The Physical Basis of Electronics

AN INTRODUCTORY COURSE

SECOND EDITION

D. J. HARRIS

Professor of Electrical Engineering, Portsmouth Polytechnic

AND

P. N. ROBSON

Professor of Electrical Engineering, University of Sheffield

PERGAMON PRESS

Oxford · New York · Toronto · Sydney

Pergamon Press Ltd., Headington Hill Hall, Oxford
Pergamon Press Inc., Maxwell House, Fairview Park, Elmsford, New York 10523
Pergamon of Canada Ltd., 207 Queen's Quay West, Toronto 1
Pergamon Press (Aust.) Pty. Ltd., 19a Boundary Street,
Rushcutters Bay, N.S.W. 2011, Australia

Copyright © 1974 D. J. Harris and P. N. Robson

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of Pergamon Press Ltd.

First edition 1963

Second edition 1974

Library of Congress Cataloging in Publication Data

Harris, Douglas James.
The physical basis of electronics.

(Applied electricity and electronics)
Published in 1963 under title: Vacuum and solid state electronics.
Bibliography: p.
1. Semiconductors. 2. Electron tubes. 3. Electronics. I. Robson, Peter Neville, joint author.
II. Title.
TK7871.85.H385 1974 621.381 74-996
ISBN 0-08-017900-2
ISBN 0-08-017901-0 (flexicover)

Printed in Hungary

Preface

THE knowledge that many students have found our earlier book *Vacuum and Solid State Electronics* of value in forming an understanding of the physical basis of modern electronics has encouraged us to revise and update the text for re-publication. The objective remains the same, namely to introduce the physical concepts basic to an understanding of electronic devices, and to apply them to the main types of device in current use. The mathematics introduced is kept to a minimum, but consistent with providing a sound foundation upon which specialist knowledge can be built.

In the ten years since the original book was first published there have been considerable advances in electronic techniques, including the introduction of field effect transistors, the increase in performance of transistors especially for use at very high frequencies, the increasing use of digital techniques, and the development of integrated circuits. The latter in particular has brought its own revolution in vastly increasing the complexity and sophistication of circuit utilization and is changing the whole approach to electronics. But in spite of the changes and advances in electronics, the physical basis on which the subject rests is largely unaltered, and the main body of the text of the original book remains relevant as an introduction to the subject.

Whilst solid state devices have largely ousted thermionic devices from conventional electronic applications, there are still areas where they have special advantages and are in widespread use, and a great deal of equipment using them is currently available. It does not seem wise at this stage to dispense with considering them in an electronics course and the thermionic device content is retained in this book, although forming now a smaller part of the whole. Because of this change of emphasis we have altered the title of the book to one which more aptly reflects its contents.

It is our hope that physics and engineering students in Universities, Polytechnics, and Technical Colleges will continue to derive help from our book, and that it might prove valuable too for any other person seeking understanding of this subject.

Formulae are generally expressed in the S.I. system of units. However, values of certain physical quantities are quoted in alternative units when this is common usage. The conversion to S.I. units in such cases is straightforward. A series of problems with solutions is included.

1. *Introduction to Electronics*

1.1. THE IMPACT OF ELECTRONICS

THE growth of the field of electronics has been a very rapid one. As in the case of many scientific developments, the starting point was a somewhat accidental discovery. In 1883 Thomas Edison was investigating a different phenomenon—the emission of light by a hot carbon filament in a vacuum. From his work followed the observation that a current was able to flow through the evacuated region between a hot filament and another conductor introduced into the vacuum and at a positive potential with respect to the hot filament. From this beginning has sprung the immense electronics industry which in the United States alone produces more than 1000 million electron devices per year. During the decade 1950–1960 a significant change came about with the development of solid state electron devices, of which the most well known is the transistor. Transistors rapidly ousted the vacuum tube from its position of importance in a wide range of applications, and now demand priority of treatment in any consideration of electronic devices.

The development of electronics has produced a social revolution, having resulted in a system of mass communication by both sound and vision. The benefits deriving from television may be debatable, but there is no doubt that the impact it has made in the more industrialized parts of the world—and indeed in many of the less developed parts—is very great. There are, of course, many other important applications of electron devices apart from domestic radio and television receivers, such as world-wide communication systems, a vast range of measuring equipment (there must be very few quantities that have not been measured by electronic methods!), medical electronics, computers, and the expanding realm of automation with its need for measurement and control, to mention but a few.

1.2. THE HISTORY OF ELECTRONIC DEVICES

The early work posed a large number of important questions to scientists at the end of the nineteenth century. It was J. J. Thomson who showed that the transfer of electric charge corresponding to the current in the evacuated region was due to the motion of small particles, each having the same mass and a small negative electric charge. These minute particles were called electrons, and their mass and charge were later measured and found to be $9{\cdot}11\times10^{-31}$ kg and $1{\cdot}60\times10^{-19}$ coulomb respectively.

Probably the first device using the free motion of electrons in an evacuated region was the cathode-ray tube demonstrated by Braun in 1897, and this was followed by the thermionic diode used to detect radio waves by Fleming in 1904. The early experimenters working on the transmission of radio waves had need of a piece of apparatus that would conduct electricity in one direction only. Since in an evacuated bulb containing two electrodes, one of which is very hot, electrons can be emitted from the hot electrode only, and a current can therefore only flow from the cold electrode to the hot one (by convention a current flows in the direction of effective transfer of positive charge), such a device was suitable for the purpose of detection. In this way the diode, the first member of the thermionic valve family, was born. Two years after Fleming's diode, de Forrest showed that the current through a valve could be effectively controlled by placing an open wire structure (called a grid) between the two electrodes of the diode, and varying its potential relative to the other electrodes. The resultant valve, known as a triode, was shown to be capable of magnifying small alternating voltages, if connected into an appropriate circuit. This amplifier action made possible the rapid development of broadcasting and other forms of communication. The triode amplifier was found to have serious limitations when used at high frequencies, but these limitations were overcome by the addition of extra grid electrodes to the valve, resulting in the tetrode and pentode valves. These were subsequently found to be capable of much greater amplification than the triode. More complex valves with many electrodes were later developed to perform specialized operations such as the mixing of two signals of different frequency, or the generation of ultra-high frequencies. Television transmission and reception followed the development of high

performance thermionic valves, cathode-ray tubes, and photosensitive devices for television cameras.

A major change took place in the electronics world with the appearance of the point contact transistor reported by Bardeen and Brattain in 1948, closely followed by the junction transistor originated by Shockley. In these solid state devices, motion of electrons takes place in the lattice of a near-perfect crystal of a semiconductor, usually silicon or germanium. If the crystal has been carefully made with very few irregularities and a high degree of purity apart from any impurities deliberately introduced, any electrons which are not tightly bound to the lattice atoms are able to move in a fairly unrestricted manner between the rows of atoms in the crystal. These free electrons may have originated from the atoms of the pure crystal material, or more likely are the result of the introduction of a controlled amount of a particular impurity. The situation is more complex than that in a thermionic valve since electron vacancies, or "holes", can also migrate through the crystal lattice. The motion of the charge carriers in the crystal lattice can be controlled by the potentials applied at various points in the crystal, electrodes being placed in contact with the crystal for this purpose. It is also possible for potentials to be applied to a thin layer having different semiconductor properties, introduced into the single crystal, as in a junction transistor. Early transistors could be used only at low frequencies to give amplification, but the maximum frequency of operation was rapidly increased to hundreds of megaherz, and amplification at a few gigaherz can now be achieved. Diodes made of silicon and germanium have also been made for rectification, and units capable of passing a current of hundreds of amperes are now available.

The main disadvantages of the thermionic valve are that a very good vacuum has to be maintained in the valve during the whole of its life, the emitting surface has to be raised to a high temperature during operation—requiring considerable power for heating the cathode, and voltages of a hundred volts or so are needed for effective operation. The advantages of the semiconductor devices are that no vacuum and no thermionic emitter are required, operation is at voltages of about ten volts, and the physical size is small. Moreover high reliability and long life make the transistor-type amplifier additionally attractive, and it has supplanted the thermionic

valve for most applications. Nevertheless, there are still applications in which thermionic valves have to be used.

1.3. BASIC ELEMENTS OF ELECTRONIC DEVICES

Despite the vast difference in physical appearance between thermionic and semiconductor valves, the basic elements are somewhat similar. The first requirement is a region through which electrons can move in a fairly unimpeded manner. This is achieved in the thermionic valve by almost completely removing any matter that might come in the way of the electron, i.e. by evacuating all the air from the valve envelope. This can be done so effectively that when the pressure is reduced to about a thousandth of a millionth of an atmosphere, the chance of a collision between an electron and a gas molecule is very remote. The solid state approach is to make a crystal so perfect, with the atoms lined up in a regular array of rows and columns, that the electrons can move readily between them. This requires very advanced manufacturing techniques. The second requirement is a source of moving charged particles. These are obtained in a thermionic valve by heating an appropriate cathode material to a high temperature, free electrons being liberated into the evacuated region. In a semiconductor the charge carriers are usually produced in the crystal itself by disturbing the lattice with a special impurity. Both types of electronic device are made to function by the controlled motion of charged particles in them, resulting from the application of electric fields. Conducting electrodes, either in the vacuum envelope or on the crystal surface, allow the fields to be set up when potentials are applied to these electrodes. They also enable currents to flow into the device from the external circuit.

Thus in spite of their dissimilar appearance it can be seen that both types of device consist of an environment in which charged particles are relatively free to move, a source of charged particles to move in this environment, and an electrode system to enable the charged particle flow to be controlled and currents to be taken into and out of the device.

The similarity between vacuum and solid state devices has been emphasized by the development of the field effect transistor or FET. Thermionic

valves are basically voltage-controlled devices with a high input impedance, i.e. they are controlled by an applied voltage but take little current from the control. Transistors, however, are basically current-controlled devices with a relatively low input impedance, and there is some interaction back from the output to the input, unlike the valve. So the transistor is inherently more complex as a circuit element than the valve. The FET, whilst a semiconductor device, is similar in behaviour to the valve, being also voltage controlled with a high input impedance. It is finding increasing use in electronics, and lends itself particularly well to integrated circuit techniques.

1.4. INTEGRATION OF ELECTRONIC DEVICES AND CIRCUITS

The increasing complexity of electronic circuits and developments in semiconductor technology has led to the production of complete circuits, including active and passive devices, on a single slice or "chip" of semiconductor. Resistors, capacitors, transistors and conductors can be fabricated on a minute scale using photographic and optical reduction techniques allied to carefully controlled manufacturing methods in which the electrical characteristics of microscopic areas and layers of the semiconductor chip can be changed to produce the required effects. In this way a complex array of circuit elements, even several hundred in number, can be produced on a single chip a fraction of an inch in dimension. This has made available a wide range of complete miniature circuits, e.g. amplifiers, which can be used directly as elements in a more complicated system. The dramatic reduction in size has been of particular value for applications such as sophisticated computers and pocket-size electronic calculators. Electronic engineers now think in terms of assembling standard circuit modules whenever these are available, to form the overall system required. Circuits need be constructed from the individual circuit elements only when a suitable circuit module is not available.

2. *Fundamentals of Vacuum Electronics*

"VACUUM ELECTRONICS" is the name given to the branch of science relating to devices in which successful operation depends upon the motion of electrons in an evacuated region, or in a very low pressure gas. A source of electrons such as a thermionic emitter is usually included to produce free electrons, and the motion of these electrons under the action of electric and perhaps magnetic fields determines the subsequent behaviour of the device. The concept of the electron in which it is considered as a particle of given mass ($9{\cdot}11\times10^{-31}$ kg) and negative electric charge ($1{\cdot}60\times10^{-19}$ coulomb) is adequate to describe its behaviour under most conditions. There are other phenomena which require a quite different concept for their explanation, in which the electron is considered to have a wave nature. Thus electrons are considered to have a dual nature in order to explain their observed behaviour, and it has to be admitted that our understanding of them is far from complete. This should introduce an element of humility into our scientific thinking from the outset! However, the particle concept is adequate to explain most phenomena that will be discussed in this book.

In order that the characteristics of individual electron tubes or valves can be discussed, it will be necessary to consider the basic processes responsible for their behaviour. These processes include the emission of electrons from solid materials, the motion of these emitted electrons under the influence of applied electric and magnetic fields, the effect of the low pressure environment on the motion of the electrons through it, and the mutual repulsion effects between electrons due to their all having a negative electric charge. The expressions for the force on charges due to applied fields and other charges are also used to determine charge motion in a semiconductor crystal. Before describing these processes the nature of the low pressure gas environment will be considered.

2.1. THE EVACUATED REGION

The electrodes of electronic valves are contained within a gas-tight envelope of glass or metal, from which nearly all of the air has been removed. Electrons are then able to move in a region where the gas pressure is exceedingly low. The behaviour of a gas is described by the kinetic theory of gases. The gas is pictured as consisting of an exceedingly large number of very small particles or molecules, these being the smallest particles of the gaseous material that can exist whilst still having the characteristics of the particular gas. The number of these molecules in a volume is obtained from Avogadro's number (see Appendix 1), and corresponds to $2{\cdot}7\times10^{25}$ molecules per m^3, at atmospheric pressure and temperature. This number is the same for all gases. The magnitude of this number can be visualized by an example. If an electric lamp bulb is full of air at atmospheric pressure and the molecules could be removed from it at a rate of a million per second it would take over 100 million years to empty the bulb! In spite of the astronomical number involved, the volume of each molecule is so small that less than one part in 1000 of the volume is actually occupied by matter, at atmospheric pressure. The remainder is void. The molecules are not at rest but are moving at high speed with a velocity approximately equal to the speed of sound in the gas. The directions of motion at any instant are completely random, with equal numbers moving in opposing directions so that there is no net drift of gas. The speed of motion is related to the temperature of the gas, increasing when the temperature is raised. Molecules are constantly colliding with each other and with the walls of the containing vessel. When an electron passes through the gas it too has collisions with gas molecules, and in order to reduce the number of these collisions and produce relatively free motion of the electrons in the gas, the gas pressure has to be reduced to a very low value. Pressures of about 10^{-7} mm Hg (one atmosphere pressure is equivalent to 760 mm Hg) are normal in conventional valves. The possibility of a collision between a particular electron and a gas molecule is then remote, but there are still more than 10^{15} molecules per m^3!

2.2. THE EMISSION OF ELECTRONS FROM A SURFACE

Emission of electrons can take place from many materials which are conductors of electricity, and it is the conduction electrons which are caused to leave the surface. In an electrical conductor each atom contributes an electron which is loosely bound to the parent nucleus, and when under the influence of the surrounding nuclei, this electron becomes free to migrate through the block of conductor. This will be considered in greater detail in Chapter 3. The conduction electrons cannot normally escape from the surface of the conductor at room temperature. A force is exerted on them whenever they tend to move away from the surface, drawing them back to the main body of the conductor. This restraining force has two components, one due to the attractive force of the surface nuclei and one due to the image charge force. When a conduction electron is situated well within the

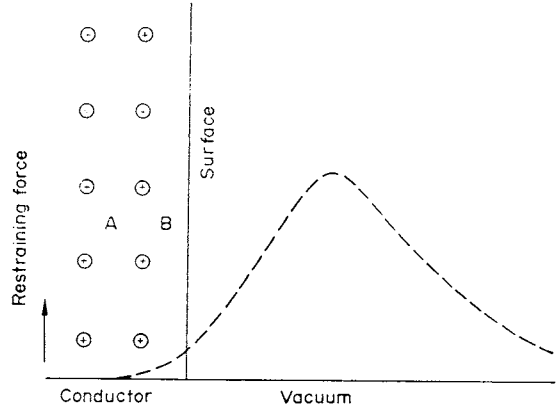


FIG. 2.1. Variation of force restraining an electron from leaving a conductor surface.

conductor, e.g. as at *A* in Fig. 2.1, it is surrounded by nuclei and the net force tends to be small. However, if an electron is close to the surface as at *B* in the diagram, the attraction forces due to the nuclei are in one direction only, tending to stop its emission from the block. Moreover, if an electron

succeeds in leaving the surface there will be the usual image charge force of attraction between the electron and its positive image in the metal. Both forces decrease with distance from the surface, the total variation being somewhat as shown in Fig. 2.1. If an electron in the conductor can be given sufficient energy it will be able to escape from the surface. The minimum energy required per unit charge to enable electrons to escape is known as the work function of the material. It is denoted by ϕ and is measured in volts, being essentially the magnitude of the potential barrier that must be surmounted by emitted electrons. The energy required by each electron for emission is then $e\phi$ joules. A material with a low work function will obviously emit electrons much more readily than one with a high work function. Values of work function vary from 1·8 V for caesium to 6·0 V for platinum.

There are several ways in which electrons in a conductor can be given energy to enable them to overcome the surface potential barrier and escape from the surface. The four main processes are thermionic emission, secondary emission, photo-electric emission and field emission. In thermionic emission, thermal energy is given to the emitting material by heating it to a high temperature. Energy is transferred to the electrons, and if the energy transfer to an electron is sufficient, emission takes place. Although high temperatures are needed, this is the most commonly used emission method. Secondary emission of electrons results from bombarding the surface with high energy particles such as electrons or positive ions (gas molecules that have lost an electron and are therefore left with a net positive charge). Energy can be transferred from the incoming particle to one or more electrons in the surface of the material being bombarded. If any of these electrons acquire sufficient energy to enable them to overcome the retarding force at the surface, they will escape from the surface. More than one electron may be emitted for each incoming particle if the energy of this incident particle is sufficiently high. The ratio of the number of emitted electrons to incident particles is known as the secondary emission coefficient δ. Devices such as the electron-multiplier tube make use of this secondary emission phenomenon. Electrons emitted from a surface by photo-electric emission have had their energy increased by the absorption of radiation falling upon the surface. This radiation may be visible light or radiation in the invisible

part of the electromagnetic spectrum, such as ultra-violet radiation or X- or γ-rays. Many photo-electric cells rely upon photo-electric emission for their operation. Field emission relies upon the presence of a very high electric field at the surface from which emission is to be obtained. A force is then exerted on the electrons in the surface tending to pull them out of the surface if the polarity of the field is of the right sign. The very high field required—of the order of a million volts per cm—illustrates the magnitude of the forces normally keeping the electrons within the bulk of the material. Field emission becomes an important process if high voltages are to be maintained across a small gap, e.g. if a thousand volts is to be maintained across a gap of a millimetre, the electric field is a million volts per metre. Irregularities of a surface tend to intensify the electric field at the surface near the irregularity.

2.3. THERMIONIC EMISSION

We would expect the number of electrons emitted from a surface at a high temperature to depend upon the amount of energy given to the surface, i.e. upon the temperature of the surface, and to be dependent also upon the work function of the surface. The total possible emission current density from a surface has been analysed theoretically and is given by the Richardson–Dushman equation:

$$J = AT^2 \exp\left(\frac{-T_o}{T}\right) \tag{2.1}$$

where J is the current density in amperes per m^2, A is a constant, T is the absolute temperature in degrees Kelvin (i.e. °C+273), and T_o is equal to 11,600 ϕ. This equation has been verified experimentally, although there is some disagreement between the theoretical and experimental values of A. Practical values for the constant A to be inserted in equation (2.1) are about 6×10^5 for most pure metals. Emission graphs are usually shown as a plot between $1/T$ and $\log_{10}(J/T^2)$, and such characteristics for tungsten, thoriated tungsten, and oxide emitters are shown in Fig. 2.2. The straight-line characteristics verify the form of the emission equation. This equation allows the maximum value of the emission current at a particular tempera-

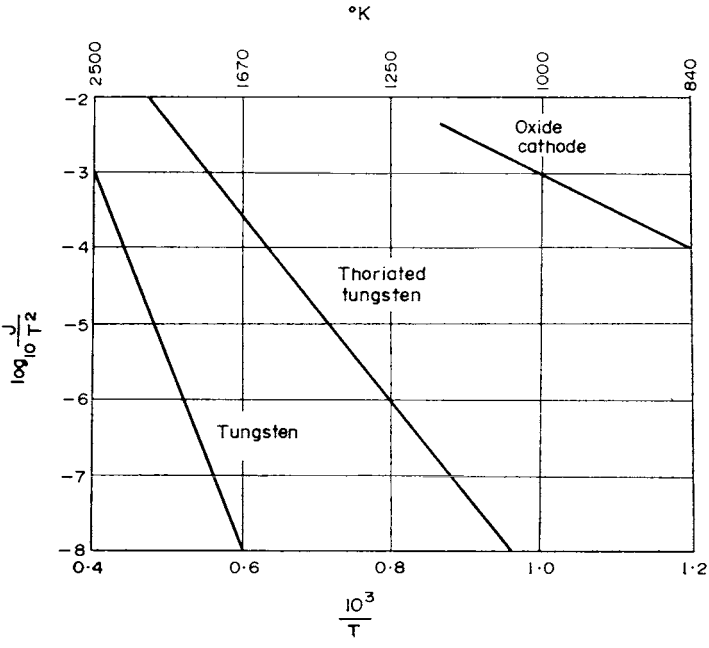


FIG. 2.2. Electron emission as a function of temperature, plotted to verify Richardson's equation.

ture to be calculated. However, all this electron current is not necessarily drawn off from the cathode as will be seen when the behaviour of the diode valve is considered.

2.3.1. *Practical Thermionic Cathodes*

In order to produce a good thermionic cathode, a material is required which has a low work function and therefore will give copious electrons at low temperature, and which has a long life under operating conditions. This latter requirement means that the emitter must be chemically stable and operate at a temperature well below its melting point. There are in fact very few suitable materials that satisfy these requirements, and the three cathode materials in common use in electron tubes are tungsten, thoriated

tungsten, and oxides of the rare earth metals usually known simply as oxide cathodes.

Tungsten is one of the few pure metals that can be used (tantalum is another), most pure metals needing a temperature around or above the melting temperature in order to give good emission. The main advantages of tungsten are that it is relatively robust and not easily damaged, and it can be heated directly by passing a current through it. Thus tungsten cathodes normally consist of a length of tungsten wire or ribbon. However, the work function of tungsten is rather high—about 4·5 V and the operating temperature for good emission is also high, being well over 2000°C. This results in a large heater power being required, and a large rise in temperature for the other components of the valve.

Thorium has a low work function and would make a good emitter except that it evaporates rapidly at the operating temperature. However, it forms a loosely bound chemical compound with tungsten which breaks up slowly at the emitting temperature of thorium. Cathodes using this compound are called thoriated tungsten cathodes. During operation the compound slowly releases the metal thorium within the volume of cathode material, and this thorium diffuses through the volume to the surface. Emission of electrons is from a thin film of thorium on the surface, the thorium being continually evaporated from the film but being replenished by that diffusing to the surface from the interior of the cathode material. The thoriated tungsten cathode operates at a temperature well below that for tungsten, but is not so robust.

The most widely used cathode materials are the rare earth metal oxides. These oxide cathodes consist of a mixture of barium and strontium oxides, which slowly decompose at the operating temperature of the cathode to produce metallic barium and strontium. A thin film of these metals is maintained on the cathode surface by diffusion through the bulk of the cathode material. The work function of the resultant film is very low—considerably less in fact than that of barium or strontium alone, and the operating temperature is only about 750°C. The main disadvantage of the oxide cathode is its chemical activity and the ease with which it disintegrates. In particular the oxide readily combines with water vapour to form a stable hydroxide, and it is therefore easily "poisoned". The oxide is very weak

mechanically and it is usually used in the form of a thin coating on a nickel tape heated by passing a current through it, or on a thin-walled nickel tube whose temperature is raised by a heating filament within it. These cathodes are known as directly and indirectly heated cathodes respectively. Oxide cathodes are made by coating the nickel base with a mixture of the carbonates of barium and strontium, and then reducing these carbonates to the oxides by heating under vacuum conditions. Once formed the oxides are poisoned if exposed to atmospheric conditions. Oxide cathodes are used almost exclusively for low-power valves. The properties of the three types of electron emitter are summarized in the table.

	Work function ϕ volts	Working temp. °C	Emission current A/m^2
Tungsten	4·5	2200	3×10^3
Thoriated tungsten	2·8	1400	3×10^4
Oxide	0·95	750	$2{\cdot}5\times10^3$

2.4. ELECTRON MOTION IN ELECTRIC AND MAGNETIC FIELDS

Once electrons have been emitted in an evacuated region, the behaviour of the particular device will depend upon the subsequent motion of the electrons under the action of applied electric and possibly magnetic fields. The laws relating to this motion will therefore need to be considered.

2.4.1. *Motion in an Electric Field*

The concept of an electric field at some point in space is essentially concerned with the force that would be exerted on a very small electrically charged particle placed at that point. The magnitude of the field is equal to the force exerted per unit charge, and the direction of the field is by convention the direction of the force acting on a positively charged particle.

This is in fact all that is meant by an "electric field". Thus the force acting on a particle of electric charge q coulombs if the electric field at the particle is E volts per metre, is given by the equation

$$F = qE \tag{2.2}$$

where F is the force in Newtons in the direction of the electric field. For an electron the charge q is symbolized by $-e$ and is equal to $-1{\cdot}60\times10^{-19}$ coulomb. The force on the electron is thus in the opposite direction to the electric field direction. There is some confusion of convention regarding the sign of e, since some texts put e as a positive number and include the negative sign in the equations, whilst others write e as a negative quantity. We shall here use the first convention, and in subsequent equations the symbol e must be replaced by $1{\cdot}60\times10^{-19}$ in making any calculations.

