertically downwards through a circular loop of wire, radius 50 cm, which lies in a horizontal plane and carries a current of 20 A. The wire has cross-sectional diameter 1 mm and Young's modulus 9×10^{10} N/m^2. What is the change in diameter of the loop when the current is switched off? (CAM. SCH.)

*14.24 A moving coil loudspeaker has a cylindrical coil of diameter 2.0 cm with 40 turns, mounted in a radial magnetic field of flux density 1.2 Wb/m^2. If the effective mass of the movement is 4.0 g, what is the amplitude of movement when a sinusoidal current of 0.10 A r.m.s. at 100 Hz flows through the coil? Neglect elastic forces due to the suspension. (C)

*14.25 A horizontal copper disc, radius 10 cm and 1 cm thick, is free to rotate about a vertical axis through its centre, and is provided with frictionless contacts on the axle and the rim, by means of which a steady current of 20 A is passed radially through it. With the disc at rest, a vertical magnetic flux of density 10^{-2} T is suddenly applied.

How long does it take the disc to rotate once? How is the energy provided for the process?
(Density of copper $= 9 \times 10^3$ kg/m^3.) (CAM. SCH.)

*14.26 Explain why, if a magnetic field acts at right angles to the current flow in a conductor, a p.d. is established across it at right angles to the direction of the field and the current. Show that the effect can be used to determine the sign and number of charge carriers for each metal atom in the conductor.

Calculate the number of charge carriers for each gold atom, if, for a gold strip of thickness 0.10 mm (in the direction of B) a p.d. of 20 μV is measured for a current of 20 A with a magnetic flux density B of 1 Wb/m^2. (OXF. SCH.)
(Relative atomic mass of gold = 197, density of gold $= 19.3 \times 10^3$ kg/m^3, Avogadro constant $= 6.02 \times 10^{23}$ mol^{-1}, $e = -1.6 \times 10^{-19}$ C.)

*14.27 In metal X of relative atomic mass 64 and density 8.0×10^3 kg/m^3, an electric current is a flow of free electrons, which are displaced through the metal with an average drift velocity of 4.0×10^{-5} m/s when there is a potential gradient of 10^{-1} V/m. Assuming one free electron per atom, taking the Avogadro constant to be $N_A = 6.0 \times 10^{23}$ mol^{-1}, calculate the resistivity of the metal X.

ABCD is a rectangular strip of X, of thickness 10^{-2} mm. A current flows from AB to CD, parallel to the sides BC and AD. Show that if a magnetic field is applied at right angles to the plane face ABCD, a p.d. would be set up between E and F, the mid-points of BC and AD respectively.

Assuming that in equilibrium the effect of the potential gradient between E and F on the electron flow balances that of the magnetic field, calculate the magnitude of the p.d. between E and F for a current of 5.0 A between AB and CD and a magnetic flux density of 0.50 Wb/m^2. Indicate the polarity of the p.d. on a diagram.

If ABCD were a rectangular trough containing a solution of an electrolyte such as copper sulphate, would you expect to find a p.d. between electrodes at E and F, with a current flowing from AB to CD and the magnetic field applied? Justify your answer. (O & C)

*14.28 A well collimated beam of mono-energetic α-particles travels a distance of 5.0 cm between two metal plates 0.50 cm apart in an evacuated vessel. When a p.d. of 10^4 V is applied between the plates, the emergent beam is found to be deflected by 2° 12′. When a magnetic field of 0.18 Wb/m^2 is also applied over the region between the plates, in a direction perpendicular to both the

normal to the plates and the original direction of the beam, the beam is restored to its original direction. Determine the charge/mass ratio for the α-particles and their velocity. (CAM. SCH.)

*14.29 An electron is injected with a velocity of 1.5×10^6 m/s into a uniform field of flux density 1.0×10^{-3} T in vacuo. The angle between the field and the initial direction of the electron is 10°. By considering the components of the velocity of the electron parallel to and perpendicular to the field, calculate the axial distance between successive turns of the helical path which the electron will follow. (C)

*14.30 Cathode rays are accelerated in a cathode ray tube by a p.d. V applied between the cathode and final anode. The tube is placed inside a long solenoid so that the cathode rays emitted are parallel to the magnetic flux, density B. An electrostatic field is now applied by small Y plates, giving the rays a component of velocity at right angles to the magnetic flux, and the flux density, which provides the necessary centripetal force on the electrons, is adjusted so that the electrons, following a helical path, make one complete revolution in the distance d between the Y plates and the screen. Show that

$$\frac{e}{m} = k\frac{V}{B^2d^2},$$

where k is a constant, and deduce the value of k. (C)

*14.31 A beam of electrons, all having the same speed, is emitted from a point source within a cone of small angle whose axis coincides with the direction of a uniform magnetic field. Describe the trajectories of the electrons, and show that the magnetic field focuses the beam on the axis of the cone. What will be the distance between the source and the first focus? (OXF. SCH.)

*14.32

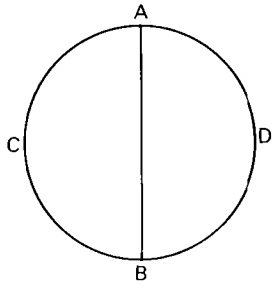


Fig. e 74

The circle of wire has radius 10 cm. The semicircle of wire ACB has resistance 1 Ω and the semicircle ADB has resistance 2 Ω. The straight wire AB has resistance 3 Ω.

What currents flow in the various parts of the circuit if a uniform magnetic field perpendicular to the plane of the circle is changed at a steady rate of 1 T/s? (CAM. SCH.)

*14.33 A pendulum consists of a uniform metal rod with a mass of 10^{-1} kg and a length 10^{-1} m. It is freely hinged at the top, and it hangs with a thin wire attached to the end dipping into a pool of mercury just below the end of the rod. The rod oscillates in a vertical plane with a maximum displacement at the end of 1 cm from the mean position. What is the maximum p.d. between the leads connected to the upper end of the rod and the mercury if a magnetic field of flux density 0.05 T exists perpendicular to the plane of oscillation?

Calculate the final static displacement of the rod if the leads are connected to a current source of 10 A when the rod is at rest.
(Moment of inertia of a rod of length l and mass m about one end is $\frac{1}{3}ml^2$.)
(O & C)

*14.34 A piece of aluminium tube 2 cm long with walls 0.15 cm thick fits closely although freely over a vertical iron cylinder 10 cm in diameter. A magnetic field is maintained normal to the surface of the cylinder, so that at all points close to the surface the flux density is 2.0×10^{-1} T.

What is the limiting velocity with which the aluminium ring will fall under its own weight?
(Density of aluminium $= 2.7 \times 10^3$ kg/m^3, resistivity of aluminium $= 3.2 \times 10^{-8}\ \Omega \cdot$m.)
(Hint: obtain a final expression for v before inserting numerical values – a very good idea in most problems.)

*14.35 A magnetic coil for damping the motion of an instrument consists of 6 turns of copper wire forming a closed circuit. The coil surrounds a cylindrical magnetic pole piece which has a radial magnetic field in the region in which the coil moves. The coil moves parallel to the axis of the cylinder. Each turn of the coil has length 8.0 cm and the wire is 6.0×10^{-3} cm^2 in cross-section.

Calculate the flux density required from the magnet in order that the coil may exert a restraining force of 0.5 N for every m/s of movement. (Resistivity of copper $= 1.75 \times 10^{-8}\ \Omega \cdot$m.)

*14.36

A thin copper ring moves freely at a velocity of 20 cm/s along a horizontal axis OX which is perpendicular to the plane containing the ring and which

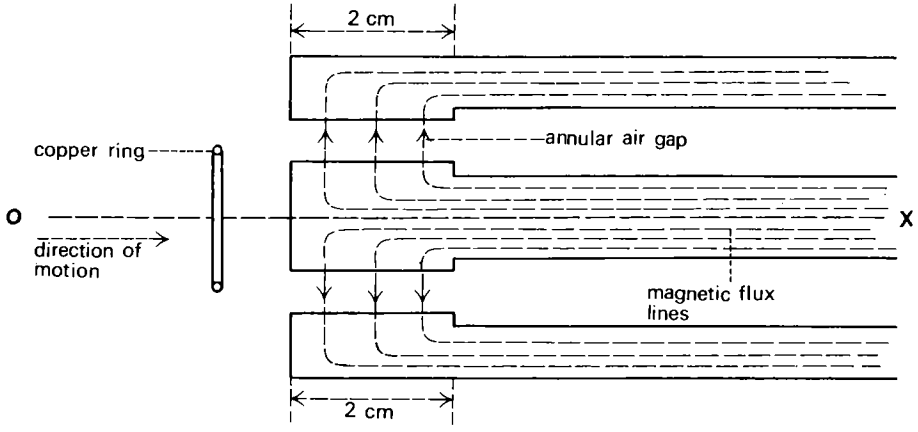


Fig. e 75

passes through its centre. The ring enters the annular air gap of a pole piece system, coaxial with OX, which produces a radial magnetic field extending uniformly over a distance 2 cm in the OX direction. The magnetic field, which is sensibly uniform over the cross-section of the ring, has a flux density of 0.02 Wb/m^2.

Considering only the force produced on the ring by the action of the given magnetic field, calculate the velocity with which the ring leaves the gap.

The following data relate to the copper ring:

mean circumference = 16 cm,

cross-sectional area = 0.01 cm^2,

mass = 1.59 g,

resistivity = $1.6 \times 10^{-8}\ \Omega \cdot \mathrm{m}$. (CAM. SCH. ENG.)

*14.37 A nichrome disc of 10 cm radius and 0.2 cm thick is mounted on an axle perpendicular to its plane, of 1 cm radius, and rotated at 3000 rev/min in a uniform magnetic field of 0.2 T parallel to the axle. Brushes connected in parallel, make contact with many points on the rim of the disc, and a similar set of brushes touches the axle.

Calculate the maximum power available in an external circuit connected to the two sets of brushes.

(Resistivity of nichrome = $1.1 \times 10^{-6}\ \Omega \cdot \mathrm{m}$,) (CAM. SCH.)

$\log_e 10 = 2.3$

*14.38 A large copper disc (5 cm diameter, 1 cm thick) is rotating freely at 1 rev/s in a magnetic field of flux density 0.5 T normal to its plane. A thick wire is connected to a slip ring on its axis and another set of brushes distributed round the periphery of the disc. The ends of these two wires are suddenly connected across a resistance of 10 Ω. Calculate (*a*) the initial current in the resistance, (*b*) the time taken for the current to fall to half its initial value, (*c*) the total amount of electrical energy dissipated in the resistance before the disc comes to rest.

(Assume that the flow of current in the disc is radial and that the electrical resistance of the disc is negligible.) (CAM. SCH.)

(The moment of inertia J of a disc about its central axis perpendicular to its plane is $\frac{1}{2}mr^2$; $\log_e 2 = 0.70$; density of copper = $9.0 \times 10^3\ \mathrm{kg/m^3}$.)

*14.39 A galvanometer coil consisting of N turns of mean area A is oscillating freely in a radial magnetic field of flux density B. The galvanometer coil has a resistance R_1, and that of the external circuit is R_2. Calculate the electromagnetic damping couple exerted on the coil at the instant when the angular velocity of the coil is ω. (CAM. SCH.)

*14.40 A circular loop of wire of radius r and resistance R is set spinning about a diameter perpendicular to a steady magnetic flux density B with an angular velocity ω. Calculate the average rate at which heat is generated in the loop. (Hint: an a.c. problem – find r.m.s. V and I.)

Given that $r = 5$ cm, $R = 2.5\ \Omega$, $B = 2.0 \times 10^{-2}$ T, and the moment of inertia of the loop about the axis of rotation is $2.5 \times 10^{-6}\ \mathrm{kg \cdot m^2}$, find the approximate time it takes for the angular velocity to halve from its initial value of 20π rad/s. ($\mathrm{Log}_e\ 2 = 0.70$) (OXF. SCH.)

*14.41 A circular loop of copper wire is rotating slowly in free space, the axis of rotation being in the plane of the loop. There is a flux density B perpendicular to the axis of rotation, so that eddy currents induced in the loop dissipate energy and slow the rate of rotation.

Show that the angular velocity decays exponentially according to the relation

$$\omega = \omega_0 \exp\left(-\frac{t}{t_0}\right),$$

where t_0 is independent of the dimensions of the loop.

Calculate t_0 from the following values: density of copper = $8.9 \times 10^3\ \mathrm{kg/m^3}$, resistivity of copper = $1.7 \times 10^{-8}\ \Omega \cdot \mathrm{m}$, $B = 4 \times 10^{-5}$ T, moment of inertia of a ring about a diameter is $\frac{1}{2}mr^2$. (OXF. SCH.)

(Method as for number 14.40.)

*14.42 A permanent horseshoe magnet produces a uniform field of flux density 1 $\mathrm{Wb/m^2}$ between its pole faces. A plane circular coil of 100 turns and mean diameter 2 cm, having a resistance of 1 Ω, is placed in the field and rotated about an axis perpendicular to the field direction. If the coil is rotated at 100 rev/s, what p.d. is generated across the coil? If the ends of the coil are joined with a 10 Ω resistance, what is the mean power dissipated in the resistance?

You may ignore the effect of the inductance of the coil. Describe qualita-

tively and carefully the source of the energy dissipated in the 10 Ω resistance. (OXF. SCH.)

*14.43 A sensitive galvanometer is connected in series with a high resistance, a source of e.m.f. E, and a switch. When the switch is closed for a short, accurately measured time t, the galvanometer shows a momentary deflection, and the highest scale reading θ attained by the pointer is recorded. What conclusion can be drawn from the following data?

E/V	0.20	0.20	0.20	0.30	0.30	0.40	0.40
t/s	0.10	0.15	0.20	0.10	0.15	0.10	0.20
θ/divs	12.0	17.9	24.2	18.1	26.8	24.0	47.9

The calibrated galvanometer and resistance are connected to a coil of 500 turns of mean area 2 cm^2 and negligible resistance, placed between the poles of a magnet. When the coil is rapidly removed, the galvanometer shows a momentary deflection of 14.9 divisions. Calculate the flux density between the poles of the magnet, explaining your reasoning.

How would you use the calibrated galvanometer to determine the time of impact when a metal ball bounces against a fixed metal plate, the duration of the impact being thought to be of the order of 10^{-3} s?

What further measurements and calculations would be needed to determine the average force of the blow? (O & C)

Chapter 15

Magnetizing coils and magnets

15.1 Rationalized flux density equations

The last chapter dealt with the properties of a magnetic field, whereby it may be detected and measured. In this chapter the origin of the field will be considered instead, and equations which enable the flux density to be calculated, once there is sufficient information available about the coil or magnet which produces it. A typical equation of this kind is the one which was quoted in the last chapter (equation 14.4) which gives the value of the flux density at a distance r from a long straight conductor, viz.

$$B = \beta \frac{2I}{r}. \tag{15.1}$$

Another gives the flux density at the middle of a long straight solenoid:

$$B = \beta\, 4\pi \frac{N}{l} I, \tag{15.2}$$

where N is the number of turns which are wound on length l of the solenoid.

It is a feature of these and all similar equations that they include the factor β, and that whereas some resemble equation 15.1, others, like equation 15.2, have the constant in the form $4\pi\beta$. When all these are collected together, including examples which will not appear in this book, some advantage is seen to be gained in dispensing with β, and substituting a new constant, μ_0, where

$$\mu_0 = 4\pi\beta \tag{15.3}$$

and

$$\beta = \frac{\mu_0}{4\pi}.$$

This practice is universally adopted when writing out equations for use with SI units; consequently it will be adopted in this book, and we will start by rewriting equations 14.1, 15.1, and 15.2 in this form:

The force of attraction between two long parallel conductors which carry current I, I'

$$F = \frac{\mu_0}{4\pi} 2II' \frac{l}{r}. \tag{15.4}$$

The flux density at distance r from a long straight conductor

$$B = \frac{\mu_0}{4\pi} \cdot \frac{2I}{r}. \tag{15.5}$$

The flux density in the middle of a long straight solenoid

$$B = \mu_0 \frac{N}{l} I. \tag{15.6}$$

These are all referred to as rationalized equations, those containing β being unrationalized.

Before discussing μ_0 further, it will be convenient first to quote the more

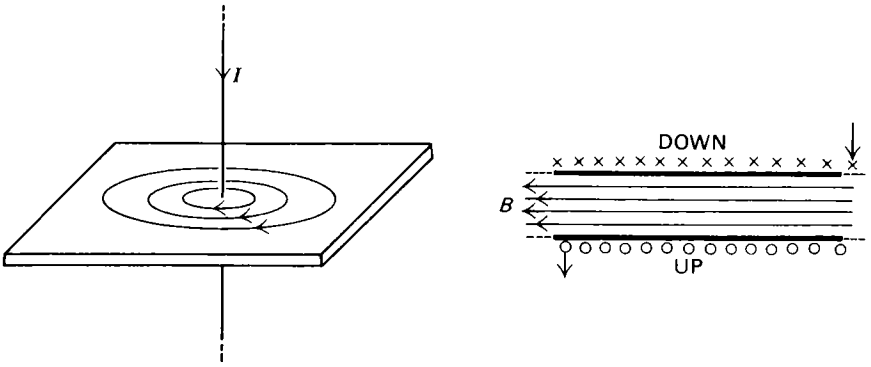


Fig. 15.1 The magnetic flux of a current (a) A long straight conductor (b) A straight solenoid

$$\text{(a) } B = \frac{\mu_0}{4\pi} \cdot \frac{2I}{r}$$

$$\text{(b) } B = \mu_0 \frac{N}{l} I$$

important equations which are similar to 15.5 and 15.6 above. This is because it is even more important to us to have these equations for use than it is to derive them.

The centre of a plane circular coil

$$B = \frac{\mu_0}{4\pi} \cdot \frac{l}{r^2} I, \tag{15.7}$$

where l is the total length of wire used in winding the coil, and r is its radius

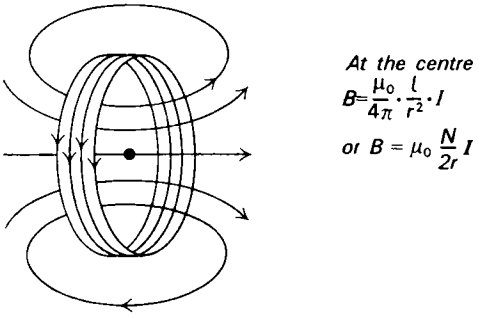


Fig. 15.2 The magnetic flux of a plane circular coil

Alternatively,

$$l = N \times 2\pi r,$$

where N is the number of turns, whence

$$B = \frac{\mu_0}{4\pi} \cdot \frac{N \times 2\pi r}{r^2} I,$$

i.e.
$$B = \mu_0 \frac{N}{2r} I. \tag{15.8}$$

The anchor ring, annulus, or toroid

This is a coil wound round a former which has the shape of a pipe curved round into a ring. (It is sometimes, also, referred to as a doughnut.) The equation for the flux density inside the hollow ring is the same as equation 15.6 for the long straight solenoid. This is consistent, since the radius does not appear in the equation, and the straight solenoid might be regarded as a circular one which has infinite radius.

$$B = \mu_0 \frac{N}{l} I, \tag{15.6}$$

where N is the number of turns which are wound on length l measured round the circumference.

Because the inside circumference of the ring is smaller than the outside, the factor N/l is greater round the inside, and so the flux density in the ring is not uniform; it is greater near the inside surface, and can be taken as constant only in the case when the radius r of the ring is large by comparison with the area A of its cross-section.

This list completes the most important of these equations. Some further ones will be developed presently. It is stressed that they apply only in the absence of magnetic materials, notably ferromagnetic ones.

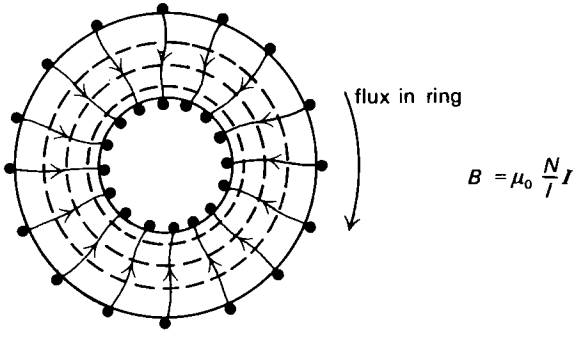


Fig. 15.3 The magnetic flux in a toroidal solenoid

15.2 The permeability of a vacuum μ_0

The constant μ_0 is termed the permeability of a vacuum. We have already stated the value of β, and of course the value of μ_0 comes from this, viz.:
using SI units,

$$\beta = 10^{-7}\ \mathrm{N/A^2},$$
$$\mu_0 = 4\pi \times 10^{-7}\ \mathrm{H/m} \tag{15.9}$$

[The reason that μ_0 is measured in henries per metre will become evident later (section 15.12).]

The convention of using rationalized equations in magnetism follows the same pattern as the similar convention in electrostatics, and the constant μ_0 may be taken to correspond to the electrostatic constant ϵ_0 which is called the permittivity of a vacuum. [In some respects, however, the equivalence is more obviously between μ_0 and $1/\epsilon_0$; an example of this is the frequent appearance in magnetism of the factor $\mu_0/4\pi$, which is β, whereas what appears in electrostatics is $1/4\pi\epsilon_0$,

which is α. What corresponds to what, however, depends rather on the particular circumstances, and the point of view adopted.]

Some justification can be found, if justification is required, for attributing the properties of permeability and permittivity to a vacuum by referring to Maxwell's equation. Electromagnetic waves are propagated by a vacuum, with a particular velocity c, shown by Maxwell to be given by the equation

$$c = \frac{1}{(\mu_0 \epsilon_0)^{\frac{1}{2}}}. \tag{15.10}$$

One important contrast between μ_0 and ϵ_0 must be stressed, and this concerns the basis on which the value of each constant is known.

The value of β, and therefore of μ_0, is defined by defining the ampere, with reference to equation 14.1 or 15.4. It follows that, in SI units, the value of μ_0 is exactly $4\pi \times 10^{-7}$, and there is no purpose whatever in devising, or trying to devise, an experiment to measure it. Nor is there any purpose in concealing the presence of π by putting in an arithmetical value for π. Rationalizing does not alter the arithmetic of an equation in any way at all.

The value of ϵ_0, on the contrary, is known only as the result of an experiment carried out to measure it, i.e. the experiment which was described in section 13.17. Alternatively, it may be deduced from equation 15.10, if we prefer to rely on an accurate measurement of the velocity c of an electromagnetic wave in a vacuum.

As a further comparison with electrostatics, rationalized equations and SI units were not the language with which electromagnetic theory was originally developed. Nor was the theory based, in its initial stages, on the magnetic effect of an electric current, but on the magnetism of the permanent magnet. Consequently, many excellent textbooks, but printed a few years ago, present electromagnetism in a markedly different way, and the treatment refers to c.g.s. units. In view of all these differences in treatment, no further reference will be made to c.g.s. units in this book, except, perhaps, for occasional mention of the gauss as a unit of flux density.

15.3 The Biot Savart equation

The equations which give the flux density in solenoids and coils are based on experimental evidence, and cannot be derived, except perhaps from each other. They are consistent with each other, however, and the experimental evidence is cumulative, as one equation after another is tested. To show their consistency, what is required is a single equation to use as a starting point from which all the others can be derived; then if these survive the experimental test, the basic equation has been tested too. In other words, something is required in electromagnetism to fill the place taken in electrostatics by the equation of the inverse square law,

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2}. \tag{15.11}$$

We can get close to the required equation if we start with the flux density at the centre of a plane circular coil (which can be tested) and work back from that. This is a particularly simple arrangement, because all portions of the wire which carries the current are equidistant from the centre of the circle where the flux density is measured.

We have

$$B = \frac{\mu_0}{4\pi} \cdot \frac{l}{r^2} I \tag{15.7}$$

and we will assume, temporarily, that this has been verified by experiment. Since, in this equation, l is the total length of wire, and all parts of it are equidistant from

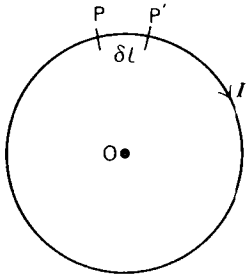


Fig. 15.4 The flux density at the centre of a plane circular coil

the centre O, we will assume that the contribution made by a small length δl, carrying current I (known as a current element) is

$$\delta B = \frac{\mu_0}{4\pi} \cdot \frac{\delta l}{r^2} I. \tag{15.12}$$

This takes us partway towards our goal, because a current element is something which might be integrated to give coils of various shapes. The problem remains, however, that the point O is situated in a special direction relative to the current element, i.e. a direction perpendicular to the element, and a more general case is required. We have to make an assumption in order to proceed any further, to give the more general relation which refers to Fig. 15.5, viz.:

$$\delta B = \frac{\mu_0}{4\pi} \cdot \frac{I\delta l}{r^2} \sin\theta. \tag{15.13}$$

Current must be brought to and taken away from a current element, and if it is considered in isolation the case is not a practical one. In such circumstances equation 15.13 can lead to anomalies, such as contradiction of Newton's third law of

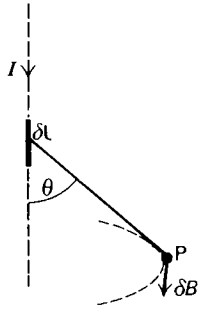


Fig. 15.5 The flux density due to a current element (The direction is given by Maxwell's corkscrew rule)

motion. Consequently, the equation is best taken to apply only when referring to a complete circuit, and quoted as an integral. In this form, it is called the Biot Savart equation.

The Biot Savart equation

$$B = \frac{\mu_0}{4\pi} \oint I \frac{\mathrm{d}l}{r^2} \sin\theta. \tag{15.14}$$

Alternatively, this may usefully be expressed, with reference to Fig. 15.6, in the form

$$B = \frac{\mu_0}{4\pi} \oint \frac{I}{r} \, d\psi, \tag{15.15}$$

where, in the diagram,

$$\delta\psi = \frac{\delta l}{r} \sin \theta.$$

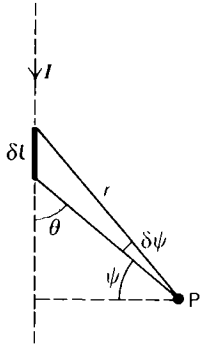


Fig. 15.6

15.4 Ampère's circuital equation

Although in theory a coil of any shape may be built up from current elements to which the Biot Savart relation can be applied, it is not always possible to integrate it, even in some of the simple cases which we require – the anchor ring is an example of this. An alternative starting point can sometimes be used successfully instead, which is accepted as having a validity which is at least equal to that of the Biot Savart relation, and this is Ampère's circuital equation.

Suppose we have a collection of electric circuits, such as those drawn in Fig. 15.7. Referring to this diagram, the dotted line represents any closed loop, taken at

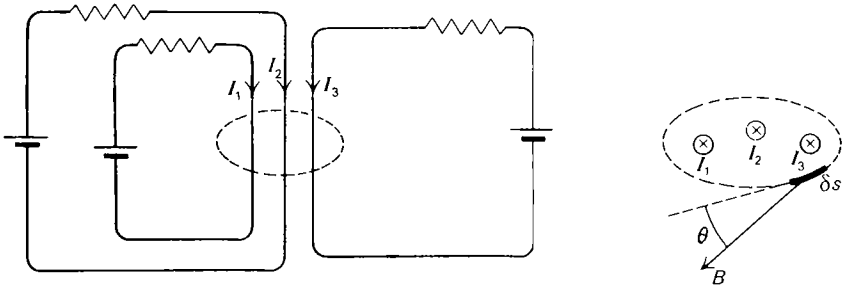


Fig. 15.7 Illustrating Ampère's circuital equation

random to enclose inside it a number of conductors, carrying currents I_1, I_2, I_3 etc. Taking a short length δS of this loop at a point P, suppose that the resultant flux density B at this point makes an angle θ with the direction of δS. Then the component of B which lies along the direction of δS is $B \cos \theta$. Taking the product,

$B \cos \theta \, \delta S$, and summing it over the whole of the closed path, gives the line integral,

$$\oint B \cos \theta \, \delta S$$

Ampère's circuital equation

$$\oint B \cos \theta \, \mathrm{d}S = \mu_0 \Sigma I, \tag{15.16}$$

where ΣI is the sum of all currents carried by conductors which are enclosed by the loop.

It will be noticed that rationalizing has removed the factor 4π from Ampère's equation, and put it into the Biot Savart equation. Ampère's equation for the magnetic field and Gauss's equation for an electrostatic field are very prominent in electromagnetic wave theory, and this is a strong reason for rationalizing, since this procedure removes the factor 4π from both of them.

15.5 Derivation of the flux density equation for a long straight conductor

We now have to derive some of the more useful flux density relations, such as equations 15.5, 15.6, and so on, using either the Biot Savart or Ampère's equation as the starting point, and also to suggest a method of testing the equation in some selected cases.

We will take the straight conductor first, because this can be treated by choosing either of the alternative starting points, and will therefore serve to illustrate that they are consistent in their results.

(a) Starting from Ampère's equation

Accepting as an experimental fact that the magnetic flux forms closed circles round the straight conductor (there would be few other possibilities, due to the symmetry), one such circle, radius r, is taken as the closed path required by Ampère's equation.

By symmetry, the flux density B must have the same value at all points on this circle, and is tangential to it.

Hence
$$\oint B \cos \theta \, \mathrm{d}S = \mu_0 \Sigma I, \tag{15.16}$$

where B is constant and $\cos \theta = 1$.

Therefore
$$B \oint \mathrm{d}S = \mu_0 I,$$

$$B \cdot 2\pi r = \mu_0 I$$

$$B = \frac{\mu_0}{2\pi} \cdot \frac{I}{r},$$

or
$$B = \frac{\mu_0}{4\pi} \cdot \frac{2I}{r}. \tag{15.5}$$

There appears to be nothing in this derivation which requires that the conductor should have infinite length. But the conductor must form part of a closed circuit for current to flow, and this condition cannot be fulfilled, whilst completely retaining the symmetry, unless the conductor is straight over a considerable distance.

(*b*) **Starting from the Biot Savart equation**

The convenient form of this equation to choose as the starting point here is

$$B = \frac{\mu_0}{4\pi} \oint \frac{I}{r} \mathrm{d}\Psi \tag{15.15}$$

which refers to Fig. 15.6. First, however, we will substitute the symbol z for r in the equation to represent the distance from the point P to the element of wire δl. This is because r is reserved to represent the perpendicular distance from P to the straight conductor of which δl is a part. So we now have

$$B = \frac{\mu_0}{4\pi} \oint \frac{I}{z} \mathrm{d}\Psi,$$

where $z = r \sec \Psi$.

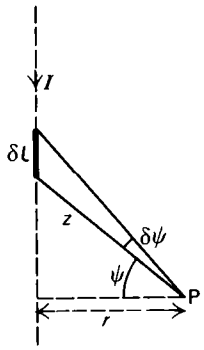


Fig. 15.8

The resultant flux at P is perpendicular to the plane of the diagram, and so it is for every small part δB of the flux density contributed by each small length δl of the conductor. Hence the vector aspect of B introduces no complication into the integration, and we can write

$$B = \frac{\mu_0}{4\pi} \cdot \frac{I}{r} \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \cos \Psi \, \mathrm{d}\Psi$$

$$= \frac{\mu_0}{4\pi} \cdot \frac{2I}{r}, \tag{15.5}$$

because $\displaystyle\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \cos \Psi \, \mathrm{d}\Psi = 2.$

15.6 The flux density at the centre of a plane circular coil

We have already arrived at the Biot Savart equation partly by working back from the case of the plane circular coil, and so there should be no difficulty about taking the derivation in the reverse direction. Thus

$$B = \frac{\mu_0}{4\pi} \oint \frac{I \sin \theta}{r^2} \mathrm{d}l, \tag{15.14}$$

where r is constant, being the radius of the coil, and $\sin \theta = 1$.

$$B = \frac{\mu_0}{4\pi} \cdot \frac{l}{r^2} I, \tag{15.7}$$

where l is the total length of wire which is used to wind the coil,

i.e. $$l = N \times 2\pi r,$$

giving the alternative form for this equation

$$B = \mu_0 \frac{N}{2r} I. \tag{15.8}$$

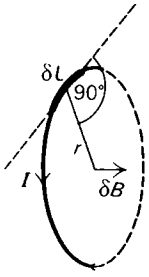


Fig. 15.9

This is an example where to start instead from Ampère's equation is not practicable, because insufficient is known about the flux density at different points in a closed path which encircles the wire, and the integral $\oint B \cos\theta \, dS$ cannot be evaluated.

An experimental test of equation 15.7 is the nearest that we can get to making a direct test of the Biot Savart equation, and so this makes a very suitable arrangement to select as an example of how such an experiment can be carried out. A similar method could be applied in other cases also; e.g. to test the relation for a straight conductor.

We have to test each of the following relations:

$$B \propto I$$
$$B \propto l$$
$$B \propto \frac{1}{r^2}, \quad l \text{ being kept constant,}$$

or, alternatively, choosing equation 15.8:

$$B \propto I$$
$$B \propto N$$
$$B \propto \frac{1}{r},$$

the number of turns N being kept constant.

The apparatus is slightly easier to arrange in the second case, and so we will choose this. All that is required is a succession of concentric plane circular coils,

having different radii, and a vibration magnetometer can be used to measure the flux density B at the centre.

Since the field of the earth may well have sufficient strength to affect the behaviour of the magnetometer, this must be allowed for, and to simplify the

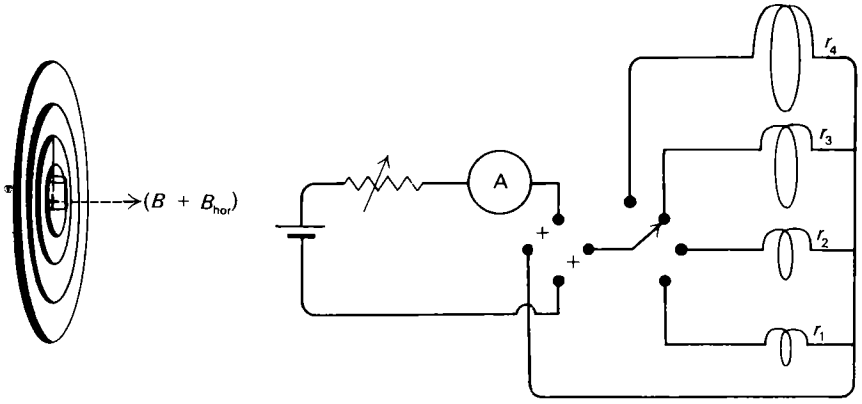


Fig. 15.10 To investigate the flux density B at the centre of a plane circular coil

procedure for doing this, the coils are set up with their planes vertical and perpendicular to the earth's magnetic meridian. The magnetometer is suspended to be at the common centre of the coils, and the direction of the current so chosen that the field of each coil and the earth's field strengthen each other. Using the same number N of turns in each coil in turn, we then proceed to test the relation

$$B \propto \frac{I}{r}, \tag{15.17}$$

where the resultant flux density is $(B + B_{\text{hor}})$.
Thus,

$$\frac{1}{t^2} \propto (B + B_{\text{hor}})$$

where t is the time of a fixed number of vibrations of the magnet. Also, as a special case, when $I = 0$,

$$\frac{1}{t_{\text{hor}}^2} \propto B_{\text{hor}}$$

$$\left(\frac{1}{t^2} - \frac{1}{t_{\text{hor}}^2}\right) \propto B, \tag{15.18}$$

and so we have to see whether

$$\left(\frac{1}{t^2} - \frac{1}{t_{\text{hor}}^2}\right) \propto \frac{I}{r}. \tag{15.19}$$

[To be precise, this procedure provides a valid test of the relation $B \propto I/r$, and could deal in a similar way with the relation $B \propto N$. The effect of varying the current I, however, is more elusive, if current is to represent the amount of charge passing a point in unit time. We could argue, however, that current I in N turns is equivalent to current NI in one turn, and test the effect of changing current by this indirect approach.]

15.7 The flux density in a toroid

The flux inside a toroidal solenoid takes the form which was shown in Fig. 15.3. By symmetry, the flux density must have the same value at all points on one circular line of magnetic flux. This is a very simple case for applying Ampère's equation. Taking a closed circle inside the solenoid, following a line of magnetic flux, we have

$$\oint B \cos\theta \, \mathrm{d}S = \mu_0 \Sigma I,$$

Where $\cos\theta = 1$, B is constant, and $\Sigma I = NI$.

Whence
$$B \oint \mathrm{d}S = \mu_0 NI,$$

$$B = \mu_0 \frac{N}{l} I \tag{15.6}$$

since
$$\oint \mathrm{d}S = l.$$

15.8 The flux density at a point inside a long straight solenoid

A straight solenoid of infinite length may be regarded as being a toroid of infinite radius, to which equation (15.8) applies.

Alternatively, Ampère's equation may be applied by taking a closed path such as PQRS in Fig. 15.11.

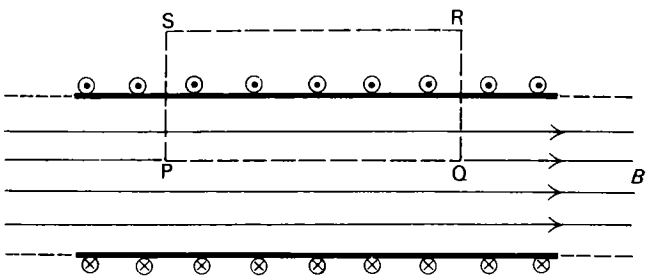


Fig. 15.11

The flux which is directed down the length of the solenoid must eventually, of course, curve round on nearing the end of the solenoid to form a number of closed loops. We have to assume that the distance PQ is far enough from the ends for the length of path RS to be in a region which is devoid of flux, in which case there is uniform flux density B directed along PQ, but no flux at all in the directions of the other three sides of the closed path PQRS, the flux being perpendicular to the sides QR, SP.

Hence
$$\oint B \cos\theta \, \mathrm{d}S = \mu_0 \Sigma I,$$
$$\therefore \quad Bl = \mu_0 NI,$$
where PQ = l and N is the number of turns of the solenoid which are enclosed by the path PQRS, i.e. the number of turns in length l.

Therefore
$$B = \mu_0 \frac{N}{l} I. \tag{15.6}$$

15.9 A straight solenoid of finite length

When, as in practice, the length of the solenoid is restricted, and an appreciable amount of returning flux lies outside the solenoid but inside any closed loop such as PQRS, then equation (15.6) is only approximate, and the exact equation is
$$B = \mu_0 \frac{N}{l} I \cdot \tfrac{1}{2}(\cos\theta_1 + \cos\theta_2), \tag{15.20}$$
where the significance of the angles θ_1, θ_2 appears in Fig. 15.12 (a).

For a point near the middle of the solenoid, the simpler equation (15.6) may be taken as a reasonable approximation (within 1% in accuracy), provided that the length l is greater than about $7d$, where d is the diam r of the circular cross-section.

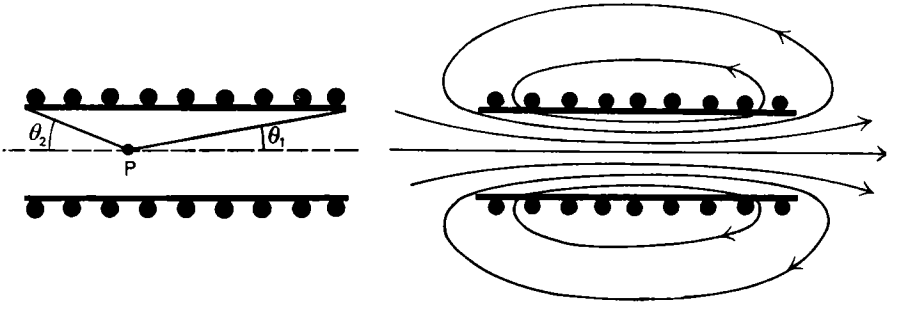


Fig. 15.12 Leakage of flux near the ends of a solenoid (a) (b)

*15.10 The Helmholtz arrangement of coils

The straight solenoid, provided that it is sufficiently long, gives the advantage that the flux density near its centre is uniform, but this is not always the easiest of places at which to site a piece of apparatus. An alternative arrangement, which is more adaptable to practical use, is the one known as a Helmholtz system. This has two plane circular coils, placed coaxially at a distance apart which is equal to the radius of each coil.

The flux due to a Helmholtz system is compared with that of a single coil in Fig. 15.13. In the case of the single coil, the flux density falls off rather rapidly for points on the axis which lie outside the coil, with the result that equation (15.8) accurately applies only at the single point which is the centre of the coil. For the Helmholtz system, on the other hand, the flux density is reasonably constant over a considerable region surrounding the point on the common axis which is midway

between the two coils. Here is a place which is readily accessible, and this arrangement is often useful in circumstances when a straight solenoid is not practicable. The field strength at this point, due to the combined effect of the two coils, is

$$B = \mu_0 \frac{8N}{5\sqrt{(5r)}} I, \tag{15.21}$$

N being the number of turns in each coil.

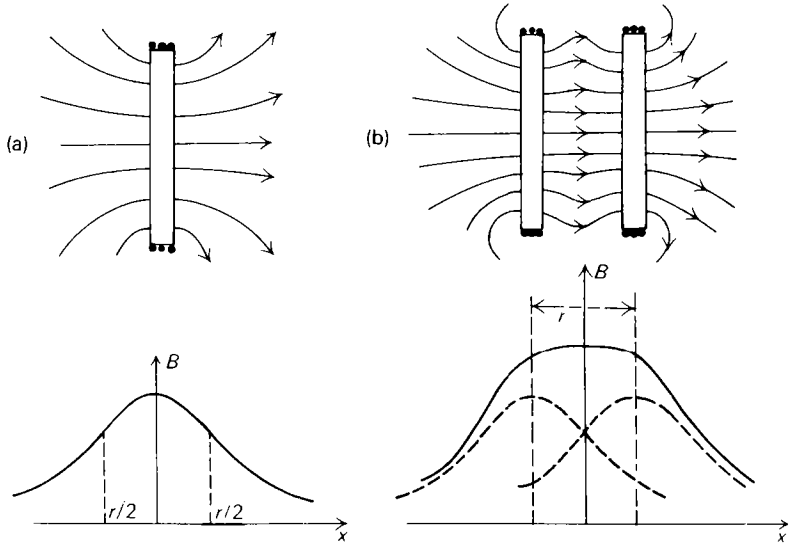


Fig. 15.13 The Helmholtz arrangment of two coils (a) The flux of a single coil (b) The flux of a Helmholtz pair of coils

The reason that the separation of the two coils must be made equal to their radius can be sought in the graphs shown in the diagram, which represent the variation in flux density B with distance along the axis. At a distance $r/2$ from the centre of one coil, its flux density is falling off at a constant rate because the curvature of the graph reverses about this point, being convex above it and concave below. It follows that at the point P midway between the coils, i.e. distance $r/2$ from both of them, the flux density due to one coil is increasing uniformly with distance along the axis at the same rate as the flux density due to the other is decreasing, and that this is so for an appreciable distance. The resultant flux density is consequently constant in the region which surrounds the point P. This assumes, of course, that the two coils are similar, and carry equal currents flowing in the same direction.

*15.11 The flux density at a point on the central axis of a plane circular coil

The last derivation we will give here is to produce the equation of the curve drawn in Fig. 15.13 (*a*), representing the flux density at a point on the central axis of a single circular coil, distance x out from its centre.

A current element δl at any point on the coil, such as S, is at right angles to the line SP. The resultant flux density at P is directed along the axis OP because, by symmetry, the components perpendicular to OP integrate to zero for the complete coil,

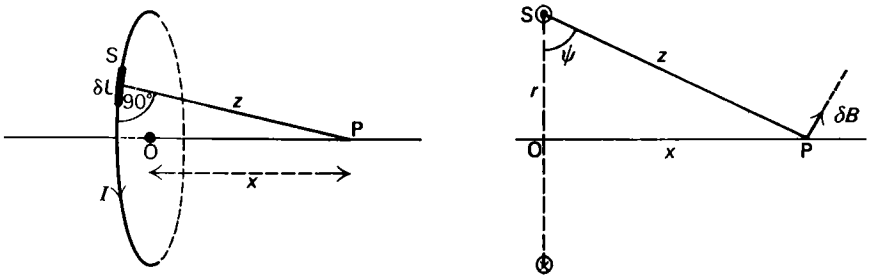


Fig. 15.14

$$\text{i.e.}\quad B = \int \mathrm{d}B \cos \Psi \text{ along the axis,}$$

$$B = \frac{\mu_0}{4\pi} \oint I \frac{\mathrm{d}l}{z^2} \sin\theta \cos \Psi, \text{ from equation 15.14,}$$

where, in this case, $\theta = 90°$, being the angle between the line SP and the current element δl.
Therefore

$$B = \frac{\mu_0}{4\pi} \oint I \frac{\mathrm{d}l}{z^2} \cos \Psi.$$

But z and $\cos \psi$ have constant values for all points such as S on the coil, and so this equation becomes

$$B = \frac{\mu_0}{4\pi} \cdot \frac{I}{z^2} \cos \Psi \oint \mathrm{d}l,$$

where $\oint \mathrm{d}l = 2\pi r N$, N being the number of turns in the coil, and $\cos \Psi = r/z$.

Therefore

$$B = \frac{\mu_0}{4\pi} 2\pi r^2 \frac{NI}{z^3},$$

i.e.

$$B = \mu_0 \frac{r^2 N}{2z^3} I, \tag{15.22}$$

or

$$B = \mu_0 \frac{r^2 NI}{2(r^2 + x^2)^{\frac{3}{2}}}. \tag{15.23}$$

The equation for the Helmholtz coils may be derived from this, and it reduces to equation (15.8) if x is made equal to 0.

15.12 Mutual inductance and self inductance

Mutual inductance M and self inductance L were defined in section 7.18, giving the equations

$$E_2 = M \frac{\mathrm{d}I_1}{\mathrm{d}t}, \tag{15.24}$$

where E_2 is the e.m.f. induced in a secondary coil as the result of the current I_1 changing in a primary coil, and

$$E = -L\frac{dI}{dt}, \tag{15.25}$$

where E is the e.m.f. induced in a coil as the result of the current I changing in the same coil.

The values of M and L may be calculated in simple cases by applying equations taken from this chapter, and the principle is illustrated in the example given below. First, it is convenient to express the definitions of M and L in a different way, by writing the two equations

$$N_2\Phi_1 = MI_1, \tag{15.26}$$

where Φ_1 is the flux which threads the secondary coil due to current I_1 flowing in the primary, i.e. $N_2\Phi_1$ is the total flux linkage due to current I_1.

Similarly,

$$N\Phi = LI, \tag{15.27}$$

where $N\Phi$ is the flux linkage with the windings of a single coil due to current I in the same coil.

It is easy to show that these two equations are consistent with the equations which they replace, for we have

$$E_2 = N_2\frac{d\Phi_1}{dt}$$

$$= M\frac{dI_1}{dt}, \tag{15.24}$$

and similarly in the case of self inductance L.

[The equations are not exactly equivalent if the coils are wound on a ferromagnetic core, because in this case the factors M, L are not constants, but vary with the current, and we have to write, for example,

$$E_2 = N_2\frac{d\Phi_1}{dt}$$

$$= N_2\frac{d}{dt}(MI_1)$$

$$= N_2\frac{MdI_1}{dt} + N_2 I_1\frac{dM}{dt}.\Big]$$

Example

A long straight solenoid is wound uniformly with a coil having N_1/l turns in unit length (the primary coil). A secondary coil having N_2 turns is wound round a small section at the middle of the primary, so that the secondary is linked by all the flux due to the current in the solenoid. If the area of cross-section of the long solenoid is A, derive an expression for the mutual inductance M.

We have

$$\Phi_1 = \mu_0\frac{N_1}{l}I_1A,$$

whence

$$MI_1 = N_2\Phi_1$$

$$= \mu_0\frac{N_1}{l}N_2AI_1,$$

$$\therefore\quad M = \mu_0\frac{N_1}{l}N_2A.$$

This is the type of equation which justifies the name, henry per metre, which was given in section 15.2 to the SI unit of μ_0, for it shows the name to be dimensionally correct.

15.13 Ampère's experiments

Magnetic flux density B was introduced in the last chapter as a means of simplifying the problem of the interaction between currents and between moving charges. The pioneer in this field in the early nineteenth century was Ampère, and he did not adopt this approach, but conducted experiments to investigate directly the effect of current in one circuit on current in another. By considering the interaction between two current elements $I\delta l$ and $I'\delta l'$, rather than that between two complete circuits, the basic relation reveals an inverse square law with respect to distance (this appears in the Biot Savart equation) and one experiment suggested by Ampère which has particular interest was to test this law. He deduced that, if the inverse square law is true, the force which acts between any two coils carrying current must be independent of the scale upon which they are built. That is, suppose that two coils are set up in a particular way, each taking a particular current, and the force between them is measured. Now another pair of coils is taken, making an exact replica of the first arrangement except that it is built on a different scale. If the currents flowing are the same as in the first case, it is a consequence of the inverse square law that the force will prove to be the same in both cases.

One example of the conclusion above can be taken from the definition of the ampere. This definition considers the case when a current of one ampere flows in each of two long conductors placed one metre apart, and specifies the force which would act on one metre of one conductor. The point here is that the lengths could just as well have been specified in some unit other than the metre – the force required would have been just the same. This is illustrated by equation (15.4)

$$F = \frac{\mu_0}{4\pi} 2II' \frac{l}{r}. \tag{15.4}$$

Only the ratio l/r is involved, and not the separate values of the two distances. Another similar example will appear in chapter 17 (section 17.2).

*15.14 The dimensions of μ_0 and ϵ_0

Maxwell's relationship was quoted in section 15.2 giving the velocity c of an electromagnetic wave in terms of the two factors μ_0 and ϵ_0, viz

$$c = \frac{1}{(\mu_0\epsilon_0)^{\frac{1}{2}}}. \tag{15.10}$$

We are rather obviously in no position to attempt to derive this by quoting Maxwell's equations, but it is interesting to trace the connection between μ_0, ϵ_0, and a velocity by considering the dimensions of the two quantities. Thus

$$F = \frac{\mu_0}{4\pi} 2II' \frac{l}{r}, \tag{15.4}$$

whence, the dimensions of current I may be expressed in terms of the dimensions of μ_0 by

$$[I]^2 = [\mu_0]^{-1}[F],$$

when, of course, $[F]$ could be broken down further, into the dimensions of mass, length, and time.

Similarly, taking Coulomb's inverse square law equation in electrostatics,

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{QQ'}{r^2},$$

the dimensions of Q may be expressed in terms of ϵ_0 by

$$[Q]^2 = [\epsilon_0]\,[F]\,[L]^2.$$

But $\quad Q = It,$

therefore $\quad [I]^2 = [Q]^2\,[T]^{-2},$

i.e. $\quad [I]^2 = [\epsilon_0]\,[F]\,[L]^2\,[T]^{-2}.$

If we accept that current must have the same dimensions, whether we regard it as something that produces a magnetic field, or whether we regard it as being moving charges, we can take the above two equations for $[I]^2$ as being equivalent, whence, in terms of dimensions,

$$[\mu_0]^{-1}[F] = [\epsilon_0]\,[F]\,[L]^2\,[T]^{-2},$$

$$\therefore \quad [\mu_0]^{-\frac{1}{2}}\,[\epsilon_0]^{-\frac{1}{2}} = [L]\,[T]^{-1},$$

i.e. the dimensions of the expression $1/(\mu_0\epsilon_0)^{\frac{1}{2}}$ are those of a velocity.

*15.15 The permanent magnet

Historically, electromagnetic theory was originally developed from the properties of the permanent magnet, because the field of a magnet was discovered long before the field of a current. Although this way of dealing with the subject is no longer consistent with our view as to what aspect is the more fundamental, the basic phenomena of magnetism are still conveniently introduced by taking the general properties of a permanent magnet first, as was done in chapter 7. This procedure soon leads to us regarding the magnet as having its attracting and repelling properties concentrated in two regions, known as magnetic poles, situated somewhere near the ends of the magnet, and equal and opposite in their effects. Once we find, however, that these poles represent, merely, the places at which lines of magnetic flux enter and leave the magnet, and that they have no more claims to emphasis than the ends of a solenoid might have, we tend to drop the idea again, basing all our equations on the field of a current and the effect of the field on a current instead. This makes it easier to develop the subject, of course, where the effects of a current are concerned, but it also deprives us of the simplest way to deal quantitatively with the effects of a permanent magnet. Because this is what we now wish to do, it will be convenient to introduce the magnetic pole again, and to treat it quantitatively. In doing so, there is no need to claim that the magnetic pole has the same degree of reality as, say, an electric charge; we may, if we choose, treat it as a useful mathematical fiction, which is all that is required of it here.

*15.16 Magnetic moment and pole strength

The magnetic moment m of a magnet was defined in section 14.7 by comparison with the magnetic moment of a coil, and we had the equation

$$\Gamma = Bm \sin\theta, \tag{15.28}$$

where Γ is the torque which acts on the magnet when it is held so that its axis makes an angle θ with a magnetic field of flux density B.

By giving a quantitative meaning to the strength, Q_m, of a magnetic pole, this

torque may be interpreted in terms of a couple, made up from equal and opposite forces acting on the two poles of the magnet.

We first define pole strength, Q_m, by the equation

$$F = BQ_m, \tag{15.29}$$

where F is the force which acts on the pole when it is placed in a magnetic field of flux density B. This force acts in the direction of the flux, and not perpendicular to it, as the force on a current element does.

Now, supposing the two poles of the magnet to be separated by a distance l, we have

$$\Gamma = BQ_m l \sin\theta, \tag{15.30}$$

whence, from equation (15.28),

$$m = Q_m l. \tag{15.31}$$

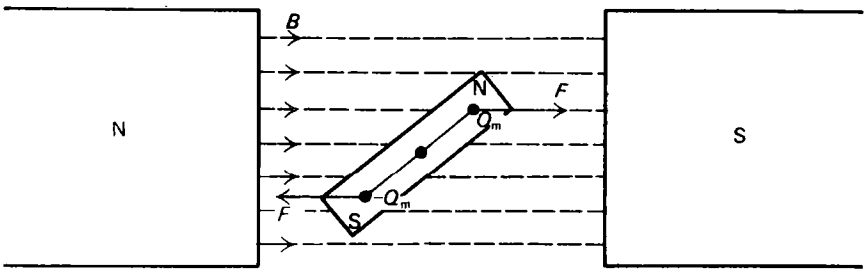


Fig. 15.15 The torque on a magnet

[It was pointed out in section 14.7 that the definition of magnetic moment m of a permanent magnet is a subject upon which there are two conflicting points of view. The alternative definition, which attaches the name and symbol to a different property of the magnet, would represent equation (15.28) – so long as the medium is non-magnetic – in a form which is equivalent to

$$\Gamma = \frac{B}{\mu_0} m \sin\theta$$

The same difference in approach applies to magnetic pole strength, Q_m, the equation which replaces equation (15.29) being, in effect,

$$F = \frac{B}{\mu_0} Q_m.$$

Little purpose will be served here by attempting to explain this difference further.]

*15.17 The flux density at distance r from a magnetic pole

Suppose that a very long magnet is placed so that it lies along the axis of a plane circular coil, the north pole of the magnet being at the centre of the coil, while the south pole is sufficiently distant to be outside the effective field of the coil (Fig. 15.16).

The flux density at the centre of the coil, due to the current which flows in it, is equal to B_1, where

$$B_1 = \frac{\mu_0}{4\pi} \cdot \frac{l}{r^2} I, \tag{15.7}$$

where l is the length of wire in the coil and r is its radius.

The resulting force which acts on the pole is

$$F = B_1 Q_m,$$

$$F = \frac{\mu_0}{4\pi} \cdot \frac{lI}{r^2} Q_m. \tag{15.32}$$

It follows from Newton's third law of motion that an equal and opposite force must act on the coil, which it does because the coil lies in the flux of the magnet,

i.e.
$$F = BIl, \tag{15.33}$$

where B is the flux density due to the magnetic pole at the distance r from it at which the coil is situated. Hence, from equations (15.32) and (15.33), we have

$$B = \frac{\mu_0}{4\pi} \cdot \frac{Q_m}{r^2}, \tag{15.34}$$

which is the equation required.

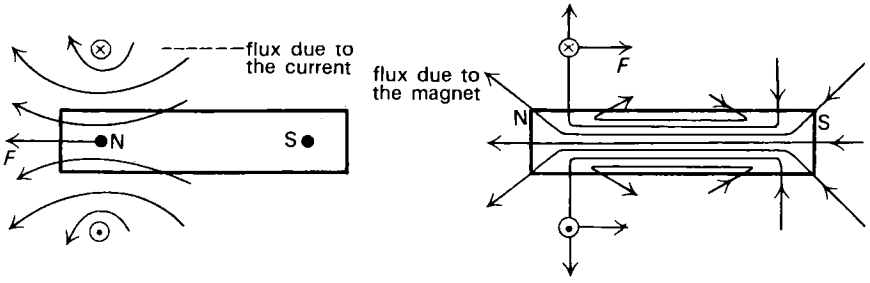


Fig. 15.16 The interaction between a magnetic pole and current I in a coil (a) The force F on the pole (b) The force F on the coil

Suppose now that two magnetic poles, having strengths Q_m, Q'_m are held at a distance r apart. The second pole, Q'_m, is in the flux due to the first, and is repelled by it (assuming the two poles to have like polarity),

i.e.
$$F = BQ'_m,$$

where
$$B = \frac{\mu_0}{4\pi} \cdot \frac{Q_m}{r^2}.$$

Therefore
$$F = \frac{\mu_0}{4\pi} \cdot \frac{Q_m Q'_m}{r^2}. \tag{15.35}$$

Thus, there is an inverse square law of force between magnetic poles, just as there is between electric charges, and it is on this equation that electromagnetic theory was originally built up.

The magnetic pole does not, however, possess the same credentials as the electric charge, and its effect is correctly given by equation (15.35) only when this is experienced at a point which lies outside the material of the magnet itself. The flux lines have been drawn in Fig. 15.16 (*b*) to illustrate this point: since lines of magnetic flux form continuous loops (which lines of force from an electric charge do not), the inverse square law of force would clearly have no meaning at a point inside the magnet itself.

*15.18 The force on a magnet, or on a coil, when placed in a non-uniform field

If a magnet is placed in a uniform field, it experiences a torque, so that, if it is pivoted, it acts as a compass and swings round so that its axis is parallel to the flux: but there is no resultant force, and we may say that the forces, $F = BQ_m$, which act on each pole are equal and opposite. If the field is not uniform, however, there is a resultant force. Assuming the flux to have a constant direction but varying strength, this force is given by

$$F = (B_2 - B_1)Q_m,$$

where B_1, B_2 are the values of the flux density at the two poles. This may be put in the more useful form

$$F = \frac{B_2 - B_1}{l} m,$$

or, if the magnet is very short

$$F = \frac{dB}{dx} m, \tag{15.36}$$

where x is measured in the direction of the flux and m is the magnet's magnetic moment.

This is the force, for example, with which two magnets attract each other, if they are placed with their axes in line and with unlike poles facing each other. If a coil is placed with its axis in the same direction as a non-uniform field, a force likewise acts on the coil, and it can be shown that equation (15.36) applies also to the coil. In this case, the moment of the coil is (section 14.6)

$$m = NAI, \tag{15.37}$$

where N = the number of turns wound on the coil and A = its area. Hence, for a coil

$$F = \frac{dB}{dx} NAI. \tag{15.38}$$

This equation is the basic equation of the current balance, which will be described in the chapter 17.4.

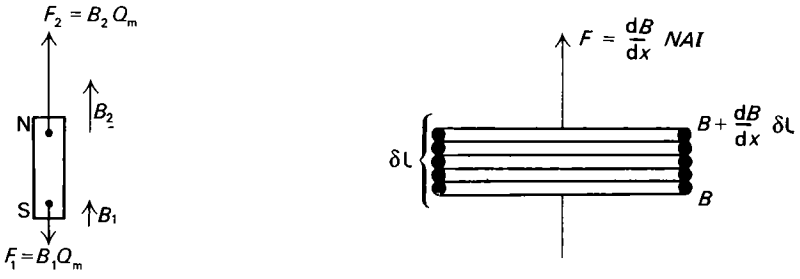


Fig. 15.17 The force on a magnet, and on a coil, placed in a non-uniform magnetic field

QUESTIONS

$$\mu_0 = 4\pi \times 10^{-7} \text{ H/m}$$

15.1 Calculate the value of the flux density B at the point P in each of the following cases. Draw a sketch in each case to indicate the direction of the line of magnetic flux through P.

(*a*) P is at the centre of a plane circular coil consisting of 500 turns, radius 15 cm, carrying a current of 4.0 mA.
(*b*) P is distance 20 cm from a system of 12 long vertical straight conductors, tied together, and each carrying a current of 2.0 A. (The return wires are sufficiently far away to be neglected.)
(*c*) P is a point on the axis of a long straight solenoid, wound with 5 turns to every cm, and carrying current 3 A.
(*d*) A solenoid consists of a coil wound on a circular closed tube, forming an annulus of radius 15 cm. The total number of turns on the coil is 90, and they carry a current of 4 A. P is a point inside the tube.

15.2 The same current is passed in turn through a long straight vertical wire and through a single circular loop of wire, radius 10 cm, placed with its plane vertical. Compare the flux density at 10 cm from the straight wire with the flux density at the centre of the circular loop.

15.3 A very long straight wire PQ of negligible diameter carries a steady current I_1. A square coil ABCD of side l with N turns of wire also of negligible diameter is set up with the sides AB and DC parallel to and coplanar with PQ; the side AB is nearer to PQ and at a distance d from it. Derive an expression for the resultant force on the coil when a steady current I_2 flows in it, and indicate on a diagram the direction of this force when the current flows in the same direction in PQ and AB.

Calculate the magnitude of the force when $I_1 = 5.0$ A, $I_2 = 3.0$ A, $d = 3.0$ cm, $N = 48$, $l = 5.0$ cm. (O & C)

15.4 Write down an expression for the magnetic flux density at a point near the middle of a long straight solenoid, and indicate by a diagram how the direction of the flux is related to the direction of the current flow. By considering the solenoid to be divided into two halves, deduce what you can about the flux density at the end of a long solenoid.

A long solenoid consists of a single layer of turns of wire in the form of a helical spring. A small circular coil is placed coaxially inside the solenoid near its mid-point, and connected to a centre zero galvanometer. Explain carefully why the galvanometer deflects when the current in the solenoid is switched on.

If the galvanometer deflects through 10 divisions to the right when the current is switched on in this way, what would happen in each case if (*a*) the current were suddenly reversed, (*b*) the circular coil were rapidly moved to one end of the solenoid, keeping it coaxial, or (*c*) the solenoid were suddenly compressed to two thirds of its original length, the current remaining constant? (O & C)

15.5 A ballistic galvanometer of resistance 15.0 Ω and sensitivity 5.0 divisions per microcoulomb is connected in series with a resistance of 100 Ω and a secondary coil of 500 turns and of resistance 50 Ω. This coil is wound round the middle of a long solenoid of radius 3.0 cm having 10 turns/cm and carrying current 0.60 A.

Assuming no damping, calculate the deflection produced in the galvanometer when the current in the solenoid is switched off. (L)

15.6 A square loop of wire of side l carries a current I_1. It is situated a distance r, which is much greater than l, from a long straight wire carrying a current I_2. The straight wire lies in the plane of the loop, and two sides of the loop are parallel to the straight wire. Find the magnitude and direction of the resultant force and/or couple which may be acting on the loop when it is in this position, and also when it is turned with its plane in each of the two mutually perpendicular positions. (O & C)

15.7 A non-magnetic wrist watch is placed at the centre of a long solenoid so that the balance wheel oscillates with its axis parallel to the field at the centre. Calculate the maximum e.m.f. generated in a radial spoke connecting the rim of the balance wheel to its centre, using the following data:

current in solenoid = 10 A,

winding of solenoid = 100 turns/m,
diameter of balance wheel = 0.50 cm,
maximum angular velocity of balance wheel = 2.0 rad/s.

What would be the effect if the watch were totally encased in mild steel? (C)

15.8 A plane circular coil, having N turns of wire, each of radius r, is placed with its plane vertical and perpendicular to the earth's magnetic meridian. Current I flows in the coil in such a direction that the flux density B at its centre due to the current is in the same direction as that of the horizontal component B_{hor} of the earth's flux density. Write down an expression for the resultant flux density at the centre of the coil.

A small magnet, freely suspended at this point, has period of oscillation t. The current I is now varied, and a series of observations of corresponding values are taken of t and I, and a graph is plotted of $1/t^2$ against I. Sketch the form which you would expect this graph to take, and explain how you would deduce from it the value of B_{hor}.

15.9 A circular coil having 9 turns each of radius 7.5 cm is placed with its plane vertical and perpendicular to the earth's magnetic meridian. A small magnet is suspended to oscillate at the centre of the coil. It is found that the period of oscillation of this magnet has the same value when the current in the coil is 0.270 A and when it is 0.195 A, the direction of the current being the same in both cases.

Expain these results, and deduce the value of B_{hor}, the horizontal component of the earth's flux density.

15.10 A plane circular coil of radius 9.0 cm and 50 turns is placed with its plane vertical and perpendicular to the earth's magnetic meridian. A small magnet is suspended at the centre of the coil. When the current flows in the coil in one direction the magnet makes 12 vibrations per minute. When the current is reversed in direction, the magnet turns round, and makes 15 vibrations per minute. If the horizontal component of the earth's field has flux density 1.8×10^{-5} T, deduce the value of the current.

15.11 A long straight solenoid, wound with 20 turns of wire in every cm, is placed with its axis horizontal and perpendicular to the earth's magnetic meridian. When current I flows in the solenoid, the direction of the resultant flux inside the solenoid is such that its horizontal component makes an angle of 60° with the earth's meridian. If the horizontal component of the earth's magnetic field has flux density $B_{hor} = 1.8 \times 10^{-5}$ T, calculate the value of I.

15.12 A vibration magnetometer is situated 8.0 cm magnetic west from a long straight vertical wire. When the magnetometer is set in vibration it makes 25 oscillations in 20.0 s when no current flows in the wire, and the same number of oscillations in 15.0 s when a current I flows. Assuming the value of the horizontal component of the earth's magnetic flux density to be 1.8×10^{-5} Wb/m^2, deduce the value of I.

15.13 A flat circular coil consisting of 150 turns of wire of mean diameter 50 cm is fixed with its plane vertical and lying in the earth's magnetic meridian, and a current of 2.0 A is passed through it. A flat rectangular coil of 100 turns measuring 2 cm by 3 cm is situated at the centre of the larger coil, with its longer sides vertical and its plane perpendicular to the meridian.

Calculate the couple acting about a vertical axis on the smaller coil when a current of 4 mA is passed through it.

15.14 A toroidal solenoid consists of 400 turns of wire wound round a closed circular tube, the cross-sectional area of which is 2.0 cm^2. If the mean radius of the toroid is 10.0 cm, estimate the self inductance L of the coil.

15.15 A long straight solenoid is wound with 7 turns of wire to every cm. Placed inside the middle of this solenoid and coaxially with it is a smaller coil, consisting of 100 turns of wire enclosing area 5 cm^2. Estimate the mutual inductance M of the arrangement.

15.16 A circular coil has 100 turns, each of radius 25 cm. A small coil having 5 turns of area 1 cm^2 is placed at the centre of the larger coil, the planes of the two

coils being inclined at an angle of 60°. Calculate the mutual inductance between the two coils.

If both coils carry a steady current of 3 A, and the outer coil is fixed, what torque does the small coil experience? (Assume both coils to be perfectly flat and situated in a field free region.) (O & C)

*15.17 An endless solenoid (toroidal coil) is uniformly wound with 1700 turns of wire on an insulating ring of mean circumference 20 cm and cross-sectional area 5 cm^2. Assuming the magnetic field inside the coil to be uniform in intensity, calculate the inductance L of the solenoid.

When a 50 Hz sinusoidal alternating p.d. of 12 V r.m.s. is applied across the coil a current of 2.4 A flows. When the frequency is doubled the current is reduced to 1.67 A for the same p.d. Calculate the actual inductance from these data, and explain why it is somewhat greater than the value calculated from the physical dimensions.

When the experiment is repeated at a much higher frequency the results are not quite consistent with those obtained at 50 Hz and 100 Hz. Explain briefly one phenomenon which could account for the inconsistency. (O & C)

*15.18 A toroidal solenoid X is formed by winding 5000 uniformly spaced turns of fine wire around a ring shaped plastic former of mean radius 10 cm and circular cross-section of 1 cm radius. A very long straight wire lies along the axis of the toroid (i.e. through the centre of its circle and perpendicular to its plane) and forms part of a closed circuit Y, the rest of which is at a large distance from the solenoid. Show that, if the field in the solenoid can be taken as uniform and equal to its value at the mean radius, the mutual inductance M between X and Y is the same as that between Y and X, the value in each case being equal to $\pi \mu H$.

If the circuit Y carries an alternating current of r.m.s. value 10 A and frequency 50 Hz, what is the r.m.s. value of the e.m.f. induced in the solenoid X? (O & C)

*15.19

In the figure, X is a long straight wire in the same plane as the rectangular metal frame Y, which has resistance 0.050 Ω. Calculate the flux linked with

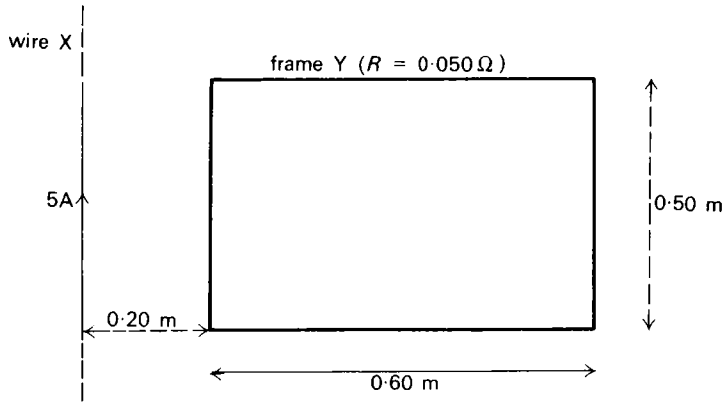


Fig. e 76

Y when 5 A flows in X, and also the charge circulating round Y when the current in X is reversed.

(Log_e 10 = 2.303.) (O)

*15.20 A pair of long parallel wires 0.10 m apart in air carry an alternating current, r.m.s. value 10 A 50 Hz, in opposite directions. A single turn of wire 5 cm square lies in the plane of the wires, midway between them. Find the r.m.s.

value of the e.m.f. induced in the turn, and indicate with sketches the relative phase differences between the currents in the parallel wires, the magnetic flux through the turn, and the e.m.f. induced. (OXF. SCH.)

*15.21 Find the maximum charge which can be stored on an isolated conducting sphere of diameter 1 cm if the insulation of the air breaks down when the electric field strength exceeds 3×10^6 V/m. Twenty such spheres, each carrying this maximum charge, are mounted on an insulating ring so that their centres are on a circle of radius 15 cm. At what speed must the ring be rotated, about an axis through its centre perpendicular to its plane, to produce a magnetic field of flux density 10^{-9} T at the centre of the circle?

Make the simplifying assumptions that the charge on each sphere acts as if it were concentrated at the centre of that sphere, and that the electric field strength at the surface of the sphere is due to the charge on that sphere alone. ($1/4\pi\epsilon_0 = 9 \times 10^9$ SI units.) (O & C)

*15.22 A small sphere carrying charge Q is rotated in a circle of radius r and makes f revolutions per second. Show that it is equivalent to a circular current of the same radius and of magnitude Qf.

According to Bohr, a hydrogen atom may be regarded as a system consisting of an electron of charge -1.6×10^{-19} C and mass 9.1×10^{-31} kg, rotating in a circle of radius 5.3×10^{-11} m with a proton as centre. Calculate the magnetic flux density experienced by the proton due to the rotating electron. (O & C)

*15.23 What is the path of a charged particle in a uniform magnetic field if its initial velocity is not at right angles to the field?

An electron of speed 2.0×10^7 m/s is emitted from a point on the axis of a very long solenoid wound with 10^3 turns per metre and carrying a current of 2.5 A. If the initial velocity makes an angle α with the solenoid axis, find at what distance from its starting point the particle next crosses the axis, and show that this distance is independent of α if α is small.
(e/m for an electron is 1.76×10^{11} C/kg.) (O & C)

*15.24 A cathode ray tube in which the distance between anode and screen is 25 cm is placed in a long solenoid having 20 turns/cm, the axis of tube and solenoid being coincident. If the final anode potential is 2.0 kV and the electrons emerging from the hole in it are slightly divergent, estimate the least solenoid current required to bring the electron beam to a fine focus on the screen. ($e/m_e = 1.76 \times 10^{11}$ C/kg.) (N)

*15.25 A plane circular coil consists of N turns, each of radius r, and carries current I. The axis OP is perpendicular to the plane of the coil and passes through the centre O of the coil. Write down the equation giving the flux density at P due to current in the coil, where the distance OP $= x$.

By differentiation, show that the variation of B with x, i.e. dB/dx, is constant over an appreciable distance when $x = r/2$. Comment on this conclusion with reference to the arrangement of two coils known as a Helmholtz pair.

*15.26 A point P is situated on the axis of a short bar magnet, magnetic moment m, at a distance d from the mid point of the magnet. By treating the magnet as effectively consisting of two magnetic poles, strength $+Q$m and $-Q$m respectively, separated by distance l, where $m = Q_m \times l$, obtain an expression for the magnetic flux density caused at P by the magnet. Simplify the equation on the assumption that $l \ll d$.

Suppose a second short magnet, magnetic moment m', is now placed with its mid point at P, so that the two magnets lie in the same straight line. Show that the resultant force between the magnets is proportional to $1/d^4$.

The figure represents two small coils X, Y placed with their windings parallel to each other, and placed a mean distance d apart, the number of turns and the current being respectively N_1, I_1, and N_2, I_2, and the areas enclosed by the turns being A_1 and A_2.

Treating such a coil as being equivalent to a short bar magnet of moment

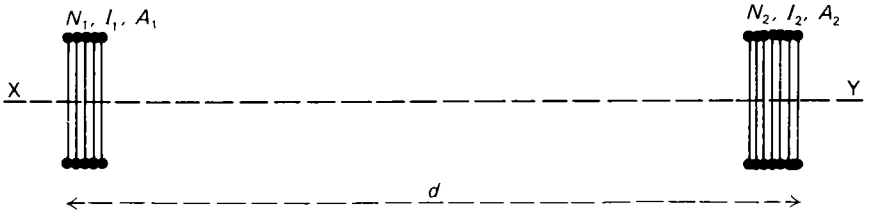


Fig. e 77

m, where $m = NAI$, obtain an expression for the force which acts between the two coils.