The motion of electrons under the action of forces is governed by the usual Newtonian laws of mechanics. The force on a body can be equated to its rate of change of momentum, i.e. $F = (\mathrm{d}/\mathrm{d}t)(mv)$, where m and v are the mass and velocity of the particle. For the case of non-relativistic velocities, the mass of the particle can be considered constant. This is not always the case in electronics since the velocity acquired by an electron in moving through a potential of tens of thousands of volts or more becomes comparable with the velocity of light and the effective mass of the moving electron is greater than that for the electron at rest. For lower voltages, however, the mass can be considered constant and the equation of motion for an electric field in the x direction becomes

$$F = -eE = m\frac{\mathrm{d}v}{\mathrm{d}t} = m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}. \tag{2.3}$$

The acceleration of the electron is therefore given by

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = \frac{-e}{m}E. \tag{2.4}$$

This equation can be integrated directly for the case of an electric field which is constant in time and space, and if the electron has a velocity u at

time $t = 0$, the velocity at time t is given by

$$\frac{dx}{dt} = \frac{-e}{m} Et + u. \tag{2.5}$$

A further integration gives the displacement as a function of time, and if $x = 0$ at time $t = 0$, i.e. if the origin of the coordinate system is the electron position at time $t = 0$, the displacement equation is obtained

$$x = \frac{-eE}{2m} t^2 + ut. \tag{2.6}$$

Elimination of t from equations (2.5) and (2.6) results in the equation

$$\frac{1}{2} m \left\{ \left(\frac{dx}{dt} \right)^2 - u^2 \right\} = -eEx. \tag{2.7}$$

The product $-Ex$ is the electric potential difference V through which the electron has moved, whilst the left-hand side of the equation represents the difference between the final and initial kinetic energy of the electron.

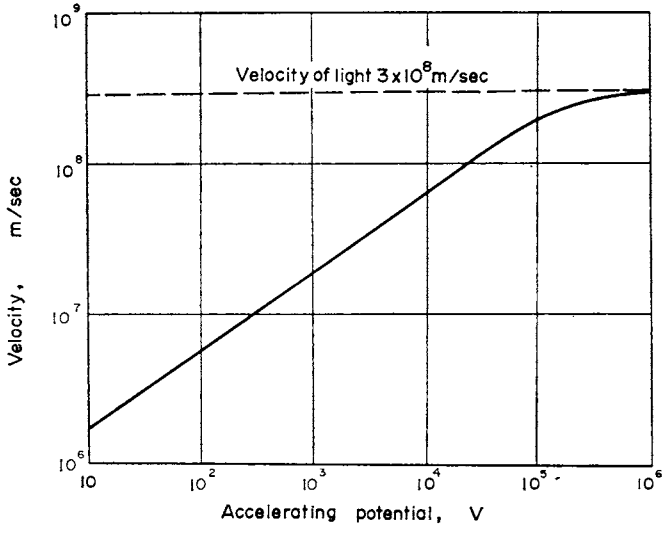


FIG. 2.3. Electron velocity as a function of accelerating potential.

Equation (2.7) therefore is an energy balance equation, equating the loss or gain of kinetic energy of the electron to its gain or loss of potential energy. If the electron starts at rest and moves through a potential difference of V volts, the final velocity is given by

$$v = \sqrt{\left(\frac{2eV}{m}\right)}. \tag{2.8}$$

Electron velocities corresponding to a range of voltages up to 100 kV are shown in Fig. 2.3.

2.4.2. *Motion in a Magnetic Field*

It can be readily verified experimentally that the force F exerted per unit length on a conductor carrying a current I in a magnetic field of flux density B, when the angle between the conductor and magnetic field is θ, is given by the equation $F = IB \sin \theta$. The direction of the force is mutually perpendicular to both current and field directions. If I and B are perpendicular, I, B, and F directions correspond to the x, y, and z directions of a set of right-handed coordinates as in Fig. 2.4 (a). A current corresponds simply to electric charge in motion, the magnitude of the current being equal to the

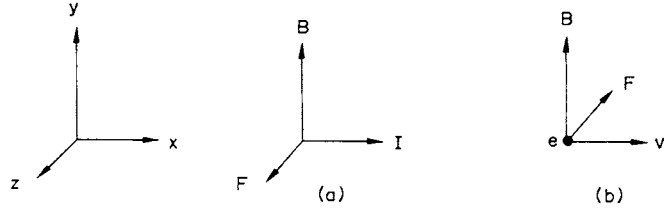


FIG. 2.4. Forces on (a) a current-carrying conductor, and (b) a moving electron in a transverse magnetic field.

rate of transfer of charge. When an electron is in motion electric charge is being transferred from one place to another, and this is equivalent to an electric current. There is thus a force on an electron if it is moving in a magnetic field with a component of the field perpendicular to the direction of motion. When a particle of charge q is moving with a velocity v, the force

exerted by a magnetic field is given by $F = qvB \sin \theta$, where θ is the angle between field and velocity directions. For an electron $F = -evB \sin \theta$. The force direction on a moving electron in a perpendicular magnetic field is shown in Fig. 2.4 (b). The reason why a force is exerted on the current-carrying conductor in a magnetic field is simply that a force is exerted on each electron moving through the conductor and this force is transferred to the conductor as a whole. The equation

$$F = -evB \sin \theta \tag{2.9}$$

is the fundamental relation from which the force equation for a current-carrying conductor in a magnetic field can be derived. No force is exerted on an electron moving in the direction of the magnetic field.

Since the force exerted on the electron by a magnetic field is always perpendicular to its instantaneous velocity, and has no component in the direction of motion, a static magnetic field can change the direction of motion of the electron but not its speed. The kinetic energy of the electron remains unchanged, and the electron will move on an arc of a circle. There are many other familiar examples in which a force acts perpendicular to the direction of motion and the moving body describes a circle, such as a stone on the end of a string, or a satellite circling the earth. The radius of gyration of the electron is obtained by equating the centrifugal force and the force due to the magnetic field, i.e. by writing

$$\frac{mv^2}{r} = evB, \quad \text{giving} \quad r = \frac{mv}{eB}. \tag{2.10}$$

The time per revolution $\tau = \dfrac{2\pi r}{v} = \dfrac{2\pi m}{eB}$ seconds (2.11)

and the angular velocity $\omega_c = \dfrac{2\pi}{\tau} = \dfrac{eB}{m}$ radians per second. (2.12)

It can be seen that both τ and ω_c are independent of the velocity of the electron, and depend only on the magnitude of the magnetic field. The parameter ω_c is an important parameter in some very high frequency valves and is known as the cyclotron (or magnetron) frequency. The value of $f_c = \omega_c/2\pi$ is 2·8 MHz per gauss, or $2{\cdot}8 \times 10^{10}$ Hz/tesla.

2.5. THE PASSAGE OF ELECTRONS THROUGH A LOW-PRESSURE GAS

Even under the conditions of the best vacuum that can be produced in practice, there are still a large number of gas molecules and hence the possibility of collisions when a beam of electrons passes through the gas. Vacuum thermionic valves operate with a gas pressure sufficiently low that only a very small fraction of the electrons have such collisions, but these collisions nevertheless occur. There are many valves in which the gas pressure is deliberately increased, and the characteristics of the valve depend upon electron–molecule collisions. The gas diode and the gas triode, or thyratron, are examples of such valves.

When an electron collides with a gas molecule one of three types of process occurs. These interactions are known as elastic, ionizing and exciting collisions.

An elastic collision is one in which the final state of the gas molecule is unchanged from its initial state apart from a small change of kinetic energy, usually a transfer of energy from the electron to the gas molecule. Such a collision obeys the usual laws of Newtonian "billiard-ball mechanics". The energy transferred in an elastic collision between an electron and a gas molecule is exceedingly small as the difference in mass of the two particles is very great (mass ratio at least 2000:1). The electron will normally rebound from the gas molecule with its velocity only slightly reduced but its direction of motion radically altered.

An ionizing collision can occur if the electron has sufficient energy, so that an electron in an outer orbit of a gas atom acquires sufficient energy during the collision to enable it to escape from its parent nucleus. The process is called ionization and the gas molecule which was initially electrically neutral is left minus an electron, i.e. it is left with a net positive charge. Such a particle is called a positive ion. There are thus three electrically charged particles left after an ionizing collision, the initial electron, the emitted electron, and the positive ion. The minimum energy needed by an electron before it can produce ionization is denoted by the Ionization Potential V_i of the gas. This is defined as the minimum energy of the electron per unit charge needed to produce ionization, and is equal to the potential

difference through which an electron has to move in order to gain the required energy. The ionization potential is normally measured in volts. Values of V_i for some gases are given in the following table:

Gas	Ar	Ne	Hg	O	H
V_i volts	15·8	21·5	10·2	13·6	13·5

If an electron with an energy greater than that necessary for ionization collides with a gas molecule, ionization is a possibility but not a certainty. Many electrons with an energy greater than V_i will in fact experience an elastic collision rather than an ionizing one. The third type of collision results in excitation of the gas molecule and subsequent emission of electromagnetic radiation, some of which might be in the visible part of the spectrum. During excitation an outer-orbit electron is given energy by the colliding electron, but not sufficient energy to enable it to escape from the parent nucleus. The transferred energy is then re-emitted by the excited molecule as radiation, the molecule returning to its initial state. A wide range of visible effects is obtainable, the frequency of radiation—or colour—depending upon the gas and gas pressure, and upon the energy of the incident electron. Light sources such as the mercury or sodium lamp, and neon-type shop-window

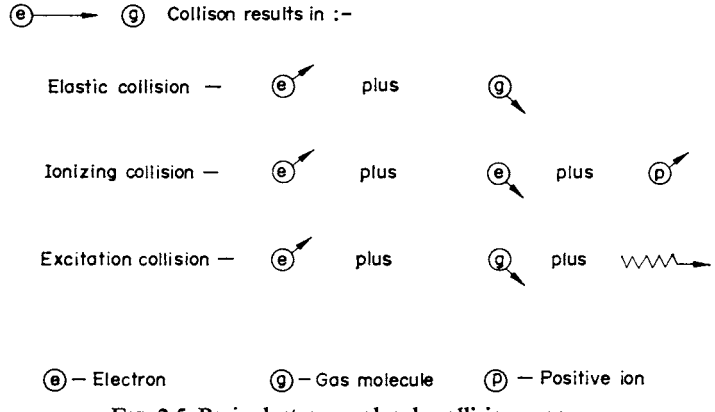


FIG. 2.5. Basic electron–molecule collision processes.

signs, make use of radiation from excitation collisions. The three types of collision are summarized in Fig. 2.5.

An important process that may occur when an electric field is imposed on a region of low pressure gas is the "electron avalanche". The process is illustrated in Fig. 2.6. Consider an electron which is accelerated by the electric field and then has a collision with a gas molecule. If the electron has

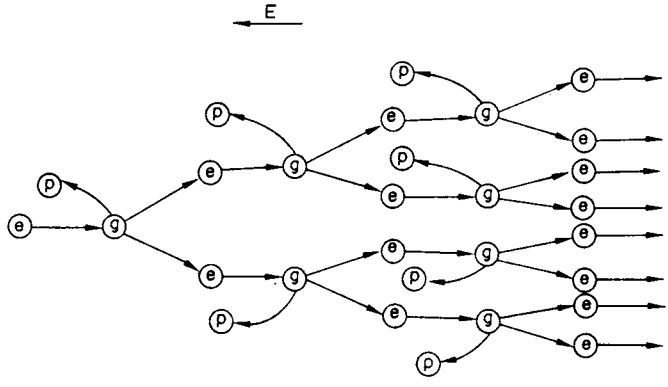


FIG. 2.6. Electron avalanche.

gained enough energy from the field it may ionize the molecule, the two resulting electrons being accelerated further by the electric field and the positive ion moving in the opposite direction. There are then two electrons which may produce ionizing collisions, and this process may be repeated many times over, the number of electrons growing rapidly across the region and the positive ions moving more slowly in the direction of the field. The electron density increases in geometric progression as the distance increases in arithmetic progression, i.e. the growth in number of electrons across the region is exponential with distance. The reason for the picturesque title of "electron avalanche" is obvious. It is essential that the electrons should gain sufficient energy to be able to produce ionization either between collisions or after a small number of collisions. At atmospheric pressure an electron will travel only about a millionth of a centimetre between collisions, and unless a very large electric field, of many thousand V/cm, is applied only elastic collisions will occur. Gas discharge valves and other

devices operate at low gas pressures of a tenth to a hundredth of an atmosphere. An electron then travels much further between collisions and sufficient energy may be obtained from the electric field to produce ionization. The average distance travelled by an electron between collisions in a gas is called the mean free path of the electron. The mean free path depends upon the gas and its pressure, and is inversely proportional to the gas density. Avalanche effects can also occur in a crystal lattice.

2.6. SPACE-CHARGE EFFECTS

The electric field has been previously assumed to be that due to the presence of various electrodes at different electric potentials, and no account has been taken of the effect that electrons will have on each other. In fact the electric field distribution can be markedly changed in a region when a high concentration of electrons or positive ions is introduced. A cloud of electrons, all having a negative electric charge, will experience mutual repulsion, and will exert a repelling force on any additional electrons that attempt to join the cloud. The force exerted on a particular electron can be calculated if the electric field set up by all the other charged particles in the vicinity is calculated. Some problems where the geometry is simple can be solved by the use of Gauss' law, which equates the total electric flux passing through a closed surface with the total electric charge within the surface.

Consider for example a cylindrical beam of electrons of radius a and of uniform charge density ϱ, as in Fig. 2.7. If the electron beam is very long

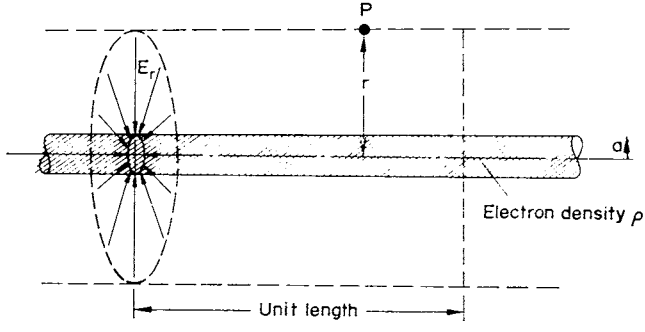


FIG. 2.7. Calculation of electric field due to cylindrical electron beam.

in comparison to the radius, all the electric flux lines will be radial and no axial force will be exerted on any electron. To find the electric field at a point distant r from the axis of the beam, e.g. at point P, a closed cylindrical surface is imagined to surround the beam, of radius r, and unit length of this surface is considered. The total electric charge within this unit length of surface is $\pi a^2 \varrho$, and the total electric flux through the surface is therefore also $\pi a^2 \varrho$. Equating the electric flux and the charge, the electric flux density D at radius r is therefore equal to $\pi a^2 \varrho / 2\pi r$. Since the electric field E is related to the electric flux density D by the relationship $D = \varepsilon_0 E$, where ε_0 is the permittivity of free space (of magnitude $1/36\pi \times 10^{-9}$ F/m), the electric field E_r at radius r is given by

$$E_r = \frac{\varrho a^2}{2r\varepsilon_0}. \tag{2.13}$$

At the outer surface of the beam this electric field is of magnitude $E_a = \varrho a / 2\varepsilon_0$, and the outward radial force on each of the electrons at the outer surface of the beam is therefore equal to $ea\varrho/2\varepsilon_0$ in magnitude.

A similar calculation can be made of the electric field due to a point charge. In this case the point is surrounded by a sphere with the point as centre. If the radius of this sphere is r, and the point charge is q_1, then q_1 is equal to $4\pi r^2 D_r$, and the electric field is given by

$$E_r = \frac{q_1}{4\pi r^2 \varepsilon_0} \tag{2.14}$$

and the force on a charge q_2 a distance r from q_1 is

$$F = \frac{q_1 q_2}{4\pi r^2 \varepsilon_0}. \tag{2.15}$$

Complex geometries cannot be treated in this simple manner but more advanced analytical techniques are available for solving many of the space-charge problems met in practice.

3. *Fundamentals of Solid State Electronics*

3.1. THE STRUCTURE OF THE ATOM

IN THIS chapter the motion of electrons within a solid is considered, rather than, as in the previous chapter, the motion of electrons in free space after they have been emitted from a hot cathode. In the earlier chapter the motion of an electron under the influence of electric and magnetic fields was considered, and it has been assumed, as is the case in most thermionic valves, that collisions between electrons and neutral molecules or positive ions are very infrequent. In solids, however, the reverse is true. Electrons moving about in the material make many collisions with atoms of the material, and their motion is profoundly affected by these collisions. It is necessary first of all to discover why materials, classed as conductors or semiconductors, possess electrons which can move from atom to atom.

3.1.1. *Bohr Model of the Atom*

The model of the atom to be discussed was proposed in 1913 by Niels Bohr. This model has now been superseded by what is known as the quantum mechanical or wave mechanical model of the atom, but the simpler Bohr model will suffice for our present considerations.

The atom is considered to consist of a positively charged nucleus and a number of negatively charged electrons revolving around the nucleus in circular orbits. The nucleus consists of neutral particles called neutrons, and a number of positively charged particles called protons. The mass of a proton is $1{\cdot}67\times10^{-27}$ kg and its charge e is $1{\cdot}60\times10^{-19}$ C. The number of protons Z possessed by the nucleus of an atom is called the atomic number of that atom. An atom is normally neutral and since the electronic

charge $-e$ is $-1{\cdot}60\times10^{-19}$ C, it follows that a neutral atom possesses the same number of protons as electrons.

The forces that keep the electron in orbit around the nucleus can best be seen by considering a simple atomic model, consisting of a nucleus with positive charge Ze and a single electron charge $-e$ in orbit around it. This is shown in Fig. 3.1. The radius of the orbit is r and the angular velocity of the electron is ω. The coulomb force of attraction between electron and

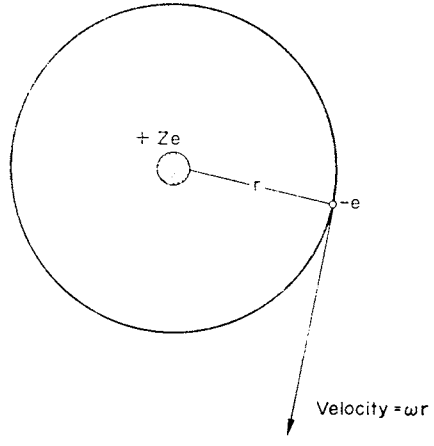


FIG. 3.1. Electron in circular orbit, radius r, around a nucleus of charge Ze.

nucleus (see Section 2.6) must be balanced by the centrifugal outward force on the electron:

$$m\omega^2 r = \frac{Ze^2}{4\pi\varepsilon_0 r^2}. \tag{3.1}$$

The total energy W of the electron is the sum of its potential energy V and kinetic energy T. Its potential energy is arbitrarily defined as being zero at infinite radius, and at radius r is given by:

$$V = -\frac{Ze^2}{4\pi\varepsilon_0 r},$$

$$\therefore\ W = T+V = \frac{1}{2}m\omega^2 r^2 - \frac{Ze^2}{4\pi\varepsilon_0 r}. \tag{3.2}$$

Substituting for $m\omega^2$ from (3.1) into (3.2)

$$W = -\frac{Ze^2}{8\pi\varepsilon_0 r}. \tag{3.3}$$

The minus sign indicates that the electron is bound to the nucleus, and thatsince its total energy is zero at infinity, an amount of energy equal to $Ze^2/8\pi\varepsilon_0 r$ has to be supplied to move the electron from its orbit of radius r to infinity.

It would appear that any orbit radius is stable. Having decided on a value of r, the corresponding angular velocity ω is given by (3.1) and energy by (3.3). However, an electron moving in orbit constitutes a current. A hypothetical observer, sitting near the orbit, would see the electron pass close to him $\omega/2\pi$ times per second, and he would assume that a pulsating current of this fundamental frequency was flowing around the orbit. Such a current is known to radiate electromagnetic energy in much the same manner as an alternating current flowing in a radio aerial radiates energy. Thus the electron must continuously be losing energy by such radiation, i.e. W must become more and more negative and reference to equation (3.3) shows then that r must decrease. The electron consequently spirals into the nucleus. In order to obtain stability, Bohr had to postulate some axioms which were contrary to classical behaviour. The necessity of these postulates was an admission that electrons in the atom did not obey classical Newtonian mechanics. The postulates are:

(1) Only certain orbits are stable. These orbits are such that the angular momentum is equal to an integer times $h/2\pi$ where h is a fundamental constant, known as Planck's constant, and equal to $6{\cdot}62\times10^{-34}$ joule seconds. Thus,

$$m\omega r^2 = \frac{nh}{2\pi}, \quad \text{where} \quad n = 1, 2, 3, \ldots. \tag{3.4}$$

n is called the principal quantum number for the orbit.

(2) In such orbits an electron does not radiate electromagnetic energy.

(3) If an electron moves from an orbit of energy W_1 to one of lower energy W_2 then it radiates electromagnetic energy of frequency f such that $f = (W_1 - W_2)/h$. This last relation is sometimes called the Einstein frequency condition.

If equation (3.4) is used to eliminate ω from (3.1), then the allowed orbit radii are:

$$r_n = \frac{\varepsilon_0 h^2 n^2}{\pi m Z e^2} \tag{3.5}$$

$$= \frac{5{\cdot}29 \times 10^{-10} n^2}{Z} \text{ m} \tag{3.6}$$

and the energy of the electron in the orbit radius r_n is from (3.6) and (3.3)

$$W_n = -\frac{me^4 Z^2}{8\varepsilon_0^2 h^2 n^2} = -\frac{13{\cdot}6 Z^2}{n^2} \text{ eV}. \tag{3.7}$$

This simple picture of one electron in orbit around a nucleus of charge Ze fits fairly well for the hydrogen atom when $Z = 1$. The lowest energy state for hydrogen is seen from equation (3.7) to be $-13{\cdot}6$ eV (i.e. $Z = n = 1$). Thus, 13·6 eV energy is required to completely ionize the hydrogen atom. The corresponding radius of this orbit is, from (3.6), equal to $5{\cdot}29 \times 10^{-10}$ m or 5·29 angstrom units. This lowes energy level is spoken of as the ground state.

3.1.2. *Complex Atoms and the Periodic Table*

With more complex atoms than hydrogen, i.e. atoms with more than one orbital electron, the above picture is only partially correct. Equations (3.6) and (3.7) give the orbit radii and orbit energies approximately. The question arises, however, as to how many electrons can have the same principal quantum number n. What is there, for example, to stop all the electrons residing in the state of lowest energy with quantum number unity?

The answer to this problem is obtained from quantum mechanics and a fundamental postulate called the "Pauli Exclusion Principle". Using these ideas it may be shown that the maximum number of electrons that can exist with principal quantum number n is $2n^2$. The $2n^2$ electrons that can possibly have the principal quantum number n are said to be members of a shell. Electrons with $n = 1$ are said to be members of the K shell,

$n = 2$ the L shell and so on. Thus:

n	1	2	3	4	5
	K	L	M	N	O

It may also be shown that the energies of electrons with a given principal quantum number are not all the same, but there are a number of discrete levels centred around the value given by equation (3.7). Thus there are, in general, several discrete closely spaced energy levels within a shell. These levels are referred to as subshells, and there are n subshells within a shell of principal quantum number n. Electrons in the same subshell have, to all intents and purposes, the same energy. Subshells can contain 2, 6, 10, 14,..., $2(2l+1)$ electrons, and are referred to as $s, p, d, f, g, \ldots$, subshells respectively. An s subshell has slightly lower energy than a p subshell, a p subshell has slightly less energy than a d subshell, and so on. Thus, to label the energy level of an electron in an atom requires that its shell and subshell be given; viz. $3p$, $4s$ (or less commonly Mp, Ns). These rules will be used by way of example to decide the number of levels and number of electrons per level for the carbon atom.

Carbon has six electrons. In the K shell ($n=1$) there are $2(1)^2 = 2$ electrons in the s subshell. In the L shell ($n = 2$) there is room to accommodate $2(2)^2 = 8$ electrons. There are, however, only four electrons left and of these two go into the s subshell and the remaining two into the $2p$ subshell. Since six electrons may be fitted into a p subshell, it follows that there are fourempty levels in the $2p$ subshell of carbon. This method of constructing the nergy level configuration of an atom assumes that electrons fill up the lewer energy levels first. In Table 1 a portion of the periodic table is showon. This shows that up to, and including, argon the table builds u regularly; i.e. the lower energy levels appear to fill up first. Above potassium irregularities in this system seem to appear; e.g. potassium has a $4s$ electron and no electrons in the $3d$ state. There is no real contradiction here, however; detailed calculation shows that the total energy of potassium is less with one electron in the $4s$ state than with it in the $3d$ level.

The radius of an electron orbit is smaller the lower its total energy; this is clearly seen from equation (3.3). Thus, the radius of an s level is less than that of a p level and so on. The atom of carbon can be drawn

TABLE 1. *Electronic Structure of Atoms*

Principal quantum number n			1	2		3			4	
Azimuthal quantum number l			0	0	1	0	1	2	0	1
Letter designation of state			$1s$	$2s$	$2p$	$3s$	$3p$	$3d$	$4s$	$4p$
Z		Element								
1	H	Hydrogen	1							
2	He	Helium	2							
3	Li	Lithium	Helium core	1						
4	Be	Beryllium		2						
5	B	Boron		2	1					
6	C	Carbon		2	2					
7	N	Nitrogen		2	3					
8	O	Oxygen		2	4					
9	F	Fluorine		2	5					
10	Ne	Neon		2	6					
11	Na	Sodium	Neon core			1				
12	Mg	Magnesium				2				
13	Al	Aluminium				2	1			
14	Si	Silicon				2	2			
15	P	Phosphorus				2	3			
16	S	Sulphur				2	4			
17	Cl	Chlorine				2	5			
18	Ar	Argon				2	6			
19	K	Potassium	Argon core						1	
20	Ca	Calcium							2	
21	Sc	Scandium						1	2	
22	Ti	Titanium						2	2	
23	V	Vanadium						3	2	
24	Cr	Chromium						5	1	
25	Mn	Manganese						5	2	
26	Fe	Iron						6	2	
27	Co	Cobalt						7	2	
28	Ni	Nickel						8	2	
29	Cu	Copper						10	1	
30	Zn	Zinc						10	2	
31	Ga	Gallium						10	2	1
32	Ge	Germanium						10	2	2
33	As	Arsenic						10	2	3
34	Se	Selenium						10	2	4
35	Br	Bromine						10	2	5
36	Kr	Krypton						10	2	6

schematically as shown in Fig. 3.2. The orbits are labelled; solid black circles represent electrons in orbit, whilst empty circles represent the levels which are there to be filled if the atom possessed more electrons. An alternative representation of the energy levels is shown in Fig. 3.3. The energy

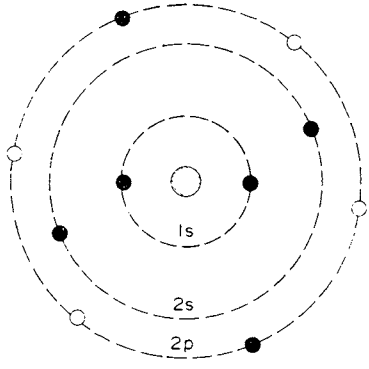


FIG. 3.2. Orbital picture of the carbon atom.

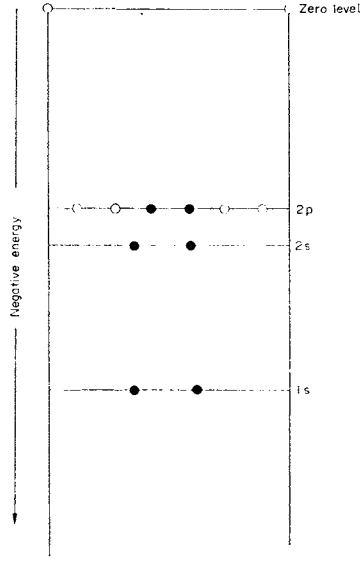


FIG. 3.3. Representation of the energy levels for the carbon atom.

levels are represented by the distance of the lines below the zero energy level (electron at infinite radius). The number of circles on each level represents allowed electronic states; those states filled by an electron are shown black. Carbon is again taken as the example in Fig. 3.3. Note that there is no significance to be attached to the horizontal length of the lines in this representation.

Considering the carbon atom once again, it has been seen that quantum mechanics and the Pauli Exclusion Principle only allow two electrons in the 1*s* subshell, for example. If an attempt were made to put a further electron in this subshell, then all efforts would be frustrated and this state could never be attained. One might suppose that very large repulsive forces acted on the electron to prevent it settling in this subshell. The nature of these forces is obscure; the reader must content himself by considering them as arising from the exclusion principle. The exclusion principle then, first suggested by Pauli, leads *inter alia*, to the conclusion that subshells can only possess 2, 6, 10, 14, etc., electrons at maximum and that a given shell, quantum number n, can only have n subshells. This limit to the number of electrons that can be crowded into a given subshell is of great importance when considering the nature of the chemical bonds between materials. There is no proof of the exclusion principle; it is justified because it predicts results in agreement with experiment.

3.2. THE NATURE OF CERTAIN CHEMICAL BONDS

When two atoms, similar or dissimilar, are brought into close proximity to each other, they may link together to form a stable molecule. Suppose atom A and atom B are the constituents of such a molecule. As A and B are brought together, the outer electron orbits of each start to overlap, and are therefore the first orbits to be perturbed. Not only are they the first to be perturbed, but they are the easiest to perturb since their electrons are furthest from the strong attractive force of the nucleus. The electrons in the outermost shell of an atom are spoken of as valence electrons, and the chemical properties of atoms are determined largely by the configuration of these electrons.

The class of bond resulting when two atoms are brought together depends on the manner in which the valence orbits are perturbed. We will limit ourselves here to three types of bonds; ionic, homopolar, and metallic bonds.

3.2.1. *Ionic Bonds*

The sodium chloride molecule is an example of such a bond. Figure 3.4 shows a sodium atom and chlorine atom in proximity. Reference to the periodic table shows that sodium has one electron in the 3*s* valence subshell, whilst chlorine has five electrons in its 3*p* valence subshell, i.e. it is one elec-

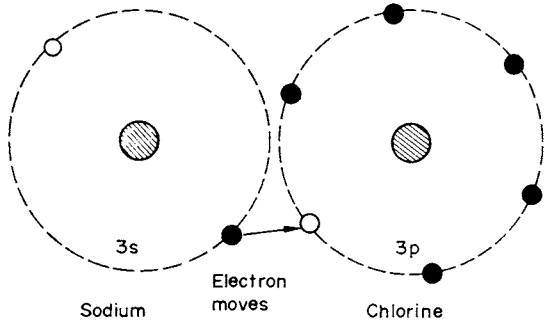


FIG. 3.4. Ionic bond between sodium and chlorine atoms.

tron short of a complete subshell. As the orbits of the sodium and chlorine atoms start to overlap, the one electron in the 3*s* sodium level fills the vacant position in the 3*p* chlorine level giving a positively charged sodium ion and a negative chlorine ion. The two ions are thereby attracted to each other electrostatically but cannot approach each other too closely since the exclusion principle forbids this. Electrostatic forces of attraction can be considered to be balanced by the repulsive forces arising from "exclusion". A stable configuration of the molecule occurs when these forces balance.

An ionic bond is characterized by one of the constituent atoms losing an electron and becoming a positive ion, the other atom gaining an electron and becoming a negative ion.

3.2.2. *Homopolar or Covalent Bonds*

Homopolar bonds occur when two or more atoms jointly share electrons. The chlorine molecule is an example of a covalent bond. A chlorine atom is one electron short of completing its $3p$ subshell. As the two atoms are brought together it might be expected that one atom would lose an electron in order to complete its neighbour's subshell, and become thereby a positive ion. The chlorine atom with the completed subshell would become negatively charged and the two atoms would be attracted electrostatically in the same manner as the ionic bond. However, there is no reason why one chlorine atom should complete its subshell at the expense of the other. The two atoms may be considered to exchange electrons continuously, and thereby attract each other electrostatically. Since the rate at which these electrons are exchanged is very rapid, it is convenient to consider that on average, one valence electron per atom spends most of its time in the region between the atoms. The situation is depicted pictorially in Fig. 3.5. This picture should

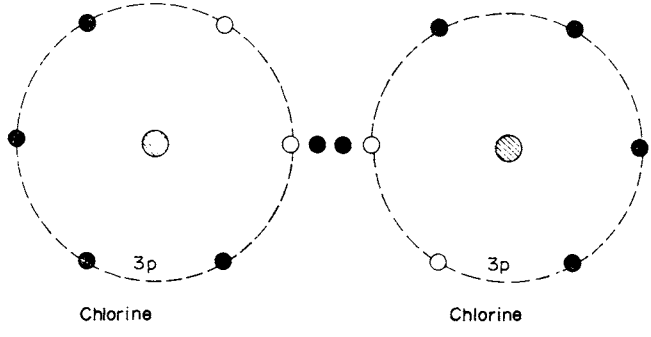


FIG. 3.5. Covalent bond between two chlorine atoms.
Inner subshells are not shown.

not be interpreted too literally, however; modern wave mechanics states that it is impossible to follow exactly the movement of single electrons. On average two electrons are most likely to be found in the region shown in Fig. 3.5.

The chlorine molecule is one example of a covalent bond. Two elements which form the basic materials used in transistors and solid state diodes, namely silicon and germanium, both form covalent bonds.

3.2.3. *Crystal Structure of Covalent Bonded Solids*

Reference to the periodic table shows that silicon has two electrons in the $3p$ level and there is room for six electrons in all in this level; the same applies to the $4p$ level in germanium. The bonding between silicon atoms in the solid state is such that each silicon atom, for example, tries to attract one electron from each of its four nearest neighbours to complete its $3p$ shell and become thereby negatively charged. However, just as in the case of the chlorine molecule, there is no preferred silicon atom which can attract electrons at the expense of neighbouring atoms. As a compromise each silicon atom tends to share its two $3s$ electrons and two $3p$ electrons

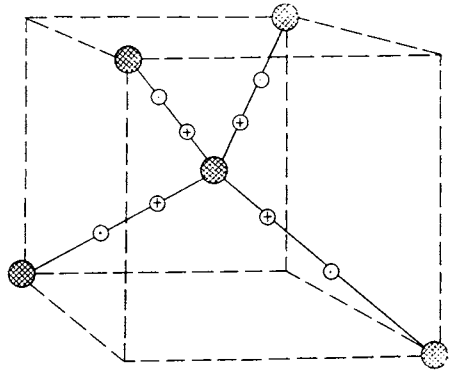


FIG. 3.6. A silicon atom with its four nearest neighbours (solid spheres). The spheres with crosses represent the four valence electrons of the central atom. The spheres with dots represent shared electrons from the four neighbours.

with its four nearest neighbours. When silicon condenses into the solid state, the atoms are arranged in a regular manner and the material is said to be crystalline. A silicon atom and its four nearest neighbours are shown diagrammatically in Fig. 3.6. The hatched spheres represent the nuclei of silicon atoms and the inner electron subshells (i.e. $1s$, $2s$, $2p$). The $3s$ and $3p$

electrons of silicon are shown as circles with crosses in them and the shared nearest-neighbour electrons are shown as circles with dots in them. The nearest-neighbour atoms lie at the corners of a regular tetrahedron and in consequence the bond is called a tetrahedral covalent bond. The whole crystal consists of this pattern repeated for each silicon atom. A simplified two-dimensional sketch is shown in Fig. 3.7. As with the covalent chlorine

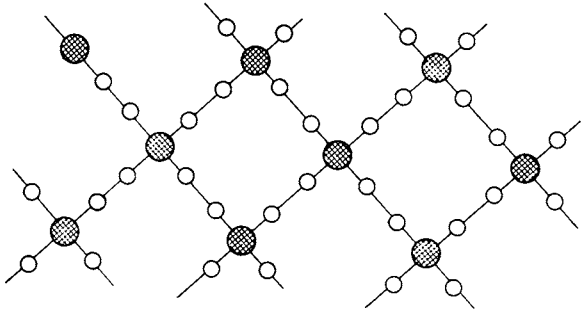


FIG. 3.7. Two-dimensional representation of covalent bonding in silicon. Hacthed circles—atom cores; open circles—electrons participating in the bond.

bond, the two shared electrons lying between the line of centres of nearest-neighbour atoms can be regarded as attracting the two positively charged silicon ions.

When such a material condenses from the liquid phase, crystals form simultaneously from many points of the melt, and rather than solidifying as a single crystal, the material consists of many small crystals of varying size; it is said to be polycrystalline. It is essential for transistor operation that only single crystals are used. Great care must be taken when preparing silicon and germanium for use in solid state devices that single crystals only are produced. Some of the techniques for growing such crystals are discussed in Chapter 8.

The structure of germanium is essentially the same as described for silicon. The $4s$ and $4p$ electrons now play the same role as the $3s$ and $3p$ electrons in silicon.

3.2.4. *The Metallic Bond*

Metals are characterized by low values of resistivity, i.e. they are good conductors. Looking back at the picture of the covalent bonds in silicon, for example, it is seen that all valence electrons (3*s* and 3*p*) participate in the bond; none of them are free to wander from their parent atoms. One should expect silicon to be an insulator since there appear to be no free electrons to carry current. As will be seen later, at normal temperatures, silicon is in fact a relatively poor conductor. It is obvious therefore that the metallic bond must be very different from the covalent bond; it must allow a high density of mobile electrons even at normal temperatures.

There are many types of unit crystal patterns found in nature, the three types most common to metals being:

(1) body-centred cubic,
(2) face-centred cubic,
(3) close-packed hexagonal.

The good electrical conductors—for example, silver, copper, gold and aluminium—belong to the face-centred cubic group. Figure 3.8 shows a face-centred unit cell; the corners of the cube and the centres of the cube faces are occupied by metallic atoms. By stacking cubes of this kind in three dimensions the metallic crystal would be obtained.

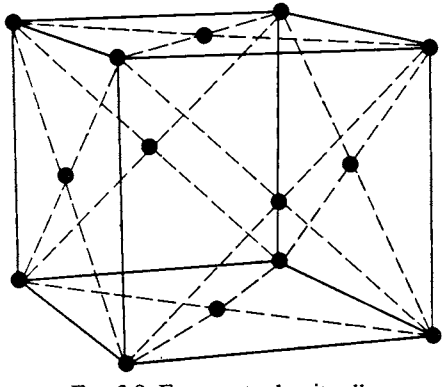


FIG. 3.8. Face-centred unit cell.

The metallic bond is not caused by nearest-neighbour atoms sharing electrons as in the covalent bond. The interatomic spacing for a metallic solid is less than for a covalent bonded solid and this means that there is a considerable overlap between the valence orbits of neighbouring atoms. The valence electrons in this case no longer become bound to their parent atoms but wander from atom to atom. A metallic solid is envisaged as a sea of mobile electrons with the resulting positive atom cores embedded in this sea. If one positive atom core moved from its equilibrium position, then an excess of negative charge would exist in the region from which it had moved. This would result in a force attracting it back to its equilibrium position; thus the atom core is bound in position in this ocean of electrons.

Whilst the atom cores are rigidly fixed, the valence electrons are now free to wander around between them. Occasionally an electron will approach very close to a positive atom core and will be attracted towards it, and thereby be deviated from its path. The electron is then considered to have undergone a collision with the atom core.

There is a great similarity between this picture of electrons moving freely through a metallic crystal lattice and molecules moving around in a gas. In a gas the molecules collide amongst themselves; in a metal the electrons collide with the positive atom cores. Collisions between electrons can be neglected. The molecules in a gas are constrained to move within the containing vessel since they cannot penetrate its walls. Likewise electrons are constrained to stay within the body of the metal as already described in Section 2.2.

The normal position of an atom core in the lattice is spoken of as a lattice site. So far such atom cores have been considered as being rigidly fixed at the lattice sites. This however is only true at the absolute zero of temperature. At temperatures above zero the atom cores vibrate about their equilibrium positions and the amplitude of oscillation is greater the higher the temperature of the substance. When a metal is heated, for example, the majority of heat energy that is given to it is stored in these oscillations, although a very small amount is also given to the free electrons.

3.3. INSULATORS, SEMICONDUCTORS AND CONDUCTORS

3.3.1. *Production of Free Charge Carriers*

We are now in a position to appreciate the differences between insulators and conductors. The factor that determines whether a material is one or the other is the nature of the bond between constituent atoms. If all the electrons of an atom are required in the bond it would appear that none are free to participate in conduction and the material should be a perfect insulator. The covalent bonds described earlier are examples of this. However, diamond, for example, has a covalent bond and a conductivity of 10^{-14} S/m. At 20°C the conductivity of pure silicon is $0{\cdot}3\times10^{-3}$ S/m, and of pure germanium 2·0 S/m. Since none of these values is zero, there must therefore be a few electrons that can participate in conduction.

It is obvious that a covalent bond can be broken since one can readily smash a sample of diamond or silicon, for example. It requires a certain amount of energy to partially break a bond and release an electron, and this energy is called the ionization energy. If the material is heated then the lattice vibrations grow in amplitude and eventually sufficient kinetic energy is given to the valence electrons to allow some of them to break free. The electrons are then able to wander about the lattice of the material and contribute to the conductivity. Eventually a free electron will come across another broken bond and be captured, i.e. drawn into the bond: it will then no longer be available for conduction. In equilibrium there will be a certain number of bonds per unit volume broken per second and the same number of broken bonds completed per second. There will be an average equilibrium value n_i of free electrons per unit volume at any given time. One might expect to find that n_i was greater:

(a) the greater the temperature, since the lattice vibrational energy increases with temperature,

(b) the smaller the energy required to break a bond, i.e. the smaller the ionization energy.

It can be shown from statistical quantum mechanics that n_i is given approximately by:

$$n_i = 2\left(\frac{2\pi mkT}{h^2}\right)^{3/2} \exp\left(-\frac{W_g}{2kT}\right)$$

$$= 5\times 10^{21} T^{3/2} \exp\left(-\frac{W_g}{2kT}\right) \quad \text{electrons per m}^3, \qquad (3.8)$$

where T is the absolute temperature of the material, W_g is the ionization energy, k is Boltzmann's constant, h is Planck's constant and m is the electron mass. It can be seen that (a) and (b) are in accordance with equation (3.8).

From equation (3.8), the number of broken bonds per unit volume at any given temperature is very dependent on the ionization energy W_g. For diamond $W_g = 6$ eV whereas for silicon $W_g = 1{\cdot}1$ eV. Thus, there are many more broken bonds to be found per unit volume in silicon than in diamond, at the same temperature. The conductivity of a material, as will be seen later, is directly proportional to the number of free carriers. The conductivity of silicon is much greater than diamond, because it is easier for the lattice vibrations to break bonds in silicon than in diamond.

The conductivities of metals lie around 10^8 S/m and are considerably greater than the values for silicon or germanium. This arises because the conduction electron density in metals is very great, even at very low temperatures, since the valence electrons from each atom become freed in the solid state. In copper, for example, the 4*s* electron is free, and, as the lattice spacing for copper is little greater than 1 angstrom unit (10^{-10} m), there are about 10^{29} electrons per cubic m (i.e. one *free* electron per atom). The value of n_i for silicon at $T = 300$ K, using equation (3.8) is found to be about $1{\cdot}5\times 10^{16}$ electrons per cubic m, which is much smaller than the free electron density in copper.

Materials with conductivities in the mid-range between metals on the one hand and good insulators (like diamond) on the other are called semiconductors. Silicon and germanium are therefore semiconductors. As will be described in the next section, semiconductors are also characterized by having negative temperature coefficients of resistance.

3.3.2. *Drift Motion of Charge Carriers in an Electric Field*

When an electron becomes freed from its bond it can move about, colliding from time to time with other atoms and eventually recombining in another broken bond. Figure 3.9 shows schematically its motion in space between being freed and recombining. The changes in direction occur due to deflecting collisions with atom cores. There is no tendency for an electron

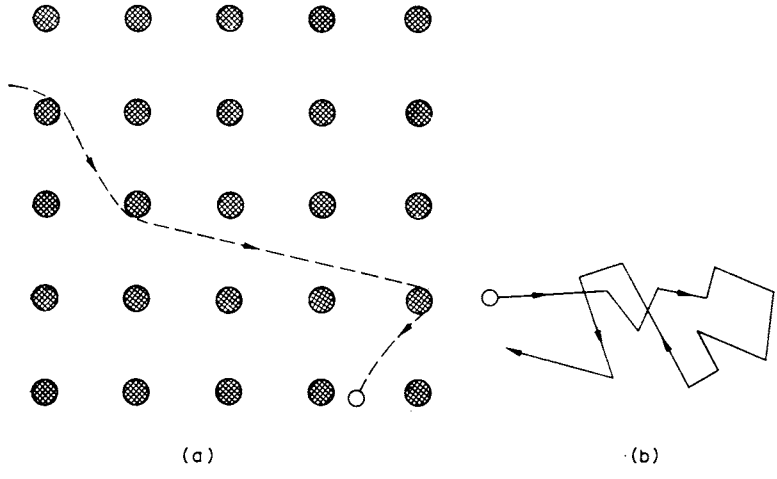


FIG. 3.9. (a) Motion of an electron as it moves through a crystal lattice. On average an electron may pass about 100 atom cores before it suffers a significant deflection.
(b) Electron trajectory over a longer period than (a).

to move in any overall direction, and at any given time there are likely to be as many electrons moving to the left say as to the right. If, however, an electric field is applied to the material, then superimposed on this random motion due to collisions, there is a net drift velocity in the direction of the force on the electron caused by the electric field. This is shown in Fig. 3.10. Suppose this drift velocity is always small compared to the random velocity of electrons between collisions. Let the electric field be E and the mean time between collisions, τ seconds. Then the force on an electron due

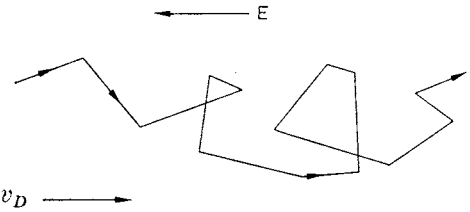


FIG. 3.10. Motion of an electron moving through a metallic lattice where an electric field E is applied in the direction shown. There is a resulting drift velocity v_D.

to the electric field is $-eE$ and its acceleration in the direction of the electric field is $-eE/m$, where m is the electron mass. The increase in velocity in the field direction between collisions is thus $-Ee\tau/m$. After collision, an electron moves off in a direction completely unrelated to its direction of motion before collision, but dependent solely on the geometry of the collision. It loses the directed component of velocity $-eE\tau/m$. Between collisions the directed component of velocity increases uniformly from zero to $-eE\tau/m$. Thus, the average drift velocity v_D is given by:

$$v_D = -\frac{eE\tau}{2m}. \tag{3.9}$$

Note that the drift velocity is proportional to the electric field and the mean time between collisions. We wish now to calculate the current produced by this drift of electrons. Suppose a plane of unit area (dashed lines in Fig. 3.11) is imagined, normal to the drift velocity, and that the number of electrons per second that pass through this plane is counted. The number that just pass through in one second must be those contained within the volume shown; the last one to pass through the surface must have been distant v_D away when counting started. The volume is v_D therefore and the number of particles passing through the surface per second is Nv_D where N is the density of free electrons. Each electron transports a charge $-e$, so the total charge passing through the surface per second is:

$$J = -Nev_D \tag{3.10}$$

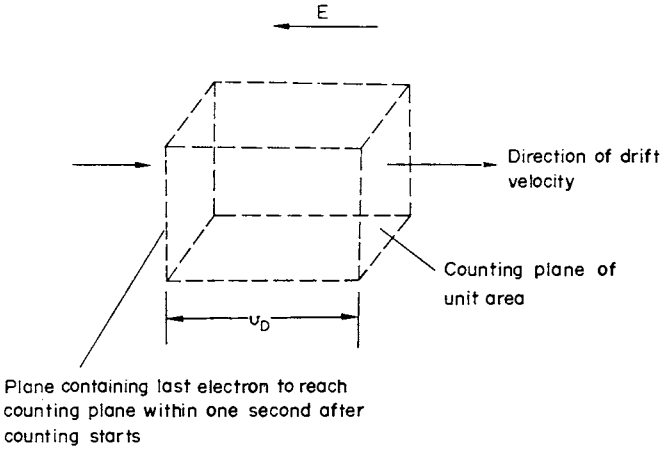


FIG. 3.11. Calculation of current density in terms of particle density and drift velocity.

where J is the current density. Now Ohm's law states that:

$$J = \sigma E \tag{3.11}$$

where σ is the conductivity of the material. Thus, from (3.10) and (3.11) it can be seen that

$$\sigma = \frac{Ne^2\tau}{2m}. \tag{3.12}$$

Thus the conductivity of a material depends on:

(a) the number of electrons free for conduction,
(b) the mean free time between collisions.

Equation (3.12) gives some insight into the reasons why metals have positive temperature coefficients of resistance whilst semiconductors and insulators have negative temperature coefficients. For a metal, N does not vary with temperature, thus, since σ decreases with increasing temperature, it follows that τ decrease with increasing temperature. The fact that an electron makes more collisions per second with the lattice atoms at higher temperatures can be explained only by wave mechanics. A partially correct

explanation, however, is that at higher temperatures the lattice atoms are vibrating with larger amplitudes of oscillation than at lower temperatures and thus present a bigger effective area for collision.

Now in the case of a semiconductor, the free electron density increases very rapidly with temperature as predicted by equation (3.8). The rapid increase in n_i with temperature completely masks the decrease in τ, thus causing a net increase of σ with temperature. A semiconductor thus has a resistance that decreases with increasing temperature.