15.27 A rectangular coil of very large dimensions is set up in a vertical plane in the laboratory in such a way that one vertical side of the coil passes through a hole in a bench top, the horizontal wires run along the floor and the ceiling, and the vertical return path, including the source of current, is a considerable distance from the bench. The coil has 20 turns, so that 20 parallel straight conductors pass through the hole in the bench, and each carries current 1.5 A downwards.

Allowing for the effect of the earth's magnetic flux, sketch the magnetic flux pattern which you would expect to exist on the bench top in the neighbourhood of the wires, and mark in the neutral point.

If the horizontal component B_{hor} of the earth's magnetic flux density is $B_{hor} = 2.0 \times 10^{-5}$ T, how far from the conductors is the neutral point?

A vibration magnetometer is placed at a point P to the magnetic west of the wires, distance r from them, and its period of vibration is t, which increases to the value t_{hor} when the current is switched off. Observations of the value of t are made for a series of values of r, and $1/t^2$ is plotted against $1/r$. Sketch the form which you would expect the graph to take.

If $t_{hor} = 6.5$ s, and $t = 4.1$ s at the point P, how far is P from the wires?

Chapter 16

Magnetic materials

16.1 The effect of a ferromagnetic material placed in a magnetic field

The equations which were discussed in chapter 15 are, if we wish to be very particular about it, free space equations; that is to say, the important factor contained in them is μ_0, the permeability of a vacuum, and there is nothing in them to allow for any effect which a medium might have. Taking a practical view, however, these equations are entirely acceptable when the only medium present is air, and in most other cases the effect of the medium is so slight that it can be neglected for most practical purposes. The exceptions to this, however, which are the classes of material known as ferromagnetic and ferrimagnetic, are of the greatest practical importance. These become so strongly magnetized when placed in a magnetic field that their presence results in a large increase in the flux density, both inside the material itself and in the space surrounding it. Consequently the equations of chapter 15 apply accurately only if such materials are removed from the vicinity.

In the simplest cases, the equations can be successfully modified to allow for the effect of a ferromagnetic medium, and such cases will be considered in this chapter. More generally, however, the problem becomes difficult, and the treatment has to be to a large extent empirical.

16.2 Iron in a uniform field

Suppose that a continuous iron core is placed in a toroidal solenoid. The arrangement is symmetrical. Except that the flux density round the inside circumference of the toroid is slightly greater than that round the outside circumference, we can say that the whole of the iron is in a uniform field, and that the lines of magnetic flux never emerge from the iron. This is the simple kind of situation in which meaningful results can be extracted. A method of investigating the effect of the iron is illustrated in Fig. 16.1. The two toroidal solenoids, 1 and 2, are identical in all respects, except that, whereas the first has an iron core, the second is wound on a ring of some substance, such as wood, the magnetic effect of which can here be treated as negligible.

Each toroid has a secondary coil wound round at some section, and the two secondary coils are connected in series through a ballistic galvanometer. Current may be sent through either one of the two primary solenoids, and its direction may be reversed by means of the reversing switch S_1.

Suppose that the flux density due to a current I in the iron filled solenoid is B, the area of cross-section being A. On reversal of this current, the change in flux threading the secondary coil is 2Φ, where

$$\Phi = BA, \tag{16.1}$$

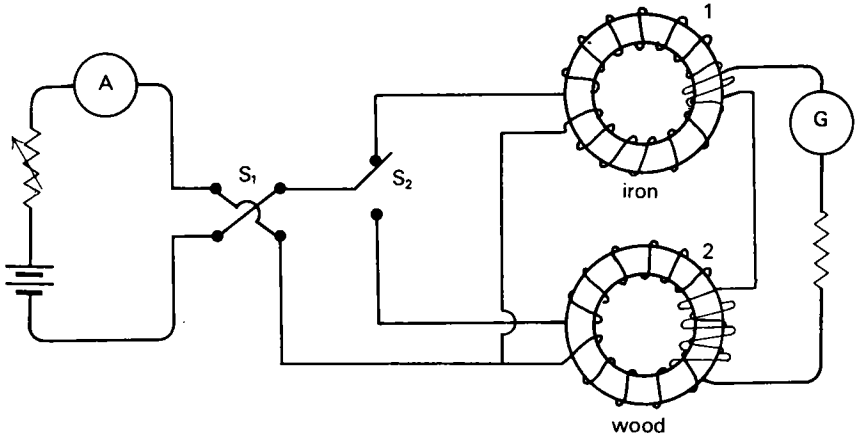


Fig. 16.1 Circuit to investigate the magnetization of an iron core

and the quantity of electric charge which flows through the ballistic galvanometer in consequence is

$$Q = N\frac{2\Phi}{R}$$

$$= N\frac{2BA}{R}, \tag{16.2}$$

where N is the number of turns wound on the secondary coil, and R is the total resistance of the secondary circuit. Similarly for the other solenoid, which lacks the iron,

$$Q_0 = N_0\frac{2B_0A}{R}. \tag{16.3}$$

If the numbers of turns N, N_0 were the same for both secondary coils, the galvanometer throws θ, θ_0, since they are proportional to Q, Q_0, would be proportional to the flux densities B, B_0. In the event, however, the iron increases the flux in the first solenoid so markedly that the number of turns N_0 wound on the wood needs to be at least 100 times the number wound on the iron, if the two galvanometer throws are to be in any way comparable.

Plotting galvanometer throw θ against the magnetizing current I in each case, graphs are obtained which have the form shown in Fig. 16.2. There are two very noticeable features in these graphs. Firstly, the iron, by becoming magnetized, greatly increases the flux, possibly by a factor which exceeds 10^3. Secondly, whereas in the absence of iron the flux density is proportional to the magnetizing current, this is far from being the case when the iron is there.

16.3 Relative permeability μ_r

The ratio of the flux density B in the iron ring to the flux density which would be present, due to the same current, in the absence of the iron, is called the relative permeability μ_r of the iron. Assuming the two toroids in Fig. 16.1 to be exactly similar, this is the ratio of B to B_0 in the graphs, and we can write

$$\mu_r = \frac{B}{B_0}, \tag{16.4}$$

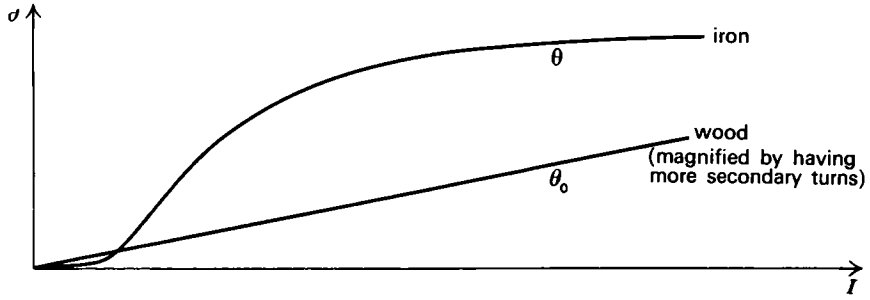


Fig. 16.2 The effect of an iron core

$$\therefore \quad \mu_r = \frac{\theta}{\theta_0} \times \frac{N_0}{N}. \tag{16.5}$$

Clearly the relative permeability of an iron specimen is not a constant quantity: data for plotting a graph to show how it varies with the magnetizing current I may be obtained by taking corresponding pairs of values of θ_0, θ from the graphs of Fig. 16.2, and making due allowance for the difference in values of N_0 and N. A typical result is shown in Fig. 16.3.

It is stressed that this definition of relative permeability, as stated in terms of equation 16.4, is valid only with reference to a specific arrangement such as that

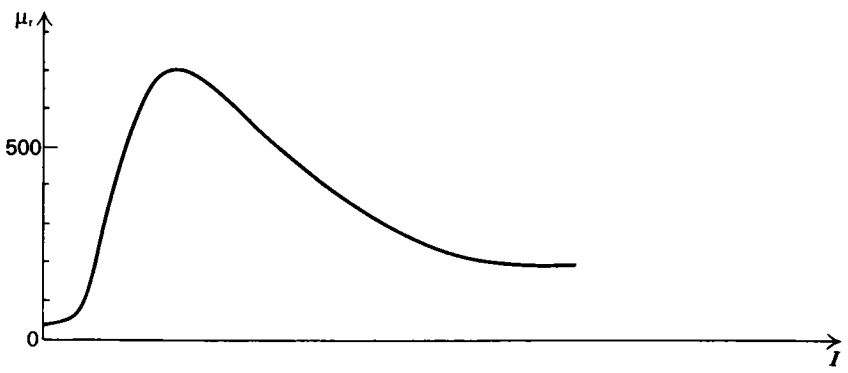


Fig. 16.3 A typical graph showing the variation of relative permeability μ_r of iron with the magnetizing current I

of the toroid, in which the flux never emerges from the iron, and the flux density is uniform (or as uniform as we can make it). It would not be correct to assume that the flux density at any point is automatically increased by the factor μ_r, merely by putting a piece of iron in a region which contains the point.

As we have just seen, the relative permeability of a ferromagnetic material is very much greater than unity (unless the magnetizing field is very large), and it varies very markedly with the strength of the magnetizing field. The apparatus we have just described is not suited to the study of diamagnetic and paramagnetic materials (section 7.6), but in these cases the value of μ_r is very nearly equal to unity, being slightly greater than unity for paramagnetic materials, and slightly less than unity for diamagnetic, and is effectively constant.

The flux density B_0 in the iron-free toroid (equation 15.6) is given by

$$B_0 = \mu_0 \frac{N}{l} I, \tag{16.6}$$

where N now represents the number of turns wound on length l of the primary solenoid. Hence, from equation 16.4, the flux density in the iron-cored toroid becomes

$$B = \mu_r \mu_0 \frac{N}{l} I \tag{16.7}$$

and this is sometimes written in the form

$$B = \mu \frac{N}{l} I, \tag{16.8}$$

where μ is called the permeability of the iron, by comparison with μ_0, the permeability of a vacuum.
We then have

$$\mu_r = \frac{\mu}{\mu_0} \tag{16.9}$$

which accounts for the name, relative permeability, which has been allotted to μ_r.

16.4 Magnetizing force H

It would be more useful, when illustrating the variation of μ_r with the strength of the magnetizing field, to plot the value of μ_r against the value of the magnetizing flux density B_0 rather than choosing the current I, as was done in Fig. 16.3. This is because such a graph would display the variation in terms of a material, indepen-

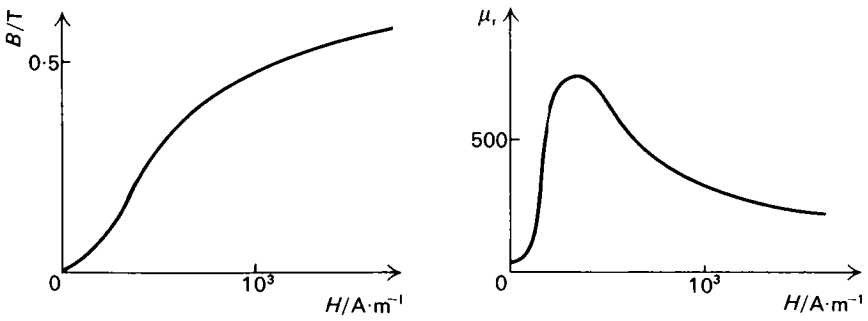


Fig. 16.4 Ferromagnetism of iron. Plotting flux density and relative permeability against magnetizing force H

dently of the actual dimensions of the chosen specimen. More often, however, such graphs are plotted omitting the factor μ_0 from the equation

$$B_0 = \mu_0 \frac{N}{l} I \tag{16.6}$$

and writing instead

$$H = \frac{N}{l} I, \tag{16.10}$$

where H is called the magnetizing force. (It is alternatively referred to as the magnetic field strength.)

In SI units, H is measured in amperes per metre.

16.5 Lines of magnetic flux form continuous loops

It is assumed throughout electromagnetic theory that lines of magnetic flux form continuous loops, and that this remains true, even when they pass in and out of a

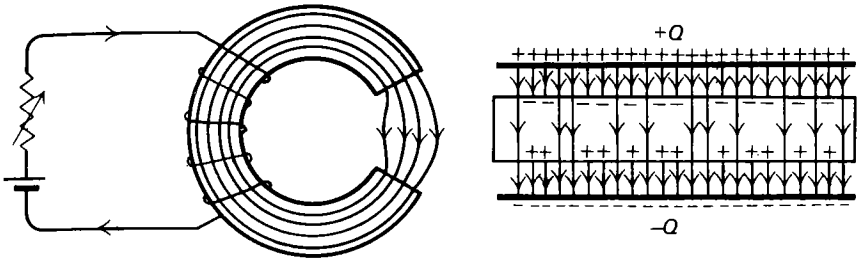


Fig. 16.5 Comparing magnetic flux lines with lines of electric force (a) Lines of magnetic flux (b) Lines of electric force

ferromagnetic material. This contrasts with electric lines of force: the electric field strength between charged parallel plates changes abruptly in value at the surface of a dielectric (section 6.21).

This continuity of magnetic flux lines may be tested by using some arrangement such as that shown in Fig. 16.6. If the current flowing in the solenoid is reversed,

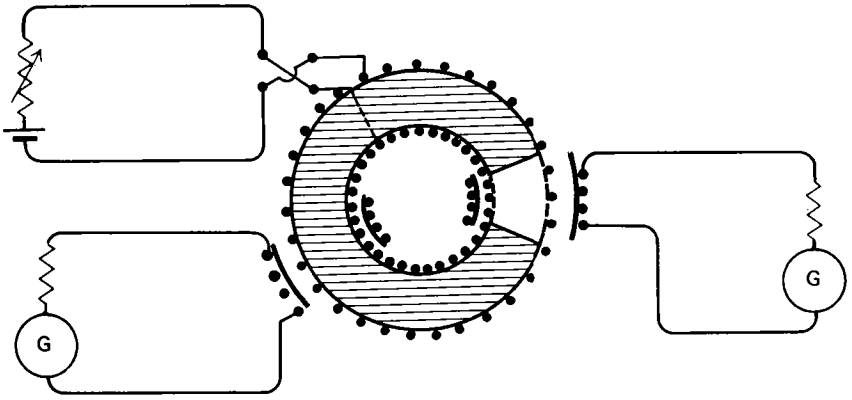


Fig. 16.6 An arrangement to demonstrate the continuity of lines of magnetic flux

the galvanometer throw proves to be the same, whether the secondary coil surrounds the iron, or whether it surrounds the air gap, which is to say that the flux threading this coil due to the current and the magnetized iron is the same in both places.

16.6 The weakening effect of an air gap

In defining relative permeability we insisted that the definition referred to flux confined to the single medium. This was because, if the circumstances were allowed to be otherwise, nothing useful would have been defined: a discontinuity in the medium, such as the air gap in the iron toroid of the last diagram, has a very

marked effect in reducing the total flux. This could be demonstrated by inserting a piece of iron into the gap, so shaped that it makes an exact fit, the result being that the galvanometer throw on reversing the current would be greatly increased. A further experiment to demonstrate this is illustrated in Fig. 16.7. A choke is

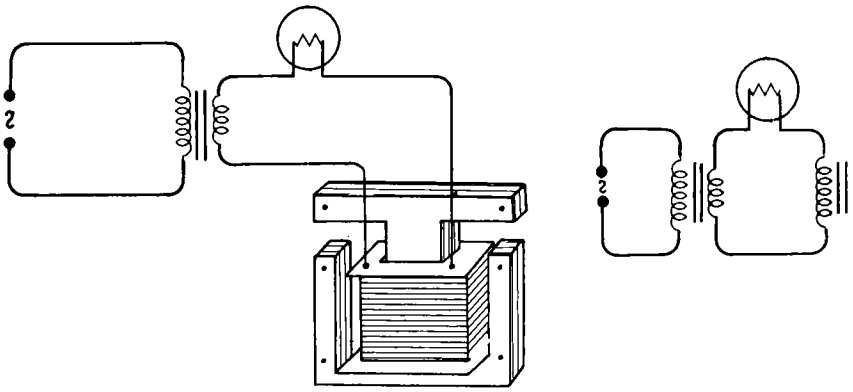


Fig. 16.7 An experiment which illustrates the effect of an air gap

constructed by having a coil of copper wire wound on a cardboard former, which can be slipped over a yoke of soft iron stampings. If the iron circuit for the magnetic flux is made continuous, this is an efficient choke. If an electric lamp is connected in series with this choke to an a.c. source, the current is small, and the lamp fails to light up. If a small air gap is introduced, however, thus interrupting the magnetic circuit, so that the flux lines are forced to emerge into air, this greatly reduces the efficiency of the choke, and the lamp partially lights.

16.7 Hysteresis

If the magnetization of an iron toroid is taken through a cycle of values by taking the magnetizing current I in its solenoid up to a maximum value, reducing it again to zero, reversing its direction, and repeating the process, a graph of the flux density B against the magnetizing current I or against the magnetizing force H takes the form shown in Fig. 16.8. The iron reveals a tendency to retain its previous

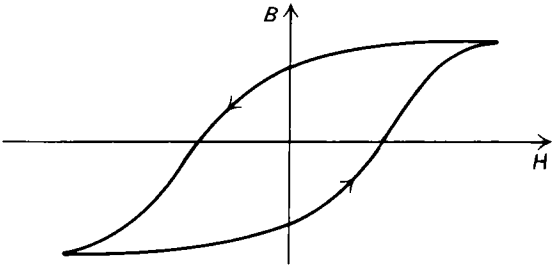


Fig. 16.8 A hysteresis curve

magnetic condition, and so the graph does not follow the same path for both directions of the cycle. This is called hysteresis. If alternating current flows in the solenoid, the iron repeatedly follows this cycle in its behaviour.

A suitable circuit for obtaining readings for a hysteresis curve is shown in Fig. 16.9. As the contact of the variable resistor chain is moved along the studs from stud **a** towards stud **h**, the effect is to reduce the magnetizing current in stages from maximum until at **h** it reaches zero. The connection to the solenoid is then

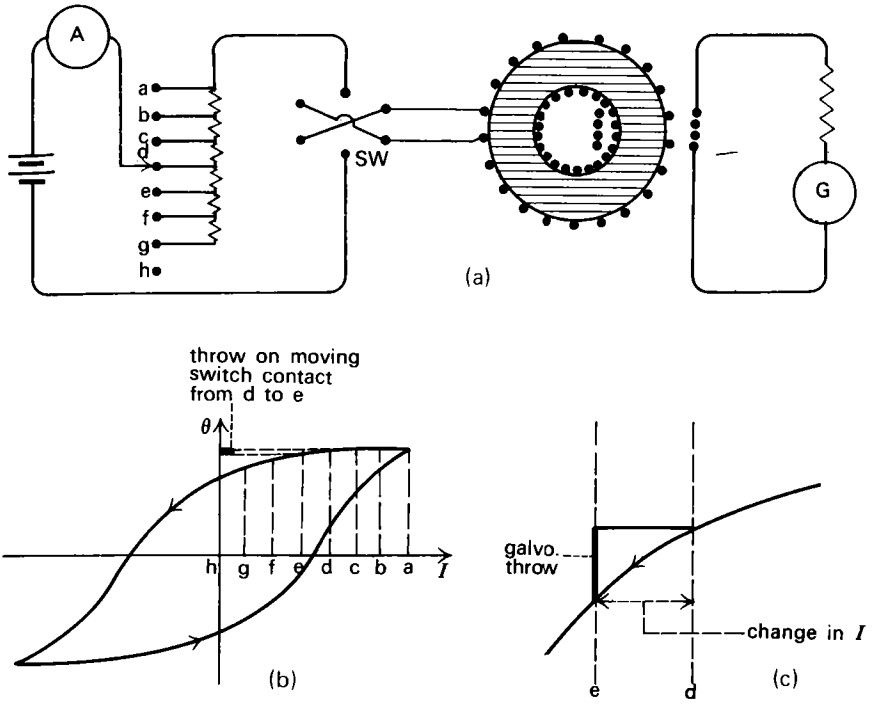


Fig. 16.9 Plotting a hysteresis cycle (a) (b) (c) A magnified section of graph (b)

reversed by means of the reversing switch SW, and the contact is moved back towards **a** to build up the current to maximum again. A cycle of values can be traversed in this way. Each time that the contact is moved, there is a change in flux, which can be measured by the corresponding throw of the ballistic galvanometer connected to the secondary coil. (Arrangements to calibrate the galvanometer would be required in practice – section 14.13.) The hysteresis loop can now be plotted by adding together the successive flux changes, in the manner suggested by the diagrams.

Hysteresis results in the conversion of some electrical energy into internal energy of the iron, causing a temperature rise, and consequently there is a loss of useful energy. Each time that the magnetizing current is taken through a complete cycle of values the total loss on this account is proportional to the area of the hysteresis loop. This is important to the choice of material for the cores of alternating current devices such as the transformer, for which a hysteresis loop enclosing only a small area is an essential requirement.

16.8 To demagnetize a magnetized material

Hysteresis is the reason that a permanent magnet comes to exist. Once the magnetism of the domains (section 16.13) has been made unidirectional, or partially so,

removal of the magnetizing flux does not serve to make their magnetism completely random again, and some residual magnetic moment remains.

The best way to demagnetize an object, should this be required, is to place it in a solenoid through which alternating current is flowing. The current is then reduced slowly to zero, or, alternatively, the object is slowly removed from the solenoid.

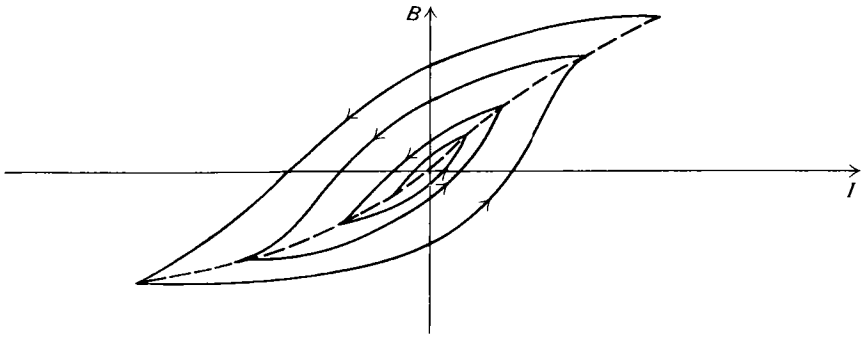


Fig. 16.10 The demagnetizing effect of an alternating field which progressively diminishes in strength

Either procedure has the effect that the magnetism in it is taken through a succession of hysteresis cycles, progressively diminishing in size, and the object finishes with no permanent magnetic moment left in it.

16.9 Permanent magnets

The properties required of a material if it is to make a satisfactory permanent magnet are revealed by studying the second quadrant of its hysteresis loop. The

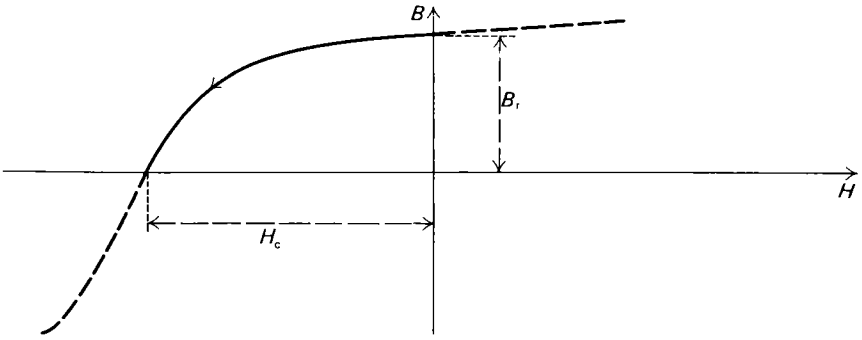


Fig. 16.11 The second quadrant of a hysteresis loop, indicating the properties of a permanent magnet

magnet needs to be both permanent in its magnetism and powerful. Its permanence can be judged by the strength of the reverse magnetizing force H_c which is needed to reduce the flux to zero, once it has been magnetized to saturation. This is called the coercivity. The strength of the magnet is measured by the flux density B_r which remains when the magnetizing field has been reduced to zero, and this is called the remanence.

Alloys which have high remanence and high coercivity are very hard, sometimes brittle. One example of a permanent magnet alloy of this kind is Alnico, which is iron alloyed with aluminium, nickel, cobalt, and copper. Owing to their hardness, permanent magnets usually have simple shapes, soft iron pole pieces being added to guide the flux.

16.10 Electromagnets and transformer cores

The core of an electromagnet requires to have a high maximum value for its relative permeability μ_r. It may be a further advantage if the coercivity is small, making its magnetism vanish when the magnetizing current is removed. A suitable alloy might be nickel-iron.

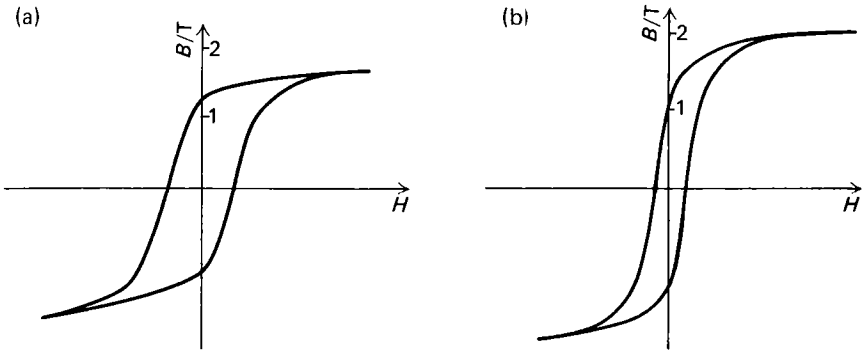


Fig. 16.12 Examples of hysteresis curves (a) Nickel-iron (b) Silicon-iron

For the core of a transformer, a small area hysteresis loop is of major importance, because this is an alternating current device, and the alternating flux would otherwise produce an excessive loss of electrical energy, and temperature rise of the core. The alloy used often contains silicon, because this increases the resistivity and so makes it easier to reduce eddy currents induced in the core. The materials which have small hysteresis effects and are used in a.c. devices, may be loosely classified as being soft iron alloys, being softer, and easier to work, than the hard steels of permanent magnets.

16.11 Ferrites

Ferrites are composite oxides of iron which include other oxides in their chemical structure, e.g. $MnFe_2O_4$. These compounds, having magnetic properties which are almost comparable with those of iron, are classified as ferrimagnetic. They are semiconductors, which means that, compared with ferromagnetic materials, they are semi-insulators. Their use, on this account, in place of iron in transformers and chokes at radio-frequencies was referred to in section 7.17.

Other ferrites are suited to other magnetic applications. Magnadur, for example, a barium ferrite, makes a good permanent magnet. It is not as powerful as alnico, but its magnetism is very difficult to destroy.

Square loop ferrites

These are so called because the shape of their hysteresis loop is almost rectangular. The result is that such a ferrite can only be in one or other of two conditions, being

magnetized in one direction, or having its magnetism reversed. This makes the square loop ferrite useful in a computer circuit, as a miniature memory store. A bead of the ferrite, threaded on a conducting wire, remains magnetized in the direction of the flux which last surrounded a current in the conductor. If a reverse current flows, the magnetism is reversed, and this happens almost instantaneously, at any rate in less than one microsecond. Between times, the ferrite bead is a permanent magnet, storing in its polarity the record of the last current which flowed.

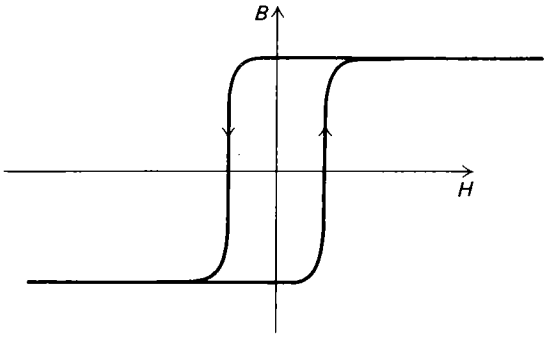


Fig. 16.13 Hysteresis loop of a square loop ferrite

16.12 The Curie temperature

The relative permeability of a ferromagnetic material falls slightly with temperature rise, until a critical value is reached, which is called the Curie temperature. Above this point, the material is only paramagnetic. The Curie temperature is 770 °C for iron, 1 100 °C for cobalt, and 350 °C for nickel.

*16.13 The atomic theory of magnetic properties

In searching for an origin of the magnetic properties of a substance, any acceptable theory must account for the small diamagnetism ($\mu_r < 1$) and paramagnetism

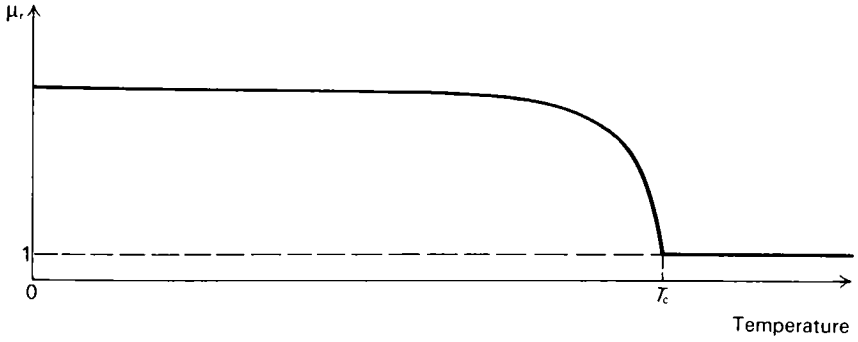


Fig. 16.14 The Curie temperature

($\mu_r > 1$) possessed by the vast majority of substances, as well as for the powerful ferromagnetism possessed by a few.

It might be expected that the source of the atom's magnetism would be what

can be described as the solenoid effect of the electrons, due to the fact that they are moving in orbits round the nucleus, giving the atom a magnetic moment resembling that due to a current in a coil. We can call this, for convenience, an orbiting magnetic moment. This is, indeed, the basis on which diamagnetism is explained, but it turns out that a contribution of at least equal importance has a different origin, and this is responsible for paramagnetism, of which ferromagnetism is a special case. The quantum theory suggests that each electron in the atom acts as a magnetic dipole in its own right, quite independently of any magnetic effect caused by its motion in an orbit. We will call the magnetic moment of this electronic dipole its spin magnetic moment, because it is associated with a mechanical property known as spin moment, i.e. the electron possesses angular momentum, as if it were spinning about an axis through itself.

Most of the electrons surrounding the nucleus of an atom fill the orbitals of completed electron shells (section 11.2). These electrons make no contribution to the magnetic behaviour of the atom, because the magnetic moments of the two electrons in each orbital, whether due to orbiting or due to spin, are exactly equal and opposite, and cancel each other out. The resultant magnetic behaviour of the atom, therefore, is due to the small number of electrons which are in an uncompleted shell, i.e. in the valency shell, or, in some cases such as iron, in the shell immediately below the valency shell in its level of energy. Even so, most of these electrons, too, are likely to be paired off, because this is the necessary condition of two electrons which occupy the same orbital. The overall effect, therefore, is that the magnetic behaviour of an individual atom is due to only one or two of its electrons.

There is an important contrast between the effects of a magnetizing field on the two kinds of magnetic moment possessed by an electron, just as there is a difference between the behaviour of a coil which is fixed in a field and that of a coil or a magnet which is pivoted. In the case of the fixed coil, if magnetic flux is introduced into it, the effect is electromagnetic induction. In accordance with Lenz' law, current is induced in the coil in such a direction that it reduces the increase in flux. If, on the other hand, a coil or magnet is pivoted, and magnetic flux is introduced, the coil or magnet turns, like a compass, to lie along the field, causing its magnetism to supplement the flux. It is on the basis of this difference in behaviour that the distinction can be drawn between diamagnetism and paramagnetism.

Diamagnetism

In a diamagnetic material, the atoms have no resultant magnetic moment due to spin, because all the electrons are paired off. Diamagnetic properties are attributed to the orbiting motion of the electrons, and in the absence of a magnetizing field there is no magnetic moment on this account either. When the substance is subjected to a powerful magnetic field, the spin magnetic moments remain paired off, but the orbital motion of the electrons is so affected by electromagnetic induction that, in accordance with Lenz' law, the field is slightly reduced, making μ_r slightly less than unity.

Diamagnetism is independent of temperature.

Paramagnetism

All atoms have this diamagnetic effect, but in paramagnetic materials it is outweighed by their paramagnetic response, which is due to electron spin. The electrons of a paramagnetic atom are not all paired off, and each individual atom possesses a resultant magnetic moment which is due to the spin of the unpaired electrons. The atoms contained in a finite sized specimen have their spin moments so orientated relative to each other that the overall magnetic moment is zero. When

subjected to a powerful magnetizing field, however, the atomic magnets tend to align themselves with the field, thus increasing the overall flux density. This effect is large enough to outweigh the diamagnetic response that also occurs, but it is prevented from being very strong by the thermal agitation of the atoms. The paramagnetic effect is consequently greater at low temperature, being described by the relation

$$(\mu_r - 1) \propto \frac{1}{T}, \tag{16.12}$$

where T is the Kelvin temperature.

Ferromagnetism

In the ferromagnetic materials, iron, cobalt, and nickel, the individual atoms are paramagnetic, having unpaired electrons providing magnetic moment due to their spin. In their case, however, there is an additional influence which causes groups of atoms to have their individual moments lined up with the same polarity, and the unmagnetized specimen consists of these strongly magnetized atomic groups, which are known as *domains*. Because the domains have random orientations, the specimen as a whole i sinitially unmagnetized. When a magnetizing field is introduced, and progressively increased in strength, the result is as shown in Figs. 16.4 and 16.15. The increasing magnetization of the specimen is the result of two distinct factors: in the weaker fields, the domains which have polarity favourable to the field grow in extent, at the expense of domains which have not. In the stronger magnetizing fields, the domains as a whole rotate their magnetization to lie along the field until, at saturation, all the individual atomic magnetic moments, or nearly all, are aligned with the field. The rotation of the domains is more difficult to reverse than their growth in size, so that hysteresis is more pronounced using strong fields.

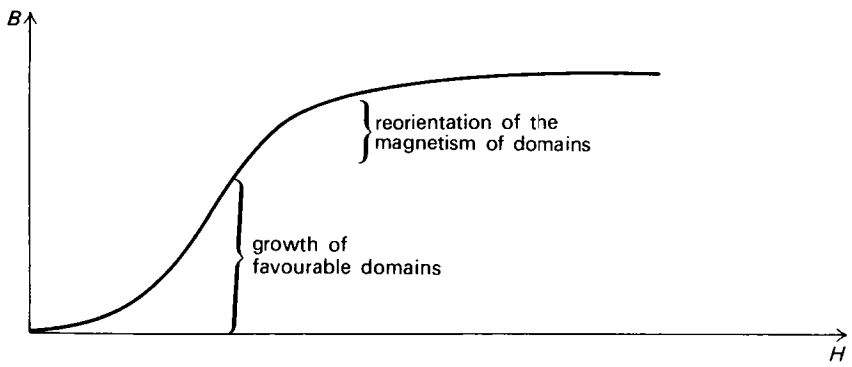


Fig. 16.15 The domain theory of ferromagnetism

A ferromagnetic material loses its tendency to form domains when it is heated to the Curie temperature, with the result that, above this temperature, it becomes paramagnetic.

The domain theory of ferromagnetism has some direct evidence to support it. At the boundary between two neighbouring domains, they produce a magnetic field which is available to exert an external influence. If the metal surface is highly polished and covered with very fine grains of magnetic material in suspension in a liquid, these grains can approach sufficiently close to the domain boundaries to

experience the field, and they cling to the surface at these boundaries, revealing their existence to a powerful microscope. This effect was first observed by Bitter in 1932, and the patterns obtained in this way are known as *Bitter patterns*.

While a ferromagnetic substance is being subjected to a growing magnetic field, the resultant flux does not increase entirely regularly, but is subject to a sequence of small but abrupt changes, due to the processes by which the domains change. An effect of this, which can be demonstrated, is called *the Barkhausen effect*. Suppose that a specimen of steel wire is placed in a solenoid which is connected to a high gain amplifier and a loudspeaker. If a strong magnet is slowly brought up towards the wire, the magnetization of the wire slowly increases. The sequence of abrupt flux changes which this involves causes clicks to be heard in the loudspeaker, owing to electromagnetic induction.

Antiferromagnetic materials

The unpaired electrons which are responsible for the paramagnetism of the individual atoms of a ferromagnetic element such as iron belong to the electron shell with energies directly below those of the valency shell, and not to the valency electrons themselves. Consequently the element retains its paramagnetism when present in a compound. In oxides of iron, the iron atoms exert influences on each other, causing the spin moments in neighbouring atoms to be aligned, just as they do in the metal, but in this case it is an alignment which causes the magnetic moments of alternate atoms to be in opposition, thus destroying the overall magnetism. Such substances are said to be antiferromagnetic.

Ferrimagnetic materials

Ferrites, being composite oxides, also have this antiparallel alignment, but the magnetic moments which point in one direction are stronger than the moments which oppose them, the effect depending on the particular crystal structure. These substances are termed ferrimagnetic. They form domains, in the same way as iron does, but are not so powerfully magnetic.

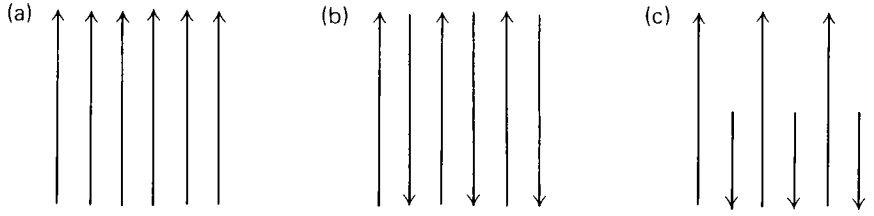


Fig. 16.16 Diagram representing the alignments of neighbouring spin moments in a domain (a) Ferromagnetic (b) Antiferromagnetic (c) Ferrimagnetic

Alloys

The magnetic properties of a ferromagnetic material are strongly influenced by discontinuities and imperfections in the crystal lattices, with the result that alloying and heat or work treatment can alter them. Steels which are mechanically hard have domains which cannot easily grow or rotate their polarity, which explains why they are difficult to magnetize or demagnetize, and so make permanent magnets. Likewise, the opposite is true of soft alloys.

Paramagnetic oxygen

A similar distinction to that between ferromagnetic and antiferromagnetic materials is to be found between molecules. Both the hydrogen and the oxygen atom, for

example, are basically paramagnetic, having electrons with unpaired spin moments. In the hydrogen molecule, however, the spin moments of the two electrons of the two hydrogen atoms are aligned in opposition to each other, whereas in the oxygen molecule the two unpaired electrons act in the same direction. Consequently, at very low temperature, when the effect of thermal agitation is less pronounced, oxygen is strongly paramagnetic.

16.14 Nuclear spin moment

The magnetic moment of the electrons accounts for the magnetic properties of the atom in all ordinary circumstances, but in fact the protons in the nucleus each possess a slight magnetic moment too, associated with spin, and so do the neutrons, despite the fact that they have no electric charge. The effect, in both these cases, of the larger mass is to make the magnetic moment many times smaller than it is for an electron, and for this reason it was correct to describe diamagnetism and paramagnetism in terms of the extra-nuclear electrons only; the contribution made by the nucleii of the atoms would be negligible. Nevertheless, nuclear magnetism has some extremely important practical applications, which, however, are outside the scope of this book.

*16.15 The magnetic circuit

It is important to the electrical engineer to be able to calculate and map the flux caused in an iron core by a magnetizing coil, and in most cases there are air gaps to complicate the issue. One theoretical device which is used to solve the problem is the magnetic circuit, which treats magnetic flux as if it were something which lives up to its name and actually flows through the material. The method seems to the writer to be empirical, and the results obtained by using it can only be approximate, particularly since the relative permeability μ_r of iron is not a constant quantity. But the method is a good deal better than none.

The cause of the flux is taken as being the product of the number N of turns of the magnetizing coil and the current I which flows in them. The product NI is termed the magnetomotive force (or m.m.f.) F_m and treated as being analogous to e.m.f. in an electric circuit, the flux Φ being taken as analogous to the current in a circuit. For each part of the magnetic 'circuit', electrical resistance is replaced by the magnetic reluctance R_m which is defined by the equation

$$R_m = \frac{1}{\mu_r \mu_0} \cdot \frac{l}{A},$$

i.e.

$$R_m = \frac{1}{\mu} \cdot \frac{l}{A}. \tag{16.13}$$

This is regarded as corresponding to the electrical resistance equation

$$R = \rho \frac{l}{A},$$

where ρ is the electrical resistivity of the material.

The application of the magnetic circuit method may be illustrated by taking the case of an iron toroid which has an air gap in it.

Suppose that the total length round the toroid is l, the air gap having length x (taking an average in both cases). The total reluctance of the magnetic circuit is

$$R_m = \frac{1}{\mu_r \mu_0} \cdot \frac{(l-x)}{A} + \frac{1}{\mu_0} \cdot \frac{x}{A}.$$

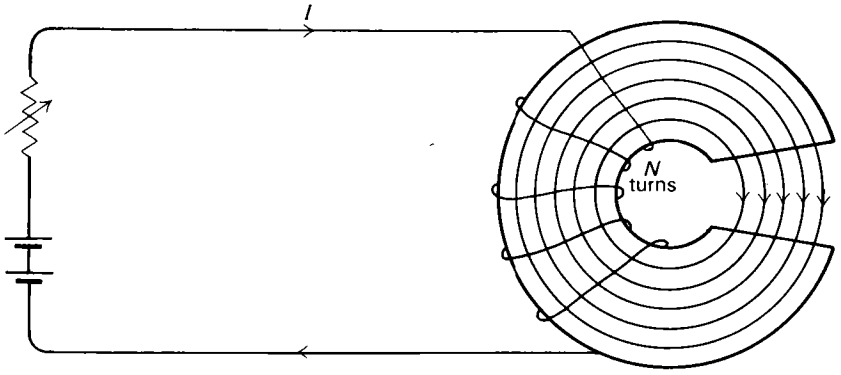


Fig. 16.17 A magnetic circuit

This assumes that the flux does not spread out appreciably in the air gap, but is restricted to area A, which is the cross-sectional area of the iron. Using a magnetic equation which is analogous to the equation of the electrical circuit

$$I = \frac{V}{\Sigma R},$$

we have

$$\Phi = \frac{F_m}{R_m},$$

i.e.

$$\Phi = \frac{NI}{R_m},$$

whence

$$\Phi = NI\,\frac{\mu_r \mu_0 A}{(l - x) + \mu_r x}. \qquad (16.14)$$

In the simple case where there is no air gap, i.e. where $x = 0$, this reduces to the familiar form

$$\Phi = \mu_r \mu_0 A \frac{NI}{l},$$

i.e.

$$B = \mu_r \mu_0 \frac{N}{l} I.$$

There is the difficulty, of course, already mentioned, that the value of μ_r is difficult to estimate, depending, as it does, on the value of I. Equation (16.14), however, is a useful answer and it does, for example, illustrate the disproportionate effect of even a small air gap in reducing the amount of flux. This may be regarded as analogous to the effect of inserting a high value resistor in series with a number of low ones in a circuit.

QUESTIONS

16.1 An iron ring has average diameter of 10 cm, and the area of cross-section of the iron is 1.2 cm^2. A solenoid with 2500 turns wound round the ring carries a current of 0.5 A. If the total magnetic flux in the iron ring in these conditions is 6×10^{-3} Wb, deduce the relative permeability μ_r of the iron.

16.2 A primary coil of 500 turns is wound uniformly on an iron ring of mean radius 10 cm, and cross-sectional area 4 cm^2. A secondary coil of 50 turns is wound over the primary and connected to a ballistic galvanometer, the total resistance of the secondary circuit being 500 Ω. When a current of 2 A is reversed in the primary the galvanometer indicates a throw of 10 divisions. When a capacitor of capacitance 2 μF is charged to a p.d. of 20 V and discharged through the same galvanometer, the throw obtained is 5 divisions. Calculate the relative permeability of the iron. (S)

16.3 An iron ring of mean circumference 60 cm and cross-sectional area 8 cm^2 is wound uniformly with a primary coil of 240 turns. A secondary coil containing 10 turns of insulated wire is wound over the primary and is connected to a ballistic galvanometer, the total resistance of the secondary circuit being 200 Ω. When a current of 2.0 A, flowing in the primary, is reversed, the galvanometer gives a throw of 60 scale divisions. When a capacitor of capacitance 2.0 μF, charged to a p.d. of 25 V, is discharged through the galvanometer, the throw is 40 scale divisions.

Assuming that the throw of the galvanometer is proportional to the quantity of electricity discharged through it, calculate (*a*) the flux density in the iron, (*b*) the flux density which would be produced by the primary current in the absence of the iron, (*c*) the relative permeability μ_r of the iron. (O & C)

*16.4 A specimen of steel in the form of a ring of radius 10 cm and cross-sectional area 2 cm^2 has a magnetizing coil wound around it. A second coil of 10 turns is also wound round the specimen, and connected to a ballistic galvanometer. The specimen is originally unmagnetized, and when a current I is switched on in the magnetizing coil, a charge of 30 μC flows through the galvanometer. The current is then reversed, and a charge of 58 μC flows in the opposite direction. When the current is finally switched off, a charge of 17 μC flows in the same direction as when the current was first switched on.

Explain these observations, and estimate the final flux density in the steel if the total resistance of the ballistic galvanometer circuit is 100 Ω. (O & C)

*16.5 A mutual inductance consists of two complete toroids of fine wire wound closely and uniformly on a non-magnetic insulating former. The primary coil has 500 turns and the secondary coil 100 turns; each coil has a cross-sectional area 1 cm^2 and the mean radius of the toroid is 10 cm. The mutual inductance between the two windings is 10 μH. When a current of 1.0 A is reversed in the primary winding, a ballistic galvanometer connected across the secondary registers a throw of 20 divisions. When the primary and secondary circuits are interchanged, reversing a current of 1.0 A in the 100 turn winding produces a throw of 10 divisions. Assuming that the galvanometer resistance may be neglected, what can you deduce about the resistances of the windings?

A circular loop of nickel wire of cross-sectional area 1 mm^2 and radius 10 cm is mounted on the axis of the toroids and the observations are repeated. When a current of 1.0 A is reversed in the primary (500 turns) the galvanometer registers a throw of 100 divisions from the secondary, and when the circuits are interchanged the reversal of 1.0 A in the 100 turn winding produces a throw of 90 divisions. What can you deduce from these observations about the magnetic properties of nickel? (O & C)

*16.6 Describe with full experimental detail how you would attempt to obtain a normal magnetization curve for a specimen of iron, showing clearly how the calculations would be carried out, and sketching the form of curve expected. (S)

*16.7 Explain what is meant by cycle of magnetization, and by hysteresis.

Sketch on the same diagram the hysteresis loops for soft iron and hardened steel, indicating on the axes the physical quantities that have been plotted. What information of practical importance can be obtained from a hysteresis loop? State, with reasons, which of the above metals would be suitable for (*a*) a permanent magnet, (*b*) the core of a transformer.

Describe, with circuit diagram, an effective electrical method of demagnetizing a steel bar magnet and explain what is happening during the process of demagnetization. (L)

*16.8 An iron ring, of mean radius 5.0 cm, is wound with several thousand turns of wire. If the relative permeability of the iron may be taken as having a constant value of 2000, by what percentage would the flux density in the iron be changed if an air gap of width 0.1 cm were made in the ring, the current flowing in the coil remaining constant? (A.E.B.)

*16.9 Define 'relative permeability' of a magnetic substance. Indicate the orders of magnitude of the relative permeabilities of ferromagnetic, paramagnetic, and diamagnetic substances respectively. Discuss in relation to its relative permeability one aspect of the behaviour of a specimen of each of these substances when placed in the same magnetic field.

An iron ring of mean circumference 30 cm and area of cross-section 1.5 cm^2 has 240 turns of wire uniformly wound on it, through which passes a current of 2.0 A. If the flux in the iron is found to be 7.5×10^{-4} Wb, find the relative permeability of the iron. (L)

Chapter 17

The absolute determination of current and resistance

17.1 The absolute determination of electric current

Although the practical units of current, potential difference, resistance, and so on were defined early on in this book, no means have yet been described to apply these definitions to actual measurement. Yet to do so is clearly of fundamental importance: when an ammeter registers a current of one ampere, we must be sure that the current flowing is what the definition intends that one ampere should be.

In the case of the ampere, there is an initial difficulty: the system of two parallel conductors, in terms of which the ampere is defined (section 1.11), is not an arrangement which is suited to accurate practical measurement. We have now developed a sufficient number of equations, however, which apply to other arrangements of conductors and are consistent with that for the parallel conductors, so that we are now in a position to select a more practical arrangement and to make an absolute determination of current with that.

The emphasis will be placed in this chapter on apparatus suitable for use in the school laboratory, which will serve to illustrate the principles involved. The accurate determination of an electric current in terms of its definition is, however, the foundation of practical electricity, and brief reference to accurate apparatus will be made.

17.2 The dynamometer

This is a small coil pivoted at the centre of a large one, in such a way that the planes of the two coils are perpendicular to one another. If the same current is made to flow in both coils by connecting them in series, a torque Γ turns the small coil, tending to bring it into the same plane as the large one. This torque may be balanced out by the effect of the weight mg of a small rider which is moved along a balancing arm to prevent the coil from turning at all. Because the planes of the

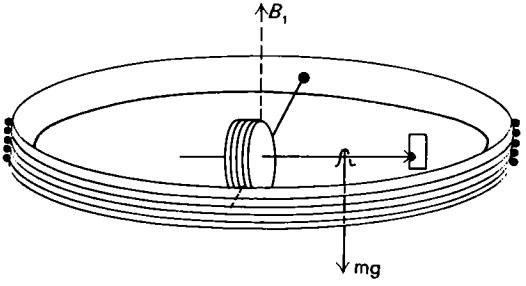


Fig. 17.1 A simple dynamometer

two coils are made in this way to retain the angle 90° between each other, we can write

$$\Gamma = B_1 N_2 A_2 I, \tag{17.1}$$

where B_1 is the flux density at the centre of the large coil due to the current flowing in it; N_2, A_2, I are, respectively, the number of turns and the area of the small coil, and the current flowing in it.

Also,

$$B_1 = \mu_0 \frac{N_1}{2r_1} I, \tag{17.2}$$

where N_1, r_1, I are the number of turns, radius, and current of the large coil.

Eliminating B_1 from these two equations gives

$$\Gamma = \mu_0 \frac{N_1 N_2 A_2}{2r_1} I^2, \tag{17.3}$$

where

$$\Gamma = mgx, \tag{17.4}$$

x being the distance of the rider from the pivot of the small coil. The final equation for the apparatus becomes

$$mgx = \mu_0 \frac{N_1 N_2 A_2}{2r_1} I^2 \tag{17.5}$$

and this enables the current to be measured without reference to any other electrical measurement; i.e., this is an absolute determination of current, measured in amperes.

The value to be inserted for μ_0 is, of course,

$$\mu_0 = 4\pi \times 10^{-7} \text{ H/m},$$

and all other quantities m, g, x, A_2, r_1 are in SI units.

[We might note, in passing, that, although it has just been specified that the unit used for measuring length must be the metre, this was done only to stress the fact that the ampere is an SI unit. To be accurate about it, the only essential thing is to measure the weight mg in newtons. This apparatus belongs to the type which was suggested by Ampere (section 15.13) to test the inverse square law of force between current elements; the system would require the same weight rider to balance the effect of a particular current if the scale of the whole apparatus, including the distance of the rider to the pivot, were increased n times. This follows from equation 17.5, in which the dimensions of length cancel out.]

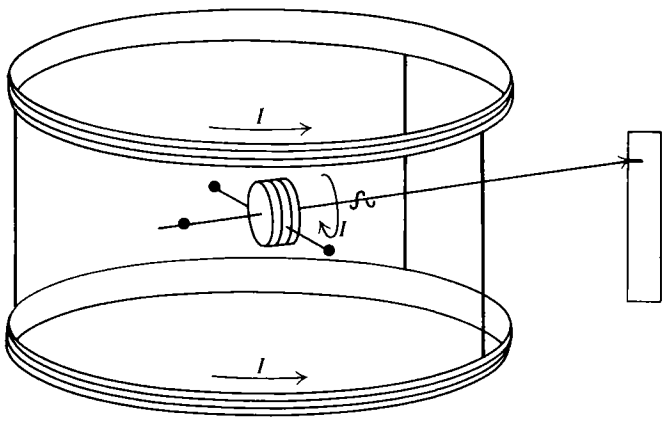


Fig. 17.2 A type of dynamometer using a Helmholtz system of coils

The apparatus shown in Fig. 17.1 can be improved upon by using a Helmholtz pair of coils (section 15.10) in place of the single large coil. In this way, a region of uniform flux density which is large enough to contain the pivoted coil and consequently to provide greater accuracy is produced. This alters the equation for B_1, of course, which becomes equation 15.21, instead of 17.2.

Accuracy is further improved by allowing for the effect of the vertical component of the earth's field, which is acting in the same, or the opposite direction, as the field of the large coils, and so contributes towards producing the torque. This effect may be eliminated by taking every reading a second time, with the currents in all three coils reversed, an operation which reverses the torque caused by the earth's field without reversing that due to the current. The average of the two measured torques then gives the correct values for insertion in the equation.

17.3 The measurement of r.m.s. alternating current

Equation 17.3 shows that the deflecting torque Γ of the dynamometer is proportional to the square of the current. The instrument may therefore be used to measure the r.m.s. value of an alternating current, quite independently of the frequency and the waveform of the current. There are basically two ways in which this can be translated into practical terms. An absolute determination of the alternating current may be made by applying equation 17.5. Alternatively, the dynamometer may first be used in series with a d.c. source and a d.c. ammeter; a calibration curve can then be drawn showing how the distance x of the rider from the pivot varies with the reading of the ammeter. If alternating current is now used instead, the calibration curve can be used for reading off the r.m.s. value of the current. Provided that the d.c. meter is accurate, the second method can be used to produce accurate results, even when the accuracy of an absolute determination would be suspect.

*17.4 The current balance

The dynamometer conveniently illustrates the principle of an absolute current determination because its equation is simply derived, but, in fact, an accurate instrument follows the slightly different pattern of a current balance. This measures a linear force between two coils, instead of a torque, as the dynamometer does, and so it may be said to measure the current by weighing its effect.

If a coil which carries current is to experience a linear force when placed in a magnetic field, there must be a gradient $\mathrm{d}B/\mathrm{d}x$ in the flux density (section 15.18), where distance x is measured in the direction of the flux. If the coil is then placed with its axis pointing along the flux, the force F is

$$F = m\frac{\mathrm{d}B}{\mathrm{d}x}, \tag{17.6}$$

where m is the electromagnetic moment of the coil,

i.e.
$$m = NAI, \tag{17.7}$$

N, A being the number of turns wound on the coil, and the area enclosed by them.

The field coils must be so arranged that they provide a uniform gradient in the flux density over a region which is sufficiently large to contain the coil which experiences the force. For such an arrangement the equation takes the form

$$\frac{\mathrm{d}B}{\mathrm{d}x} = k\mu_0 I \tag{17.8}$$

where k is a constant dependent on the exact configuration of coils adopted; its value may be determined from measurements such as the radius, number of turns, etc.

Combining equations 17.6, 17.7, 17.8 gives

$$F = k_1 \mu_0 N_2 A_2 I^2, \tag{17.9}$$

where the subscripts 1, 2, refer to the field coil and the balanced coil respectively, and I is the current, which is common to both.

A suitable design for the school laboratory is shown in Fig. 17.3.

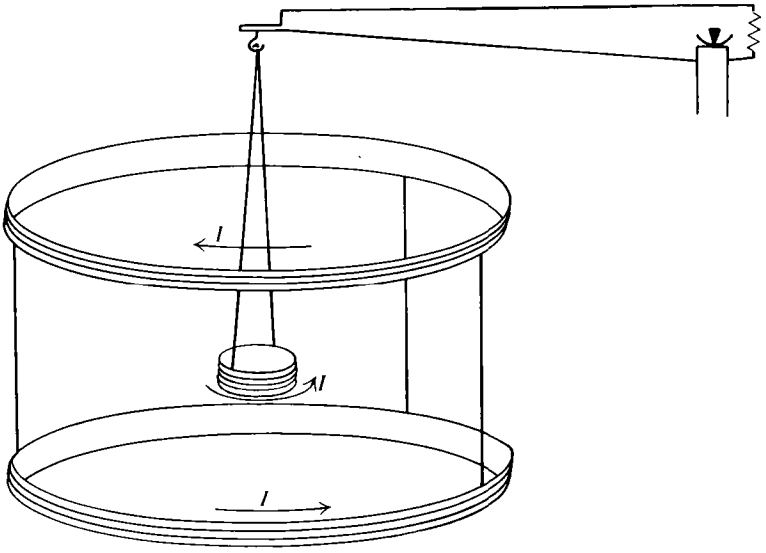


Fig. 17.3 A simple form of current balance

The field coils have the geometry of a Helmholtz system, being a pair of horizontal circular coils which are separated from each other by a distance equal to their common radius. The difference from the Helmholtz system is that the current I circulates in opposite directions in the two coils. This has the result that the magnetic flux along their common vertical axis is directed along this axis, but varies rapidly in strength, i.e. there is a steep gradient in flux density dB/dx. The separation r of the coils is chosen because it produces the required region where this gradient is uniform; this is illustrated in Fig. 17.4, which might be compared with Fig. 15.13.

It is not difficult to derive the equation which gives the value of the constant k, but it is not worth while to do so here.

The sensitivity of this apparatus may be doubled by having a similar arrangement for the other arm of the balance, with the currents so directed that the magnetic interaction makes the coil suspended from one arm heavier, and that from the other arm lighter; this also eliminates the effect of the vertical component of the earth's field.

The instruments described above, both dynamometer and current balance, are designed merely to illustrate principles. In a proper standardizing instrument, of course, accuracy is the important thing, and whether or not the final equation is more complicated is almost irrelevant. Indeed, it is fundamental to electric

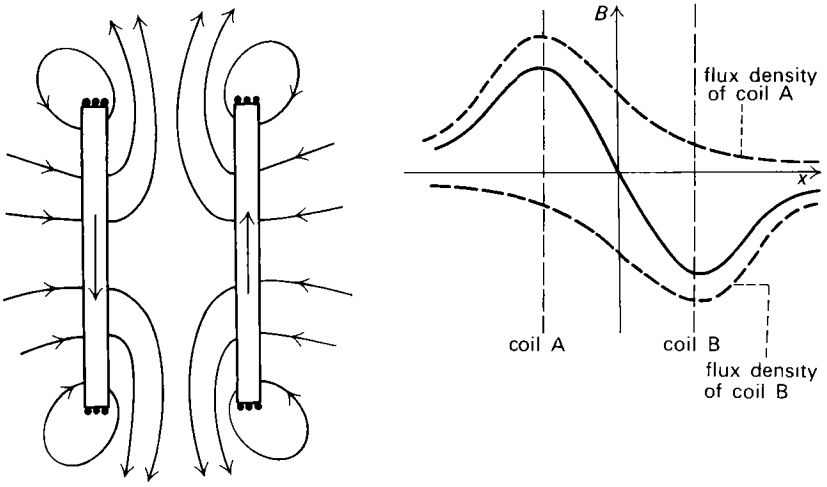


Fig. 17.4 The field coils of one form of current balance

measurement to have an accurate current balance available, because all the SI units in electricity, which are the practical units, are based on the definition, and therefore on the measurement, of the ampere. The flat coils used in the designs described above would need to have more than a single layer of wire, and this would reduce the precision with which the dimensions of the apparatus can be measured. For such reasons as this, the accurate form of current balance has single layer coils wound on coaxial cylinders instead; the accuracy obtained is something like one part in 10^5.

17.5 The absolute determination of resistance

Having made an absolute determination of current, a p.d. might then be measured in terms of the definition of the volt, by determining the relevant quantities for insertion in the defining equation

$$V = \frac{W}{Q},$$

i.e.

$$V = \frac{W}{It}. \tag{17.10}$$

Using SI units, energy W is measured in joules, current I in amperes, and time t in seconds, giving the p.d. in volts. Because the current could in effect be determined by using a current balance, this would be an absolute measurement of the p.d.

The most obvious experiment to carry out would be to insert a resistor in a material of known specific heat capacity c, and to measure the heat released in a given time as the result of passing a measured current. Equation 17.10 would then become

$$V = mc\frac{(T_2 - T_1)}{It} \tag{17.11}$$

where m, c are the mass and specific heat capacity of the material, and $(T_2 - T_1)$ is its rise in temperature, m being in kg, c in J/(kg·K) and T_2, T_1 in K.

Alternatively to measuring p.d., this experiment could, of course, be used to make an absolute determination of the resistance of the heating resistor, putting the equations in the form

$$R = \frac{W}{I^2 t},$$

whence

$$R = mc\frac{T_2 - T_1}{I^2 t}. \tag{17.12}$$

Experiments of this kind are more normally carried out with the object of measuring the specific heat capacity c. Nevertheless, this can alternatively be done independently of electrical measurements, the heat being generated by conversion from mechanical energy. Having determined the specific heat capacity c by such a means, the electrical measurement becomes a legitimate means of making the absolute determination of p.d. or of resistance, as suggested above. In practice, however, heating experiments are not capable of very great accuracy, and an alternative method is chosen instead.

This alternative is based on the equation of electromagnetic induction

$$V = \frac{\mathrm{d}\Phi}{\mathrm{d}t}. \tag{17.13}$$

This equation was derived directly, in section 14.8, from the definition of p.d., using conservation of energy, and so it provides a legitimate starting point. The principle of making an absolute determination of p.d. measured in volts consequently becomes that of using some form of dynamo. This will presently turn out to be an absolute determination of resistance instead, measured in ohms, but this is equally satisfactory.

To obtain an e.m.f. which may be calculated, using equation 17.13, without reference being made to any other electrical measurement, except, perhaps, that of an electric current, what is required is to produce a uniform magnetic flux in such a way that the flux density may be calculated from measurements of the apparatus used to produce it. Precision apparatus of this kind is based on a method devised by Lorenz, and an experiment suitable for the school laboratory which shows the principle of Lorenz's method will now be described.

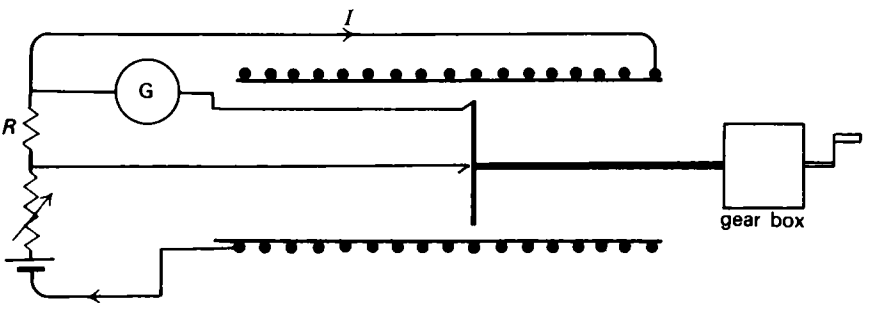


Fig. 17.5 The absolute determination of resistance

In this apparatus, the uniform flux is produced by using a long straight solenoid, and the e.m.f. is induced in a Faraday's disc inserted in the middle of it. We have

$$B = \mu_0 \frac{N}{l} I \tag{17.14}$$

and

$$V = BAf, \tag{17.15}$$

whence
$$V = \mu_0 \frac{N}{l} IAf, \tag{17.16}$$

where N is the number of turns in length l of the solenoid, A is the area of the disc, and f is the number of revolutions made by the disc in unit time. Using SI units, this gives the p.d. between axle and circumference of the disc, measured in volts.

Equation 17.16 could be used as it stands, the need to measure current I being accepted because its absolute determination has already been carried out. The rotating disc would then be regarded as a standard source of a small e.m.f.

It is more convenient, however, to balance the p.d. V across the radius of the disc against the p.d. across a resistance R which is connected in series with the solenoid. We then have

$$V = RI$$

and
$$V = \mu_0 \frac{N}{l} IAf,$$

whence the current I cancels out, giving the equation

$$R = \mu_0 \frac{N}{l} Af, \tag{17.17}$$

from which an absolute determination of the resistance can be made. The procedure is to turn the disc at the appropriate frequency f to balance the galvanometer.

Taking typical readings which might be obtained using a simple form of this apparatus, we might have

$$\frac{N}{l} = 5 \textit{ turns}/\text{cm} = 500 \textit{ turns}/\text{m},$$

$$A = 75 \text{ cm}^2 = 75 \times 10^{-4} \text{ m}^2 \text{ (corresponding to a disc radius of about 5 cm)},$$

$$f = 15 \textit{ rotations}/\text{s},$$

whence, using equation 17.17,

$$R = 4\pi \times 10^{-7} \times 500 \times 75 \times 10^{-4} \times 15 \ \Omega,$$

i.e.
$$R = 7 \times 10^{-5} \ \Omega, \text{ approximately.}$$

This value of resistor is too small to be useful directly, and a means is required to compare it accurately with a resistance which is considerably larger. This may be done by inserting a suitable resistance network in place of the single resistor R in the circuit, as shown in Fig. 17.6.

A resistance R of reasonably high value, say about 1.0 Ω, is connected in series with the solenoid, and the p.d. across this resistor is reduced to a suitably low value for comparison with the p.d. across the Faraday's disc by means of the potential dividing arrangement, R_1 and $10^4 R_1$, in series. If the resistance of this potential divider is sufficiently high, the current flowing in it can be neglected, and we can write

$$V_{PQ} = 10^{-4} V_{AB}$$
$$= 10^{-4} RI,$$

whence
$$10^{-4} R = \mu_0 \frac{N}{l} Af. \tag{17.18}$$

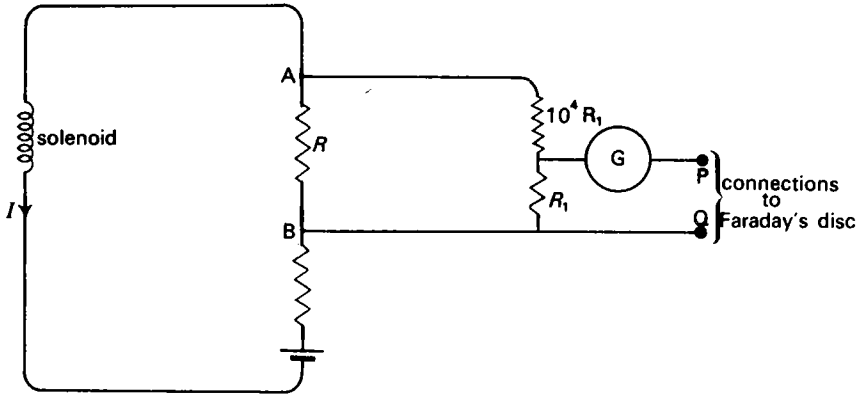


Fig. 17.6

17.6 The preservation of standards

It is not possible to preserve a standard current, obviously. The procedure is to check the value of a resistor, using an accurate form of the measurement just described. A current is then sent through this standard resistor, and adjusted in value until the p.d. across it balances the e.m.f. of a standard cell. This current I can then be measured by means of a current balance, which enables the e.m.f. of the standard cell to be deduced:

$$V = RI.$$

The standards can then be preserved, in the standard resistor and the standard cell.

QUESTIONS

$$\frac{\mu_0}{4\pi} = 10^{-7}\ \text{H/m}$$

17.1 A long solenoid wound with 20 turns of wire to every cm carries a current of 4.0 A. A square coil, side 2.0 cm, with 40 turns, is suspended inside the solenoid. Calculate the maximum torque exerted on this coil if the current in it is 0.25 A, and give a diagram showing the position of the coil when it experiences the maximum torque.

Discuss whether the torque on the coil would have the same mean value in the same position in the solenoid, if the currents in the solenoid and coil were both alternating at the same frequency, with r.m.s. values 4.0 A and 0.25 A respectively.

17.2 Explain how an absolute determination of resistance may be made by using the Lorenz disc apparatus shown in figure e 78.

The solenoid is 1 m long, has 2000 turns, and carries current 2.5 A. What is the flux density B inside the solenoid?

The brass disc rotates at a steady speed of 600 rev/min; its radius is 2.0×10^{-2} m, and the radius of the axle is 2×10^{-3} m. What is the e.m.f. generated? If, at this speed, the sensitive galvanometer reads zero, find the value of the resistance R. (O)

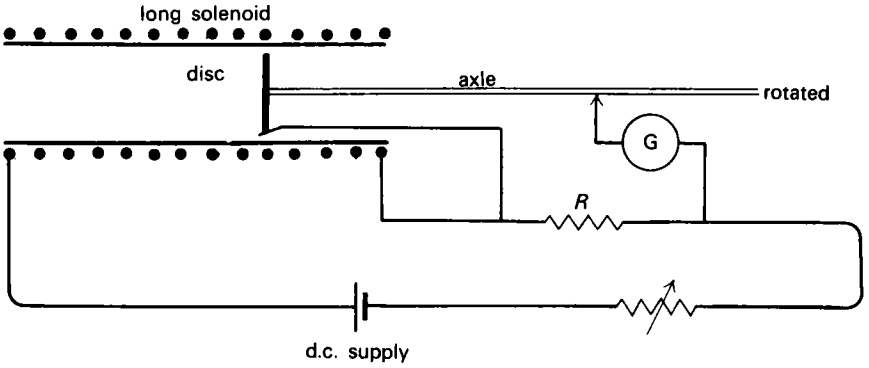


Fig. e 78

17.3 A copper disc of diameter 20 cm is arranged coaxially and centrally inside a long solenoid wound with 8 turns/cm, carrying a current I. Calculate, in terms of I, the induced e.m.f. in each radius of the disc when it is being turned about the axle at 10 rev/s. (Neglect the radius of the axle.)

It is found that this e.m.f. will just balance the p.d. across the resistor R_2 in the network shown in diagram e 79, which is joined in series with the solenoid. Assuming that $R_1 \ll R_2$, deduce the value of R_1.

17.4 A flat circular coil, placed with its plane horizontal, has radius 20 cm and is wound with 500 turns. A second, small coil, with 500 turns, enclosing area 3 cm^2, is pivoted at the centre of the first with its axis horizontal, i.e. perpendicular to the axis of the first coil. A current is sent through the two coils in series, with the result that a torque Γ acts on the small coil, where $\Gamma = 1.2 \times 10^{-5}$ N·m. Deduce the value of the current.

The vertical component of the earth's magnetic flux density is 4×10^{-5} T, and this has been neglected. Does this cause an appreciable error?

Why would the apparatus be more accurate if the large coil were replaced by two large coaxial coils, each of radius 20 cm, placed 20 cm apart?

17.5 A circular coil of radius 15 cm was mounted with its plane horizontal and another circular coil of 100 turns each of radius 2.0 cm was pivoted at its centre so that it could turn about a diameter of the large coil. The small coil carried a light counterbalanced pointer fixed at right angles to its plane. The coils were connected in series and a current passed through them. A rider of mass 0.04 g was adjusted on the pointer to keep the plane of the small coil vertical. The following results were obtained for d, the distance from rider to pivot, and N, the number of turns in the large coil.

d/cm	14.7	13.4	12.0	10.6	9.3	8.0
N	100	90	80	70	60	50

Represent these results graphically, and use the graph to calculate the current in the coils and the strength of the earth's vertical component. State clearly any formulae used in the calculation, and list the probable sources of error in this experiment.
(Take $g = 9.8$ m/s^2.) (S)

17.6 Describe an absolute method of measuring electric current. Give a reasoned estimate of the order of magnitude of current which is most suitable for measurement with the apparatus you describe.

If you were provided with this apparatus together with one standard resistor, the value of which had been determined by absolute measurement, and such uncalibrated batteries, rheostats, and galvanometers as you required, how would you attempt to measure the e.m.f. of a standard cell in absolute volts?

Discuss briefly whether the earth's magnetic field would have any effect on the measurements you have described. (O & C)

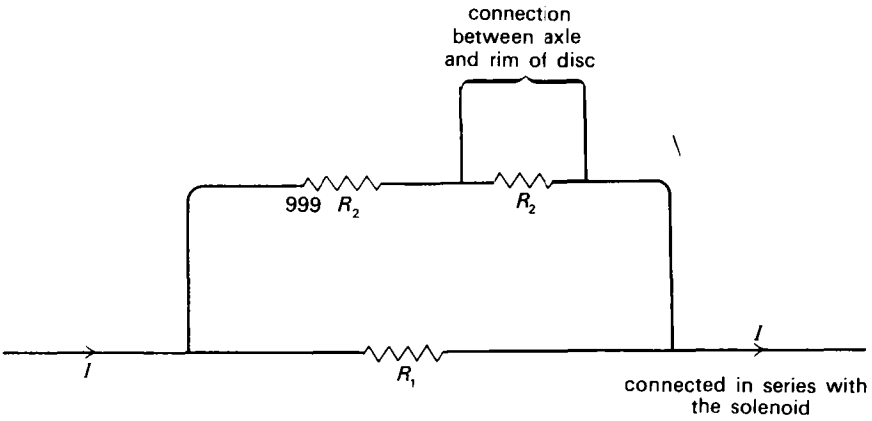


Fig. e 79

Chapter 18

Current measuring instruments

18.1 The moving coil meter

When it comes to the everyday measurement of a current or p.d., where absolute determination is not a relevant consideration, the most popular instrument is the moving coil meter. The basic instrument is a milliammeter, or, at its most sensitive, a microammeter, and a shunt or a series resistor is added, as the case may be, to make an ammeter or a voltmeter (sections 3.14 and 3.15).

[All measuring instruments which are based on an electromagnetic force are primarily current measuring devices, because a magnetic field is the result of a current. The alternative of measuring an electrostatic force, as in the gold leaf electroscope, produces a voltmeter, because an electrostatic field is a potential gradient. Electrostatic instruments, however, although they have the merit that they draw no current, are either insensitive or difficult to use, and we will not consider them further.]

The action of the moving coil meter is based on the torque produced when current flows in a coil placed in a magnetic field. The current is measured by using this torque to twist an elastic suspension or spring, and so the current is, in effect, weighed by the spring. This is similar to the action of the dynamometer, with the difference that, in the moving coil meter, the field is the strong flux density between the poles of a permanent horseshoe magnet, and the effect of external fields, such as that of the earth, is negligible.