The Joule heating effect that occurs in a conductor or semiconductor when current passes is readily explained on the above picture. An electron gains kinetic energy from the electric field equal to $\frac{1}{2}m(eE\tau/m)^2$ between collisions. On collision it gives all this up to the lattice atom it collides with. The electron makes $1/\tau$ collisions per second, and there are N electrons per unit volume, so the total kinetic energy per unit volume per second imparted to the atomic lattices by collision is:

$$\frac{1}{2}m\left(\frac{eE\tau}{m}\right)^2\frac{N}{\tau} = \frac{Ne^2\tau}{2m}E^2 \quad \text{W/unit vol.}$$

From equation (3.12) this becomes equal to σE^2. This is the usual relationship for the Joule heating effect.

From equation (3.9) the drift velocity is proportional to the electric field. The constant of proportionality, or the drift velocity per unit electric field strength, is called the mobility μ, thus:

$$v_D = \mu E.$$

From equation (3.9),

$$\mu = -\frac{e\tau}{2m}. \tag{3.13}$$

Later, when considering semiconductors in more detail it will be seen that there are charge carriers other than electrons. Suppose a mobile charge carrier has a charge q_1, mass m_1, collision mean free time with the lattice τ_1, then its mobility μ_1 is given by:

$$\mu_1 = \frac{q_1\tau_1}{2m_1}. \tag{3.14}$$

If an electric field E is applied to the material, then the current density J_1 is given by:

$$J_1 = N_1 q_1 \mu_1 E \tag{3.15}$$

where N_1 is the number of carriers per unit volume.

Equation (3.14) is a generalization of equation (3.13) and equation (3.15) comes from (3.10) and (3.13). When several different types of charge carrier are present the resultant current density J is given by

$$J = N_1 q_1 \mu_1 E + N_2 q_2 \mu_2 E + \ldots$$

where the suffixes 1, 2, ..., etc., stand for the different groups of particles. Such currents are usually called *drift currents.*

3.3.3. *Diffusion of Charge Carriers*

A current will exist in a material because of the drift velocity imposed on all current carriers by an external electric field. Currents may also exist without an external field if there is an unequal concentration of charge carriers within a material. Let us consider first of all an example of this in a gas of neutral particles, e.g. a molecular gas. Figure 3.12 shows a container filled with gas in which the pressure is greater to the right than to the left.

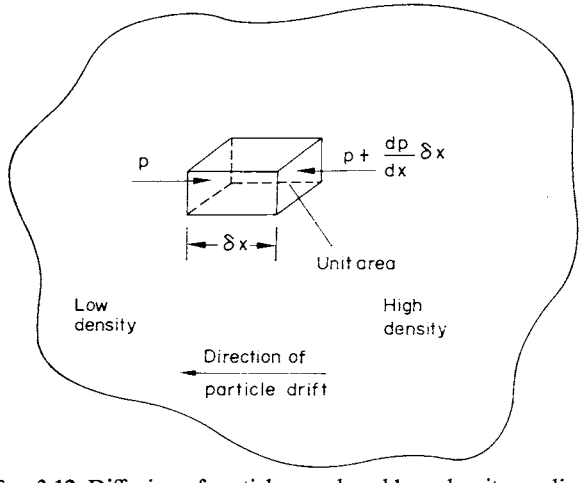


FIG. 3.12. Diffusion of particles produced by a density gradient.

Suppose that the pressure gradient exists in the x-direction only. Imagine a volume of molecules of length δx and unit cross-sectional area. This is shown in Fig. 3.12. The force on the left-hand face of the cylinder is p and on the right-hand face is $p+(\mathrm{d}p/\mathrm{d}x)\delta x$, where p is the pressure. Thus the overall force on the cylinder is $-(\mathrm{d}p/\mathrm{d}x)\delta x$ to the right. If there are N molecules per unit volume, then the total number in the cylinder is $N\delta x$. The average force per molecule is therefore

$$\frac{-\dfrac{\mathrm{d}p}{\mathrm{d}x}\delta x}{N\delta x} = \frac{-\dfrac{\mathrm{d}p}{\mathrm{d}x}}{N}.$$

This force also results in a drift velocity v_D. The argument now closely parallels the one used for calculating mobility except that the above force replaces the electric field force $-eE$. If τ is the mean time between collisions, the drift velocity is found to be:

$$v_D = -\frac{\tau}{2mN}\frac{\mathrm{d}p}{\mathrm{d}x}. \tag{3.16}$$

Now for any gas having N molecules per unit volume at temperature T, kinetic theory gives:

$$p = NkT$$

where k is Boltzmann's constant. Assuming that the temperature is everywhere the same,

$$\mathrm{d}p = kT\,\mathrm{d}N.$$

Substituting this in (3.16)

$$v_D = -\frac{\tau kT}{2mN}\frac{\mathrm{d}N}{\mathrm{d}x}. \tag{3.17}$$

This equation is the required diffusion equation and shows that there will be a net drift of molecules from regions of high density (and therefore pressure) to regions of lower density. The average drift velocity per particle is, from equation (3.17), equal to

$$(\text{constant})\times\frac{\text{density gradient}}{\text{density}}.$$

The constant is called the diffusion constant D.

Thus
$$D = \frac{\tau kT}{2m} \tag{3.18}$$

and
$$v_D = -\frac{D}{N}\frac{\mathrm{d}N}{\mathrm{d}x}. \tag{3.19}$$

The same effect occurs in any gas where there is a concentration gradient; the fact that the particles may be charged does not affect this analysis.

There is a relation between the mobility of a gas of charged particles and the diffusion constant of such a gas. Since from equations (3.18) and (3.14)

$$D = \frac{\tau kT}{2m}$$

and
$$\mu = \frac{q\tau}{2m}$$

it follows:
$$D = \frac{kT}{q}\mu \tag{3.20}$$

This equality is known as the Einstein relation.

If now the specimen is subject to an electric field E_x along the x-axis, and there is also a density gradient $\mathrm{d}N/\mathrm{d}x$, the resultant current density J_x arising from drift current and diffusion current is seen from equations (3.10) and (3.19) to be

$$J_x = Nq\mu E_x - Dq\frac{\mathrm{d}N}{\mathrm{d}x}. \tag{3.21}$$

If
$$E_x = \frac{D}{\mu N}\frac{\mathrm{d}N}{\mathrm{d}x} \tag{3.22}$$

then
$$J_x = 0.$$

It follows then that a suitable value of electric field can maintain a concentration gradient of charged particles, i.e. the tendency for particles to drift in one direction because of the concentration gradient can be balanced by an applied electric field trying to force the particles to move in the opposite direction. This interplay between concentration gradients and electric fields is very important to an understanding of transistor and diode action.

3.4. CONDUCTION IN SEMICONDUCTORS

3.4.1. *Intrinsic Semiconductors*

We have seen that at absolute zero temperature, silicon and germanium are perfect insulators. The average number of bonds broken per unit volume increases with temperature and is proportional to

$$T^{3/2} \exp\left(\frac{-W_g}{2kT}\right).$$

W_g is the ionization energy introduced in Section 3.3.1. Each time a bond is broken an electron is ejected from the bond and becomes free to move through the lattice until it eventually recombines in another broken bond. A broken bond is shown schematically in Fig. 3.13 where the freed electron is seen moving off through the lattice. A vacant site for an electron is thus

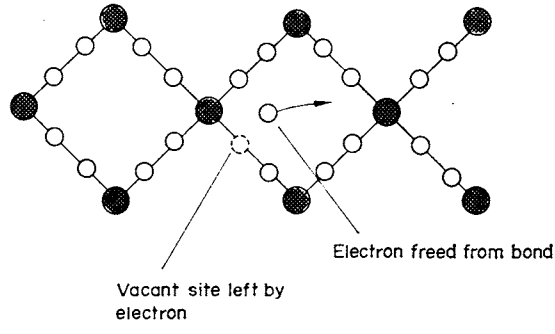


FIG. 3.13. A broken bond in silicon or germanium. The freed electron is seen wandering off through the lattice.

left in the bond and it might be supposed that after some time this site would be filled with a wandering electron from another distant broken bond. However, before this has time to happen, a neighbouring valence electron may move into this vacant site, complete the bond and leave a partly complete bond behind it. This is shown in Fig. 3.14 where a neighbouring electron is seen moving into the original vacant electron site. This transference of an electron from a neighbouring bond happens quite spontaneously and is independent of any applied electric field. Provided there is a vacant site in

the bond, a neighbouring electron is just as likely to jump into this bond as stay in its own bond. The vacant electron site in consequence moves around until finally a free electron wandering through the lattice fills the vacancy and reforms the stable bond.

Every time the vacant electron site moves, say, to the left, an electron must have moved to the right. Now an electron moving to the right is equiv-

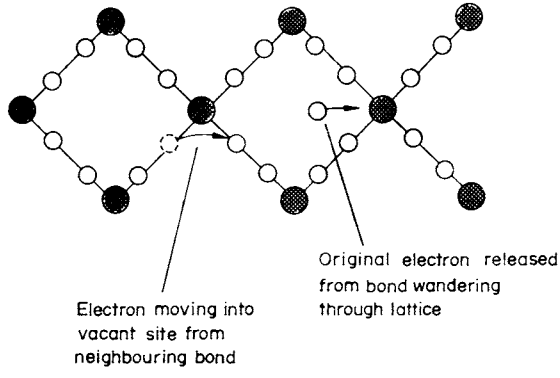


FIG. 3.14. Electron moving into a broken bond and leaving an incomplete bond behind it.

alent to a similar positively charged particle moving to the left; both particles produce the same resultant current. Thus, a vacant electron site can be considered as a fictitious positively charged particle, or positive hole as it is called. The current produced by the motion of a positive hole is in reality produced by an electron, moving in the opposite direction to the motion of the vacant electron site.

Positive holes can be endowed with the same parameters as electrons; viz. charge, mass, mobility, and diffusion constant. Whilst the charge of a positive hole has the same magnitude as an electron (but different sign of course), its mass, mobility, and diffusion coefficient are not necessarily the same as that of an electron. To add to the confusion, the mass of an electron as it wanders about the lattice is not necessarily the same as the value such an electron would have in free space. As an electron moves through the lattice, it comes sufficiently close to an atom core to count as a

collision fairly infrequently ($1/\tau$ times/sec). In the time between collisions it has steered itself past many atom cores but the continuous, slight reaction on it of the electric fields from these atom cores perturbs its motion. One can, however, conveniently forget this effect and use Newtonian laws of motion for an electron moving through the lattice if it is considered to have an *effective* mass different from its real mass. Lattice reaction is then automatically taken into account by the effective mass.

The values of effective mass, mobility and diffusion constant for electrons and positive holes at temperature 300 K in both germanium and silicon are given in Table 2.

TABLE 2

Silicon

Particle	$\frac{m^*}{m}$	μ m²/volt sec	D m²/sec
Electron	0·98	0·12	31×10^{-4}
Hole	0·49	0·025	$6{\cdot}5\times10^{-4}$

Germanium

Particle	$\frac{m^*}{m}$	μ m²/volt sec	D m²/sec
Electron	1·57	0·36	93×10^{-4}
Hole	0·28	0·17	44×10^{-4}

The effective mass m^* is divided by the free electron mass m.

The conductivity of a semiconductor is now seen to be due to two types of charge carrier, electrons and positive holes. The current J for an applied electric field E is thus:

$$J = eE(p\mu_p - n\mu_n) \tag{3.23}$$

where p is the density of holes, n the density of electrons, μ_p the mobility of holes, μ_n the mobility of electrons. Reference to equation (3.14) shows

that the mobility of electrons is negative, thus (3.23) may be written

$$J = eE(p\,|\,\mu_p\,|+n\,|\,\mu_n\,|) \tag{3.24}$$

where $|\,\mu_p\,|$ and $|\,\mu_n\,|$ are the absolute magnitudes of the two mobilities. Equation (3.24) is often written without including the moduli signs. The conductivity is given from (3.24):

$$\sigma = e[p\,|\,\mu_p\,|+n\,|\,\mu_n\,|]. \tag{3.25}$$

Pure silicon and pure germanium are called intrinsic semiconductors since their conductivity arises from a mechanism that is intrinsic in the nature of the atomic bonds in these materials. At any given temperature there is a certain average number of broken bonds per unit volume giving rise to an equal number of electrons and also positive holes. For every bond that is broken one free electron and one free positive hole is produced. It is usual to refer to this process as the thermal generation of intrinsic electron–hole pairs. Under these conditions it is obvious that in equations (3.24) and (3.25)

$$n = p = n_i$$

where n_i is the intrinsic electron density given by equation (3.8).

It is possible, however, to produce electrons and positive holes in silicon or germanium by the intentional addition of certain impurities. Addition of such impurities is known as doping. Doped semiconductors are sometimes called extrinsic, or impurity semiconductors.

3.4.2. *Impurity Semiconductors*

The principal elements used for doping silicon and germanium are taken from the third and fifth columns of the periodic table; the former are called group III elements and have three valence electrons, whilst the latter, called group V elements, have five valence electrons. Typical group III elements used are Boron (B), Aluminium (Al), Gallium (Ga) and Indium (In). Group V elements used are Phosphorus (P), Arsenic (As) and Antimony (Sb).

The proportion of impurity atoms added to the base material (i.e. silicon or germanium) is very small, being typically less than one part per million. The addition is usually done whilst the base material is molten and then

a single crystal is carefully grown from the melt. Thus a single crystal of germanium, say, with impurity atoms distributed very sparsely throughout the crystal lattice is obtained. It is found that the impurity atoms take up positions in the lattice that would normally be occupied by a parent atom. Figure 3.15 shows schematically a phosphorus impurity atom in a germanium lattice. Now phosphorus has five valence electrons, as compared with

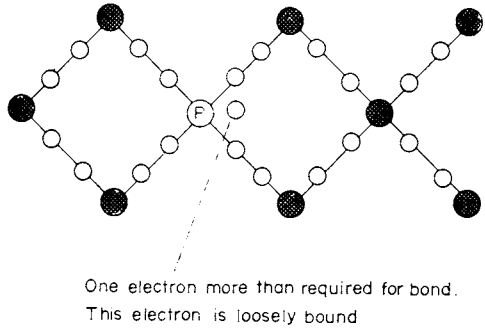


FIG. 3.15. A phosphorus (P) impurity atom in a germanium lattice. The black circles represent germanium atoms.

germanium's four. One of the phosphorus valence electrons is redundant as far as contributing to the covalent bond with its four surrounding germanium atoms. This extra electron is found to be very loosely bound to the phosphorus impurity atom and little energy is required to release it. As in the case of intrinsic semi-conductors, the thermal vibrations of the phosphorus atom and its neighbouring germanium atoms is usually sufficient to eject the superfluous electron, which then wanders about through the lattice. The phosphorus impurity; produces one free electron and is known therefore as an *n*-type impurity, *n* because it gives rise to a negatively charged particle (i.e. an electron). All the elements from group V are *n*-type impurities. The impurity atoms from this group are sometimes referred to as *donor* impurity atoms since they donate free electrons to the lattice. After an electron has been ejected from a donor atom, it must leave this atom positively charged. The positively charged atom, however, is rigidly bonded to its nearest four germanium atoms and cannot move or contribute to conduction in any way.

The ionization energy W_i, or energy required to release the donor electron, is given below for some n-type impurities in silicon and germanium. Note that the ionization energies in silicon are somewhat higher than in germanium.

TABLE 3

Impurity	P	As	Sb
Ionization energy in germanium (eV)	0·012	0·013	0·010
Ionization energy in silicon (eV)	0·045	0·049	0·039

It was stated earlier that the surplus electron per donor impurity was ejected because of the thermal vibrations of the donor impurity atom. The kinetic energy of oscillation of the lattice atoms is of the order kT per atom. Now at room temperature (300 K), kT is 0·025 eV. This is comparable with the energy required to release the surplus donor electron as the above table shows. More detailed calculation shows that at room temperature practically all the impurity atoms have lost their surplus electrons which are thus wandering through the lattice. All the donor atoms are said to be ionized.

Suppose a specimen of intrinsic germanium is considered. The number of free electrons per unit volume due to broken bonds is given by equation (3.8)

$$n_i = 5\times 10^{21} T^{3/2} \exp\left(-\frac{W_g}{2kT}\right).$$

Putting in the value $W_g = 0{\cdot}75$ eV for germanium, $T = 300$ K (room temperature), gives

$$n_i \approx 10^{19} \text{ electrons/m}^3.$$

Now suppose this material is doped with phosphorus atoms such that there is one impurity atom per million germanium atoms. There are approximately 10^{28} germanium atoms per m^3 in the solid state, thus the density of phosphorus atoms is $10^{28}/10^6 = 10^{22}$ atoms/m^3. Assuming all the donors are ionized, there are now 10^{22} free electrons per unit volume compared

with 10^{19} for the intrinsic case. The conductivity of the extrinsic germanium must, therefore, be some 10^3 times greater than that of the intrinsic germanium even though the doping is only one part per million.

Suppose, instead of doping with a group V element, a group III element such as boron is used. Then, as with phosphorus, the boron atoms take the place of germanium atoms in the lattice. However, since boron has only three valence electrons, there is one electron short in forming the complete covalent bond between the boron atom and its four neighbouring germanium atoms. This state is shown schematically in Fig. 3.16. Now an electron from a neighbouring complete germanium bond will fill this vacant electron site if a certain amount of energy can be imparted to it. Under these conditions

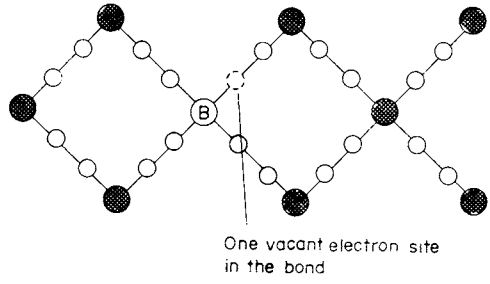


FIG. 3.16. A boron (B) impurity atom in a germanium lattice. The dotted circle shows a vacant electron site in the bond.

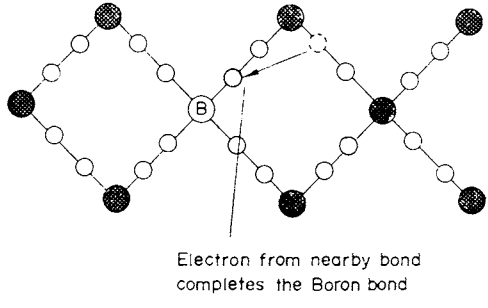


FIG. 3.17. An electron from a nearby germanium bond completing the boron–germanium bond if sufficient energy is supplied to it. A vacant electron site (shown dotted) is left in the germanium bond.

the boron atom will accept the neighbouring electron and, as shown in Fig. 3.17, a positive hole is created in the germanium bond that has given up its electron. Since boron and the other group III elements accept electrons to complete their bonds they are said to be *acceptor* impurities. The amount of energy required to transfer an electron into the incomplete bond is called the acceptor ionization energy. A list of these is given in Table 4 for silicon and germanium.

TABLE 4

Impurity	B	Al	In
Ionization energy in germanium (eV)	0·010	0·010	0·011
Ionization energy in silicon (eV)	0·045	0·060	0·160

The acceptor ionization energies are of similar order to the donor ionization energies and it follows from the argument used previously for donor impurities that at room temperature practically all the acceptor atoms are ionized. Note that an acceptor atom becomes negatively charged by addition of the electron required to complete its bond. The acceptor atoms are, however, rigidly fixed in position in the lattice and cannot participate in conduction. The positive holes produced, however, are free to wander about the lattice. Conduction in a semiconductor doped with a group III element then is largely by positive holes and the material is said to be *p*-type. Material doped with a group V element is called *n*-type.

It must be emphasized that addition of a donor impurity only produces electrons for conduction; no holes are created in this process. After the surplus electron from the donor atom has been released to wander through the lattice, the donor atom is left with four valence electrons; just the number required to form a stable covalent bond. There is no vacant electron site in the bond and, consequently, no positive hole is produced. Similarly acceptor impurities produce only positive holes and no free electrons are created.

3.5. BEHAVIOUR OF MINORITY AND MAJORITY CARRIERS

Let us return to the example of germanium at room temperature doped with phosphorus impurity atoms such that there is one impurity atom per million germanium atoms. It may be assumed that all the donor atoms are ionized and there are, in consequence, about 10^{22} free electrons/m^3. Now also a certain number of electrons and holes are produced by broken germanium bonds; i.e. produced by the intrinsic processes. This figure for pure germanium has been seen to be about 10^{19} electrons/m^3 and the same density for positive holes also. It might seem reasonable then to expect that in the doped specimen there were:

$$\text{Impurity electrons} + \text{Intrinsic electrons} = 10^{22} + 10^{19} \text{ electrons/m}^3 \approx 10^{22} \text{ electrons/m}^3.$$

$$\text{Intrinsic positive holes} \approx 10^{19} \text{ holes/m}^3.$$

It will be seen later that this result is not quite correct; but that the number of electrons is still far greater than the number of positive holes. The electrons are therefore spoken of as the *majority* charge carriers and the positive holes as *minority* charge carriers. In a *p*-type semiconductor the converse is obviously true, positive holes are the majority carriers whilst electrons are minority carriers.

Since usually in an impurity semiconductor one type of charge carrier is considerably more numerous than the other type, it might be thought that the minority carriers could be neglected. As will be seen later, minority carriers play a very important role in diodes and transistors and it is imperative to consider them.

3.5.1. *Production and Recombination Rates*

The addition of an impurity element to a pure piece of semiconductor has the effect of reducing the total number of broken intrinsic bonds. The total number of intrinsic electrons and positive holes decreases as the proportion of impurity element present increases. That this is so can be shown by the following argument.

Suppose the intrinsic electron density in the pure semiconductor before impurities have been added is n_i and the corresponding positive hole density is p_i. The subscript stands for intrinsic. Then $p_i = n_i$. Let the rate at which electron–hole pairs are created per unit volume be g. Now it is reasonable to suppose that electrons and holes recombine at a rate proportional both to the number of electrons present per unit volume and also the number of holes present per unit volume. Let the recombination rate per unit volume then be written as:

Recombination rate/unit volume $= rnp$ electron–hole pairs/unit vol/sec,

where n is the electron density, and p is the positive hole density, and r is a recombination constant. For the intrinsic conditions, equating creation rate with recombination rate gives

$$g = rn_i p_i. \tag{3.26}$$

Now for the doped semiconductor, the rate at which electron–hole pairs are created per unit volume (i.e. the rate at which intrinsic bonds are broken per unit volume) is not altered by the impurity. If the electron and positive hole densities are n and p respectively in the doped material, then

$$g = rnp. \tag{3.27}$$

From equations (3.26) and (3.27)

$$n_i p_i = np.$$

Since

$$n_i = p_i,$$

$$\therefore \quad n_i p_i = n_i^2 = p_i^2 = np. \tag{3.28}$$

Thus, as n say goes above the intrinsic value because of doping, it follows that p must decrease below the intrinsic value in order to keep the product np constant.

Applying equation (3.28) to the example of phosphorus doped germanium, the intrinsic density was found to be 10^{19} electrons or positive holes/m^3.

The electron density produced by doping was 10^{19} electrons/m^3. Thus, from equation (3.28),

$$10^{19} \times 10^{19} = 10^{22} \times p,$$
$$\therefore \quad p = 10^{16} \text{ positive holes/m}^3.$$

Thus, the minority carrier density is even less than the value of 10^{19} which might have been expected at first sight.

The conductivity of a specimen of doped semiconductor is given by equation (3.25) as

$$\sigma = e(n|\mu_n| + p|\mu_p|)$$

where $|\mu_n|$ and $|\mu_p|$ are the mobilities of electrons and holes respectively. The conductivity then is essentially determined by the number of majority carriers, since μ_n and μ_p are of similar magnitude. For an n-type semiconductor then, if μ_n is known and σ can be measured it is possible to calculate

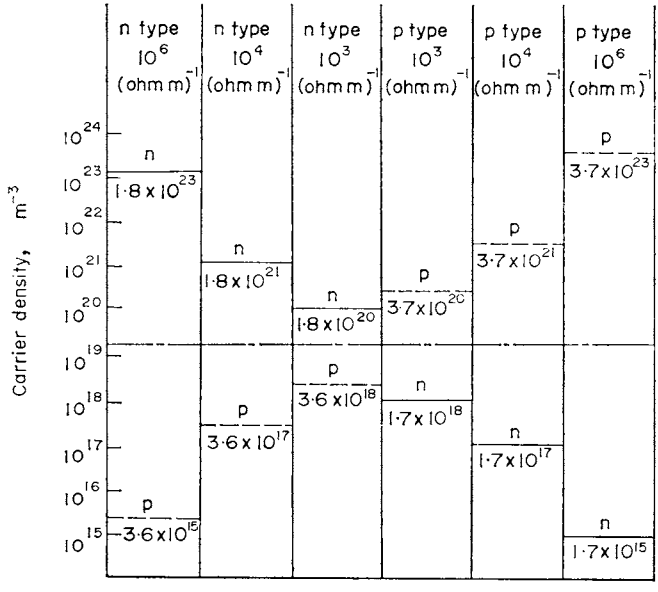


FIG. 3.18. Carrier densities at 300 K for germanium with different degrees of doping.

the number of electrons per unit volume n. If the intrinsic densities n_i and p_i are known then p may be found using equation (3.28).

Figure 3.18 is a table showing the majority and minority carrier densities for doped germanium at 300 K for varying conductivities (i.e. for various concentrations of impurity). Note that the product of $n \times p$ is always constant and equal to the intrinsic density squared ($n_i^2 = 6{\cdot}25 \times 10^{38}$ m^{-6} for germanium at 300 K).

3.5.2. *Minority Carrier Injection*

Consider the block of n-type semiconductor shown in Fig. 3.19. To begin with, the electron and hole densities n and p are assumed to be uniform

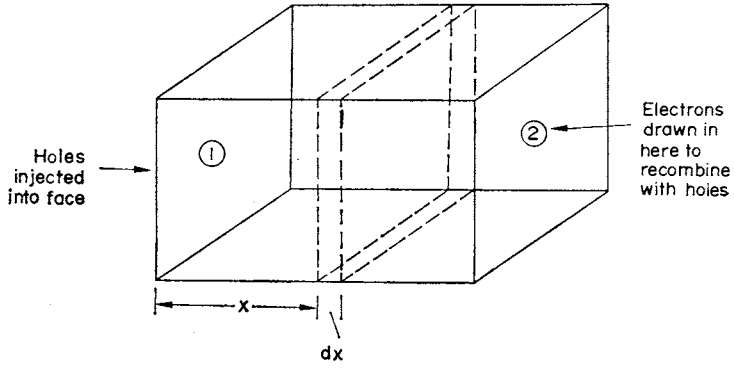


FIG. 3.19. Injection of minority carriers (holes) into a block of semiconducting material (n-type).

throughout the volume and equal to their equilibrium values n_n and p_n respectively. Since the material is assumed to be strongly n type:

$$n_n \gg p_n$$

and from equation (3.28)

$$n_n p_n = n_i p_i \tag{3.29}$$

where n_i and p_i are the intrinsic electron and hole densities.

Suppose now an attempt is made to inject *minority* carriers (holes) into the block continuously through the face 1 and to extract them through

the opposite face 2. The manner in which this might be done is described in detail in Chapter 4.

Injecting holes through the left-hand (L.H.) face will cause the hole density near this face to rise above the initial value p_n. These holes will start to diffuse away towards the right-hand (R.H.) face of the block because of the increased hole density near the L.H. face. As the holes diffuse across the block they will tend to recombine with electrons after a short distance because of the high electron density. The average distance that a hole moves before it recombines with an electron is called the *diffusion length* for holes L_p. It is typically about 10^{-3} m in the *n*-type materials used in diodes and transistors.

Thus, if the length of the block shown in Fig. 3.19 is much greater than 10^{-3} m, any attempts to inject holes through the L.H. face and collect them at the R.H. face will be completely unsuccessful. This does not mean, however, that no effects will be noticed at the R.H. face 2. Since holes are being continuously "pumped" as it were into face 1 and recombine in the block, electrons must continuously flow into the block through face 2 if a steady state is to exist. These electrons make good the loss of electrons by recombination with injected holes. It is assumed that both faces 1 and 2 have contacts made to them.

It is necessary to obtain an expression for the variation of hole density p throughout the block under the above conditions. If the density of holes at face 1 is p_{n0}, then intuitively the variation of density with distance x is as shown in Fig. 3.20a. The hole density must fall almost to the equilibrium value p_n after a distance of approximately L_p.

It will be supposed that hole flow is one-directional, i.e. along the direction of the x-axis. The cross-sectional area of faces 1 and 2 is A.

Consider a thin slab, width $\mathrm{d}x$, shown dotted in Fig. 3.19. Suppose the hole density in this slab is p. The recombination rate for hole–electron pairs in this slab, from equation (3.27) is:

$$Arn_n p \, \mathrm{d}x. \tag{3.30}$$

But the production rate, by the intrinsic process, is $gA \, \mathrm{d}x$ and this may also be written, using equations (3.26) and (3.29), as

$$Arn_n p_n \, \mathrm{d}x. \tag{3.31}$$

Thus, the net number of electron–hole pairs recombining in the dotted region per second is, from (3.30) and (3.31)

$$A(rn_n p - rn_n p_n)\,\mathrm{d}x. \tag{3.32}$$

If the hole current density flowing into the left-hand face of the dotted volume is J_p, and the hole current density flowing out of the right-hand face of this volume is $J_p + \mathrm{d}J_p$, then the number of holes lost per second in the volume is

$$-\frac{A\,\mathrm{d}J_p}{e} \tag{3.33}$$

where e is the charge on a positive hole. This must just be the number lost by recombination. Thus, equating (3.32) and (3.33)

$$\frac{A\,\mathrm{d}J_p}{e} = -Arn_n(p-p_n)\,\mathrm{d}x$$

or

$$\frac{\mathrm{d}J_p}{\mathrm{d}x} = -ern_n(p-p_n). \tag{3.34}$$

If there is no electric field present, the hole current J_p must arise solely from the hole density gradient. Thus, from equation (3.21)

$$J_p = -D_p e \frac{\mathrm{d}p}{\mathrm{d}x} \tag{3.35}$$

where D_p is the diffusion constant for holes.

Thus, from (3.35) and (3.34)

$$\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} = \frac{rn_n}{D_p}(p-p_n). \tag{3.36}$$

Writing

$$\frac{D_p}{rn_n} = (L_p)^2 \tag{3.37}$$

(3.36) becomes

$$L_p^2 \frac{\mathrm{d}^2 p}{\mathrm{d}x^2} = (p-p_n). \tag{3.38}$$

Now the correct solution to this equation, remembering that $p = p_{n0}$ at $x = 0$ and $p = p_n$ as x becomes very large, is

$$p = (p_{n0} - p_n) \exp\left(\frac{-x}{L_p}\right) + p_n . \tag{3.39}$$

This may be verified by substitution of (3.39) into (3.38). The curve in Fig. 3.20 (a) is thus given by the above equation.

Suppose now that instead of injecting holes into face 1, some method can be found for extracting holes through this face. The density will now fall near this face below the equilibrium value p_n and in consequence holes will continue to diffuse out through this face. The density of holes will vary with distance as shown in Fig. 3.20 (b). The distance over which the density

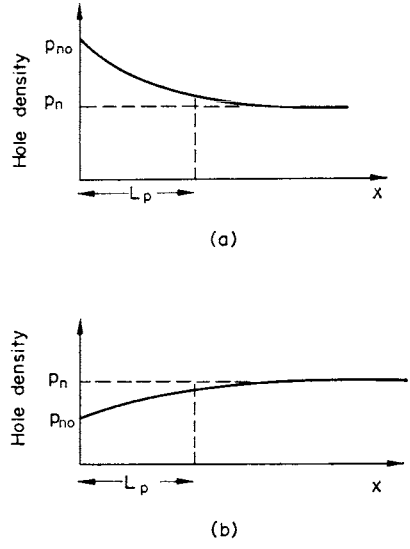


FIG. 3.20. Variation of hole density in the block shown in Fig. 3.19.
(a) Holes injected at face $x = 0$. (b) Holes extracted at face $x = 0$.

changes, however, must again be about L_p since holes cannot move further than this distance before they recombine with electrons. The density variation is in fact still given by (3.39), but now $p_n > p_{n0}$. Note that in the region where $p > p_{n0}$ there is a net production rate for electron–hole pairs. This

follows from equation (3.32) since now $rn_n p_n > rn_n p$. The electrons produced in the region near face 1 move out of the block through face 2. Hole injection and extraction are shown schematically in Fig. 3.21.

It is not evident at first sight that the assumption made in deriving equation (3.35) is necessarily true. The positive hole density near face 1 of the

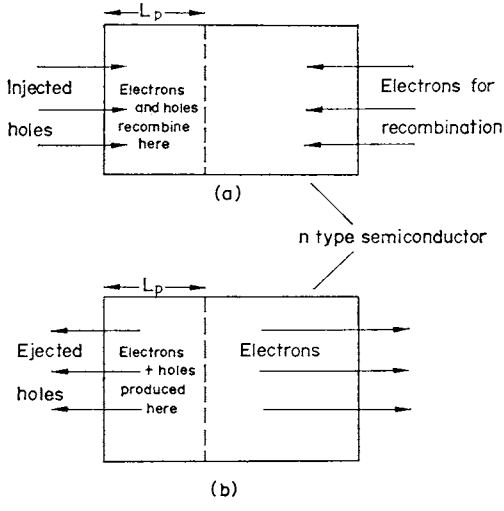


Fig. 3.21. Representation of electron–hole recombination (a) and production (b).

block has increased above the equilibrium value p_n over a region of length about L_p (Fig. 3.20). Unless the increased positive charge density in this region is counterbalanced by a similar increase in electron density, an electric field must result which will produce a component of hole drift current. This component would have to be added to the diffusion current given by equation (3.35).

The increase in minority carrier density above the equilibrium value, however, produces a similar increase in the majority carrier density in an attempt to restore charge neutrality. This is shown (not to scale) in Fig. 3.22. However, the increase in electron density is slightly less than the increase in hole density and an electric field is produced in the direction shown in Fig. 3.22. It is this field that is responsible for drawing in electrons from the

right-hand face of the cube towards the left-hand face. However, more detailed calculations show that the hole drift current produced by this small electric field is usually very much less than the hole diffusion current and may safely be neglected.

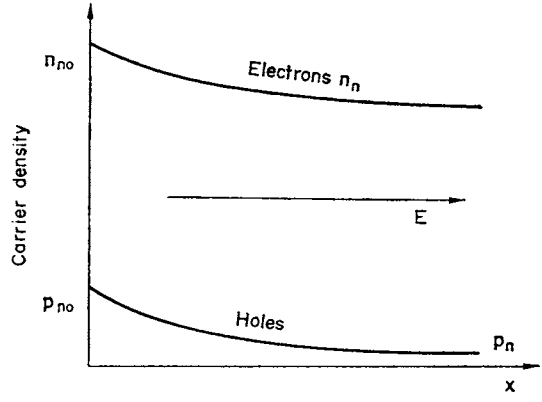


FIG. 3.22. Variation of hole and electron density in block shown in Fig. 3.19. The direction of the small resulting electric field is shown also.

This very important result, which will not be proved, may be stated as follows: The influence of small concentrations of minority carriers in the presence of much larger concentrations of majority carriers produces only a very small change in the local electric field. The effect of this field on the motion of minority carriers may be safely neglected.

4. *The Semiconductor Junction Diode and Transistor*

IN ORDER to understand the operation of a transistor, it is first necessary to be familiar with the mechanism of charge transport across a junction formed between a piece of *n*-type semiconductor and a piece of *p*-type semiconductor. Such a junction is called a *p–n* junction and exhibits rectifying properties, i.e. it appears to have a low resistance to current flow in one direction and a high resistance in the other direction. A *p–n* junction is shown schematically in Fig. 4.1 (a). Electrical connections are made to both ends.

A transistor is essentially a sandwich of two *p–n* junctions. A thin wafer of *n*-type material between sections of *p*-type material constitutes a *p–n–p*

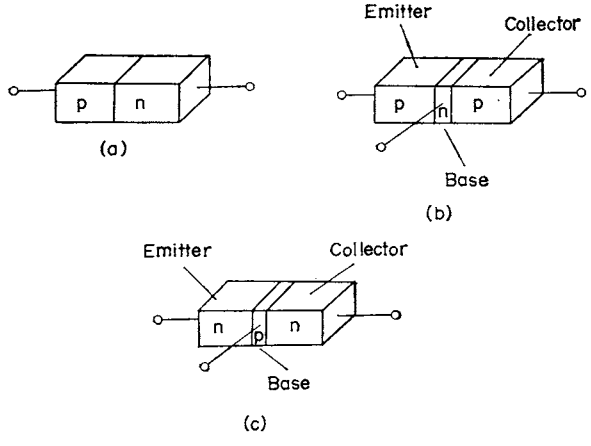


FIG. 4.1. (a) A *p–n* junction diode. (b) A *p–n–p* junction transistor. (c) A *n–p–n* junction transistor.

transistor, whilst the converse holds for an *n–p–n* transistor. Electrical connections are made to all three sections as shown in Figs. 4.1 (b) and 4.1 (c).

It is essential to realize that in all these three structures it is not sufficient to place the various doped pieces of semiconductor in contact with each other. The whole block constituting a *p–n* junction, for example, must be a single crystal of silicon or germanium, one-half of the crystal being doped with donor impurity atoms, the other half with acceptor atoms. If separate pieces of *p*- and *n*-type material were just placed in contact, then the discontinuity in the lattice structure at the junction would produce effects which in all probability would mask the desired ones. A transistor is one single crystal with three regions of different doping. The techniques used for manufacturing transistors and *p–n* junctions are deferred until Chapter 8.

4.1. THE MECHANISM OF CARRIER EXCHANGE AT A NON-RECTIFYING METAL-SEMICONDUCTOR CONTACT

Consider first of all a piece of *n*-type semiconductor with metallic contacts made to its ends. If a battery is connected across the material in the sense shown in Fig. 4.2, then electrons will flow in the direction indicated. The carriers in an *n*-type semiconductor are to all intents electrons; the number

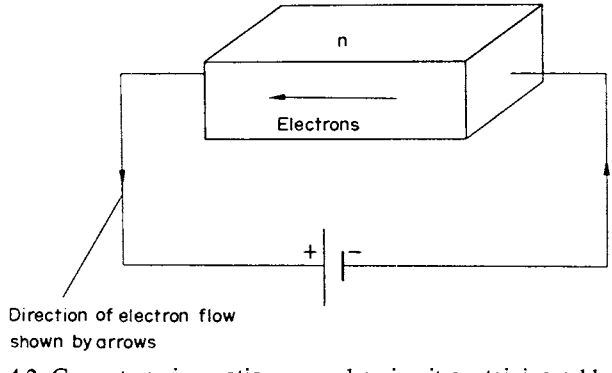


FIG. 4.2. Current carrier motion around a circuit containing a block of *n*-type semiconductor.

of positive holes, as seen in the previous chapter, is much smaller. The positive contact on the semiconductor will attract electrons from the body of the semiconductor towards it. This would leave the body of the semiconductor positively charged if it were not for the fact that electrons are emitted from the negative contact into the semiconductor at just sufficient a rate to balance the number leaving at the positive contact. There is a drift of electrons through the semiconductor and the total current passing through any transverse plane of the semiconducting block is constant. Also the electron density at all points within the body of the semiconductor remains constant at the value pertaining when no current is flowing. As quickly as electrons drift away from any volume of the specimen because of the applied voltage, new electrons drift into this volume, thereby ensuring the electron density is kept constant.

The situation is different if the specimen is *p*-type, since now conduction is predominantly by positive holes. However, in the connecting wires the conduction must be by free metallic electrons. Some mechanism must exist to allow the current to change its mode of transportation at the metal–semiconductor junction. Figure 4.3 shows a piece of *p*-type semiconductor

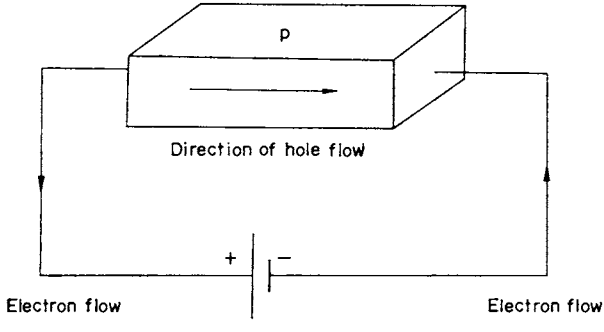


FIG. 4.3. Current carrier motion around a circuit containing a block of *p*-type material.

connected to a battery. The direction of electron flow in the metal wires is shown, and also the direction of positive hole flow in the semiconductor. These two directions are obviously opposite to each other, since holes and electrons have opposite charge. Positive holes in the semiconductor are

attracted towards the negative contact and tend to build up positive charge at this contact; this charge will attract electrons from the wire into the semiconductor and the excess positive holes collecting at the contact will recombine with the electrons flowing in. Thus, as many positive holes per second arrive at the negative contact from the body of the semiconductor as electrons arrive from the external lead, and they recombine in the semiconductor very close to the contact. At the positive contact, since holes have moved in the semiconductor to the negative contact, there must be a deficit of holes and the semiconductor must have a net negative charge in this region due to the unneutralized acceptor atoms. (The acceptor atoms have an electron stolen from another bond in order to make up a complete covalent bond.) The effect is to cause the stolen electrons from the acceptor atoms to be ejected into the lead wire in an attempt to bring about charge neutrality. Thus, the acceptor atoms near this contact once again become bonded with only three electrons to their neighbouring atoms and conditions are suitable for the creation of further new holes. The rate then at which electrons are freed from the acceptor atoms near the positive contact and ejected in the external lead wire is just equal to the rate of formation of holes in the semiconductor, which then drift along to the negative contact. To reiterate, holes are created at the positive contact, drift towards the negative contact and there they recombine with electrons moving into the semiconductor. The creation of holes at the positive contact causes electrons to be ejected into the positive contact wire. Throughout the circuit the total current is constant. Minority carrier injection and extraction was described quantitatively in Section 3.5.2.

The metal–semiconductor contacts described above are non-rectifying since current can pass across them equally in either direction. Certain metal–semiconductor junctions are rectifying but these will not be dealt with in this text.

4.2. THE *p–n* SEMICONDUCTOR DIODE

4.2.1. *Space Charge Layers*

Suppose a *p–n* junction is considered with its connecting wires open circuited. The *p*-region contains many positive holes (majority carriers) and relatively few electrons (minority carriers). The converse is true for the *n*-region.

Figure 4.4 shows the variation of carrier densities on either side of the junction. The *p*-type material is assumed to have a conductivity of 10^4 S/m,

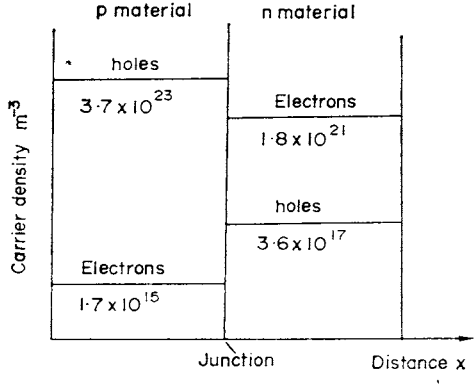


FIG. 4.4. Carrier densities on either side of a *p–n* junction before equilibrium.

whilst the *n*-type has a conductivity of 10^2 S/m. The corresponding carrier densities are obtained from Fig. 3.18 of Chapter 3.

The state of affairs, as depicted in Fig. 4.4, however, could not last for long, since there is seen to be a very large concentration gradient both of holes and electrons across the junction. These gradients must cause holes to diffuse into the *n*-type material and electrons to diffuse into the *p*-type. Figure 4.5 (a) shows the conditions at some slightly later time than Fig. 4.4 as holes have diffused into the *n*-region and electrons into the *p*-region. The holes which have diffused out from the *p*-region must leave behind negative acceptor atoms, since these atoms are bound in the lattice structure and are not able to follow the holes. Similarly, electrons diffusing into the *p*-region must leave behind positively charged donor atoms in the *n*-material. A double

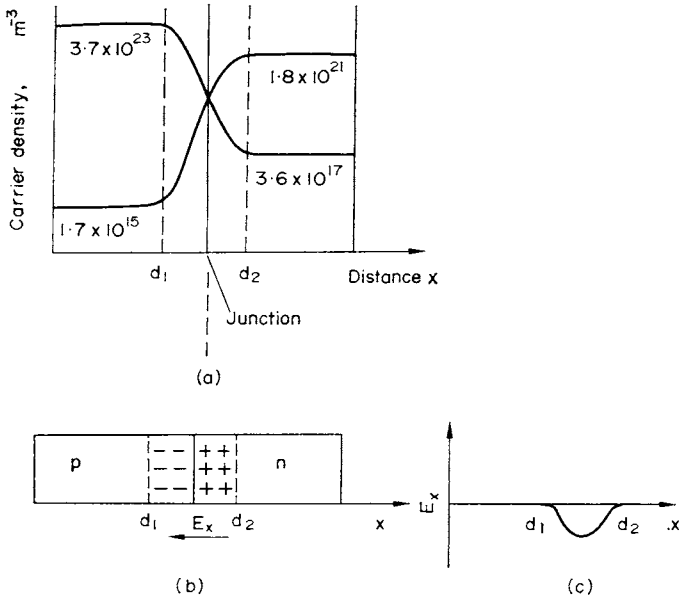


FIG. 4.5. (a) Distribution of carrier densities in equilibrium. The transition occurs over the distance d_1d_2. (b) Double charge layer set up at a p–n junction in equilibrium. (c) Variation of electric field E_x within the double charge layer.

charge layer forms at the junction as shown schematically in Fig. 4.5 (b) and this must give rise to an electric field across the junction, shown in Fig. 4.5 (c). The direction of the electric field is seen to be such as to try and prevent any further diffusion of holes and electrons. Diffusion will continue until the electric field force produced by the double charge layer just counterbalances the effective force on both holes and electrons produced by their respective concentration gradients. Reference to equation (3.22) shows that if the electric field at a point in the double charge layer is E_x, then for electric and diffusion forces to balance both for holes and electrons:

$$E_x = \frac{kT}{e}\frac{\mathrm{d}p}{p\,\mathrm{d}x} = \frac{kT}{-e}\frac{\mathrm{d}n}{n\,\mathrm{d}x} \tag{4.1}$$

where p and n are the hole and electron densities at the point where E_x is measured.

The double charge layer which forms on either side of a p–n junction is usually referred to as a space-charge layer, barrier layer or depletion layer. The name depletion layer arises since in this region the semiconductor is deplete in its normal majority carrier density. This is readily seen by comparing Figs. 4.4 and 4.5(a). The thickness of the depletion layer is shown in Fig. 4.5 (a) as the distance d_1d_2 and must extend over the region where the carrier densities are changing.

The thickness of the space-charge layer depends on the degree of doping on both sides of the junction. A typical width may be around 10^{-6} m; thus, the layer is very thin usually in comparison with the overall length of the diode. Referring back to Fig. 4.5 (b) once more, it can be seen that the edge of the space-charge layer d_2 in the n-region is more positive than the edge d_1 in the p-region. The redistribution of carriers across the junction has set up a potential difference across the junction. This potential difference is called a contact potential or diffusion potential and will be denoted by V_{pn}. Its magnitude depends on the degree of doping in the n- and p-regions but would typically be of the order of a volt or so.

Everywhere in the space-charge layer the diffusion force on the holes is finely balanced by the electric field force. Suppose that the restraining electric field could somehow be removed; there would then be a large diffusion current of holes from the p-region to the n-region. It is instructive to estimate the hole current density that might flow under these conditions. This can be done using equation (3.21) and the values of hole densities shown in Fig. 4.5 (a). The junction will be assumed to be about 10^{-6} m thick. If the rather crude assumption is made that the positive hole density varies linearly across the space-charge layer, then the density gradient $\mathrm{d}p/\mathrm{d}x$ is given by

$$\frac{\mathrm{d}p}{\mathrm{d}x} \approx \frac{3{\cdot}7\times10^{23}}{10^{-6}}.$$

Using equation (3.21), the hole diffusion current density J_p is seen to be

$$\begin{aligned} J_p &= eD_p\times(\text{concentration gradient}) \\ &= \frac{1{\cdot}6\times10^{-19}\times44\times10^{-4}\times3{\cdot}7\times10^{23}}{10^{-6}} \\ &= 2{\cdot}5\times10^{8}\ \mathrm{A/m^2}. \end{aligned}$$

The positive hole charge is $1{\cdot}6\times10^{-19}$ C and the diffusion coefficient for holes has been assumed to be 44×10^{-4} m^2/sec.

Now, $2{\cdot}5\times10^{8}$ A/m^2 is a very large current density; in actual operating conditions the current density across a *p–n* junction when an external voltage is applied is usually less than say 10^5 A/m^2. When no voltage is applied to the *p–n* junction this tendency for a large diffusion current to flow is counteracted by the electric field set up across the space-charge layer. It is not difficult to imagine from the above example that only a small unbalance is needed between diffusion force and electric field force to cause quite large current densities to flow. This is in fact what happens when a suitable voltage is applied across the *p–n* junction from an external source.

The diffusion current density has been calculated for holes only. A similar calculation can be made for electrons diffusing from the *n*-material to the *p*-material. Again the tendency for a large electron diffusion current to flow is counter-balanced by the electric field across the space-charge layer.

4.2.2. *The p–n Junction with an Applied Bias*

Suppose now a battery, voltage V, is connected across the *p–n* junction with the positive terminal of the battery connected to the *p* side of the junction. This is shown in Fig. 4.6. If the battery voltage were zero, then at first

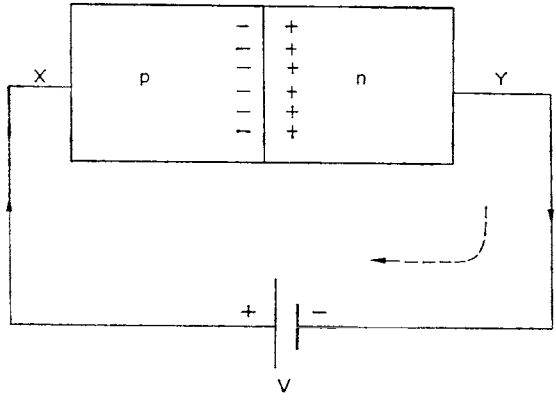


FIG. 4.6. A *p–n* junction with forward applied voltage V. Solid arrows show direction of conventional current flow.

sight it might appear that the contact potential at the junction between *p*- and *n*-type materials would drive a current round the external circuit in the sense depicted by the dotted arrow. This does not happen, however, because contact potentials also appear at the metal–semiconductor junctions *X* and *Y* (Fig. 4.6) due to the carrier concentration gradients at these junctions (the electron density in the metal contact, for example, is much greater than the electron density in either *n*- or *p*-type materials). The sense of the contact potentials appearing at *X* and *Y* are such as to cancel the contact potential appearing across the junction, and thus no current flows around the circuit.