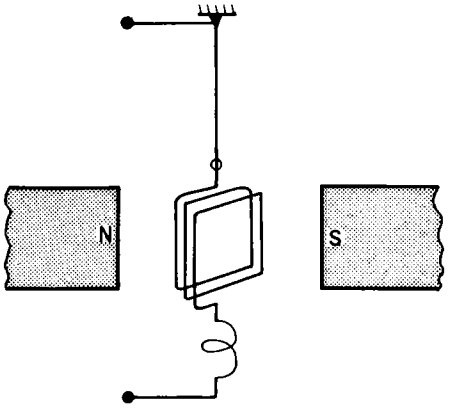


Fig. 18.1 The basic construction of a moving-coil meter (suspended type)

Assuming that the plane of the coil is parallel to the flux of the magnet so that the deflecting torque has its maximum value $BNAI$, we can write

$$BNAI = k\theta, \tag{18.1}$$

where θ = the angle through which the elastic suspension or spring is twisted,
k = a constant, which measures the elastic 'stiffness' of the suspension or spring,
B = the flux density of the magnet,
N = the number of turns of the coil,
A = the area enclosed by the coil.

Thus, the current sensitivity $S_I = \theta/I$ is given by

$$S_I = \frac{BNA}{k}. \tag{18.2}$$

Two types of instrument are illustrated in Fig. 18.2. In the free suspension type, the coil is freely suspended from a fine conducting wire or ribbon, usually made from phosphor bronze on account of its good elastic properties. This provides both the elastic restoring torque which is used to measure the current and also one of the two connections bringing current to and from the coil, the other being made by means of a loosely wound spiral of silver wire, intended to be torsionless. In the pivoted type, shown in the second diagram, the coil is balanced to revolve about two sharp steel pivots resting in jewelled bearings, and the restoring torque is provided by two fine springs, which are usually coiled flat like the hair spring of a watch (although they are not drawn like this in the diagram).

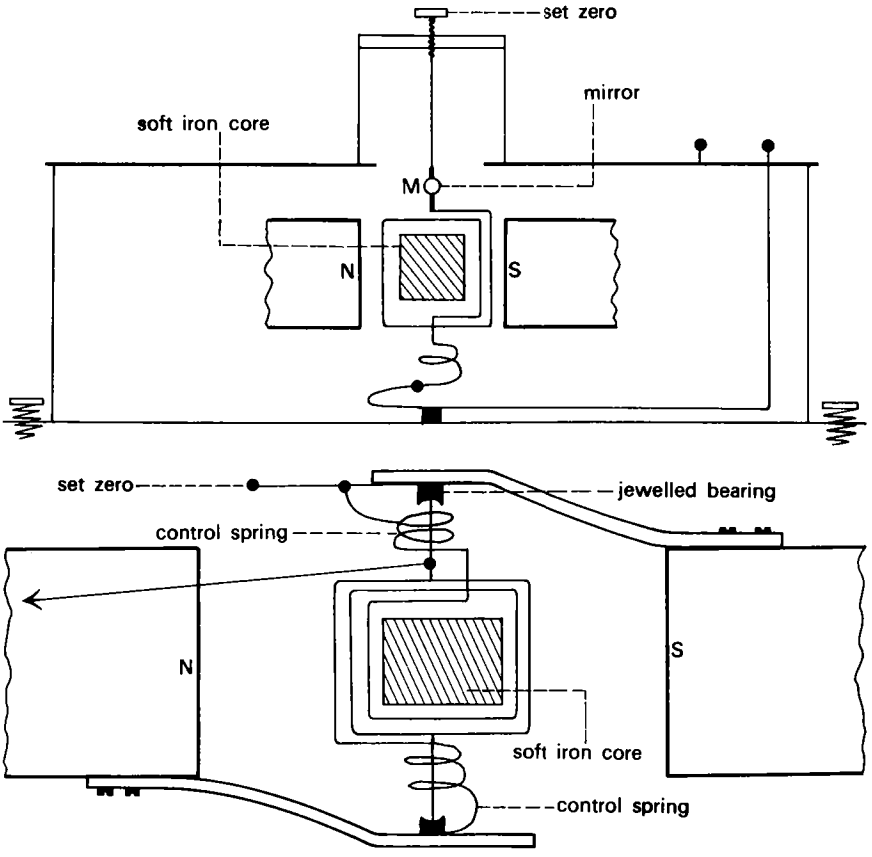


Fig. 18.2 Two types of moving-coil meter (a) Free suspension type (b) Pivoted type

Clearly the meter is more sensitive if the spring is easy to twist, and this is the chief characteristic of the free suspension type of meter. The sensitivity is usually increased still further by using a beam of light as the pointer, since this can be of considerable length; a galvanometer mirror M is fixed to the moving coil, and the light beam is directed on to this.

The pivoted type is more portable, and does not require careful levelling, as the suspension type does, and it is from this that ammeters and voltmeters are made. The basic instrument in this case seldom gives full scale deflection with a current smaller than 100 μA, because a more sensitive pivoted meter than this would be too fragile.

[There is a third type, which is a compromise between the other two, and has a taut suspension both above and below the coil, thus eliminating the levelling up process needed with the free suspension type.]

The linear scale

An advantage of the moving coil meter as a measuring instrument is the fact that it has a linear scale, i.e. the deflection θ is directly proportional to the current I. At least this is so provided that equation 18.2 is correct, and if the factors B, N, A are all constant.

The equation was derived on the assumption that the plane of the coil is parallel to the magnetic flux, thus making the deflecting torque directly proportional to the current and independent of the angle of deflection. For this to remain true, even when the coil is deflected, the flux must be radial; the coil would otherwise make an angle with the flux once it was deflected. The features of the design which produce this radial field are shown in Fig. 18.3; they also fulfil the other necessary condition of making the flux uniform, so that the factor B in the equation is constant. Soft iron pole pieces are added to the magnet to provide a cylindrical cavity which is coaxial with a soft iron cylindrical core, and the coil moves in the small air gap between.

[If all that is required is current detection, and not its measurement, a linear scale is not an essential requirement, and in some cases is deliberately dispensed with, in favour of having a meter which has lower sensitivity when the current is comparatively large. This is an advantage for work with a bridge or a potentiometer.]

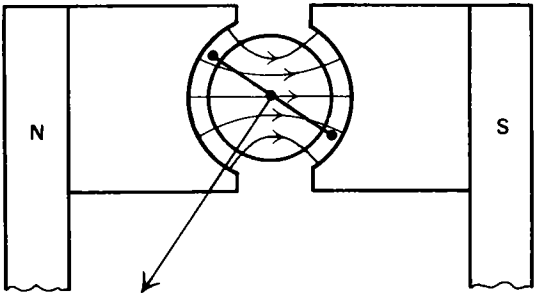


Fig. 18.3 The moving-coil meter – radial field

Accuracy

The accuracy of the meter obviously depends on it being correctly calibrated, but it also helps to have a large scale, and a linear one; this is because the position of the pointer can then be more accurately located and interpreted – half way between two divisions means 0.5 div if the scale is linear, not otherwise.

Sensitivity

It is not as important as it used to be to design galvanometers with the greatest possible sensitivity, because transistor circuits can be designed to act as d.c. amplifiers, to amplify the current before the meter is called upon to detect it. Nevertheless, only a sensitive meter can make a good voltmeter or ammeter, because otherwise the resistance of one is too low, and of the other too high.

To make the current sensitivity (equation 18.2) high, the elastic constant k is made as small as possible, so that the control spring or suspension is very easy to twist, and the flux density B is made as large as possible, which is a reason for making the air gap small. To follow the guidance of the equation further, the area of the coil and the number of turns would be large, but there are practical objections to this, because of the inertia and weight. In practice, the coil is kept small and light, in order to take advantage of having the weakest possible spring; if the coil were large as well, the coil movement would be too slow and the friction at the bearings would become noticeable (or with the suspended type, the suspension would not bear the weight).

[The number of turns could be increased without increase in weight by using thinner wire for the coil, and this would increase the current sensitivity, an advantage for some purposes. But this would be at the expense of having a meter with a higher resistance, and its voltage sensitivity would be low. For some applications, this is more important than current sensitivity, and the connection between the two is

$$S_V = \frac{S_I}{R}, \tag{18.3}$$

where $S_V = \theta/V$, and $S_I = \theta/I$.]

Damping

If it is to be useable, the meter must take up its reading in a reasonably short space of time, say within a second or so, and its motion should cease when it arrives there. This involves two factors, the period of oscillation of the coil, and the damping of this oscillation. The period must clearly be reasonably short, to provide the rapid response, and the damping must be sufficient to bring the coil to rest when it has reached full deflection, without being so overdone that the movement becomes sluggish. The effects of three degrees of damping are illustrated in Fig. 18.4, and it can be seen that the ideal is the critical condition.

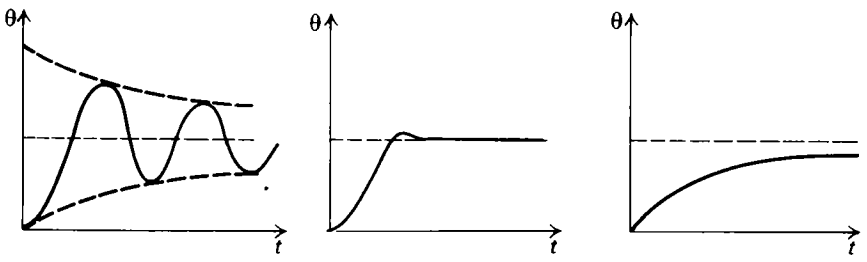


Fig. 18.4 The effects of damping (a) Insufficient damping (b) Nearly critically damped (c) Over damped

The speed of the critically damped motion is only a small amount slower than that of the undamped oscillation, and the period of this oscillation therefore governs the response time of the meter. This period is

$$T = 2\pi\left(\frac{\mathcal{J}}{k}\right)^{\frac{1}{2}} \tag{18.4}$$

where $\mathcal{J}$ is the moment of inertia of the moving coil about the axis of oscillation, and k is the elastic constant of the springs, already referred to as 'the stiffness'. This equation illustrates the point made when discussing sensitivity: if k is to be made very small in order to increase sensitivity, then the moment of inertia $\mathcal{J}$ must be kept small too to prevent the coil motion becoming too slow.

The required damping is provided by electromagnetic induction. The usual device is to wind the coil on a metal frame: in this case, so long as the coil is in motion, this frame is cutting flux and a current is induced in it which converts energy to heat; since this energy derives from the kinetic energy of the coil and frame, the result is to damp the motion.

[An additional contribution is made to the damping by the coil itself; so long as this is connected in a circuit, the motion causes induced current to flow in the coil, just as it flows in the frame. This additional damping may be a nuisance, because its effectiveness depends on the resistance of the circuit, and the result may be that the coil is inadequately damped when it is open circuited, yet severely over-damped when connected in a low resistance circuit. For this reason, some sensitive meters are provided with a resistance network which is permanently connected across the coil, the effect being to keep the damping reasonably constant.]

Set zero

The zero setting of the suspension type of meter may be adjusted by twisting the top end of the suspension. With the pivoted type, one of the two control springs, or both, has its anchored end fixed to a small lever, and this can be moved to set the coil correctly. These two hair springs are coiled in opposite directions in order to balance out any tendency which they might have to coil or uncoil a little with temperature rise.

The balance of the coil

It is important that the moving parts of the pivoted type of meter shall be correctly balanced, in order to reduce the friction at the bearings, and this may be adjusted by having small balancing nuts, as shown in Fig. 18.5, largely to counterbalance the weight of the pointer.

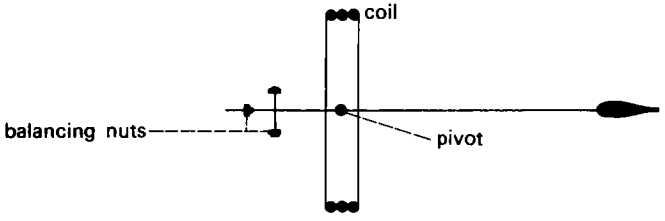


Fig. 18.5 Balancing the coil and pointer

*18.2 The moving coil meter used ballistically

The ballistic use of the moving coil meter to measure a transient flow of electric charge has already been described (sections 8.13 and 14.13) but the theory of the instrument itself in this respect was omitted, and this can now be given.

A steady current I flowing in the coil produces a steady deflection θ, which is, for the properly designed meter, proportional to the current.

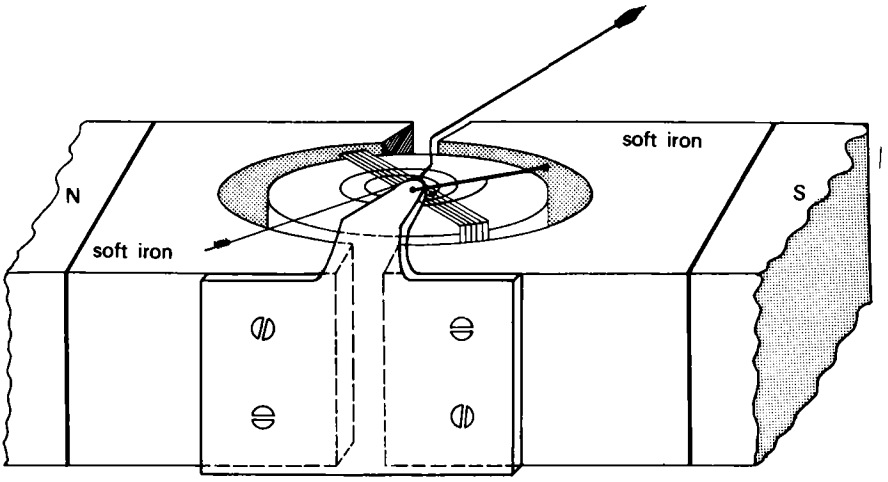


Fig. 18.6 The movement of a moving-coil meter

$$\theta = S_I I, \tag{18.5}$$

where S_I is the current sensitivity.

A rapid discharge, $Q = It$, on the other hand, causes an impulsive torque to act momentarily on the coil, which consequently suffers a momentary deflection or throw Ψ, after which the coil returns to its zero position or oscillates about it, dependent on the damping. This throw Ψ is proportional to the total charge which has flowed through the coil, provided that the impulse resulting from it is almost spent by the time that the motion of the coil has become appreciable. We can write

$$\Psi = S_q Q, \tag{18.6}$$

where S_q is the quantity sensitivity.

Owing to the condition stipulated above, a meter which is specifically intended for ballistic use is constructed to have an unusually large period of oscillation, say about 10 s, to allow greater tolerance in the matter of the time for which the impulse may last. Usually the damping device is also omitted, and the coil is left free to oscillate in the field; if this is done, a shortcircuiting switch is connected across the terminals, to provide critical damping when the reading is to be brought back to zero (fig. 18.7). Nevertheless, an ordinary moving coil meter can be put to

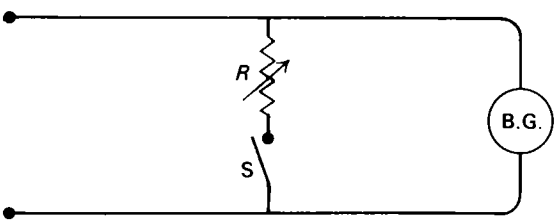


Fig. 18.7 Arrangement to damp the oscillations of a ballistic meter. The value of the resistor R is adjusted to provide critical damping

ballistic use successfully, provided that the flow of charge is sufficiently short-lived.

The theory of the ballistic meter is as follows. The discharge through the coil causes an angular impulse, given by $\int \Gamma \mathrm{d}t$, where Γ is the deflecting torque at the

instant t. This produces angular momentum $\mathcal{J}\omega$ of the coil, where $\mathcal{J}$ is its moment of inertia about the axis of rotation and ω is the angular velocity which it acquires. Thus,

$$\int \Gamma \mathrm{d}t = \mathcal{J}\omega, \text{ where } \Gamma = BNAI.$$

Therefore
$$BNA \int I\mathrm{d}t = \mathcal{J}\omega,$$

and
$$BNAQ = \mathcal{J}\omega, \tag{18.7}$$

where $Q = \int I\mathrm{d}t$ is the total charge which flows through the coil.

The coil is now in motion, its initial kinetic energy being

$$E_\mathrm{k} = \tfrac{1}{2}\mathcal{J}\omega^2. \tag{18.8}$$

Assuming that there is no damping, the coil will reach the extreme of its oscillation at a momentary deflection Ψ such that all its energy has temporarily become stored in the twisted suspension or control springs, in which case it is elastic potential energy

$$E_\mathrm{p} - \tfrac{1}{2}k\Psi^2, \tag{18.9}$$

k being the elastic 'stiffness' constant of the suspension.

Therefore
$$\tfrac{1}{2}\mathcal{J}\omega^2 = \tfrac{1}{2}k\Psi^2. \tag{18.10}$$

Eliminating ω between equations 18.7 and 18.10 gives

$$BNAQ = (k\mathcal{J})^{\frac{1}{2}}\Psi \tag{18.11}$$

and this gives us an expression for the quantity sensitivity S_q, viz.

$$S_\mathrm{q} = \frac{\Psi}{Q}$$
$$= \frac{BNA}{(k\mathcal{J})^{\frac{1}{2}}}. \tag{18.12}$$

Comparing this with equation 18.2, which gives the current sensitivity S_I, viz.

$$S_I = \frac{BNA}{k},$$

we have
$$S_\mathrm{q} = S_I\left(\frac{k}{\mathcal{J}}\right)^{\frac{1}{2}},$$
$$= S_I\frac{2\pi}{T}, \tag{18.13}$$

where T is the period of oscillation of the coil. This last step follows from equation 18.4

$$T = 2\pi\left(\frac{\mathcal{J}}{k}\right)^{\frac{1}{2}} \tag{18.4}$$

The theory above has assumed complete absence of damping, and the observed throw Ψ' would, in practice, be less than Ψ. The task of correcting this discrepancy can be avoided in experiments which are merely making comparisons, provided that the circuit resistance can be kept the same for both the measured deflections;

the damping correction term is then the same for both readings, and we can write

$$\frac{\Psi_1'}{\Psi_2'} = \frac{\Psi_1}{\Psi_2} = \frac{Q_1}{Q_2}. \tag{18.14}$$

18.3 The use of the moving coil meter to measure alternating current

The moving coil meter cannot be used directly to measure alternating current because its deflection is in direct proportion to the current, and the direction of the deflecting torque reverses with every reversal of the current. So, if response to alternating current is required, the current must be rectified before flowing through the meter, and a bridge circuit (section 8.8) is normally used for this, since the output in this case is steady enough not to require smoothing. The detail of the circuit depends on whether it is p.d. or current that is to be measured and the two arrangements are shown in Fig. 18.8; to measure current, the a.c. flows in the primary coil of a small transformer, and the output from the secondary is rectified and measured.

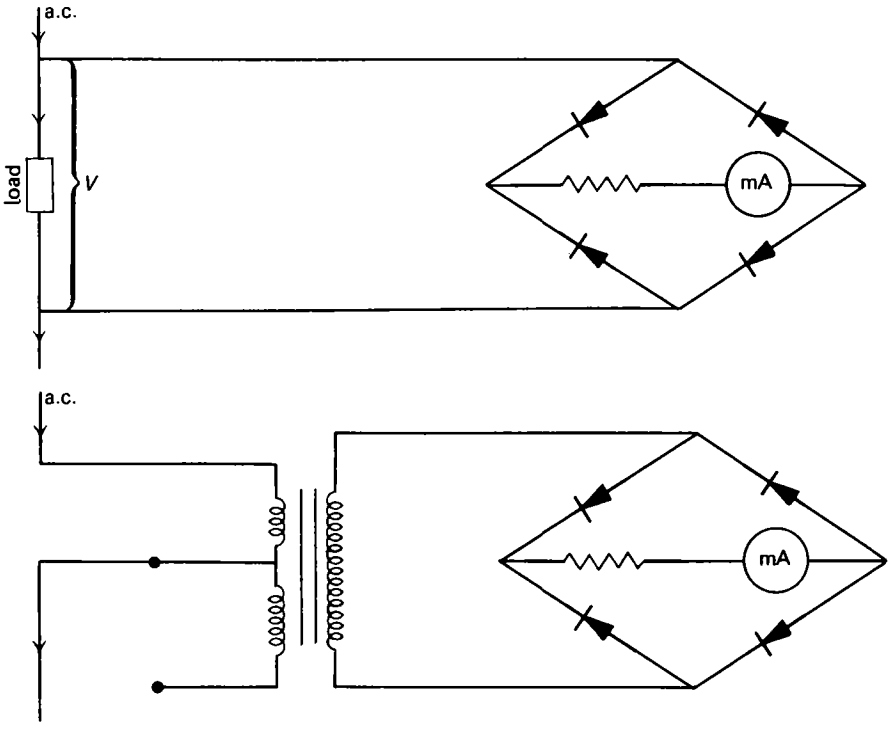


Fig. 18.8 A moving-coil meter, adapted to measure an alternating p.d. or current (a) An a.c. voltmeter (b) An a.c. ammeter (Changing the transformer tapping alters the range

The thermomilliammeter

An alternative device is more indirect. A proportion of the current is made to flow through a small heating resistor, which raises the temperature of an antimony-bismuth thermojunction placed in contact with it, and the resulting current in the

thermocouple is measured on the meter. This instrument, which is called a thermo-milliammeter, is useful for work at radio-frequencies, but it is easily destroyed if overloaded.

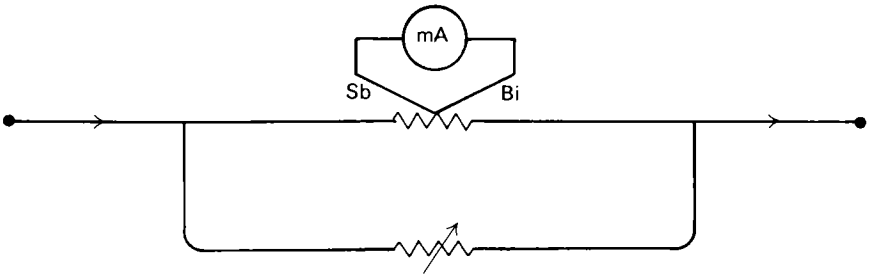


Fig. 18.9 A thermomilliammeter

18.4 The moving iron meter

Although not as important as the moving coil instrument, the moving iron meter has the advantage that it can be directly applied to the measurement of alternating current without the need for rectifying circuits and the like, and it is used quite extensively for this purpose. One design of such an instrument is shown in Fig. 18.10. When current is sent through the fixed coil C, the piece of soft iron F is magnetically attracted towards it, and the force is still attraction if the current is reversed, which accounts for its response to alternating current. In this particular design the centre of gravity of the moving parts lies below the pivot and the deflecting torque is balanced against a torque due to gravity, but a control spring could equally well be used instead. The gravity controlled instrument needs to be levelled, and in this way the zero setting is adjusted. In another type of meter, two pieces of soft iron are arranged side by side inside a solenoid, one of them being fixed and the other attached to a pivoted arm which includes the pointer; the current in the solenoid magnetizes both pieces of iron with the same polarity, and the deflection in this case is due to their repulsion.

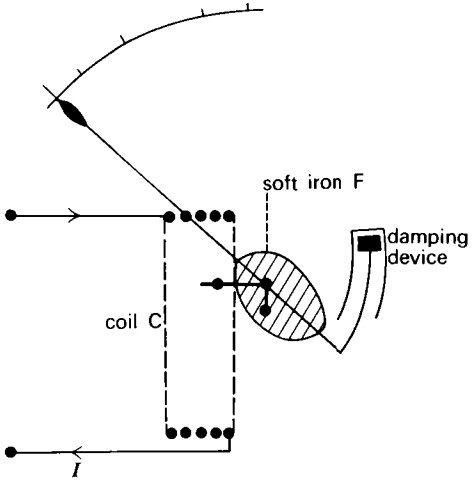


Fig. 18.10 A moving-iron meter – attracted iron type

In no type of moving iron meter is the torque proportional to the current; it is more nearly proportional to the square of the current, at any rate over the lower part of the scale, consequently the scale is not linear, which is a disadvantage. There is no permanent magnetic flux to provide electromagnetic damping, and a dashpot is provided instead, consisting of leaky piston and cylinder. Because the flux due to the current is comparatively weak, the reading can be affected by external magnetic fields, and this tendency is avoided by the use of an iron case to act as a magnetic shield.

Moving iron meters have not the sensitivity or the accuracy of moving coil ones, but they are robust, and are the cheapest reliable instruments available for measuring alternating current and p.d. at 50 Hz.

18.5 Calibrating an alternating current meter

If a meter has been calibrated using direct current, it is not usual to find that the calibration is relevant also to its use with alternating current; there is no particular reason why it should be. This is a difficulty, because a d.c. calibration can be carried out by reference to a standard cell and a standard resistor, but this cannot be done using a.c.

However, a dynamometer or a current balance may be used to solve the problem (section 17.3), because of their characteristic behaviour. The response of these instruments is proportional to the square of the current, with the result that, if calibrated using d.c., the same calibration represents r.m.s. values when a.c. is flowing, and this is exactly what is required.

The easiest way to use a dynamometer as a means of calibrating an a.c. meter is illustrated in Fig. 18.11. The dynamometer is first calibrated using d.c., either by

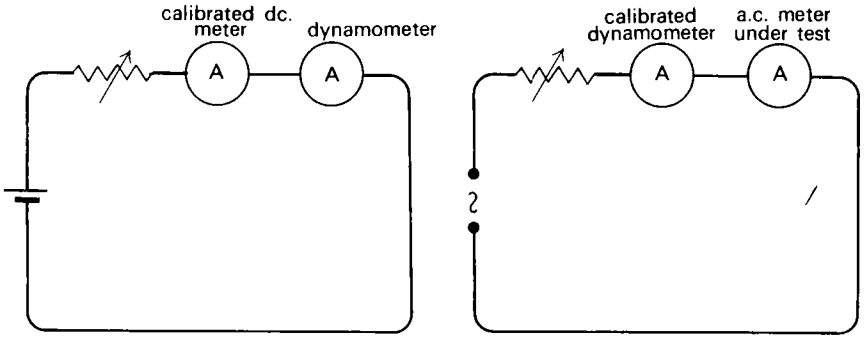


Fig. 18.11 Calibrating an a.c. meter

direct reference to a standard cell, or by connecting it in series with a reliable d.c. meter. It is then connected in series with the a.c. meter under test and with an a.c. source, and used to calibrate the a.c. instrument.

18.6 A brief survey of measuring devices

The meters described in this chapter, being electromagnetic in their action, are all basically current measuring instruments; if used to measure a p.d. they do so by measuring the current which flows when the p.d. is applied across a suitable resistor. This is an intrinsic defect in the instrument when it is used as a voltmeter, since the ideal voltmeter would draw no current at all. In cases where this defect is

unacceptable, the traditional procedure was formerly to use an electrostatic instrument instead. The modern technique, however, is to use some form of valve voltmeter (the term may refer to a thermionic valve or a transistor).

As a current measuring device, the most sensitive type of moving coil meter can detect and measure a current rather less than 1 μA; the range can be extended to below 10^{-9} A without the meter itself requiring unusual sensitivity with the aid of a d.c. amplifier, which amplifies the effect of a current electronically before measuring it. (This is also about the range of a pulse electrometer, section 8.21.)

There are some applications, however, such as the measurement of ionization current in a gas, where even greater sensitivity is needed. In such cases, the current to be measured is made to flow through a very high value resistor, and the p.d. across this resistor is measured. A current of 10^{-12} A, for example, would produce a potential difference of 1 V across a resistor of 10^{12} Ω. But this requires a voltmeter of very special properties, for the act of connecting it across a 10^{12} Ω resistor must not appreciably lower the effective value of this resistor. Such a voltmeter must, therefore, have almost infinite resistance and yet be sensitive to a P.D. of less than 1 volt. The instrument used is a d.c. amplifier with a special form of valve (or its solid state equivalent) connected across its input, known as an electrometer valve. The input impedance of this valve is very high indeed. When measuring currents as small as this, all the leads require screening from the effects of stray charges.

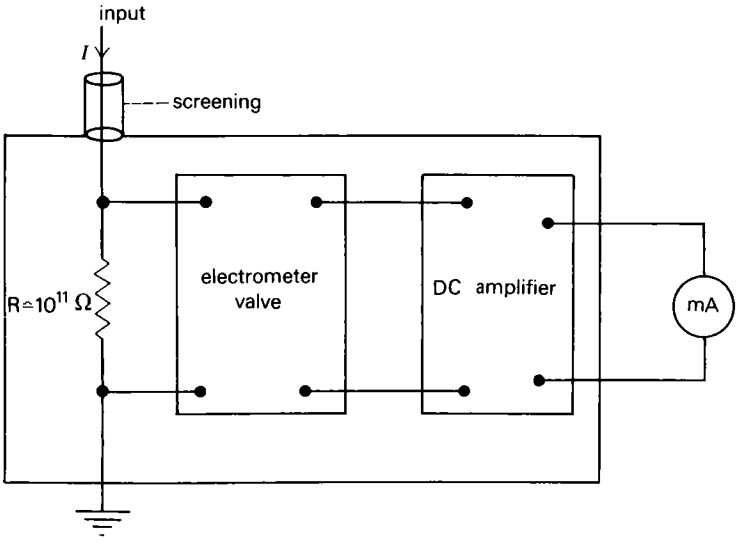


Fig. 18.12 Measuring an extremely small current, about 10^{-11}A

QUESTIONS

18.1 A moving coil galvanometer with a linear scale has a coil consisting of 50 turns of wire wound on a former 2 cm square and pivoted in a magnetic field of flux density 0.3 T. A current of 10 μA produces a deflection of 5°. Calculate the torque required to twist the suspension through an angle of 1 rad. (O & C)

18.2 A suspension type of moving coil galvanometer has a coil with 300 turns of

mean area 1 cm^2, moving in a radial field of flux density 0.2 T. The galvanometer mirror reflects a beam of light onto a curved scale 20 cm distant from the mirror. The sensitivity of the instrument is represented by a deflection of 1.2 cm on the scale for a current of 1 μA. What torque would be required to twist the coil suspension through an angle of 1 rad?

18.3 A moving coil microammeter of a particular design has a rectangular coil of N turns enclosing area A which moves in a radial field of flux density B. The resistance of the coil is R.

A similar meter is constructed, on exactly the same design as the first, but having $10N$ turns in the coil. In order to keep the time of oscillation the same as before, this coil is wound using thinner wire, in such a way that the total mass of copper is the same in the coils of both meters.

Compare (*a*) the current sensitivities, (*b*) the voltage sensitivities of the two meters.

18.4

The circuit in the diagram is used to find the sensitivity and resistance of a highly sensitive galvanometer. R_1 and R_2 are both resistance boxes capable of up to 10000 Ω by 1 Ω steps. The resistance of the galvanometer is, in fact, 1000 Ω. It is found that if R_1 is set at 4999 Ω, and R_2 at zero, the deflection of the galvanometer is 10.0 cm. What is its sensitivity, in μA/mm?

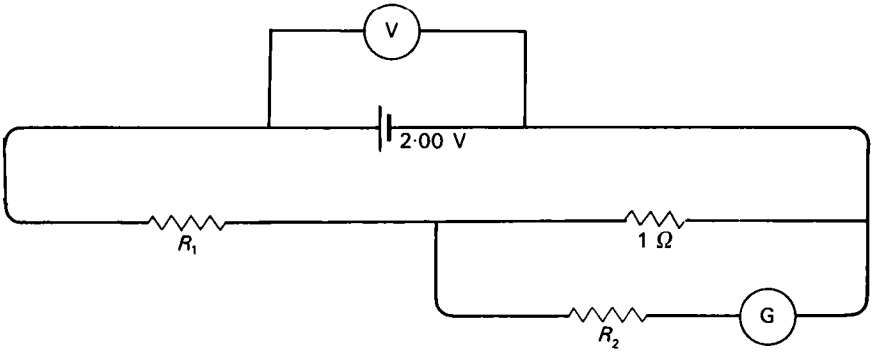


Fig. e 80

Describe how the apparatus could have been used to find the resistance of the galvanometer if it had not been known, explaining how the result is obtained from the readings taken. (S)

*18.5 Compare the sensitivity (*a*) for current and (*b*) for voltage of two galvanometers, one of which has a coil and former n times smaller in all linear dimensions than the other. The number of turns in the coil and the magnetic flux density are the same in both cases, and the gauge of wire is the same, but the torsional constant of the suspension is adjusted so that the time of swing is the same. (OXF. SCH.)

*18.6 A fixed length of wire l is to be made into a square coil of one or more turns, and when carrying a current I it is to be placed in a magnetic field of flux density B acting in the plane of the coil. Show that the torque experienced will be a maximum when the coil consists of a single turn, and find the magnitude of this torque. In view of this result, why are the coils of moving coil galvanometers wound with many turns of wire? (O & C)

*18.7 A certain moving coil galvanometer, used with a lamp and scale, is found to give a steady deflection proportional to the current through it; 10 μA give a steady deflection of 60 cm on the scale. The galvanometer has a period of oscillation of 15 s. When it is connected to the secondary coil of a mutual inductance of 4 mH, a ballistic throw of 10 cm occurs on reversing a current

of 1 A in the primary circuit. Find the total resistance of the secondary circuit. (OXF. SCH.)

*18.8 The current sensitivity of a moving coil galvanometer is 100 mm/μA, and its periodic time is 10 s. A 1 μF capacitor is charged to 1 V and discharged through the galvanometer. The ballistic throw observed is 60 mm. The secondary of a mutual inductance of 1 mH is then connected to the galvanometer terminals, and the total resistance comprising the secondary and the galvanometer is 1 000 Ω. A current of 0.5 A is reversed through the primary of the mutual inductance and a ballistic throw of 40 mm is observed. Comment on these results. (CAM. SCH.)

*18.9 A moving coil meter which has been calibrated at 15 °C is used with an external circuit consisting of a constant voltage source and copper conductors. For the copper coil of the meter, the temperature coefficient of resistance is $+4.0 \times 10^{-3}$ °C^{-1}; for the control spring, the temperature coefficient (i.e. the fractional change in torque per unit rise in temperature) is -3.0×10^{-4} °C^{-1}. Will the meter read too high or too low at temperatures above 15 °C? Is the error significant at 25 °C?

If the meter is to be converted to read as an ammeter, which would be the better material for the necessary shunt – one with negligible temperature coefficient of resistance, or one with the same temperature coefficient of resistance as the coil? (O)

18.10 Explain the principle of the moving coil galvanometer with the aid of a careful diagram, and derive expressions for (*a*) the current sensitivity, and (*b*) the voltage sensitivity, in terms of the torque k per unit deflection, stating clearly the meaning of any other symbols you use. Explain why the coil is often wound on a conducting frame from which it is insulated.

Two galvanometers A and B are alike in all respects except that B has twice as many turns in its coil as A, and the resistance of B is three times that of A. Compare (*a*) their current sensitivities, (*b*) their voltage sensitivities. (S)

18.11 Describe briefly some form of moving iron ammeter, and explain why it may be used for measuring alternating current.

Describe fully how you would calibrate a moving iron ammeter to read the r.m.s. value of sinusoidal currents alternating at 50 Hz.

Such an instrument is used to measure the current flowing in an air-cored solenoid connected to a 50 Hz alternating source. State, with reasons, what changes in current you would expect to observe if (*a*) a laminated soft iron cylinder, (*b*) a copper cylinder were inserted in the solenoid. Why should the ammeter not be placed near the solenoid when measuring the current? (O & C)

18.12 Describe the structure of a pointer type moving coil galvanometer. State what you mean by the current sensitivity of such an instrument and discuss the factors on which it depends.

Explain carefully how a linear scale is achieved.

*18.13 Give an account of the ballistic galvanometer. What factors determine (*a*) the charge sensitivity, (*b*) the period of swing, (*c*) the damping? (O)

Chapter 19

Applications of electromagnetism

19.1 The moving coil loudspeaker

In its basic construction the moving coil loudspeaker is very simple, which makes it a good subject for starting this catalogue of instruments. The moving coil consists of a few turns of copper wire wound round a very light cylinder which is placed in the cylindrical cavity separating the central from the surrounding pole of a powerful permanent pot magnet. This cylinder forms the apex of a fibre cone which is fixed round its open circumference AB (Fig. 19.1). The output current

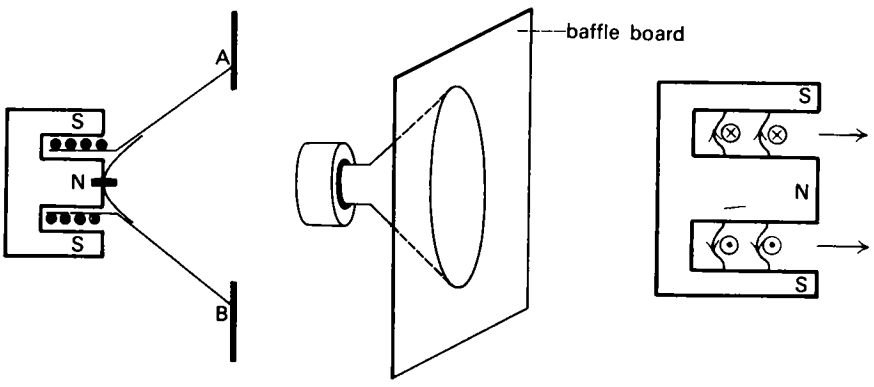


Fig. 19.1 A moving-coil loudspeaker (a) (b) (c)

from an amplifier flows in the coil, and the resulting motor force moves the coil in and out in sympathy with the current, thus vibrating the cone and launching sound waves into the air. This is a direct application of the equation

$$F = BIl \tag{19.1}$$

and the force is proportional to the current.

The vibrating cone launches two waves, one in front, and one behind, and these must be prevented from meeting, since they are clearly out of phase. The best way of achieving this is to clamp the cone to a large baffle board, as shown in the diagram.

19.2 The moving coil microphone

The energy conversion in the loudspeaker is reversible and it is possible to use it as a microphone; if sound waves set the cone in vibration, the resulting movement

of the coil causes it to cut flux and an e.m.f. is induced in it, the equation in this case being

$$V = Blv, \tag{19.2}$$

i.e. the induced e.m.f. V is proportional to the velocity with which the coil moves.

This is the principle of the moving coil microphone, which differs from the loudspeaker only in some details of its construction.

19.3 The telephone receiver

The receiver used in the telephone system works on a different principle, using the attraction between a magnet and a ferromagnetic diaphragm. The speech current flows in coils wound on two soft iron pole pieces P placed on the poles of a permanent bar magnet. The diaphragm AB, made of permalloy, is clamped round its circumference, while its centre is attracted towards the pole pieces without actually touching them. The varying speech current causes variation in the force attracting the diaphragm, causing it to respond by vibrating at the same frequency.

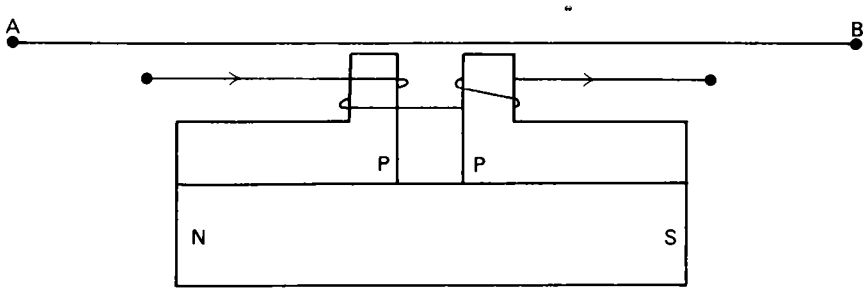


Fig. 19.2 The telephone receiver

The function of the permanent magnet is twofold: without it, the diaphragm would be attracted twice for every single cycle of the energizing current, thus putting the signal up one octave in pitch; with the magnet present, however, the diaphragm is attracted all the time, and the variations in strength of the attracting force provide the correct frequency of response (Fig. 19.3).

The second function of the permanent magnet is to magnetize the diaphragm by induction, with the result that its response to the signals is more pronounced.

19.4 The carbon microphone

The type of microphone used in the telephone system is based on the rather surprising fact that the electrical resistance of a box of carbon granules changes appreciably if they are pressed further together, owing to the better contact between them, and returns to its original value when the pressure is released. Moreover, this behaviour persists, even when the pressure is varied at audio-frequency, and the small effect of a sound wave on a diaphragm is sufficient to produce it. As compared with the sound powered moving coil, this microphone is battery powered, since it does not generate its own current, the basic equation being

$$I = \frac{V}{R_0 + R}, \tag{19.3}$$

where R, the resistance of the carbon granules, varies with the signal.

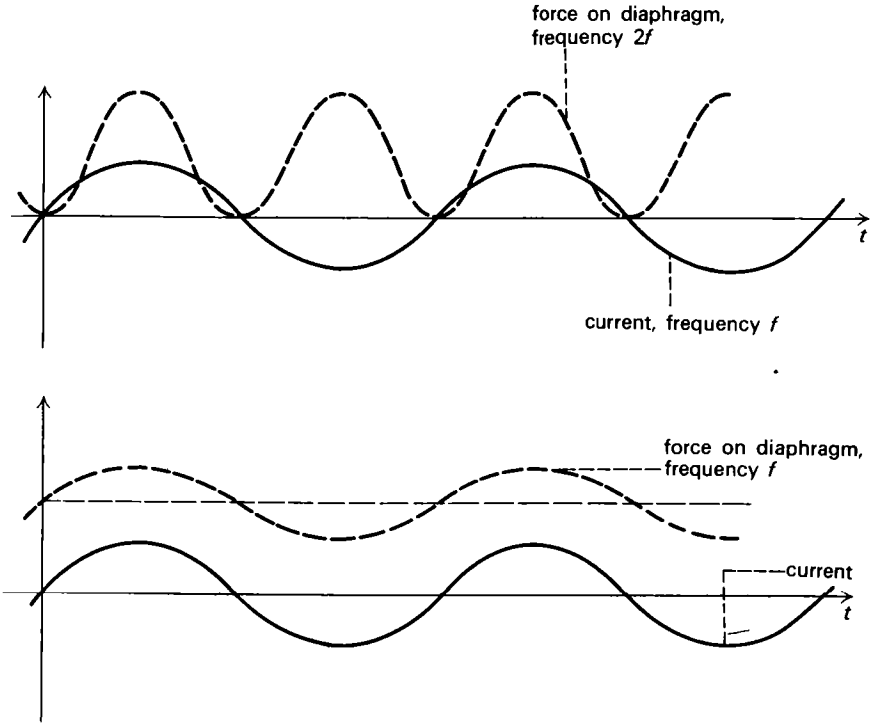


Fig. 19.3 The function of the permanent magnet in a telephone receiver (a) Without the permanent magnet (b) The effect of the permanent magnet

Since it is current variation which is really required, not resistance variation, and the two are not in direct proportion, it is clear that the carbon microphone is not a very faithful performer, but it has the advantage of providing signals which do not require amplification. This is because the energy derives from a battery, and not from the original sound. In this sense, the microphone is itself an amplifier.

Among the unwanted components, the battery supplies a direct current, which must not be allowed to dissipate energy uselessly along the telephone lines. For this reason, and for matching purposes, the line is fed through a transformer (Fig. 19.4).

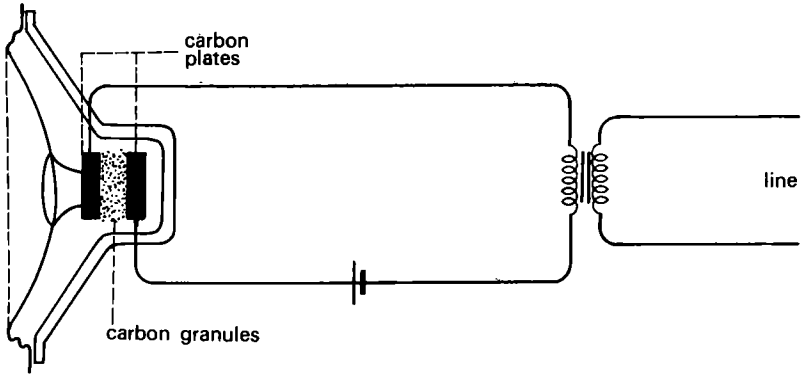


Fig. 19.4 The principle of the carbon microphone

There are other equally useful types of microphone, but these will not be described here, as it is the intention in this chapter to be selective rather than exhaustive, and we will instead take up the new topic of generators and motors.

19.5 A commutator d.c./a.c. motor

Every type of electric motor has, in principle, the possibility in it of being used as a generator, because the energy conversion is a reversible one, and it is with this aspect, rather than the technology of various types, that we are concerned here. Consequently, since it happens to be convenient, the stress will be laid on a commutator motor, although in practice other types particularly suited to a.c. operation are more commonly used.

When a motor is running, the coil of its armature is cutting magnetic flux, inducing in it, as if it were a dynamo, an e.m.f. which opposes the p.d. applied from the source of electric power. The equation which expresses the power conversion in all types of motor is basically the same, viz.

$$EI = \Gamma\omega, \tag{19.4}$$

where E is the back e.m.f. induced in the armature coil, owing to the cutting of flux, I is the armature current, Γ is the torque which drives the armature and the load placed on it, ω is the angular velocity with which the armature rotates, i.e. $\omega = 2\pi f$, where f is the frequency of rotation.

With a dynamo, the equation is the same, but in this case Γ is the torque exerted by the mechanical source of power which drives the armature, E is the e.m.f. of the dynamo, which acts as a source of electric power, and I is the current taken from it.

A very simple form of two pole commutator motor is illustrated in Fig. 19.5.

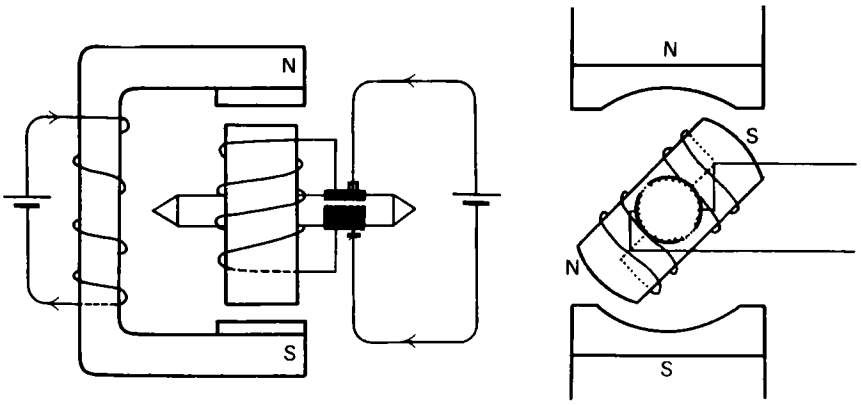


Fig. 19.5 Two views of a simple commutator motor (a) (b)

The armature is an electromagnet, which draws its current from the source by way of a commutator on its shaft, made from a copper sleeve split into two halves, each half making contact with a carbon brush. The armature revolves between the two poles of a horseshoe electromagnet, which is the field magnet, activated by current flowing in a field coil. Thus, as the armature revolves, which it does through the interaction of the two magnets, the commutator automatically reverses the armature polarity twice in each revolution, and this reversal can be arranged to occur at the appropriate instants to maintain unidirectional rotation. Although this

explanation has been framed in terms of a force between magnets, the motor is in fact a practical application of the motor force equation

$$F = BIl,$$

the current I being the armature current and B the flux density caused by the field magnet.

Such a simply constructed motor would in practice be both inefficient and jerky in its motion, since the driving torque has the form of a succession of impulses. A more practical motor is wound in a complicated manner aimed at providing a continuous torque, and its armature has, in effect, a large number of poles, the commutator being split into a corresponding number of narrow strips, but these are details which we need not discuss further.

The reason why the commutator motor will work also with alternating current is that it does not affect the action if both armature and field currents reverse direction together; the motor action requires the reversal of one at the appropriate instants, not of both.

Shunt and series connection

In practice, both armature and field coils draw their currents from the same source, and the usual way to arrange this is to connect them in parallel, making a shunt wound motor, and most of the theory which follows will refer to this. Since the function of the field current is merely to energize the field magnet, and not to contribute to the motor action, the principle is to pass a small field current through a large number of turns.

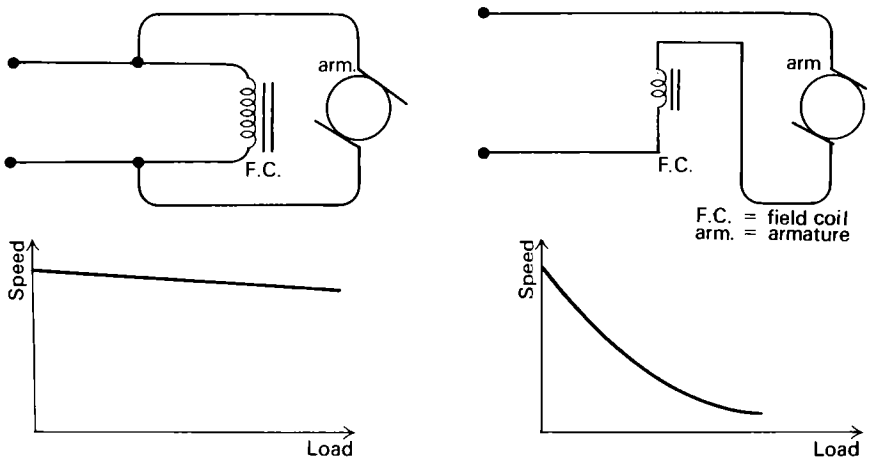


Fig. 19.6 Shunt and series wound motors (a) Shunt wound (b) Series wound

In the series wound motor, on the other hand, where the field coil and armature take the same current, this must be large, and the field coil consists of a comparatively small number of turns with low resistance.

The advantage of the shunt wound motor lies in the fact that its running speed is fairly constant, being only slightly dependent on the load put on it. The series wound motor, lacking this feature, makes compensation by providing a large driving torque while it is starting up, and is preferred on this account for some applications.

Eddy currents

Since the flux in the iron is continually changing, electromagnetic induction sets up e.m.f.s in it, which would cause eddy currents to flow, dissipating energy. For this reason, both armature core and the field magnet are laminated. The direct effect of eddy currents in this case would be that of an electromagnetic brake, making rotation more difficult.

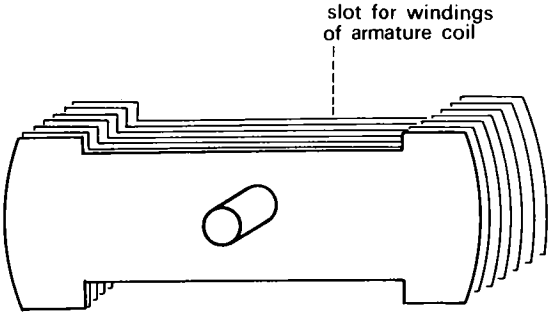


Fig. 19.7 The laminated armature core of the simple two pole motor

19.6 The equations of the shunt wound motor

The theory of the shunt wound motor is relatively simple because the field current I_f, and consequently the flux density B, are effectively constant, being unaffected by the motion of the armature. Referring to Fig. 19.8, we have

$$I_f = \frac{V}{R_f}, \tag{19.5}$$

$$I = \frac{V-E}{R}, \tag{19.6}$$

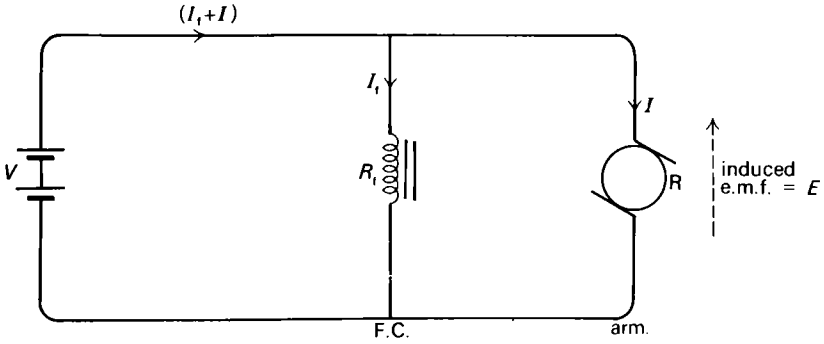


Fig. 19.8

where I, R refer to the armature, and E is the e.m.f. induced in it owing to its rotation in the field. Also following the usual equations giving a motor force or torque,

$$\Gamma \propto BI, \tag{19.7}$$

and from Faraday's law of electromagnetic induction

$$E \propto Bf, \tag{19.8}$$

where f is the frequency of rotation of the armature.

These last two equations are consistent with the relation for power conversion which has already been quoted,

$$EI = \Gamma \cdot 2\pi f. \tag{19.4}$$

The total power taken from the source, however, includes the power which is wasted as heat in the armature and field windings, viz.

$$V(I_f + I) = I_f^2 R_f + I^2 R + EI, \tag{19.9}$$

an equation which could be derived from 19.5 and 19.6.

From these equations we can draw some conclusions with respect to the characteristic behaviour of the motor, as follows.

(a) The variation of speed with load

The motor adjusts itself automatically to the changed circumstances if the load is increased, providing the extra torque that is needed by the effect of a small reduction in speed. This is because the speed reduction reduces the back e.m.f. E (equation 19.8), with the result that the armature takes a larger current (equation 19.6), and this increases the torque (equation 19.7).

Example

A small 200 V *shunt wound motor is running under a particular load at a speed of* 10 rev/s *and the armature current is* 4.0 A. *The load is now increased to five times its previous value. If the armature resistance is* 0.5 Ω, *what is the new running speed?*

Under the original load, we have

$$4.0 = \frac{200 - E}{0.5}, \tag{19.6}$$

Therefore the back e.m.f. $E = 198$ V.

To provide sufficient torque for the increased load, the armature current must be increased five times, since the torque is proportional to the current (19.7).

Therefore $\quad I' = 20$ A,

$$20 = \frac{200 - E'}{0.5}, \tag{19.6}$$

i.e. $\quad E' = 190$ V.

But the e.m.f. is proportional to the running speed, and so the frequency f' of rotation is now given by

$$\frac{f'}{10} = \frac{190}{198},$$

i.e. $\quad f' = 9.6$ rev/s,

which is only a very small reduction.

[The reason for this is that the armature current, and therefore the torque, is proportional to $(V - E)$, whereas the speed of rotation is proportional to E, and a small change in speed, which alters E only a little, nevertheless results in a large change in $(V - E)$.]

(b) A starting resistance

So far, we have assumed the armature of the motor to be revolving, and nothing has been said about the initial condition when the motor is stationary and the p.d. is first applied. At this instant, there is no back e.m.f. at all, and the armature current is given by

$$I = \frac{V}{R}.$$

Since the armature resistance is deliberately made as small as possible in order to reduce the energy loss, and it is the back e.m.f., not the resistance, which normally controls the size of the armature current, this starting current would be intolerably large. (In the example above, it would be 400 A.) For this reason, every motor except the smallest model one has a starting resistance connected in series, which can be progressively reduced in size as the armature gathers speed.

(c) Controlling the speed by varying the field current

A variable resistor may be connected in series with the field coil of a shunt wound motor, and used as a speed control. If the motor is running, and the flux density B is reduced, the rather surprising result is an increase in speed. The reason is that the immediate effect of the reduced flux is to reduce the back e.m.f. E and so the armature current increases. This more than compensates for the reduction in the flux density B in its effect on the driving torque, and the net effect on this torque is to increase it, and so the motor speeds up.

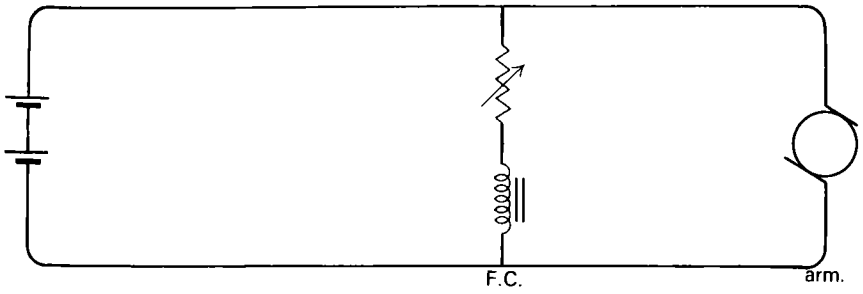


Fig. 19.9 A shunt wound motor – a speed control

19.7 The equations of a series wound motor

Most of the equations stated in the last section apply also to a series wound motor, but their practical consequence is more complicated, owing to the factor that the field current, being identical with the armature current, is not constant, and the flux density B varies with it. Consequently the motor cannot adjust itself to a changed load in the same way as the shunt wound motor can, without considerable changes in its speed.

19.8 The d.c. dynamo

Figure 19.5 may be accepted as a diagram of a simple d.c. dynamo, provided that the external connections are appropriately altered, for the dynamo merely reverses the energy conversion of the motor. In this case, instead of being driven round electrically, the armature is driven round by an external source of mechanical

power, with the result that its coil cuts the flux of the field magnet, and an e.m.f. is induced in it. This e.m.f. is available as a source of electrical power; mechanical energy is fed in, and electrical energy is taken out.

The e.m.f. induced in the armature is an alternating one, and the function of the split ring commutator is now to rectify its output. (With the very simple armature and commutator which we have been describing, the output current would have the waveform which results from unsmoothed full-wave rectification, but a more complicated winding would leave merely a slight ripple, known as commutator ripple.)

If the ends of the armature winding were connected to two continuous slip rings, instead of to a split commutator, the output would, of course, be alternating, but this is not a principle which is put to much practical use, because of preference for the type of a.c. generator which will next be described.

Whether d.c. or a.c. is taken out, the field coils must be energized with d.c., which, in the case of the d.c. dynamo, is part of the generated output. There is the same choice as with the motor between having a shunt or series connection, and the characteristic behaviour of the two types is shown below.

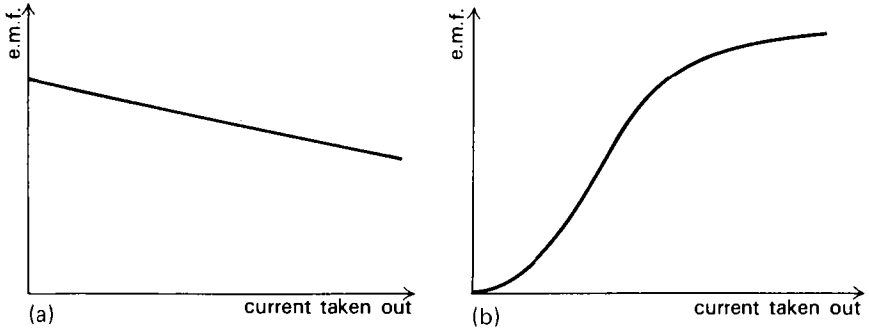


Fig. 19.10 The characteristics of shunt and series wound dynamos (a) Shunt wound (b) Series wound

Power conversion

The equation representing the power conversion in the dynamo is, of course, the same as it is for the motor, viz.

$$\Gamma \cdot 2\pi f = EI. \tag{19.4}$$

The interpretation, in this case, is that, when current is not being taken out, the only torque required to maintain the armature rotation is that demanded by friction. As soon as current is taken, however, this produces a motor effect opposing the motion, and it is mechanical power taken up in opposition to this that is converted to electrical power.

19.9 An a.c. generator

The practical a.c. generator is based on the same principle as the bicycle dynamo, in which the rotor is a revolving magnet which induces an alternating e.m.f. in a stationary coil by varying the flux which threads it. The advantage of this is that no brushes and slip rings are needed, because the power is taken out from stationary coils.

It should be clear from the diagram that the flux in the coil C alternates with the same frequency as the frequency with which the magnet revolves. This is a bicycle

dynamo; with a large generator the revolving magnet is an electromagnet with a considerable number of alternately north and south poles, and these induce an e.m.f. in a number of coils each surrounding a pole piece and connected in series; but the principle is the same.

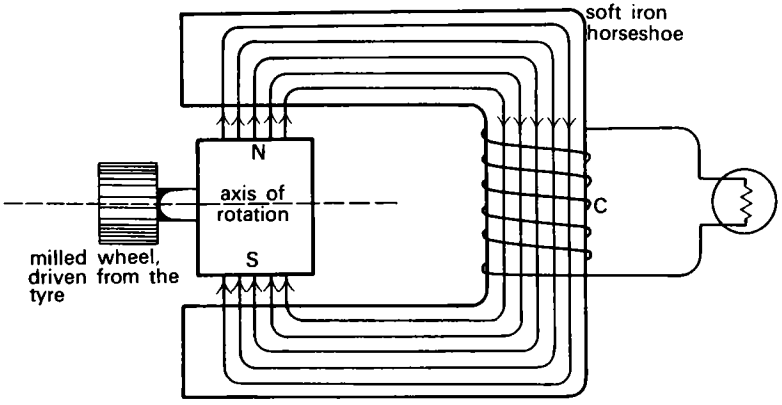


Fig. 19.11 A bicycle dynamo

19.10 The transformer

The transformer is an application of mutual induction, and is used to change the voltage of an alternating supply. It consists of two coils wound round the same soft iron core; thus, the input current results from an alternating p.d. which is applied across the primary coil, and the output is taken from the secondary coil, which has an e.m.f. induced in it by the alternating magnetic flux in the core.

The core is so constructed that the leakage flux is kept to a minimum, and we will assume that both this leakage flux and the resistance of the coils are negligible.

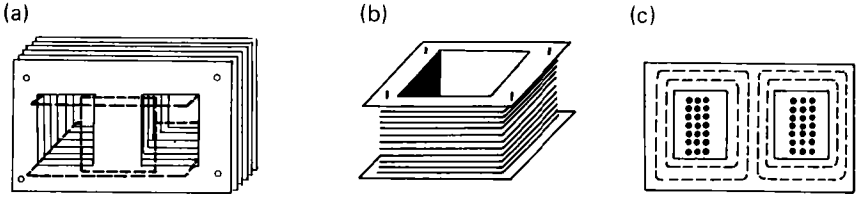


Fig. 19.12 A transformer (a) The soft iron core, made from flat sheets clamped together (b) The coils, wound on a former (c) The flux

The voltage ratio

Supposing that at any instant the flux in the core is Φ, the e.m.f.s induced in primary and secondary respectively are

$$E_1 = -N_1 \frac{d\Phi}{dt} \tag{19.10}$$

and

$$E_2 = -N_2 \frac{d\Phi}{dt}, \tag{19.11}$$

where N_1, N_2 are the numbers of turns of the two coils.

Assuming that resistance is negligible, the p.d. applied across the primary coil

must be balanced by the e.m.f. induced in it, and the size of the primary current depends on this; if the two were not balanced, the current would change, thus altering the flux, until balance was obtained.

Thus
$$V_1 = -E_1$$
$$= N_1 \frac{d\Phi}{dt}. \tag{19.12}$$

It follows that, at any instant, the ratio of the induced e.m.f. in the secondary to the p.d. applied across the primary is

$$\frac{E_2}{V_1} = \frac{-N_2}{N_1}, \tag{19.13}$$

and this is consequently the ratio of the r.m.s. values also,

i.e.
$$\left(\frac{E_2}{V_1}\right)_{\text{r.m.s.}} = \frac{N_2}{N_1}. \tag{19.14}$$

19.11 Energy considerations in the transformer

So long as current is not being taken from the secondary coil, the primary behaves as an inductor, and is designed to have high reactance at the frequency of the supply. Thus it takes only a small primary current I_L, and this is 90° out of phase with the p.d.

Now suppose that current, with r.m.s. value I_2, is taken out from the secondary coil, and that this is in phase with the e.m.f. In this case, power equal to $E_2 I_2$ is being taken out. But power cannot be taken out of the transformer without it having been fed in, and we deduce that the result of taking current from the secondary must be that extra current flows in the primary, and that this current I_1 is in phase with the applied p.d.,

i.e.
$$V_1 I_1 = E_2 I_2. \tag{19.15}$$

This effect may be demonstrated by the circuit shown in Fig. 19.13. When the switch S in the secondary circuit is open, the lamp in series with the primary coil fails to light, because the coil is acting efficiently as a choke. When the switch is closed, however, thus short circuiting the secondary coil, the lamp glows brightly.

The cause of this extra flow of current in the primary can be traced to a disturbance of the flux. The strength that the flux must have is fixed by the value of the applied p.d., since the back e.m.f. induced in the primary must balance this p.d. Now when current flows in the secondary coil, this contributes to the flux and therefore alters its value, which upsets the balance in the primary circuit. The additional current which now flows in the primary serves to destroy the flux which the secondary current created, leaving only the original flux, and so the balance is restored. Both this argument and equation 19.15 lead to the same conclusion as to what the ratio of these two currents must be, viz.

$$\frac{I_2}{I_1} = \frac{N_1}{N_2}. \tag{19.16}$$

19.12 The primary coil as a choke

Although the inductive current I_L in the primary coil plays no part in the transmission of energy from primary to secondary, it has the important function that it is entirely responsible for the magnetic flux. Since there must be some resistance

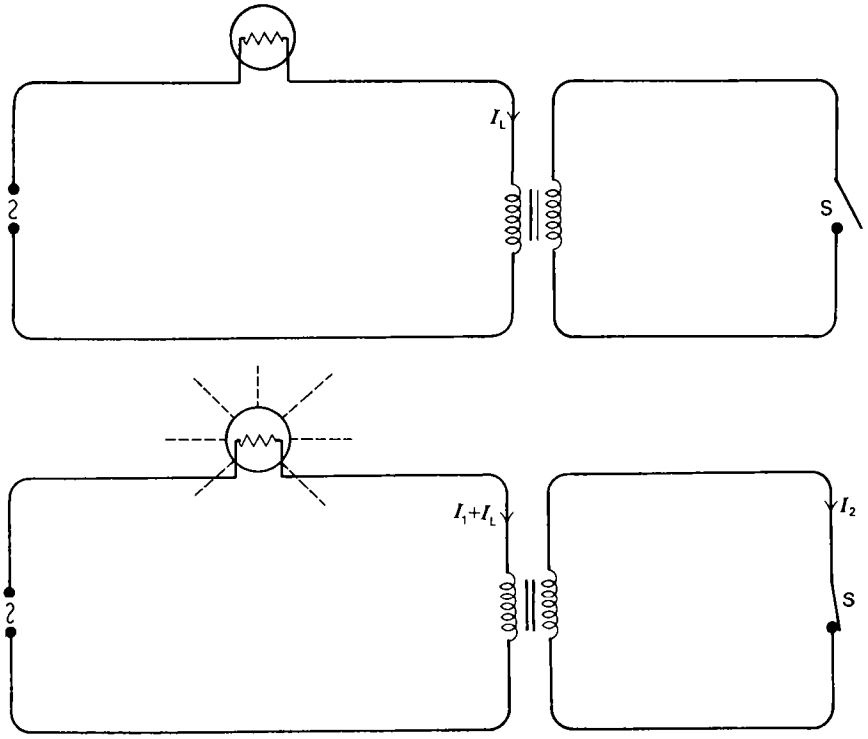


Fig. 19.13 The effect in the primary circuit of taking current from the secondary (a) The secondary open-circuited (b) The secondary short-circuited

(although so far we have neglected it), it is essential that this current, so long as it fulfils its purpose, shall be small in order to avoid overheating. For this reason, the reactance of the primary coil must be made reasonably high, and this is given by the equation

$$X = 2\pi f L, \tag{19.17}$$

where L, the inductance of the coil, depends on the number of turns N_1, the area of cross-section A of the core, and the relative permeability μ_r of the iron. Thus if the frequency of the supply is high, the physical size of the transformer may be small, i.e. if f is large, A may be small.

19.13 Energy losses in the transformer

Measured as a fraction of the whole, the energy losses in a transformer are not high, but energy lost is energy misdirected, and partly on this account the losses must be kept to a minimum, since they tend to cause overheating; a large transformer, in fact, requires to have special cooling arrangements.

Copper loss

The most obvious factor here is the resistance of the coils, causing Joule heating, which can be reduced only by increasing the cross-section of the wire. Fortunately, the coil which carries the heavier current is the coil with the smaller number of turns (equation 19.16) and heavier gauge wire can be used for this without taking too much space.

Hysteresis and eddy currents

Further losses due to hysteresis and eddy currents take place in the core, and these do not depend on whether or not any current is being taken out. To reduce the hysteresis, the material is chosen for its high permeability and the small area of its hysteresis loop. Silicon-iron has the added advantage that its resistivity is comparatively high, and this helps to keep down eddy currents, but the main contribution in this respect is to laminate the core, which is made from thin metal stampings. The geometry of these is such that their planes are parallel to the magnetic flux but perpendicular to the e.m.f. induced in the core. Thus the eddy currents would have to flow from one lamination to another, crossing the insulated surfaces.

19.14 The resistance transformer

One specialized application of the transformer is to change the effective value of a resistance, thus making it possible to match the resistance of a load to that of an alternating source, with the object of obtaining the maximum power from it (section 3.23).

The principle is this. Suppose that a load resistor of value R_2 is connected across the transformer secondary in series with a switch S, the turns ratio of the transformer being $N_2:N_1$. When the switch is closed, a current I_2 given by

$$I_2 = \frac{E_2}{R_2} \tag{19.18}$$

flows in the secondary circuit, and the effect of this in the primary circuit is that a current I_1 flows in it, where

$$I_1 = I_2 \frac{N_2}{N_1}. \tag{19.16}$$

But the p.d. across the primary coil is V_1, and the effect in the primary circuit of closing the switch in the secondary is consequently the same as it would be if this p.d. had been applied directly across a resistance R_1 given by

$$R_1 = \frac{V_1}{I_1}. \tag{19.19}$$

From these three equations, we have

$$\begin{aligned} \frac{R_1}{R_2} &= \frac{V_1}{I_1} \cdot \frac{I_2}{E_2} \\ &= \frac{V_1}{E_2} \cdot \frac{I_2}{I_1} \\ &= \frac{N_1}{N_2} \cdot \frac{N_1}{N_2}, \end{aligned}$$

i.e.

$$\frac{R_1}{R_2} = \left(\frac{N_1}{N_2}\right)^2. \tag{19.20}$$

For example, if the ratio N_1/N_2 were 25/1, and a resistor of value 4 Ω were connected across the secondary, its effect on the source connected across the primary would be that of a resistance of $4 \times 25^2 = 2500\ \Omega$.

*19.15 a.c. theory applied to the transformer

The treatment of the transformer can be developed more precisely by applying the methods of a.c. theory, as follows.

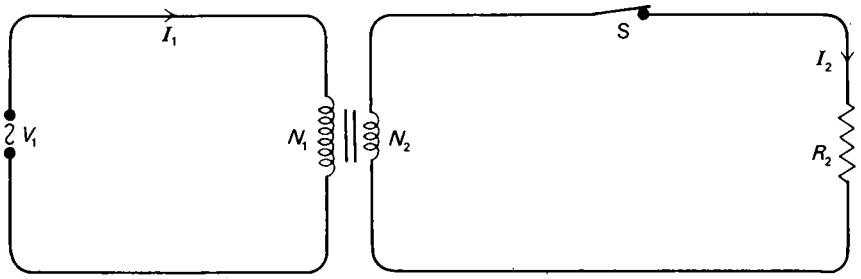


Fig. 19.14 A matching transformer

Assuming that both the applied p.d. and the resulting flux have good waveform, the instantaneous value of the flux is

$$\Phi = \Phi_0 \sin \omega t, \tag{19.21}$$

where $\omega = 2\pi f,$

and the p.d. applied across the primary, being equal and opposite to the induced e.m.f., is

$$V_1 = N_1 \frac{d\Phi}{dt}$$

$$= N_1 \omega \Phi_0 \cos \omega t. \tag{19.22}$$

Similarly, the induced e.m.f. in the secondary is

$$E_2 = -N_2 \frac{d\Phi}{dt}$$

$$= -N_2 \omega \Phi_0 \cos \omega t. \tag{19.23}$$

The corresponding r.m.s. relationships, therefore, are

$$(V_1)_{\text{r.m.s.}} = N_1 \omega \Phi_{\text{r.m.s.}}, \tag{19.24}$$

$$(E_2)_{\text{r.m.s.}} = N_2 \omega \Phi_{\text{r.m.s.}}, \tag{19.25}$$

whence

$$\left(\frac{E_2}{V_1}\right)_{\text{r.m.s.}} = \frac{N_2}{N_1}. \tag{19.14}$$

The inductive part, I_L, of the primary current, i.e. the part which flows whether current is taken from the secondary or not, is responsible for the flux. Neglecting hysteresis effects, this current is therefore in phase with the flux, and we can write

$$I_L = (I_L)_0 \sin \omega t, \tag{19.26}$$

and it follows, by comparing equation 19.22, that the current I_L is 90° out of phase with the applied p.d., which is the normal effect of inductance.

When current is taken from the secondary, assuming it to be in phase with the e.m.f. in the secondary, we have

$$I_2 = (I_2)_0 \cos \omega t, \tag{19.27}$$

and this current tends to produce its own contribution to the flux, viz.

$$\Phi' = kN_2 I_2$$

$$= kN_2 (I_2)_0 \cos \omega t. \tag{19.28}$$

The appearance of this flux would upset the balance in the primary between applied p.d. and induced e.m.f., and the result is that an extra current, I_1, now flows in the primary, which produces flux equal and opposite to Φ', and thereby destroys it,

i.e. $$-\Phi' = kN_1(I_1)_0 \cos \omega t, \tag{19.29}$$

where $$I_1 = (I_1)_0 \cos \omega t.$$

Thus, whereas I_L is 90° out of phase with the applied p.d., I_1 is in phase with it and accepts energy from the source.

Also, equating 19.29 and 19.28 gives

$$N_2(I_2)_0 = N_1(I_1)_0,$$

from which equation 19.16 follows:

$$\frac{I_1}{I_2} = \frac{N_2}{N_1}. \tag{19.16}$$

These results are summarized diagrammatically in Fig. 19.15.

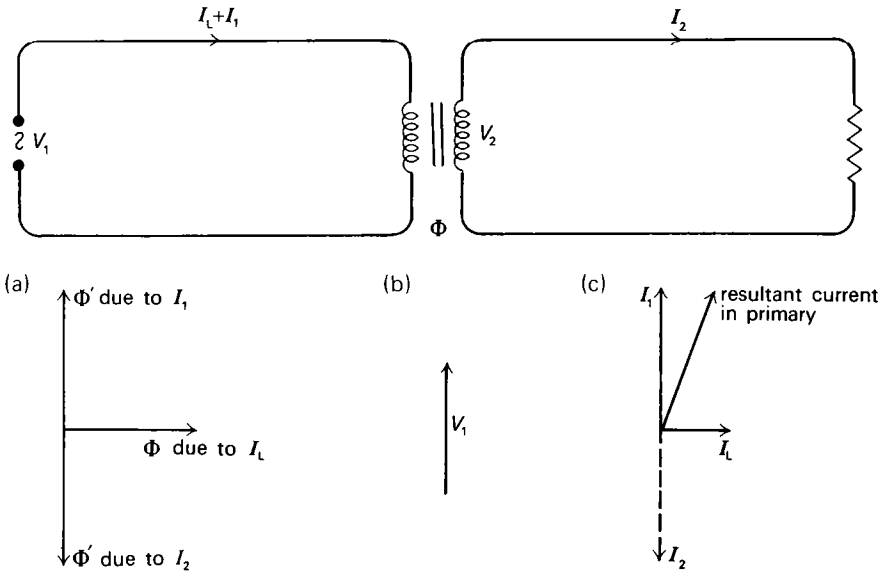


Fig. 19.15 The transformer. Phase relationships (a) The flux (b) The applied p.d. (c) The current

*19.16 The induction coil

The induction coil is a further application of mutual induction, having two coils wound round a straight core consisting of soft iron wires. Using a low voltage d.c. source connected across the primary, the output is a succession of very high voltage pulses in the secondary. Whereas the transformer depends for its action on alternating flux, the induction coil has a make and break device connected in series with the d.c. supply, and the necessary change in flux occurs each time that the primary circuit is broken.