Suppose now that the battery voltage is some finite value *V*. Conventional current flows from the positive battery terminal through the *p–n* junction to the negative terminal. Let us follow this process more carefully in order to see the carriers actually responsible for conveying the current round the circuit. Electrons will flow from the negative terminal of the battery along the lead wire and through the *n*-type material as majority carriers. Somewhere near the *p–n* junction this electron flow must be converted into positive hole current since any electrons flowing into the *p*-type material would very soon recombine with positive holes. The current is thus carried through the *p*-type material by positive holes (flowing from left to right). It is then transferred from positive holes back to electrons at the metal-*p*-type semiconductor junction in the manner described in Section 4.1. Figure 4.7 shows the direction and nature of charge transport around the whole circuit. The most important part of the circuit is the region around the junction between *p*- and *n*-type materials. In this region, as can clearly be seen from Fig. 4.7, electron current is converted into positive hole current. As we will attempt to show, the number of electrons per second per unit area that move into the junction area from the *n*-region and combine with a similar number of holes moving in from the *p*-region depends on the applied voltage *V*. Whilst the current depends on the applied voltage it is not a linear function of *V* as in the case of an ohmic resistance.

The applied voltage *V* causes a current to flow around the circuit. The current will produce a small ohmic resistance drop in the *n*-type and *p*-type materials. This drop, however, is usually negligible and practically all the applied voltage *V* appears across the space-charge depletion layer. Without

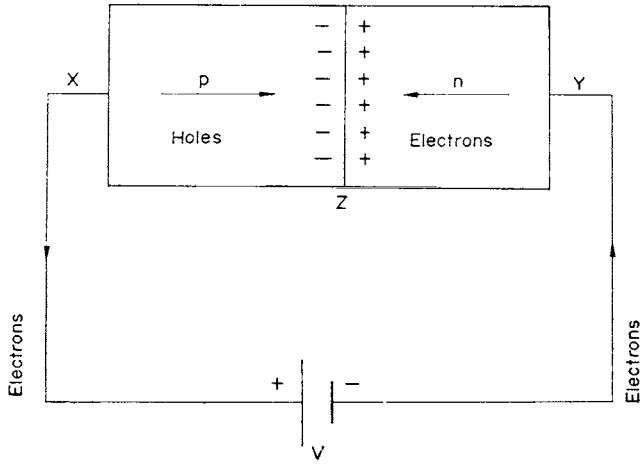


FIG. 4.7. Direction of motion of current carriers around a p–n junction biased with the p-type material positive (forward direction).

any applied voltage, as we have previously seen, the n-type material becomes V_{pn} volts positive with respect to the p-type. Since, to all intents and purposes, the whole applied voltage V may be considered to be dropped across the junction, the potential across the depletion layer is now V_j, where $V_j = V_{pn} - V$ volts. As the voltage across the depletion layer has fallen, it appears reasonable to believe that the electric field across the layer has also fallen. This means that the concentration gradient force tending to push holes to the right and electrons to the left (in Fig. 4.7) is greater than the electric field force across the layer tending to do the reverse. As a result of this, electrons are injected from the n-region into the p-region, whilst holes are injected from the p-region into the n-region. A current therefore flows across the p–n junction in the sense to be expected from the applied voltage V.

In Fig. 4.8 the majority and minority carrier densities are shown for the case when there is no applied bias ($V = 0$). The depletion layer is shown by the dotted lines and is very thin compared with the length of both p- and n-regions. Details of the transition of carrier densities in the layer have not been shown for the sake of clarity. The hole and electron densities in the p-material are p_p and n_p respectively and in the n material are p_n and n_n

respectively. The carrier densities are altered by the application of the bias voltage V. Since electrons are injected into the p-region, the electron density near to the junction in the p-region must rise above the equilibrium value n_p to some higher value n_{p0} say. Similarly, since holes move over into the n-region, the positive hole density in the region near to the junction must

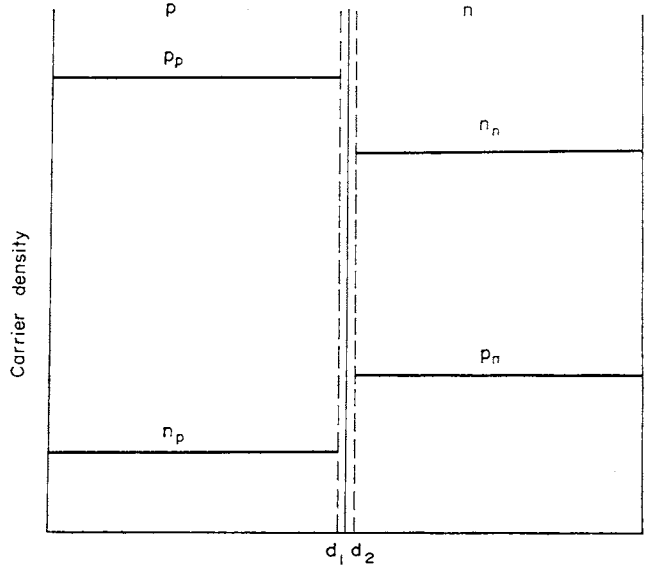


FIG. 4.8. Symbols for carrier densities with no applied bias.

also rise above the equilibrium positive hole density p_n to a value p_{n0} say. As the electrons injected into the p-material move towards the positive electrode they recombine with the vastly more numerous positive holes. The electrons injected into this region are minority carriers now. The electron density then falls from the value n_{p0} near the junction to the equilibrium value n_p. A similar effect occurs with the positive holes injected into the n-region. The positive hole density falls from the value p_{n0} at the depletion layer to the equilibrium value p_n as holes moving through the n-material towards the negative contact combine with electrons. The carrier density distribution under these conditions is shown in Fig. 4.9.

It will be noticed in Fig. 4.9 that the majority carrier densities p_p and n_n have not changed near to the barrier. Considering only holes for the moment, there is a very large reservoir of holes in the p-region and the effect of injecting some of these through the depletion layer into the n-region does

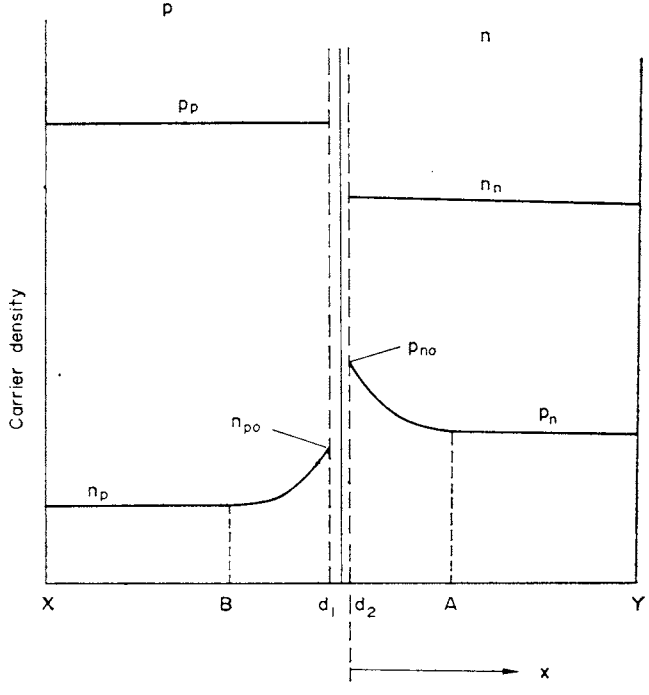


FIG. 4.9. Carrier densities in a forward biased p–n junction.

not perturb their density to any degree. It does of course affect the minority hole density in the n-region since there holes are scarce. Similar conclusions may be drawn regarding the electron density in the p-region.

Returning to Fig. 4.9, in the region Ad_2 the hole density is above the equilibrium value p_n. In the equilibrium region AY the rate at which electron–hole pairs are produced by the intrinsic process is g (see Section 3.5); i.e. g intrinsic bonds are broken per second per unit volume. The recombination rate of holes and electrons is given by rn_np_n (see equation (3.27)). Now these

values must be the same in the equilibrium region, otherwise there would be a net production or loss of electrons and holes. Thus,

$$g = rn_n p_n.$$

In the non-equilibrium region Ad_2, intrinsic hole–electron pairs are being produced at the same rate g, but since the hole density is greater than p_n it follows that the recombination rate for holes and electrons is now greater than $rn_n p_n$. Thus, more electron–hole pairs recombine per second in the region Ad_2 than are produced by the intrinsic process. In order to maintain a steady state in the region Ad_2 it follows that both holes and electrons must flow into it continuously. The holes move in from the p-region across the depletion layer as described above and the electrons move in from the region AY of the n-material.

Similar reasoning applied to the p side of the junction shows that electrons move into the region Bd_1 from the n-material through the depletion layer whilst holes move from the body of the p-region BX into the region Bd_1.

The conditions that apply in a non-equilibrium region have been discussed in Section 3.5.2 of Chapter 3 and the continuity equation derived there can be applied to the regions Ad_2 and Bd_1.

Concentrating for the moment on the holes injected into the n-region, the density variation with distance may be written, using equation (3.39), as:

$$p = (p_{n0} - p_n) \exp\left(\frac{-x}{L_p}\right) + p_n \tag{4.2}$$

where x is measured from the edge d_2 of the depletion layer in the sense shown in Fig. 4.9. It is readily verified that this equation gives the correct limiting values. At $x = 0$ equation (4.2) gives the value p_{n0} and as $x \to \infty$ gives the value p_n. The distance d_2A is seen to be approximately L_p in equation (4.2). This distance, the diffusion length for holes, is about 10^{-3} m both in silicon and germanium.

A similar law to (4.2) holds for the minority electron density in the p-type material. The corresponding diffusion length for electrons L_n is also about 10^{-3} m.

Suppose the electric field at any point x in the depletion layer is E_x and the density of electrons say at that point is n, then the concentration gradient

is dn/dx. The electric field force must balance the concentration gradient force almost exactly, otherwise, as has already been seen, currents of the order of millions of A/m^2 would flow through the depletion layer. The condition for balance is given by (4.1) and it will also be assumed to be true when a voltage V is applied to the p–n device.

From (4.1)

$$eE = \frac{kT\dfrac{dp}{dx}}{p}.$$

Multiplying both sides by dx and integrating across the depletion layer:

$$+e\int_{d_1}^{d_2} -E\,dx = -kT\int_{p_p}^{p_{n0}} \frac{dp}{p}$$

$$= -kT\log_e\left(\frac{p_{n0}}{p_p}\right). \tag{4.3}$$

Now $-\int_{d_1}^{d_2} E\,dx$ is equal to $V_{d2}-V_{d1}$, where V_{d1} is the potential at d_1 and V_{d2} the potential at d_2. The potential drop across the depletion layer $V_{d2}-V_{d1}$ is equal to $V_{pn}-V$. From (4.3)

$$p_{n0} = p_p \exp e\left(\frac{V-V_{pn}}{kT}\right) \tag{4.4}$$

$$= p_p \exp\left(\frac{eV}{kT}\right)\exp\left(\frac{-eV_{pn}}{kT}\right). \tag{4.5}$$

Now it is known that when $V = 0$, $p_{n0} = p_n$,

$$\therefore\ p_p \exp\left(\frac{-eV_{pn}}{kT}\right) = p_n. \tag{4.6}$$

Substituting (4.6) into (4.5)

$$p_{n0} = p_n \exp\left(\frac{eV}{kT}\right). \tag{4.7}$$

Thus, the minority carrier density of holes p_{n0} just outside the depletion layer increases exponentially with applied voltage.

If a value of T around room temperature (290 K) is taken, then

$$\frac{e}{kT} \approx 40 \text{ (volts)}^{-1},$$

$$\therefore \quad p_{n0} = p_n \exp (40\, V).$$

Suppose $\quad V = \frac{1}{10}$ volt,

$$\therefore \quad p_{n0} = p_n \exp (4)$$

$$= 50 p_n.$$

Thus, with an applied voltage as low as $\frac{1}{10}$ volt the minority carrier density is increased 50-fold above its equilibrium value due to the injection of holes from the p-region. Similar reasoning applied to electrons injected from the n-region gives:

$$n_{p0} = n_p \exp \left(\frac{eV}{kT}\right). \tag{4.8}$$

4.2.3. *Current in the Forward Direction*

The minority carrier densities n_{p0} and p_{n0} on either side of the barrier have been calculated when a bias voltage was applied to the junction by assuming that the current across the junction was zero. It may seem that this already precludes the possibility of calculating the current across the junction. However, more detailed calculation shows that equations (4.7) and (4.8) give, to a high degree of accuracy, the correct results even when allowance is made for current flow. This is to be expected from earlier deductions, namely that in all practical diodes the current flow across the junction is produced by the very small unbalance between electric field forces and concentration gradient forces. The current density is the small difference between the two large terms in equation (3.21).

Consider the n-region in Fig. 4.9 again. In the recombination region Ad_2 there is seen to be a concentration gradient of minority hole carriers. There is no electric field in this region since all voltage drops occur in the depletion layer $d_1 d_2$. The flow of electrons then in this region is produced by the con-

centration gradient alone. (See Section 3.5.2.) Note that the gradient is in such a sense as to cause holes to diffuse from left to right which, it is known, produces the correct direction of current flow.

From equation (3.21) the current density J_x of particles of charge q produced by a density gradient $\mathrm{d}N/\mathrm{d}x$ is

$$J_x = -Dq\frac{\mathrm{d}N}{\mathrm{d}x}.$$

Applying this equation to the holes in region Ad_2 and calling the diffusion current density J_{px} we have:

$$J_{px} = -D_p e p_n \frac{\mathrm{d}}{\mathrm{d}x}\left[\left\{\exp\left(\frac{eV}{kT}\right)-1\right\}\exp\left(\frac{-x}{L_p}\right)+1\right]$$

where the hole density is taken from equation (4.2) and p_{n0} from equation (4.7).

Performing the differentiation:

$$J_{px} = \frac{D_p e p_n}{L_p}\left[\left\{\exp\left(\frac{eV}{kT}\right)-1\right\}\exp\left(\frac{-x}{L_p}\right)\right]. \tag{4.9}$$

In particular, at $x = 0$ (i.e. at the edge of the depletion layer d_2)

$$J_{p0} = \frac{D_p e p_n}{L_p}\left[\exp\left(\frac{eV}{kT}\right)-1\right]. \tag{4.10}$$

Note that when $V = 0$, $J_{p0} = 0$, which is known to be correct.

This analysis may be repeated for the electrons diffusing from right to left in the p-type material. In particular, the electron diffusion current J_{n0} at the barrier is given by:

$$J_{n0} = \frac{D_n e n_p}{L_n}\left[\exp\left(\frac{eV}{kT}\right)-1\right]. \tag{4.11}$$

Now the total current density J_0 crossing the depletion layer is the sum of the electron current and the hole current,

i.e.
$$J_0 = J_{p0}+J_{n0}$$
$$= \left(\exp\frac{eV}{kT}-1\right)\left(\frac{D_p e p_n}{L_p}+\frac{D_n e n_p}{L_n}\right).$$

Writing
$$\frac{D_p e p_n}{L_p} + \frac{D_n e n_p}{L_n} = J_s, \tag{4.12}$$

$$\therefore \quad J_0 = J_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right]. \tag{4.13}$$

If the cross-sectional area of the junction is A, then the total current crossing the barrier is given by:

$$I_0 = AJ_0 = AJ_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$
$$= I_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right]. \tag{4.14}$$

J_s is called the saturation current density and I_s the saturation current. The current density at the barrier has been evaluated since this is the simplest place to use. However, it is known from Kirchhoff's law that the current must be constant at all cross-sections along the p–n device and equal to I_0.

The reader may notice that equation (4.9) shows the hole diffusion current falling with distance from the barrier layer as holes combine with

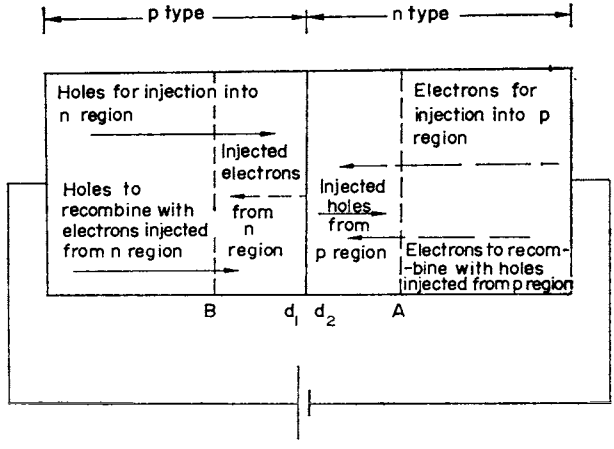


FIG. 4.10. Nature of charge transport in a forward biased p–n junction. Solid arrows (→) represent direction of positive hole flow, dashed arrows (⇢) represent direction of electron flow.

electrons in this region. Since electrons are combining with holes, more electrons must be moving in from the right to compensate for those lost in recombination otherwise a steady state would not persist. Thus, in the n-region, hole current injected through the depletion layer is eventually converted to electron current, whilst by a similar argument in the p-region, injected electron current eventually becomes hole current. These effects occur within the regions Ad_2 and Bd_1 respectively.

When the p side of the junction is connected to the positive terminal of the supply voltage V the current is given by equation (4.14). The current increases almost exponentially with applied voltage and conduction through the device is high. A voltage applied in this sense then is said to bias the junction in the *forward direction*. Figure 4.10 shows the nature of current in the various regions of the forward biased p–n junction.

4.2.4. *Current in the Reverse Direction*

If now the junction is biased in the opposite direction, i.e. the negative terminal of the supply is connected to the p side of the junction, then it would appear reasonable to replace V by $-V$ in equation (4.14). The circuit is shown in Fig. 4.11. Thus,

$$I_0 = -I_s\left(1-\exp\frac{-eV}{kT}\right). \tag{4.15}$$

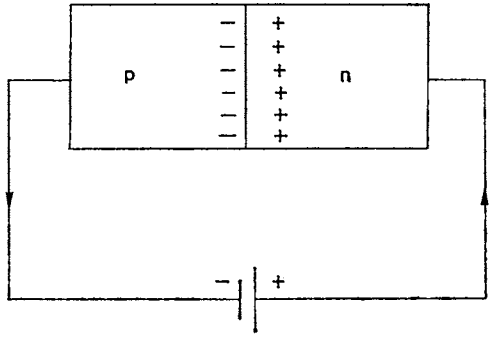


FIG. 4.11. A p–n junction with reverse bias. Arrows show direction of conventional current flow.

Now $\exp(-eV/kT)$ is always less than or equal to unity, thus the current given by equation (4.15) is always negative, i.e. always flows from the right to the left in Fig. 4.11. This, of course, is to be expected when the voltage is applied in the sense shown in Fig. 4.11.

The current is plotted in Fig. 4.12 for both senses of applied voltage using equations (4.14) and (4.15). Note that the current is always much lower for a given voltage when the junction is biased as in Fig. 4.11 and, moreover, it saturates for small voltages at the value I_s. If $V = -0{\cdot}1$ for example, $I_0 = -I_s(1-\frac{1}{50})$. The junction is said to be biased in the "reverse" or "back" direction. Experimental V/I graphs show good agreement with equation

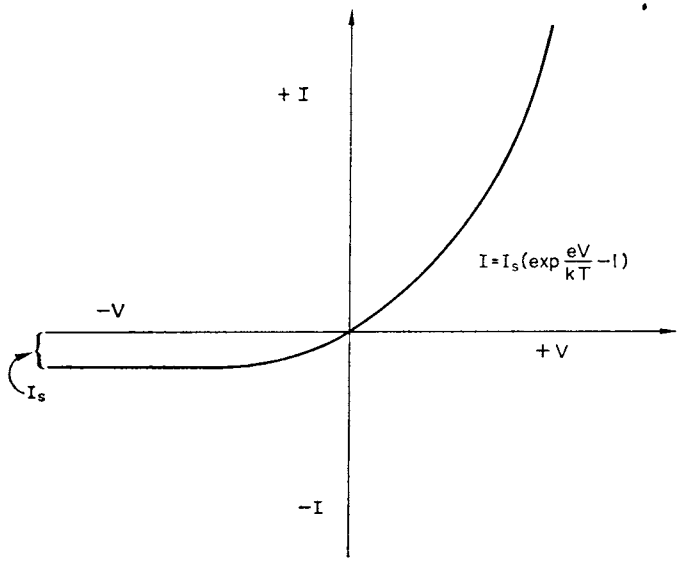


FIG. 4.12. The volt–ampere characteristic for a *p–n* junction diode.

(4.14) and (4.15). The rectifying properties of a *p–n* junction are apparent from Fig. 4.12.

The close agreement between experiment and theory gives good reason to believe that the analysis previously done for a forward biased diode also holds under reverse bias conditions with V replaced everywhere by $-V$.

In particular, the minority electron density just inside the p-region is from equation (4.7):

$$n_{p0} = n_p \exp\left(\frac{-eV}{kT}\right) \tag{4.16}$$

and the minority hole density just inside the n-region is from equation (4.7):

$$p_{n0} = p_n \exp\left(\frac{-eV}{kT}\right). \tag{4.17}$$

The variation of minority hole density with distance in the n-region from equation (4.2) is

$$p_n\left[\left\{\exp\left(\frac{-eV}{kT}\right)-1\right\}\exp\left(\frac{-x}{L_p}\right)+1\right] \tag{4.18}$$

and a similar relation holds for electron density in the p-region. The variation of density for holes and electrons is shown in Fig. 4.13 using equations (4.16), (4.17), and (4.18). As V becomes greater than kT/e,

$$n_{p0} \to 0 \qquad \text{from (4.16)}$$

$$p_{n0} \to 0 \qquad \text{from (4.17)}$$

This is the case shown in Fig. 4.13. A physical explanation of the effects depicted in Fig. 4.13 is as follows.

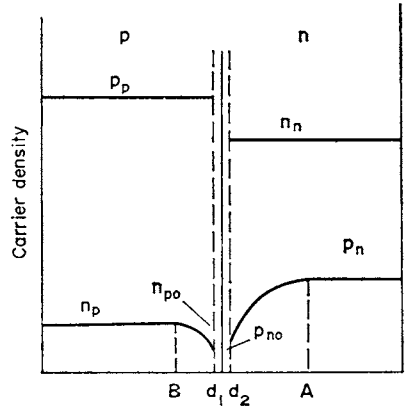


FIG. 4.13. Carrier densities in a p–n junction with reverse bias.

The bias applied in Fig. 4.11 appears across the barrier layer. The fine balance between diffusion forces and electric field forces in the barrier is slightly upset and positive holes are injected into the negatively biased *p*-region from the *n*-region. Likewise electrons are injected through the barrier into the *n*-region from the *p*-region. This causes the minority carrier densities to be reduced below the equilibrium values n_p and p_n near the barrier layer as borne out by Fig. 4.13. Concentrating attention on the *n* side of the junction, the holes near the barrier layer are swept from the *n*-material into the *p*-material and the density of holes then decreases near the barrier. A density gradient is produced which maintains a continuous drift of holes towards the barrier. Since there is a continuous flow of holes from the *n*-type material through the barrier to the *p*-type, it is necessary that within the *n*-region holes must be generated continuously to make good the loss of those injected through the barrier. The formation of hole–electron pairs by the intrinsic process takes place in the region marked d_2A in Fig. 4.13 at a greater rate than holes and electrons recombine in this region. The reason for this is just the opposite of that explained in Section 4.4. Thus, there is a continuous supply of positive holes for injection into the *p*-region and also a supply of electrons, since for every hole produced in this region a corresponding electron must appear. These electrons drift to the right

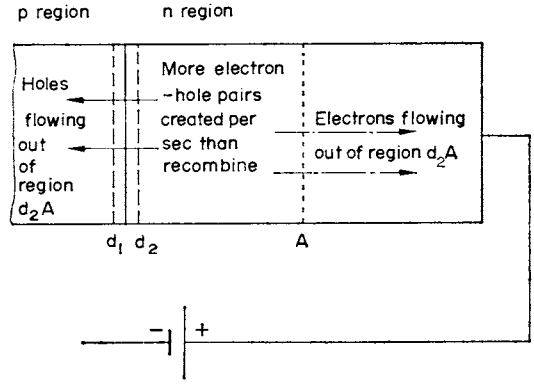


FIG. 4.14. Hole–electron pair production in the region d_2A. The generated holes move into the *p*-material; the electrons move towards the positive contact.

towards the positive terminal. The current to be injected through the barrier from *n*-material to *p*-material is carried by positive holes in the region near to the barrier, i.e. d_2A in Fig. 4.13. Moving away from the barrier d_2 to the right, the carriers of this current gradually become electrons. This is shown schematically in Fig. 4.14. As well as the current carriers just discussed, there are electrons injected from the *p*-material through the barrier into the *n*-material. Since the likelihood of electrons recombining in the *n*-region is small (there are few positive holes), this current flows through the *n*-region without significant change.