Now that the electricity supply is universally an alternating one, the induction coil has largely lost its previous important place in the laboratory, being superseded

by an E.H.T. unit, made up from transformer and rectifier. It still has an essential function, however, in the motor car, because the sparking plugs require voltages much in excess of the 12 V d.c. supplied by the car battery. The make and break in this case, or contact breaker, is geared to the cycle of the engine, to provide the electrical impulses at the appropriate instants, and each reaches the appropriate plug via the distributor.

The principle of the construction is to have a comparatively small number of turns to act as the primary coil, wound round a bundle of soft iron wires (made in this form to reduce eddy currents), and the secondary coil is a large number of turns of fine wire wound outside the primary. The secondary is made up as a series of flat coils (Fig. 19.16) in order that adjacent coils shall have only small potential differences between them, thus avoiding internal sparking.

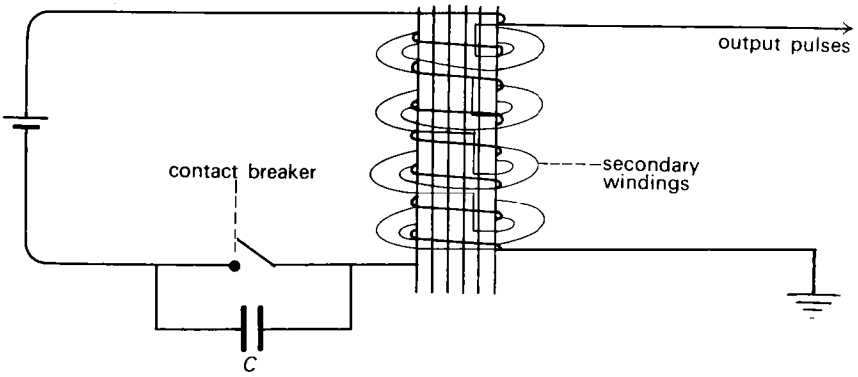


Fig. 19.16 An induction coil

The action of the induction coil is very similar to the behaviour of an inductor which has been connected in series with a d.c. source, and this was fully described in section 8.6. Thus, when the circuit is made, the primary current grows comparatively slowly to its final steady value owing to the effect of self inductance, but when the circuit is broken the current dies rapidly. Assuming that the flux varies in the same manner, being caused by this current, a large e.m.f. is induced, both in secondary and primary, at break, and only a small one at make, and the effect is particularly large in the secondary, being proportional to the number of turns, viz.

$$V_2 = N_2 \frac{d\Phi}{dt}. \tag{19.30}$$

There is good reason for using a core which allows the flux to emerge into air, because it is essential to the action that the flux shall vanish rapidly at break, and this would be less likely to happen if the iron made a continuous magnetic circuit for the flux, as it does in the transformer.

In practice, a capacitor is connected across the contact breaker, and this has the effect of accelerating the change in flux at break, and so increasing the induced e.m.f. The reason is that, in vanishing, the flux has energy $\frac{1}{2}LI^2$ to dispel, which it does by causing arcing at the contact, thus slowing down its decay. With the capacitor in place, however, this energy flows rapidly into the capacitor instead, and a high frequency damped oscillation results (Fig. 19.17).

The very marked difference between the behaviour of an induction coil and that of a transformer is due to the different circumstances in which they are used. The

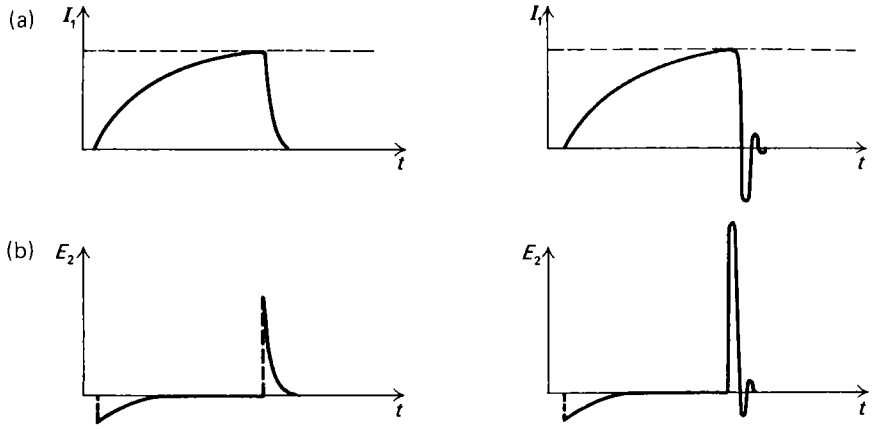


Fig. 19.17 Characteristics of the induction coil (a) With no capacitor (b) With a capacitor connected across the contact breaker

induction coil depends on transient effects, occurring when the circuit is made or broken, whereas the useful behaviour of the transformer concerns a continuous flow of alternating current; at the instant when the current is switched off, the transformer produces effects resembling those of the induction coil.

*19.17 Superconductivity

The high permeability of ferromagnetic materials is a boon to the electrical engineer, enabling him to obtain high flux densities without requiring magnetizing currents of prohibitive size. There is a limit, however, beyond which ferromagnetism ceases to have any further useful effect, and this is reached at a flux density of about 3 T (30 kilogauss). This is because, once saturation has been reached, the iron can make no further contribution to the flux, and the relative permeability falls.

Magnetic fields in excess of 3 T are required in some spheres of research, and these must be obtained by using a different technique, because the iron core might as well be dispensed with. In this case, the flux must derive entirely from the magnetizing current, and this involves a very heavy current and a coil with many turns. This may be illustrated by taking a long air-cored solenoid as an example, for which we have

$$B = \mu_0 \frac{N}{l} I.$$

Putting in the values $B = 3$ T and $\mu_0 = 4\pi \times 10^{-7}$ H/m, we have

$$\frac{N}{l} I = \frac{3 \times 10^7}{4\pi}$$

$$= 2 \times 10^6 \text{ ampere turns per metre.}$$

Using copper conductors having normal resistivity, there are two possible techniques: either a transient field is accepted, the magnetizing current being provided over a brief space of time by the discharge of a large bank of charged capacitors, or elaborate cooling arrangements are made to permit the very large current to flow continuously without undue overheating. There is a limit to what such cooling systems can do, however, because the power to be dissipated may exceed, say, 1000 kW.

What makes the maintenance of large magnetic fields possible over more than a brief period is the application of superconductivity, since this enables large currents to flow without any heating effects.

Superconductivity

Superconductivity was discovered by Kammerlinghe Onnes, not long after he had succeeded in liquefying helium for the first time. He found that mercury, below the temperature of 4.3 K, has literally no electrical resistance at all, and he presently discovered that other materials too, among them lead and tin, react to a very low temperature in the same way. It is as if the conduction electrons retain their mobility, while the vibrating ions lose their power to resist them. Not all metals, however, become superconducting, and copper, although its resistance becomes very low, is among the metals which do not.

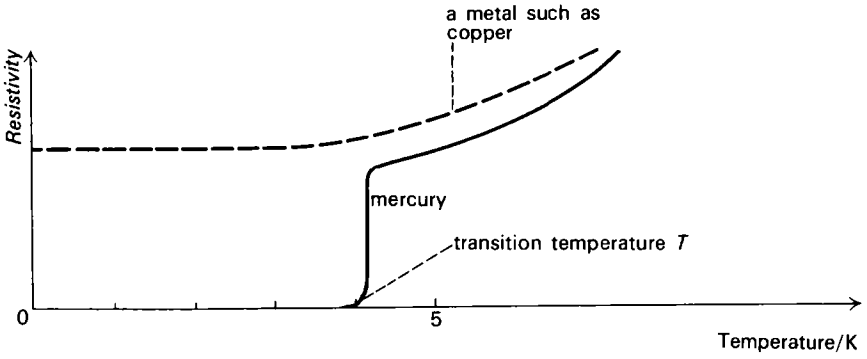


Fig. 19.18 Superconductivity. The variation in resistivity with temperature

If magnetizing coils can be made from superconducting metal cooled down by liquid helium, the problem of producing ultra high magnetic fields is solved, because, in the absence of resistance, the magnetizing current produces no heat. The same technique might be available to solve the problem of heat dissipation in other cases too. Onnes realized the existence of these possibilities, but he failed to develop them, and success has only been achieved since his time. This is because, with the metals used by him, the super conductivity is destroyed by a strong magnetic field, and cannot, therefore, be used to produce one.

The transition temperature T at which superconduction first occurs, depends on the flux density B at the surface of the metal (Fig. 19.19). The transition temperature of mercury, for example, is 4.3 K if there is no magnetic flux at all, but it reverts to normal conduction at this temperature if there is flux at its surface. There is a limiting value for this flux density, above which the mercury never becomes superconducting at all. A further difficulty is that, if 'hot spots' appear in the metal, owing to irregularities which make its conductivity normal at these places, a return to normal resistive behaviour tends to spread rapidly. Because of such effects as these, there is a critical value for the current density, above which super conduction is destroyed.

An aspect of the above effects is the fact that metals such as mercury, tin, and lead are completely diamagnetic below their transition temperature. When superconducting, the metal completely expels any magnetic flux from its interior, and the current can flow only in the outer shell of the metal. Above the transition

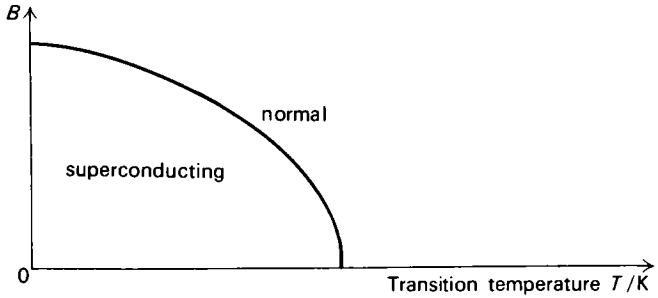


Fig. 19.19 Superconductivity. The dependence of transition temperature T on flux density

temperature, the power to exclude flux disappears, and the metal reverts to its normal behaviour.

Superconducting magnets can now be produced commercially because another class of superconducting materials, which behave rather differently from the metallic elements used by Onnes, has been discovered. These are alloys. They have comparatively high transition temperatures, up to 20 K in some cases, and, what is of more practical importance, they retain their superconductivity in the presence of high magnetic fields. Some of these alloys tend to be brittle, and consequently unsuited to winding in coils, but magnetizing coils are now made, using alloys of niobium and tin, or of niobium and zerconium.

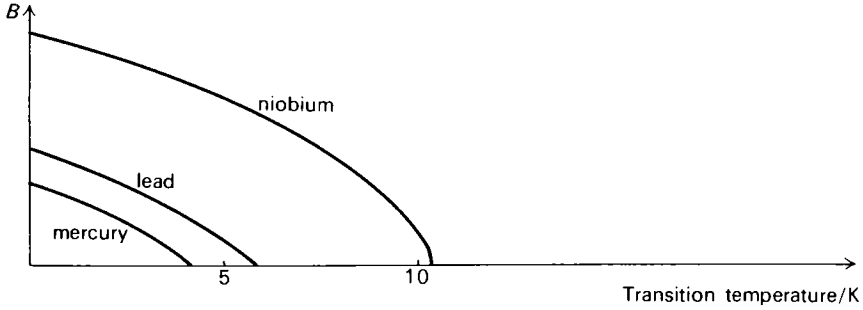


Fig. 19.20 Superconductivity. A comparison of the superconduction range of three elements

QUESTIONS

19.1 (*a*) A shunt wound motor running at 3000 rev/min is taking a current of 20 A from a 100 V d.c. supply. Assuming the resistances of the armature and the field coils to be respectively 0.5 Ω and 100 Ω, calculate (*a*) the back e.m.f. induced in the armature, (*b*) the torque exerted by the armature.
(*b*) A shunt wound d.c. dynamo has armature and field coil resistances of 0.5 Ω and 100 Ω respectively. If the p.d. across the output terminals of the dynamo is 100 V when it is delivering a current of 5 A to an external load, what is the e.m.f. of the dynamo?

19.2 When a shunt wound d.c. motor is connected across a 240 V d.c. supply, its unloaded speed is 12 rev/s and the armature current is 0.30 A. When a

certain load is placed on the motor, the armature current increases to 16.0 A. If the resistance of the armature is 3.0 Ω, calculate the running speed in this condition.

19.3 A small shunt wound d.c. motor drives a pulley of 5 cm diameter. Over the pulley is a cord, at one end of which is hung a 0.60 kg weight, the other end being anchored to a spring balance. When the motor is connected to a 13 V supply the current taken by the armature is 2.0 A, and the balance reads 0.10 kg weight, while the pulley rotates at 30 rev/s.

Find (*a*) the power which is being taken from the motor, (*b*) the back e.m.f. induced in the armature, (*c*) the armature resistance.
(Take $g = 10$ m/s^2.)

19.4 A shunt wound d.c. motor takes a current of 10 A from the 100 V d.c. mains when the rotation speed of the armature is 3000 rev/min. The armature and field windings have resistances of 1 Ω and 100 Ω respectively. Calculate (*a*) the back e.m.f. in the armature, (*b*) the rate at which energy is dissipated as heat in the windings, (*c*) the torque exerted by the armature on the load. (O & C)

19.5 The resistances of the armature and the field windings of a shunt wound machine are 0.20 Ω and 48 Ω respectively. When the machine is used as a generator, running at 300 rev/min, its output is 12.0 kW at 240 V. If the machine is used as a motor with a power consumption of 12.0 kW when connected to the 240 V d.c. mains, at what speed will it run? (L)

*19.6 A 100 V shunt wound d.c. motor having armature resistance 0.20 Ω runs at 500 rev/min when unloaded, and is taking an armature current of 0.50 A.

(*a*) What power is the armature taking from the supply? What power is being dissipated as heat in the armature resistance? What becomes of the remainder of the power taken from the supply?

(*b*) A load is now placed on the armature, and the current taken by it increases to 10 A. At what speed is it now running?

Assuming that the torque required to overcome friction at the bearings is constant, calculate the torque imposed by the load.

*19.7 An electric lift is operated by a shunt wound d.c. motor. The resistance of the armature and brushes is 0.8 Ω, and the brushes are connected across a 400 V d.c. supply. When the load in the lift cage is 800 kg, it is raised at a uniform speed of 2 m/s. The current in the armature is 50 A. With a reduced load in the cage, the armature current is 37.5 A when the speed is uniform. At what speed is this load being raised, and what load is the cage carrying?

The armature supply voltage, the motor field current, and the frictional resistance to motion, are the same in both operations.
($g = 10$ m/s^2) (CAM. SCH. ENG.)

19.8 Explain the importance of back e.m.f.s in the theory of the d.c. electric motor. Describe the main features and application of series and shunt wound d.c. motors. (O & C)

19.9 Describe the construction of a simple form of a.c. transformer.

If the secondary coil is on open circuit, explain, without calculation, the effect on the current flowing in the primary of (*a*) a fall in the supply frequency, (*b*) a reduction in the number of primary turns. (O & C)

*19.10 Explain the construction and action of an a.c. transformer. Describe the principal features of a transformer designed to work off the 240 V a.c. main supply, whose secondary is to deliver 100 A into a load whose resistance is 0.02 Ω.

A long coil of wire in the form of a solenoid is connected in series with an a.c. ammeter to the a.c. main supply. State what happens to the ammeter reading and explain why it occurs (*a*) when a thick cylinder of iron is placed inside the coil, (*b*) when the inside of the coil is filled with a bundle of iron wires, and (*c*) when a thick copper ring is placed around the centre of the coil. (O & C)

*19.11 Outline the principles of an alternating current transformer. An ideal transformer has N_1 primary turns and N_2 secondary turns. Show that, if a resistance R is connected across the secondary terminals, the current flowing in the primary circuit is the same as if the transformer were replaced by a resistance $R' = (N_1/N_2)^2 R$ connected across the primary terminals.

State precisely the conditions which have to be satisfied for such ideal behaviour, and explain why these conditions cannot be met in a practical transformer.

A transformer designed to operate at a frequency of 500 Hz is connected by mistake to a 50 Hz supply (of the correct voltage). Even with the secondary on open circuit, it is found that the transformer overheats. Suggest an explanation of this behaviour. (O & C)

*19.12 A transformer takes a current 5.0 A from 210 V mains when the secondary is supplying 18.0 A through a resistive load, the p.d. across the load being 50 V. In these circumstances, the power factor of the primary circuit is 0.99. The resistance of the primary winding is 2.0 Ω, and that of the secondary winding is 0.10 Ω.

Calculate (*a*) the efficiency of the transformer at this load, (*b*) the power loss in the iron core. (C)

*19.13 A hollow metal cylinder of radius r, length l, and wall thickness dr ($\ll r$) is placed with its axis parallel to a uniform alternating magnetic field of flux density B, where

$$B = B_0 \sin \omega t.$$

If the resistivity of the metal is ρ, and the self inductance of the cylinder can be neglected, show that the mean power dissipation in the cylinder is given by

$$P = \pi \frac{B_0^2 \omega^2 r^3 l \, dr}{4\rho}.$$

Hence find an expression for the mean power dissipation due to eddy currents in a solid cylinder of length l and radius r, with its axis parallel to a uniform field B, where

$$B = B_0 \sin \omega t.$$

Explain why the power dissipation can be reduced by laminating the cylinder, and indicate a suitable arrangement of the laminations. (O & C)

*19.14 A coil A passes a current of 1.25 A when a steady p.d. of 5.0 V is maintained across it, and an r.m.s. current of 1.0 A when it has across it a sinusoidal p.d. of 5.0 r.m.s. at a frequency of 50 Hz. Explain why the current is less in the second case, and calculate the resistance and inductance of the coil.

The same coil A, which has 100 turns, has a second coil B, with 500 turns, wound on it, so that all the magnetic flux produced by A is linked by B. Find the r.m.s. value of the p.d. that appears across the open circuited ends of B when a sinusoidal alternating current of 1.0 A at a frequency of 50 Hz is passed through A. Why is the ratio of this e.m.f. to the r.m.s. p.d. across A not the same as the ratio of the number of turns in B and A, i.e. 5:1?

Explain why the insertion of an iron core into the coils would decrease the current in A and increase the p.d. across B, if the alternating p.d. across A were kept unchanged. The effects of hysteresis and eddy currents in the iron may be neglected. (O & C)

Chapter 20

Electromagnetic radiation and the special theory of relativity

20.1 Induction and radiation fields

So far, in discussing electric and magnetic fields in this book, we have been concerned with the effects of static charge and of steady current, or else with a current which varies comparatively slowly. In these circumstances, the lines of force of every electric field originate at a charge Q, and the lines of flux of every magnetic field either thread a permanent magnet or they encircle a conductor which carries a current I. If we encounter either type of field, we can look for a charge, a magnet, or a current not very far away which is the cause of it. If, now, we have an electric charge which is not stationary, or slowly moving, but one which suffers rapid changes in its motion, a new phenomenon results, and energy is radiated away from the charge, in the form of rapidly alternating electric and magnetic fields which can be detected at places far distant from the charge which launched them. These fields are termed radiation fields, and they are additional to the shorter range fields which are subject to the laws of Coulomb and Ampere, which we distinguish by naming them induction fields.

Analogies to the effect of an accelerating charge can be drawn from mechanics, and we will take the case of a displacement propagated along a stretched cord. Suppose that we hold one end of a long horizontal cord, the far end of which is fixed to a point on a distant wall. If the near end of the cord is slowly raised to a new level and held there, farther parts of the cord are raised to some extent in consequence, the effect progressively decreasing with distance until, at some comparatively short range, it has become scarcely noticeable. The displacement does not appear at all parts of the cord simultaneously, but the disturbance requires time to travel. If now, instead of slowly raising the free end of the cord, we raise it and lower it again in rapid succession, the effect is markedly different. This time, a hump is launched, which can be seen to travel the full distance down the cord being reflected from the far end, as a trough, which comes all the way back again. Thus the abrupt disturbance is appreciable over a much greater range than the almost static one, but of course the static one is permanent, so long as the static displacement which causes it is maintained, whereas the pulse travels to and fro, due to reflections, until it is finally damped out.

20.2 An electromagnetic wave

The process by which an electric or a magnetic field is first set up has characteristics which are similar to those of the disturbance in the cord. If the current is suddenly switched on in a conductor, the magnetic flux does not appear everywhere instantaneously but spreads outwards from the conductor with a finite speed, although a very rapid one, and this applies to both the induction field and the

radiation field. If the current in the conductor is oscillating, particularly at a very high frequency, the result is similar to what the result of oscillating the free end of the analogous cord would be, and the radiation field takes the form of an oscillating field spreading outwards as a wave.

This oscillating magnetic field which is propagated as a wave is associated with an electric field, in a manner which does not apply in the case of the shorter range induction fields. Whereas a stationary electric charge produces only an electric field, and a steady current in a conductor produces only a magnetic one, the oscillating magnetic field of the wave is inevitably accompanied by an oscillating electric field which has the same frequency and is propagated with the same speed, and the ratio in strength of the two at any point, in a vacuum, is always the same. For this reason the wave is referred to as an electromagnetic wave.

Figure 20.1 represents an electromagnetic wave which is travelling in the direction OZ. This particular example would be described as a vertically polarized wave, because its electric field is in a vertical plane; the associated magnetic field is always perpendicular to the electric one. An observer stationed at the point Z

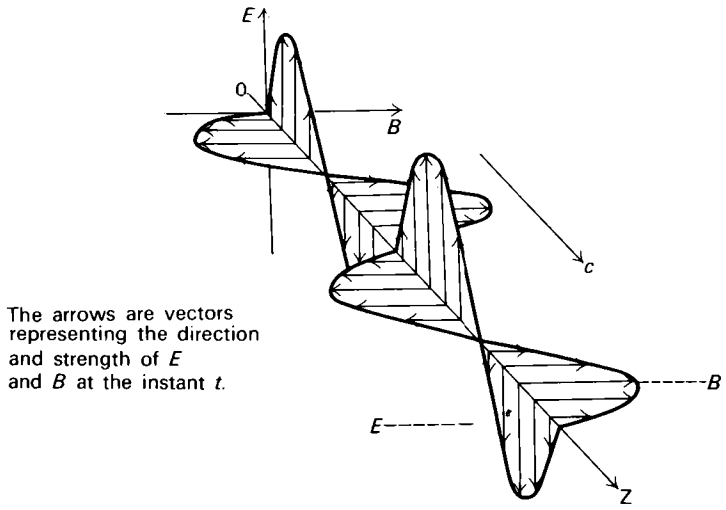


Fig. 20.1 A vertically polarized electromagnetic wave

would observe a rapid oscillation in the value of the fields E and B as the wave travels through Z with speed c. As with all types of wave, the wavelength λ is related to the frequency f and speed c by the equation

$$c = f\lambda. \tag{20.1}$$

20.3 Light as an electromagnetic wave

The equations of the electromagnetic wave were first derived by Maxwell in 1865, and such waves were produced from an electric oscillating circuit for the first time by Hertz in 1887. These first radio waves belonged to the waveband which is now classified as VHF, having a wavelength a little shorter than one metre.

Maxwell's equations showed the speed of the waves in a vacuum to be given by the equation

$$c = \frac{1}{(\mu_0 \epsilon_0)^{\frac{1}{2}}}, \tag{20.2}$$

μ_0, ϵ_0 being the permeability and the permittivity of a vacuum. This works out to the value 3.0×10^8 m/s, which is the same as the speed of light. Thus it seemed probable from the start that light must be an electromagnetic wave. Its place in the electromagnetic spectrum is shown in Fig. 20.2.

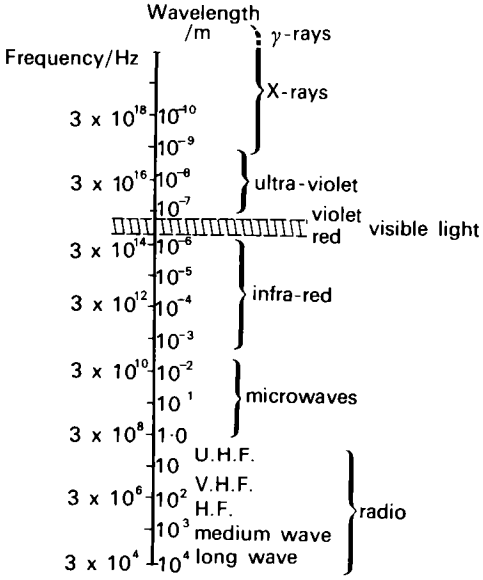


Fig. 20.2 The electromagnetic spectrum

The final acceptance of the wave theory of light, largely as the result of the work on interference and diffraction by Young and Fresnel, had preceded Maxwell by half a century, but the work of Maxwell and Hertz resolved two outstanding problems that had remained, namely how it comes about that light can be polarized, and how it contrives to cross a vacuum. Before Maxwell, light was supposed to involve mechanical vibrations of some kind, and this necessitated postulating that space was filled with the aether, a material capable of sustaining these vibrations. The aether might be supposed to have the elastic properties of a gas, but scarcely those of a solid. Yet the elasticity possessed by a gas, which is resistance to compression, permits only longitudinal waves, and these cannot be polarized. Maxwell's theory resolved both these difficulties since electromagnetic waves, being transverse (Fig. 20.1), can be polarized and they do cross a vacuum. Yet his theory does not entirely remove the significance of the aether, because space is still demonstrated to possess certain properties, albeit not mechanical ones. It must have electrical properties analogous to elasticity and inertia, permitting the propagation of alternating fields at a particular speed, and these fields carry energy with them. It is these properties which are represented by the two constants ϵ_0 and μ_0, upon which the speed of the wave depends.

20.4 The basis of Maxwell's equations

Maxwell based his equations on the postulate that electric and magnetic radiation fields could exist in a vacuum in addition to the familiar induction fields of previous theory, and that the electric and magnetic components sustained each other and so were never found independently of each other. A better analogy to electro-

magnetic radiation in space is the propagation of a sound wave in air, rather than a wave in a cord: the energy in the sound wave is shared between the kinetic energy due to the inertia of the moving gas and the potential energy due its elasticity when it is compressed or rarefied; and the speed with which the sound wave travels in the gas depends on its density and on its elasticity. On this analogy, a vacuum has the two properties which correspond to these electrically, μ_0 taking the place of the density and ϵ_0 (or, more accurately, $1/\epsilon_0$) the elasticity, and it is in sustaining both a magnetic and an electric field that the vacuum can transmit a wave.

The lines of force of an induction electric field originate and terminate on electric charges, but the lines of force of a radiation field do not do this; they encircle lines of magnetic flux instead, but these must be lines of changing flux. Similarly the lines of magnetic flux in an induction field encircle a conductor which carries current, but lines of flux of a radiation field are different; they encircle electric lines of force instead, but these must represent a changing electric field. Thus, so long as the source is an accelerating charge, thereby providing in the first place both a varying electric field and a varying magnetic field, a situation is created where the complementary electric and magnetic fields postulated by Maxwell can maintain each other as they carry their energy away into space.

[The equations which summarize the characteristics of these different fields are as follows:

Induction fields

Gauss' equation (section 13.18)
$$\oint \epsilon_0 E dA = \Sigma Q \tag{20.3}$$

Ampère's equation (section 15.4)
$$\oint \frac{B}{\mu_0} ds = \Sigma I \tag{20.4}$$

Radiation fields

Maxwell's equations
$$\oint E ds = A \frac{dB}{dt} \tag{20.5}$$

$$\oint \frac{B}{\mu_0} ds = A\epsilon_0 \frac{dE}{dt} \tag{20.6}$$

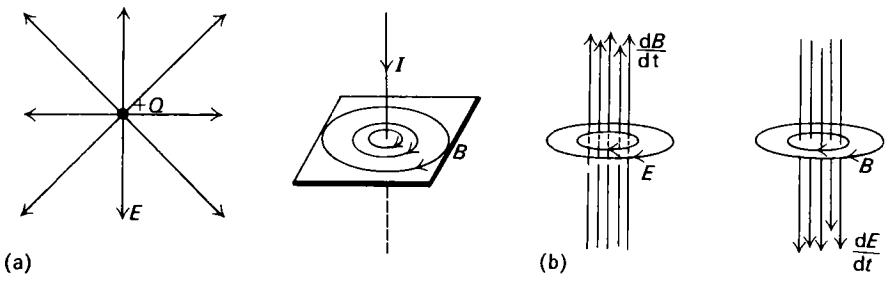


Fig. 20.3 Induction and radiation fields (a) Induction fields (b) Radiation fields

20.5 The relationship between electric and magnetic fields in a wave

(a) The effect of a varying magnetic field

The concept of an electric field consisting of lines of force which encircle a varying magnetic field is an extension of Faraday's law of electromagnetic induction.

Suppose that a magnetic field having flux density B threads a conducting loop which encloses area A of the field. If the flux is changing, an e.m.f. V is induced in the conducting loop and, by Faraday's law, we can write

$$V = \frac{d\Phi}{dt},$$

i.e.

$$V = A\frac{dB}{dt}. \tag{20.7}$$

This e.m.f. is distributed round the loop and represents an electric field, or potential gradient, in it. If we represent the strength of this field at any point by E, we have

$$E = \frac{dV}{ds},$$

whence

$$V = \oint E ds, \tag{20.8}$$

the integral being taken over the complete loop. Eliminating V between equations 20.7 and 20.8, gives

$$\oint E ds = A\frac{dB}{dt}. \tag{20.5}$$

It would seem reasonable to expect that, if the conducting loop were removed, the induced electric field E would still be present, giving the result that a varying magnetic flux is in all circumstances encircled by an electric field, given by equation 20.5. The existence of such a field is, in fact, demonstrated in the action of a type of accelerator, not described in this book, called the betatron, in which it is utilized to accelerate electrons to a very high energy in an evacuated circular track.

(b) The effect of a varying electric field

We now consider the circuit of Fig. 20.4, where a parallel plate capacitor is in process of accepting charge from a battery.

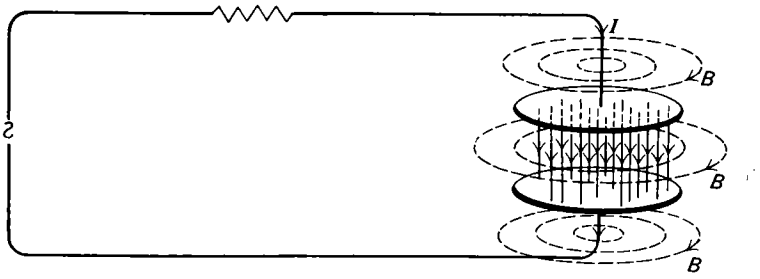


Fig. 20.4 The magnetic field encircling the increasing electric field of a charging capacitor

At the particular instant under consideration, a current I is flowing in the connecting wires because the capacitor is in process of charging up. Consequently, lines of magnetic flux encircle these wires, and Ampère's equation applies:

$$\oint \frac{B}{\mu_0} ds = \Sigma I. \tag{20.9}$$

If we now direct our attention to the space separating the capacitor plates, there is no current here and the circuit is, in this respect, incomplete. There is, however, an electric field E between the plates, which is increasing in value as a direct result of the current flowing in the connecting wires. The relation between the two is as follows.

The field strength E between the plates, due to the charges $Q, -Q$ which already reside on them, is

$$E = \frac{1}{\epsilon_0} \cdot \frac{Q}{A},$$

where A is the area of each plate. But the current I is increasing the charge held by the capacitor, and we have

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t},$$

whence, from these two equations,

$$I = A\epsilon_0 \frac{\mathrm{d}E}{\mathrm{d}t}. \tag{20.10}$$

(Maxwell called this quantity $A\epsilon_0 \mathrm{d}E/dt$ a displacement current.)

Now it would seem reasonable, if magnetic flux lines encircle the current in the conductors, to assume that they also surround the growing electric field that exists between the plates. Combining equations 20.9 and 20.10, we then have

$$\oint \frac{B}{\mu_0} \mathrm{d}s = A\epsilon_0 \frac{\mathrm{d}E}{\mathrm{d}t}. \tag{20.6}$$

20.6 Transmitting and receiving aerials

Any single conductor which carries alternating current radiates energy away, provided that the frequency is high enough. Other things being equal, the power radiated is proportional to the square of the frequency, which is therefore a very important factor, and radiation at a frequency of 50 Hz, for example, is negligible. (Reference to equation 20.1 would show that a wave of this frequency would, in any case, have the wavelength of 6000 km, which is about equal to the earth's radius.)

A twin conductor, used to convey both outward and return currents, does not radiate efficiently unless the wavelength is extremely short, because the currents in the two conductors, and therefore their fields, are out of phase with each other, and so the transmitting aerial is opened out and preferably made about a half wavelength long, forming a half-wave antenna, as in Fig. 20.5. This shows such an antenna being supplied at its centre with high frequency alternating current from an oscillating source, and this current charges the ends of the antenna in the same way as it would charge the opposite plates of a capacitor; indeed, there is capacitance between the two extremes of the antenna, and an electric field links them when it is charged up. The diagram shows the radiated electric field, but not the magnetic field which goes with it.

For the medium and long wave broadcasting bands, a half-wavelength antenna would be too long to be practicable, and for these long wavelengths the radio frequency signal is applied between the lower end of a vertical conductor and the earth, so that only a quarter wavelength of conductor is required, the other quarter wavelength being represented by its image in the conducting earth. Even the remaining quarter wavelength is too long, and this is shortened still further by

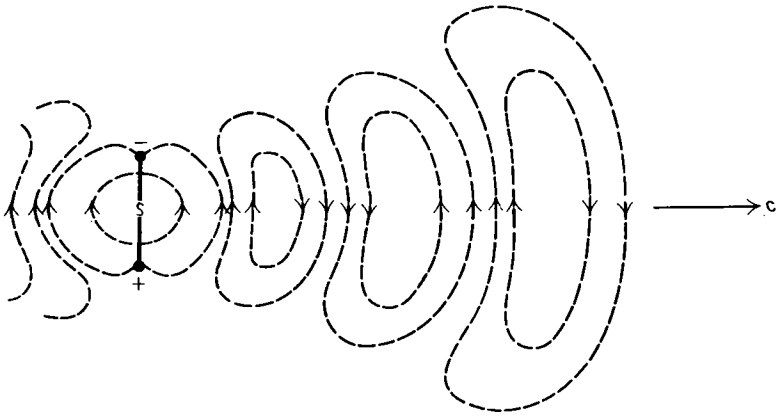


Fig. 20.5 The electric field radiated from a half-wave aerial antenna (There is also a magnetic field which is perpendicular to the plane of the diagram)

attaching a system of horizontal conducting wires at the top end to increase its capacitance, and by including an inductance in series at the lower end.

The receiving aerial

The receiving aerial may be similar to that used for transmission, in which case the energy is accepted by the agency of the electric field, which causes an oscillating

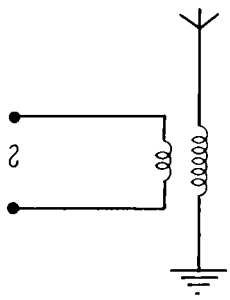


Fig. 20.6 A transmitting aerial for medium and long wavelengths

p.d. to appear across the aerial. Alternatively, the magnetic field of the wave may be collected by using electromagnetic induction to produce an e.m.f. in a coil, and this is the method normally adopted for medium and long wave reception. In this case, a horizontal ferrite rod is placed at right angles to the direction of the wave, in the plane of the magnetic field, so that the varying magnetic flux of the wave shall point along it, and the aerial coil is wound round this rod. The high permeability of the ferrite rod has the effect of concentrating inside it the magnetic flux of the radio wave in the area surrounding it.

It might be thought that, for maximum effect, both the electric field and the magnetic field should be separately picked up, and the two results added together. This is not the case, however; although only half the energy is in the electric field, all the energy can be taken from it, because if the whole of the electric field were to vanish into an aerial, the magnetic field would automatically vanish also, and both would part with their energy.

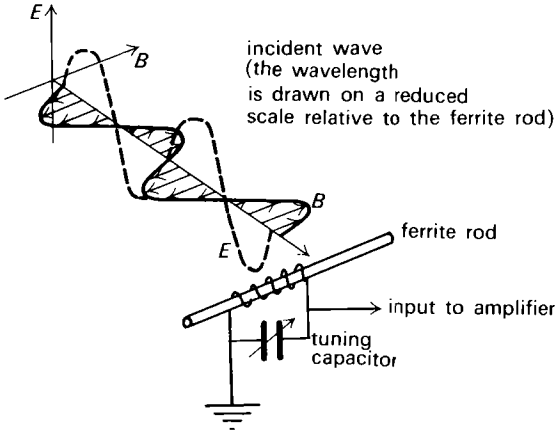


Fig. 20.7 A ferrite aerial rod

20.7 Detection and modulation

The input circuit of the receiver is an L–C circuit tuned to resonate to the frequency of the wave, for example, as shown in Fig. 20.8. To detect the wave, a suitable meter may be connected to this circuit, and this could be a thermocouple type or a moving coil meter in series with a rectifying diode. The oscillating p.d. or current can, of course, be amplified before detection.

The capacitor C_2 shown in the circuit is a smoothing or decoupling capacitor, which short circuits the radio-frequency component of the rectified current, leaving the direct current component to flow through the meter.

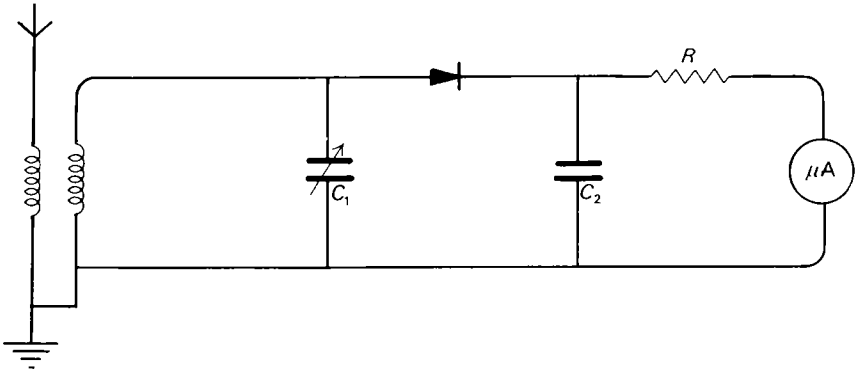


Fig. 20.8 The detection of a radio wave

Amplitude modulation

A broadcast radio wave, however, is required to carry with it a signal at audio-frequency, which is a much lower frequency than that of the radio wave itself, and this is done by modulating the wave. Suppose, for example, that an audible tuning note, having frequency 500 Hz, is to be broadcast on a wavelength of 330 m, which is a frequency of 908 kHz. The audio-frequency signal is superimposed on the radio-frequency as a variation in the amplitude of the radio wave, a procedure known as amplitude modulation. The frequency of the modulation is the frequency,

500 Hz, of the audio signal, and the depth of the modulation governs its intensity.

Using the detecting circuit shown in Fig. 20.8, the decoupling capacitor C_2 would have only small capacitance, sufficient for it to decouple the radio-frequency, but not the audio-frequency component of the current. The current through the

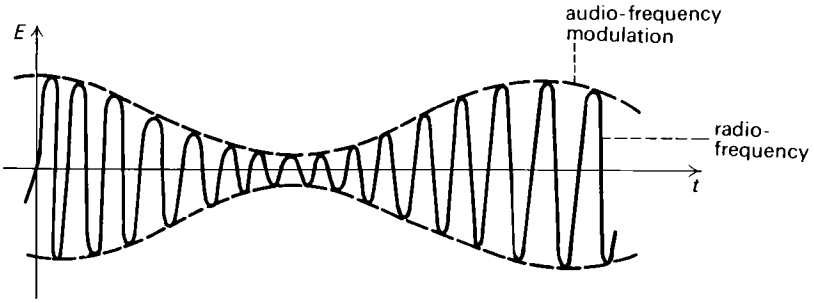


Fig. 20.9 Amplitude modulation

meter is consequently modulated at the audio-frequency, and there is an oscillating p.d. at audio-frequency across the resistor R. This p.d. can be amplified, or a telephone may be substituted for the meter, providing a simple crystal receiver.

Frequency modulation

An alternative way to superimpose the audio-frequency signal is by frequency modulation, which affords better reception because it is less subject to interference and other annoyances. Like television signals, however, this can only be used at V.H.F. and U.H.F. because it takes up a wide band of frequencies, and this is acceptable only where the frequency is very high. In frequency modulation, instead of modulating the amplitude, the frequency is repeatedly varied over a range, the repetition rate of this sweep being the audio-frequency signal.

20.8 The special theory of relativity

It is hardly necessary to point out that, in bringing energy to us across a vacuum, electromagnetic radiation is for us of unique practical importance. Because it is a wave propagation in a vacuum, it is also of unique importance in fundamental physics, in the development of Einstein's special theory of relativity.

Making due allowance for the rotation of the earth about its axis and for its orbit round the sun, we inhabit an inertial system. This has the speciality that our frame of reference, e.g. the walls of the laboratory, is not accelerating, although it may, for all that we can tell, be moving through space with a constant velocity, possibly a very high one. There is no way for us to tell what this velocity is, because in all inertial systems the observed laws of physics are the same. Newton's laws of motion, for example, apply in them all: it does not require a force to maintain a body's motion at constant velocity, any more than it requires a force to keep a body at rest.

According to classical physics, there is one exception to this universal consistency in behaviour, and this concerns an electromagnetic wave, such as a light wave. The velocity of a wave depends entirely on the medium which propagates it, and the velocity of its source makes no contribution to this; the ripples launched by a duck in still water spread out in circles in the water from their points of origin, unaffected by the motion of the duck. Likewise, since the propagating medium is a vacuum,

light waves should spread in circles in the stationary vacuum, thus marking out reference points of absolute rest in space. It should be possible on this basis, although not easy (owing to the high speed with which light travels) to detect our passage through this stationary vacuum as we orbit round the sun. Several careful attempts to do this were made towards the end of the last century, notably by Michelson and Morley, and they should have been successful, but they were not. By the start of the present century, it had become clear, without reasonable possibility of doubt, that the measured speed of light was the same whether the observer was moving towards it or moving away from it; the possibility of escaping from a beam of light by moving away from it does not exist.

Einstein solved this paradox by accepting it as a basic physical law. His special theory of relativity states that all physical laws which apply in one inertial system apply equally in any other, including the fact that an electromagnetic wave in a vacuum is observed to have the same velocity c. This statement is not consistent with some of the derived equations of Newtonian mechanics, nor is it consistent with classical methods of reasoning; if we move with velocity v towards a beam of light which is approaching with velocity c, we expect to find that the relative velocity is $(c+v)$, not that it remains equal to c just as if we had not moved at all.

Einstein showed that the flaw in the classical point of view is an unjustified assumption of the absolute nature of time. If, for example, a time t elapses between two events which occur in succession at the same place, the time interval would not measure up as being equal to t in the experience of an observer who was moving past with velocity v. He would judge the time interval to be t', which is slightly larger than t, being given by

$$t' = t\left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}}. \tag{20.11}$$

According to Einstein's theory, there is no such thing as absolute rest, or, if there is, we have no means to recognize it. The only distinction that we can measure between two inertial systems is the velocity of one relative to the other.

20.9 Relativity in mechanics

If classical ideas are faulty about the measurement of time, this must affect the validity of some, at least, of the classical equations which are concerned with time, particularly the equations of dynamics, and discrepancies might be expected between classical theory and practical experience. The fact that such discrepancies have taken so long to appear and that Newton's equations are so successful is due to the very large value of the speed of light; equation 20.11 is typical of the type of the correction required by relativity, where it can be seen that in practice, if the velocity v under consideration falls very short of the velocity of light c the size of the correction becomes negligible.

Where relativistic dynamics is relevant to electrical problems is in the action of particle accelerators, designed to bring charged particles up to high speeds. An accelerating p.d. of 3 kV is sufficient to bring electrons within one tenth of the speed of light, but even heavier particles such as protons reach relativistic speeds in a modern accelerator.

Anything more than a very brief reference to relativistic dynamics would be out of place here, but we do require some acquaintance with the relativity aspects of mass, momentum, and energy. The most important consideration with respect to these is that the fundamental physical laws of conservation of momentum and conservation of energy retain their importance in relativity, and conservation of

mass becomes identified with conservation of energy. An explanation of these statements can usefully start with Newton's second law of motion.

Newton's second law of motion

According to this law, if we choose an appropriate unit of force F, the force acting on a mass m is equal to the rate at which it changes the momentum p of the mass,

i.e.
$$F = \frac{dp}{dt}, \tag{20.12}$$

where
$$p = mv. \tag{20.13}$$

This relation is also valid in relativistic dynamics and leads to the principle of conservation of momentum.

In classical mechanics, however, it is assumed that the mass m of a body is a constant quantity, independent of its velocity, which leads to Newton's second law being alternatively expressed in the form

$$F = ma, \tag{20.14}$$

where $a = dv/dt$ is the acceleration of the body. Thus, since the acceleration caused by a constant force is constant, the conclusion is that the body can gain speed indefinitely under the influence of the force, eventually, perhaps, reaching a value which exceeds that of light.

This is where relativity leads to a different conclusion, which is the one borne out in practice by the behaviour of accelerated electrons. According to relativity, it is impossible for the body quite to attain the speed of light, and its acceleration begins to fall off as it approaches this speed, until eventually the force fails to cause any further acceleration at all. Thus, we now have a body which suffers little or no acceleration although it is acted upon by a force, and the interpretation put on

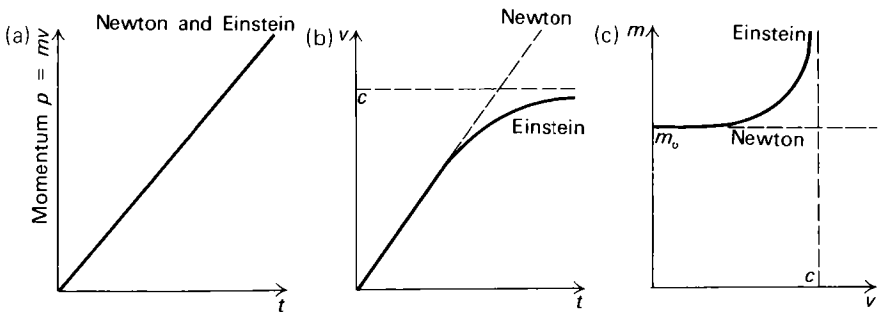


Fig. 20.10 Newton's second law of motion – the effect of a constant accelerating force (a) Momentum (b) Velocity (c) Mass

this is that its inertial mass m must have grown very large. To summarize, the relativistic point of view is that equation 20.12 is valid, and a force does increase the momentum of a body indefinitely, but the mass m of the body, which features in equation 20.13, is not constant, and at very high speeds an increase in momentum is an increase in mass rather than in velocity. The variation of mass with velocity is given by the equation

$$m = m_0\left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}}, \tag{20.15}$$

where m_0 is the rest mass of the body.

Kinetic energy

The relativistic equation for kinetic energy is

$$E_k = (m - m_0)c^2, \tag{20.16}$$

and this reduces to the familiar Newtonian expression

$$E_k = \tfrac{1}{2}m_0 v^2 \tag{20.17}$$

in the ordinary circumstances when the velocity v is much less than c. (This follows by substituting for m from equation 20.15.)

Mass and energy

Equation 20.16 is interpreted as meaning that a mass m_0 at rest is already in possession of an amount of energy W_0 given by

$$W_0 = m_0 c^2, \tag{20.18}$$

i.e. that mass and energy are, at any rate in theory, interchangeable. If a body loses an amount of energy ΔW, we would expect its mass to decrease in consequence by an amount Δm, where

$$\Delta m = \frac{\Delta W}{c^2}. \tag{20.19}$$

In ordinary circumstances, the factor $1/c^2$ is small so that this change in mass is much too small to detect, but it is of great importance in the equations of nuclear reactions.

20.10 Relativity and electromagnetism

Applied to electromagnetism, the special theory of relativity does not reduce classical equations to the status of being very good approximations, as happens in dynamics. Instead, all the equations are left untouched and are shown to be already relativistic equations, and the theory offers further explanations of them, interpreting magnetic forces as being due to the effect of relative motion on electrostatic forces. A simple example of this is the magnetic field surrounding current in a long straight conductor, and the basis of the argument in this particular case is discussed below.

Suppose that the metal conductor shown in Fig. 20.11 is carrying current I, caused by electrons drifting along it with the mean drift speed v given by

$$I = NAev, \tag{20.20}$$

N being the number of conducting electrons in unit volume and A the area of cross-section of the conductor.

Two charges $-Q$,$-Q'$ are shown in the diagram, both at the same distance r from the conductor, but the first charge $-Q$ is stationary, whereas the second charge $-Q'$ is supposed to be moving parallel to the conductor in the same direction as the conduction electrons are moving and with the same velocity v as their mean drift velocity.

The practical experience is that the force F experienced by the stationary charge Q is zero, while the moving charge Q' experiences a force F' attracting it towards the conductor which is given by

$$F' = BQ'v, \tag{20.21}$$

where

$$B = \frac{\mu_0}{4\pi} \cdot \frac{2I}{r}.$$

This force is judged to be magnetic in origin, and there is no electrostatic force acting on either charge because the conductor is electrically neutral, having stationary positive ions in it which neutralize the charge of the drifting electrons.

Now let us suppose that an observer proceeds to move in a path parallel to the conductor with velocity v, so that he keeps pace with the drifting electrons in the conductor and the charge $-Q'$ has the appearance to him of being stationary. The current which flows in the conductor is, in his estimation, caused by the positive ions in it, which appear to be moving, but the charge $-Q'$ experiences no magnetic force because it is stationary. But this is not to say that all force acting on the charge has miraculously vanished merely because the observer of it has started to move; what happens is that he still observes the force but he changes his interpretation of it. In his view, owing to the relative motion, the conductor now holds an excess concentration of positive charges and is exerting on the negative charge $-Q'$ a force of electrostatic attraction. Thus a force which appears to a stationary observer to be caused by magnetism is interpreted by him instead, when he is moving appropriately, as being electrostatic in origin, i.e. a magnetic force is a relativistic edition of an electrostatic one.

A similar argument applies to the observer's interpretation of the absence of force on the other charge $-Q$. When he is stationary, he sees a stationary charge in a magnetic field and expects no force to act on account of that; nor is there any electrostatic force, because the conductor is electrically neutral. When the observer is moving, however, in the same manner as before, and the conductor appears to be positively charged, he finds the negative charge $-Q$ to be attracted by it. At the same time, he sees this negative charge to be moving in the magnetic field of the conductor in such a direction that it is magnetically repelled from the conductor, and the zero force is the result of these equal and opposite electric and magnetic forces cancelling each other out.

The view taken by the observer of the metallic conductor remains to be explained; why is it apparently neutral in one case, and positively charged in the other? This is the consequence of a relativistic contraction in length. If the mean separation of the positive ions in the metal is l and the observer is moving parallel to the conductor with velocity v, this separation appears to him to have contracted to the smaller distance l', where

$$l' = l\left(1 - \frac{v^2}{c^2}\right)^{\frac{1}{2}}. \tag{20.22}$$

At the same time the electrons in the metal have a velocity v relative to the observer when he is stationary, and it is consequently towards the stationary observer that they show this apparent contraction; towards the observer who is moving

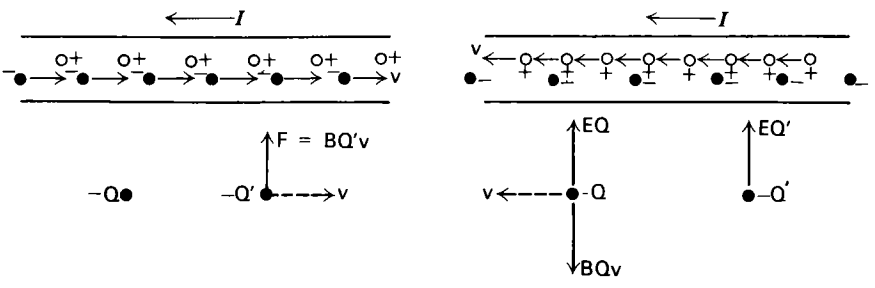


Fig. 20.11 The effect of current in a straight conductor (a) The view taken by a stationary observer (b) The view taken by an observer who is moving towards the right with velocity v

with them they show their normal separation. It follows that the moving observer sees a conductor in which the positive ions are put closer together due to the relative motion and the electrons are put further apart, thus giving it an excess of positive charge in any given length.

A very remarkable aspect of this relativistic account of electromagnetic forces is the fact that the relative velocity v which is responsible is only very small. Current in a metallic conductor seldom involves electron drift speeds which are greater than a fraction of 10^{-2} m/s. The relativity effect would, in consequence, be trivial, were it not for the extremely large size of electrostatic forces, as is indicated, for example, by the very large value of the electrostatic constant $\alpha(=1/4\pi\epsilon_0)$. It has been remarked before that the balance of positive and negative charges in nature is very finely kept, so that the full strength of this force is seldom revealed (except that it holds rigid bodies together), but it is large enough for the slightest relativistic disturbance of this balance to have appreciable effect.

Chapter 21

Conduction by ionized gases. The electron

21.1 Charge carriers in a gas

A gas such as air, hydrogen, nitrogen, consisting of neutral molecules, has no free charge carriers and is normally an almost perfect insulator. Before it can conduct, an appreciable number of atoms or molecules must first become ionized by losing one or two electrons. This ionization occurs if a very strong field is applied, as, for example, in a lightning flash, or in corona discharge from a high voltage cable, the value of the breakdown field in air at atmospheric pressure being about 3×10^6 V/m. The effect of this strong field is due to a self ionizing process; starting with a few ions, the field produces many more. There are at all times a few ionizing radiations or ionizing particles in the atmosphere, which are responsible for a very slight ion concentration in the gas, too small to provide more than an almost undetectable current. In a strong field, however, these sparse ions and released electrons gain kinetic energy. If this energy is sufficient, the electrons make inelastic collisions with gas molecules, causing further ionization, and in this way the ion concentration builds up. The gas can now carry an appreciable current.

If the pressure of the gas is reduced, the mean free path of the ions and electrons increases, with the result that they gain more energy in the field between collisions, and the field strength required to cause further ionization is consequently reduced. For this reason, the breakdown field strength is directly proportional to the pressure, unless this has been reduced to a very low level indeed. There must clearly be a lower limit to this proportionality, however, since a vacuum cannot conduct at all; at the very lowest pressures, the gas begins to resist strong fields again.

The effect on a gas of reducing the pressure may easily be demonstrated in the case of air by setting up a simple discharge tube connected to a vacuum pump and

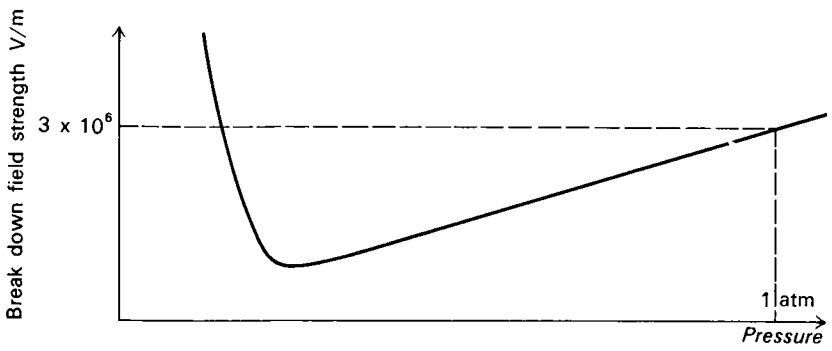


Fig. 21.1 The breakdown field strength for air – variation with pressure

applying a high p.d. across the electrodes from an induction coil or an E.H.T. unit. At atmospheric pressure there is no conduction, but, if the tube is now slowly exhausted, a silent, luminous discharge soon begins.

Conversely, if the pressure of a gas is increased, it can withstand an even stronger field than it did before, and application of this in the operation of a Van de Graaff generator has already been referred to (section 6.14).

21.2 Conduction at atmospheric pressure – the effect of ionizing influences

The ion concentration in the air may be temporarily raised above its normal very low level by passing ionizing particles or radiations through it, and in this condition the air becomes slightly conducting even when the field is small. If, for example, a charged electroscope is placed near an X-ray tube or near radioactive material, the electroscope slowly discharges, which was one of the early ways of detecting radioactive substances. The variation of current with p.d. between two electrodes placed in air ionized in this way is shown in Fig. 21.2. The horizontal part of this graph represents saturation, the ions being carried across between the electrodes as frequently as they are produced; below saturation, the very small current is described by Ohm's Law. When the p.d. across the electrodes is raised sufficiently, the breakdown value is eventually reached, when a rapid increase in ion concentration results, and a considerable current flows.

Fig. 21.2 Air exposed to ionizing influences. The variation of current with applied p.d.

[To obtain readings for plotting the part of this graph which lies below the saturation level, something more sensitive than a discharging electroscope or a pulse electrometer would be required because the currents to be measured may be less than, say, 10^{-11} A. The procedure is to connect a very high resistance in series with the ionization chamber, and measure the p.d. across this by using an electrometer valve and a d.c. amplifying circuit (section 18.6).]

21.3 Conduction by gas at reduced pressure

When a gas is enclosed in a discharge tube and the pressure is reduced, ionization, and consequently conduction, occur more readily, as we have already pointed out. The luminous effects are also quite spectacular. If a long discharge tube containing

air is connected to a vacuum pump and a p.d. of about 1 000 V is applied across the electrodes, these luminous effects go through the sequence sketched in Fig. 21.3.

In the condition shown in diagrams (*b*) and (*c*), the emitted light is quite bright and has the line spectrum which is typical of the gas, so that the colour depends on the gas. This is the condition of the conducting gases used in advertising signs and in the discharge tubes used for studying spectra. At all the pressures which are represented in the diagrams, the discharge follows the path provided by the gas, between cathode and anode, even if the tube is twisted into complicated shapes as with advertising signs.

The most interesting effect, however, although not the most spectacular, occurs when the pressure is reduced even further, below the value represented by diagram (*e*); this led to the discovery of the electron, and is discussed in more detail below.

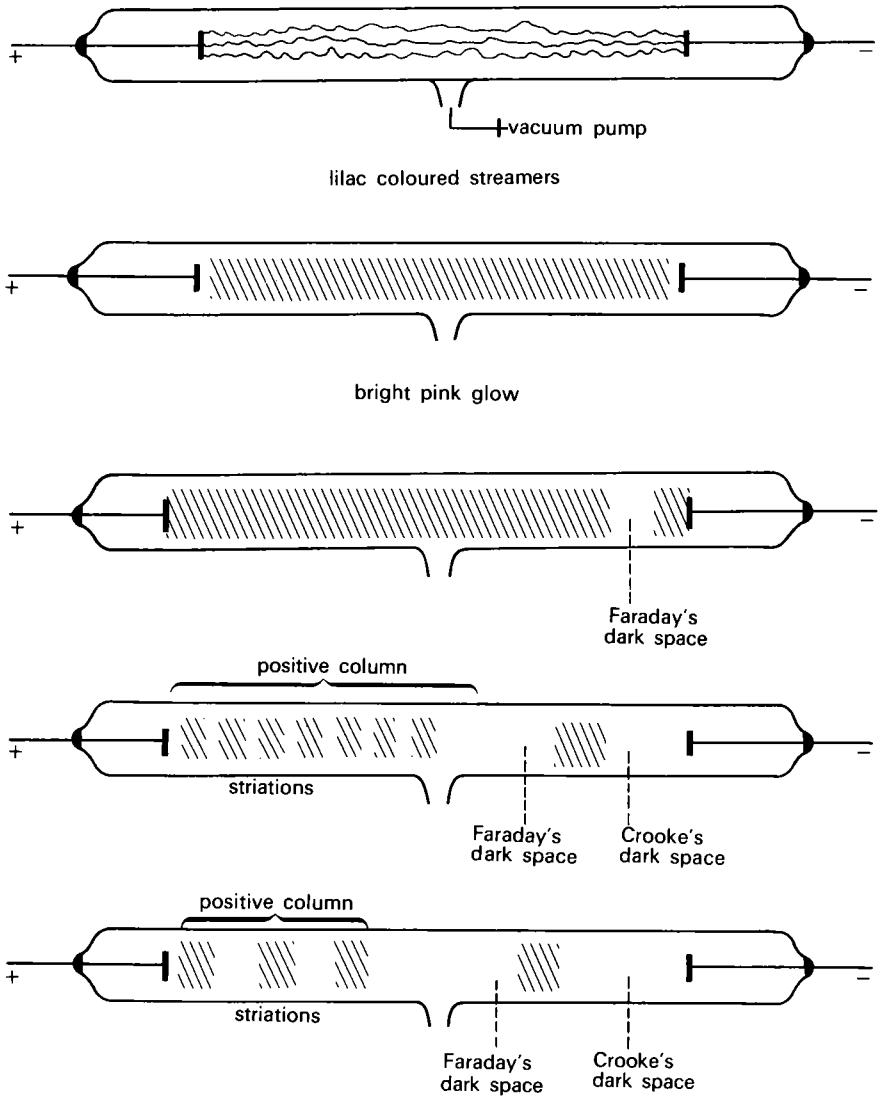


Fig. 21.3 The luminous discharge in air at reduced pressure (The luminous parts of the gas are shown as shaded areas) (a) P ≈ 10 mm of mercury (b) P ≈ 5 mm (c) P ≈ 1 mm (d) P ≈ 0.5 mm (e) P ≈ 0.1 mm

21.4 Cathode rays

By the time that the pressure corresponding to Fig. 21.3(*e*) has been reached, the conditions which favour conduction have been passed and conduction has started to become more difficult again (Fig. 21.1). As the tube is still further exhausted, the Crooke's dark space grows in extent until it eventually fills the entire tube, the positive column and Faraday's dark space having contracted into the anode. Unless the pressure is very low indeed, there may be a misty blue glow remaining in the tube; apart from this, there would be nothing visible to indicate the slight discharge that is still taking place, were it not for the fact that the glass now begins to fluoresce – usually with a greenish yellow colour, but the colour depends on the glass. Fluorescent chemicals, such as zinc sulphide, coating the glass, would glow brightly.

The most noticeable effect about this fluorescence is that the anode casts a dark shadow in it on the far end of the tube. If the anode is repositioned down a side tube, and some other obstacle (traditionally a metal plate in the shape of a maltese cross) is inserted in the discharge tube, this obstacle casts its shadow instead. The anode, away to the side, takes no visible part in the proceedings.

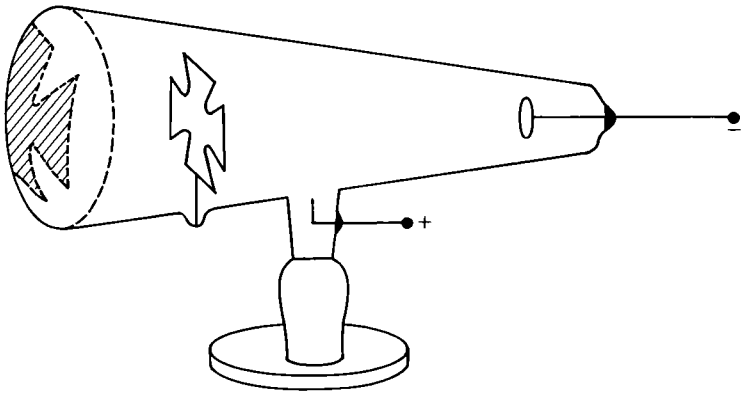


Fig. 21.4 Cathode rays. The Maltese Cross experiment

The first conclusion to be drawn from this casting of shadows is that rays, cathode rays, are travelling in straight lines away from the cathode, making no concession to the shape of the tube or to the exact whereabouts of the anode. Furthermore, if a magnet is held near to the discharge tube, producing a horizontal magnetic field which is perpendicular to the path of the rays, the position of the shadow becomes vertically displaced, in the direction which suggests that the cathode rays carry negative electric charge. A further experiment, first carried out by Perrin, strengthens this conclusion. A hollow metal cylinder is inserted in the tube to collect the rays, and it is connected to an electroscope, the result being that the leaf charges up negatively, and rises in consequence.

[It is rather remarkable that an attempt to deflect the rays electrostatically by the influence of a charged body held nearby would not be successful, unless the pressure in the tube were exceptionally low, and J. J. Thomson was the first to achieve this. The reason seems to be that the cathode rays are electrically screened by positive ions which surround them.]

This was the guise in which electrons were first discovered, and their identity

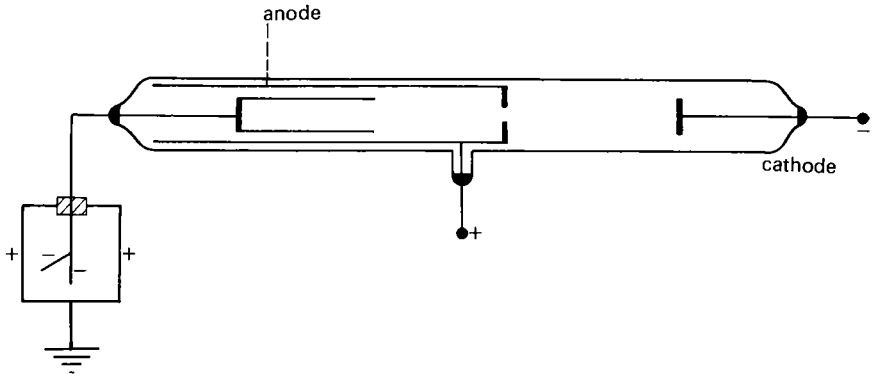


Fig. 21.5 Cathode rays. Perrin's apparatus to demonstrate that they carry negative charge

as negative particles was finally established by J. J. Thomson. The initial conduction in a cold cathode tube of this kind is due to ions from the very low pressure gas, and the presence of some gas is consequently essential for the tube to function. But the greater part of the current is due to electrons which are released from the cathode as a result of bombardment by positive ions from the gas. It is the electrons released from the cathode in this way which form the cathode ray stream, and these increase in momentum so rapidly in the electric field near the cathode that they follow straight paths perpendicular to the cathode, no matter what the lines of force eventually may do (Fig. 21.6).

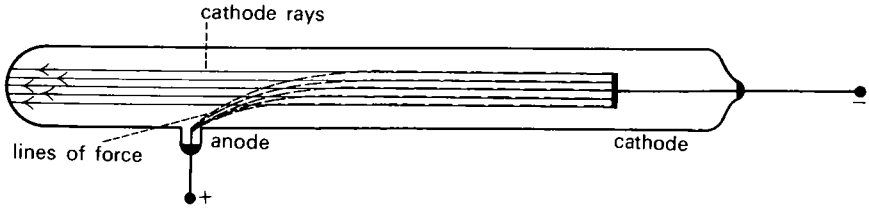


Fig. 21.6 The straight path followed by cathode rays

The cold cathode discharge tube has considerable historical importance, being the type of tube with which Thomson discovered the electron, and with which, also, Röntgen discovered X-rays, but for most purposes today we would choose the hot cathode of the modern cathode ray tube. As a comparison, this needs no gas left in the tube to work it because the electrons are produced by thermionic emission, and if we wish, we can use a p.d. lower than the very high one which the cold cathode tube requires before it will function at all.

Gas discharge tubes, too, may be designed using a hot cathode, and mercury and sodium vapour lamps are examples of this. Such a tube has the peculiarity that, once its striking voltage has been reached, the p.d. across the tube stabilizes, and, if the current is to be controlled, a resistor (or, using a.c., a choke) must be connected in series with it.

21.5 The experiments of J. J. Thomson

The most famous of J. J. Thomson's experiments were carried out before 1897, using apparatus rather similar to a modern cathode ray tube but having a cold

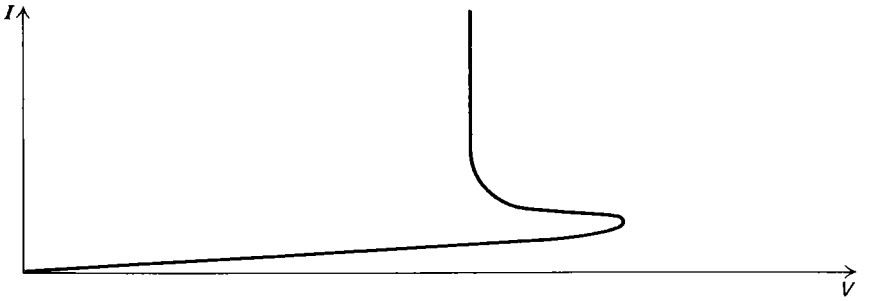


Fig. 21.7 The characteristic of a hot cathode gas or vapour discharge tube

cathode (Fig. 21.8). He deflected a narrow beam of cathode rays in electric and magnetic fields, and thereby deduced their speed and their charge/mass ratio e/m. This latter deduction involved assuming that the cathode rays were a stream of particles, and his accumulated results were accepted as justifying this assumption. His most significant discovery was that the particles appeared to be subatomic, and contained in the atoms of different elements, for the ratio e/m always came out to effectively the same value no matter what he used as the source of the cathode rays, and this value strongly indicated that the particles were much less massive than the lightest atom. The name, electron, which was adopted by Thomson for these particles, had been suggested previously by the Irish physicist, Johnstone Stoney, who believed that the phenomena of electrolysis necessitated a particle theory of electricity to explain them.

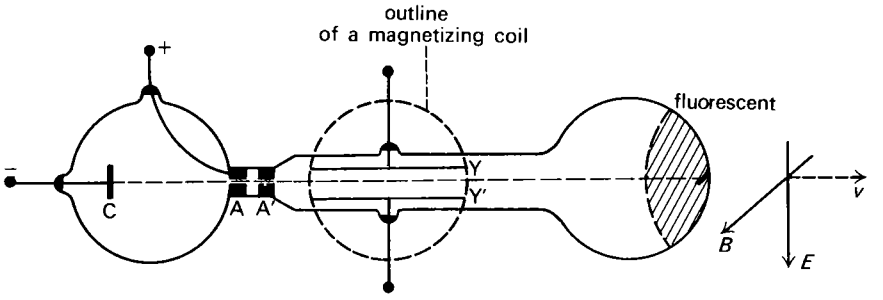


Fig. 21.8 Apparatus used by J. J. Thomson

The basis of Thomson's experiments was the deflection of the cathode rays in a uniform magnetic field of flux density B perpendicular to their path. Assuming the rays to be electrons, each having charge e, these electrons would be deflected, while in the field, into a circular path, radius r (section 14.5), where

$$\frac{mv^2}{r} = Bev, \tag{21.1}$$

whence

$$v = Br\frac{e}{m}. \tag{21.2}$$

The magnetic field was produced by means of a Helmholtz arrangement of two coaxial coils (section 15.10), one on either side of the tube. (The outline of the coils is indicated as a dotted circle in the diagram.)

The result of this single experiment was the deflection of the collimated cathode ray stream as one undispersed whole, merely moving the position of the bright line which marked their arrival at the face of the tube. By measuring the vertical deflection, the radius r of the curved path was deduced. This behaviour suggested that all the particles moved with the same speed v, assuming that they were all similar particles. But this is as far as the single experiment would go.

Thomson now applied a p.d. V across the two deflector plates Y, Y′ to produce a uniform electric field $E = V/d$ perpendicular to the magnetic field B and extending over the same length of path. Giving this p.d. the appropriate polarity, this produced a deflecting force in opposition to the magnetic deflecting force. This is because the force on charges moving across a magnetic field is perpendicular to the flux, whereas an electrostatic force is in the direction of the electric field. The p.d. V was now adjusted to return the deflection of the cathode rays to zero. The two deflecting forces must now be equal in size, and in opposition to each other, giving

$$F = Ee$$

and

$$F = Bev,$$

whence

$$v = \frac{E}{B}. \tag{21.3}$$

Substituting this value in equation 21.2 gives the charge/mass ratio, e/m, for the electrons.

This is as far as Thomson could go. He could not separate the value of the charge e, in his calculations, from that of the mass m, nor could he isolate an electron. But he experimented using different gases in his cold discharge tube and using cathodes of different materials, and always obtained the same value for e/m. He later experimented with electrons emitted thermionically and by photoelectric emission (not, in the latter case, using quite the same technique) and obtained the same results.

Thomson arrived at a conclusion regarding the mass of the electron by comparing his result for the ration e/m with the corresponding charge/mass ratio for the hydrogen ion, obtained by electrolysis. If the charge e of the electron is assumed to be identical with that carried by a monovalent ion, then the masses of the electron and the hydrogen atom may be compared.

Using results obtained from similar determinations more recent than Thomson's, the calculation may be made as follows.

The charge carried across a voltameter by hydrogen ions is

$$9.649 \times 10^4 \text{ C/mol},$$

this being the value of the Faraday constant F (section 5.4). The relative atomic mass of hydrogen is 1.008, and so 1 mole of hydrogen atoms has mass M_H, where

$$M_H = 1.008 \times 10^{-3} \text{ kg}.$$

Hence, the charge/mass ratio for the hydrogen atom is

$$\frac{e}{m_H} = \frac{9.649 \times 10^4}{1.008 \times 10^{-3}} \text{ C/kg}$$

$$= 9.57 \times 10^7 \text{ C/kg}.$$

[Alternatively, this could be directly determined by carrying out an electrolysis which liberates hydrogen at the cathode.]

This compares with the ratio, e/m, for the electron, which is

$$\frac{e}{m} = 1.76 \times 10^{11} \text{ C/kg},$$

whence, if the two charges e are taken to be identical,

$$\frac{m_H}{m} = \frac{1.76 \times 10^{11}}{9.57 \times 10^{7}}$$

$$= 1840,$$

i.e. the electron must be a subatomic particle, about 2000 times lighter than the hydrogen atom, which appears to be a component part of all atoms.

One point of detail about Thomson's calculations should not be overlooked, and this is the fact that equations 21.2 and 21.3 assume ideal conditions which are unlikely to be completely realized in practice. Equation 21.3, in particular, is based on the assumption that both the fields E and B are uniform, with sharply defined boundaries, and that they extend over exactly the same distance. This is what examiners like to describe as being 'uniform and coterminous'. Corrections would obviously need to be made for departures from this condition.