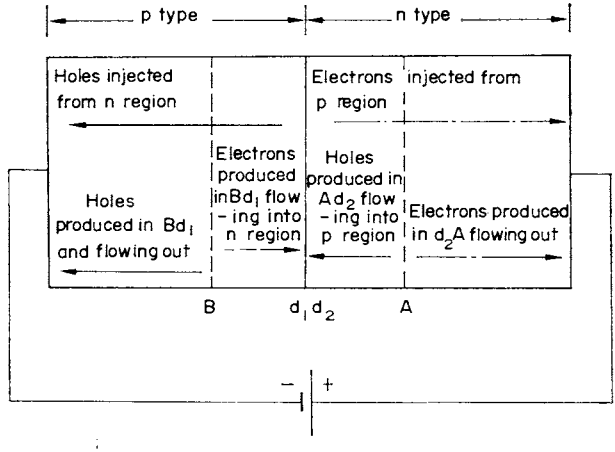


FIG. 4.15. Nature of charge transport in a reverse biased *p–n* junction. Solid arrows (→) represent direction of positive hole flow, dashed arrows (– – →) represent electron flow.

The above argument can be applied to the *p*-region, only the roles of holes and electrons are now interchanged. The complete picture with all the current components is shown in Fig. 4.15. Note that in the *n*-region the current is eventually carried by majority carriers (electrons) and in the *p*-region again by majority carriers for this region (holes).

4.2.5. *Relative Magnitudes of Hole and Electron Currents*

It is of interest later when transistors are discussed to have an idea of the relative magnitudes of the hole and electron currents that cross the barrier. Returning to equation (4.12), the ratio of

$$\frac{\text{Hole current injected in } n\text{-material}}{\text{Electron current injected in } p\text{-material}}$$

for a forward or reverse biased p–n junction is seen to be equal to

$$\frac{J_{p0}}{J_{n0}} = \left(\frac{D_p p_n}{L_p}\right)\left(\frac{L_n}{D_n n_p}\right)$$

$$= \left(\frac{L_n}{L_p}\right)\left(\frac{D_p}{D_n}\right)\left(\frac{p_n}{n_p}\right). \tag{4.19}$$

Now $$n_n p_n = n_p p_p = n_i^2 = p_i^2$$

where n_i and p_i are the intrinsic densities,

$$\therefore \frac{p_n}{n_p} = \frac{p_p}{n_n}.$$

Also from the Einstein relation (3.20)

$$\frac{D_p}{D_n} = \left(\frac{kT}{e}\mu_p\right)\left(\frac{e}{kT\mu_n}\right)$$

$$= \frac{\mu_p}{\mu_n}.$$

$$\therefore \left(\frac{D_p}{D_n}\right)\left(\frac{p_n}{n_p}\right) = \frac{\mu_p p_p}{\mu_n n_n}. \tag{4.20}$$

The conductivity of the p and n regions σ_p and σ_n are

$$\sigma_p = e p_p \mu_p, \qquad \sigma_n = e n_n \mu_n,$$

$\therefore$ from (4.19) and (4.20)

$$\frac{J_{p0}}{J_{n0}} = \left(\frac{L_n}{L_p}\right)\frac{\sigma_p}{\sigma_n}.$$

Now $L_n \approx L_p$, thus,

$$\frac{J_{p0}}{J_{n0}} \approx \frac{\sigma_p}{\sigma_n}, \tag{4.21}$$

i.e. the ratio of injected hole current to electron current is equal to the ratio of the conductivities of the *p*-material to the *n*-material. If the *p*-material is chosen to have a conductivity of 10^4 S/m and the *n*-material a conductivity of 10^2 S/m then the hole current injected into the *n*-material is a hundred times greater than the electron current injected into the *p*-material. The current across the depletion layer is thus almost completely carried by holes.

The same analysis leading to equation (4.21) applies also for reverse bias. Consider a reverse biased *p–n* junction in which the *n* material has still a conductivity of 10^2 S/m but the *p*-region is much less heavily doped and has a conductivity of 10 S/m. The positive hole current injected into the *p*-region from the *n*-region is only $\frac{1}{10}$ of the electron current injected into the *n*-region from the *p*-region. The current is carried across the junction largely by electrons. These examples are chosen because they are representative of the conditions that apply in a *p–n–p* transistor.

4.2.6. *Summary of Diode Action*

The action of a junction transistor is easily understood once the mechanism of charge transport in forward and reverse biased *p–n* junctions is understood. The reader is advised to re-read the previous sections of this chapter until he feels that he has grasped these fundamentals. The most significant features are summarized below:

(a) In a *p–n* junction with no bias voltage applied the equilibrium carrier densities in the *p* and *n* sections change abruptly over a fairly thin region (10^{-6} m) known as the depletion layer. A potential difference, called the contact potential, appears across this barrier. The *p*-material becomes negatively charged with respect to the *n*-material.

(b) In the forward bias condition the *p*-material is connected to the positive terminal of the source and the negative terminal is connected to the *n*-material. The applied voltage appears almost completely across the de-

pletion layer because of the high conductivity of the p- and n-materials and in this case reduces the voltage across the layer. The balance of diffusion and electric field forces in the depletion layer is altered, the diffusion force now being slightly greater than the electric field force. This causes a continuous flow of holes from the p-region into the n-region and conversely a flow of electrons from the n-region into the p-region.

(c) In the reverse bias condition the p-material is connected to the negative source terminal. The voltage across the barrier is thus increased and the electric field force in the barrier is now everywhere greater than the diffusion force. Electrons move through the barrier from the p-material to the n-material and conversely holes from the n-material move across the barrier to the p-material. The rate at which electrons can be produced in the p-material limits the maximum electron current that can be injected into the n side of the junction and similarly the rate at which holes can be produced in the n-material limits the maximum hole current injected into the p side. Thus, the reverse current density in a p–n junction saturates with increasing bias at the value

$$(D_p e p_n/L_p + D_n e n_p/L_n) \text{ A/m}^2.$$

4.3. THE JUNCTION TRANSISTOR

The junction transistor can take the form of a p–n–p or n–p–n combination. The p–n–p transistor consists of a thin wafer of n-type material sandwiched between p-type material and the converse for the n–p–n transistor. These are shown in Figs. 4.1 (b) and 4.1 (c). The thin region in the centre is called the transistor base; and the regions on either side are called the emitter and collector respectively. The emitter is usually characterized by having a conductivity greater than the base which in turn has a conductivity greater than the collector. These are all produced by different degrees of impurity doping. Some typical orders of magnitude are

conductivity of emitter region $= 10^4$ S/m

conductivity of base region $= 10^2$ S/m

conductivity of collector region $= 10$ S/m

The width of the base region is usually about 10^{-5} m and the diffusion length for holes, it will be remembered, is much greater than this distance, being about 10^{-3} m.

The circuit representation of a *p–n–p* and a *n–p–n* transistor is shown in Fig. 4.16. It will be seen that the direction of the arrows on the emitter leads is the only means of differentiation. This convention arises since the transistors are so biased under operating conditions that conventional emitter current always flows in the direction of the arrow.

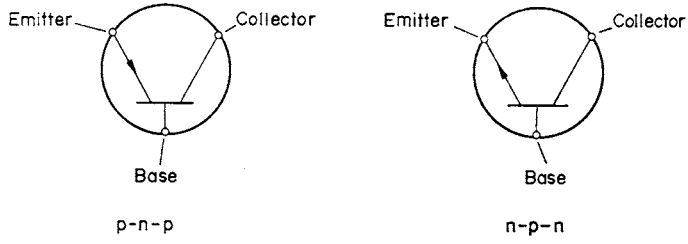


FIG. 4.16. Symbols for a *p–n–p* and *n–p–n* transistor.

Discussion will be restricted to a *p–n–p* transistor although the same arguments apply equally well to a *n–p–n* transistor with the role of electron and holes interchanged.

4.3.1. *Current Distribution in a Transistor*

To begin with it will be assumed that the base width w is much greater than the diffusion length for holes and electrons, and then the effect of decreasing the base width to a value much below the diffusion length will be examined. In the first instance there are essentially two *p–n* diodes back to back as shown in Fig. 4.17. Under actual operating conditions the emitter–base junction is always biased in the forward direction and the base–collector junction biased in the reverse direction. Since the conductivity of the emitter is a good deal greater than that of the base, the current across the emitter–base depletion layer is essentially a hole current produced by holes injected from the emitter into the base. This was shown to be the case in Section 4.2.5. It is usually sufficient to neglect the small electron current injected

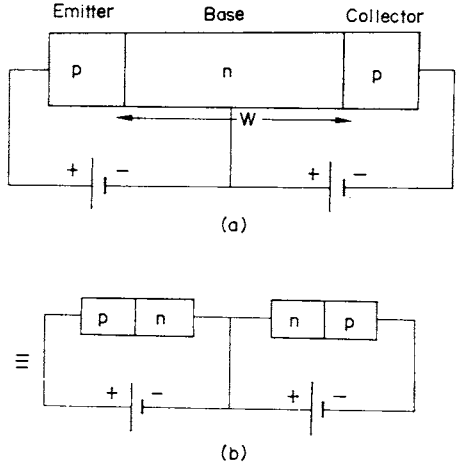


FIG. 4.17. A transistor with a wide base (a) is equivalent to two diodes connected back to back (b).

from the base into the emitter. Similarly since the base conductivity is greater than the collector conductivity, the flow of current across the base–collector barrier layer is to all intents and purposes electron current injected from the collector into the base. Current due to holes injected into the collector from the base can be neglected since it is very small. With these slight restrictions the current flow in the model having a very wide base region is shown in Fig. 4.18. The emitter–base current I_e is carried across the emitter–

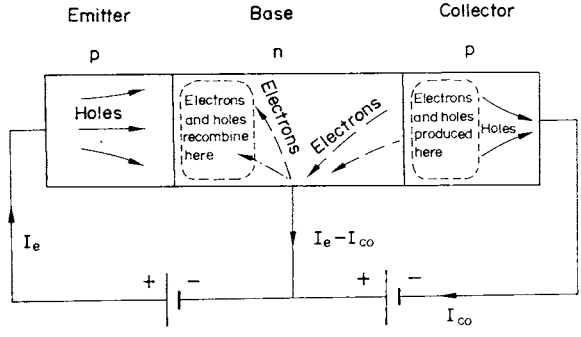


FIG. 4.18. Current carriers in a *p–n–p* transistor with a wide base region.

base depletion layer by holes injected from the *p*-region into the *n*-region. These holes recombine in the *n*-region thereby drawing electrons into the *n*-region from the base wire connection. The collector current I_{co} is carried by electrons injected from the *p*-type material into the *n*-type and these electrons then flow through the *n*-type material to the base wire connection. Since the emitter is forward biased and the collector reverse biased, $I_e \gg I_{co}$. Note also that of the two types of carriers injected into the base it is only the holes coming from the *p*-region that recombine to any degree in the base; the electrons injected from the collector are majority carriers in the base and since they find relatively few positive holes they are unable to recombine at all. The current leaving the base is $I_e - I_{co}$.

The variation of carrier charge density with distance is shown in Fig. 4.19. This is seen to be identical with two *p*–*n* junctions connected back to back.

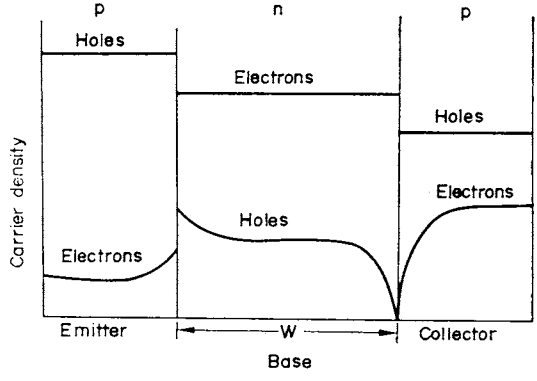


FIG. 4.19. Carrier density in a *p*–*n*–*p* transistor with a wide base. The depletion layers have not been included for clarity.

Now instead of making the base width *w* much greater than the diffusion length for holes, suppose it is made much smaller. The effect this would have on the electrons injected from the collector into the base is negligible since there is no electron recombination in the base. The current I_{co} circulating around the collector base circuit would be unaltered. The holes injected into the base from the emitter would not have much chance to recombine since the base is now shorter than the recombination length for holes. These

holes continue to drift across the base region under the hole concentration gradient force in the base until they reach the collector–base depletion layer. The electric field across the collector–base depletion layer is such as to drive these holes into the collector region (the collector is negative with respect to the base). Thus the positive hole current I_e injected from the emitter, which

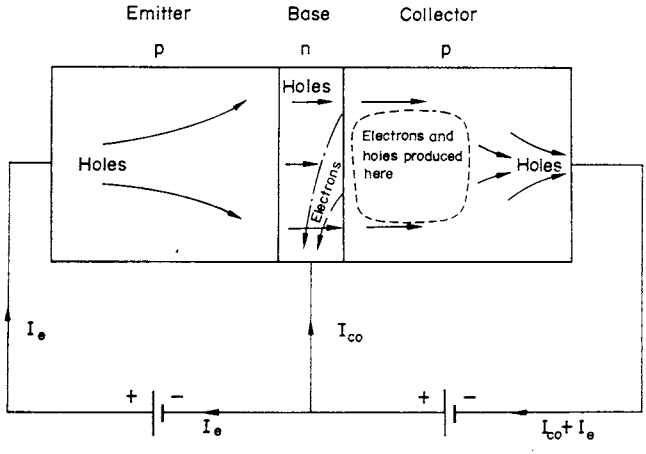


FIG. 4.20. Current carriers in a *p–n–p* transistor when the base region is very thin. There is no hole recombination in the base and all holes ejected by the emitter diffuse across to the collector.

previously flowed around the base–emitter circuit when the base region was wide, now continues on into the collector circuit. This situation is shown in Fig. 4.20. The carrier charge density distribution is shown in Fig. 4.21. This may be derived from Fig. 4.19 by imagining the base region to grow gradually narrower.

It has been assumed in Fig. 4.20 that no holes recombined in crossing the base. This will not be completely true unless the base region is infinitely thin. In practice, some of the holes injected from the emitter do recombine in the base and the hole current reaching the collector is less than the hole current at the emitter. Electrons must therefore flow into the base to replace those which have combined with holes. As will be seen later, for good transistor

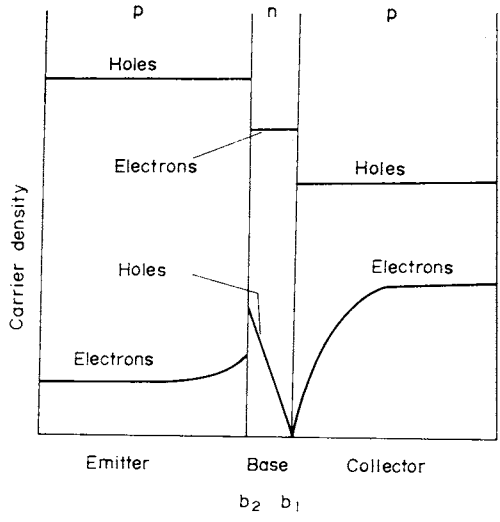


FIG. 4.21. Carrier density in a *p–n–p* transistor with a very narrow base. This figure may be derived from Fig. 4.19. by imagining the base region to contract.

action it is imperative that as much of the emitter hole current as possible should reach the collector. The ratio

$$\frac{\text{Emitter hole current reaching the collector}}{\text{Total hole current injected by emitter}}$$

is called the hole transport factor β. It is required to make β close to unity. From the two limiting cases considered it is obvious that $\beta \to 0$ as the base width w becomes much greater than the diffusion length for holes, whilst $\beta \to 1$ as $w \to 0$. In the former case all the injected hole current recombines in the base; in the latter case the base is sufficiently thin so that there is no recombination. The two methods of making β approach unity are:

(a) make the base very thin,
(b) make the free electron density in the base small so that the likelihood of recombination between holes and electrons is small. This amounts to making the conductivity of the base region small.

The small electron current injected from the base into the emitter and previously neglected flows around the base–emitter circuit. The ratio

$$\frac{\text{Emitter hole current into base}}{\text{Emitter hole current into base} + \text{Base electron current into emitter}}$$

is called the emitter efficiency γ and, as has been seen earlier, can be made almost unity if the conductivity of the emitter is made considerably greater

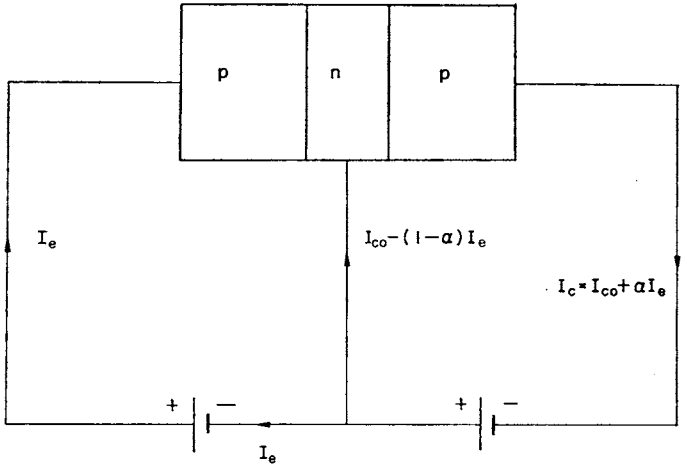


FIG. 4.22. Currents in a *p–n–p* transistor circuit. Direction of arrows shows conventional current flow.

than the base. If the total emitter current is I_e, from the definition of γ it follows that the hole current injected into the base is γI_e. A fraction of this current $(1-\beta)\gamma I_e$ is lost in the base by recombination and the remainder $\beta\gamma I_e$ flows into the collector. The total collector current I_c is thus:

$$I_c = \beta\gamma I_e + I_{co}.$$

Writing $\beta\gamma = \alpha$,

$$I_c = \alpha I_e + I_{co}. \tag{4.22}$$

The distribution of current under these conditions is shown in Fig. 4.22.

It will be seen that the emitter of the transistor and the cathode of a triode play a comparable role; whilst the base of the transistor and the grid of the triode are also somewhat analogous. A voltage applied between the base and the emitter controls the amount of current flowing into the collector circuit just as the voltage applied between grid and cathode of a triode controls the total current in the anode circuit. An added complexity arises in the transistor, however, since some minority carriers recombine in the base region and thereby cause current to flow into the base. The input impedance of the control circuit for a transistor is thus very much lower than for a triode. The field effect transistor has a high input impedance however.

Suppose that the base-collector bias voltage V_{cb} is kept constant and the emitter current is increased a small amount δI_e from its original value I_e. This could be done by increasing the emitter–base voltage. The collector current I_c will increase now by a small amount δI_c say. From equation (4.22)

$$\delta I_c = \alpha \delta I_e + \delta I_{co}. \tag{4.23}$$

But I_{co} will not change since it depends on the base–collector bias only and it has been stipulated that this is to be kept constant; thus $\delta I_{co} = 0$. The ratio

$$\left| \frac{\delta I_c}{\delta I_e} \right|_{V_{cb} \text{ constant}}$$

is called the current amplification factor or current gain and is seen from equation (4.23) to be equal to α.

From previous considerations it is obvious that α is less than unity. Typical values of α range from 0·95 to 0·99 for practical transistors.

The component of collector current I_{co} is called the collector leakage current. It is the current that flows around the reverse biased collector–base junction and is usually small in comparison with the emitter current I_e. When the collector–base voltage is more negative than about $\frac{1}{10}$ volt, I_{co} reaches its saturation value. Since a transistor is usually operated with the base–collector bias voltage much more negative than $\frac{1}{10}$ volt, it follows that I_{co} is independent of even quite large changes in this voltage.

Although the collector leakage current is usually only a small fraction of

the emitter current, there are certain transistor circuit arrangements, discussed in the next chapter, in which it is imperative to allow for it. In consequence I_{co} will be included in all instances.

4.4. THE TRANSISTOR AS AN AMPLIFIER

The current gain of a transistor, defined as the ratio of collector current change to emitter current change, is less than unity, and it is therefore not readily apparent that any useful gain can be obtained from the device.

Two possible ways in which voltage gain might be achieved, however, are:

(i) Inclusion of an impedance in the collector circuit which is larger than the impedance of the emitter circuit. It will be seen in the next chapter that this is possible. Then, even though the time-varying component of current in the collector circuit is not very different from that in the emitter circuit, the resultant voltage across the collector circuit impedance will be greater than that producing the current change in the emitter circuit. This is used in the grounded base amplifier.

(ii) By using the characteristic that the base current is small compared with the collector current. A change of voltage between emitter and base results in a change of both collector and base current, but the relatively large change of collector current may produce a larger voltage change in a suitable value impedance connected in the collector circuit than the voltage change in the emitter–base circuit. This is used in the grounded emitter amplifier.

These possibilities are considered and analysed in the next chapter.

5. *The Transistor as an Amplifier*

THERE are three usual methods of connecting a transistor, known respectively as:

(a) grounded (i.e. earthed) base,
(b) grounded emitter,
(c) grounded collector.

Figures 5.1 (a), (b) and (c) are simplified circuit diagrams for each of the above arrangements. The transistor is assumed to be a *p–n–p* type in each case. Note that in all these circuits the emitter–base junction is always forward biased (easy current flow) whilst the collector–base junction is always

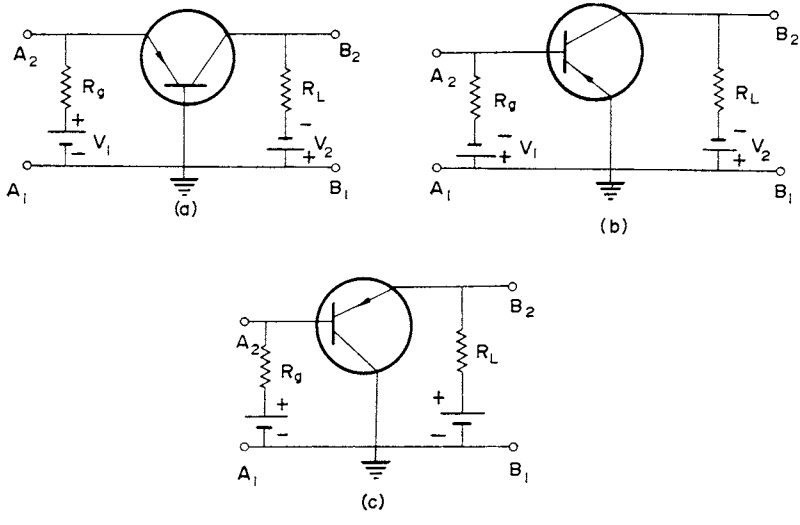


FIG. 5.1. (a) Grounded base *p–n–p* transistor. (b) Grounded emitter *p–n–p* transistor. (c) Grounded collector *p–n–p* transistor.

reverse biased (difficult current flow). The input terminals are A_1 and A_2, the output terminals are B_1 and B_2. R_g is the input resistor across which the signal voltage is applied and R_L is the load resistance across which the output voltage is taken. The grounded collector circuit is seldom used since it may be shown to have a voltage gain less than unity. The grounded base and grounded emitter circuits can have voltage gains much greater than unity and are therefore used more often. The grounded collector circuit will not be discussed any further in this chapter.

The grounded base circuit is described first of all because it is simpler to understand. It will be shown later that for many requirements the grounded emitter circuit has characteristics that make it preferable to the grounded base connection. The grounded emitter circuit is the most commonly used in practice.

5.1. THE GROUNDED BASE CIRCUIT

This circuit is shown in Fig. 5.2 for a *p–n–p* junction with the applied voltages and corresponding currents labelled. Note the direction of the current arrows shows the direction of conventional current flow; the heads

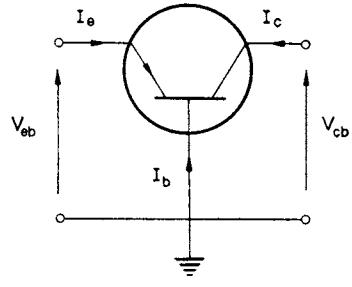


FIG. 5.2. Current and voltage labelling for a grounded base *p–n–p* transistor.

of the voltage arrows denote the positive voltage terminals. In practice the current in the collector circuit of a *p–n–p* transistor flows in the opposite direction to that shown in Fig. 5.2 since the collector is biased negatively with respect to the base, thus I_c and V_{cb} must be negative quantities. If, however, the transistor in Fig. 5.2 were a *n–p–n* type then the directions of

V_{cb} and I_c as drawn would be correct but the directions of current flow and bias voltage for the emitter circuit would be wrong.

In order to preserve some consistency, both emitter and collector currents are therefore considered to flow into the transistor, and emitter and collector voltages are measured with respect to earth (i.e. the base).