21.6 Determining e/m in a school laboratory

When an approximate determination is made in a school laboratory of the ratio e/m, it is not convenient to go all the way in imitating the method of Thomson. Whether the cathode ray tube chosen is an ordinary commercial type, or one specially manufactured for this particular experiment, it will be a hot cathode tube since this brings nothing but advantages. The principle of using a magnetic field to deflect the electrons can be followed, and so we still use the equation

$$v = Br\frac{e}{m}. \tag{21.2}$$

In practice, although not in theory, there is an easier way than Thomson's to eliminate the velocity v from this equation, and this was devised by Kauffmann in the same year as Thomson's original experiment. The p.d. V applied between the cathode and anode of the cathode ray gun is measured, and this is taken to be entirely responsible for the velocity with which the electrons emerge from the pierced anode. This gives the energy equation

$$\tfrac{1}{2}mv^2 = eV, \tag{21.5}$$

and between this and equation 21.2, v may be eliminated to give the value of e/m.

In determining the radius r of the cathode ray track in a magnetic field, and in

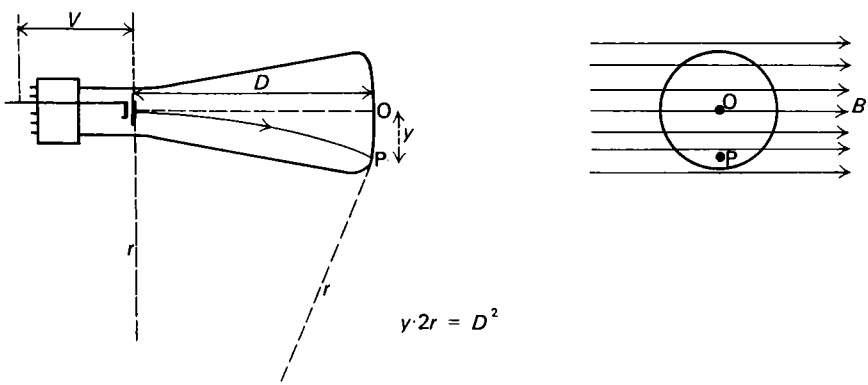


Fig. 21.9 The deflection of cathode rays in a uniform magnetic field

determining the flux density B which causes it, the detail of the method depends on the detail of the cathode ray tube available, and there would be little advantage here in describing any particular one. The principle is the same in all cases, however, and this consists of applying a uniform magnetic field from a Helmholtz pair of coils, for example, perpendicular to the whole length of path of the cathode rays. The radius of the circular track may then be deduced from the deflection caused by the field, and one way to do this is illustrated in Fig. 21.9.

21.7 Millikan's determination of the electronic charge e

The experiment which finally established beyond doubt the particle nature of electricity was Millikan's famous oil drop experiment, concluded by him in 1910. Not only did he measure the electron's charge e with considerable accuracy, but he also demonstrated clearly that this was a definite unit, for he observed the same small oil drop gain and lose charge repeatedly, always in small integral multiples of the same basic quantity e. Although more accurate determinations of this charge have been made since, none of these later experiments has been so elegant.

Having found the value of e, this can be combined with that of the ratio e/m, to give the mass of the electron. Combining it with the value of the Faraday constant, $F = N_A e$, obtained by electrolysis, would give the value of the Avogadro constant N_A. But this last deduction is nowadays applied in the reverse direction, the value of N_A being determined by X-ray diffraction, thus obtaining a value for e by a method which is independent of Millikan (section 5.4).

Millikan's experiment was a brilliant development from ideas which had been tried out by others. After the identity of the electron had been established, J. J. Thomson and other workers in the Cavendish Laboratory at Cambridge made various estimates of the electronic charge, which they largely did by making observations on a mist of charged water droplets. The most promising of these experiments involved measurements of the movement of the top surface of such a mist, as the drops slowly settled down under gravity, and then, again, as they moved upwards in the influence of a vertical electric field.

It was in the process of investigating the flaws in such experiments, and searching for means of eliminating them, that Millikan discovered that he was able to observe the motion of a single illuminated oil droplet, selected from a spray of such drops. He was able to do this over considerable periods of time, alternately allowing the droplet to fall slowly through the air under gravity, and raising it up again, since it was electrically charged, by applying a vertical electric field. By exposing the air to X-rays, he made it possible for the droplet to gain or lose charge from time to time, from or to the ions in the air, and it was in this way that he demonstrated the existence of a basic unit of charge.

Millikan's experiment can be repeated, in principle, in the school laboratory. The X-rays can be left out of it, because oil drops from an oil spray are for the most part already charged. A suitable arrangement is shown in Fig. 21.10.

Two brass plates A, B are mounted horizontally, one above the other, about 5 mm apart, the lower plate B being insulated, and the upper plate A being earthed. The plates are connected across a switch S, so that they may be short circuited to each other, or given a p.d. of about 200 V.

Vacuum oil can be directed into the space above the plates from a scent spray, and some considerable number of very small droplets find their way down the central tube, through a small hole drilled in the plate A, into a space between the plates. This space can be sealed off to prevent air currents by the plunger P. A strong beam of light is focused on a point between the two plates, directly below

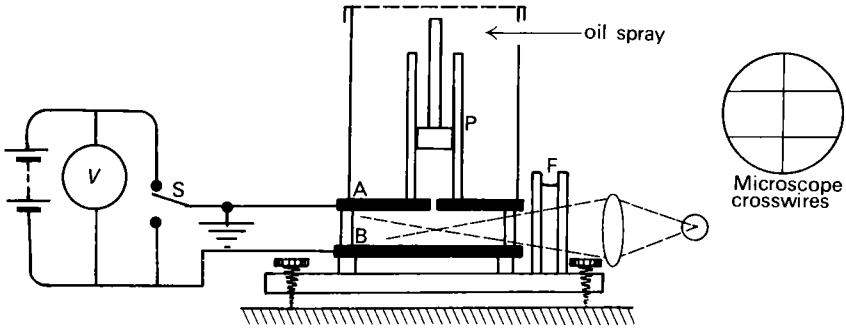


Fig. 21.10 An oil drop apparatus suitable for the school laboratory

the inlet hole, and the brightly illuminated droplets can be observed by means of a microscope, which is focused on the concentrated beam. (This microscope does not appear in the diagram because it would be placed almost at right angles to the plane of the paper.) The brass plates are sealed off by glass windows, and the light passes through a water filter F to absorb the unwanted infra-red, which would otherwise heat up the apparatus and might cause convection currents.

If the oil drops are observed for a minute or so, falling under gravity, the heavier drops soon fall out of the field of view. If the electric field is now applied, it can be seen that many of the droplets are electrically charged, since some are speeded in their fall by the field, some are slowed down, and some reverse their motion, and proceed to move upwards. If one of this last group is selected, it may be observed over a considerable period of time, because it can be allowed to fall under gravity, and brought back in the field of view by applying the electric field.

Because the droplet is so small, its motion is the constant speed of its terminal velocity against the viscous resistance of the air. Its movement is defined by the following equations.

Under gravity:

$$mg = kv_0, \tag{21.6}$$

where v_0 is the speed of its fall.

In the field:

$$EQ - mg = kv, \tag{21.7}$$

where $E = V/d$ is the field strength between the plates, d is the separation of the plates, Q is the charge on the drop.

Stoke's equation for the viscous force on a moving sphere gives

$$k = 6\pi\eta r, \tag{21.8}$$

where r is the radius of the droplet, and η is the viscosity of the air.

Finally, the connection between mass m and radius r of the drop is

$$m = \frac{4}{3}\pi r^3\rho, \tag{21.9}$$

where ρ is the density of the oil.

The basis of the algebra required to deduce the charge Q from these equations is that equations 21.6, 21.8, and 21.9 are required to determine the radius, and hence the weight mg of the drop, and equation 21.7 is then applied to obtain its charge Q.

Millikan, using X-rays to ionize the air and hence to change the charge on the

drop, had the advantage of measuring a succession of different charges held by the single drop at different times. We would have to measure the charges on a succession of different drops. By either method, it is found that all the different values of Q are small integral multiples of a single basic charge, which we take to be the electronic charge e, e.g. $Q = \pm e, \pm 2e, \pm 3e$ and so on.

21.8 Positive rays

The positive ions which are released in the gas of a cold cathode discharge tube normally finish up at the cathode; indeed, it is their energy which releases from it the electrons which form the cathode rays. If a hole is pierced in the cathode, however, some of the positive ions pass through it and emerge from the far side as a stream of positive rays (originally called canal rays).

Thomson applied a technique of magnetic and electric deflection to these positive rays, and recorded their arrival on a photographic plate placed perpendicular to their path to receive them. The actual procedure could not be the same in detail as it had been for the cathode rays because the positive ions, being much more massive, are not deflected so easily, and because they do not all travel at the same velocity. However, he evolved a method which separated out the ions on the photographic plate according to their charge/mass ratio, a process of positive ray analysis which represented the first mass spectrograph. This principle was developed into an accurate science largely by Aston, working in Cambridge, but we will describe a more recent instrument here, on account of its basic simplicity.

21.9 The Bainbridge mass spectrograph

The source of the ions is above the point C in the diagram. They are produced by passing electrons from a hot cathode through the gas or vapour under investiga-

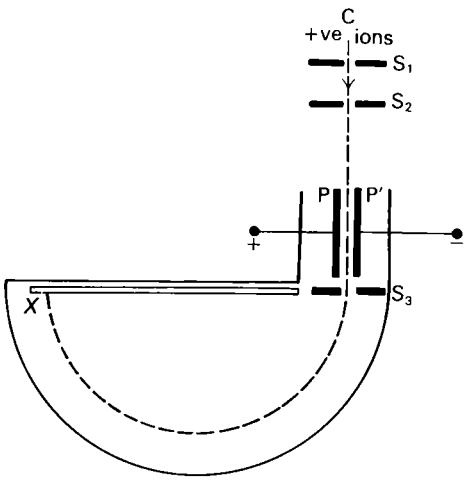


Fig. 21.11 The Bainbridge mass spectrograph

tion, and are then directed vertically downwards towards the point C by the attraction of a strongly negative electrode. The part of the apparatus between S_2 and S_3 is a *velocity selector*, and only ions which have a selected and known velocity v emerge at S_3. This is achieved by applying a p.d. between the two metal plates P,P′ to produce a transverse electric field E. At the same time, a uniform

magnetic field with flux density B is applied, over the area of the entire apparatus, perpendicular to the plane of the diagram. In this way, many of the ions are deflected, either by the electric field or by the magnetic one, and fail to reach the slit S_3. The only ions which pass through S_3 are those for which the two deflecting forces cancel each other, and these ions all have the velocity v, given by the now familiar equation

$$v = \frac{E}{B}. \tag{21.3}$$

These ions which emerge from slit S_3 are now subject to the effect of the magnetic field only, and their track is consequently curved into a circular one, of radius r, given by

$$\frac{mv^2}{r} = BQv,$$

whence

$$r = \frac{vm}{BQ}. \tag{21.10}$$

It follows that all ions which carry the same charge are curved into circular tracks with radius in proportion to their mass, and the sequence of lines mapped out on the photographic plate X are mass spectrum lines, with separation proportional to the differences in mass.

QUESTIONS

$$e = 1.6 \times 10^{-19}\ \text{C}$$

21.1 A simple apparatus is set up in a school laboratory to obtain an approximate measurement of the specific charge, e/m, of an electron. A small cathode ray tube is placed in the middle of a plane circular coil of large radius in such a way that the axis of the cathode ray gun, i.e. the path of the undeflected electron beam, is in the plane of the coil and along a diameter. When a current is sent through the coil, it may be assumed, to a first approximation, that there is a uniform magnetic flux density due to the current, perpendicular to the path of the cathode rays and equal in magnitude to the flux density at the centre of the coil.

Use the following data to obtain a value for e/m:

coil radius = 50 cm,
number of turns = 100,
current = 2 A,
accelerating potential of C.R.O. = 800 V,
radius of the path followed by the deflected electrons = 38 cm.

21.2 An electron, mass m_e, charge e, moves with velocity v in a direction perpendicular to uniform magnetic flux of density B, and is found to describe an arc of a circle of radius r. When $B = 1.2 \times 10^{-3}$ T, $r = 38.0$ cm.

A uniform electrostatic field E is then superimposed on the magnetic field so as to exert a force on the electron in opposition to the magnetic force. When the electrostatic field is adjusted to a strength of 9.6×10^4 V/m, the path of the electron is a straight line.

Assuming that E and B extend over exactly the same distance deduce the value of e/m_e.

21.3 In a Millikan type apparatus, the horizontal plates are 1.5 cm apart. With the electric field switched off, an oil drop is observed to fall with the steady velocity 2.5×10^{-2} cm/s. When the field is switched on, the upper plate being

positive, the drop just remains stationary when the p.d. between the two plates is 1 500 V.

(*a*) Calculate the radius of the drop.

(*b*) How many electronic charges does it carry?

(*c*) If the p.d. between the two plates remains unchanged, with what velocity will the drop move when it has collected two more electrons as a result of exposure to ionizing radiation?

The viscosity of air $= 1.8 \times 10^{-5}$ N·s/m^2 in the condition of the experiment; take $g = 10$ m/s^2; density of the oil $= 900$ kg/m^3. (O)

21.4 In an oil drop experiment, two drops A and B both remain stationary at the same time in the electric field. When the field is switched off, the times taken to fall the same distance are 50 s for A and 38.2 s for B. Which drop has the greater charge? If the larger charge is $12e$, what is the smaller one? (O)

*21.5 In a certain experiment, a very small drop of oil was introduced between two horizontal and parallel plates separated by a vertical distance of 20 mm, the apparatus being open to the air. The drop was observed to fall at a constant velocity over a distance of 15 mm in 15 s, but when a p.d. of 7 500 V was applied across the plates, it was found to rise, covering the same distance in times 65.4, 65.4, 52.2, 52.2 s in four successive measurements. Given that under the circumstances of the experiment the viscosity of air was $1\,830 \times 10^{-8}$ kg/m·s, and the density of the oil was 0.923×10^3 kg/m^3, deduce all you can from the observations.

(A sphere of radius r moving with velocity v through a fluid of viscosity η experiences a retarding force $F = 6\pi\eta rv$.) (CAM. SCH.)

21.6 An ion of mass m and charge Q is accelerated by a p.d. V and enters a region containing a uniform magnetic field of flux density B. In this field the ion moves in a semicircle and strikes, at a distance d from the point of entry, a plane photographic plate which lies along the diameter of the semicircle.

Show that the mass of the ion is given by

$$m = k\frac{B^2 d^2 Q}{V}$$

and deduce the value of k.

Indicate the importance of this experiment in connection with a study of isotopes. (C)

21.7 In a Bainbridge mass spectrograph, the electric field between the plates of the velocity selector is 12×10^4 V/m, and the magnetic flux density is 0.60 T over the whole of the path followed by the ions, and perpendicular to the plane in which they move. A stream of singly charged neon ions, Ne^+, moves in the field in a circular path of radius 7.28 cm after leaving the velocity selector. Deduce the mass of the isotope.

21.8 Singly charged ions from a magnesium source follow a semicircular path after leaving the velocity selector of a Bainbridge mass spectrograph, before arriving on a photographic plate. Three spectrum lines are formed on the plate, the line which is nearest to the velocity selector being a distance 15.94 cm from it. The other two lines, which are less intense, are respectively 0.66 cm and 1.32 cm away from the first one.

If the first line is due to the magnesium isotope $^{24}_{12}Mg$, what are the nucleon numbers of the other two kinds of ion in the beam?

21.9 Explain one method of ionizing a gas at normal pressures, and describe how you would measure the ionization current.

Assuming that the ionization continues steadily and that the pressure is kept constant during the measurement, how would you expect the ionization current to depend on the p.d. applied across the gas?

In a cold cathode discharge tube to which a constant voltage is applied, the current passing through the gas under a given p.d. is found to increase initially as the pressure of the gas is reduced from atmospheric. What explanation can be offered for this? (O)

21.10 State briefly your reasons for believing that the electron (*a*) is a fundamental particle, and (*b*) carries the fundamental unit of electric charge.

Describe J. J. Thomson's experiment for determining the value e/m_e for the electron.

The theory of Thomson's method leads to a formula for e/m_e in which the velocity of the electrons does not appear. It seems therefore that it is immaterial whether or not the beam contains electrons with a wide range of velocities. How, if at all, would you expect the spread of velocities to affect the accuracy of the results? (O)

21.11 Compare the processes of electrical conduction in (*a*) a metal, (*b*) a solution of an electrolyte, (*c*) a gas at atmospheric pressure.

21.12

The diagram shows a simplified version of the apparatus used to investigate 'positive rays'. Explain how the positively charged particles which emerge through the small hole in the cathode are produced and why they do not all have the same velocity.

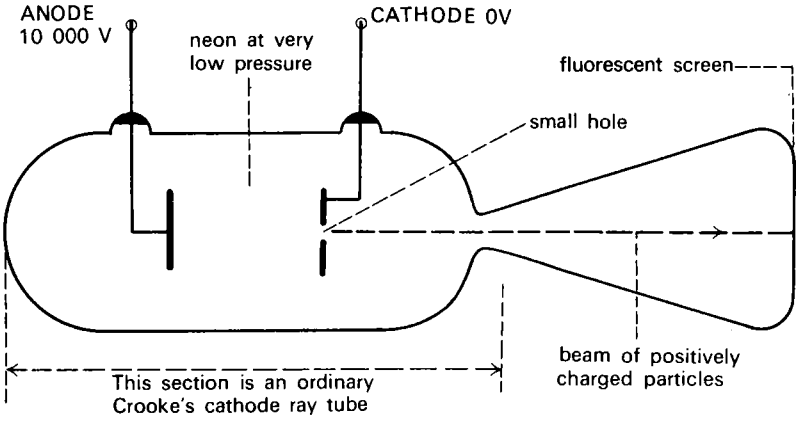


Fig. e 81

What information about neon (and other gases) is obtained by deflecting beams of the particles electrically and magnetically (simultaneously) and examining the pattern they then make on a fluorescent screen? (S)

Chapter 22

The quantum theory. Photocells and the X-ray tube

The X-ray tube and the photocell are put together in this chapter because they have this in common: when it comes to explaining their characteristic behaviour conventional theory fails to do it, whereas an obviously unconventional application of quantum theory serves to account for both, and the argument is much the same in both cases. Since the theory was originally applied by Einstein to explain photoemission, this will be described first.

22.1 Photoemission – a photocell

Photoemission is the emission of electrons from a cold metal surface as the result of radiation of appropriate wavelength, in most cases ultraviolet, falling on it; thus it is a process resembling thermionic emission, with the difference that the escaping electrons acquire the necessary energy, not thermally, but by taking it from the incident radiation.

Photoemission may be demonstrated by directing the radiation from a mercury vapour lamp onto the surface of a plate of zinc, but this must be freshly cleaned because the surface is all important (Fig. 22.1). A p.d. is applied between the zinc, which is made negative, and a metal gauze anode placed parallel to it. The current, which is only very small, is detected by using a d.c. amplifier connected across the very high value load resistor R. (Section 18.6).

When the radiation falls on the zinc, electrons are emitted from its surface and some of these reach the metal anode, causing a small current to flow. If the polarity of the applied p.d. is now reversed, the current ceases to flow. It also ceases if a glass filter is placed in the path of the radiation, because this absorbs the ultraviolet, and visible light cannot produce photoemission from zinc; if an ultraviolet filter, which absorbs the visible light but transmits the ultraviolet, is interposed instead, the current is unaffected.

One of the remarkable features of photoemission is that it is more likely to occur when the frequency of the incident radiation is high, and the only pure metals for which frequencies in the visible range are high enough are the alkali metals, lithium, sodium, potassium etc. The reason for choosing zinc in the demonstration just described is that it works at a low enough frequency to include the mercury's ultraviolet without being too corrosive to keep its clean surface when exposed to the atmosphere, as an alkali metal would be. In the commercial type of cell (Fig. 22.2), the cathode is specially coated to make it photosensitive over the visible range, usually with a mixture of caesium and antimony, or caesium and oxidized silver. The cell shown in the diagram is a form of vacuum diode, and it is operated, like any other diode, with a p.d. applied to attract the electrons to the anode, but it is run at saturation, and in this condition the current is directly

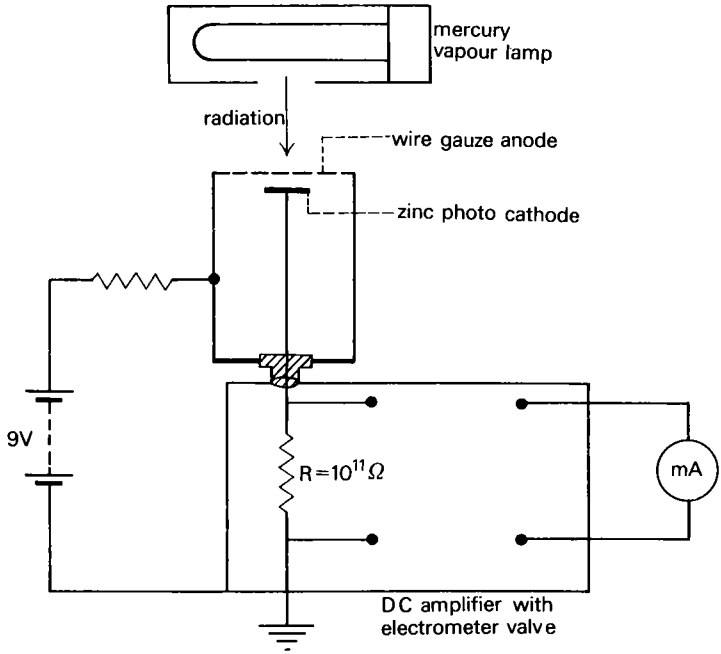


Fig. 22.1 An arrangement to demonstrate photoemission

proportional to the intensity of the radiation which falls on the photosensitive cathode. The characteristics shown in the diagram are typical, and it will be noticed that the current is only small. (There are also gas filled types designed to give larger currents, but in these the advantage of having direct proportion between current and light intensity is lost.)

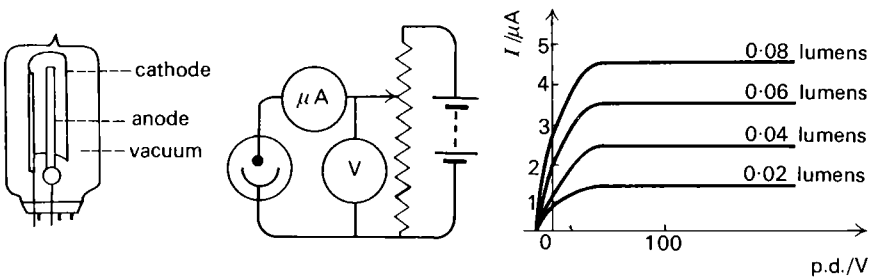


Fig. 22.2 A photoemissive cell (a) One type of cell (b) Circuit for plotting characteristics (c) Characteristic curves

22.2 Quantum theory – Einstein's equation

It is a curious feature of photoemission that the energy with which an individual electron leaves the surface of the metal is quite independent of the intensity of the incident light; the increased current which results when the intensity is increased is due to a rise in the number of electrons emitted, not to an increase in their individual energies. This is the behaviour which Einstein explained by quantum theory when classical arguments had failed to do it.

The quantum theory was first suggested by Planck in 1901. It represented a

departure from conventional ideas, devised for the specialized purpose of accounting for the energy distribution in the spectrum radiated from a heated black surface. The theory, as propounded by Planck, assumed that the atoms in the black surface could radiate and absorb energy in discrete units, or quanta, the energy of each quantum being given by

$$E = h\nu, \tag{22.1}$$

where ν is the frequency of the radiation and h is a constant, known as *Planck's constant*, the value of which has since been measured and is

$$h = 6.63 \times 10^{-34}\ \mathrm{J \cdot s}, \text{ approximately.} \tag{22.2}$$

In order to explain the characteristics of photoemission, Einstein transferred the idea of quantization to apply to the radiation, and postulated that it must be a property of electromagnetic radiation that it carries its energy in quanta, or photons, each photon having energy E, where $E = h\nu$.

Thus, each electron in a photosensitive surface which accepts energy from the incident light gains the energy of a photon, irrespective of the intensity of the light. The effect of an increase in intensity is to increase the number of photons arriving, but not their individual size; consequently it increases the number of electrons released, but not their individual energies, and this correctly describes what Einstein set out to explain.

The theory takes us further, however. Since the energy of a photon is proportional to its frequency, there is likely to be a minimum frequency for every metal below which photoemission does not occur; in the case of zinc, for example, a photon of ultraviolet light from mercury vapour has enough energy, but a photon of visible light has not. This aspect of photoemission is summed up in *Einstein's equation*

$$\tfrac{1}{2}mv^2 = h\nu - E, \tag{22.3}$$

which is explained as follows.

Every conduction electron in a metal needs to receive an appreciable addition of energy before it can escape from the surface, because the positive ions left behind in the crystal lattice combine forces to hold it back. Some electrons need to receive more energy than others, but all of them must receive more than a minimum amount E, the value of which depends on the metal. If the energy of each incident photon is $h\nu$, none of the escaping electrons will have kinetic energy exceeding the difference $(h\nu - E)$ between this energy and the minimum amount E taken up in allowing it to escape. Hence the expression $\frac{1}{2}mv^2$ in the equation represents the maximum kinetic energy possessed by any of the emitted electrons; many will have less.

Writing the energy E in the form

$$E = e\Phi, \tag{22.4}$$

where e is the electronic charge, the quantity Φ, being energy divided by charge, is a potential and is measured in volts. It represents the p.d. through which an electron must be attracted if it is to acquire an amount of energy equal to E. The potential Φ is termed the *work function* of the metal.

If the frequency of the incident radiation is the threshold frequency ν_0, below which no photoemission occurs at all, then we have

$$E = h\nu_0 \tag{22.5}$$

and so an alternative form for Einstein's equation is

$$\tfrac{1}{2}mv^2 = h(\nu - \nu_0). \tag{22.6}$$

Example

If photoemission is to result from the incidence of orange light, wavelength $\lambda = 600 \times 10^{-9}$ m, *what is the maximum value of the work function which the photocathode may possess?*

From equation 22.5, we have

$$E = h\nu_0$$

$$= h\frac{c}{\lambda_0},$$

where $h = 6.6 \times 10^{-34}$ J·s, $\lambda_0 = 6.0 \times 10^{-7}$ m, $c = 3.0 \times 10^8$ m/s.

Substituting these values, $$E = 6.6 \times 10^{-34} \times \frac{3 \times 10^8}{6 \times 10^{-7}} \text{ J}$$

$$= 3.3 \times 10^{-19} \text{ J}.$$

Alternatively, $$E = e\Phi,$$

Therefore $$\Phi = \frac{3.3 \times 10^{-19}}{1.6 \times 10^{-19}} \text{ V}$$

$$= 2.0 \text{ V},$$

i.e. the maximum value for the work function is 2.0 V.

The work function of most pure metals is in the region of 4 V or more, and orange light cannot release electrons from them, the only exceptions being the alkali metals.

It will be recognised that Einstein's theory and his equation are not classical physics, and they are not consistent with any aspect of wave behaviour which a conventional treatment of wave propagation, such as Maxwell's theory, would predict. There is no contradiction involved, from a practical point of view, however, in applying Maxwell's equations to the comparatively low frequencies of radio, or even of microwaves, because the size of the photon in such cases is exceedingly small, but at the higher frequencies of ultraviolet light and X-rays we have a situation where conventional methods are applied to explain some things, such as interference, while quantum theory must be adopted for others, such as photoelectric effects. We accept this situation because the experimental basis of both is very accurately established.

Einstein's equation was first verified experimentally in 1916 by Millikan, who used a specially constructed photocell with interchangeable cathodes of lithium, potassium, and sodium. The principle of such an experiment is to apply a small reverse p.d. across the photocell, i.e. to make what would normally be the anode negative, so that it repels the electrons emitted from the photosensitive surface instead of attracting them. This p.d. is varied until it is just sufficient to prevent any current from flowing, in which case

$$eV_0 = \tfrac{1}{2}mv^2, \tag{22.7}$$

where V_0 is the value of this stopping potential, and $\frac{1}{2}mv^2$ is the maximum kinetic energy with which the incident radiation is releasing electrons. Substituting for $\frac{1}{2}mv^2$ from equation 22.3 gives

$$eV_0 = h\nu - E$$

$$= h\nu - e\Phi,$$

$$\therefore \quad V_0 = \frac{h}{e}\nu - \Phi. \tag{22.8}$$

The stopping potential is measured, using a number of different monochromatic wavelengths, both in the visible and in the ultraviolet range, and a graph of V_0 against the frequency ν of the radiation is plotted. The experiment requires a very sensitive microammeter, such as a d.c. amplifier, if the stopping potential is to be identified with any precision.

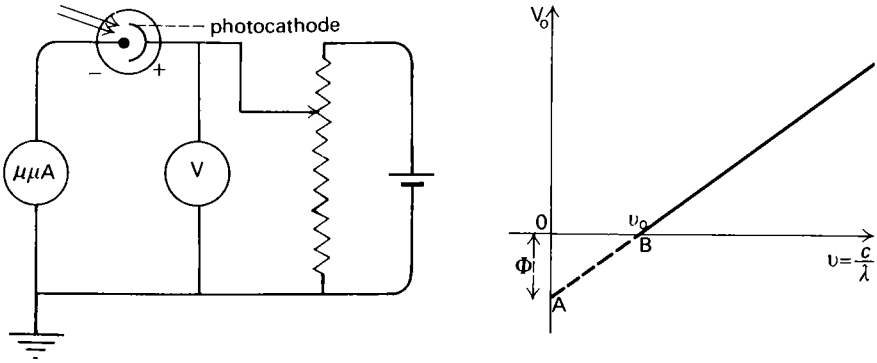


Fig. 22.3 Arrangement to test Einstein's equation (a) (b)

Referring to the graph shown in the diagram and equation 22.8 which it represents, we have:

When $\nu = 0,\ V_0 = -\Phi,$

i.e. the work function of the metal Φ is equal to the intercept OA, except that the intercept is negative.

When $V_0 = 0$, the intercept OB gives the threshold frequency ν_0 for the metal.

Finally, the slope of the graph, i.e. the ratio $(V' - V)/(\nu' - \nu)$, is equal to h/e, from which the value of Planck's constant h can be calculated.

22.3 Photoconduction

Another photoelectric effect, which is applied in several different types of photocell, occurs in semiconductors, and this is the effect of incident photons in releasing electrons and holes from their bonding in a crystal lattice. This is the principle of the photoconducting type of cell, for example, which is a high resistance semiconductor, usually made of selenium or of cadmium sulphide, across which a p.d. is applied from an external source in series with a meter. The effect of the incident photons lowers the resistance, causing an increase in the current. A state of dynamic equilibrium is set up between charge carriers which are continually being released by incident photons and electron-hole pairs which are recombining, and the overall effect is appreciable only if the rate of recombination is comparatively slow. For this reason, the response of the cell falls off if the frequency of the radiation is too high, because the photons in this case are so energetic that all the electrons and holes are released in the same region close to the surface and promptly recombine. If, on the other hand, the wavelength is too long, the photons lack sufficient energy to release the charge carriers at all. These two factors account for the shape of the response curve shown in Fig. 22.4.

It is typical of the photoelectric effects in semiconductors that they are likely to occur at low enough frequencies to include the visible range, and in many cases part, at least, of the infra-red. This is because the energy needed to release charge

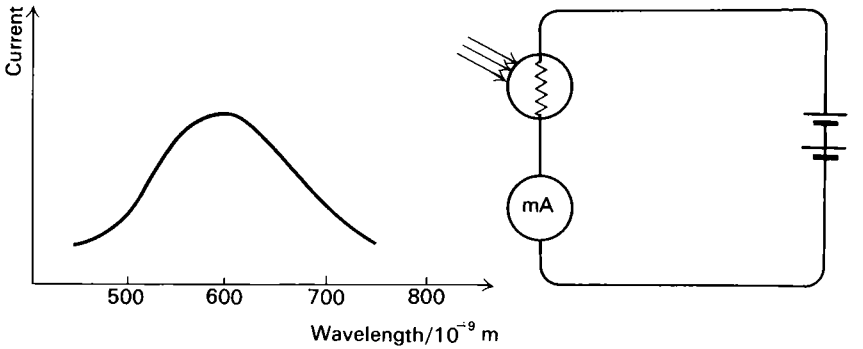


Fig. 22.4 A cadmium sulphide photocell – response curve

carriers in a crystal lattice is not usually as great as that required to release electrons from the surface of a metal.

The photodiode and the phototransistor

The photodiode works on the same principle as a photoconducting cell, the high resistance in this case being that of the depletion layer at the *n–p* junction of the diode. This is operated with reverse bias applied from a battery and the incident radiation is allowed to fall on the junction, or close to it. The minority holes and electrons released by this radiation in the proximity of the depletion layer greatly reduce its resistance because the bias is favourable to them. A proper photodiode is so constructed that the light can fall on it to best effect, but any junction diode is photosensitive if the light excluding cover can be removed from it.

In the phototransistor, the photoelectric effect takes place at the emitter-base junction, where the incident photons cause charge carriers to flow into the base, instead of an input signal doing so. The normal reverse bias is applied between base and collector, and transistor action amplifies the effect.

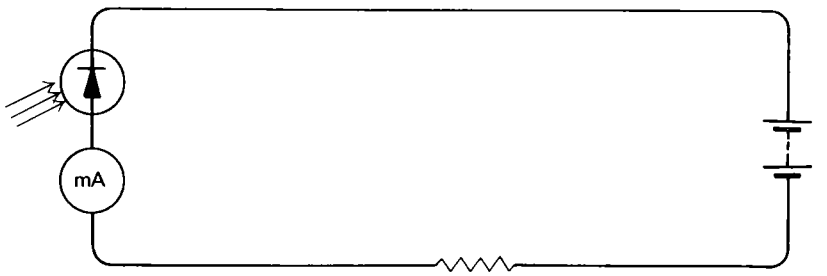


Fig. 22.5 A photodiode

22.4 The photovoltaic cell and the solar cell

In the photovoltaic cell, an external battery is not required, since the cell supplies the current. The potential barrier which exists across the depletion layer of a diode becomes a source of continuous current, the energy being supplied by the incident photons. These release minority charge carriers on the two sides of the junction, and towards these the barrier p.d. provides a favourable potential gradient, with

the result that a current flows through a microammeter connected across the diode (positive current flowing from the p region to the n in the external circuit).

In the solar cell, the photovoltaic action is used to generate a useful amount of power. This is a heavily doped n-type silicon crystal, a few cm^2 in area, one face of which has been made p-type to a very shallow depth, so that the incident photons of sunlight can reach the vicinity of the junction. A large number of these cells are used in series and parallel arrays to make up a solar battery.

22.5 An X-ray tube

If cathode rays are accelerated through a p.d. which exceeds, say, 1 kV before they are stopped by striking an obstacle, there is a chance that part of their energy on being stopped will be converted into X-radiation. This is electromagnetic radiation with frequency many times greater than that of light, corresponding to wavelengths of 10^{-9} m down to 10^{-12} m and even shorter. They were named X-rays by Röntgen, who discovered them in 1895 while studying cathode rays in a cold cathode discharge tube. All the early X-ray tubes, with which much of the pioneer work was done, were cold cathode tubes, but the modern one uses thermionic emission.

The theory of the X-ray tube is very similar to that of the photocell. When an electron accelerated through a p.d. V is brought to rest, there is a chance that the whole of its energy will be converted into a single X-ray photon, and when this happens the relation is

$$Ve = h\nu. \tag{22.9}$$

This equation therefore gives the highest frequency obtainable by the use of any selected p.d. If the tube is working at 1 kV, for example, the shortest wavelength possible is about 10^{-9} m, from which we deduce that an X-ray tube must be operated at a higher p.d. than this. The X-rays are more likely to be produced if the electrons are brought very suddenly to rest, and so the target used to stop them is made from a heavy metal such as tungsten. A large proportion of the energy, nevertheless, is converted to heat, and so the tungsten is imbedded in a large block of copper, which is also the anode, to act as a heat sink. This describes the arrangement in the typical X-ray tube shown in Fig. 22.6, in which electrons from a directly heated tungsten cathode are accelerated through a p.d. of about 100 kV. The distance which separates cathode from anode is unimportant, because the energy of the electrons depends only on the p.d. which accelerates them.

Because an X-ray tube is a form of thermionic diode, it is often used as its own rectifier, the p.d. being applied across it directly from a high voltage transformer, and this is the arrangement shown in the diagram. In this case, of course, the tube is in operation only during part of every alternate cycle of the alternating current. The copper anode is connected to earth for safety reasons, since this is the exposed part of the tube.

22.6 The spectrum from an X-ray tube

Equation 22.9 gives the highest frequency X-ray which can be radiated from the tube, on the assumption that the whole of an electron's energy is converted into a single photon. The greater part of the radiation has lower frequency, i.e. longer wavelength, than this limiting photon, and the spectrum obtained from an X-ray tube is shown in Fig. 22.7. Thus, the average wavelength emitted can be controlled by varying the p.d. applied across the tube, very high voltages being required if the shorter wavelengths are to be obtained. In the early cold cathode tubes this control

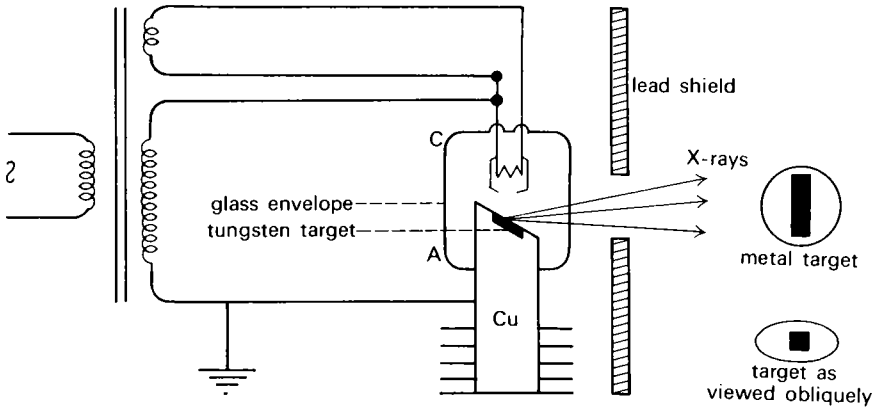


Fig. 22.6 A modern type of X-ray tube (a) The tube (b) Oblique projection

of p.d. was not available, because the gas contained in the tubes made them work as constant potential devices. Consequently, to obtain short wavelengths, the tube required a harder vacuum, and short and long wavelength X-rays are still often referred to as being hard rays and soft rays respectively because of this history.

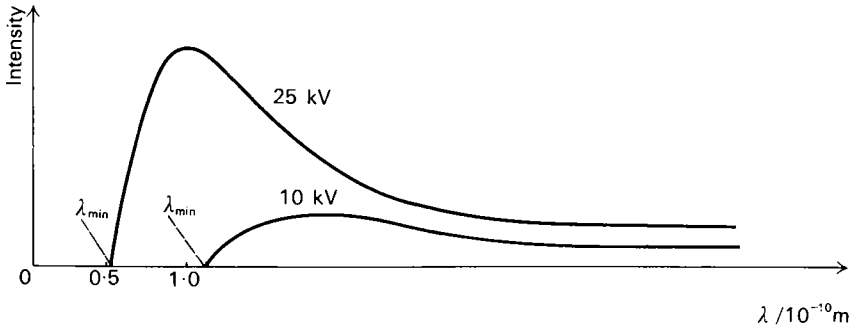


Fig. 22.7 The continuous X-ray spectrum from an X-ray tube

Equation 22.9 is the basis of a method of measuring the value of Planck's constant h which is more accurate than the use of a photocell; the limiting wavelength, λ_{min} which corresponds to a particular p.d. V, is measured by an X-ray spectrometer, and we have

$$Ve = \frac{hc}{\lambda_{min}}. \tag{22.10}$$

22.7 The dangers and medical uses of X-rays

A further precaution when operating an X-ray tube is to have a thick lead shield, as shown in Fig. 22.6, protecting the operator from the radiation, because of its ill effects on the human body. This is particularly necessary when using a high voltage to obtain hard X-rays, because these are the most damaging. The danger is all the greater because the effects are not immediately apparent; indeed in some cases they are genetic, and the most delayed effects are cumulative, so that exposure to even the softest rays, as used for photography, should be as brief as possible and never suffered carelessly.

It will be noted that a cold cathode discharge tube, when used to demonstrate cathode rays, has a very low pressure inside it and requires a p.d. of several thousand volts to work it, so that it may be radiating quite penetrating and damaging X-rays. Television tubes, likewise (but not the much lower voltage oscilloscope) are potential X-ray sources, although the radiation is soft and not intense.

On the credit side, the useful applications of X-rays in medicine are for X-ray photography, and for the destruction of malignant growths. The most effective rays for this second purpose are exceptionally hard, such as the γ-rays are which are emitted from some radioactive isotopes, and these may be used instead of an X-ray tube. The radiation from a radioactive source, however, spreads out in all directions, and cannot be concentrated, so that an X-ray tube producing rays of equal hardness is preferable. This is not a tube of conventional design because the energy of the bombarding electrons, in order to produce such energetic photons, must be several million volts; these are produced by using the principle of a modern type of particle accelerator.

X-rays cannot be focused by a lens, and X-ray photographs are shadows, which require a point source if they are to be sharply defined. This is the reason for the shape of target shown in Fig. 22.6. The bombarding electrons must not be concentrated on too small an area because of the excessive local heating which would result, and so the principle of oblique projection is used, making a comparatively large area source appear to be small.

22.8 The diffraction of X-rays

Many of the characteristics of X-rays were discovered by Röntgen. He detected them in the first place by their effect in causing fluorescence in suitable materials placed outside a discharge tube which was giving cathode rays. This fluorescence persisted even when the discharge tube was covered with black paper, and it occurred even at a distance of a meter or two from the tube. He concluded that the rays must be some form of wave propagation, having tried without success to deflect them as cathode rays are deflected, by a magnetic field; nor has any later experiment produced either magnetic or electrostatic deflection. He attempted to refract them, by passing them through a variety of different substances, and in this, also, he had no success. He discovered, instead, their property of passing through all but the densest materials without deviation, and used them to cast shadows of concealed objects, including the bone structure of his hand. He also found that X-rays render gases slightly conducting when they pass through them, an effect which is due to their ionizing action, and he designed the first X-ray tubes. To prove that X-rays are waves, however, and to measure their wavelength requires the principle of interference; an optical diffraction grating, although used successfully by special methods comparatively recently, is too coarse to be a very likely prospect. The first direct evidence of interference was produced in 1912 and resulted from the suggestion by Von Laue that the regular arrangement of atoms in a crystal lattice might provide a workable diffraction grating for such short wavelengths as he believed X-rays to possess. The crystal, being, in effect, a three-dimensional arrangement of opaque dots, is not, however, as simple as the plane diffraction grating with its single layer of equispaced apertures. The first successful application of Von Laue's suggestion was made by Friedrick and Knipping, who passed a narrow pencil of X-rays through a thin crystal of zinc sulphide onto a photographic plate which they exposed for several hours. The result was a regular pattern of dots on the plate, surrounding the strong impression made by the direct pencil of rays. Such a pattern is known as a Laue pattern. Perhaps the term

diffraction has been inappropriately chosen, since a full explanation would show that the dark spots result, not from diffraction, but from constructional interference between waves reflected from successive crystal planes in the lattice – but their significance with respect to the wave nature of X-rays is nevertheless valid.

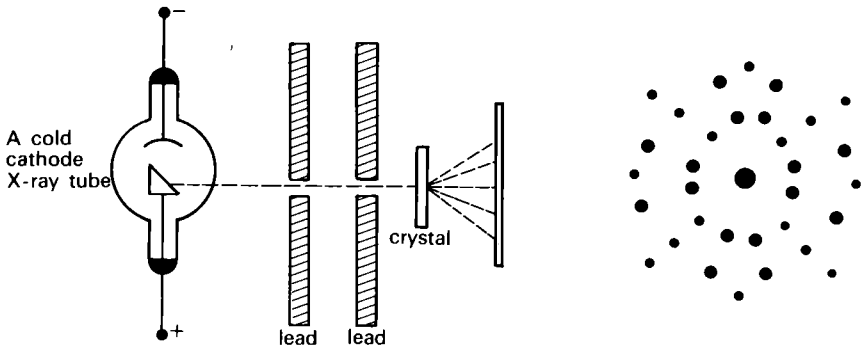


Fig. 22.8 Friedrich and Knipping's experiment (a) Friedrich and Knipping's apparatus (b) A Laue pattern

22.9 The X-ray spectrometer

The first X-ray spectrometer was designed by W. H. Bragg in 1912. With it, he was able to use a rocksalt crystal to analyse the wavelengths emitted from an X-ray tube. A narrow beam of the rays was directed on to the face of the crystal, and the reflected beam was detected by using an ionization chamber.

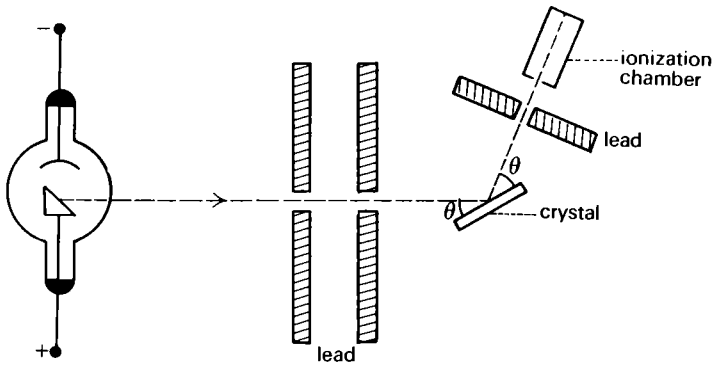


Fig. 22.9 The principle of the Bragg X-ray spectrometer. The angles θ marked in the diagram represent angles made with a plane of atoms in the crystal lattice. The face of the crystal has been assumed, in the diagram, to be parallel to this

With reference to Fig. 22.10, let us suppose that a beam of X-rays from the X-ray tube has a single wavelength λ, and neglect the effects of other wavelengths which are also emitted in practice. We further suppose that the beam enters the crystal making angle θ with successive parallel planes of ions in the lattice, **aa′**, **bb′**, **cc′** etc. The crystal behaves towards the X-rays as a multiple layer, parallel plate interferometer. A very weak reflected beam comes off each plane of ions at the angle of reflection; any one of this succession of reflected beams would be too weak

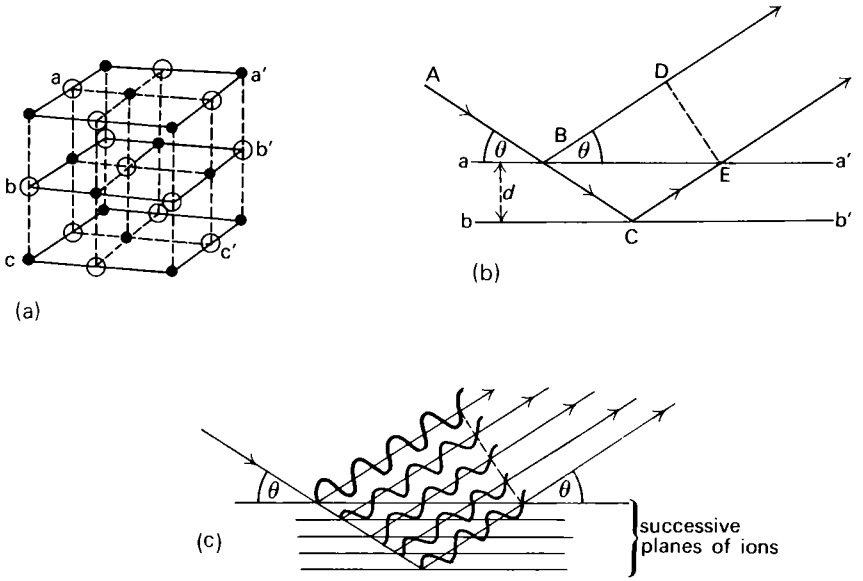


Fig. 22.10 Reflections from planes of ions in a crystal lattice (a) The rocksalt crystal (b) Reflection from successive planes of ions in the crystal (c) Successive reflections at a Bragg angle. The reflected waves from successive planes are all in phase with each other, causing constructive interference

to detect, but the crystal consists of a stack of reflecting planes, one ion layer after another. If the reflections from all these planes happen to emerge from the crystal in phase with each other, then a reflected beam of appreciable intensity, which can be detected by the ionization chamber, results. This constructive interference occurs, however, only if the incident angle θ is appropriately selected, such as to satisfy the equation

$$2d \sin \theta = n\lambda, \tag{22.11}$$

where d is the separation of successive planes of ions. This is called *the Bragg equation.*

The derivation of this equation is identical with that applied in optics to the constructive interference between waves reflected from two parallel surfaces of an air film. The same equation applies in this case, except that it is the convention in optics to put it in terms of the angle of incidence i, where $i = (90 - \theta)$.

Referring to Fig. 22.10(*b*), the path difference between the two rays BD, BCE is equal to the difference (BCE – BD), so that the condition for these two rays to emerge in phase with each other is

$$(\text{BCE} - \text{BD}) = n\lambda,$$

from which equation 22.11 follows by simple trigonometry.

The practical significance of the Bragg equation is that, if X-rays of a single wavelength were used, strong reflection would occur only if the angle θ made with a plane of ions were correctly chosen to satisfy the equation. The design of Bragg's spectrometer allowed for this, and the angle could be varied. There is, in fact, more than one suitable angle, allowing for values of $n = 1, 2, 3$, etc. An X-ray tube, however, gives a continuous spectrum of wavelengths (Fig. 22.7), and consequently Bragg detected some reflected radiation for all values of θ over a con-

siderable range, and his results enabled him to plot the spectrum distribution curve.

The above treatment has neglected several complicating factors, one of these being the fact that the chlorine and sodium ions which make up the rocksalt crystal are not equally effective sources of reflected wavelets. A more interesting consideration, however, is that the ions of the crystal lattice can be grouped in a variety of different ways when selecting what we choose to regard as crystal planes.

We could, for example, choose the planes shown in Fig. 22.11(*b*), instead of those in Fig. 22.11(*a*); these also give a reflected beam in accordance with the Bragg equation, although not quite so strongly, because the concentration of ions

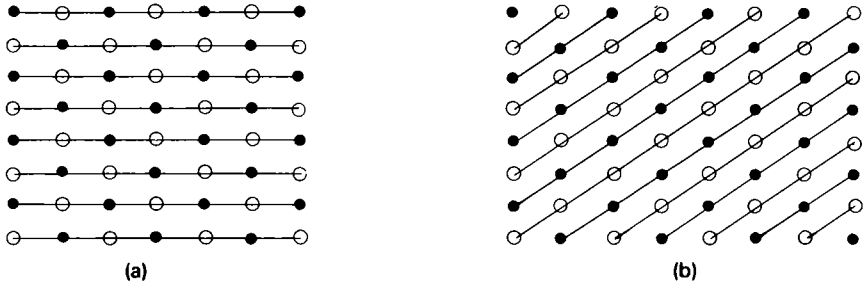


Fig. 22.11 Two possible sets of reflecting ion planes in a crystal lattice (a) (b)

in each plane is less. It is this fact which enabled Bragg to establish the geometry of the crystal lattice. Having done this, he could deduce the separation d of the chief planes of ions, using a knowledge of the density of the crystal and of the Avogadro constant N_A. The wavelengths of the X-rays could now be obtained from equation 22.11.

The determination of electronic charge e

It was later found that X-rays could be analysed using a conventional reflecting type of diffraction grating at grazing incidence. In this case, the element spacing d is known, giving the wavelength of the X-rays without reference to the value of the Avogadro constant N_A. This means that the crystal in the X-ray spectrometer may be used to determine N_A, as an alternative to obtaining it from measurements of the Faraday constant and electronic charge e (chapter 21.7), and so the value of e may be calculated too.

22.10 X-ray spectrum lines

The spectrum distribution curves obtained by Bragg revealed more than the continuous spectrum which was drawn in Fig. 22.7. Provided that the p.d. applied across the X-ray tube is sufficiently high, say 50 kV for one particular target element, two or three relatively intense spectrum lines are superimposed on the background of the continuous X-ray spectrum.

The diagram shows the spectrum lines as they appear if an ionization chamber is used to detect them, the spectrum distribution being deduced from this. The greater resolution afforded by a more refined type of X-ray spectrograph, using photography for detection, shows that there are three distinct groups of these lines, which are known as the K, L, M series. The wavelengths making up one group are relatively close together, but widely separated from those of each other group. The lines shown in the diagram are the K lines for molybdenum, and these have the shortest wavelengths, the L and M series being very soft X-rays.

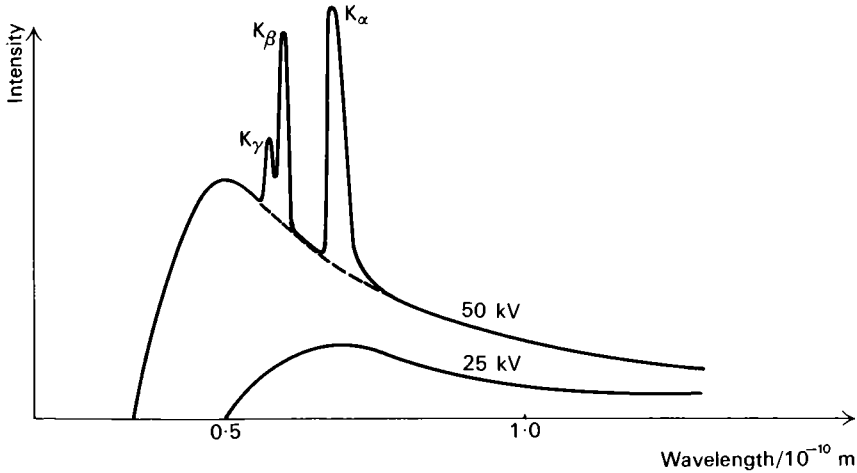


Fig. 22.12 The X-ray spectrum from a molybdenum target, using two different p.d.s in the X-ray tube

By using different elements as the target in his X-ray tubes, and by using crystals containing different elements with which to analyse the X-rays, Bragg showed that the wavelengths of these X-rays lines do not depend in any way on the analysing crystal, or on the p.d. applied across the tube, but are characteristic of the element chosen as the target. For any selected target, however, there is, for each series of spectrum lines, a minimum p.d. below which the lines do not show up at all; when this p.d. is exceeded, all the lines of the one series appear together. At low p.d.s, only the continuous spectrum appears.

The explanation, which was developed later, is that a wavelength in the K series results from an electron being expelled from the electron shell of a target atom which is closest to its nucleus, i.e. the K shell. The energy in a photon of this characteristic radiation does not derive directly from an electron of the bombarding cathode ray stream, as photons in the background continuous spectrum do. It is energy which is released when an electron from one or other of the shells further from the nucleus of the target atom moves in to the K shell to occupy the vacancy. The separate lines in the K series result from such transitions occurring from the L, M, and N shells respectively. This is why the minimum p.d. which is required to excite one K line is sufficient to excite them all, because it represents the minimum energy which the bombarding electrons must have to expel an electron from the K shell in the first place. Similarly, the L series results from electrons being expelled from the L shell of target atoms, and so on.

QUESTIONS

$$h = 6.63 \times 10^{-34}\ \mathrm{J \cdot s}$$
$$c = 3.00 \times 10^{8}\ \mathrm{m/s}$$
$$e = 1.6 \times 10^{-19}\ \mathrm{C}$$
$$m_e = 9.1 \times 10^{-31}\ \mathrm{kg}$$

The electronvolt. This is a unit of energy, being the energy possessed by a particle carrying charge e, when it is at a potential of 1 V. Thus

$$1\mathrm{eV} = 1.6 \times 10^{-19}\ \mathrm{J}, \text{ approximately.}$$

22.1 Calculate the energy of a photon in each of the following cases, expressing this (i) in joules, (ii) in electronvolts.

(*a*) Infra-red radiation, $\lambda = 10^{-5}$ m.

(*b*) Yellow light, $\lambda = 5 \times 10^{-7}$ m.

(*c*) Long wavelength ultraviolet, $\lambda = 10^{-7}$ m.

(*d*) A soft X-ray, $\lambda = 10^{-9}$ m.

(*e*) A γ-ray, $\lambda = 10^{-12}$ m.

22.2 Light of wavelength 500×10^{-9} m ejects electrons from a certain metal with a maximum kinetic energy of 2.0×10^{-19} J. What will be the maximum kinetic energy of electrons ejected from the same metal by light of wavelength 400×10^{-9} m? (S)

22.3 The maximum kinetic energy of the electrons emitted from a metallic surface is 1.6×10^{-19} J when the frequency of the incident radiation is 7.5×10^{14} Hz. Calculate the minimum frequency of radiation for which electrons will be emitted. (N)

22.4 In an experiment on the photoelectric effect using radiation of wavelength 4.00×10^{-7} m, the maximum electron energy was observed to be 1.40×10^{-19} J. With radiation of wavelength 3.00×10^{-7} m, the maximum energy was 3.06×10^{-19} J. Derive a value for Planck's constant.

Mention one other physical phenomenon involving Planck's constant. (N)

22.5 Einstein's equation for the photoelectric emission of electrons from a metal surface can be written

$$h\nu = \tfrac{1}{2}mv^2 + \Phi e,$$

where Φ is the work function of the metal, and consistent energy units are used for each term in the equation. Explain briefly the physical process that this equation represents.

Outline an experiment by which you could determine the values of h (or of h/e) and Φ.

For caesium, the value of Φ is 1.35 V.

(*a*) What is the longest wavelength that can cause photoelectric emission from a caesium surface?

(*b*) What is the maximum velocity with which photoelectrons will be emitted from a caesium surface illuminated with light of wavelength 400×10^{-9} m?

(*c*) What p.d. will just prevent a current passing through a caesium photocell illuminated with light of wavelength 400×10^{-9} m? (O)

22.6 A rubidium surface was illuminated by light of different wavelengths. Values of the maximum kinetic energy of the emitted electrons were determined at each wavelength. The results obtained are tabulated below:

Wavelength/10^{-10} m	3650	4047	4358	4797	5461
Max K.E./eV	1.32	0.98	0.77	0.50	0.19

Find the maximum wavelength for which photoelectric emission can be observed from rubidium. (C)

Also determine the value of Planck's constant, h, and of the work function of rubidium.

22.7 When ultraviolet light of wavelength 150×10^{-9} m falls in a vacuum on the clean surface of a well insulated sheet of a certain metal, the metal attains a potential of +6.6 V relative to its surroundings. For wavelengths of 300×10^{-9} m and 450×10^{-9} m, the corresponding potentials are +2.4 V and +1.0 V respectively. What theory can be advanced to explain these data?

Explain what is meant by threshold wavelength and work function, and use the above data to determine the numerical value for these quantities and for Planck's constant.

Show how the theory put forward explains (i) the effect of increasing the

intensity of the light falling on the metal surface, (ii) the effect of illuminating the surface simultaneously with light of all three wavelengths, the surface being initially at earth potential in each case. (O & C)

22.8 An effective point source emits monochromatic light of wavelength 450×10^{-9} m at a rate of 0.11 W. How many photons leave the source per second? Light from the source is emitted uniformly in all directions, and falls normally on the cathode, of area π cm^2, of a photocell at a distance of 50 cm from the source.

Calculate the photoelectric current, assuming that 10 per cent of the photons incident on the cathode liberate electrons. (O & C)

*22.9 Describe the construction of an alkali metal photocell (and also a suitable circuit for amplifying and measuring the current that passes when the sensitive surface is exposed to light).

The figure represents the current-voltage characteristic of such a cell; the current is the actual (unamplified) photocurrent. The cell is illuminated with a beam of monochromatic light of constant intensity and wavelength 4.0×10^{-7} m.

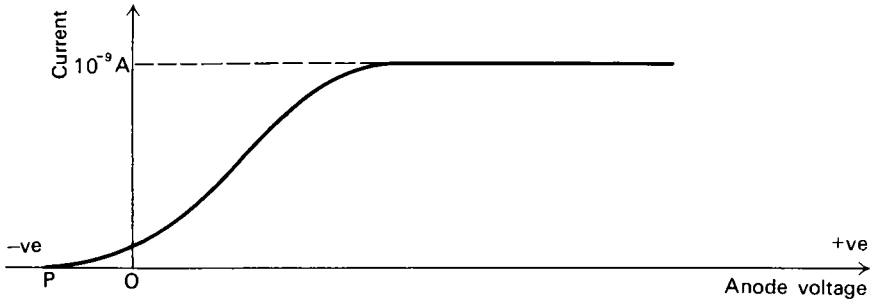


Fig. e 82

(*a*) Compare the general form of the curve with the characteristic of a thermionic diode.

(*b*) Explain the intercept OP on the graph (here shown exaggerated) and estimate its actual value, given that the work function of the cell surface is 2.2 V.

(*c*) How many electrons cross the cell per second when the photocurrent has its saturation value of 10^{-9} A?

(*d*) Assuming that each quantum of light energy absorbed liberates one electron, find the power in watts received by the sensitive surface of the cell. (O)

*22.10 Sodium, relative atomic mass 23 and density 970 kg/m^3, requires 3.66×10^{-19} J for the photoelectric ejection of an electron from one of its atoms. Making any reasonable simplifying assumptions about the arrangement of atoms in a lump of sodium, calculate the area which the cross-section of a single sodium atom presents to incident light. Hence calculate how long it would take for the energy needed to eject an electron to arrive on a single atom from a 100 W lamp placed 1 m from the sodium. Assume that 1 per cent of the energy supplied to the lamp is radiated as light capable of causing photoemission.

Discuss your calculated value for this time, referring to the known behaviour and uses of photoelectric devices, and explain any discrepancy.

Discuss whether 1 per cent appears to be reasonable for the proportion of the energy supplied which is radiated as light capable of causing photoemission.

What difference would it make to the current from a photocell if it were covered with a sheet of polaroid?
(Avogadro constant $N_A = 6.0 \times 10^{23}$ mol^{-1}.) (S)

22.11 X-rays are emitted from a target bombarded by electrons which have been accelerated from rest through 10^5 V. Calculate the minimum possible wavelength of the X-rays, assuming that the corresponding energy is equal to the whole of the kinetic energy of one electron. (O & C)

22.12 The density of sodium chloride is 2.17 g/cm^3, and the mass of 1 mole is 58.5 g (taking each molecule to consist of two ions, a sodium ion and a chlorine ion); in the solid state, the ions are arranged alternately in a regular cubic arrangement. How many ions are there in a crystal in the form of a cube of side 1 cm? What is the distance between adjacent ions? Find the smallest angle with a plane of ions at which X-rays of wavelength 1.54×10^{-10} m could be reflected from sodium chloride. How many orders of reflection could be observed from these planes of ions?
(Avogadro constant $= 6.023 \times 10^{23}$ mol^{-1}.) (C)

22.13 The wavelength of the K_β line in the characteristic X-ray spectrum of molybdenum is 6.3×10^{-11} m. What is the glancing angle in the first order with the planes of ions in a crystal where the spacing between the ions is 2.8×10^{-10} m?

Explain why the continuous spectrum of the radiation from an X-ray tube has a short wavelength limit, the value of which depends on the p.d. across the tube. What is the value of this limit for a p.d. of 40 kV? (O)

Would this p.d. of 40 kV necessarily cause the K_β line to appear in the spectrum? Explain your answer.

22.14 Describe a modern form of X-ray tube, and explain briefly the energy changes that take place while it is operating.

Calculate the energy in electronvolts of a quantum of X-radiation of wavelength 1.5×10^{-10} m.

An X-ray tube is operated at 8000 V. Why is there no radiation of shorter wavelength than that calculated above (while there is a great deal of longer wavelength)? When the voltage is increased considerably above this value why does the spectrum of the radiation include a few (but only a few) sharp strong lines, the wavelengths of which depend on the material of the target? (O)

Chapter 23

Electrons in the atom

23.1 The atom's nucleus

Until the discovery of the electron, the atomic theory had effectively been that of John Dalton, based on such quantitative chemical evidence as the law of multiple proportions. Its applications in physics concerned such matters as the kinetic theory of gases, in which atoms and molecules were treated as being very small solid particles in a continual state of random motion, and possessing perfect elasticity. Electricity entered the picture with the ionic theory of electrolysis. It is worth noting that our present knowledge of the internal structure of the atom does not destroy the relevance of the solid particle point of view, as applied, for example, in the kinetic theory of gases, but merely throws a little more light on it. The elastic collisions between molecules are seen to be collisions between like charges, as the electrons surrounding the nucleii of the different atoms approach each other. This does not alter the basic mechanics of the problem, but it leads us to expect that at higher temperatures the collisions might sometimes be inelastic, causing excitation or ionization of the atoms.

Thomson's discovery that the electron is a subatomic particle and a component part of different atoms, indicated that the atoms must have some kind of internal structure; it did not, however, provide the means of investigating what this structure might be. This came from the discovery of radioactivity by Becquerel in 1896. The original aim of Becquerel's experiments had been to test a theory about the origin of the X-rays radiated from a discharge tube. Röntgen had observed these to come from the target which halted the passage of the cathode rays in the discharge tube, and this place is often marked by fluorescence. Becquerel set out to investigate whether the essential feature of the X-ray source might not be this fluorescence, the use of cathode rays to produce it being incidental. We now know that this is not the case; it is the fluorescence that is, in fact, incidental, and it does not always occur. So Becquerel was bound to fail in his original expectation, but it led him to experiment with a variety of fluorescent materials, and he discovered that the fluorescence of uranium salts was unusual, being caused by penetrating radiations emitted by the salts themselves. Failing to establish his original premise about the origin of X-rays, Becquerel had discovered instead something of greater importance, the radioactivity of uranium.

All the atoms of the heaviest elements are radioactive, meaning that their nucleii are to various degrees unstable, and when such an atom breaks down, its nucleus ejects either an α- or a β-particle, dependent on the element. The β-particle is a high speed electron, and it is not of particular interest in the present context, but the α-particle is a helium nucleus, He^{2+}, which is ejected from the nucleus of the radioactive atom with very considerable energy, in the region of 5 MeV (8×10^{-13} J). Having such mass and energy, and being a bare nucleus unprotected by any satellite electrons, the α-particle does not rebound from an atom or molecule as if

this were impenetrable, but it cleaves its way through the electron shells of the atom, forcing the electrons out of its path.

Alpha particles from a radioactive source were used by Rutherford in this way as projectiles, and by this means he discovered the existence of the atom's nucleus, showing it to be positively charged and extremely dense, holding in its very small bulk almost the entire mass of the atom. By far the greater part of the atom's territory, it appeared, lay outside the nucleus, and was a space in which only the electrons of the atom would be found. Atomic physics is concerned with the behaviour of these electrons, and, in particular, with the levels of energy which they occupy, the main source of information being the spectrometer, used to analyse the electromagnetic radiations by which the electrons signal their behaviour.

The electronvolt

Since in this chapter and the succeeding ones, particles of nuclear and subnuclear size will be much under discussion, it will be convenient to describe their energies in terms of a unit which is much smaller than the joule. This practice has already been introduced, earlier in this paragraph, in describing the energy of an α-particle. The unit, the electronvolt, is particularly appropriate in this context and has been generally used for this purpose up to the present time, and will be used here. It is the energy possessed by any particle carrying charge e when placed at a potential of 1 V. Thus,

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J, approximately,}$$

where the approximate value of the charge e is 1.6×10^{-19} C.

Perhaps it is unnecessary to point out that the electronvolt, although convenient and descriptive, is not a unit which belongs to the SI system, in which the only proper unit of energy is the joule. If used in calculations, the electronvolt demands some careful consideration: a helium nucleus, He^{2+}, for example, is doubly charged, and if a p.d. V were applied to accelerate it to an energy of 5 MeV, this would not require the value of V to be 5×10^6 V, but 2.5×10^6 V.

23.2 The discovery of the nucleus

The experiments which established the existence of the nucleus were carried out by Geiger and Marsden between 1910 and 1913, working under Rutherford at Manchester. In the first of these, they studied the effect when a pencil of α-particles fell on a very thin metal foil, so thin that most of the particles passed through it. This required the foil to be very thin indeed, not more than about 1 000 atoms thick, because the penetrating range of α-particles is only small; they cause very intense ionization of the atoms which lie in their path, and lose their own energy in the process. Using such a foil, Geiger and Marsden found that the large majority of the α-particles passed through it, suffering only slight deflections or no deflection at all; a small number, however, behaved quite differently, coming through with a very large deflection, or in some cases bouncing back from the foil, as if they had met an obstruction which the other particles had missed.

The separate atoms in a metal foil are, presumably, packed close together, and on this basis the diameter of the atom works out to be of the order of 10^{-10}m. If the mass of the atom, and its charge, were distributed throughout its volume, it should not have been possible for any of the α-particles to bounce back from the foil, as Geiger and Marsden discovered that one occasionally did. Rutherford, therefore, deduced that the mass and positive charge must all be concentrated in a much smaller nucleus, and that the large deflection resulted when the occasional particle happened to pass very close to this nucleus.

On the basis that the concentrated positive charges on the atomic nucleii repelled the α-particles in accordance with the inverse square law, Rutherford derived an equation which would give the angular distribution of scattered α-particles, and this was tested by Geiger and Marsden in 1913. Their results agreed closely with those predicted by the equation, which also gave them a means of estimating the charge carried by the nucleus, and its approximate diameter. This worked out to be a little more than 10^{-14} m, as compared with the value of about 10^{-10} m for the complete atom.

The principle of the apparatus used by Geiger and Marsden is illustrated in Fig. 23.1. The method used to detect the α-particles was effectively the same as that which had led Becquerel to discover radioactivity in the first place, namely

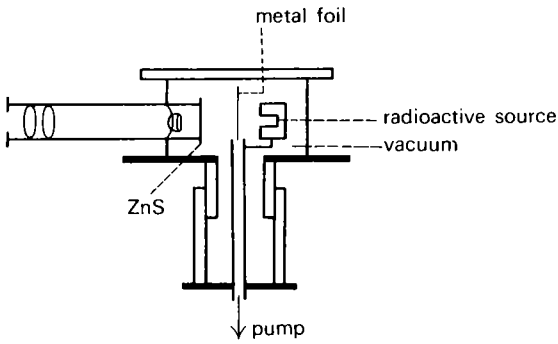


Fig. 23.1 The apparatus of Geiger and Marsden

by the fluorescence produced when the particles struck a suitable material, in this case zinc sulphide; observed in a darkened room, each separate α-particle could be detected by the small flash of light produced as it reached the screen. The radioactive source and the metal foil were fixed in position inside an evacuated metal box, and the box itself, carrying the zinc sulphide screen and the observing microscope, could be revolved about an axis containing the foil, the angle being measured on a circular scale.

23.3 Problems relating to Rutherford's atom model

It was now necessary to account for the fact that the electrons of the atom, being negatively charged, and attracted by the positively charged nucleus, do not fall into the nucleus. The simplest explanation would be to picture the electrons as moving in orbits round the nucleus, like satellites launched to travel round the earth. A satellite can be put in orbit at any height, given the appropriate velocity when it gets there; gravitational attraction then provides the centripetal force required to hold it in its path. Similarly, we might think, does an electron behave, the centripetal force, in this case, being due to the electrical attraction of the nucleus. There is a difficulty about this picture of the atom, however, which renders it untenable, and this is due to its electrical nature. An electron, circling round the nucleus in this way, is an electric charge in a condition of high frequency oscillation, and in this respect its behaviour resembles that of the electrons in a very high frequency radio transmitting aerial. Accepting Maxwell's electromagnetic theory, such an oscillating charge must inevitably radiate energy into space, in the form of an electromagnetic wave, having frequency equal to that of the oscillating charge. The very high frequency involved should make this radiation all the more

intense because, according to Maxwell, the power radiated is proportional to the square of the frequency. All this would have practical results which do not, in fact, occur. In losing its energy in this way, the electron would be pulled in towards the nucleus, gaining in frequency of rotation as it fell. This means that the atom would radiate waves which progressively increased in frequency until, in a very short space of time, the electron disappeared into the nucleus.

It is undoubtedly true that, when an electron in the atom does lose energy, it radiates, the waves having the frequency of X-rays, ultraviolet, visible light, or infra-red, according to the circumstances. It is, indeed, by studying the spectrum of this radiation that most of the evidence about the electron's behaviour is obtained. Nevertheless, an atom does not radiate in all circumstances, as Maxwell's theory, applied to the simple model, suggests that it should; nor is the spectrum, at any rate in the case of a gas or vapour, the continuous sequence of wavelengths which would result if indeed the electrons were falling in to the nucleus all the time. Before the gas or vapour can emit radiation at all it must first be sufficiently energized, and the spectrum which results consists of a series of isolated wavelengths, or spectrum lines. These difficulties were first tackled with any success in 1913 by Bohr, who developed a very successful model of the hydrogen atom by introducing some ideas based on quantum theory and combining these with some of the classical laws. This is still the most satisfactory model to accept if a simple picture is needed, but it proves to be an inadequate basis for describing accurately the behaviour of atoms more complicated than hydrogen.

23.4 Bohr's model of the hydrogen atom

The visible spectrum of atomic hydrogen, as obtained from a hydrogen discharge tube, for example, is a series of spectrum lines (Fig. 23.2) known as the Balmer

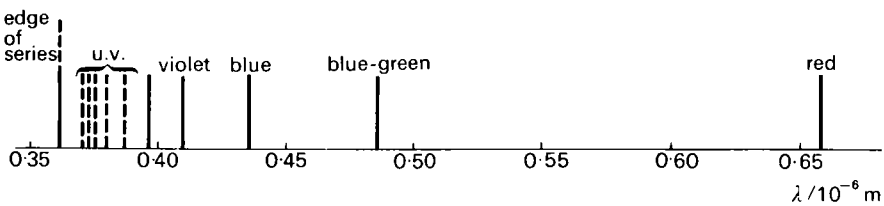


Fig. 23.2 The visible spectrum of atomic hydrogen — the Balmer series

series. There is a simple mathematical relation between these lines, the equation for which, discovered by Balmer in 1885, is

$$\frac{1}{\lambda} = R\left[\frac{1}{2^2} - \frac{1}{n^2}\right], \tag{23.1}$$

where n is an integer having the successive values $n = 3, 4, 5$ etc. for each successive wavelength, and R is a constant known as the Rydberg constant. Its value is

$$R = 1.097 \times 10^7 \text{ m}^{-1}.$$

There is a similar series of longer wavelength spectrum lines, all lying in the infra-red, which is called the Paschen series, and the equation for these, found by Paschen in 1908, is

$$\frac{1}{\lambda} = R\left[\frac{1}{3^2} - \frac{1}{n^2}\right]. \tag{23.2}$$

It was this information about the radiation from hydrogen which Bohr's model had to fit.

The rules imposed by Bohr on the single electron of the hydrogen atom involved extending the application of quantum conditions to govern what the orbiting electron was permitted to do, as well as applying to the radiation which it emitted. He accepted that the electron's orbit was in accordance with the classical equation, which identified centripetal force with electrical attraction between electron and nucleus, but he freed the electron from the other classical rule, of Maxwell, that it should be radiating energy away, and he accepted that it did not. Instead, he postulated that the electron was denied the continuous range of possible orbits and corresponding energies which conventional theory would allow, and that it was permitted to take up only one or other of a limited selection from among these orbits, the selection being made in accordance with a quantum rule. Thus, the permitted orbits of the electron (assuming them to be circular) are specified by two equations. The first of these gives the centripetal force

$$m_e \frac{v^2}{r} = \frac{1}{4\pi\epsilon_0} \cdot \frac{e^2}{r^2}, \tag{23.3}$$

where m_e, e are respectively the mass of the electron and the charge of both negative electron and positive hydrogen nucleus. This is a slightly simplified version of Bohr's equation, since the nucleus has been assumed here to be too massive to contribute to the relative motion.

The second equation, involving a quantum rule, and not classical physics, is

$$m_e vr = n\frac{h}{2\pi}$$

where h is the Planck constant, and n is a positive integer, 1, 2, 3 etc., which is now known as the principal quantum number. The product $m_e vr$ is the angular momentum of the electron about the axis of its supposedly circular orbit, and the equation is stating Bohr's postulate that this is quantized; the orbits permitted to the electron are limited to those in which its angular momentum is an integral multiple of the amount $h/2\pi$.

An alternative form for this equation, which will presently acquire a special significance (section 23.6), is

$$2\pi r = n\frac{h}{p}, \tag{23.4}$$

where $p = mv$ is the linear momentum of the electron.

The total energy of the orbiting electron, kinetic and potential combined, E, may be deduced from these two equations, giving an expression with the form

$$E = -\frac{k}{n^2}, \tag{23.5}$$

where k is a function of the quantities h, e, m_e, and ϵ_0. The value of k works out to be

$$k = 21.8 \times 10^{-19}\ \text{J}$$

or, in units which are not SI, $k = 13.6$ eV

i.e. $= 13.6$ V per electron. (23.6)

[The negative sign appears in equation 23.5 because the electron has been taken to have zero energy where completely removed from the influence of the nucleus.]

In the normal, or ground state, of the hydrogen atom, its electron has the smallest

amount of energy permitted to it by equation 23.5, and this places it in the orbit which is closest to the nucleus, corresponding to the value $n = 1$.

The spectrum lines emitted by hydrogen are accounted for by following Planck, not Maxwell. Supposing that, by some means, the hydrogen atom accepts extra energy, this will put it in an excited state, having its electron in one of the orbits of larger radius and greater energy, represented by inserting $n = 2,3,4$, as the case may be, in equations 23.4 and 23.5. The atom can remain in this excited state for only a small fraction of a second before the electron falls back to the ground state again. Whenever the electron loses energy in this way, by falling from a higher energy level to a lower one, the surplus energy is released as a single photon of radiation, in accordance with the equation

$$E' - E = h\nu. \tag{23.7}$$

Combining this with equation 23.5, we have

$$h\nu = k\left[\frac{1}{n^2} - \frac{1}{(n^1)^2}\right],$$

whence

$$\frac{1}{\lambda} = \frac{k}{hc}\left[\frac{1}{n^2} - \frac{1}{(n^1)^2}\right], \tag{23.8}$$

where c is the velocity of light. Not only does this equation have the same form as the equations of the Balmer and the Paschen series; more impressively, the value of the constant term, k/hc, works out to be equal to that of the Rydberg constant R to a remarkable degree of accuracy.

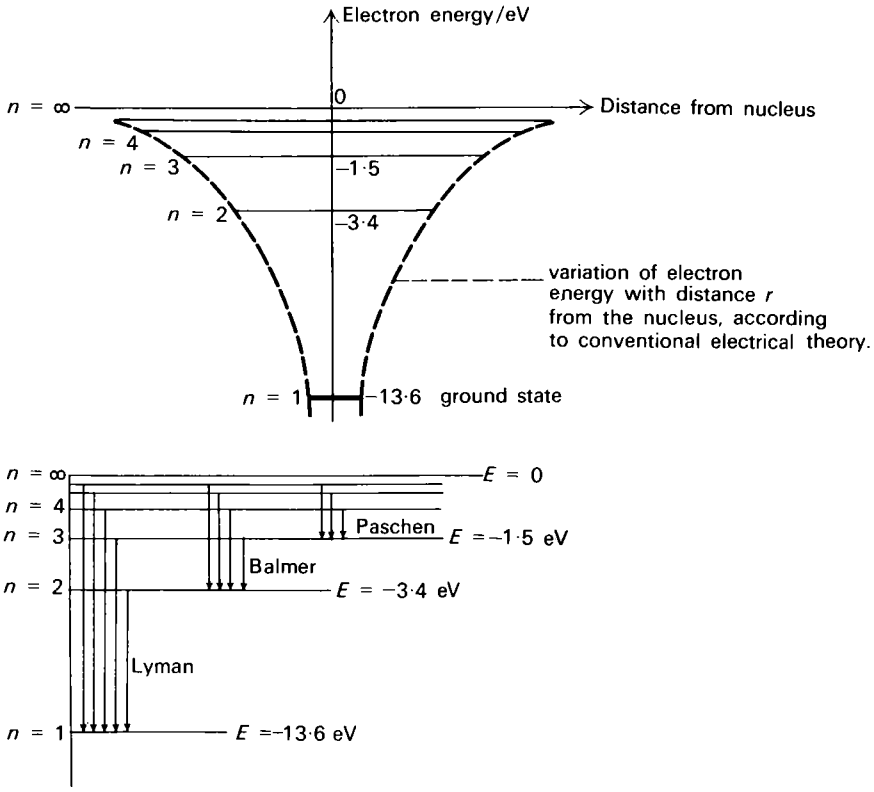


Fig. 23.3 Diagrams representing the energy levels of the electron in the hydrogen atom

The energy aspect of Bohr's model is pictured in Fig. 23.3 in two slightly different ways. It will be noted that when the atom is in the ground state the energy of the electron is –13.6 eV, and it must acquire extra energy at least equal to 13.6 eV if it is to escape from the influence of the nucleus entirely. For this reason, the ionization potential of hydrogen is said to be 13.6 V.

Bohr's theory had the essential quality of a good theory; that it led to new discoveries as well as explaining old ones. One such possibility is very evident in the energy level diagrams: the visible spectrum of hydrogen does not result from transitions of electrons to the ground state, but to the energy level for which the principal quantum number is $n = 2$. If an excited electron were to return directly to the ground state, it would radiate a photon of such energy that its wavelength would lie in the short ultraviolet range. Similarly, a series of long infra-red wavelengths would be expected to result from transitions to the level for which $n = 4$. Both these series of spectrum lines were predicted by Bohr, and were later discovered, the ultraviolet series by Lyman in 1914, and the infra-red by Brackett in 1922.

Where Bohr's model proved inadequate was in the attempts to develop it to account for the more complicated behaviour of more complicated atoms than hydrogen, where there is more than one electron to consider. Nevertheless, all gases and vapours, when dissociated into atoms, give line spectra, and these are accounted for in terms of Bohr's original postulate, that radiation of photons results from electron transitions between different energy levels in accordance with equation 23.7. From this starting point modern atomic theory has developed, producing the periodic table with which we are familiar, the electrons in the atoms being placed in their appropriate shells and subshells and orbitals.

23.5 Moseley's experiments. Atomic number Z

In describing Bohr's theory, we have assumed it to be obvious that the hydrogen atom has but one electron, and that the positive charge of its nucleus is $+e$. This significant aspect of the meaning of atomic number would not have been quite so certain in 1913, although the methods of Geiger and Marsden were later adapted to give an estimate of nuclear charge. Previously, atomic number had referred only to the place occupied by an element in the periodic table, drawn up by placing the elements in order of their atomic weights. A more fundamental interpretation of it resulted from experiments carried out by Moseley in 1913 on X-ray spectra.

Moseley measured the wavelengths of the characteristic X-rays radiated from the target of an X-ray tube (section 22.10), using in succession a large number of different elements as the target, and he observed the characteristic X-ray lines photographically using an X-ray spectrograph. His results showed a very remarkable correlation between the spectra of the different elements. Each of the K series of lines, for example, which is the series having the shortest wavelength, could be clearly identified for each element, and the wavelength became progressively longer as atomic number decreased; the K series for hydrogen, indeed, has such long wavelengths that this does not consist of X-rays at all, but is the ultraviolet series later discovered by Lyman. The same correlation appears between the L lines.

When the periodic table had been drawn up by arranging the elements in order of their atomic weights, this had revealed a periodic variation in the chemical properties of the elements, but there had been one or two discrepancies; in order to follow correctly the sequence of the chemistry, it had been necessary to displace one or two elements from the orderly sequence of atomic weights. Taking atomic number Z to refer to the place of the element in this corrected table, Moseley

plotted the square root of the frequency against atomic number for each sequence of corresponding X-ray lines, and obtained in each case a very convincing straight line graph (Fig. 23.4).

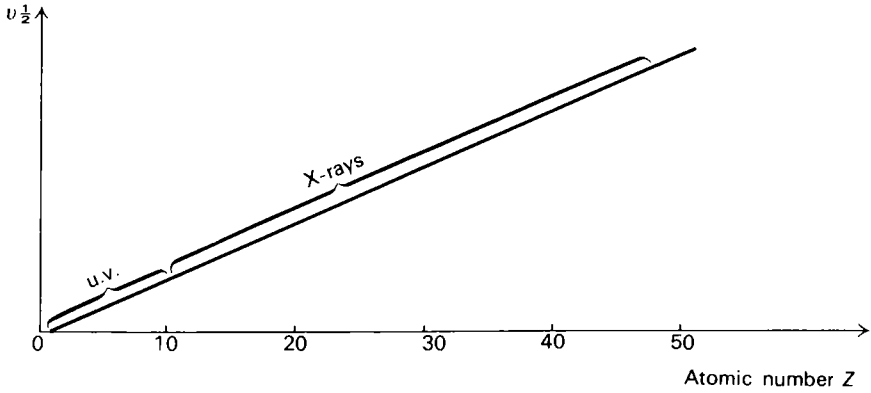


Fig. 23.4 Characteristic X-ray spectra. Moseley's graph for one of the K lines

Taking advantage of the hindsight which makes things clearer to us than presumably they were at the time, we can accept that, when an atom has a number of electrons, these are to be found in the sequence of K, L, M shells, etc. Whereas the visible spectrum of such an element arises from energy transitions of an electron from the outer shell, the X-ray spectrum is concerned with one or other of the innermost shells, and this is why Moseley's results with X-rays gave the information which they did, which is information about the atom's nucleus. When he plotted atomic number Z on his graph, what he was really using was a number which represented the amount of charge held by the nucleus.

If the graph of $\nu^{\frac{1}{2}}$ against Z for one of the K lines is extrapolated back to the value $\nu^{\frac{1}{2}} = 0$, we find that the intercept on the Z axis is $Z = 1$. This can be explained on the basis that the K shell holds two electrons, and the X-ray photon has resulted from the expulsion of only one of them. The effective charge, therefore, which controls the radiated frequency is the charge on the nucleus, Ze, reduced by an amount e, to allow for the effect of the single electron left in the K shell. Thus, we have

$$\nu^{\frac{1}{2}} \propto (Z-1)e. \tag{23.10}$$

[The Lyman series for hydrogen does not exactly fit this scheme, as was suggested earlier, because the hydrogen atom has only one electron in its K shell in the first place.]

Taking Moseley's results, coupled with the results of such experiments as those of Geiger and Marsden, these make up strong evidence that atomic number Z counts the number of electronic charges, positive in sign, held by the nucleus, and consequently the number of electrons outside the nucleus of the neutral atom.

23.6 Wave mechanics

The empirical quantum rules of Bohr, proving inadequate to describe the spectra of more complicated atoms than hydrogen, led to the development of wave mechanics, or quantum mechanics. This gives a mathematical account of the atom which is both consistent with the practical evidence available, notably from the spectrometer, and affords a strong basis for the chemistry of the elements in terms

of the arrangement of their electrons. If we require a simple picture of the atom, however, wave mechanics does not lend itself to accurate interpretation in such a way, and we return to something like an extension of Bohr's model, such as the diagrams of electron shells which were drawn in chapter 1 (Fig. 1.1).

Wave mechanics was developed from a suggestion made by de Broglie in 1924. Quantum theory attributes some of the properties of a particle to an electromagnetic wave, and we are forced by the practical evidence to accept this, although it cannot be explained on the basis of Maxwell's classical theory of electromagnetic radiation. De Broglie suggested that, in certain circumstances, the inverse might be true, and particles might be found to share the properties of a wave, having a wavelength associated with them, and exhibiting such wave effects as diffraction. This theory was preceded by the discovery, by A. H. Compton, of a further particle behaviour on the part of radiation, known as *the Compton effect*. According to Einstein's special theory of relativity, energy and mass are equivalent to each other, in accordance with the equation

$$E = mc^2, \tag{23.11}$$

c being the velocity of light in a vacuum. If we apply this equation to the energy of a photon, and substitute the expression $h\nu$ for the energy E, an expression results giving the mass of a photon, viz.

$$E = h\nu,$$

$$h\nu = mc^2,$$

whence

$$\frac{h}{\lambda} = mc, \tag{23.12}$$

which follows because $c = \nu\lambda$. Compton postulated that equation 23.12 represents the momentum, $p = mc$, possessed by a photon, and that the ordinary laws of mechanics should apply, in particular the law of conservation of momentum, to collisions between photons and particles. It will be observed that the effective mass of a photon is inversely proportional to its wavelength, and we can obtain an idea of how large this mass might be by inserting the value $\lambda = 10^{-10}$ m in equation 23.12. This is, as we have already seen, the wavelength of a fairly short wavelength X-ray, and any wavelength longer than this would have a smaller mass. Thus

$$m = \frac{h}{\lambda c}$$

where $c = 3.0 \times 10^8$ m/s, $h = 6.6 \times 10^{-34}$ J·s, $\lambda = 10^{-10}$ m,

whence

$$m = \frac{6.6 \times 10^{-34}}{3.0 \times 10^{-2}} \text{ kg}$$

$$= 2.2 \times 10^{-32} \text{ kg}.$$

and this is only one fortieth of the mass of an electron.

What Compton's theory meant in practical terms was that if a photon collided elastically with a free particle, there should be some exchange of momentum between photon and particle, with the result that the deflected photon should have a momentum, and consequently a wavelength, which was different from its original value. In accordance with mechanical laws, however, the change would be appreciable only if the mass of the photon were comparable in size with that of the particle with which it collided. The calculation given above suggests that this would be the case if the photon were a hard X-ray and the particle an electron.

Compton tested his theory by using the free electrons contained in graphite to scatter a beam of very short wavelength X-rays, and observed the changes in wavelength which he had expected, in complete agreement with the equations which he had worked out. This is what is known as *the Compton effect*.

De Broglie's suggestion with respect to the wave properties of a particle took over equation 23.12, and substituted in it the momentum $p = mv$ of the particle in place of the momentum $p = mc$ of the photon. That is, he postulated that, in certain aspects of its behaviour, a particle having momentum mv would be found to resemble a wave having a wavelength λ, given by the relation

$$\lambda = \frac{h}{mv}. \tag{23.13}$$

This equation was tested by Davisson and Germer, in America, using relatively slow moving electrons, and by G. P. Thomson in Scotland, using fast ones. The corresponding wavelengths are the same as those of very short X-rays, and the methods adopted resembled those of X-ray diffraction, except that the diffracting crystals were crystals of metal, and these produced diffraction patterns, not of X-rays, but of electrons used, in a narrow beam, to bombard the metal.

Taken together, equations 23.12 and 23.13 obviously contradict the more clear cut ideas of classical physics. We are accustomed to the view that waves and particles are quite distinct in their behaviour. Waves involve interference and diffraction and the wide distribution of energy; particles follow well defined tracks and carry their energy with them. Quantum theory and wave mechanics blur out these simple distinctions. Nevertheless, they do not completely remove them, and fundamental differences remain. The electromagnetic wave in vacuo can have only one velocity, the velocity c; a particle, on the other hand, can move with any velocity, provided that this is less than c. A further aspect worth noting is that the wavelength associated with a particle, which is in inverse proportion to its momentum, is, in the vast majority of cases, extremely short. Only for an electron, or a slowly moving nucleon, is the momentum sufficiently small, and the wavelength in consequence sufficiently long, to produce any appreciable wave-like effects.

The interpretation that is given to this now established fact that small particles display some wave-like properties is that this concerns probabilities. The wavelength associated with a beam of electrons represents in some way an indication of the probable whereabouts of the electrons. The diffraction patterns obtained in practice are a demonstration of this probability working out, so that more electrons arrive at one place on a photographic plate than on another. Conversely, the same interpretation can be placed on the fact that waves have a particle aspect to their behaviour, as suggested by the quantum theory relation, $E = h\nu$, and by Compton's momentum equation $p = h/\lambda$. An interference or diffraction pattern produced with electromagnetic waves may, alternatively to the normal wave approach, be regarded as a probability pattern for the paths taken by photons. In cases where either the wave or the particle point of view may be adopted to solve a problem, they both give the same answer, but the two approaches cannot usefully be combined with each other.

De Broglie's equation may be applied to Bohr's quantum rule for the electron orbits of hydrogen (equation 23.4) to give this some theoretical basis. Thus

$$2\pi r = n\frac{h}{p}, \tag{23.4}$$

where r is the radius of the electron's supposedly circular orbit, and p is its momentum mv. But we now have, also, de Broglie's equation

$$\lambda = \frac{h}{p}, \tag{23.13}$$

whence, combining the two equations,

$$2\pi r = n\lambda, \tag{23.14}$$

i.e. treating the electron as being associated with a wavelength λ, the orbits permitted to it by Bohr's quantum rule are those which have the appropriate lengths, $2\pi r$, to accommodate an integral number of electron wavelengths. This seems to be reasonable; if the wavelengths did not fit the orbit in this way, it would be difficult to see that wavelength could, in this context, have any useful interpretation.

23.7 Pauli's exclusion principle

The application to Bohr's hydrogen model is the nearest that we can get to fitting the concept of electron wavelength into an attempt to picture actual orbits. Wave mechanics, which incorporates de Broglie's equation, provides a description of the electron distribution in atoms more complicated than hydrogen, but the treatment is mathematical, and cannot be accurately interpreted in terms of a physical model. The main feature which emerges is that the energy which an electron in the atom is permitted to possess is governed, not by one quantum rule, as Bohr had suggested, but by four, and this means that each available energy level is specified by the values of four quantum numbers. The most important of these in its effect, the principal quantum number n, corresponds to the single quantum number which was postulated by Bohr, and provides the electrons of all atoms, not only those of hydrogen, with a sequence of K, L, M shells, and so on, which they can occupy. The effect of the other three quantum numbers is to break down each shell into a number of subshells, and each subshell into a number of orbitals, all representing energy levels which differ from each other, although in many cases only slightly. Additionally, following Heisenberg's uncertainty principle (section 11.2) which is an expression of the probability aspect of wave mechanics, we have to accept that, in placing an electron's energy level precisely, we lose any clear sight of its movement.

When the atom is in the ground state, each electron occupies the lowest level of energy which is available to it. The four quantum numbers identify the energy levels, as we have already said, but whether or not an electron can be placed in one is decided by using an important extra rule which was suggested by Pauli, on empirical grounds, in 1925. This is known as Pauli's exclusion principle. It states that each electron in the atom must differ from all the others in at least one of the four quantum numbers allotted to it; so long as one particular energy state is already taken up by an electron, all other electrons are excluded from it.

The diagram of the sodium atom in Fig. 23.5 is a reproduction from Fig. 1.1, included here as a simple, if inaccurate, illustration of Pauli's exclusion principle. The actual circular orbits, of course, have little significance, but the placing of the eleven electrons of sodium in the separate K, L, M shells is relevant to modern theory. These shells correspond, respectively, to the values 1, 2, 3 of the principal quantum number n. Once all the four quantum rules and the exclusion principle are satisfied, the total number of electrons which can be placed in the shell with principal quantum number n is equal to $2n^2$, with the result that the maximum complement in the K and L shells is 2 and 8 respectively, which is the distribution shown in the diagram. Hence the eleventh electron of the neutral sodium atom has to occupy a place in the M shell.

Since any one shell comprises, with its subshells and orbitals, quite a number of

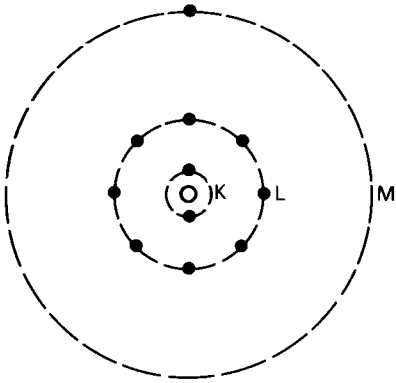


Fig. 23.5 A simple model representing the neutral sodium atom

different energy levels, an atom such as sodium, when in an excited state, can have an electron in quite a variety of different circumstances, which accounts for the rather complex spectrum which most gases and vapours emit; each spectrum line, however, corresponds to an electron transition from one energy state to another, in accordance with Bohr's original equation

$$h\nu = E' - E. \tag{23.7}$$

The two familiar yellow lines, for example, which make up the visible sodium spectrum, both result from an electron transition from one subshell in the M shell to another of slightly lower energy in the same shell.

*23.8 Energy levels in a solid material

To take a description of subshells and orbitals any further would be an unnecessary duplication of what is to be found in the chemistry books. An aspect which is more relevant to electricity is the effect which neighbouring atoms have upon each other when they are arranged close together in a solid. An electron in the outer shell of an isolated atom cannot accept an addition of energy unless this is the appropriate amount to put it in a vacant place in a higher energy orbital of the atom, or unless the energy is sufficient to ionize the atom by freeing the electron completely.

In the case of the sodium atom, for example, pictured in Fig. 23.5, the single valency electron normally occupies an orbital in the M shell, and the minimum increment of energy which this electron is permitted to accept is about 2 eV, which would raise it to a higher energy orbital in the same shell; to release the electron entirely from the atom's influence would require 5.1 eV, in other words, the ionization potential of sodium is 5.1 V. In a solid piece of sodium, however, the atoms are close enough together for their fields to interact, and a valency electron can become free to wander about in the metal, without having sufficient energy to allow it to leave the metal surface. This is because the electron is free of the overriding attraction of the single nucleus, yet it is still attracted by the collection of positive ions which make up the mass of the metal.

In accordance with quantum rules, however, the electron cannot accept the extra energy which the drift motion associated with conduction would involve unless this places it in a vacancy among the permissible energy levels. If the gap to be crossed were still the 2 eV that it is for the isolated sodium atom, sodium would be an insulator or a semiconductor. What makes metallic conduction

possible is the multiplication of energy levels which results when atoms are packed closely together in the solid.

If there are N atoms in the solid, each energy level in the valency shell of the single atom splits into N separate levels, and these are placed very close together.

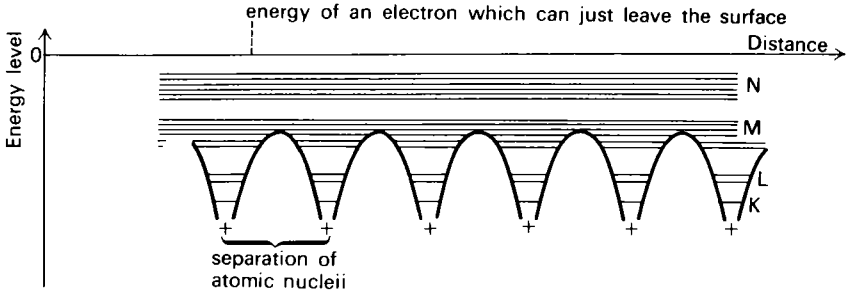


Fig. 23.6 Diagram representing bands of permitted electron energy levels in a lattice of atoms

So many of them are there that the separate levels of the M shell of sodium all merge together to form what is effectively almost a continuous band of permitted energy, and this is known as the valency band. The same effect occurs with the vacant N shell, and with all the shells of higher energy; the same is true, also, of all other metals. There may, however, be a band of forbidden energies separating one permitted band from the next, and whether this is so or not depends on the substance. The electrons in the K and L shells are unaffected by this liberating effect; they are too close to their individual atomic nucleus to escape its overriding influence.

These bands of energy levels are represented in Fig. 23.6. The diagram cannot be entirely realistic, however, if the existence of discrete levels inside the band is to be represented at all. This is because there are, in fact, so many of the levels that they could not all be drawn without merging into one another. The hoop-like curves included in the diagram are a development from Fig. 23.3(*a*). They represent the influences of the individual nucleii, as affected by their neighbours, being a plot of potential energy of an electron against distance.

Although there are so many energy levels in the one valency band, these must accommodate the valency electrons from all the atoms in the solid, and these electrons remain subject to Pauli's exclusion principle. It follows, therefore, that a large proportion of the lower levels in the band are occupied; it may be that the whole of the valency band is filled up in this way, although this is not the case with sodium. In the example chosen in the diagram, the electrons at all but the very lowest levels in the valency band would be free to wander in the metal. If there are vacant levels in this band, as there are, for example, in sodium or in copper, such electrons are able to accept very small increments of energy, because this places them at vacant levels, and so they become charge carriers. In such a case, the valency band is also termed the conduction band, because the electrons in it can conduct.

In the covalent or the electrovalent crystal, the valency band is completely filled, and a forbidden energy gap of appreciable size separates this band from the conduction band, which is normally empty. Consequently, an electron in such a substance cannot become a charge carrier unless it first receives sufficient energy to lift it across the gap. For this reason, such crystals are not good conductors;

whether the substance is an insulator or a semiconductor at ordinary temperatures depends on the width of the forbidden gap. This is illustrated in Fig. 23.7. The filled energy levels are represented in each diagram by being blacked in.

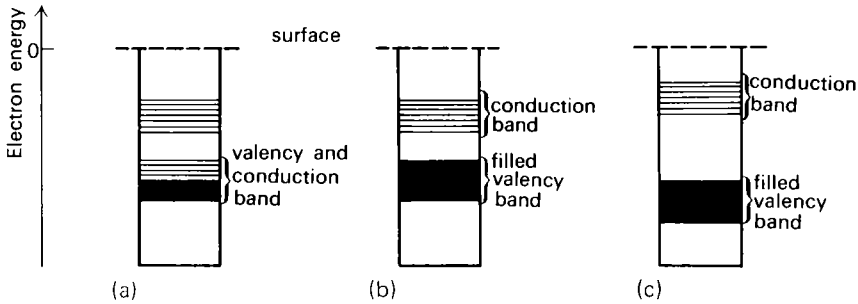


Fig. 23.7 Bands of energy levels in a solid (a) A metal (b) A semiconductor (c) An insulator

*23.9 Semiconductor

When an electron in a semiconducting crystal is in one of the bonding orbitals of the crystal lattice, its energy puts it in the valency band, which is completely filled by such electrons. Consequently, it cannot accept a small increment of energy because there is no vacant energy level to accommodate it; the electron is locked in place in the crystal. To become a free charge carrier, the electron must first receive sufficient energy to take it across the forbidden gap into the empty conduction band. It can now accept further energy, even in very small amounts, because there are plenty of empty energy levels which it can occupy. This is its condition when it has broken free of the valency bonding, and can wander through the crystal, and can act as a charge carrier.

The intrinsic semiconductors

In the case of an intrinsic charge carrier, every electron which reaches the conduction band leaves a vacant place behind it among the energy levels of the valency band, and this represents the energy of the charge-carrying positive hole. The electrons which move to shift these positive holes about do not require much extra energy to permit them to do so; they are at an energy level of the valency band (Fig. 23.8).

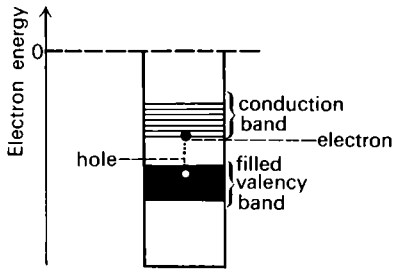


Fig. 23.8 An intrinsic semiconductor. Energy levels of an electron – hole pair

The width of the forbidden gap which separates the valency and conduction bands depends, of course, on the substance, and this controls the temperature at which semiconduction becomes appreciable. The values of this energy gap for carbon (in a diamond lattice), silicon, and germanium are respectively 5.2 eV, 1.1 eV, and 0.7 eV. Perhaps the differences between these three are not surprising, since the carbon atom, with the lowest atomic number, has its valency electrons in the lowest energy shell of the three, the L shell.

Extrinsic semiconductors

The conduction electrons in an *n*-type semiconductor are surplus to the valency bonding, belonging to the foreign donor atoms which have been intruded into the lattice. These donor atoms have not quite the same quantum conditions as native atoms in the crystal, and the extra electron occupies an energy level which places it in the forbidden gap, not far below the conduction band. For such an electron to break free of its parent atom and become a free charge carrier, it only requires to gain the comparatively small amount of energy which will place it in the conduction band. Unlike an intrinsic charge-carrying electron, it does not leave a hole behind it in the valency band.

In a similar way, the acceptor atoms which have been introduced to make a *p*-type crystal produce electron vacancies in the bonding. These do not represent vacant places in the energy levels of the valency band, however, but levels placed in the forbidden gap, a little way above this band. For an electron in the valency bonding to move into this vacancy, it requires the small increment of energy which will place it at this level. It then leaves a positively charged hole behind it in the valency bonding. This hole is a free charge carrier; its energy level is in the valency band, and valency electrons require only a minute amount of extra energy to allow them to move into it, thus shifting the hole about in the crystal.

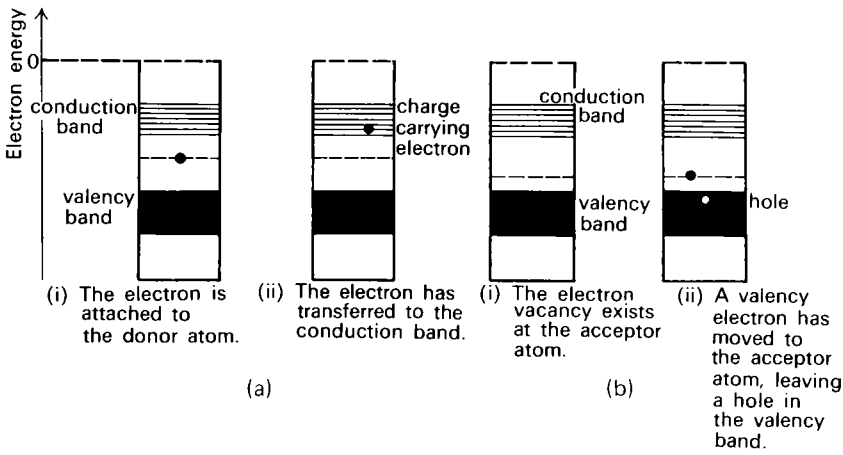


Fig. 23.9 Extrinsic semiconductors (a) An *n*-type semiconductor (b) A *p*-type semiconductor

23.10 Photoconduction and photoemission

A photoelectric action is one means by which an electron may acquire the energy needed to raise it from one energy band to another, and this accounts for the photosensitivity of semiconducting devices such as diodes and transistors. Photoconduction results when an electron in the valency band of a semiconductor

absorbs a sufficiently energetic photon to transfer it to the conduction band, and the longest wavelength radiation which can cause this is therefore given by the equation

$$h\nu = \frac{hc}{\lambda}$$
$$= Ve, \tag{23.15}$$

where Ve is the width of the forbidden energy gap. [Using SI units, this is expressed in joules, and it corresponds to the amount of energy which an electron would possess if raised to a negative potential V.]

Since there is already a large number of electrons in the conduction band of a metal, photoconduction is not a phenomenon which is encountered with metals.

In the case of photoemission, the energy of the photon must be at least sufficient to remove an electron from the highest occupied energy level so that it can escape from the solid entirely, and this is the energy which is associated with the work function, Φ_0, of the substance. Thus the longest wavelength which can cause photoemission is given by

$$h\nu = \frac{hc}{\lambda}$$
$$= \Phi_0 e. \tag{23.16}$$

Because the energy gap between the valency and conduction bands of a semiconductor is in many cases smaller than the work function of any substance is found to be, photoconduction is available at longer wavelengths than photoemission, and photoconducting cells are very useful in the infra-red region; the substance used may be selected to have an energy gap suited to the required frequency.

Fig. 23.10 Photoemission and photoconduction (a) Photoemission. This diagram represents a metal, but the effect can also occur in a semiconductor (b) Photoconduction, in a semiconductor

QUESTIONS

$$h = 6.63 \times 10^{-34}\ \mathrm{J \cdot s}$$
$$c = 3.00 \times 10^{8}\ \mathrm{m/s}$$

*23.1 The wavelengths λ of the lines in the visible spectrum of the hydrogen atom are given by the Balmer series formula

$$\frac{1}{\lambda} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right),$$

where R is the Rydberg constant, $n_1 = 2$, and n_2 is an integer such that $2 < n_2 \leqslant 6$.

The longest wavelength member of this series is the red C line, for which $\lambda = 6.56 \times 10^{-7}$ m. Find (*a*) the value of R, (*b*) the wavelengths of the other visible lines, (*c*) the wavelength of the series limit in the ultraviolet, (*d*) the energy in joules of the second and third levels of the hydrogen atom. (OXF. SCH.)

*23.2 Outline the evidence for the existence of electron energy levels within the atom.

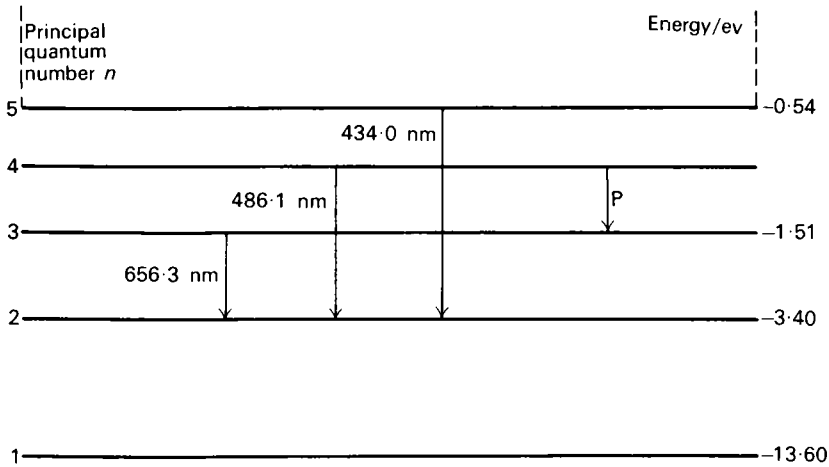


Fig. e 83

The figure represents a typical energy-level diagram for the hydrogen atom, with some details omitted. Energies are given in electronvolts, with the zero of energy corresponding to the complete removal of the electron from the atom; wavelengths of the emitted spectral lines are given in nanometers (1 nm $= 10^{-9}$ m).

(*a*) Explain why the arrows shown indicate possible emission of radiation.

(*b*) Why does the spectrum of the radiation emitted take the form of sharp lines of definite wavelength?

(*c*) Find the energy associated with the level $n = 4$, and the wavelength of the line indicated by the transition P.

(*d*) Show, without any detailed calculations, that all lines resulting from transitions that end on the level $n = 1$ must be in the ultraviolet. (O)

23.3 The copper target of an X-ray tube contains manganese and certain other impurities, and the emission spectrum shows a strong K_α line due to copper at a wavelength of 1.541×10^{-10} m, and a weak K_α line of manganese at 2.102×10^{-10} m. Weak K_α lines of the other impurities are observed at 1.789, 1.936, and 2.289×10^{-10} m. Find the atomic numbers of the impurities. (Atomic number of manganese = 25, of copper = 29.) (C)

*23.4 An electron, charge $-e$, mass m_e, moves in a circular orbit around a nucleus of charge $+e$, which may be assumed stationary. Obtain an expression for the total energy, potential and kinetic, of the electron when the radius of the orbit is r.

Taking the angular momentum of the electron in this orbit to be equal to $h/2\pi$, obtain an expression for the energy E in terms of e, m_e, h, and ϵ_0. Hence find the energy change which occurs if the electron abruptly moves into a new orbit with angular momentum h/π. (C)

*23.5 Calculate the de Broglie wavelengths of each of the following particles. (The effect of relativity in each case may be neglected.)

(*a*) An electron, moving at 10^7 m/s.

(*b*) A thermal neutron, with energy 0.04 eV.

(*c*) A fast neutron, with energy 2 MeV.

(Rest mass of electron $= 9.1 \times 10^{-31}$ kg,
rest mass of neutron $= 1.67 \times 10^{-27}$ kg.)

Chapter 24

Radioactivity

24.1 The beginnings of nuclear physics

All the early discoveries in nuclear physics stemmed from Becquerel's discovery of radioactivity, and there are basically two reasons why this was so. One is because radioactivity is the result of nuclear instability, and in ejecting an α- or a β-particle the nucleus of one element changes to the nucleus of another, thus providing evidence about nuclear structure. The other is because α-particles have sufficient energy to allow them actually to penetrate the nucleii of the lighter elements, which lack sufficient positive charge to repel them effectively, and by this means nuclear changes can be produced.

One of the first results of the discovery of radioactivity was a search for further elements having the same property as uranium, and this led the Curies to isolate the new elements, polonium and radium, both of which, particularly radium, are much more intensely radioactive than uranium. The basis of their systematic experiments was an extensive application of large-scale chemical analysis. The sought-for element revealed its presence by its radioactivity, owing to the ionizing effect of α- and β-particles, and this enabled it to be tracked down, much in the way that a radioactive tracer is used today. The search began with an examination of the various compounds of uranium taken from natural sources in all parts of the world, and the discovery that samples of pitchblende, obtained from particular areas, were more radioactive than their uranium content demanded. Since all the other elements known to be in the ore were not radioactive at all, the inference was that it must also contain some previously undetected element. Since this had not previously been discovered, its amount must presumably be very small, and since its radioactivity was nevertheless effective, it must be extremely radioactive.

24.2 Cloud chambers

The methods used to detect α- and β-particles in all the early experiments were those which had been used by Becquerel, being photography, fluorescence, or the slow discharge of an electroscope connected to an ionization chamber. An exceptionally valuable method was added to these in 1911, when C. T. R. Wilson invented the cloud chamber. This has the unique advantage that it provides visual evidence of the actual tracks followed by ionizing particles; if a collision causes a particle to change direction, this is immediately revealed, and the angle of deflection can be measured. Wilson's cloud chamber had widespread application, and remained basically unaltered in design for more than thirty years.

The principle of the cloud chamber concerns the conditions in which a moisture laden atmosphere can deposit water drops when it suffers a fall in temperature. Very small liquid drops have an abnormally high vapour pressure owing to the effect of surface tension, and for this reason there is a tendency for saturated air to

become supersaturated when it cools down, unless there are dust particles present, on which the water drops can condense without initially being very small. If any charged ions happen to be present in such a supersaturated atmosphere, however, these have the same effect as dust particles would have had, and the water vapour uses the ions as nucleii for condensation, with the result that the location of the ions is revealed by the visible water droplets which form on them. Since α- and β-particles and the like travelling through air leave a trail of ions in their wake, their passage through the supersaturated air of the cloud chamber is marked by a trail of droplets. The reason that water condenses on ions in this way is the fact that the energy stored by a charged sphere is smaller when its radius is larger.

The expansion cloud chamber

In Wilson's type of cloud chamber, the fall in temperature which is needed to supersaturate the air is effected by means of a sudden expansion, the result being that the track of an ionizing particle which crosses the chamber immediately after the expansion is marked by a trail of water droplets, one on every ion which the particle has left in its wake. The space is initially rendered free of dust by arranging one or two preliminary expansions, and allowing the resulting mist to settle, taking down the dust with it. Some careful adjustment of the expansion is necessary, since too large a fall in temperature would result in condensation even in the absence of ions, whereas too small a fall would be only partially effective. When the particle tracks are to be used in a quantitative way, by measuring the angles of deflections, two photographs are taken, from different viewpoints, in order to obtain a stereoscopic effect.

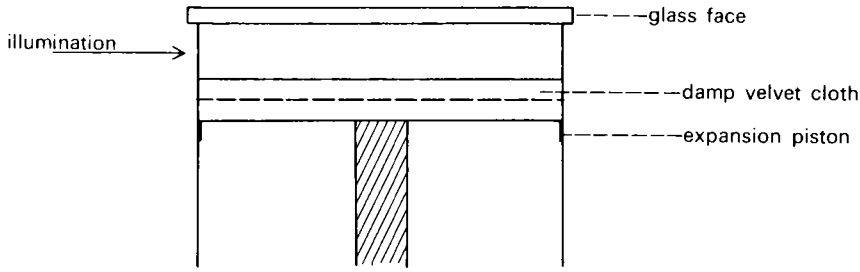


Fig. 24.1 The principle of Wilson's expansion cloud chamber (The dark velvet cloth serves as a background against which the illuminated track of water droplets is viewed. It is also kept damp, to saturate the air in the expansion chamber. The source of ionizing radiations may be placed inside the chamber, or a thin window in the vertical wall may be provided, to allow all but the least penetrating particles to enter the chamber from outside)

The diffusion cloud chamber

The expansion type of chamber has the disadvantage that it displays particle tracks only during a very short period after an expansion has been carried out. This has now been largely superseded by the diffusion chamber, in which a supersaturated space is maintained by having a permanent temperature gradient, instead of the temperature fall caused by an expansion. The vapour used in the diffusion chamber is usually alcohol, put there by having a felt rim round the perimeter of the top viewing window and saturating this with methylated spirits. The base of the chamber is a blackened metal plate which is cooled down by contact with dry ice – solid carbon dioxide – packed underneath it. In this way, a vertical temperature gradient is set up in the chamber. The warmer air at the upper surface is kept saturated with

vapour from the liquid in the felt, and the vapour molecules continuously supersaturate the cooler space below by diffusing into it. One difficulty which arises from the fact that this chamber works continuously must be overcome; although a track of droplets slowly disperses if left to itself, some means is required of clearing the field quickly if a fresh track is to be clearly observed. This can be done by applying a p.d. between the metal base plate and a second electrode; this produces an electric field to sweep the unwanted ions from the chamber.

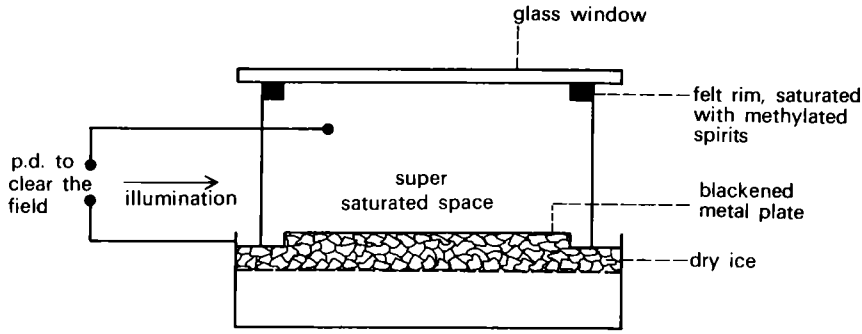


Fig. 24.2 A diffusion cloud chamber

A bubble chamber

A further development, used to study the effects of the extremely high energy particles produced in powerful accelerators, is the bubble chamber. This uses a liquid which is superheated above its normal boiling point, and bubbles of vapour form in the wake of particles passing through it. The liquid most often used is liquid hydrogen, because the purpose is to combine the tracking effect of the bubbles with the use of the hydrogen atoms as targets for the particles. The application of this instrument, however, is in a field which lies outside the scope of this book; only the particles studied in high energy physics would be capable of penetrating the liquid.

24.3 Identifying the radiations

When the nucleus of a naturally occurring radioactive isotope decays, it ejects either a helium nucleus or a negative electron, dependent on the identity of the particular isotope. In either case, a photon of very short wavelength X-radiation may be emitted as well. A uranium salt is usually emitting all three, because it contains traces of a variety of radioactive elements, the products of earlier decays. As soon as it was recognised that there were three distinct types of radiation, these were named α-, β-, and γ-rays; it is for this reason that energetic helium nucleii are now called α-particles, high speed electrons are β-particles, and very short wavelength X-rays are γ-rays.

The distinction between α- and β-rays was first made by Rutherford, on the basis of the relative ease with which α-rays could be stopped. Whereas these fail to penetrate a few thicknesses of paper, and have a range in air of only a few centimetres, the β-rays can penetrate paper quite easily, and it requires a metal sheet a millimetre or so thick to stop them completely. We now know that the shorter range of α-rays is due to the much heavier ionization which they cause; this does not deflect them from their straight path because an α-particle is so massive compared with the electrons which it usually encounters, but it does use up their

energy and so stops them. The difference in behaviour between α- and β-particles in their passage through air is well illustrated by cloud chamber photographs. Whereas the α-particle trails are continuous lines of droplets, owing to the large number of ions, those left by the β-particles are tortuous and the water droplets are comparatively sparse, showing the results of individual ionizing events.

Fig. 24.3 The tracks of α and β particles in a cloud chamber (a) α-particles (b) β-particles

The next step was made by Becquerel, who found that β-rays could be deflected quite easily by a transverse magnetic field (an effect which is due to their negative charge and very small mass). Gamma rays were then identified by Villard as being rays with a very long range in air, which a magnetic field failed to deflect. Using a photographic plate to detect them, he showed that they could penetrate a thin sheet of lead. Their identity as short wavelength X-rays was eventually established by the technique of X-ray diffraction.

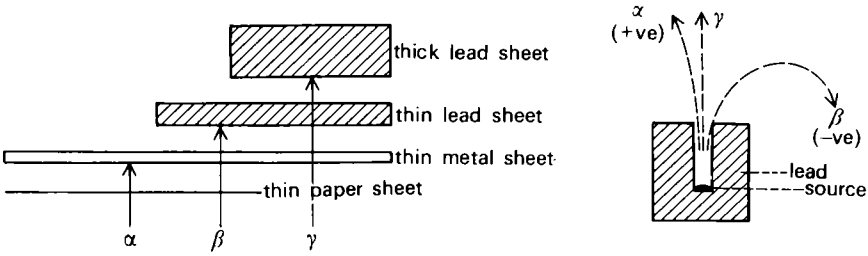


Fig. 24.4 Diagrams illustrating two simple distinguishing features of α, β, and γ rays (a) Absorption (b) Deflection in a transverse magnetic field. (The deflection of the α-particles has been greatly exaggerated by comparison with that of the β-particles)

These two simple distinguishing features of the three types of radiation are illustrated in the next diagram, but the effect of the magnetic field has deliberately been drawn out of proportion, for purposes of the illustration; the momentum of an α-particle is so great that its deflection by even a very strong magnetic field is only slight, and it has been exaggerated in the diagram.

24.4 β-particles

Becquerel's experiment is illustrated in Fig. 24.5. He put a small quantity of a uranium compound at the bottom of a cavity in a block of lead, and placed this on top of a photographic plate in a vacuum. In this way he obtained a collimated beam of β-rays, which he subjected to the influence of a transverse magnetic field. After exposing the plate for some time, he developed it, and found that the field

had bent the path of the β-rays, causing them to produce a diffuse patch on the plate. This diffuseness was due to the fact that, as they leave the radioactive source, β-particles have a spread of energies, and the more energetic particles are deviated less by the field, a conclusion which he verified by repeating the experiment after placing aluminium foil above the photographic plate to stop the slower particles.

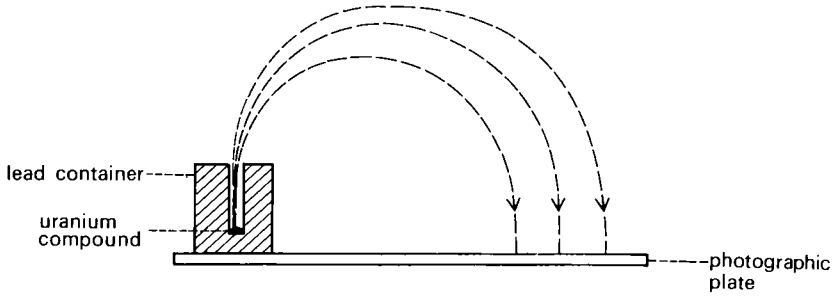


Fig. 24.5 The deflection of β-particles in a magnetic field

The identity of the β-rays as being high speed electrons was finally established in 1902 by Kauffmann, who measured their ratio of charge to mass, e/m, by a method of simultaneous electric and magnetic deflection.

24.5 α-particles

Alpha-rays were finally identified by Rutherford as being doubly charged atoms of helium, and therefore more appropriately named α-particles. The fact that they were helium nucleii, being the atoms stripped of their only two electrons, naturally followed when the structure of the atom had been worked out. Unlike β-particles, all the α-particles from a single radioactive element are ejected with the same velocity, and consequently have the same range in air, as illustrated in the cloud chamber diagram, Fig. 24.3(*a*).

Rutherford first showed that α-rays could be deflected by a powerful magnetic field and that they carried positive charge. He next devised a method to measure their charge/mass ratio, and found this to be the same for particles ejected from different radioactive sources, and this value proved to be about half the charge/mass ratio determined by electrolysis for the hydrogen ion H^+. Thus he discovered in the case of α-particles what Thomson and others had earlier discovered for electrons, the significant fact that the same α-particles were ejected from different elements. Furthermore, in view of their ratio of charge to mass, they were probably doubly charged helium atoms, since the helium atom has four times the mass of hydrogen.

In the next experiment, a count was made of the separate α-particles ejected from a radioactive source in a measured solid angle in a measured time, and the total charge which was produced by collecting them was also measured. Thus the charge carried by a single particle could be deduced, and from this, also, the mass of the particle, since the ratio of charge to mass had already been determined. This particular experiment was carried out by Rutherford and Geiger in 1908, and the instrument used by them to detect the arrival of each α-particle was a specially sensitive form of ionization chamber, the earliest type of Geiger tube. The same experiment was carried out independently in Germany at the same date by Regener, who counted the α-particles by observing their separate arrivals on a fluorescent

screen. By 1908, a sufficiently good estimate of the size of the electronic charge e was available to make it evident, as a result of these experiments, that the particle's charge was equal to $2e$.

In the culminating experiment, carried out by Rutherford and Royds in 1909, α-particles from a radioactive source were allowed to accumulate in a tube until it was eventually possible to demonstrate that helium had been produced.

It would occupy too much space in this book to describe all these experiments in detail, and the accounts which follow are restricted to the first of them and the last.

24.6 The deflection of α-particles in a magnetic field

The basic arrangement used by Rutherford for this experiment is shown in Fig. 24.6. The α-particles were allowed to enter the base of a gold leaf electroscope

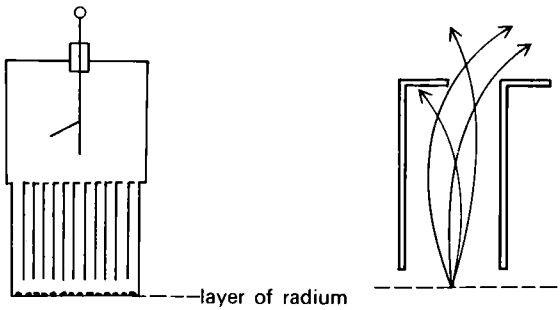


Fig. 24.6 Apparatus devised by Rutherford to show the deflection of α-particles by a magnetic field (a) (b) Showing the effect of reversing the field

through narrow channels between a series of parallel vertical metal plates about 1 mm apart, their arrival inside the case of the previously charged electroscope being registered by a slow fall in the leaf, due to ionization occurring in the air. When a strong magnetic field perpendicular to the plane of the diagram was applied, this deflected the path of the α-particles sufficiently to direct a proportion of them on to the vertical metal plates, thus preventing them from reaching the electroscope, with the result that the rate of fall of the leaf was reduced. By putting a lip on one side of each of the plates at the emergent end, Rutherford increased the collecting power of each plate with respect to deflection in one direction, without affecting the collecting power if the deflection was reversed. The effect of this is indicated in the diagram. By this means, he was able to deduce the direction of the deflection in relation to that of the magnetic field, and so to establish that the charge on the particles was positive.

24.7 To establish the identity of α-particles

The source of the α-particles in this experiment was a small quantity of the radioactive gas radon, which is emitted from radium as the product of its decay. The gas was compressed into the fine-walled capillary tube A, which was then inserted into the evacuated tube T (Fig. 24.7). The α-particles were ejected from the gas contained in the capillary tube, and the object of the experiment was to collect them in the outer tube T and to test whether helium gas had accumulated there as a result. Yet α-particles can be stopped by a few sheets of thin paper, so that it

was evidently essential for the wall of the capillary tube to be exceptionally thin, and it was, in fact, less than 10^{-2} mm in thickness.

When the α-particles had been allowed to accumulate as gas in the tube T over a period of several days, this small amount of gas was compressed by mercury into the top of the tube, which was constructed in the form of a discharge tube. This was found to give the spectrum lines of helium when an electric discharge was passed through it, although there had been no trace of a helium spectrum from it at the outset of the experiment. As a further check, the experiment was repeated putting helium in the thin-walled capillary tube in place of the radon, and no helium lines were obtained from the discharge tube in this case, even after a period of eight days. The spectrum must, therefore, have been due to α-particles, which alone were capable of penetrating the thin glass to reach the discharge tube, where they picked up the electrons needed to convert them to neutral helium atoms.

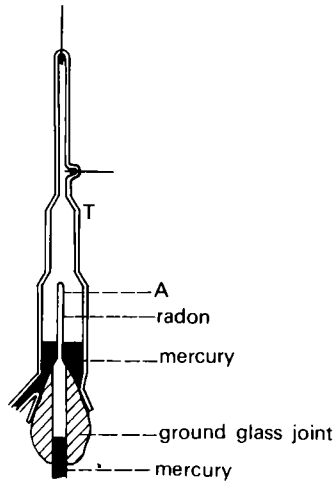


Fig. 24.7 The apparatus used by Rutherford and Royds to establish that α-particles are charged atoms of helium

24.8 Radioactive transformations

When a radioactive atom ejects an α- or β- particle, this comes out of the nucleus, which changes, in the process, to the nucleus of another element. This statement is easy to accept today, but this was not so early in the century, when the existence of the nucleus was as yet unknown, and for one element to change to another was a concept which belonged to the alchemists. In giving an account of the changes involved in radioactivity, we have largely to neglect an appreciation of history, since what was at the time an intricate puzzle has now become a relatively simple one.

The heaviest element which exists naturally, except in minute traces, is uranium, of which the most abundant isotope is uranium 238, ${}^{238}_{92}\mathrm{U}$. That is to say that the atomic number Z of uranium is 92; its nucleus contains 92 protons, giving it electrical charge equal to $+92e$, where the electronic charge is equal to $-e$; the neutral atom has 92 electrons outside its nucleus. The nucleon number, or mass number A, of this particular isotope is 238, which means that altogether its nucleus contains 238 nucleons, made up of the 92 protons already referred to, and 146 neutrons, i.e.

$$ {}^{238}_{92}\mathrm{U} = 92\,{}^{1}_{1}\mathrm{H} + 146\,{}^{1}_{0}\mathrm{n}, $$

an equation which we have written to refer to the content of the nucleus, without any implication being intended that it ever splits up into all these parts.

To try to keep the history in proportion, we have to remember that Rutherford discovered the nucleus in about 1910, and showed in 1919 that protons could be knocked out of other nucleii, of nitrogen in the first place. The neutron was not identified until 1933.

Most nucleii appear to have about the same density, and this is a very high density indeed; the relation between nuclear radius and atomic mass number A is approximately given by the empirical equation

$$r = A^{\frac{1}{3}} r_1, \tag{24.1}$$

where $r_1 = 1.2 \times 10^{-15}$ m.

Using this equation, the radius of the uranium-238 nucleus is about 7.5×10^{-15} m. The mutual electric repulsion between the protons in the nucleus must clearly be very great; yet the nucleus does not fly apart, from which we deduce that the nuclear force, which holds neutrons and protons together and is not electrical in origin, is even greater. This is an extremely short range force, however, and does not extend its influence beyond the boundary of the nucleus.

Despite the cohesive effect of the nuclear force, uranium-238 is an unstable nucleus, as are all the other isotopes of uranium and all the isotopes of all elements with atomic number higher than 82, which belongs to lead, with the exception of a single isotope of bismuth, ${}^{209}_{83}\mathrm{Bi}$. In a specimen containing uranium, whether this is present as the uncombined element, or in a compound, an individual uranium nucleus will at some unpredictable instant reveal its instability by ejecting a helium nucleus, ${}^{4}_{2}\mathrm{He}$, endowed with considerable energy (4.2 MeV, in this particular case), and this is an α-particle. The equation representing the transformation is

$${}^{238}_{92}\mathrm{U} \rightarrow {}^{234}_{90}\mathrm{Th} + {}^{4}_{2}\mathrm{He}.$$

At the same time, further energy is released in the form of a photon of γ-radiation, i.e. a very short wavelength train of X-rays, and this energy comes from the nucleus, also.

The instant at which a uranium nucleus will break down in this way is unpredictable and uncontrollable, and depends in no way on any chemical or physical factors, such as chemical combination or temperature. The overall effect in a mass containing many uranium atoms is a statistical one, and the rate at which nucleii break down, $-\mathrm{d}N/\mathrm{d}t$, is, on average, proportional to the number N of nucleii which, at the relevant instant, remain, i.e.

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -\lambda N, \tag{24.2}$$

where λ is called the decay constant of the isotope. Thus, the decay constant represents the probability that an atom will decay in the next unit of time.

Integrating equation 24.2, we obtain

$$N = N_0 \exp(-\lambda t), \tag{24.3}$$

where N_0 is the number of undecayed nucleii present at time $t = 0$. Putting $N = \frac{1}{2}N_0$ in this equation, we have

$$\tfrac{1}{2}N_0 = N_0 \exp(-\lambda T),$$

whence

$$T = \frac{1}{\lambda} \log_e 2. \tag{24.4}$$

This time T is called the half-life of the isotope, being the time which elapses

before half the nucleii originally present will have decayed; this is quite independent of the instant chosen as a starting point.

The half-life of uranium-238 is equal to 4.5×10^9 years, which accounts for the fact that this isotope is still found today in the earth's crust.

It is stressed that equation 24.2 represents an average behaviour; a specimen of uranium compound does not eject particles in a regular manner, but unpredictably.

24.9 A radioactive series

The product of the radioactive decay of uranium-238 is an isotope of thorium, ${}^{234}_{90}\text{Th}$, and this, too, is radioactive. Its nucleus ejects a high energy negative electron, i.e. a β-particle, together with a photon of γ-radiation, the equation being

$$ {}^{234}_{90}\text{Th} \rightarrow {}^{234}_{91}\text{Pa} + {}^{0}_{-1}\text{e}. $$

The half-life is 24 days.

Yet again, the protoactinium isotope, ${}^{234}_{91}\text{Pa}$, which is the product in this case, is radioactive, ejecting a further β-particle, with half-life of 1.2 minutes, and so the sequence of radioactive disintegrations continues, until it ends with the formation of the stable isotope of lead, ${}^{206}_{82}\text{Pb}$.

The complete uranium series of transformations is shown in table 3. Reference to the half-lives recorded in this table shows that it is the extremely slow decay of the parent element ${}^{238}_{92}\text{U}$ which accounts for the continued existence in uranium ores of the other radioactive isotopes in the series, despite their more frequent disintegrations. Among these are radium and polonium, the two radioactive elements which were first extracted from uranium ore by the Curies.

Table 3. *The uranium series*

Element	Isotope	Emission	Half-life
uranium	${}^{238}_{92}\text{U}$	α, γ	4.5×10^9 a
thorium	${}^{234}_{90}\text{Th}$	β, γ	24 d
protactinium	${}^{234}_{91}\text{Pa}$	β, γ	1.2 min
uranium	${}^{234}_{92}\text{U}$	α, γ	2.5×10^5 a
thorium	${}^{230}_{90}\text{Th}$	α, γ	8.0×10^4 a
radium	${}^{226}_{88}\text{Ra}$	α, γ	1620 a
emanation (radon)	${}^{222}_{86}\text{Em}$	α	3.8 d
*polonium	${}^{218}_{84}\text{Po}$	α	3.1 min
lead	${}^{214}_{82}\text{Pb}$	β, γ	27 min
*bismuth	${}^{214}_{83}\text{Bi}$	β, γ	20 min
polonium	${}^{214}_{84}\text{Po}$	α	1.6×10^{-4} s
lead	${}^{210}_{82}\text{Pb}$	β, γ	19 a
*bismuth	${}^{210}_{83}\text{Bi}$	β	5.0 d
polonium	${}^{210}_{84}\text{Po}$	α	138 d
lead	${}^{206}_{82}\text{Pb}$	stable	

* A small proportion of the isotope marked in this way undergoes a different form of decay.

When uranium ore has been undisturbed and equilibrium has been set up, this means that, for any element in the radioactive series except the parent element, uranium-238, and the end product, lead-206, the nucleii are disintegrating at the same rate as fresh nucleii are being formed, and we can write

$$ \frac{dN_1}{dt} = \frac{dN_2}{dt} = \frac{dN_3}{dt}, \text{ etc.} \qquad (24.5) $$

Whence, from equation 24.2,

$$\lambda_1 N_1 = \lambda_2 N_2 = \lambda_3 N_3, \text{ etc.,} \tag{24.6}$$

and, from equation 24.4,

$$\frac{N_1}{T_1} = \frac{N_2}{T_2} = \frac{N_3}{T_3}, \text{ etc.,} \tag{24.7}$$

i.e. the number of atoms of any isotope present in a given amount of ore is directly proportional to the half-life period of this isotope.

The uranium compounds which were studied in the early experiments were found, as we have seen, to emit all three types of radiation, but this was because these materials were to some extent mixtures, including in them the products of radioactive transformations. When a single uranium-238 nucleus breaks down, it ejects a single α-particle, and a pure specimen of uranium-238 would not emit β-particles at all. The direct result of its decay, however, is to product thorium-234, a single nucleus of which ejects a β-particle, so that the uranium specimen would in due course be found to emit β-particles after all. (Gamma radiation, it will be noticed, may accompany either type of emission.) This is an example of the kind of complication which had to be resolved by the early experimenters, in developing a table such as table 3, and one of the first steps was made by Crookes in 1900. By chemical means he separated out a precipitate from a solution of the original uranium salt which appeared to take all the radioactivity with it, leaving the uranium with none. This was because, by using photography for detection, Crookes was observing only the β-radiation, the α-particles which were in fact emitted by the uranium being absorbed and not noticeably affecting the photographic plate. What he had done was to separate out the thorium-234, which was responsible for the β-radiation, and he found that over a period of two or three months the uranium regained its β-activity while the precipitate lost it; this was due, of course, to the fact that the uranium had accumulated a fresh amount of thorium-234, whereas that contained in the precipitate had decayed.

A further feature of a radioactive transformation which is evident in the series shown in the table is the result of conservation of charge and conservation of nucleons. When an α-particle, ${}^{4}_{2}\text{He}$, is ejected from a nucleus, it takes with it four nucleons, and consequently reduces the nucleon number A by 4. Because the α-particle is doubly charged, two of these nucleons must be protons, the other two being neutrons, and so the atomic number Z of the nucleus is reduced by 2, i.e. the product of the decay is two places lower down (nearer to hydrogen) in the periodic table.

The ejection of a β-particle is rather more puzzling, as it may seem remarkable that a nucleus should eject an electron when it is said not to contain one, but to consist entirely of protons and neutrons. It is a conclusion of wave mechanics, however, that an electron would take up too much space inside the nucleus for its existence there to be possible. Ejection of a β-particle is assumed to be the result of a transformation of a neutron within the nucleus; the neutron changes to a proton, which remains in the nucleus, and into an energetic electron, which is ejected from it:

$${}^{1}_{0}\text{n} \rightarrow {}^{1}_{1}\text{H} + {}^{0}_{-1}\text{e}.$$

When outside the nucleus, a neutron is, in fact, a radioactive particle, undergoing the above transformation with a half-life of 13 minutes, although it is usually stable when inside the nucleus. (To be more correct, this explanation should include reference to a further particle known as an antineutrino, but at this point the complication is unwelcome, and we can afford to neglect it.)

We can now say that the ejection of a β-particle from a nucleus reduces the number of neutrons in it by one, and increases the number of protons by one, with the result that the nucleon number A is unchanged, while the atomic number Z is increased by one, i.e. the product of the decay is one place higher in the periodic table.

To sum up, each time that a radioactive nucleus ejects a particle, it becomes the nucleus of a different element. Loss of an α-particle, in reducing the nuclear charge by two units, moves it two places down the periodic table, while loss of a β-particle, which increases the nuclear charge by one unit, moves it up one place. This explanation, which was first suggested by Soddy, involves accepting that each separate element does not possess one unique value for its atomic mass, because it leads to the conclusion that nucleii exist having the same atomic number, but with different nucleon numbers and hence different atomic masses. It was for this reason that Soddy suggested that there were such things as isotopes, the different elements being necessarily different only in their atomic numbers. This conclusion was presently verified by Aston, who designed the first accurate mass spectrograph, and found that specimens of chemically pure elements often contained atoms which differed from each other in mass. This had already been suggested by the less accurate experiments of Thomson in positive ray analysis.

Since ejection of an α-particle reduces the nucleon number A by 4, and the ejection of a β-particle does not change it at all, it follows that the isotopes in the uranium series given in table 3 can have only the values of A given by $A = 238$, 234, 230, 226, and so on. Values lying between these, such as 237, 236, 235, are missed out. There are, in fact, three further radioactive series, additional to the uranium series which we have been discussing, and these account between them for the missing numbers.

The actinium series has uranium-235, ${}^{235}_{92}\mathrm{U}$, as the parent element, with half-life equal to 7.1×10^8 years, and the stable end product is lead-207, ${}^{207}_{82}\mathrm{Pb}$.

The thorium series starts with thorium-232, ${}^{232}_{90}\mathrm{Th}$, with half-life 1.4×10^{10} years, and ends with lead-208.

The neptunium series, which is the fourth series, does not occur in nature, because the longest half-life included in it, which belongs to neptunium-237, ${}^{237}_{93}\mathrm{Np}$, is 2×10^6 years, and this is not long enough to have ensured survival. (The age of the earth's crust is about 4.5×10^9 years.) Neptunium is a transuranic element, i.e. its atomic number places it beyond uranium in the periodic table, and it does not occur naturally, except for slight traces. It is produced in a nuclear reactor, however, as the result of neutron-induced transformations.

24.10 Artificially induced radioactivity

With one or two isolated exceptions, the naturally occurring isotopes of elements which have a lower atomic number than lead are stable. A radioactive isotope of almost any element can, however, be produced artificially, and this has either an excess or a deficiency of neutrons in its nucleus. The usual way to obtain the isotope is in a nuclear reactor, either as a product of the nuclear fission by which the reactor operates, or by exposing substances to irradiation by neutrons in the reactor. An alternative method is to bombard the substance with charged nucleii such as protons which have been speeded up in a relatively small accelerator such as a cyclotron.

The first example of artificially induced radioactivity was discovered by the French physicist Joliot and his wife, Madame Joliot-Curie. They bombarded aluminium with α-particles, and found that a product of this bombadrment was radioactive, ejecting positrons, i.e. positively charged electrons, ${}^{0}_{1}\mathrm{e}$, and continuing

to do so, with an exponential decay, after the α-particle bombardment had ceased, the half-life being about 3 minutes. The Joliots worked out that the reaction had been

$$^{27}_{13}\text{Al} + ^{4}_{2}\text{He} \rightarrow ^{30}_{15}\text{P} + ^{1}_{0}\text{n}.$$

The phosphorus-30 is the radioactive isotope, ejecting a positron by the reaction

$$^{30}_{15}\text{P} \rightarrow ^{30}_{14}\text{Si} + ^{0}_{1}\text{e}.$$

To verify these equations, they actually separated out some of the radioactive phosphorus by chemical means.

Positron emission does not occur among the heavy isotopes in a natural radioactive series. Such a nucleus either contains too many nucleons altogether for stability, in which case it ejects an α-particle, or else it contains an excess of neutrons, with the result that a neutron changes to a proton inside the nucleus by ejecting a negative β-particle. In a substance such as the radioactive phosphorus, however, the nucleus apparently contains too few neutrons for stability, and it is a proton inside the nucleus which changes to a neutron, with the ejection of a positive β-particle:

$$^{1}_{1}\text{H} \rightarrow ^{1}_{0}\text{n} + ^{0}_{1}\text{e}.$$

The small number of radioactive isotopes of comparatively low atomic number which occur naturally do not owe their continued existence to the radioactive decay of a long-lived parent element, but to the fact that they are continuously being produced by cosmic bombardment. The most well known of these is radioactive carbon, $^{14}_{6}\text{C}$, which ejects negative β-particles, with a half-life of 5600 years. This is produced in the atmosphere from nitrogen nucleii, owing to neutrons which themselves are formed by cosmic ray bombardment.

$$^{14}_{7}\text{N} + ^{1}_{0}\text{n} \rightarrow ^{14}_{6}\text{C} + ^{1}_{1}\text{H}.$$

Additionally, there are now some radioactive atoms in the atmosphere, such as strontium-90, $^{90}_{38}\text{Sr}$, which are present as the result of nuclear explosions, and have sufficiently long life to persist. The half-life of strontium-90 is 28 years.

24.11 Some applications of radioactive isotopes

Both naturally occurring and artificially produced radioactive isotopes have many applications, a selection from which is described below.

Carbon-14 dating

We can assume that by this time the amount of carbon-14 present in the atmosphere is constant, since equilibrium between the rate at which it is being formed from nitrogen nucleii and the rate at which it is decaying must have been reached. The result of this equilibrium is that the proportion of radioactive carbon, $^{14}_{6}\text{C}$, to stable carbon, $^{12}_{6}\text{C}$, in the atmosphere, e.g. in atmospheric carbon dioxide, is constant. This ratio is a very small one, so that atmospheric carbon is only slightly radioactive, one gram of carbon emitting about 12 β-particles per minute. This proportion remains the same in the carbon of a living plant or animal, since this is exchanging carbon with the atmosphere so long as it is alive. So soon as the plant or animal dies, however, the exchange ceases, and the radioactive content of the dead material is no longer replenished, and so becomes subject to exponential decay, the activity being halved in the period of 5600 years. It follows that any vegetable or animal material which has been dead for a long period, but for less than about

12000 years, can be dated by measuring the activity of the $^{14}_{6}C$ which remains in it. Because the activity is very small in any case, this requires very careful procedures.

Measuring geological times

If a rock contains uranium, the age may be estimated on account of radioactivity. Since it may be assumed that equilibrium conditions have by now been reached among the successive radioactive isotopes present in the rock, the only net change which is still proceeding can be taken to be the very slow decay of the parent element, $^{238}_{92}U$, resulting in the equally slow growth of the stable end product, $^{206}_{82}Pb$. If we are entitled to make the further assumption that all the lead now present in the rock has been produced over millions of years by this radioactive decay of uranium, an estimate of the age of the rock can be made, by measuring the relative proportions of lead and uranium in the rock.

A check of this estimate is possible on account of the fact that the range of α-particles in a solid is extremely small, and when they get their extra-nuclear electrons back they become helium atoms. Consequently, such radioactive rock has helium accumulated in it, which can be driven out by heating, and measured. This represents a count of all the α-particles which have been emitted in the specimen, in the process of decay from uranium to lead, and so, again, the total time involved can be estimated.

Radioactive tracers

A very extensive application of radioactive isotopes, in medicine, industry, and the chemistry and physics laboratories, is to adopt the technique which the Curies used in the first place to track down polonium and radium. By using a radiation detector, even the minutest amount of a radioactive material can be traced, yet the chemistry of a radioactive isotope is identical with that of the stable atom. It follows that a very small quantity of a radioactive isotope added to the stable atoms in a chosen element takes its part in all chemical and physical activities, and so, wherever this element goes, the radioactivity betrays it. If a short-lived isotope is chosen, this has the obvious advantage that the radioactivity persists long enough to be useful, but can be relied upon to have vanished when it might otherwise have become a hazard. Such radioactive tracers may be used, for example, to trace leaks in buried water pipes, or the spread of waste products put into a river. An example in the physics laboratory is their use in friction experiments, to see where traces of one solid adhere to another when two surfaces slide across each other.

Radiation therapy

The dangers from exposure to γ-radiation will be discussed in the next paragraph. Like the radiation from an X-ray tube, it has beneficial application also, in the destruction of malignant cells, the radioactive isotope having particular advantages in the treatment of deep seated tumours.

24.12 Radiation dangers

The danger from radioactive materials lies chiefly in the γ-radiation, because of its great penetration. Being short wavelength X-rays, γ-rays have the same damaging effects. With the radioactive material, however, there is the additional hazard that the source cannot be switched off like the X-ray tube, and, in addition to this, a very small amount can be dangerous, and it could get spread about. For such reasons, only the weakest of sources are allowed in a school laboratory, and these should be stored in suitable containers, made of lead in the case of a γ-ray source, and a

check should be kept on their whereabouts. Short-lived isotopes, of course, are not such a menace as the others, because their activity does not persist.

The dangers incurred by exposure to radiation, whether this comes from a radioactive source or from an X-ray tube, must never be underestimated, and are rendered all the greater by the fact that the effects are not immediately apparent. A strong exposure results in the destruction of living cells, and causes deep seated burns to develop after a comparatively short period of time. A greater hazard, however, lies in the very long term effect of repeated exposures, even when the individual doses are relatively small, because this effect is cumulative, and causes changes in living cells, giving rise to diseases such as leukaemia; most serious of all is the danger of changes being produced in genes, thus affecting future generations. The human body is, in any case, unavoidably subject to a low intensity exposure due to cosmic radiation, artificially supplemented at the present time as the result of nuclear test explosions. The term 'safe', when applied to an artificial dose, is therefore only a relative one, and is used to describe an exposure which is slight compared with the unavoidable background radiation.

The distinction between the different types of radiation with respect to their dangerous nature is basically this. The damage which they cause is related to their ionizing effect, and so we might expect it to be greatest from α-particles. To do the damage, however, the particles first have to reach the vulnerable places, and both α- and β-particles are too easily stopped to be dangerous, provided that all traces of their source are kept from entering the body, whether by reaching the lungs or the stomach. For this reason, the sources must not be handled, and should be kept at a distance. The only defence against γ-radiation, on the other hand, is to avoid exposure to it, and the source is kept in its lead container.

Units

It must first be said that neither the röntgen nor the curie, defined below, is an accepted SI unit, and consequently both may eventually go out of general use.

The röntgen (R)

This measures the intensity of X-rays and γ-rays, which is judged in terms of the ionizing effect. An exposure dose of 1 R produces 8.38×10^{-3} J of energy by ionization in 1 kg of air at N.T.P. (The reason for this peculiar figure is that the unit was originally defined in terms of c.g.s. units of charge carried by the ions.) Its value in SI units is 1 R = 2.58×10^{-4} C/kg.

The curie

The curie is a unit which measures the activity of a radioactive source. The disintegration rate of a source is equal to 1 Ci if 3.7×10^{10} nucleii are, on average, disintegrating in every second. (As with the röntgen, the unit was originally defined differently, as being the disintegration rate of 1 g of radium.)

1 g of radium would be a very powerful and dangerous radioactive source. The activity of any source which is used in a school laboratory is below 10 μCi.

The SI unit of radioactivity measures, simply, the average number of disintegrations in unit time, and is therefore second^{-1}. 1 Ci = 3.7×10^{10} s^{-1}.

24.13 Radiation detectors

All types of radiation detector basically depend on one or other of two effects, either the ionization or the excitation of atoms which lie in the path of the radiation, coupled with a means of detecting or measuring the ionization or excitation produced. However, the detector will fail to respond if the radiation is absorbed

at the wrong place, and this may lead to some difference in detail between a device designed for α-particles and one which responds to γ-rays. Although the ionization produced by α-particles is intense, they may all be absorbed, for this very reason, without entering the detector. Conversely, in other cases, it is γ-rays which escape detection, by passing right through the instrument without having appreciable effect. The behaviour of a cloud chamber is an example of these two extremes. The tracks of α-particles are very clear (Fig. 24.3), but the source must be placed inside the chamber because the particles cannot penetrate the walls. On the other hand, γ-rays from an external source can pass right through the chamber; the only traces left by them are the tracks of high energy electrons which their occasional ionizing action produces, (Fig. 24.8), and low intensity γ-radiation escapes detection.

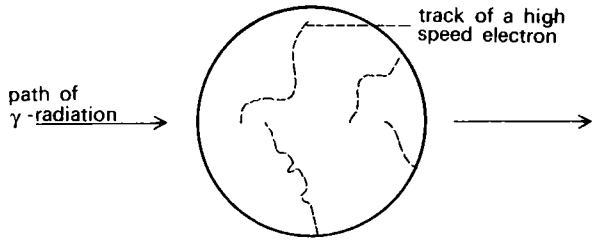


Fig. 24.8 The trace left in a cloud chamber by a narrow beam of γ-rays

A further distinction which can be made between types of radiation detector depends on the measuring aspect. In some cases the separate particles are counted as they arrive, and this applies also to γ-rays, which are counted as photons; (a proportional counter compares the energy of the separate particles too, but this type will not be described here). In other cases, the overall effect is integrated over a period of time as, for example, when the total ionizing effect is measured.

24.14 An ionization chamber

The principle of the ionization chamber has been referred to already (section 21.2). The ionization caused by the passage of radiations through air renders it slightly conducting, and the intensity of the radiation can be measured by this effect. In many of the early experiments, the method used was to observe the slow discharge of an electroscope due to ionization of the air contained inside the case of the instrument itself. More usually, however, the current detector is separate from the ionization chamber, which consists simply of two electrodes, separated by an air space, and perhaps enclosed in a thin-walled container; the electrodes are then connected in series with a voltage source and a highly sensitive current measuring device, which registers a current so long as ionizing radiation is entering the chamber. If the value of the p.d. is sufficiently high, all the ions formed are carried across to the electrodes before they can recombine, and the current reaches its saturation value (Fig. 24.9), in which case it represents a measure of the total ionization produced; even so, the current may be as small as, say, 10^{-12} A.

One type of ionization chamber is specifically designed as a fitting to a d.c. amplifier or to a pulse electrometer, which is used as the current detector, and one application of this will be described later in this chapter. Such an arrangement is not sufficiently sensitive to detect the sparse ionization caused by γ-radiation unless the source is so powerful that it would not be permitted in a school laboratory. α- and β-radiation can be detected, however,

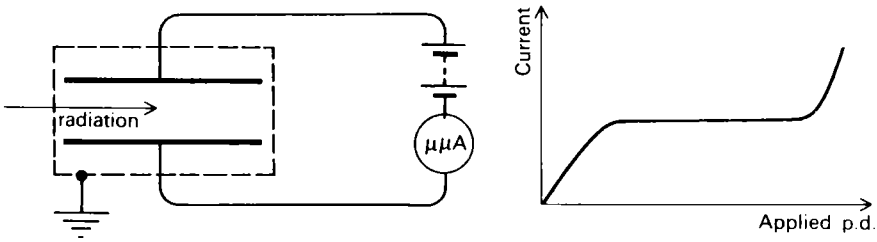


Fig. 24.9 The circuit of an ionization chamber

provided that the α-particles are allowed free access to the chamber, or the source is placed inside it.

24.15 The spark counter

If the electric field in an ionization chamber is made sufficiently powerful, the effect of ionizing radiation is greatly increased by the process known as gas-multiplication. In this process, the ions originally produced by the incident radiation gain sufficient energy in the field to release further ions from the gas by collision, so that a small effect builds up to a much larger one. The instrument can now be used as a counter, responding to the arrivals of individual particles.

Fig. 24.10 A spark counter

A very simple device which works on this principle is the spark counter. This consists simply of a fine wire stretched out between two insulating supports, so that it lies a short distance above a metal base plate. A p.d. from an E.H.T. unit is applied between wire and plate, with the result that a slight brush discharge occurs in the vicinity of the wire, where the electric field is particularly strong. If a source of α-particles is brought near, each separate particle which reaches the counter causes an audible spark to pass. This is effective with α-particles, because no absorbing material other than the air lies in their path, but it is not sufficiently sensitive to detect β- or γ-radiations.

Alternatively to watching or listening to the sparks, a high value resistor may be connected in series with the spark counter; the pulse of current which accompanies each spark in this case produces a voltage pulse across the resistor, and this can be recorded electronically by a device such as a scaler.

24.16 The Geiger-Müller counter

The Geiger-Müller counter, more briefly referred to as the Geiger tube, is a more sophisticated and sensitive instrument than the spark counter, and operates on the same principle of gas multiplication. It represents a development from the detector which was originally devised by Rutherford and Geiger for counting α-particles (section 24.5).

As with the spark counter, one electrode of the Geiger tube takes the form of a very fine tungsten wire, the other electrode in this case being a metal cylinder surrounding the central wire. Thus, the electric field strength in the immediate neighbourhood of the fine wire, where the lines of force come very close together, is particularly high, and it is here that the gas multiplication occurs. The tube contains an inert gas, argon or neon, at a reduced pressure of about 100 mmHg. When in circuit, the fine wire is used as the anode, with the result that the ions which are attracted to it are electrons.

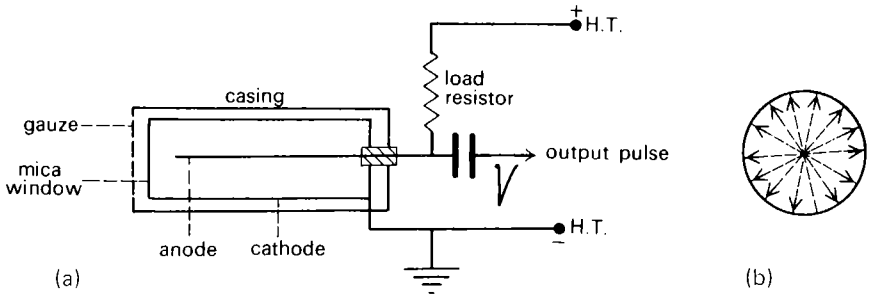


Fig. 24.11 A Geiger-Müller tube (a) The tube circuit (b) The electric field

When an ionizing particle enters the tube, the electrons which reach the neighbourhood of the anode as the result of the initial ionization are sufficiently accelerated in the powerful field to produce an avalanche of ions, which runs along the whole length of the wire, and an output voltage pulse of considerable size is produced across the load resistor.

For the instrument to be effective as a counter, a very rapid response is essential, to allow a rapid succession of pulses to be separately recorded. This is the reason that the tungsten wire is essentially connected to be the anode, since the electrons move towards it much more rapidly than the more massive positive ions move towards the cathode.

A description of the essentials is not yet complete, however, owing to the behaviour of the positive ions, including those produced by the avalanche along the length of anode. These ions, travelling comparatively slowly towards the cathode, gain energy in the process. If they were allowed free access to the cathode, they would strike it with sufficient energy to release electrons from it, and these would be attracted to the anode, to cause a fresh avalanche, and so a succession of discharges would ensue, all as the result of a single ionizing event. One method commonly adopted to prevent this occurrence is to add to the gas in the tube a small proportion of a suitable foreign vapour or gas, which acts as a quenching agent, taking up the energy of the positive ions before they reach the cathode, so that the secondary emission of electrons does not occur. Tubes of this type are said to be self quenching. One type contains argon with ethyl alcohol as the quenching agent; this requires a p.d. of 1 – 2 kV to operate it. Another contains a mixture of argon and neon and is quenched by a halogen, chlorine or bromine. This has the advantages that it operates at a smaller p.d. of a few hundred volts and has a long operating life; its maximum count rate, however, does not exceed a few thousand pulses in each minute, owing to the comparatively large recovery time required between successive pulses.

Unlike the spark counter, the simplest type of Geiger tube is best suited for counting β-particles. This is because, although sensitive, it is closed in, to contain

the low pressure gas. Beta particles are admitted through a thin window of mica or glass which closes the end of the tube, but an α-particle detector requires to be specially constructed, having an exceptionally thin window. The ordinary type of tube will also respond to a small proportion of incident γ-ray photons, which cause a discharge by releasing electrons from the walls, but the remainder of the γ-radiation passes straight through.

The characteristic behaviour of a Geiger counter

We suppose that a Geiger tube is set up, connected to a suitable arrangement for counting its output pulses, and that it is exposed to the radiation from a fixed radioactive source. In this case, the number of pulses recorded in a given time, i.e. the count rate, would not be exactly constant, but would fluctuate about a mean value, owing to the fact that the radioactive atoms in the source do not disintegrate at regular intervals. This is the characteristic behaviour of a radioactive source.

In addition to this random variation in the count rate, there is also a systematic variation, which depends on the value of the p.d. applied across the tube, and this represents the characteristic behaviour of the Geiger tube. This is illustrated in Fig. 24.12, from which it can be seen that there is a range of values of the p.d. which do not affect the count rate very much. The horizontal portion of this characteristic curve, which represents this condition, is called *the plateau region*, or *the Geiger region*, and for normal operation the p.d. is selected to be in the middle of this plateau.

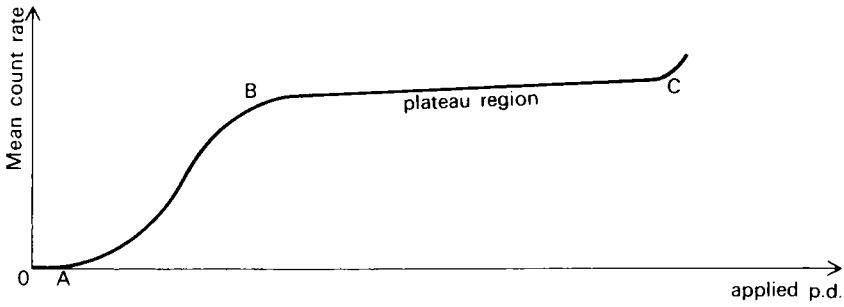


Fig. 24.12 The characteristic of a Geiger-Müller tube

When operated in the plateau region, the counter responds to every particle, but the size of the pulse is independent of the energy of the particle. This is because even a small initial ionization produces a full avalanche. Consequently, all the particles are counted, without discrimination between them. This statement must be slightly qualified, however, since in practice the plateau of the characteristic curve is not quite horizontal, meaning that the count rate is to some small degree dependent on the value of the applied p.d. For this reason, the tube is operated near the middle of the plateau, and the p.d. is kept at a constant value throughout a single experiment.

The p.d. should not be allowed to reach a value beyond the limit C of the plateau. In this condition, the action of the quenching agent fails, and the discharge may become continuous, causing damage to the tube.

24.17 Recording the pulses

Pulses can be observed in a school laboratory experiment by connecting a pulse electrometer directly in series with the tube, but this is an elementary method of

using it since only the very slowest of count rates can be measured. To obtain full use from the tube requires some form of recorder which can respond to fast count rates and which has a dead time of its own not exceeding that of the tube. Using a halogen quenched tube, the voltage pulses developed across a load resistor can be applied directly to an oscilloscope, or through a simple amplifier to a loudspeaker. For counting purposes, however, the pulses are applied to a scaler or a ratemeter, after passing through a suitable pulse shaping circuit.

A scaler

A scaler is an electronic pulse counter, and if this name had not already been allotted to the tube which produces the pulses, perhaps the scaler would have been called a counter. The display may take one of several different forms, a common one being that given by a dekatron tube.

A dekatron is a form of neon tube, of which the circular glass top surface is presented to the observer. This displays a central metal disc, the anode, and has the appearance of having ten separate rods, equally spaced round the perimeter, pointing towards the observer. These are cathodes, and are numbered in the sequence 0, 1, 2, 3, . . . 9.

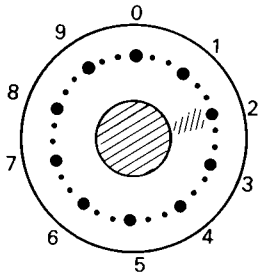


Fig. 24.13 The face of a dekatron tube

Before any pulse has reached the scaler, it can be adjusted so that the gas discharge in the dekatron occurs between the anode and the zero cathode, with the result that a neon glow registers zero. This occurs because the p.d. between the anode and this particular cathode exceeds that provided at any other cathode. The arrival of the first pulse is made to alter the dekatron potentials in such a way that the discharge transfers itself to the number one cathode, and so on, until the tenth pulse arrives, which brings the discharge to the zero cathode once more.

The zero cathode is coupled to a second dekatron in such a way that this receives a pulse on each occasion that the zero cathode of the first dekatron receives one. Thus, the second dekatron is made to count tens, and similarly a third one can count hundreds. By this stage, the count rate required is sufficiently slow to register thousands on an electromechanical counter, which works on electromagnetic principles.

A more complete account of the operation of the dekatron would explain the presence of two extra electrodes separating each pair of cathodes; these are called guide electrodes, having an essential function in the process of transferring the discharge from one cathode to the next.

A ratemeter

The ratemeter records the average count rate of pulses as a reading on a meter. It consists essentially of a pulse shaping circuit, a circuit which integrates the pulses, and a microammeter.

The simplest form taken by an integrating circuit is shown in the diagram (Fig. 24.14). The pulse shaping circuit first ensures that pulses of equal amplitude and suitable wave form reach the top plate of the capacitor C, with the result that it receives a number n of increments of charge Q in time t, n/t being the mean count rate, where Q represents the charge delivered with each pulse from the shaping circuit. Initially, the capacitor proceeds to charge up, but it quickly settles down at a steady p.d. such that the mean current flowing in the resistor R and recorded on the microammeter is

$$I = \frac{nQ}{t}.$$

The time constant, RC, of the resistor and capacitor is selected to be large compared with the interval between successive pulses in order to provide a proper integrating effect. The meter necessarily fluctuates to some extent in its reading, however, owing to the random nature of the pulses caused by the radioactive source.

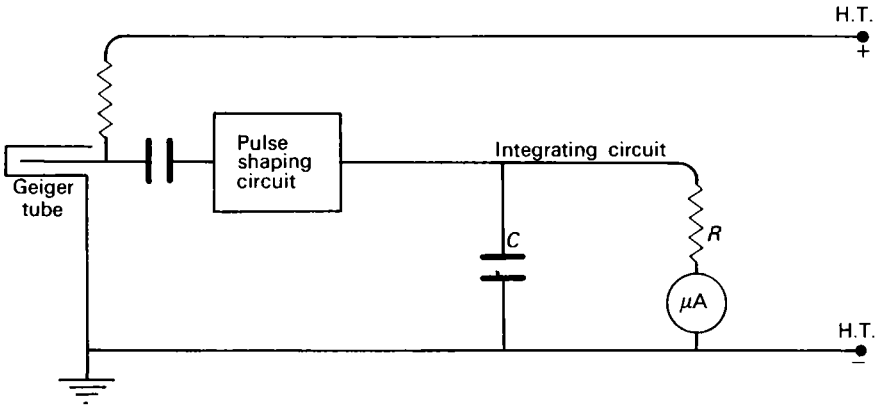


Fig. 24.14 The principle of the ratemeter

Background radiation

A Geiger counter connected to a scaler or a ratemeter is sufficiently sensitive to record the background radiation to which we are continuously exposed, and this must be allowed for when readings are being taken. Since the background count, like all radioactive effects, is irregular, it should be determined over a sufficiently long period of time to average out the irregularity. The mean background count rate is then subtracted when measuring the count rate due to a radioactive substance.

24.18 Solid state radiation detectors

The account which follows of solid state types of counter is brief, because they are not normally used in elementary experiments; in a more advanced field they are very important instruments.

The spinthariscope

This simple device was used to count α-particles in many of the famous early experiments, each separate particle signalling its arrival by causing a faint flash of light on a small screen. This is a screen coated with a thin layer of zinc sulphide,

and the effect of the α-particle is to put electrons in the zinc sulphide into higher energy levels, the light being emitted as the electrons fall back to the ground state. In the spinthariscope, the screen is observed through a magnifying glass, and must be viewed in a darkened room, after allowing time for the eye to adjust itself to the gloom.

Like the spark counter, the spinthariscope will detect only α-particles. Fluorescence can, of course, be produced also by a strong beam of electrons or of X-rays; otherwise, television and visual X-ray pictures would not be possible. The reason that the spinthariscope fails to work with β- and γ-rays is lack of intensity; a single α-particle affects a large number of electrons, whereas a single β-particle or γ-ray photon does not.

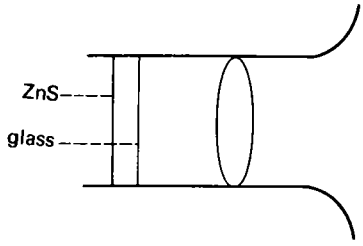


Fig. 24.15 A spinthariscope

The scintillation counter

In this instrument, which is a modern development from the spinthariscope, the scintillations produced in a selected phosphor are not detected by the human eye but by a highly sensitive form of photoemissive cell known as a photomultiplier. The small flashes of light produced by incident radioactive particles or photons are directed on to a photosensitive surface, which consequently emits electrons, and the number of these is enormously increased by a series of secondary emissions, in a manner which is effectively a chain reaction. The electrons are accelerated towards one electrode after another in an evacuated tube, each electron ejecting several further electrons at the next electrode. In this way, each light flash results in a detectable pulse of current flowing in a load resistor connected in series with the final anode.

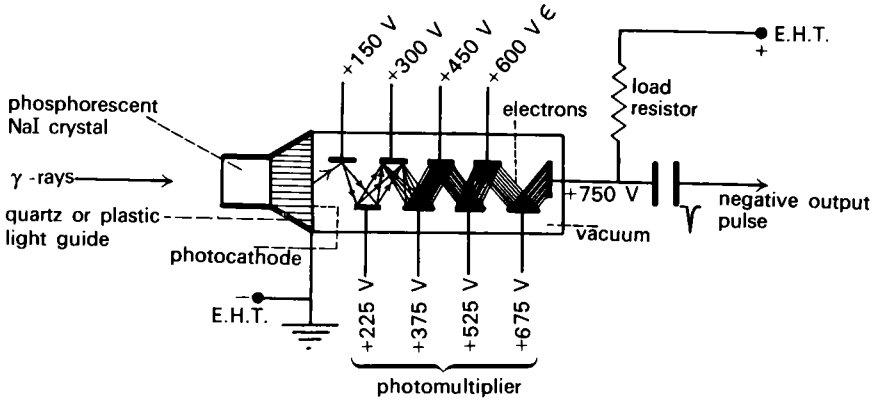


Fig. 24.16 A type of scintillation counter for γ-rays

Scintillation counters can be made which function with α-, β- or γ-radiation, by selection of a suitable phosphor in each case. They have the advantage of being capable of a very high count rate indeed.

A semiconducting counter

This is a reverse-biassed silicon junction diode, and its reaction to radioactive radiation is similar to that of a photodiode to visible photons; the energy of the incident radiation is taken up in releasing charge carriers in the depletion layer of

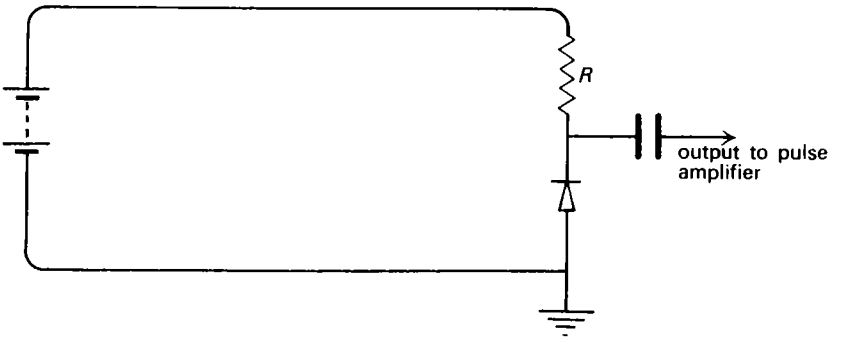


Fig. 24.17 The basic circuit of a semiconducting diode used as a particle detector

the diode, and the separate arrival of each particle or photon produces a voltage pulse across a load resistor. For use as a particle counter in this way, the diode is so formed that its depletion layer is very close to a surface of the crystal, so that the particles shall reach it.

The photographic emulsion

Another of the earliest detecting devices which has been adapted to greater effect is the photographic plate. Special plates are now made in which a thicker layer of emulsion than normal is applied, which contains an unusually high density of silver bromide grains. When an ionizing particle enters such an emulsion, and this is subsequently developed, the track followed by the particle is revealed as a trail of free silver. The path of an α- or β-particle through such a dense material is, of course, extremely short, and this technique is particularly suited to the study of the very high energy particles produced by cosmic radiation or by a modern accelerator, since the tracks of such particles are longer.

24.19 Absorption experiments

Provided that it can be accepted that the only types of radiation possible from a given radioactive source are α-particles, β-particles, or γ-rays, the easiest way to identify them is by studying their absorption, since the penetrating powers of the three differ so markedly. The two experiments of this kind described below can serve to illustrate the use of apparatus found in a school laboratory.

The absorption of α- and β- particles – an experiment using an ionization chamber and a pulse electrometer

The ionization chamber commonly used in this experiment takes the form of a cylindrical metal can, which is used as one electrode, and surrounds a vertical rod which is connected to make an extension to the central rod of a pulse electrometer. This is used as the second electrode, and in this way the ionization chamber is

automatically connected in series with the electrometer which is used to measure the ionization current, represented by the pulse rate of its leaf. A p.d. of 2 or 3 kV is applied across the chamber and electrometer in series, since this high voltage is required to operate the electrometer. A gauze window set into the top surface of the metal can allows access for α-particles to enter the chamber. The radioactive source is mounted a short distance above the gauze window, and the ionization current can now be measured, in arbitrary units, by counting the pulse rate. The effect of inserting absorbing screens directly between the source and the top of the ionization chamber can now be observed. These screens may be either sheets of very thin paper, such as cigarette paper, or of very thin aluminium foil (about 10^{-2} mm).

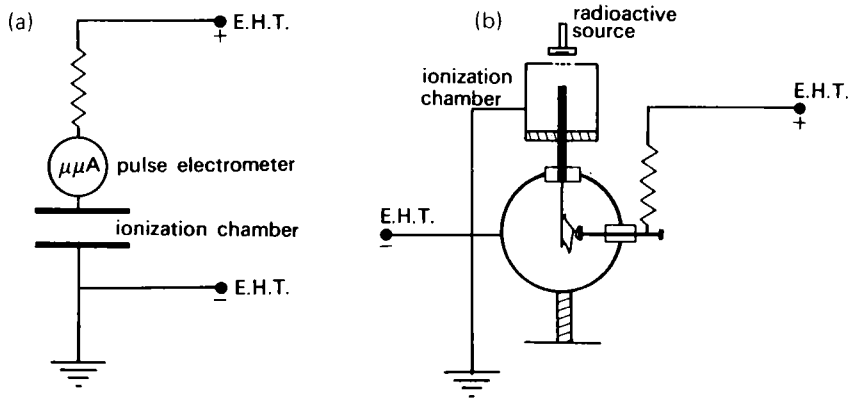


Fig. 24.18 Absorption experiment using a simple ionization chamber and pulse electrometer (a) The circuit (b) The actual arrangement

If the source is emitting α-particles, the effect of these absorbing sheets is very evident; the first three or four thicknesses progressively reduce the response of the electrometer to a very marked degree. Any ionization current that remains, however, is only slightly affected by the addition of further sheets, and would require a more effective absorber, such as several mm of aluminium, to remove it completely. If any such residual current does flow, this indicates that the source is also emitting β-particles. Whether or not there is some γ-radiation as well cannot be determined by this apparatus, unless the γ-ray intensity is very dangerously high, because a pulse electrometer cannot detect the small ionization current which would result.

If ionization current, i.e. pulse rate, is plotted against the number of thicknesses of paper or foil inserted, the graph takes the form shown in Fig. 24.19. Pulse rate is

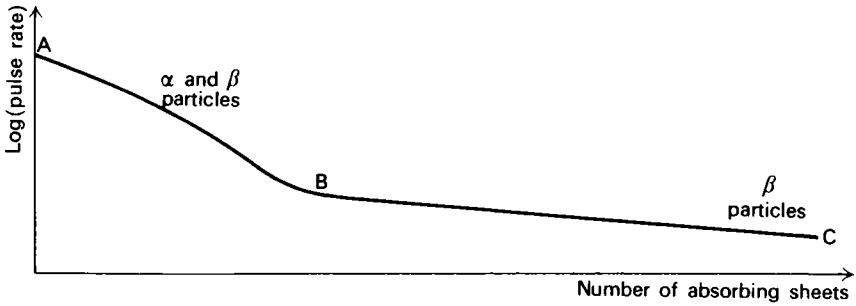


Fig. 24.19 Absorption curve for α and β particles

best plotted on a logarithmic scale, in order to display both the large effect of the first few absorbing sheets and the much smaller effect which occurs in the region BC, where only β-particles are getting through. This second part of the graph is nearly straight, because the absorption of β-particles approximately follows an exponential law.

The absorption of β-particles and γ-radiation – an experiment using a Geiger-Müller tube

If γ-radiation is to be detected, an ordinary halogen quenched type of Geiger tube may be used. This would, however, have been an unsuitable alternative in the last experiment, because only a specially constructed type of tube will admit α-particles, and these would have been missed.

The Geiger tube is connected to its power supply, and the output pulses across the load resistor are passed in to a scaler or a ratemeter (Fig. 24.11). The p.d. is then adjusted to a value corresponding to the middle of the plateau region for the tube (Fig. 24.12). This can be done by observing the effect upon the count rate of varying the p.d. when the radioactive source is placed at a fixed distance from the tube.

Next, the source is removed, and the backgruound cont rate is measured; this will be subtracted from all subsequent readings if it is large enough to make an appreciable contribution to the total.

The source is now placed in position again, at a fixed distance of several centimetres from the Geiger tube, and the effect of interposing aluminium sheets which are thick enough (each, say, 1 mm thick) to produce a marked absorption of β-particles is observed.

Using this apparatus, the count rate represents an actual count of the number of β-particles reaching the Geiger tube, e.g. number of particles per minute, plus a proportion of the γ-ray photons which enter it. This is in contrast with the significance of the pulse rate measured in the last experiment, which represented the ionization current. Consequently, when using the Geiger counter, an effect is very evident which is absent with the ionization chamber, and this is the random way in which particles and photons arrive, demonstrating the essentially random nature of radioactive disintegrations. It is therefore necessary to make counts over a time interval which is large enough for a reasonable approach to the mean count rate to be obtained.

If log (count rate) is now plotted against the number of absorbing screens interposed, a graph which bears a close resemblance to that of the last experiment is obtained. In this case, however, it is the β-particles which have been completely absorbed by a few mm of aluminium, and the radiation which passes through these is γ-radiation. There is no limiting thickness of an absorber which will completely stop γ-radiation, and it would require a few cm of lead to reduce its intensity to negligible proportions. The part BC of the curve, where only γ-radiation is having effect, is a straight line, because the absorption in this case is accurately subject to an exponential law.

A note on absorbing processes

When α-particles are passing through a medium, each particle continuously loses energy to the electrons in the medium, until finally it has lost its ionizing power and accepts two electrons to become a neutral helium atom. Because all the α-particles from a particular radioactive isotope are ejected with the same amount of energy, the range of all these particles in an absorbing medium is the same, being a few centimetres in the case of air (Fig. 24.3). Thus, absorption does not diminish the number of particles in the beam until this limiting range has been reached,

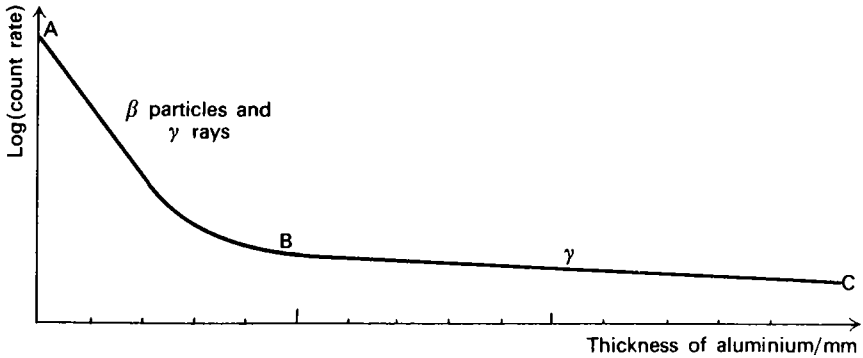


Fig. 24.20 Absorption curve for β and γ radiation

beyond which point the ionizing effect ceases. The response of an ionizing chamber (Fig. 24.19) is reduced with each additional absorbing screen inserted in the path of α-particles, because the remaining ionizing activity of the particles has thereby been reduced, not because, before the limiting thickness has been reached, they have diminished in number.

The absorption of γ-radiation follows a different pattern; energy is lost only when a photon of the radiation causes ionization, and this single photon then loses all its energy, and vanishes, while the remainder of the radiation passes on with its energy undiminished. From considerations of probability, therefore, we can assume that the number of photons which lose their energy over a length δx of path is proportional to the number of photons which have survived so far, whence we can write

$$\delta I = -aI\delta x,$$

from which integration gives

$$I = I_0 \exp(-ax), \tag{24.8}$$

where I_0 is the initial intensity, which falls to the value I after passage of the radiation through a thickness x of the medium.

The constant a is called the linear absorption coefficient of the medium for the particular γ-radiation.

The absorption of β-particles is due to the same continuous ionizing processes as occur with α-particles, except that they ionize atoms less readily and their range in a medium is consequently greater. The β-particles ejected from a radioactive isotope, however, possess a large range of different energies, and their passage through a medium, even through a gas, is tortuous (Fig. 24.3), the net effect being that, over a limited range, an exponential equation similar to equation 24.8 applies approximately to β-radiation.

The inverse square law

In circumstances when the effect of absorption can be neglected, and the size of the radioactive source is small, the intensity of the radiation diminishes with distance by an inverse square law, i.e.

$$I \propto \frac{1}{x^2}. \tag{24.9}$$

This is the natural consequence when radiation is spreading out from a point source into three dimensions.

This law holds for all types of radiation in a vacuum, and it applies in the case of γ-radiation passing through the air, because absorption is negligible.

24.20 Radioactive decay – a measurement of half-life

To make a half-life determination conveniently in the school laboratory, a substance which decays in a reasonably short period of time is required; yet such an isotope is fated, by definition, to disappear within a short space of time, so that a simple means of replenishing the supply is desirable. This requirement is met by choosing the radioactive gas, radon, $^{220}_{86}Rn$, which is also known as thoron. This gas belongs to the radioactive series which starts with thorium, $^{232}_{90}Th$, and any specimen of a thorium salt which has been accumulating its decay products over a period of several years is producing a small quantity of thoron gas continuously. A small quantity of such a salt can be enclosed in a flannel bag contained in a polythene squeeze bottle, and the air in the bottle is consequently contaminated with thoron. A specimen can be obtained as required by squeezing the bottle, the supply in which becomes replenished by further decay of the thorium salt.

The thoron is an α-particle source, and consequently a suitable arrangement for studying its radioactivity is an ionization chamber and a d.c. amplifier (or alternatively a pulse electrometer with an E.H.T. unit as the source of current). For this experiment there would be no gauze window in the ionization chamber, which is completely enclosed, except for inlet and outlet tubes to allow access for thoron from the squeeze bottle.

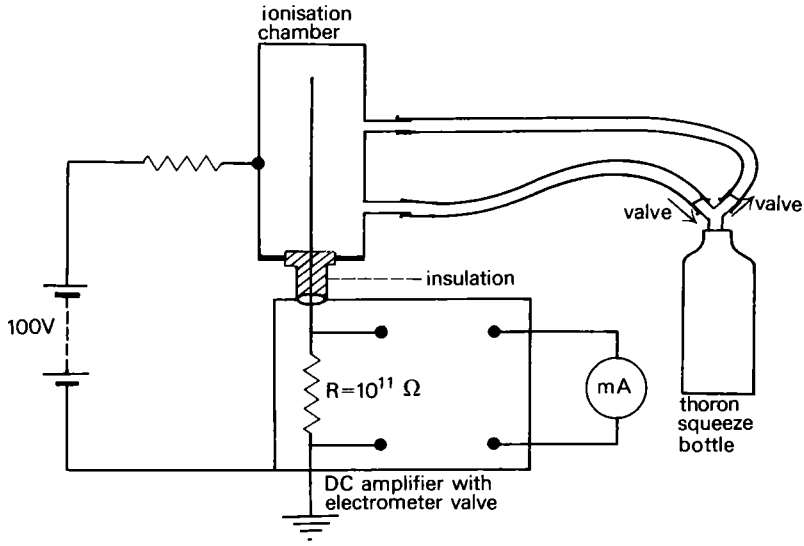


Fig. 24.21 Apparatus to measure the half-life of thoron

Once some thoron has been injected into the chamber, ionization current flows into the 10^{11} ohm load resistor connected in series with it, and this current is measured after amplification by the d.c. amplifier. The current progressively dies away as the thoron decays, falling to almost negligible proportions in a matter of five minutes or so, because the half life of thoron is less than one minute.

The graph of current against time is an exponential curve. The half-life period T may be deduced directly from it by measuring the time which elapses before the current halves in value; this result is quite independent of the particular instant chosen from the graph as the starting point, so that an average can be taken from a number of similar measurements.

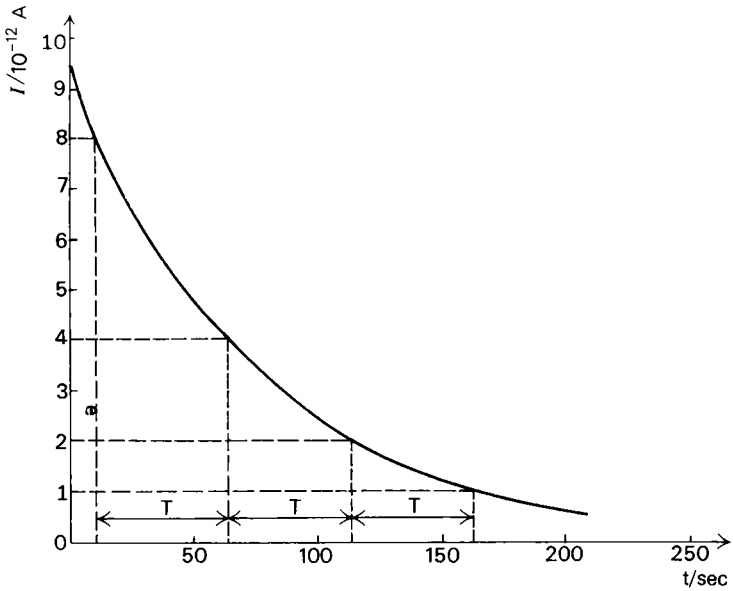


Fig. 24.22 Radioactive decay of thoron. The total ionization current is plotted against time

A more accurate method of calculation consists of plotting log I against t, where I is the ionization current, thus obtaining a linear graph. We have

$$I = I_0 \exp(-\lambda t) \tag{24.10}$$

whence

$$\log I = \log_{10} I_0 - \lambda t \log_{10} \mathrm{e}. \tag{24.11}$$

Thus, both the decay constant λ and the half-life period T may be deduced.

Any two points on this graph which represent a fall in current from any arbitrary value I to the value $I/2$ are separated by a period of time which is equal to the half-life T. The change in value of $\log_{10} I$ which separates these two points is

$$\begin{aligned} y &= \log_{10} I - \log_0 \frac{I}{2} \\ &= \log_{10} 2 \\ &= 0.30. \end{aligned}$$

The half-life period may consequently be read off the graph by the simple procedure which is illustrated in Fig. 24.23. The unit which is chosen for the current I when plotting log I is immaterial.

There are further reasons which make thoron a fortunate choice for this experiment. A radioactive gas which emits α-particles is normally more dangerous than a solid source would be, for the reason that some gas might accidentally be breathed into the lungs. The short life of thoron, however, removes the danger from this, and the products of its decay, which are all solids, are relatively harmless.

A further reason which makes the choice a good one is the relative simplicity of the effect observed. Most cases of radioactivity are complicated by the presence in the source of several different radioactive substances, products of each other's decay. This is the reason that a single substance is often apparently a source of both α- and β-particles, although any single isotope emits only one of the two. The

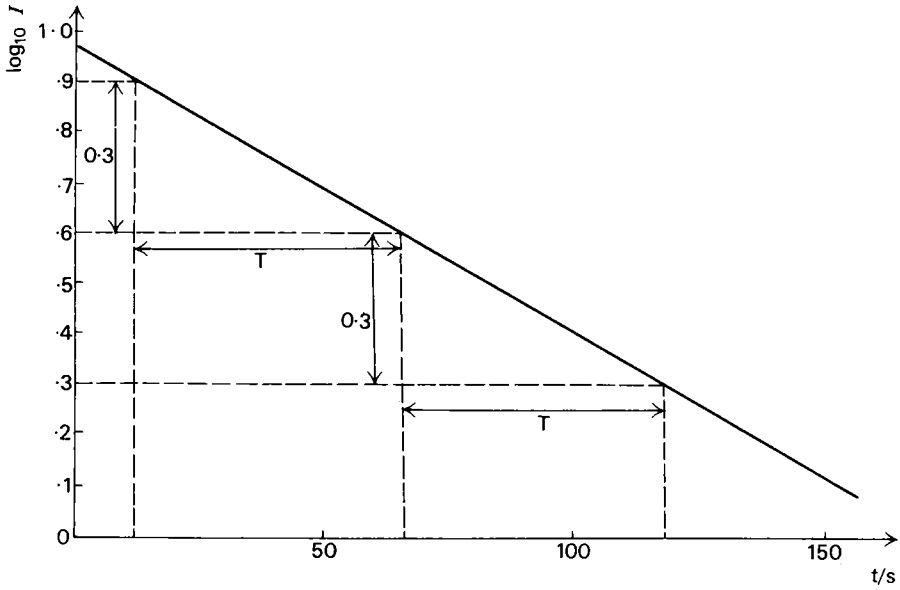


Fig. 24.23 Radioactive decay. Measurement of half-life from a straight line graph

thoron, however, which is introduced into the ionization chamber, is free from its parent substance, and the products of its radioactivity decay either much too quickly or much too slowly to mask the behaviour of the thoron.

Determination of half-life when the decay is slow

In the case of a very long half-life, a different technique would clearly be necessary, since there would be no observable change in the rate of decay over any reasonable period of time. In this case, a means must be devised to determine the total activity of a specimen of the isotope, in terms of the total rate at which it is emitting α- or β-particles, $-\mathrm{d}N/\mathrm{d}t$. The total number N of radioactive atoms present in the specimen must also be estimated, and the decay constant λ can then be deduced from the equation

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -\lambda N. \tag{24.2}$$

Having obtained the value of λ, the half-period is calculated from equation 24.4

$$T = \frac{1}{\lambda} \log_e 2. \tag{24.4}$$

QUESTIONS

$$\log_e 2 = 0.693$$
$$\log_{10} e = 0.4343$$

24.1 (*a*) The parent isotope of the uranium series is $^{238}_{92}\mathrm{U}$, and the final stable product is $^{206}_{82}\mathrm{Pb}$. By the time that a uranium nucleus has undergone the entire sequence of transformations, how many α-particles, and how many β-particles have been ejected?

(*b*) The parent isotope of the thorium series is $^{232}_{90}Th$, and in the course of the complete series of decays, 6 α-particles and 4 β-particles are ejected. What are the nucleon number A and the atomic number Z of the final stable isotope?

In the first three transformations, one α-particle and two β-particles are ejected. What change has resulted in the number of neutrons and the number of protons contained in the nucleus?

(*c*) The half-life of thoron is 52 s. Calculate the value of its decay constant λ in min^{-1}.

24.2 A radioactive element has a half-life of 30 min. What is the average rate of decay per million atoms at any instant? (S)

24.3 The radioactive isotope of phosphorus, $^{32}_{15}P$, decays by the emission of an electron to the isotope of sulphur, $^{32}_{16}S$. Interpret this phenomenon in terms of the neutron-proton model of the nucleus.

The stable isotope of phosphorus is $^{31}_{15}P$. Show that this is consistent with your interpretation. (N)

24.4 The range of an α-particle in Ilford G 5 photographic emulsion is related to its original energy E by the equation

$$E = 0.25R$$

where E is the energy measured in MeV, and R is the range in micrometres.

Radium emits α-particles which are observed to have a range of 19.2 μm in this emulsion. Calculate their initial velocity.

If an α-particle moving at this speed makes a perfectly head-on collision with a stationary proton, calculate the direction and length of the subsequent track of the α-particle.

(Take the mass of a proton to be $m_p = 1.7 \times 10^{-27}$ kg and the mass of an α-particle to be $4m_p$.

24.5 1 g of radium, $^{226}_{88}Ra$, emits 3.7×10^{10} α-particles per second. Estimate (*a*) the number of atoms of radium in 1 g (Avogadro constant $= 6.0 \times 10^{23}$ mol^{-1}.) (*b*) the decay constant λ of radium, in s^{-1}, (*c*) the half-life period T.

24.6 The following readings of count rate dN/dt after time t were taken using a counter at a fixed distance close to a container of radioactive gas:

$\frac{dN}{dt}$ / s^{-1}	30.0	21.5	15.0	12.0	8.75	6.5
t/s	0	15	30	45	60	75

With the gas removed, the counter gave 101 counts in 100 s. Determine the half-life of the gas. (C)

24.7 Two radioactive sources A and B initially contain equal numbers of radioactive atoms. Source A has a half-life of 1 h, and source B a half-life of 2 h. What is the ratio of the rate of disintegration of source A to that of source B (*a*) initially, (*b*) after 2 h, (*c*) after 10 h? (O & C)

24.8 The decay of a certain radioactive source is found to be exponential. Explain this statement.

The disintegration of a radioactive atom is considered to be a matter of chance. How do you reconcile this idea with the observed regular law of decay?

To determine the half-life of the radioactive gas thoron an experimenter uses an ionization chamber into which he introduces air loaded with thoron. The ionization current is 90×10^{-12} A initially, and falls to half this value in 60 s, which is taken to be the half-life of thoron. Unknown to the experimenter, the ionization chamber is contaminated, from previous use, with a radioactive substance whose half-life is 11 h, which is responsible for 4×10^{-12} A of the ionization current at the outset of the experiment. Calculate a more accurate value for the half-life of thoron.

What experimental procedure would have revealed this error? (O & C)

24.9 Describe, with the aid of a diagram, the structure of a Geiger-Müller tube.

Why does the tube contain a small quantity of a halogen or an organic gas?

Explain in what important respect a Geiger-Müller tube suitable for detecting β-particles differs from a tube used for detecting γ-radiation. State, giving your reasons, whether either tube would be suitable for detecting α-particles.

A Geiger-Müller tube in conjunction with a scaler was used to investigate the rate of decay of a radioactive isotope of protoactinium. Counts were made over periods of 10 s, each count starting 30 s after the previous count was completed. The results, corrected for background radiation, were recorded as follows:

Time interval from start/s	0–10	40–50	80–90	120–130	160–170	200–210
Count	3410	2310	1620	1110	770	505

Determine, graphically or otherwise, the half-life of the isotope. (L)

24.10 Give a brief account of the properties of α-particles, with reference to one experiment that gives information about their nature.

Thoron is drawn into a metal air-filled ionization chamber which contains an insulated electrode, and a highly sensitive current-measuring instrument is connected in series with a high tension battery, between the electrode and wall of the chamber. The chamber is isolated from the source of thoron and observations of the current I are made at intervals of time t as follows:

t/s	0	10	20	40	60
I (arbitrary units)	100	88.1	77.6	60.2	46.7
t/s	80	100	120	160	200
I (arbitrary units)	36.2	28.0	21.8	13.2	7.9

Assuming the rate of decay of thoron atoms by the emission of α-particles to be proportional to the number present, find the constant of proportionality. Explain in detail why the value of the current in this experiment is related to the rate of decay of thoron. (L)

*24.11 A sample of aluminium $^{27}_{13}$Al is irradiated by α-particles. What does the symbol $^{27}_{13}$Al signify? The aluminium target gives off a neutron for each $^{27}_{13}$Al undergoing reaction with the α-particles. The product nucleus is radioactive and decays by emission of a positron (a positively charged electron) to give the stable nucleus $^{30}_{14}$Si. Construct a diagram of nucleon number against atomic number to represent this two-stage process.

A Geiger counter measured the positron activity as a function of time after removal from the cyclatron, with the following results:

Time/min	Counting rate
1	5000
2	3841
3	2915
4	2220
5	1724
6	1293
7	977
8	745
9	617
11	361
13	209
15	163
17	110
27	75
37	62
47	57

Construct a graph, choosing your own system of coordinates. Find the half-life of the decay. How would you improve the accuracy of your results? (OXF. SCH.)

*24.12 A piece of timber has been recovered from an archaeological excavation, and it is required to find its approximate age by measuring the radioactivity of the carbon-14 contained in it. For this purpose it may be assumed that the proportion of the carbon-14 in the natural carbon of living wood is everywhere and at all times the same and that it begins to decay at death. If the number of disintegrations observed from 5 g of carbon prepared from the specimen is 21 per minute, how old is the specimen?
(The proportion by mass of carbon-14 to natural carbon in living wood is $1.25/10^{12}$ and the half-value period of carbon-14 may be taken to be 5600 a. The nucleon number of natural carbon is 12.) (N)

24.13 'Gamma rays obey the inverse square law.' What does this mean? What conditions must be satisfied for the statement to be valid?

Describe an experiment designed to verify the statement for γ-rays in air.

The window area of a γ-ray detector is 5.0 cm^2. The window is placed horizontally and lies 80 cm vertically below a small source of γ-rays. 60 photons per minute from the source are incident on it. Estimate the rate of emission of photons from the source. (N)

24.14 A parallel beam of monochromatic X-rays having an intensity of 5000 units was allowed to fall on different thicknesses of steel, and the intensity of the transmitted radiation was measured. The following readings were obtained:

Thickness/cm	1.00	2.00	3.00	4.00	5.00
Intensity	2 810	1 580	885	497	297

Deduce as much as you can from these observations. (N)

24.15

The apparatus in the diagram includes a metal cylinder A about 8 cm in diameter, having a close fitting telescoping lid B. The base of the cylinder has an insulating plug through which protrudes a metal rod whose other end

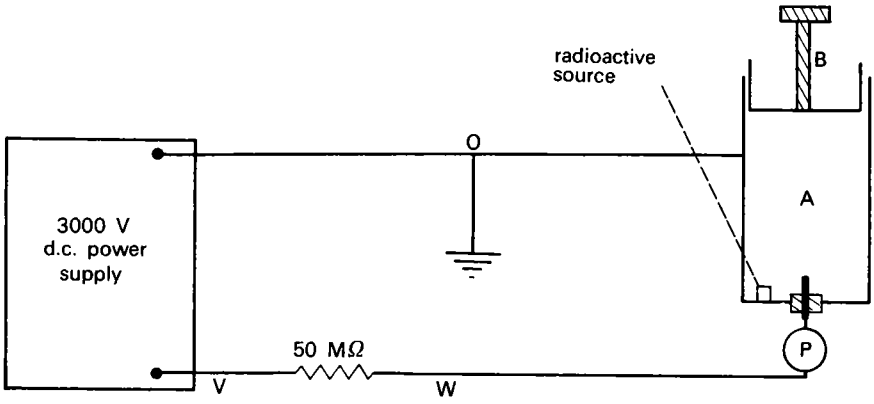


Fig. e 84

is connected to a very sensitive device P for indicating extremely small currents. (This device could be a pulse electrometer, but its exact nature does not matter.) Normally, the current is negligible.

Explain in some detail why the presence of a radioactive source in A causes a steady current to flow. Give an upper limit to the size of this current.

If a source emitting only α-particles is placed in A, a weak current flows,

increasing as B is raised, but never exceeding the value attained at a certain position of B, no matter how much further B is raised. Explain this, and indicate what information can be obtained from the experiment.

If, instead of an ordinary solid source, a certain amount of a particular radioactive gas is pumped into A, the current reading falls, rapidly at first, more slowly later, but quite smoothly, from an initial reading of 112 units to one of 7 units in 208 s. Explain this. Define and calculate the distinctive numerical property of the gas which may be obtained from these observations.

Why has the circuit been earthed at the point O, rather than at some point such as V or W?

The 50 MΩ resistor (usually installed inside the power supply unit) is described as a 'safety resistor'. Explain what protection it gives and how it acts. (S)

24.16 Give an account of the properties of the particles and radiation emitted from radioactive substances.

Describe, in brief, experiments to show how the nature of the particles and the radiation may be established.

How is (*a*) the atomic number, (*b*) the nucleon number of a radioactive nucleus changed by emission of (i) an α-particle, (ii) a β-particle, and (iii) a γ-ray? (O & C)

24.17 Given a standard set of radioactive sources, which include nearly pure α, β, and γ emitters, and also a thorium hydroxide preparation, describe with full experimental details how you would demonstrate two of the following:

(i) that each source emits ionizing radiation;
(ii) the characteristic differences between α, β, and γ radiations;
(iii) that radioactivity involves a decay process, the half-life of which can be measured;
(iv) that radioactive decay is a random process. (O)

24.18 Explain the term 'isotope', and give a brief account of the early observations of the naturally-occurring radioactive series that led to the recognition that isotopes could exist.

Describe the mass spectrometer.

Which of the following processes would you expect to give rise to an isotope of a given nuclide: (i) emission of an α-particle, (ii) emission of a β-particle, (iii) emission of α-particles and β-particles in succession? (O)

Chapter 25

Nuclear reactions

25.1 The ejection of protons from nucleii

Until thirty years after the discovery of radioactivity, the value of the α-particle as a means of piercing the atom was unique, and the first experiments described below belong to this period. No other particle with similar properties was available, with the single exception of those produced in the atmosphere by the effects of cosmic radiation, but these appear at unpredictable places and at unpredictable times. More effective instruments did not appear until about 1930, when successful particle accelerators were constructed, and the neutron was discovered.

Although an α-particle easily enters the electron shells of an atom, it is repelled by the nucleus, because both the nucleus and the particle hold positive charge, and the potential barrier which the α-particle must surmount if it is actually to penetrate inside the nucleus is very high. We can take the nucleus of uranium as an example, since its radius was worked out by an empirical equation in section 24.8. Assuming the potential at the surface of the nucleus to be given by the ordinary electrostatic equation for the potential at the surface of a charged conducting sphere, we have

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{Ze}{r},$$

where, using SI units,

$$\frac{1}{4\pi\epsilon_0} = 9.0 \times 10^9 \text{ N}\cdot\text{m}^2/\text{C}^2,$$

$$Z = 92,$$

$$e = 1.6 \times 10^{-19} \text{ C},$$

$$r = 7.5 \times 10^{-15} \text{ m},$$

whence

$$V = 18 \text{ MV}.$$

Since an α-particle carries charge $2e$, the energy which it must possess, to carry it over this potential barrier, is equal to about 36 MeV, which is nearly ten times the energy with which an α-particle is typically ejected from its radioactive source. Hence we deduce that an α-particle from a natural radioactive source cannot penetrate inside the nucleus of a uranium atom, being successfully repelled before it can reach it, and the same is true in the case of the atoms in the metal foils which were used by Rutherford as targets when he discovered the nucleus.

In the case of an element with low atomic number, however, the potential barrier is not so high, and an α-particle which chances to make a head-on collision with a nucleus of such an element can sometimes enter the nucleus and disturb its stability. The first experiment to have this result was carried out by Rutherford in 1919, when he bombarded nitrogen gas with α-particles, and showed that protons were ejected from nitrogen nucleii in consequence, the equation being

$$^{14}_{7}\text{N} + ^{4}_{2}\text{He} \rightarrow ^{17}_{8}\text{O} + ^{1}_{1}\text{H}.$$

Rutherford's apparatus consisted of a cylindrical metal chamber, which could be filled with nitrogen, air, or other gases, and an α-particle source was mounted inside it. A window set in one end of the chamber could be covered with variable thicknesses of metal foil, and any particles which succeeded in passing out of the chamber through the foils were detected outside by means of a spinthariscope. Initially the chamber was filled with air, and scintillations were observed in the spinthariscope even when the foil was too thick to allow penetration by α-particles, the evidence suggesting that these were caused by protons. Substitution of oxygen or carbon dioxide for the air in the chamber reduced the number of scintillations observed, whereas substitution of nitrogen increased it, thus indicating that the source of the protons might be nitrogen atoms bombarded by the α-particles.

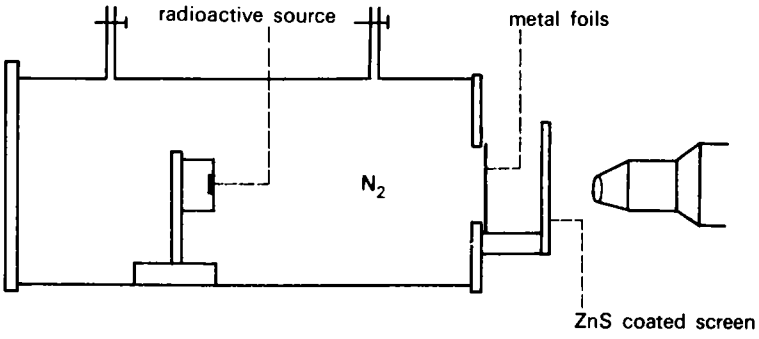


Fig. 25.1 The ejection of protons from nitrogen nuclei. Rutherford's apparatus

The result of this experiment was subsequently confirmed by Blackett, using a cloud chamber containing nitrogen. It required photographs of a very large number of α-particle tracks to produce any in which the necessary head-on collision between α-particle and nitrogen nucleus occurred, but several of these were obtained, the result in each case being a fork in the track, showing that two particles had been produced by the collision. This was significant evidence, indicating that the α-particle had actually entered the nitrogen nucleus, producing the transformation given by Rutherford's equation; if, instead, the α-particle had rebounded, after knocking a proton out of the nucleus, three particles would have been produced, not two.

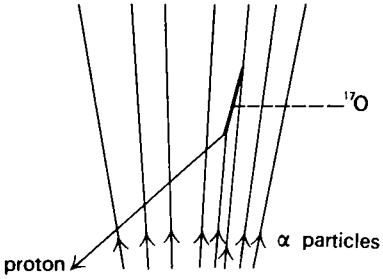


Fig. 25.2 Tracks in a cloud chamber of α-particles passing through nitrogen

25.2 The discovery of the neutron

Having established that the nucleus itself could be broken down into smaller components, one of these being the proton, it was necessary to postulate the

existence of some other nucleon also, in order to account for the fact that, for all elements except hydrogen, the atomic mass number A, now called the nucleon number, is greater than the atomic number Z. For some years this other nucleon was supposed to be the electron, a theory which would account for the ejection of β-particles from radioactive nucleii. There were objections to this theory, however, on the grounds that, according to wave mechanics, the electron would be too large to fit into the nucleus. This difficulty was resolved in 1932, when Chadwick established the existence of the neutron, a particle which carries no charge and has atomic mass number equal to unity, i.e. ${}^{1}_{0}\text{n}$.

This was yet another discovery which resulted from using the technique of bombardment by α-particles. In this case, two German physicists, Bothe and Becker, had found that the bombardment in this way of the light element, beryllium, produced, not protons, but some form of radiation which showed only very slight ionizing effect, and was consequently exceptionally penetrating. Similar results were obtained by using boron or lithium as the target. In view of its behaviour, they wrongly assumed this radiation to consist of very short wavelength γ-rays, and, by studying the absorbing effect of sheets of lead, they estimated the energy of the photons to be about 7 MeV.

This discovery was further investigated by Joliot and Mme Joliot-Curie, who tested the effects of placing various absorbing media in the way. A screen of paraffin wax, or of some other substance rich in hydrogen, gave the surprising result that the ionization produced beyond the screen was considerably increased by its presence. By further experiments using a cloud chamber they showed that high speed protons had been ejected from the paraffin wax, having energies up to 4 MeV, and ranges in air up to 40 cm. For such protons to be released by the effect of γ-radiation would have required photons with energies in excess of 50 MeV, and energy and momentum could not have been conserved in the process of producing them.

The problem was placed in a different light by Chadwick, when he suggested that the radiation emitted from the bombarded beryllium did not consist of γ-rays at all, but of neutrons, particles whose existence had already been suspected by Rutherford. Chadwick's suggested equation was

$$ {}^{9}_{4}\text{Be} + {}^{4}_{2}\text{He} \rightarrow {}^{12}_{6}\text{C} + {}^{1}_{0}\text{n}. $$

The failure of the radiations to cause appreciable ionization, and their resulting high power of penetration, were then explained by the fact that the neutron carries no charge. Their effect in ejecting high speed protons from a target of paraffin wax was explained by the fact that a neutron and a proton have very nearly equal mass; by the ordinary laws of mechanics, such a neutron, on making a head-on collision with a stationary hydrogen nucleus, would surrender to it the whole of its energy.

In justifying his conclusion, including the equation above, Chadwick repeated the experiments and measurements made by Joliot and Joliot-Curie, placing aluminium sheets in the path of the protons as a means of deducing their energies. He also passed neutrons from bombarded beryllium through nitrogen gas in a cloud chamber, and deduced the velocities imparted to the nitrogen nucleii by measuring their tracks. By careful calculations and measurements of mass, momentum, and energy, he demonstrated the consistency of his neutron theory, and he also applied it successfully to another nuclear reaction, bombarding the boron isotope, ${}^{11}_{5}\text{B}$, with α-particles:

$$ {}^{11}_{5}\text{B} + {}^{4}_{2}\text{He} \rightarrow {}^{14}_{7}\text{N} + {}^{1}_{0}\text{n}. $$

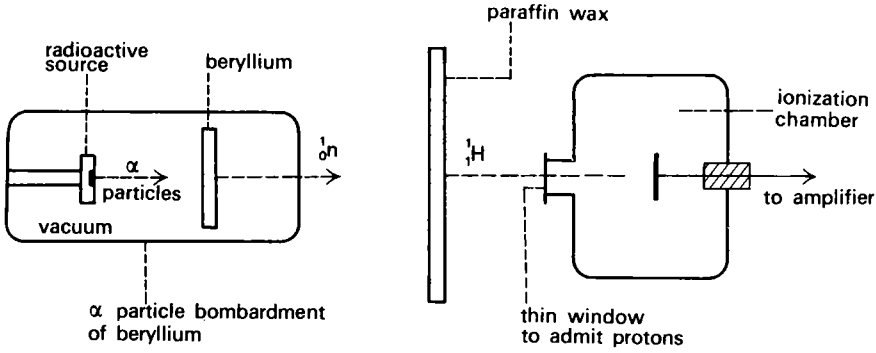


Fig. 25.3 The discovery of the neutron. One of Chadwick's experiments

25.3 The detection of neutrons

The only reason that neutrons produce ionization at all is by the ionizing effect of nucleii which have taken some or all of their energy. Collisions which have this effect are not very frequent in the ordinary way, and the majority of neutrons escape detection by the normal type of ionization chamber or counter. To make these instruments effective, materials from which the neutrons are most likely to eject ionizing nucleii are included. Fast neutrons can be detected by the method that the Joliots and Chadwick used, of putting hydrogen atoms in their path; this can be done by having hydrogen gas in an ionization chamber, or by using a plastic containing hydrogen atoms as the window of a counter. For the detection of slow neutrons, a nuclear reaction which results in the production of ionizing nucleii is used. A Geiger tube, for example, will respond to slow neutrons if it has boron trifluoride gas in it, the reaction being

$$^{10}_5B + ^1_0n \rightarrow ^7_3Li + ^4_2He.$$

25.4 Cockroft and Walton's experiment

When the neutron was discovered, the α-particle lost its unique status as a nuclear projectile, the uncharged neutron being much more effective for this purpose. At about the same time, a rival to the radioactive source was produced, in the form of the first successful particle accelerators. The machines produced in America between 1930 and 1932 by Lauritsen, Lawrence, and Van de Graaff could accelerate atomic nucleii such as protons and helium nucleii to energies approaching those of α-particles from natural radioactive sources, and they had the advantage over a radioactive source that intense and controllable beams were provided. The first important success, however, was gained in England by Cockroft and Walton, who had built a rather less powerful accelerator, which they only later developed to produce energies greater than 1 MeV. Their achievement was to produce the first nuclear reaction to be entirely artificially contrived. Whereas when Rutherford produced the transmutation of nitrogen atoms his bombarding α-particles came from a natural radioactive source, Cockroft and Walton used protons accelerated by their own apparatus to produce a transmutation of atoms of lithium, contained in a target of lithium oxide.

$$^7_3Li + ^1_1H \rightarrow 2^4_2He.$$

This was a result which, by the conventional laws of physics, their experiment should have failed to achieve, because their protons lacked sufficient energy to

surmount the potential barrier surrounding a lithium nucleus. According to the theory of wave mechanics, however, there would be some small chance that such a proton could penetrate this barrier, without actually having to surmount it, and the proton beam was sufficiently intense for this small probability to have some practical effect, with the result that the reaction duly occurred. In their first experiments, the helium nucleii produced by the reaction were observed by means of a spinthariscope; in accordance with the law of conservation of momentum, the two particles should have been ejected simultaneously in almost opposite directions, and two spinthariscopes, suitably placed, were used to view the arrivals of both of them. In a later experiment, the target end of the accelerator projected into a cloud chamber, which revealed the two tracks of the particles.

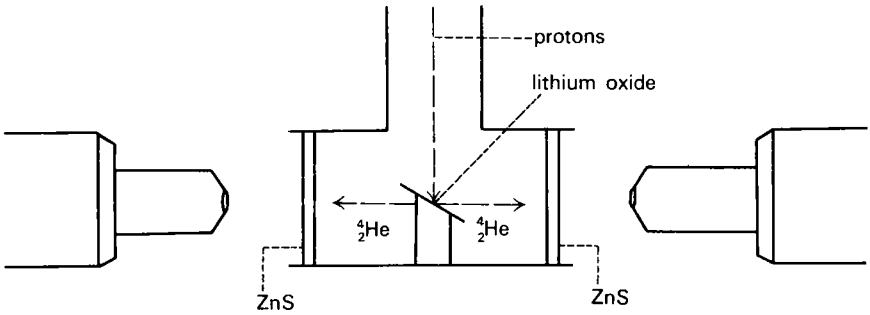


Fig. 25.4 Cockcroft and Walton's experiment

This reaction had the special significance that it released some energy from the nucleii, the two α-particles being ejected from the lithium with combined energy which exceeded that possessed by the incident proton. Compared with the masses of the nucleii involved, this surplus energy is sufficient to represent a significant reduction in mass, in accordance with Einstein's equation

$$E = mc^2. \tag{25.1}$$

Thus, the combined rest mass of the two helium nucleii is slightly less than the sum of the nuclear masses of lithium and hydrogen.

25.5 Some mass and energy considerations

This energy aspect of Cockroft and Walton's experiment is typical of a nuclear reaction in the practical relevance which it gives to Einstein's relation between energy and mass. As a theoretical matter, equation 25.1 is of general application, and whenever a system gains or loses energy its mass increases or decreases by a corresponding amount. Thus, even the energy put into a spring when it is stretched represents an increase in its mass. In cases such as this, however, although the *energy* change may be very noticeable, the change in *mass* is many orders too small to detect; insertion of appropriate values in equation 25.1 would show that, to increase the mass by 10^{-6} g would require a gain of 9×10^7 J in energy. Similarly, the energy changes in a chemical reaction are accompanied, in theory, by corresponding changes in mass, but even the amount of energy released in a chemical explosion is too small for the resulting loss of mass to be detectable. For this reason, conservation of mass can be treated by the chemist as a principle kept distinct from conservation of energy, without any contradiction from Einstein.

The energy changes involved in nuclear reactions, however, relating to the amount of material involved, are of a different order of magnitude, being so large that equivalence of mass and energy becomes a consideration of practical significance.

It is worth pointing out early in this discussion that only in rare instances (which will appear in the next chapter) does conversion of mass to energy involve an actual conversion of matter; when energy is released by nuclear fission in a reactor, for example, this energy has come from the nucleii of the atoms, and there has been an overall reduction in their mass in consequence, but the total number of nucleons has not been changed. Similarly, when a spring is stretched, it has theoretically gained in mass, but this has not increased the number of atoms contained in the spring.

25.6 The binding energy of the nucleus

If we add together the masses of all the nucleons which make up a particular nucleus, and compare the result with the mass of the nucleus, the two figures do not exactly tally, the mass of the nucleus being slightly smaller than the combined total of its component nucleons. This difference is accounted for by the effect of the nuclear forces which bind the nucleons together inside the nucleus, with the result that they are in a state of lower energy when in the nucleus than when they are separate from each other, and the difference in energy is so large that its mass equivalent can be measured. This is usually expressed in energy units, however, and is termed *the binding energy*, ΔE, *of the nucleus*. Thus:

$$\Delta E = [\Sigma m_n + \Sigma m_p - m_N]c^2, \tag{25.2}$$

where m_n, m_p, m_N are the masses, respectively, of the neutron, the proton, and the relevant nucleus, and c is the velocity of light.

The term 'binding energy' appropriately describes the significance of this quantity, ΔE, for the following reason. Because the nucleons in the nucleus are bound together by nuclear force, they possess less energy when in this situation than when they are free, and the more strongly the nucleons are bound together, the greater is this reduction, ΔE, in their total energy.

The quantity which best compares the stability of different nucleii measures the average loss in energy by each nucleon due to being bound in the nucleus, viz.

$$\text{the binding energy per nucleon} = \frac{\Delta E}{A},$$

where A is the nucleon number of the particular nucleus.

Thus, if a nuclear reaction occurs in which the quantity $\Delta E/A$ has, on average, increased, this means that surplus energy has been released.

When a phrase such as 'releasing nuclear energy' is used, this does not imply that a nucleus locks up energy in it which would be released by its complete disruption into its component nucleons, still less that the energy comes from the destruction of nucleons. On the contrary, if a hypothetical experiment were ever to succeed in separating the nucleons, a large amount of energy would need to be *supplied*. When a nuclear reaction releases energy, this is because the nucleons are so rearranged in forming the product nucleii that they are effectively more strongly bound than they were before, thus being in a lower energy state, and the surplus energy has been released. This is similar to the cause of energy changes in a chemical reaction, except that the chemical reaction concerns atoms bound together, whereas the nuclear reaction concerns nucleons, and the energy exchanges in the latter case are far the greater.

25.7 The significance of nuclear fission and nuclear fusion

It will be observed from the graph that the binding energy of a nucleon is greatest, and its actual energy is therefore least, when it belongs to a nucleus with a nucleon number in the neighbourhood of 50 or 60. Thus any nuclear change which brings nucleons towards this condition is likely to be accompanied by a release of energy. The ejection of an α-particle from a heavy nucleus is an example of this, and the energy freed with the α-particle, about 5 MeV, represents about

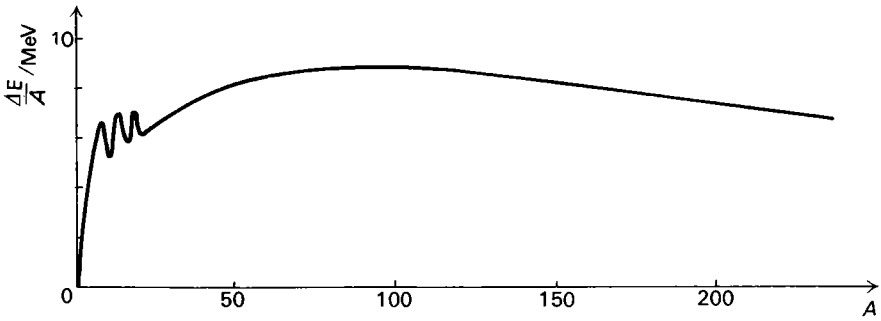


Fig. 25.5 The binding energy per nucleon plotted against nucleon number A for different elements

one million times the amount which is released for each atom involved in the combustion or explosion of a conventional chemical fuel. If an event were to occur which split the heavy nucleus into approximately equal halves, the released energy would be even greater, about 200 MeV for every nucleus which was split up. This is the event which is known as nuclear fission.

For the light atoms, as the graph shows, the energy changes go the other way; if two light nucleii can be induced to fuse together, the release of energy is even greater than that produced by fission of a heavy nucleus. This process is nuclear fusion.

25.8 The neutron in collision with a nucleus

As a means of causing nuclear reactions, the neutron is much more effective than an α-particle or a particle from an accelerator. This is because it carries no charge, and consequently the electric potential of the nucleus sets up no barrier against the neutron's approach, and it can enter a heavily charged nucleus such as uranium at least as easily as it can enter the nucleus, for example, of nitrogen. Once the neutron reaches the boundary of a nucleus, the strong nuclear force is available to attract it in. An added advantage is the fact that the neutron has a negligible ionizing effect, so that the electrons near which it passes do not drain away its energy; unless the neutron happens to encounter a nucleus, it is not stopped at all.

Because it is unaffected by any potential barrier against its approach, the neutron does not require the energy which a charged particle needs to allow it to enter a nucleus, and in many cases a slow moving neutron is the more effective.

In describing nuclear reactions which are induced by neutrons, these are often classified in terms of the neutron energies. A fast neutron possesses the same order of energy as an α-particle, about 1 MeV upwards. A thermal neutron has been

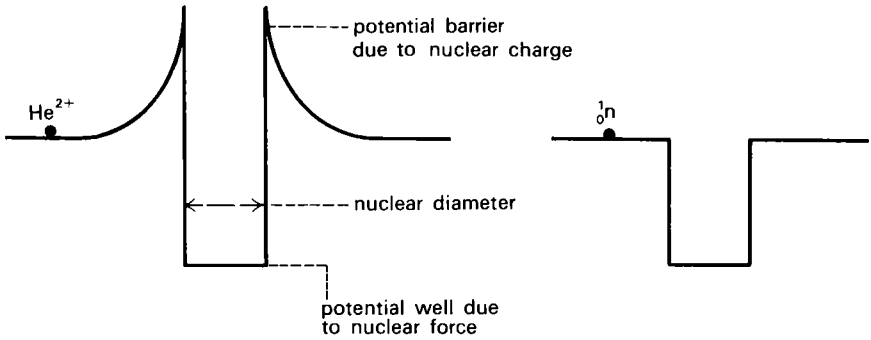


Fig. 25.6 Diagram comparing the energy required by an α-particle with that required by a neutron, if they are to enter a nucleus (a) The α-particle (b) The neutron

slowed down to be in thermal equilibrium with its surroundings, having the kinetic energy E_k given by the kinetic theory relation

$$E_k = \frac{3}{2}kT,$$

where k is the Boltzmann constant, $= 8.6 \times 10^{-5}$ eV/K. At room temperature, this gives a neutron energy of about 0.04 eV.

25.9 Nuclear fission

As soon as the neutron was discovered, its potentialities as a means to effect nuclear transformations were recognized, and in 1934 Fermi embarked on a series of experiments, with the aim of extending the periodic table of the elements. If a neutron were to enter the nucleus of a uranium atom, this would increase the atomic mass of the uranium; if the resulting instability were to cause the ejection of a β-particle, this would increase the atomic number, producing the nucleus of an element beyond uranium in the periodic table, i.e. the transuranic element which is now known as neptunium, $_{93}$Np.

Following the same line of research as Fermi, the German physicists, Hahn and Strassmann, observed the first case of nuclear fission, in 1939. When uranium salts in solution were irradiated with slow neutrons, it was shown by chemical means that one of the products was barium. The atomic number of barium is 56, whereas that of uranium is 92, and so this was fission; on capturing a neutron, the uranium nucleus had not merely ejected a β-particle, or even a few nucleons, but had split into two almost equal parts, the other product of the split being krypton, with atomic number 36.

The significance of nuclear fission with respect to the large release of energy which it involves has already been discussed. This would have no great practical application as an energy source, however, any more than in the case of ordinary radioactivity, unless a way were found to instigate the fission on a sufficiently large scale. The problem is similar to that of lighting a fire, which is successful only because the flames spread, the combustion of one part setting light to the remainder. If it were necessary to provide from a separate external source one neutron for every single nucleus which was split up, there would be no unusual outcome from fission, particularly since many neutrons would, in any case, miss their targets. The analogy would be that of lighting each separate fragment of an ill laid fire, using a vast box of inefficient matches. What presently altered the prospect was the

discovery, later in the same year, that fission of a uranium nucleus, brought about by a single neutron, resulted in the ejection, among the fission fragments, of several more neutrons. This immediately provided the possibility of a chain reaction occurring: if but two of the released neutrons were to cause the fission of two further nucleii, the neutrons from these might split four more, and so on; the fission would rapidly spread through the uranium fuel, releasing a vast amount of energy. The process would start automatically, since there are always a few stray neutrons present owing to cosmic radiation, and these can set the process off.

25.10 Types of neutron collision – controlling factors in a chain reaction

In view of the possibility of a chain reaction taking place, it might at first appear both fortunate and surprising that Hahn and Strassmann survived their experiment, and escaped the experience of a man who inspects his petrol tank with the aid of a lighted match. Their protection was that the fissile material in their experiments was much too loosely packed, so that local fission was unable to spread. This is true, also, of pure uranium extracted from a natural ore, which consists of three isotopes, one of them in such a small proportion that we can neglect it. The important two are $^{238}_{92}U$ and $^{235}_{92}U$. More than 99% of the uranium is uranium-238, and this does not support a chain reaction, although in favourable circumstances it can undergo fission. The nucleii of the isotope do, however, capture neutrons, thus helping to suppress the spreading fission of the uranium-235. Thus, although this rarer isotope can support a chain reaction, it does not do so in natural uranium because too many of the neutrons either escape or are captured by the abundant uranium-238.

When a neutron strikes a nucleus, there are several possible consequences. The collision may be an elastic one, neutron and nucleus sharing their energies, and momentum being conserved. Such a collision is more probable with a fast neutron than with a slow one, and the neutron loses more of its energy to the target nucleus if this has mass of comparable size. This is the normal mechanical effect of collisions between particles, and is the reason that a neutron can transfer all its energy to a proton on striking a target of paraffin wax. An elastic collision of this kind is sought for in a nuclear reactor when the object is to slow down fast neutrons to thermal energy, and the material chosen for the purpose is called *a moderator*.

If a neutron collision is not elastic, the reason is that it has entered the target nucleus, an event known as *neutron capture*. This is, generally speaking, more likely to happen to a slow neutron than to a fast one, and the result is to create a nucleus which is radioactive, having too large a ratio of neutrons to protons for its stability. Radioactive isotopes are formed in this way by neutron irradiation in a nuclear reactor. Neutron capture is also the process by which the chain reaction in a reactor can be deliberately controlled, cadmium or boron control rods being inserted to remove neutrons, preventing the reaction from spreading too fast.

Some nucleii capture neutrons preferentially, having an exaggerated tendency to accept those possessing a particular energy, and this is the case with uranium-238, which captures neutrons having a certain medium range of energies more effectively than it captures either fast or slow ones.

Nuclear fission is an alternative result of neutron capture, and occurs only with certain of the heaviest nucleii. In this case, as already explained, the addition of the extra neutron puts the nucleus in such an unstable condition that it breaks into two fragments of almost equal size. The fission does not always produce the same two fragments, and barium and krypton are not the only possible results of uranium

fission, with the consequence that many radioactive isotopes are produced in a reactor as the direct result of fission. Taking the production of barium and krypton from uranium-235 as a typical example, however, the equation is

$$^{1}_{0}n + ^{235}_{92}U \rightarrow ^{141}_{56}Ba + ^{92}_{36}Kr + 3^{1}_{0}n.$$

In this case, the neutrons are ejected as fast neutrons, the average energy being about 1 MeV.

Uranium-238 undergoes fission only if the neutron which it captures brings with it an amount of energy considerably in excess of 1 MeV. Since the neutrons ejected as a result have less energy than this, they cannot produce further fission, and a chain reaction does not occur. It can occur, on the other hand, with uranium-235, because in this case the capture of any neutron, even a thermal one, causes fission.

25.11 A thermal reactor

Although natural uranium in ordinary circumstances cannot sustain a chain reaction, it is possible, nevertheless, to create conditions in which it does, and it can be used as the fuel in the type of nuclear reactor which is known as a thermal reactor. The fuel is contained in rod shaped cans which are inserted into a large volume of a suitable moderator, into which many of the neutrons escape. The majority of them do not escape further, however, but are rapidly slowed down to thermal speed by elastic collisions, so that a concentration of thermal neutrons accumulates in the moderator, and a proportion of them continuously re-enter the fuel. Owing to the selective nature of neutron capture by uranium-238, which renders thermal neutrons comparatively immune, these are left available to maintain a chain reaction with the uranium-235, and it is because the active neutrons are thermal ones that the reactor gets its name.

Having now arranged for the fission to spread, it is crucially important to keep it under control, and this is the function of cadmium or boron control rods, the action of which in capturing neutrons has already been referred to. These rods can be automatically adjusted in their effect by varying their degree of insertion into the moderator. When lowered in completely, which happens in any emergency, they close down the reactor.

To prevent danger arising from the escape of neutrons and radioactive radiations, the whole reactor is surrounded with thick concrete walls, which serve to absorb them. The energy released in the reactor produces heat, which is extracted for use by circulating a suitable coolant.

If the fuel is natural uranium, the proportion in it of the fissile uranium-235 isotope is so small that much is demanded of the moderator in avoiding wastage of neutrons, and it must not have any tendency to remove them by neutron capture. The material used is either pure graphite, in tightly packed blocks, or it is heavy water, D_2O. To avoid the escape of too many neutrons, the volume of the moderator has to be large, and the size of such a reactor, especially one using graphite, is consequently considerable.

25.12 Plutonium

The large proportion of uranium-238 isotope in a natural uranium fuel is not entirely useless. Although not itself capable of sustaining a chain reaction, the result of neutron capture produces a substance which can do so, and this is plutonium-239, $^{239}_{94}Pu$, a radioactive isotope of the transuranic element plutonium,

with the long half-life of 24000 years. This is formed from the uranium-238 after a sequence of rapid radioactive transformations:

$$^{1}_{0}n + ^{238}_{92}U \rightarrow ^{239}_{92}U \rightarrow ^{239}_{93}Np + ^{0}_{-1}e,$$
$$^{239}_{93}Np \rightarrow ^{239}_{94}Pu + ^{0}_{-1}e.$$

Plutonium-239 resembles uranium-235 in being able to sustain a chain reaction by capture of neutrons, whether these are fast or slow, but it has the advantage of being a different element from uranium and consequently easier to separate out from the uranium-238.

In many types of reactor, the natural uranium fuel is enriched by the addition of plutonium-239 or of uranium-235, thus reducing the necessity to conserve neutrons. This allows greater versatility, enabling smaller sized reactors to be built, and the choice from a wider range of moderators, including ordinary water. If the fuel is very strongly enriched, a moderator may be dispensed with altogether, and the reactor is then termed a fast reactor, since it operates with fast neutrons.

In a fast reactor, many of the neutrons escape, and the fuel is surrounded with a covering of uranium to capture these, thus producing further plutonium. In favourable conditions, the reactor produces more plutonium in this way than the chain reaction is consuming, and it is called a breeder reactor (meaning, of course, that it breeds plutonium).

25.13 The atomic bomb

Even a pure specimen of uranium-235 or of plutonium-239 cannot sustain a chain reaction if the specimen is smaller than a particular critical size, because too large a proportion of the neutrons escape, and so fail to spread the fission. If the critical size is exceeded, however, and there are no control rods which can capture the neutrons, then the uncontrollable reaction proceeds with explosive violence, one of the most disastrous effects ever achieved by mis-applied science. In an atomic bomb, two separate masses of the pure isotope, each separately rather smaller than the critical size, are brought rapidly into intimate contact by use of a conventional explosive, and this sets off the reaction. Incidentally, the critical size is only about that of a cricket ball, yet the energy released is enormous.

25.14 Nuclear fusion

The other possibility for extracting energy from the nucleus is by the fusion together of light nucleii (section 25.7), and to bring this about the two nucleii, which are, of course, both positively charged, must first be brought together, which means that they must be supplied with energy amounting to several MeV. Such energy is readily available to the nucleii in a suitable accelerator, and isolated cases of fusion can be produced by this means, but the method offers no prospect of providing an overall energy gain, because a great deal of energy is expended in operating the accelerator, in order to obtain only very small results. The alternative is to energize the nucleii thermally, i.e. simply by raising their temperature; if fusion were initially achieved in this way in a part of the nuclear fuel, the chain reaction which served to spread it would be similar to that of an ordinary fire, where the combustion of one part raises the temperature of the remainder so that it ignites as well. Thermal energy of a nucleus at room temperature, however, amounts to only about 0.04 eV, which means that to provide the energy of several MeV which is required for fusion the temperature must be raised to more than 10^6 K. Fusion reactions which occur in such conditions are referred to as thermo-

nuclear reactions, and these are responsible for the energy released by the sun, the result of hydrogen being converted to helium.

The only thermonuclear reaction achieved so far by man is the catastrophically explosive one of the hydrogen bomb, where the high temperature initially required is produced by means of a nuclear fission explosion. There is at least some limit, albeit an exceedingly high one, to the power of a fission explosion, because neither of the two components which are to be brought together must separately exceed the critical size. There is no such limit to the size of a fusion bomb.

Controlled nuclear fusion is a prospect which still lies unpredictably in the future, and the problems are very great. To solve them is an important aim, since the earth's sources of energy are running short, and nuclear fission adds only uranium to the list. A controlled fusion reactor, however, could take its fuel from the hydrogen of sea water.

25.15 Some definitions

Before considering the arithmetic of a few typical nuclear equations, it will be useful to gather together certain relevant definitions and allied information. Some of these are included here for completeness although they have appeared already, in earlier chapters.

Atomic number Z

This is, by definition, a whole number, being the number of protons contained in the nucleus. It therefore measures the nuclear charge.

Nucleon number A

This is also, by definition, a whole number, being the total number of nucleons contained in the nucleus, counting both protons and neutrons. Because the masses of the proton and the neutron are almost equal, and that of the electron is, by comparison, almost negligible, the nucleon number gives an approximate comparative measure of nuclear or of atomic masses.

The unified atomic mass constant m_u

This is one twelfth part of the mass of the neutral atom of the carbon-12 isotope, ${}^{12}_{6}C$.

Since the value of the *Avogadro constant* N_A gives the number of atoms of carbon-12 which are contained in exactly 12×10^{-3} kg of this isotope, it follows that the value, in kg, of the atomic mass constant is given by

$$m_u = \frac{10^{-3}}{N_A} \text{ kg.}$$

Relative atomic mass A_r

This was previously known as the atomic mass of the element, although it lacks the dimensions of mass, being a ratio. It compares the exact mass of the neutral atom with that of the unified atomic mass constant, m_u.

Thus,
$$A_r = \frac{m_a}{m_u}$$

where m_a is the mass of a single neutral atom of the element under consideration. The relative atomic mass of the carbon-12 isotope is, therefore, exactly equal to 12.

It will be noticed that, in most cases, the above definitions refer to the neutral atom, rather than to its nucleus only. This is because the instrument used to

measure the masses and the relative masses of atoms is the mass spectrograph, and this measures ions, not nucleii. That is to say that the spectrograph measures atoms which have lost only one or two electrons, and not atoms which have been stripped of them all. For this reason, too, *the binding energy*, ΔE, of a nucleus, which was defined in equation 25.2 in terms of nuclear masses, is more usefully treated in terms of neutral atoms, and we write, as alternative to equation 25.2:

$$\Delta E = [\Sigma m_n + \Sigma m_H - m_a]c^2, \tag{25.3}$$

where m_n, m_H, m_a are the masses, respectively, of the neutron, the neutral hydrogen atom, and the neutral atom under consideration. Both equations yield the same result, because the electron masses which are included in the second one cancel out.

The mass of the unified atomic mass constant m_u

We have
$$m_u = \frac{10^{-3}}{N_A}\ \text{kg}$$

where $N_A = 6.023 \times 10^{23}\ \text{mol}^{-1}$.

$$\therefore\quad m_u = 1.6604 \times 10^{-27}\ \text{kg}$$

and the energy equivalent to this is 1.4922×10^{-10} J, or, using the non SI, but sometimes useful, electronvolt, this energy is equal to 931.4 MeV.

The proton and the neutron

The mass of the proton $= m_p = 1.0073 m_u$
The mass of the neutron $= m_n = 1.0087 m_u$

The mass of the neutral hydrogen atom

$$m_H = 1.0078 m_u$$

25.16 Examples

1. The binding energy of the helium nucleus

The atomic mass of helium is given by

$$m_{He} = 4.0026\ m_u$$

and its nucleus is made up from two protons and two neutrons.

Adding their separate masses together, and using the atomic mass of hydrogen rather than the proton's mass, in accordance with equation 25.3, we have

$$\begin{aligned} 2m_H &= 2.0156 m_u \\ 2m_n &= \underline{2.0174 m_u} \\ &\ 4.0330 m_u, \end{aligned}$$

whence, the discrepancy in mass between the helium nucleus and its component nucleons is

$$\begin{aligned} \Delta m &= \Sigma m_n + \Sigma m_H - m_{He} \\ &= (4.0330 - 4.0026) m_u \\ &= 0.0304\ m_u, \end{aligned}$$

i.e.
$$\begin{aligned} \Delta E &= 0.0304 \times 931.4\ \text{MeV} \\ &= 28.3\ \text{MeV} \end{aligned}$$

and the binding energy per nucleon is

$$\frac{\Delta E}{A} = \frac{28.3}{4}$$

$$= 7.1 \text{ MeV per nucleon.}$$

2. The energy released in Cockroft and Walton's experiment

The equation in this case is

$$^{7}_{3}\text{Li} + ^{1}_{1}\text{H} \rightarrow 2\,^{4}_{2}\text{He}.$$

As already explained, the atomic masses quoted below include the electrons of the neutral atoms, which balance each other out on the two sides of the equation, and we have:

$$m_{\text{Li}} = 7.0160, m_{\text{u}}$$

$$m_{\text{H}} = \frac{1.0078\, m_{\text{u}}}{8.0238\, m_{\text{u}},}$$

$$2m_{\text{He}} = 2 \times 4.0026\, m_{\text{u}}$$

$$= 8.0052\, m_{\text{u}}.$$

Hence $\quad m_{\text{Li}} + m_{\text{H}} - 2m_{\text{He}} = 0.0186\, m_{\text{u}}$

and the amount of energy released $= 0.0186 \times 931.4$ MeV

$= 17.3$ MeV.

The incident protons in this experiment had been accelerated through a p.d. of about 0.3 MV, giving each proton the energy of 0.3 MeV. Adding this to the figure calculated above, and dividing the total energy available equally between the two α-particles which were ejected from the lithium target, the energy of each α-particle works out to be

$$E = \frac{17.3 + 0.3}{2} \text{ MeV}$$

$$= 8.8 \text{ MeV.}$$

Cockroft and Walton took measurements from which they deduced the actual energy to be 8.6 MeV.

The α-particles necessarily accepted the energy equally, and moved off in very nearly opposite directions, in order to conserve momentum, since the momentum of the incident proton was comparatively only small. This is what the α-particles were in the event observed to do.

QUESTIONS

25.1 What is meant by the statement that the stable isotope of gold has an atomic number of 79 and nucleon number of 197? A sample of pure gold is irradiated with neutrons to produce a small proportion of the radioactive isotope of gold of nucleon number 198. What experiments would you perform to examine the radiation emitted by the sample, to establish whether it was α, β, or γ radiation? If chemical analysis of the sample subsequently showed that it contained a trace of mercury (atomic number 80), what would you conclude from this about the nature of the radiation from the radioactive gold? What would you expect the mass number of the isotope of mercury present in the gold to be? (O & C)

25.2 An α-particle is emitted with a kinetic energy of 2.87 MeV from a stationary $^{209}_{83}Bi$ nucleus. Find the approximate velocity with which the residual thallium nucleus recoils.
(The atomic mass constant $m_u = 1.66 \times 10^{-27}$ kg.) (C)

25.3 Calculate the values of each of the following (i) in joules, (ii) in electronvolts:

(*a*) the energy equivalent of the electron's rest mass,

(*b*) the thermal energy of any particle at a temperature of 300 K,

(*c*) the energy in one photon of γ-radiation, wavelength 10^{-10} m.
($m_e = 9.1 \times 10^{-31}$ kg, Boltzmann constant $= 1.38 \times 10^{-23}$ J/K, $h = 6.63 \times 10^{-34}$ J·s.)

25.4 Using the information on atomic masses given below, show that a nucleus of uranium-238 can disintegrate with the emission of an α-particle according to the equation

$$^{238}_{92}U \rightarrow ^{234}_{90}Th + ^{4}_{2}He.$$

Calculate (*a*) the total energy released in the disintegration, (*b*) the kinetic energy of the α-particle, the nucleus being at rest before disintegration.

mass of $^{238}U = 238.0508\ m_u$,
mass of $^{234}Th = 234.0436\ m_u$,
mass of $^{4}He = 4.0026\ m_u$,

where $m_u = 1.66 \times 10^{-27}$ kg. (N)

25.5 Copy and complete the following table to show the number of protons, neutrons, and electrons in the given atoms.

Isotope	Atomic mass	p	n	e^-
$^{6}_{3}Li$	6.01513 m_u			
$^{2}_{1}H$	2.01410 m_u			
$^{4}_{2}He$	4.00260 m_u			

In certain circumstances lithium hydride, LiH, can be transmuted into helium. Write down a nuclear equation symbolizing this statement using the isotopes featured in your table, and write down another similar equation using the values of the atomic masses in the second column.

Explain carefully how conservation laws are involved in your equations.

Calculate (*a*) the loss in mass, per kg of lithium hydride so transformed, (*b*) the maximum energy, in joules, which can be released from 10.0 kg of lithium hydride.
($c = 3.00 \times 10^8$ m/s.) (C)

25.6 Calculate the binding energy for $^{12}_{6}C$, assuming the following data:

$$^{1}_{0}n = 1.0087\ m_u,\ ^{1}_{1}H = 1.0078\ m_u,\ ^{12}_{6}C = 12.000\ m_u,$$

where the above are masses of the neutral atoms.
Avogadro constant $N_A = 6.023 \times 10^{23}$ mol^{-1}, the velocity of light $c = 3.00 \times 10^8$ m/s.
Give your answer (*a*) in joules, (*b*) in MeV, (*c*) in MeV per nucleon.
(The value in kg, of the atomic mass constant m_u may be deduced from the Avogadro constant.) (C)

25.7 The following equations represent two nuclear reactions:

(*a*) $^{235}_{92}U + ^{1}_{0}n \rightarrow ^{236}_{92}U \rightarrow ^{141}_{56}Ba + ^{92}_{36}Kr + 3^{1}_{0}n$,

(*b*) $^{2}_{1}H + ^{2}_{1}H \rightarrow ^{4}_{2}He$.

Explain what these equations represent and why the reactions given may be used as a source of energy.

How is it possible to control the reaction represented by equation (*a*)? To what use has the controlled reaction been put? Why is the reaction represented by equation (*b*) difficult to initiate? Explain briefly what attempts have been made to do this. (L)

Chapter 26

Cosmic radiation and particle accelerators

26.1 Cosmic radiation

Radioactive elements are not quite the only natural sources of energetic nuclear particles, for there are also the cosmic radiations which reach us from space, and the energy of these is very high indeed. The primary cosmic radiation consists largely of protons, and the effect of the earth's magnetic field upon them, due to their positive charge, is to deflect those of smaller energy away, so that the protons which reach the earth's atmosphere have energies exceeding 10^3 MeV, i.e. 1 GeV. These expend themselves by nuclear collisions high up in the atmosphere, and what reaches us at ground level, much lower in intensity, consists of the more penetrating among the products of these high altitude collisions.

Effects due to cosmic radiation were first detected in 1900 by C. T. R. Wilson, who observed that a charged electroscope, however careful its insulation, was subject to a very slow discharge, demonstrating that the air was slightly conducting. This effect, now more familiar as the background count in a Geiger tube, was later traced to origins outside the atmosphere, by sending specially constructed electroscopes up in balloons, and lowering them down mines and deep lakes. In this way it was shown that the activity was much greater at high altitudes, yet some part of the radiation was so penetrating that even depths of water could not remove it.

The modern technique for studying cosmic ray activity is that of collecting particle tracks in photographic emulsions. The effect of nuclear collisions made by particles of such great energy is to produce new kinds of elementary particle, π-mesons, μ-mesons and the like; quite a variety of these in all could be listed, but a description of them would not properly belong in this book. All such particles, if they are charged at all, carry charge e, equal in magnitude to that of the electron and of the proton, and positive or negative in sign; but they differ widely in their mass, mesons, for example, lying between the electron and the proton, with mass about 200–300 m_e. The only stable particles to add to the electron and the proton and the relatively stable neutron are the positive electron (the positron), and the neutrino. All the others have very short half-lives, 10^{-6} s or less, as they rapidly change in identity, the stable end products usually being positive or negative electrons.

26.2 The positron

It was in the course of investigating the properties of cosmic radiations that the positron was discovered by Anderson in 1932. Its existence had previously been predicted, on theoretical grounds, by Dirac, as being the antiparticle to the negative electron (which is alternatively named the negatron). The positron is a stable particle, which has exactly the same rest mass as the negatron, and exactly equal charge, but positive in sign.

Anderson was carrying out experiments in collaboration with Millikan to estimate the energy of cosmic particles, to which end they had constructed a vertical form of expansion cloud chamber, which might record tracks in the wake of cosmic particles coming downwards through the atmosphere. This cloud chamber was placed in the field of a powerful electromagnet so that the speed of the particles might be estimated from the curvature of their tracks. It took much patient experimenting to obtain useful photographs from such a cloud chamber, because it would record only events which chanced to occur immediately after an expansion had been made in the chamber, and the arrival of cosmic particles was quite unpredictable. The first photograph obtained which was relevant to the discovery of the positron displayed the result of a collision with a nucleus, two particles coming downwards following the collision, their tracks curved opposite ways, so that they must have carried charges of opposite sign. The positive one was initially thought to be a proton, knocked out of the target nucleus, but another photograph, later obtained by Anderson, showed that the two tracks must have been those of a positron and a negatron. In obtaining this second photograph, Anderson had placed a sheet of lead across the horizontal diameter of the cloud

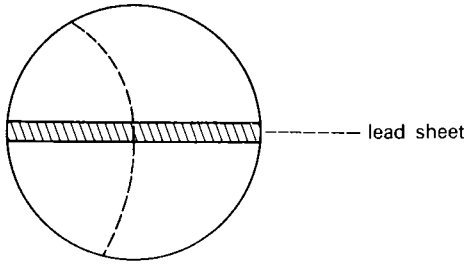


Fig. 26.1 The positron. The appearance of a positron track in Anderson's photograph. That the particle is moving upwards is evident from the sharper curvature of the track when it is above the lead. The direction of the deflection then shows the charge to be positive

chamber, to tell him which direction a particle was following – whether, for example, it was a negative particle moving downwards, or a positive particle moving upwards; this the lead would do by slowing the particle down, giving its track a sharper curvature in the magnetic field after passing through the lead. This particular photograph displayed the track of a particle (which happened to be moving upwards) and the appearance and range of the track was that of a high energy electron; but the direction of its deflection showed that it carried positive charge, and it must be a positron.

It was two years after this discovery that Joliot and Joliot-Curie conducted their experiment in which a radioactive isotope which ejected positrons was produced. This type of radioactivity is no longer uncommon, being produced by bombarding a substance with positively charged nucleii from a cyclotron. Unlike the effect of neutron bombardment, this tends to put an excess of protons in the nucleus, which gives rise to positron radioactivity.

26.3 Pair production and pair annihilation

What Anderson and Millikan had recorded in the first of the photographs referred to above was a case of pair production, and they performed further experiments to verify this. This occurs when a very high energy γ-ray photon passes close to an atomic nucleus. The energy of the photon is converted into the masses of an

electron pair, a negatron and a positron. No γ-ray photon can undergo this transformation unless its energy is sufficient to create the necessary mass, in accordance with Einstein's mass-energy relation

$$E = mc^2. \tag{26.1}$$

Thus, the energy required to create the electron pair is

$$E = 2m_e c^2$$

and this minimum photon energy is just greater than 1 MeV, the corresponding frequency and wavelength being

$$\nu = 2.4 \times 10^{20}\ \text{Hz},$$
$$\lambda = 1.25 \times 10^{-14}\ \text{m}.$$

If the photon which produces the electron pair has greater energy than this, the surplus is taken up by the motion of the particles, their masses then becoming greater than their rest masses. The presence of a nearby nucleus is essential, in order to conserve momentum.

Although the positron is a stable particle, it does not, in practice, have a long life, because of the abundance of negatrons. When the two meet, pair annihilation results, their mass being converted back to the energy of γ-radiation. This event does not require the proximity of a nucleus, as momentum is conserved by the production of a pair of photons, the energy being equally divided between them.

Unlike a nuclear reaction, such as those involved in nuclear fission and nuclear fusion, pair production and pair annihilation represent the actual conversion of radiation energy to matter, and vice versa, actual particles being created from this radiation energy, and back again.

If the energy of the initial γ-ray photon is particularly great, a rapid succession of these transformations can result in the production of a whole cascade of positrons, negatrons, and lower energy γ-ray photons, and these are sometimes observed to occur in the atmosphere, where they are known as Auger showers, after the French physicist who discovered them in 1937. The process starts with the particle known as a neutral π-meson, one of the products of a nuclear collision with a primary cosmic particle high up in the atmosphere. The π-meson has an extremely short half-life, decaying into two very high energy γ-ray photons, and the cascade is initiated by pair productions from these.

26.4 The neutrino and the antineutrino

The pretext for including some reference here to the neutrino and the antineutrino is that they complete the list of particles which are stable; but they are very elusive particles. The neutrino has zero rest mass (or, at most, a mass too small to detect), zero charge, and no magnetic properties, but it takes up energy, and acquires mass, by moving with the speed of light, and on this account its appearance is sometimes postulated, to account for conservation of energy, in a nuclear reaction. The antineutrino is its antiparticle, just as the positron and negatron are antiparticles of each other, and when a neutrino and an antineutrino come together, they annihilate each other, the energy turning into γ-ray photons.

When a radioactive transformation produces a positron, a neutrino is believed to be ejected as well, whereas ejection of an antineutrino accompanies negatron radioactivity. An example of the latter case is the radioactive decay of a neutron, which should properly be represented by the equation

$${}^{1}_{0}\text{n} \rightarrow {}^{1}_{1}\text{H} + {}^{0}_{1}\text{e} + \tilde{\nu},$$

where $\tilde{\nu}$ is the symbol for an antineutrino (that for a neutrino being ν).

The reaction of these particles with matter is so slight that they can even pass through the earth; evidence of their existence has, however, been found by experiments in a nuclear reactor, where their presence must be abundant, owing to the concentration of neutrons.

26.5 Particle accelerators

The earliest particle accelerators, including that of Cockroft and Walton, were constructed in order to usurp the function of the radioactive source by producing intense beams of protons and other light nucleii with energies, in some cases, up to 10 MeV. Electrons were accelerated to similar energies, also, in order to provide intense beams of γ-radiation on the principle of a very high voltage X-ray tube. Electron accelerators of similar types are still used for the same purpose today, and in more than one electronic type of microscope as well.

The object of the modern accelerator, which produces nucleii with energies of 30 GeV and more, is evidently not to compete with the α-particle. The purpose now is to produce the types of event initiated by cosmic particles, with the advantages of having an intense and a controllable beam for the purpose. By this means, information is sought about the short-lived fundamental particles, of which mesons are examples, and about the nature of the forces which operate inside the nucleus. High energy electrons also have a special part to play in this, because they are not affected by the strong nuclear force which binds the neutrons and protons, and they are consequently available to provide information which might otherwise be masked by this force.

The modern accelerator, at its most ambitious, is a highly expensive machine, both to build and to operate, and the largest one in Western Europe is internationally controlled by CERN, and located at Geneva.

Since the uses to which such accelerators are put lie outside the scope of this book, this might reasonably be expected to apply also to the machines themselves. The pretext for including some account of them lies in the fact that their basic principles are the simple laws of electricity, and the ingenuity with which these are applied holds considerable interest.

26.6 The basic accelerators

The obvious means to accelerate charged particles (but not, of course, neutrons) is the means by which electrons are accelerated in a cathode ray tube or an X-ray tube. The particles are introduced into one end of an accelerator tube, and a high p.d. is applied between two electrodes placed in the tube. If the potential drop is V, the energy gained by a singly charged particle, such as an electron, a proton, or a deuteron, is eV. To produce such particles, for example, in possession of 10 MeV, i.e. 1.6×10^{-12} J, of energy would require the application of a p.d. equalling 10 MV, and the problem becomes that of providing this exceedingly high potential. The methods available for doing this are identical with those used for the E.H.T. units of a school laboratory, either that of the Van de Graaff generator, which works on electrostatic principles, or an arrangement, such as that used by Cockroft and Walton, based on the transformer or the induction coil. The system devised by Cockroft and Walton was a form of voltage multiplying circuit; a bank of capacitors connected in parallel was charged from a high voltage source, and then put in series instead by the use of diode switching devices. In this way they finally obtained a p.d. of about 2 MV. The Van de Graaff generator, which was described in an earlier chapter, can be made to go higher than this, the maximum being about 14 MV, but the limit of achievement by these direct methods has then been reached.

The focusing electrodes

If charged particles are to be accelerated on the above principle of applying a very high p.d. between two electrodes, some measure of focusing must be provided in the accelerating tube, in order to maintain a concentrated beam. This is effected by having a succession of cylindrical electrodes down the length of the tube, charged to intermediate potentials. Since the electric field strength is zero inside these cylinders, as it always is inside a charged conductor, the particles are accelerated only while they are crossing the spaces between the electrodes, and here the lines of force are curved, giving a focusing effect.

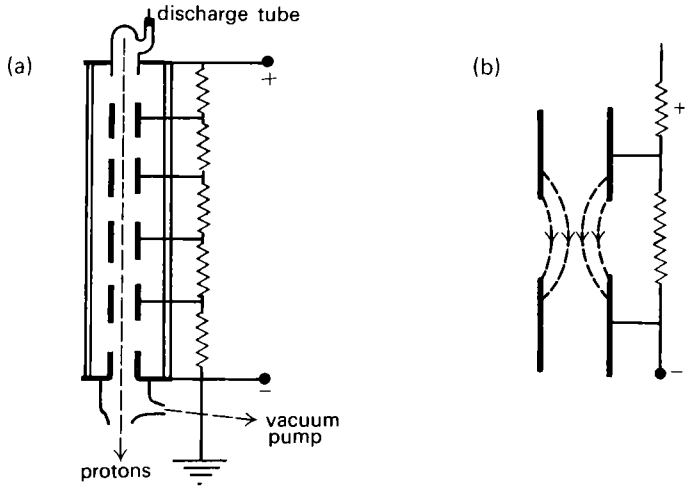


Fig. 26.2 Focusing electrodes in an accelerator tube (a) The accelerator tube (b) The focusing field between successive electrodes

26.7 Synchronous acceleration – a linear accelerator

Although potentials which are much in excess of 10 MV would not seem to be practicable, this does not set a corresponding limit to the energies which can be given to particles. This is due to a principle first suggested by Wideroe in 1928, as the basis of a proposed linear accelerator. The charged particles are repeatedly subjected to the accelerating peaks of an alternating potential, while being protected from the retarding effects of each opposing half cycle of the alternations. In this way, if a particle carrying charge e is accelerated through the same peak p.d. V, n times in succession, it will acquire a total energy equal to neV.

The principle of the linear accelerator is illustrated in Fig. 26.3. The ions are first brought up to high speed in a subsidiary machine, such as a Van de Graaff, and are then injected into a long evacuated track, where they pass through a succession of cylindrical electrodes which are alternately connected to the output terminals of a high frequency source of alternating potential. The gaps between the electrodes are quite small, but it is in these gaps that the accelerating field appears, a particle being in zero electric field whilst it is moving inside the electrodes themselves.

This is called synchronous acceleration, for the reason that only those particles which find themselves between the appropriate electrodes at the right instant are accelerated. As indicated in Fig. 26.3(*a*), they must then contrive to reach the next gap in time to collect the benefit of the accelerating potential here also, which

appears half an oscillation later, owing to the reversal in polarity of the alternating p.d.

Because the particles gain speed as they travel down the tube, yet must reach the electrode gaps at equal intervals of time, it follows that the electrodes must be

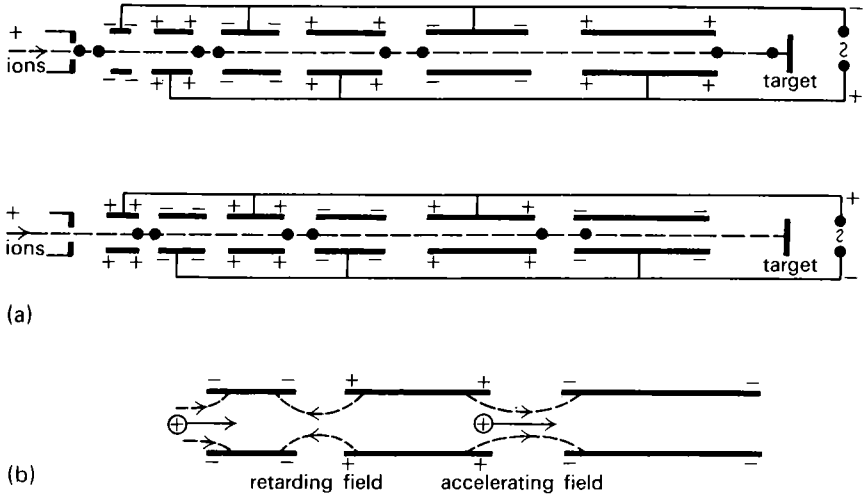


Fig. 26.3 A linear accelerator – synchronous acceleration (a) Two successive half-cycles of alternating p.d. (b) The electric field

made progressively greater in length, in the correct proportion, as illustrated in the diagram. The length of an electrode must be a little less than the distance travelled by the particle in time of one half cycle of the applied p.d.

i.e.
$$l \approx \frac{v}{2f}, \tag{26.2}$$

where v is the particle velocity at that point. Fortunately for the successful operation of a synchronous accelerator, the particles which do not initially happen to be quite at the right place at the right time to be properly accelerated are automatically pulled into step, an important factor which was recognized before such an accelerator was ever constructed.

Since the aim is to produce very energetic particles, and the velocity of these is extremely high, the problem is to keep the overall length of the accelerator within reasonable limits. A 1 MeV proton, for example, is moving at about 1.4×10^7 m/s. This means that it is essential to build up the energy as rapidly as possible, before the particles have had time to travel too far, which means that the successive impulses must be as large as possible, and applied at the maximum available repetition rate. For this reason the construction of a successful proton accelerator had to wait until the technique which would provide the required high potential at very high frequency had been developed, largely during the second world war.

26.8 A linear electron accelerator

Owing to its very small rest mass, an accelerated electron is soon brought up to a speed which approaches that of light, after which any further increase in its

energy is an increase in mass, rather than any further change in speed. To cope with this extremely high velocity, microwave frequencies are used for the linear acceleration of electrons. At a frequency of 3000 MHz, for example, the electrodes of a linear accelerator would require to be placed at intervals of a little more than 5 cm, in accordance with equation 26.2. In practice, however, it is unnecessary, at the short wavelengths possessed by microwaves, to use a sequence of separated electrodes at all; a continuous hollow metal pipe which acts as a waveguide takes their place. Microwave radiation is launched from an oscillator into one end of the pipe, which guides the waves down it, much as a sound pipe carries sound waves. If the cross-section of such a guide is too small relative to the wavelength with which it is intended to operate, the waves are rapidly attenuated, and fail to reach the far end at all; this is the reason that waveguides can only be used at microwave frequencies.

The successful operation of this accelerator depends on the electrons being carried along by the travelling wave in the guide, so that they are continuously pulled forwards by bunching together where the field is favourable, and travelling with it. There is fortunately the same tendency as with the proton accelerator for

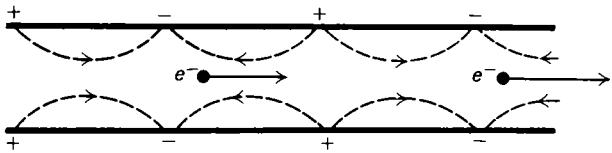


Fig. 26.4 An electron accelerator. The electric field of a travelling wave in a waveguide

the particles to be pulled into phase. It will be noted that, since the electrons are experiencing a forward force while they are moving, they are taking energy from the field, even when they are no longer gaining speed.

If the electrons are to 'ride the waves' successfully in this way, they must be injected into the waveguide when already travelling at an appropriate speed. This is a simpler problem than it might appear, because the velocity of the wave in the guide can be reduced at the injection end by the insertion of diaphragms; in this way, the required initial velocity can be reduced to about $0.4c$, where c is the velocity of light in free space, and this corresponds to an electron energy of about 50 keV, which can be provided by a more conventional method.

Electron linear accelerators can be used to produce intense beams of high energy photons, on the principle of an ultra high potential X-ray tube, and one important application is in radiotherapy.

It will be noted that waveguide accelerators can only be used with electrons, because the particles must initially be travelling at a very high speed, and only electrons are sufficiently small in mass for this to be a practical proposition.

26.9 The cyclotron

The long track required by a linear proton accelerator can be avoided by using a magnetic field to curve the path followed by the particles into a circular one. One pair of electrodes is then sufficient to provide synchronous acceleration, the charged particles being helped on their way twice in each revolution, as they repeatedly cross from one electrode to the other. This is the principle used in the cyclotron, which was invented and first constructed in America by Lawrence in 1930.

The electrodes of the cyclotron are referred to as 'dees' because of their shape. They take the form of two halves of a hollow circular metal box, about one metre across, cut along a diameter and separated from each other. The two dees are enclosed in an evacuated container, which is placed between the poles of a powerful electromagnet.

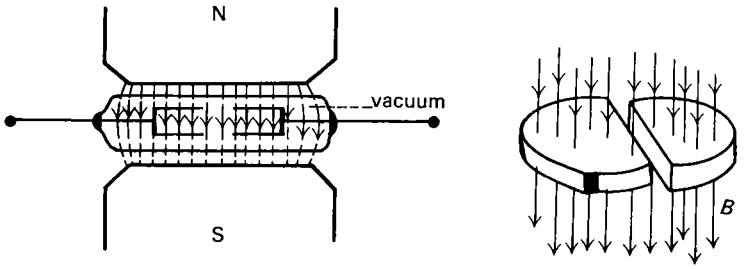


Fig. 26.5 The cyclotron (a) The magnet and dees (b) The dees

The uniform magnetic field B is so directed that it is perpendicular to the plane in which the charged particles are moving, with the result that the particles are subject to a centripetal force, $F = BQv$, which curves their path into a circle, radius r, given by

$$BQv = \frac{mv^2}{r}. \tag{26.3}$$

The frequency f with which they repeat their path, therefore, is

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

$$= \frac{BQ}{2\pi m}. \tag{26.4}$$

This equation contains the essential principle upon which the action of the cyclotron is based, for, although the electrodes are being used as the means to speed up the particles, this acceleration does not affect the frequency with which they complete their circuit; as their speed increases, the radius of their path increases also, and the time which they take to go round remains the same. This means that the p.d. which is applied across the dees to produce the acceleration may be taken from a high frequency oscillating circuit of constant frequency. To obtain synchronization, this frequency must be tuned to the frequency f which is given by equation 26.4, and which therefore depends on the selected value of the magnetic flux density B.

The cyclotron is essentially a device which is suitable for accelerating ions or nucleii, such as protons or deuterons. The effects of relativity prevent it from operating usefully in the case of the much less massive electron; the cyclotron action breaks down before the electron has gained very much energy. This is because the electron, owing to its small mass, requires a comparatively small acquisition of energy to bring its velocity close to that of light. If further energy is supplied to it, the effect is to increase its mass, rather than to increase its speed. Although equation 26.4 is valid in terms of relativity, the frequency f is no longer constant, since the mass m has become a variable. The circuiting electrons begin to go round too slowly, and the synchronization which is essential to the cyclotron operation is lost.

The same thing would occur to a proton too, in the end, of course, but it would have gained considerably in energy first.

The protons, or similar nucleii, start their journey in the cyclotron at the centre of the evacuated container, in the gap which separates the dees. They are introduced at the centre in the first place, either by having a small discharge tube there, with a vertical slit in it from which the protons can emerge, or by putting a small quantity of the appropriate gas into the vacuum, and having a heated filament at the centre as an ionizing device.

Once the accelerating electric field which crosses the gap between the dees has had effect, the protons follow a semicircular path inside one dee until they reach the gap again. This is because there is no electric field inside the dee, and the protons are subject only to the centripetal force caused by the magnetic flux. Provided that the frequency of the applied p.d. is tuned to resonance, the favoured protons will always arrive at the gap at the appropriate moment to encounter an accelerating electric field there, and so their circular track repeatedly increases in radius. This progress is illustrated in Fig. 26.6. Finally, when they have reached

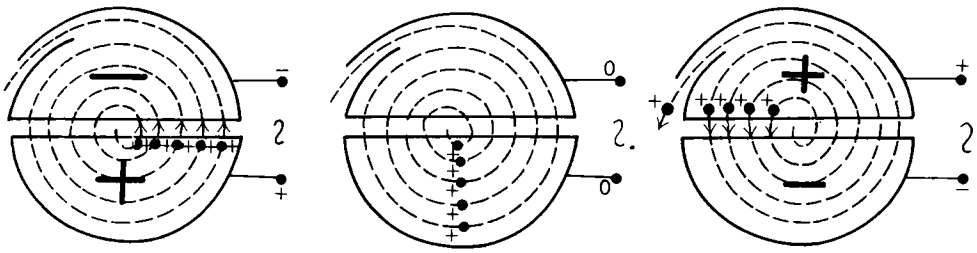


Fig. 26.6 The cyclotron. The synchronous accelerating field

such a speed that the magnetic flux density is only just strong enough to hold them inside the circumference of the dees, the protons are deflected out of the system.

The final speed reached by the accelerated proton depends on the value of the flux density B and on the maximum radius r_{max}, available for the proton's path. Equation 26.4 gives

$$v_{max} = Br_{max}\frac{Q}{m}. \tag{26.5}$$

This speed does not, in theory, depend on the magnitude of the accelerating p.d. The reason for this is that, if the p.d. were reduced, the protons would receive smaller impulses, but they would make a larger number of circuits before reaching the radius r_{max}, and by this time their final speed would have the same value v_{max} as before. A high potential is applied in practice, however, because the beam of protons tends to go out of focus if the distance travelled by them is too great.

The limit of the cyclotron is reached, in the case of protons, at an energy of about 25 MeV. This is partly because, with this energy, the speed of a proton is equal to about $0.2c$, so that its mass is beginning to increase appreciably, and partly because the protons begin to move out of the plane in which they are required to remain.

Assuming the flux density B to be about 1.0 T, and applying equations 26.4 and 26.5 to the case of a proton, the resonant frequency needed for the accelerating potential proves to be about 15 MHz, and the final radius of the proton's path, on reaching an energy of 25 MeV, becomes approximately 70 cm.

26.10 The synchrotron

None of the types of accelerator which have been described so far, whether for protons or for electrons, can match the energies which cosmic particles possess. This goal is reached by the use of the more costly machines known as synchrotrons.

Far from relativistic increase in mass placing a limit on the performance, as it does in the case of the cyclotron, the successful functioning of the synchrotron depends upon it. If a particle can first be brought close to the speed of light by means of a subsidiary accelerator, further additions of energy, causing it to increase in mass, will have negligible effect on its speed. Consequently, the principle of synchronous acceleration can now be applied while the particle is being forced to trace out repeatedly the same closed path; it does not have to spiral outwards as its speed increases, which is its behaviour in the cyclotron, because the term 'acceleration', (the wrong term to use, perhaps), no longer implies any change in speed, and the particle will always complete the same circuit in the same time no matter how fast its energy is increasing. The advantage of this is that the track can, if necessary, be a very long one, since the magnetic flux which is applied to curve the particle's path, although it must be powerful, need extend only across the narrow width of the track, and the total deflection may be made up by contributions from a sequence of magnets placed at convenient points round the track.

The principle is most easily applied, in practice, in the case of electrons, because the initial task of bringing their velocity close to that of light is comparatively easily achieved.

The electron synchrotron

The electrons are injected into the synchrotron from a linear accelerator, in which they have already acquired very considerable energy, in the region of 40 MeV, and

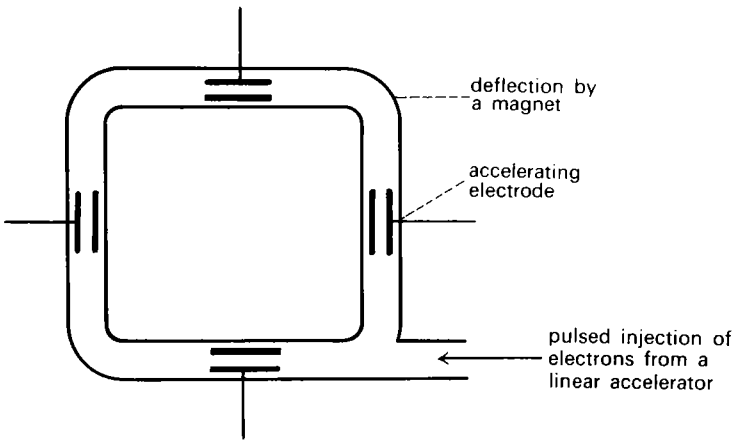


Fig. 26.7 The electron synchrotron (In practice, there are considerably more magnets and accelerating electrodes)

this easily satisfies the requirement with respect to their speed. Provided that the magnets can successfully curve the path of the electrons and keep them repeatedly circuiting round the same fixed track, the electrons will complete successive laps at a constant repetition rate, whether energy is being added to them from the electrodes or not. If they are gaining energy, however, this means that they are also

gaining momentum, and consequently becoming more difficult to deflect, and the magnetic flux which deflects them must by some means be progressively increased in density as the momentum increases. This problem is solved by using alternating magnets energized by 50 Hz alternating current, and pulsing the injection of electrons to synchronize with this in such a way that the electrons enter the synchrotron while the flux density is increasing in the appropriate direction, and are ejected well before the flux begins to die away again. The circumference of the track is considerable, about 200 m, but the electrons are tracing it out with the speed of light, so that a single lap is completed in less than 10^{-6} s. This allows an electron to go round about 10^4 times before it is necessary to eject it.

The accelerating electrodes are placed in straight sections of the track, and connected to the output of a very high frequency oscillating circuit, and conditions are so arranged that each pulse of electrons is attracted towards an electrode as it approaches it, and is repelled by the electrode as it leaves it. The object of the accelerator is to provide electron energies of 1 GeV or more, which means, since each electron completes about 10^4 laps, that it needs to gain at least 10^5 eV each time that it goes round. Although quite a number of electrodes is used, this still means that the potential applied from the oscillator must be a high one.

Radiation loss

There is one disadvantage in a machine using a closed track, which becomes important only in the case of electrons, and this is the large amount of energy which is lost as electromagnetic radiation, ranging in frequency from ultraviolet to X-rays, owing to the fact that the electrons are subject to centripetal acceleration. This effect is in inverse proportion to the fourth power of the rest mass, and is almost negligible in the case of protons, but it puts an upper limit to the performance of an electron synchrotron, and is the reason that the effective diameter of the track is made so large.

The proton synchrotron

Several of these very costly machines have been constructed, but the only one in Western Europe is at Geneva, and operates up to energies of 30 GeV, now in process of being exceeded. The basic principle is the same as that of the electron synchrotron, but the application is complicated by the fact that no subsidiary accelerator can bring the proton velocity very close to that of light. The speed of the protons when injected into the synchrotron fall well short of this, with the result that frequency modulation of the accelerating potential is necessary, to allow for the fact that in their early laps the protons are still gaining in speed. In addition to this complication, such massive particles are very difficult to deflect, and very powerful magnets are required.

QUESTIONS

26.1 Calculate, in MeV, the energy which must be made available if a proton and an antiproton are to be created by a process of pair production. (The mass of a proton is 1.67×10^{-27} kg.)

26.2 A uniform magnetic flux density B is applied perpendicular to the planes of the dees of a cyclotron. Show that the period of motion t of a particle of mass m and charge Q moving with velocity v in an orbit of radius r in a plane normal to B is independent of r and v, being given by

$$T = \frac{2\pi m}{BQ}.$$

By writing down the relation between the kinetic energy of the particle and the radius of its orbit, calculate the change per revolution in the orbit radius and the particle energy if there exists a p.d. V between the two dees and the sign of V is changed so that the particle is accelerated each time it crosses the gap.

At what frequency must the sign of V be changed, and why is the particle only accelerated at the gap?

Calculate the maximum proton energy (in electronvolts) obtainable in a cyclotron of radius 0.5 m and magnetic flux density 1.3 T. What voltage on the sphere of a single stage Van de Graaff accelerator would be required to give the same proton energy? Comment on this value.
(Q/m for protons $= 9.6 \times 10^7$ C/kg.) (OXF. SCH.)

26.3 Assuming that you are designing a cyclotron to accelerate protons to an energy of 30 MeV, and you require the radius of the dees to be not greater than 0.5 m, deduce (*a*) the minimum flux density which will retain protons of this energy within the radius of the dees; (*b*) the frequency of the V.H.F. oscillating p.d. which must be applied across the gap between the dees; (*c*) the peak value of this p.d., if the protons are to acquire 30 MeV from rest after 500 circuits of the cyclotron.
(Q/m for a proton $= 9.6 \times 10^7$ C/kg.)

Answers

3.5 (*a*) +2 V, (*b*) −6/7 V, (*c*) +2 V, +10/17 V, (*d*) +5 V, +25/21 V
3.9 +0.005 V
3.11 76.8 m, 4.9×10^{-4} m/s
3.13 1.5×10^{-7} Ω·m
3.15 2.0 A
3.17 5.26×10^{-3} °C^{-1}
3.19 300 Ω at 20 °C, 400 Ω at 100 °C, $\alpha = 4.55 \times 10^{-3}$ °C^{-1}
3.21 (*a*) (i) 4.0 V (ii) 4.0 V (*b*) (i) 4.0 V (ii) 1.7 V
3.23 50 V
3.25 Increase value of shunt to 2.94×10^{-3} Ω.
3.27 1.5 Ω in series with meter and 76.7 Ω in parallel with the two.
3.31 Input p.d. = 100 kV, output p.d. = 99 kV, output power = 19800 kW; input p.d. = 10 kV, output p.d. = 0, output power = 0
3.35 (*a*) 2.17 V, 1.42 Ω
3.37 4.65 V
3.39 (*a*) (i) 66.7 V, (ii) 33.3 V, (iii) 16.7 V, (iv) −16.7 V, (*b*) 20 V, (*c*) 9.1 V
3.41 28 Ω, 1.46 V
3.43 217 W, 6/19 and 13/19, 85.3%, 16 V
3.45 1.77 V
3.47 95 V, 38 Ω, 49.5 W
3.49 1.43 Ω
3.51 $2\epsilon/l$
3.53 (*b*) 286 Ω; (*c*) 0.98 °C/s
3.55 1.24 : 1, 0.484 : 1
4.7 0.3 cm, 0.49 : 1
4.9 6.75×10^{-2} V
4.11 8.91 mV
4.13 $l = 44.4$ cm, put 120.5 Ω in parallel
4.15 32.05 Ω
4.17 $R = 15.8 + 0.084x$; $R = 15.8$ Ω, 20.0 Ω, 24.2 Ω
4.21 1.023 V, 2.32 mA
4.23 $R_2/R_1 = 2$
4.25 318 Ω
4.27 0.33×10^{-6} A
6.15 6.4×10^{-19} C
6.17 -8.0×10^{-19} C, 4.8×10^{-14} N downwards
6.21 3.77×10^{-2} m^2, 6×10^{-5} C
7.11 2 A/s, 4 A, 400 V
8.5 10^{-3} μF
8.7 (*a*) 280 μC, 120 μC, 80 μC, 40 μC, (*b*) 0.014 J
8.9 (*a*) 17 μC on each (*b*) 24 μC on C_2 (*c*) 12 μC on C_1, 9.6 μC on C_2
8.11 (*a*) 20 V, 10 V, 8 V; 7.6×10^{-4} J, (*b*) 14.3 μC, 5.43×10^{-4} J
8.13 0.1 μF, 100 Hz
8.15 20×10^6 Ω
8.17 0.2 F, 0.04 F, 0.0033 F
8.19 1130 Hz, 10^{-2} J, 1.41 A; 1.0 H, 1000 V, 0.14 A
8.21 2.64 V
8.23 92.3 μC, 30.8 V, 46.1 V, 23.1 V; 92.3 μC, 2130 μJ, 79.1 μC, 7910 μJ
8.25 12
8.27 0.7 s
8.29 18 W, 545 Ω
8.33 0.32 s
9.5 0.20 A, 2.4 J
9.7 2.5 W
9.9 (*a*) $2.4 \times 10^{-4}\cos\theta$ Wb (*b*) $3.77\sin\theta$ V (*c*) 2.67 V (*d*) 0.2 A (*e*) 0.4 W

9.11 4.7 V, 283 V, 4.5×10^{-3} J
9.13 5000 Ω, 1.01 V, 0.71 V
9.15 1.88×10^{3} Ω; 150 V, 0, 150 V
9.17 $V_C = 100 \pm 1.6$ V, $V_R = 0 \pm 10$ V
9.19 0.7
9.21 225 Hz, 165 V, 170°
9.23 $L = 0.88$ H, $R_1 = 118$ Ω
9.27 $10/\sqrt{3}$ mA, $10/\sqrt{3}$ mA
9.29 $V = 1.77$ V
9.31 2.8×10^{15} electrons
10.11 0.35 mm/V
12.13 0.46×10^{6} Ω, 1.5 mA
13.5 (*a*) 1.44×10^{11} V/m, 14.4 V (*b*) 28.8 V, 7.2×10^{15} Hz (*c*) 9.0×10^{-13} m (*d*) 3.4×10^{6} V, 10.9×10^{-13} J (*e*) 1.9×10^{6} V
13.9 255 V/m, 76.4 V, 68.2 V, 1.64×10^{-9} J
13.12 $r = \dfrac{1}{4\pi\epsilon_0} \cdot \dfrac{4Q^2}{mv^2}$
13.13 $\frac{1}{2} \times 10^{-6}$ J, $1/6 \times 10^{-6}$ J
13.15 8.89×10^{-12} F/m
13.17 25×10^{-12} F
13.19 2×10^{-3} N
13.21 1.05×10^{-3} V, $\epsilon_r = 2.4$
13.23 8.0×10^{-11} C
13.25 $1/6 \times 10^{-6}$ J, $1/18 \times 10^{-6}$ J, $1/3 \times 10^{-10}$ F, $\frac{1}{2} \times 10^{-10}$ F
13.27 $10/3 \times 10^{-11}$ F, $5/3 \times 10^{-7}$ J; 5×10^{-11} F, $10/9 \times 10^{-7}$ J
13.29 2.0×10^{-9} C, 860 V
13.31 6000 V
13.32 1.035×10^{4} V
13.37 8.0(3) mm, 2.1×10^{7} V
14.5 3.7×10^{-3} V
14.6 4 410 V
14.9 78.5×10^{-6} V, 3.1×10^{-8} J
14.11 2.61×10^{-7} A, 1.02×10^{-11} W
14.13 5×10^{-5} C, 10^{-1} T
14.15 4.0×10^{-5} T
14.17 6×10^{-3} C
14.19 88 s, 7.5×10^{-5} T
14.21 10^{5} N/m²
14.23 2.8×10^{-2} cm
14.25 $3\pi\sqrt{2}$ s
14.27 2.1×10^{-7} Ω·m, 21 μV
14.29 5.3 cm
14.33 303 μV, 0.50 cm
14.35 0.25 T
14.37 127 W
14.40 0.88 s
14.41 3.8×10^{-5} s
14.43 0.25 T
15.7 7.9×10^{-9} V
15.9 1.75×10^{-5} T
15.11 12.4 mA
15.13 1.81×10^{-7} N·m
15.15 44 μH
15.17 9.08×10^{-3} H, 9.87×10^{-3} H, 3.92 Ω
15.19 6.93×10^{-7} Wb, 2.77×10^{-5} C
15.21 6.75 m/s
15.23 11.4 cm
15.27 30 cm, 20.7 cm

16.2 $\mu_r = 500$
16.5 10^{-3} T, $\mu_r = 401$; 0.2×10^{-3} T, $\mu_r = 801$
16.9 2.5×10^3
17.3 0.316 Ω
17.5 1.0 A, 4.2×10^{-5} T
18.2 2×10^{-7} N·m
18.4 4×10^{-3} μA/mm
18.7 2010 Ω
18.10 (*a*) A/B = ½, (*b*) A/B = 3/2
19.3 23.6 W, 11.8 V, 0.6 Ω
19.5 272 rev/min
19.7 2.06 m/s, 575 kgf
19.12 (*a*) 86.5%, (*b*) 57.6 W
21.3 (*a*) 1.5×10^{-6} m (*b*) 8 (*c*) 6.25×10^{-5} m/s upwards
21.5 $Q_2 - Q_1 = 1.62 \times 10^{-19}$ C $= e$, $Q_1 = 21e$, $Q_2 = 22e$
21.7 3.7×10^{-26} kg
22.5 921×10^{-9}m, 7.9×10^5 m/s, 1.76 V
22.7 700×10^{-9}m, 6.72×10^{-34} J·s
22.9 0.9 V, 625×10^7 s^{-1}, 3.1×10^{-8} W
22.11 1.25×10^{-11} m
22.13 6° 27′, 3.1×10^{-11} m
23.3 27, 26, 24
23.5 7.3×10^{-11} m, 1.43×10^{-10} m, 2.03×10^{-14} m
24.4 1.5×10^7 m/s, 6.9 μm
24.5 1.4×10^{-11} s^{-1}, 1600 a
24.8 56 s
24.9 75 s
24.11 2.55 min
24.13 1.61×10^4 s^{-1}
24.15 $\lambda = 0.0134$ s^{-1}, $T = 51.8$ s
25.4 6.87×10^{-13} J; 6.75×10^{-13} J
25.5 3×10^{-3} kg, 2.7×10^{15} J

Index

Index

Page numbers in bold type indicate major text references

Absolute determination 77
  of current 400–404
  of resistance 404–406
  of r.m.s. a.c. 402
Absorption α, β, γ rays 527–530
Accelerators 126, 516, 527, 538, 541, 554, **557–564**
Acceptor atoms 260
Accumulator, lead 54, **105–106**
  nickel-iron (NIFE) 106–107
Aerials 449
α particles 448–490, **506–533**, 538–541
  charge 510–511
  charge to mass ratio 510
  detection 506–508, 519–527
  identification 508–512
  as projectiles 488–490, 506, 538–541
Alternating current 173–177, **209–234**, 435–437
  measurement 402, 417–419
  generators 341–342, 431–432
  oscilloscope demonstrations 243–250
Ammeter 47–48, **410–419**
  calibration 81
  shunt 43–45
Ampère, the unit of current **17–18**, 329, 361
  absolute determination 400–404
Ampère's circuital equation 363–364
Ampère's experiments in electromagnetism 373
Amplification factor of transistor 286
Amplifier, d.c. 196, **295**, **420**, 459, 472, 476, 520, 531
Amplifying 276, **281–184**, **294–295**
Amplitude modulation 451–452
Antineutrino 515, **556–557**
Armature, motor 426–431
Atom models 13, 490, **491–494**
Atomic bomb 548
Atomic mass, constant (unified) m_u 549
Atomic mass number (nucleon number) A, 14, 512, 513, 516, 540, **549**
Atomic mass, relative (atomic weight) A_r **97–98**, 494, **549**
Atomic number Z 14, **494–495**, 512, 516, 540, **549**
Auger showers 556
Avogadro constant N_A **98**, 466, 483

B, *see* 'magnetic flux density'
Back e.m.f. **24**, 100–103, 152–153, 426
Background radiation 525, 529
Ballistic meter **184–185**, **343–344**, 383–384, 388–389
  theory of 414–417
Band, conduction and valency 500–503
Barrier, potential 270–272, 276
β particles 488, **506–530**
  detection 506–508, 519–520, 521–527
  identification 508–510
Binding energy **543–544**, 550–551
Biot-Savart equation 360–363, 366–367
Biot's hemispheres 125
Bohr's atom model 13, **491–494**
Boltzmann constant 454
Bottoming, transistor 287
Bragg equation 482

Crystal lattice 15, 253–257, 480–483
Curie, the 519
Curie temperature 392, 394
Current, absolute determination 400–404
  related to moving charges 16, 28–29, 40–42, 99–100
Current balance 402–404
Cyclotron 516, **560–562**

Damping of meter movements 413, 419
Dangers from radioactive materials 518–519
  from x-rays 479–480
Decay constant, radioactive 513–514
Depletion layer **270–272**, **277–280**, 527
Detection of radiowaves 451
Detectors, radiation 506–508, 519–523, 525–527
Diamagnetism 146, 393
Dielectric **130–135**, 148, 315–317
Differentiating circuit 191–192

Diffraction, of electrons 497
of X-rays 480
Diffusion cloud chamber 507–508
Diode 269–276
Dip, angle of 144
Discharge tube 458–462
Donor atoms 259
Doping 259
Dynamo, a.c. 341–342, **431–432**
d.c. 430–431
Dynamometer 400–402

F, *see* 'electric field strength'
Earth's magnetic field 144
Eddy currents 156–158, 428, 435, 438
Einstein, equation for photoemission 474
special theory of relativity 474
Electric field strength, *E* 110, **120–122**, 151–152, **300–315**, 317–320
definition of 110, 300–301
unit of 121, 301
Electrode potential 102–3
Electrolysis 96–103
Electrolyte 96
Electromagnet 391
Electromagnetic force, unrationalized constant, β, 17–18, 329, 358
Electromagnetic induction 152–160
laws 155
derivation 336–338
experimental evidence 349, 405–406
Electromagnetic waves 361, 373–374, **444–452**, 478, 490–491
Electrometer, attracted disc 313–314
pulse, **195–196**, 520–521, 523, 527–528
Electrometer valve 420, 459, 476, 531
Electromotive force, e.m.f. **24**, **54**
of cells 100–101, 102–103
comparison of 77–78
induced **152–155**, **336–338**, 424, 426, 431–434
polarization 100–101
standard 77, 104, 107
thermoelectric 263
Electron 11–12, 13–16
acceleration in a vacuum 27
in the atom 13–15, 490–500
charge of 466–467, 483
charge to mass ratio 462–466
diffraction 497–498
discovery 462–465
photoemission 472–473
secondary emission 525
thermionic emission 274–276
wavelength 497–498
Electron forces described 11
Electron gun 241
Electron lens 241–242
Electron-volt 489
Electrophorus 119–120
Electroscope 20–21, **111–117**, **123–126**, **129–130**, 506, 520, 554
as a voltmeter 116–117
to measure a small current 194–196
Electrostatic force, unrationalized constant α 12, 18, 299–300
Energy 20–27
binding 543–544, 550–551
equivalence to mass 455, 496, 542–543, 555–556
of photon 496
nuclear 538–551
unit 21, 489
Energy bands 499–503
Energy levels in the atom 489, 492–494, 498–503
Equipotential surfaces 115–116
Exclusion principle—Pauli 498–500
Extrinsic semiconductor 258–260, 502

Farad, the unit of capacitance 129
Faraday constant, *F* 98–99, 464
Faraday dark space 461
Faraday's disc 155–256, 405 340
Faraday's ice-pail experiments 111–114
Faraday's law of electromagnetic induction 155
Faraday's laws of electrolysis 97
Feedback, negative 290
positive 277, 291
Ferrimagnetism 395
Ferrites 146, 158, **391–392**, 395, 450
Ferromagnetic materials 142, 146–147, **383–391**, 394–395
Fission, nuclear 516, **543**, **544–548**, 556
Force, electric 11
magnetic 16–17, 142
magnetic and electric forces compared 329 flyback 245
relativity aspect 455–457
Fourier's theorem 212
Frequency of a.c. 210, 212
measurement by oscilloscope 244–245
Frequency modulation 452
Fuel cell 107
Fusion, nuclear 544, **548–549**, 556

G, the constant of gravitation 12
Galvanometer 410–414
γ rays **508–509**, 513–515, 518–523, 527–530, 555–556
detection 519–623, 527
identification 508–509
wavelength 446
Gas multiplication of ions 521

Gauss's theorem 303–305, 321–322
Gauss, the c.g.s. unit of flux density 332
Geiger-Müller tube 510, **521–3**, 529, 541
Geiger region 523
Generator a.c. 341–342, 431–432
  d.c. 430–431
Germanium 253, 256–260, 502
Guard ring 311

Half-life **513–518**, **531–533** 553
  of a neutron 515
Hall effect 345–348
  coefficient 347
Hard X-rays 479
Heat sink 273
Heating effect of current, irreversible **33**, **35**, 210, 414–415
  reversible 262, 266–267
Helmholtz arrangement of coils **369–370**, 402–403, 463
Henry, the unit of inductance 157–158
Hertz, the unit of frequency 210
Holes, charge carrying **257–260**, 268, 269–274, 276–291, 476–478, 501–502
Hysteresis, magnetic **388–392**, 435

Ice-pail experiments 111–114
Impedance 224–225
Induced charge 111, 117–120
Induced magnetism 142–143
Inductance, mutual 157–158, 371–373
  self 158, **164–166**, 168–175, 371–373
Induction coil 437–438
Induction, magnetic *B*, *see* 'magnetic flux density'
Inductor 158–159, 164–175
Insulator 130, 194, 254, 501
Integrating circuit 191–192, 524–525
Internal resistance 54–56
Intrinsic semiconductor 256–258, 501–502
Inverse square law of force, current elements 373
  electric charges 12, 299, 319–320, 490
  gravitation 12
  magnetic poles 376
Inverse square law, radiation intensity 530
Ionic bonding 96, 254–255
Ionization chamber 506, 510, **520–521**, 527–528, 531
Ionization potential 494, 499
Ions 14, **96–107**, **458–462**, 468–469
Isotopes 516–517

Joule, the unit of energy 21
Joule's law 35

Kirchoff's laws 60–62, 87–89

Law patterns 480–481
Lead accumulator 105–106
Leclanché cell 103–104
Left-hand rule **150**, 331–332
Lenz's law **155–156**, 158, 338
Light 445–446
Lightning conductor 126
Linear accelerators 558–560
Linear timebase 245
Lines of force, electric **110–114**, 135, 148, **303–305**, 387, 447–448
Lines of magnetic flux 16, **143–148**, 387, 447–449
  compared with electric lines of force 148, 387
Lines, mass spectrum 469
Lines, optical spectrum 491–494
Lines, X-ray spectrum 483–484, 494–495
Lissajous figures 244–245
Lorenz absolute determination of resistance 405–407
Loudspeaker 423

Magnet, permanent 142–143, 147, 374–377, 390–391
Magnetic circuit 396–397
Magnetic field 16–17, 142–147, 150–151
Magnetic flux Φ **151**, 154–156, 337, 339, 342–343
  definition 151, 332
Magnetic flux density or induction, *B* 150–151, 330–349, 358–377, 383–386
  definition 150–151, 330–331
  measurement 342–347
  unit of 331
Magnetic materials 383–397
  antiferromagnetic 395
  diamagnetic 146, 393
  ferrimagnetic 146, 395, 450
  ferromagnetic 142, 147, 383–391, 394–395
  paramagnetic 146, 393–395
Magnetic poles 142–143, 374–376
Magnetic pole strength 374–375
Magnetic reluctance 396–397
Magnetizing force *H* 386
Magnetometers, Hall 345–347
  vibration 344–345
Magnetomotive force 396
Mass, equivalent to energy 455, 496, 542–543, 555–556
  increase with velocity **454**, 561–563
Mass spectrograph **468–469**, 516, 550
Matched load 56–57
Maxwell's electromagnetic field equations 445–447
Maxwell's method of measuring capacitance 185–186

Maxwell's screw rule 14
Mesons 554
Microphones 423–425
Millikan's oil drop experiment 446–468
Moderator in a nuclear reactor 546–548
Modulation of radio waves 451–452
Mole, definition 98
Moment, electromagnetic, of a coil 335
  of a magnet 336, 374
Momentum of a photon 496–497
Motor effect 149–150, 331–336, 423, 426
Motors, electric 426, 430
Mutual inductance M 157, 371–373
Mutual induction 157, 160, 432

n type semiconductors **259–260**, 267–269, 277–278, **502**
Neutrino 554, 556–7
Neutrons 13–14, 512–517, **538–541**
  detection 541
  discovery 539–541
Nickel-iron (NIFE) cell 106–107
Nuclear explosion 517, 548
Nuclear fission 516, **543**, **544–548**, 556
Nuclear force 14, 513, 543
Nuclear fusion **544**, **548–549**, 556
Nuclear number, or atomic mass number A 14, 512, 513, 516, 540, **549**
Nuclear reactions 542–543
Nuclear reactors 547–548
Nuclear transformations 512–516, 538
Nucleus of the atom 13, 488–491

Ohm, absolute determination 405–407
  definition 33
Ohm-meter 48–49
Ohm's law 35, **40–42**
Orbital 254–255, 498–499
Oscillatory circuit 171–173, 193, 227, 277, 291
Oscilloscope 240–250

p-type semiconductors **260**, 267–269, 277–278, **502**
Pair production and annihilation 555–556
Parallel plates, capacitance of 130, **310–311**
  force of attraction 312, 313
Paramagnetic materials 146, 393–396
Pauli's exclusion principle 498–500
Peak value of a.c. **210**, **212–214**, 233–234
  measurement 243, 246–247
Peltier effect 265–268
Periodic table 494–495, 516
Permanent magnet 142–143, 147, 374–377, 390–391
Permeability of a vacuum, μ_0 358, **360–361**
  dimensions of 373–374
Permeability, relative, μ_r **384–386**, 392–394
Permittivity, relative ϵ_r 130–131
  measurement 189
Permittivity of a vacuum, ϵ_0 **130**, **299–300**
  compared with μ_0 361
  dimensions of 373–374
  measurement 320–321
Phase angle 222–223
  measurement by oscilloscope 248–250
Photocells 260, **472–473**, **476–478**
Photoconduction **476**, 502–503
Photoemission of electrons **472–476**, 502–503, 526
Photographic emulsion 527
Photomultiplier 526
Photon **474–478**, **493–494**, 508, 520
  momentum 496–497
Planck constant h 474–479
  measurement 475–476, 478–479
Plutonium 547–548
Point discharge 126
Polarization of a cell or voltameter 100–103
  of a dielectric 134–135
  of an electromagnetic wave 445–446
Pole, magnetic 142–143, 374–376
Pole strength 374–375
Positive rays (canal rays) **468**, 516
Positron (positive electron) 516–517, **554–556**
Potential, electric 20–23, 115–117, 123, 300–315
  definition 23, 301
Potential barrier, at n-p junction 270–272
  of a nucleus 535
Potential difference, p.d. **20–27**, 57–62
Potential drop in cables 49–52
  in cells 52–54
Potential gradient **120–123**, 301
Potentiometer 77–85, 87–88
Power **23–25**, 33
  in a.c. circuits 210, 215–218, 224–225
Power loss in cables 49–52
  in cells 52–54
Preferential discharge 100
Proton 13–14, 512–516, 538–541
Pulse electrometer **195–196**, 520–521, 527–528

Quantum theory **473–476**, 478, **492–503**
Quenching agent 522–523

Radiation, electromagnetic 444–452, 478, 490–491
Radiation field 444–445, 447
Radio 444–452

Radio reception and transmission 449–451
Radioactive isotopes 517–518
Radioactive radiations **508–511**, 527–530
  background radiation **525**, 529
  radiation detectors 506–508, 519, 520–523
  radiation hazards 518–519
Radioactive series 514, 516
Radioactive tracers 518
Radioactivity 506–533, 555
  discovery 488
Ratemeter **524–525**, 529
Rationalized equations, in electromagnetism 358–361
  in electrostatics 299–300
Reactance 217–218
Reactors, nuclear 547–548
Rectification of alternating current 173–177
Rectifiers 269
Relativity **452–457**, 561
Reluctance, magnetic 396–397
Remanence 390–391
Resistance 33–62
  definition 33
  of electrolyte 102
  internal 54–56
Resistance measurement, absolute determination 404–406
  by bridge 85–87
  by potentiometer 85
  of very high resistance 194–198
Resistance transformer 435
Resistance with a.c. 210, 214–215, 231
Resistivity 38, 253
Resistor 33
  load 282–283, 288
  standard **81–82**, 158–159, 407
  starting 430
Resonance, electrical 227–228
Rest mass 454–455
Right-hand rule 152–153
Röntgen, the 519
Root mean square value 210–214

SI Units 18
Scaler 523–524, 529
Scintillation counter 526
Search coil 342–343
Seebeck effect 262–265
Self inductance L 158, 371–373
Semiconductors 253, 501–502
  extrinsic and intrinsic 256–260, 501–502
  Hall effect in 345–348
  photoelectric effects 476–478
  thermoelectric effects 267–268
Shells, electron **14**, 96, **254–256**, 483–484, 494–495, **498–500**, 538
Silicon 253, 256–257, 502
Smoothing devices 173–177
Soft X-rays 479
Solar cell 477–478
Solenoid 146, 358, 368–369
Space charge 276
Spark counter 521
Spectrograph, mass **468–469**, 516, 550
  X-ray 483, 494
Spectrum, electromagnetic 446
Spinthariscope **525–526**, 539, 542
Standard cell 77, 104–105
Standard resistor 77, 158
Strontium-90 517
Superconductivity 439–441
Synchronous acceleration 558, 560
Synchrotron 563–564

Telephone receiver 424
Temperature, Curie 392, 394
  transition 440–441
Temperature, variation of resistance with 38–39, 42, 102
Temperature coefficient of resistance 39
Tesla, the unit of magnetic flux density 331
Thermistor 40, **261–262**
Thermocouples 262–268
Thermoelectric effects 262–268
Thermionic emission 274–276, 292
Thomson effect 266–268
Transformer 159–160, 391, **432–437**
Transistors 176–291
Transit time 295
Transuranic elements 516, 545, 548
Triode, thermionic 276, 292–295

Unified atomic mass constant, m_u 549–550

Valency 14, 96, 254–259, 499–500
Valency bond 500–503
Valve, electrometer 420, 459, 476, 531
Van der Graff generator **126–129**, 541, 557–558
Van der Waal's forces 256
Vector treatment of a.c. 220–228, 233–234, 437
Velocity of cathode rays 464, 465
Velocity of charge carriers **27–29**, **40–42**, 99–100, 261, 295, 345–347
Velocity of electromagnetic waves 361, 373–374, 445, 452–456
Velocity selector 468–469
Vibration magnetometer 344–345
Volt, the unit of potential **22**, 404–405
Voltameter 97, 100–102

Voltmeter, conversion from microammeter 45–48
Voltmeter, electrostatic 116–117, 313–314, 420
Voltmeter calibration 81

Watt, the unit of power 21, 23–24
Wave, electromagnetic 361, 373–374, **444–452**, 478, 490–491
Waveform of a.c. 209–210, 245
Waveguide 560
Wavelength of a particle 497–498
Wave mechanics **495–498**, 540, 542
Weber, the unit of magnetic flux 332
Weston cadmium cell 104–105
Wheatstone's bridge 85–89
Wilson's cloud chamber 506–507
Wimshurst machine 120
Work function **474–476**, 503

X-ray diffraction 480
  spectrum 478–479, 483–484, 494–495
  spectrum lines 483–484, 494–495
X-rays 478–484
  hard and soft, 479
  hazards 479–480
X-ray spectrograph 483, 494
  spectrometer 481–482
X-ray tube 478, 560

Z, atomic number **14**, **494–495**, 512, 516, 540, 549

Macmillan

SBN 333 12799 4