Four variables, the emitter–base and collector–base voltages V_{eb} and V_{cb}, and the emitter and collector currents I_e and I_c, are required to show the behaviour of a transistor.

The relationship between these four quantities is most easily seen by plotting two sets of characteristic curves. The first set, called the collector characteristics for grounded base (or sometimes output characteristics) is a family of curves relating I_c and V_{bc} for fixed values of emitter current I_e.

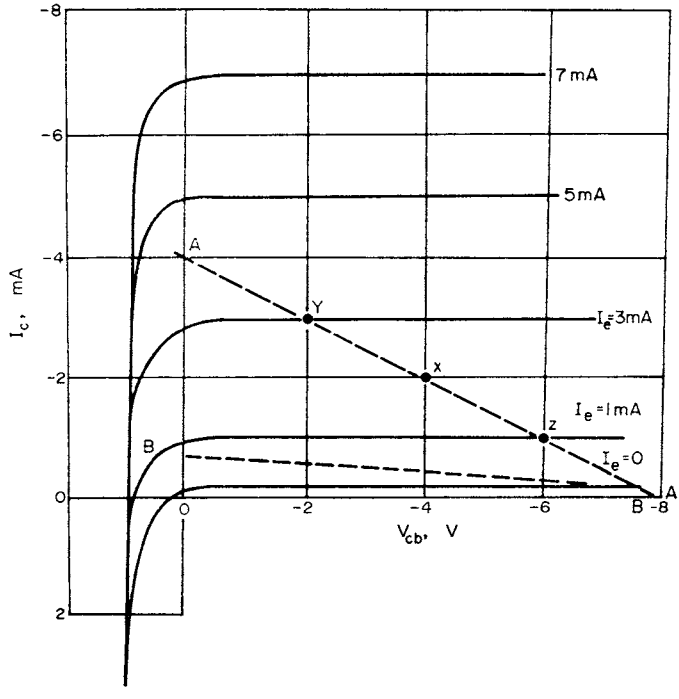


FIG. 5.3. Collector characteristic curves. Load line *AA* with collector load of 2 kΩ, *BB* with collector load of 10 kΩ.

A typical collector characteristic is shown in Fig. 5.3. The second set, or input characteristics, represents the variation of emitter current I_e as the emitter–base voltage is varied, keeping the collector–base voltage V_{bc} constant. A set of input characteristics is shown in Fig. 5.4.

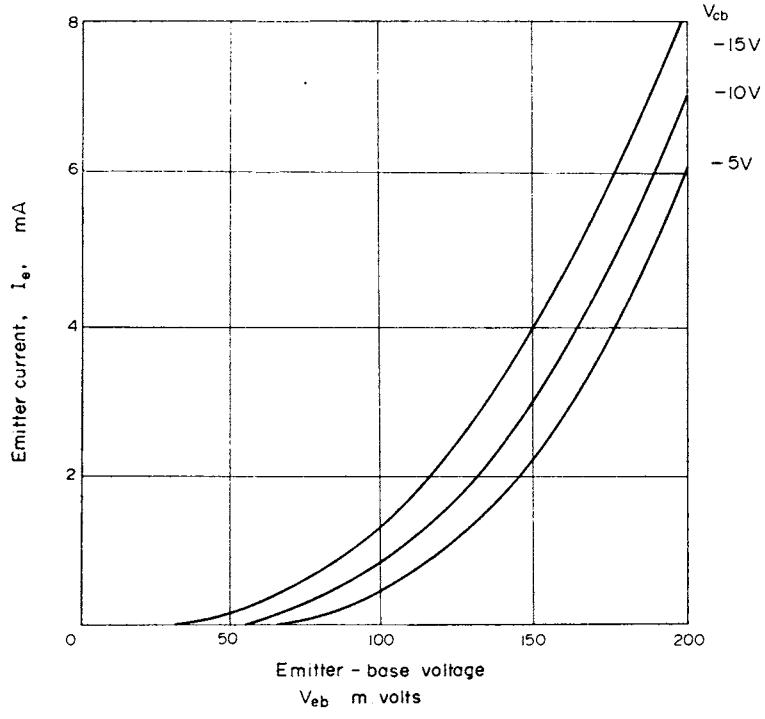


FIG. 5.4. Input characteristics for a *p–n–p* transistor in grounded base.

Consider, first of all in Fig. 5.3 the characteristic $I_e = 0$. This is the case when the emitter–base is open circuited. The characteristic is then just the V/I characteristic of a biased p–n junction. The region of interest is, of course, the reverse biased region and this is plotted to the right in Fig. 5.3. Note the collector current very soon saturates at a low value of reverse collector voltage. The value of collector current for $I_e = 0$ is the quantity previously referred to as I_{co}. Using equation 4.22, it is readily seen that the characteristic for any value of emitter current I_e is parallel to the $I_e = 0$ characteristic and

displaced upwards from it by an amount αI_e. A complete set of collector characteristics for reverse bias is obtained in this manner. The forward bias collector region is of little practical interest and it is left as an exercise for the reader to convince himself that the behaviour shown in Fig. 5.3 is as might be expected.

5.1.1. *Characteristics of a Grounded Base Amplifier*

The production of voltage gain can be illustrated by a simple approximate example.

Consider the amplifier circuit of Fig. 5.5. The emitter bias voltage V_e is, say, 2 volts and the collector bias voltage V_c is -8 volts, say. The input voltage is V_{in} and a resistance R_g is put in series with this voltage. The collector load resistance is R_L, and the output voltage V_{out} is taken across R_L.

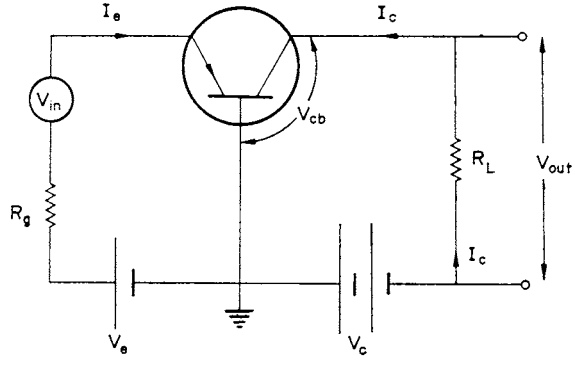


FIG. 5.5. Simple amplifier circuit for a *p–n–p* transistor in grounded base.

Considering the collector circuit first, the collector voltage is given by:

$$V_{cb} = V_c - I_c R_L.$$

Putting $V_c = -8$ volts

then
$$V_{cb} = -8 - I_c R_L. \tag{5.1}$$

Suppose $R_L = 2$ kΩ. The load line AA representing equation (5.1) can be plotted on the collector characteristic curves. This is shown as the dotted

line in Fig. 5.3, and I_c and V_{cb} must satisfy both the load line characteristic and the transistor characteristic.

Consider the emitter circuit now. Since the emitter is biased in the forward direction its resistance is low, typically around 50 Ω. Suppose that R_g is made much greater than the emitter resistance, say 1 kΩ. Then when $V_{\text{in}} = 0$ the emitter current I_e is approximately equal to

$$\frac{V_e}{R_g} = \frac{2}{1000} = 2 \text{ mA}.$$

The resistor R_g is chosen so as to set the transistor operating conditions at a convenient part of the characteristic, i.e. a linear part of the characteristic with an appreciable allowable input voltage swing and not too large an operating current.

Thus, the point marked X on Fig. 5.3 represents the conditions with zero input voltage V_{in}. Suppose now that V_{in} varies between $+1$ and -1 volt. Then the emitter current varies between $(2+1)/1000 = 3$ mA and $(2-1)/1000 = 1$ mA. These limits are shown by the points Y and Z in Fig. 5.3. The collector current is seen to vary from almost -3 mA to -1 mA, or more correctly -3α mA to -1α mA. The voltage across the load resistance R_L changes by

$$\frac{(3\alpha - 1\alpha)(2000)}{1000} = 4\alpha \text{ volts}.$$

The input voltage therefore has changed by 2 volts, the output by 4α volts. The voltage gain then is 2α which, since α is almost unity, is nearly 2. By putting $R_L = 10$ kΩ rather than 2 kΩ, the voltage gain would be 10. The load line *BB* for this condition is also shown dotted in Fig. 5.3. It is obvious that the emitter bias voltage V_{eb} would have to be much less than 2 volts in this case and also the applied voltage V_{in} could not be varied through anything like the range ± 1 volt without considerable distortion.

Transistors are, in general, current driven, i.e. it is necessary to produce a current which is proportional to the voltage to be amplified. This is most conveniently done in the circuit of Fig. 5.5 by putting in series with the input voltage a resistance large compared with the emitter–base input resistance. Suppose in Fig. 5.5 the resistance R_g were left out. Then the total

emitter–base voltage V_{eb} would be $V_e + V_{in}$. The emitter–base combination is a forward biased junction and the relationship between emitter current I_e and emitter voltage V_{eb} is characteristic of such a junction. Reference back to Section 4.2.3, equation (4.14) shows that I_e and V_e are related in the following manner:

$$I_e = I_s\left[\exp\left(\frac{eV_{eb}}{kT}\right) - 1\right].$$

The emitter current is not linearly dependent on V_{eb} and in consequence the output voltage of the transistor is not proportional to the input voltage. Distortion would result.

The reader may wonder why the collector characteristic curves are plotted for I_e constant rather than V_{eb} constant. This of course could be done, but the curves plotted in Fig. 5.3 are the most useful and the most easily interpreted. The family of curves would not be equally spaced if they were plotted for equal incremental changes in V_{eb} since V_{eb} and I_e are not linearly related.

The input characteristics. These curves have the typical exponential variation of I_e with V_{eb} that would be expected for a forward biased p–n junction

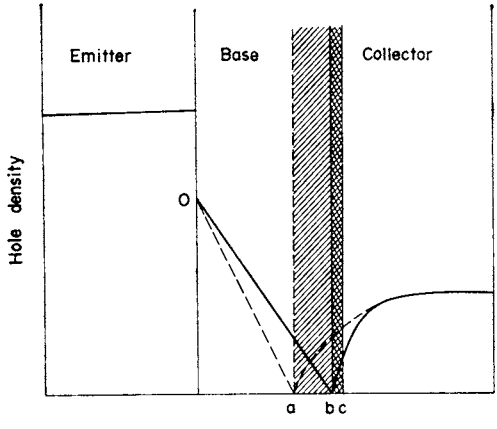


FIG. 5.6. Variation of hole-density gradient in the base region due to the "Early" effect. As the collector-base voltage is made more negative, the collector-base depletion layer widens from say *bc* to *ac*. The base hole density changes from the solid line *Ob* to the broken line *Oa*.

(i.e. the emitter–base junction). It is seen, however, that for a given value of V_{eb}, I_e increases slightly as V_{cb} is made more negative. The simple theory outlined in Chapter 4 suggests that I_e should depend only on V_{eb} and be independent of V_{cb}. Reference to Section 4.3.1 and Fig. 4.21 shows that the minority carrier density in the base at the emitter–base depletion layer depends on V_{eb}, whilst at the base–collector depletion layer it is almost zero. The hole density gradient in the base thus depends on V_{eb} and the width of the base region. The emitter current depends directly on this gradient through equation (3.21). If I_e is to increase then the concentration gradient in the base must increase and this can only happen, with V_{eb} kept constant, if the base region grows narrower, as shown in Fig. 5.6. More detailed treatment shows in fact that as V_{cb} becomes more negative the collector–base depletion layer widens and penetrates further into the base thus causing a corresponding reduction in base thickness. This phenomenon is called the "Early" effect after J. M. Early who first appreciated the mechanism responsible for the behaviour shown in Fig. 5.6. Reference to Fig. 5.4 shows, however, that the variation of I_e with V_{cb} is fairly small.

5.1.2. *Parameters for the Grounded Base Transistor Circuit*

It is convenient to define parameters relating the change in variables. The following parameters are found to be useful:

(i) *Collector resistance* r_c

This parameter defines the rate of change of collector current with base–collector voltage, with the emitter current kept constant, i.e.

$$r_c = \left| \frac{\delta V_{bc}}{\delta I_c} \right|_{I_e = \text{constant}} . \tag{5.2}$$

It is thus the reciprocal of the gradient of the collector characteristics. Reference to Fig. 5.3 shows that r_c is large, a typical value being of the order of $10^6\ \Omega$ (megohm).

(ii) *Grounded base current gain* α

This parameter defines the rate of change of collector current with emitter current, with the collector–base voltage kept constant, i.e.

$$\alpha = \left| \frac{\delta I_c}{\delta I_e} \right|_{V_{cb} = \text{constant}} . \tag{5.3}$$

The ground base current gain α would be unity if there were no recombination of minority carriers in the base. The value of α could be obtained from the collector characteristics by erecting an ordinate at the required constant value of V_{cb} and measuring the change in I_c for a given change in I_e. Strictly α, as defined above, is the modulus or size of the ratio and thus is always positive. However, as reference to Fig. 5.3 shows, when I_e increases, I_c decreases (i.e. becomes more negative); this point must be borne in mind when the equivalent circuit is derived in Section 5.1.4.

(iii) *Emitter resistance* r_e

This parameter defines the rate of change of emitter current with emitter–base voltage, with the collector–base voltage kept constant, i.e.

$$r_e = \left| \frac{\delta V_{eb}}{\delta I_e} \right|_{V_{cb} = \text{constant}} . \tag{5.4}$$

The emitter resistance is the reciprocal of the gradient of the input characteristic. It is not very dependent on V_{cb} but varies considerably with I_e as Fig. 5.4 shows. A typical value of r_e might be around 25 Ω.

It should be remembered that these three parameters depend on the bias conditions. The values of α and r_c, as can be seen from Fig. 5.3, change little with bias whilst r_e is particularly dependent on bias conditions, as Fig. 5.4 shows.

5.1.3. *Equivalent Circuit and Gain Calculations for Grounded Base*

Suppose now an alternating voltage, amplitude v_{eb}, is applied across the input terminals as shown in Fig. 5.7. The arrows on all a.c. quantities show their directions during a given half-cycle. In the next half-cycle, the directions

of all these arrows would have to be changed. A large capacitor C has been shown in series with the input connection. This prevents the emitter bias resistor R_g being short-circuited by the source impedance and yet provides a negligible impedance to the a.c. current. The a.c. voltage v_{eb} appears

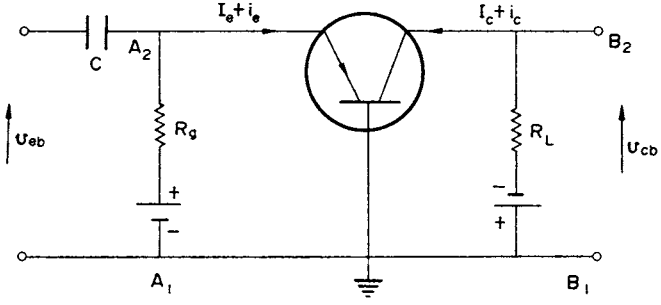


FIG. 5.7. Grounded base p–n–p transistor with a.c. input voltage v_{eb}. The output voltage is v_{cb}.

between the emitter and the base; thus the total emitter-base voltage is the sum of d.c. and a.c. voltages

$$V_{eb}+v_{eb}. \tag{5.5}$$

The a.c. emitter voltage causes the emitter current to become:

$$I_e+i_e \tag{5.6}$$

and the collector current to become:

$$I_c+i_c. \tag{5.7}$$

The voltage drop across R_L was originally $R_L I_c$ but is now

$$R_L(I_c+i_c). \tag{5.8}$$

Thus, the a.c. output voltage v_{cb} appearing across terminals $B_1 B_2$ is

$$v_{cb} = -R_L i_c. \tag{5.9}$$

Now there are two causes for the variation of collector current i_c:

(a) The emitter current has changed by i_e.

(b) The collector–base voltage has been changed by v_{cb}.

Thus,

$$i_c = -\alpha i_e + \frac{v_{cb}}{r_c} \tag{5.10}$$

i.e. if the collector–base voltage had not changed, then i_e itself would have produced an alternating collector current $-\alpha i_e$. Similarly, if the emitter current had been kept constant, a collector–base voltage v_{cb} would have produced an alternating collector current v_{cb}/r_c. Since these two effects occur simultaneously, equation (5.10) gives the correct relation.

Combining equations (5.9) and (5.10):

$$v_{cb} = \frac{\alpha i_e r_c R_L}{(r_c + R_L)}. \tag{5.11}$$

The collector circuit of the transistor is thus seen to be equivalent to a current source αi_e in parallel with both the load resistance R_L and the collector resistance r_c as shown in Fig. 5.8 (a). It will be seen that currents and voltages in this figure are related as in equation (5.11).

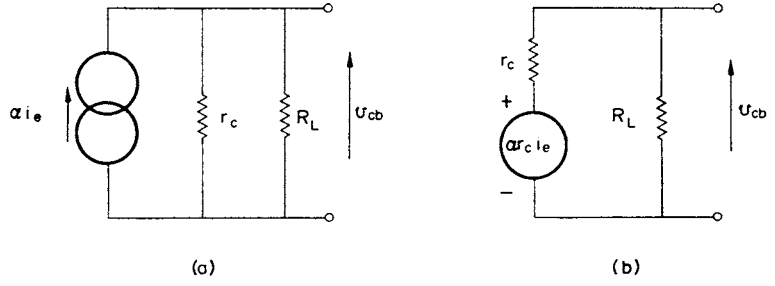


FIG. 5.8. A.C. equivalent circuits for a transistor. (a) Equivalent current generator circuit; (b) equivalent voltage generator.

An alternative equivalent circuit which will give this result is shown in Fig. 5.8 (b). A voltage generator $\alpha r_c i_e$ with series resistance r_c when connected across a load resistance R_L produces an output voltage v_{cb} such that

$$v_{cb} = \frac{\alpha r_c i_e R_L}{(r_c + R_L)}. \tag{5.12}$$

This is the same as equation (5.11).

Only the output circuit has been considered so far. For the input circuit, the emitter–base voltage drop v_{eb} may be conveniently divided into two components v'_{eb} and v''_{eb},

viz. $$v_{ec} = v'_{eb} + v''_{eb}.$$

The first component v'_{eb} is the voltage drop that would arise due to the emitter current i_e if the collector–base voltage V_{cb} were kept constant. However, the emitter current causes an a.c. voltage v_{cb} to be developed across the load resistance R_L. Reference to Fig. 5.4 shows that for constant I_e, V_{eb} changes slightly if V_{cb} changes, due to the "Early" effect. Thus an a.c. component of emitter–base voltage v''_{eb} arises from the a.c. collector voltage v_{cb}.

Thus $$v''_{eb} = -\left|\frac{\delta V_{eb}}{\delta V_{cb}}\right|_{I_e \text{ constant}} \times v_{cb}$$

Reference to Fig. 5.4 shows, however, that

$$\left|\frac{\delta V_{eb}}{\delta V_{cb}}\right|_{I_e = \text{constant}}$$

is small and in this instance will be neglected.

Assuming then that only v'_{eb} is significant:

$$v_{eb} \approx v'_{eb} = \left|\frac{\delta V_{eb}}{\delta I_e}\right|_{V_{cb} \text{ constant}} \times i_e$$
$$= r_e i_e. \qquad (5.13)$$

The complete equivalent circuit for the transistor in grounded base can now be constructed and is shown in Fig. 5.9. Fig. 5.9 (a) is the current source equivalent circuit; Fig. 5.9 (b) the voltage source equivalent circuit. The quantity $r_m = \alpha r_c$ is sometimes called the collector transfer resistance. Note that the load resistance R_L has been omitted from these equivalent circuits. This is because it is a feature external to the transistor. In deriving the expression for gain the load resistance has to be connected across the output terminals.

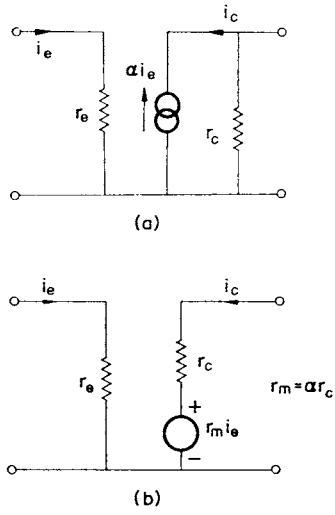


FIG. 5.9. (a) A.C. current source equivalent circuit including emitter circuit. (b) A.C. voltage source equivalent circuit including emitter circuit.

It is left as a simple exercise using either equivalent circuit in Fig. 5.9 to show that the following results hold:

$$A_v = \text{Voltage Gain} = \frac{v_{cb}}{v_{eb}} = \frac{\alpha r_c R_L}{(r_c + R_L) r_e}, \tag{5.14}$$

$$A_i = \text{Current Gain} = \frac{\text{current through load resistance}}{\text{input current}}$$

$$A_i = \frac{\alpha r_c}{(r_b + R_L)}. \tag{5.15}$$

$$A_p = \text{Power Gain} = \frac{\text{a.c. power into load}}{\text{a.c. power into transistor}}$$

$$= \frac{(\text{load current})(\text{load voltage})}{i_e v_{eb}}$$

$$= \frac{\alpha^2 r_c^2 R_L}{(r_c + R_L)^2 r_e} \tag{5.16}$$

Now it was stated earlier that r_c is about 1 MΩ for a typical transistor. It is generally true that $r_c \gg R_L$. Making these assumptions: from (5.14), (5.15), (5.16)

$$\text{Voltage gain} = \alpha(R_L/r_e) \tag{5.17}$$

$$\text{Current gain} = \alpha \tag{5.18}$$

$$\text{Power gain} = \alpha^2(R_L/r_e) \tag{5.19}$$

Taking, for example, $R_L = 1$ kΩ, $r_e = 25$ Ω say, $\alpha \approx 1$, the voltage gain and power gain are both around 40 whilst the current gain is just less than unity.

The ground base amplifier is characterized by having:

(a) high voltage and power gain,
(b) low current gain,
(c) low input impedance (25 → 100 Ω),
(d) high output impedance (1 MΩ).

Note also that the input and output voltage are in phase in a grounded base stage, i.e. when the input voltage increases, the output voltage also increases.

The equivalent circuits given in Fig. 5.9 are the simplest possible. There are equivalent circuits to be found in more advanced texts on transistor electronics containing many more elements than those in Fig. 5.9. These equivalent circuits represent the behaviour of a transistor more accurately than the circuits which have been derived, particularly at higher frequencies. The circuits of Fig. 5.9 will, however, be taken one stage further by including the base resistance and collector depletion layer capacitance. This latter parameter will be discussed when the frequency characteristics of transistors are considered.

The equivalent circuit of Fig. 5.9 (a) may be drawn as in Fig. 5.10 (a). A schematic diagram of the transistor itself is shown alongside. Now the base current i_b flowing along CD is given approximately by $-(1-\alpha)i_e$ and is a small fraction of i_e since α is near unity. This small current has to flow through the base region, however, which is very thin in a good transistor to minimize hole recombination. The ohmic resistance of the base therefore is quite high, possibly around 500 ohms. This can be allowed for by including a resistance r_b in the lead CD of the equivalent circuit. The modified

current and voltage source equivalent circuits are shown in Fig. 5.10 (c) and 5.10 (d).

Typical values for a low frequency transistor are

$$r_e = 25\ \Omega,$$
$$r_b = 500\ \Omega,$$
$$r_c = 1\ \text{M}\Omega.$$

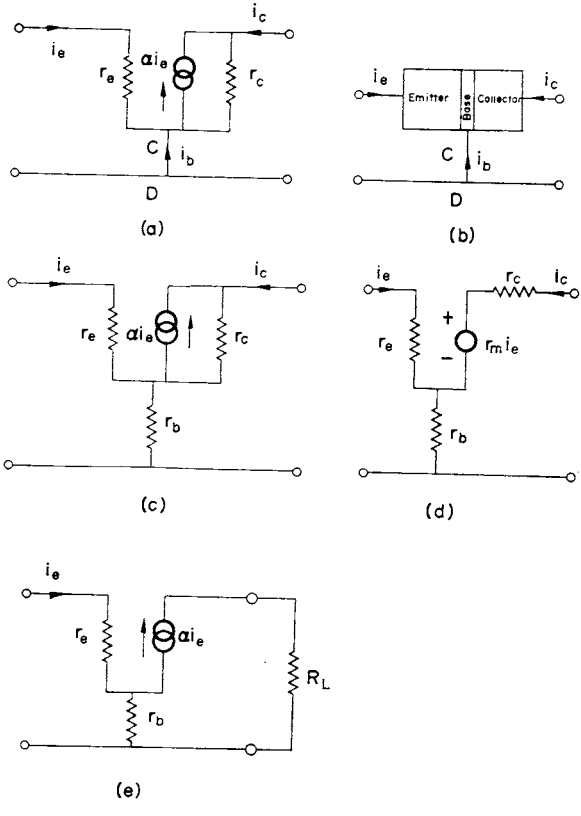


FIG. 5.10. Development of transistor equivalent circuit in grounded base. (a) Without base resistance. (b) Schematic layout of transistor. (c) Equivalent current generator circuit including base resistance. (d) Equivalent voltage generator circuit including base resistance. (e) Simplified current generator circuit to calculate voltage and current gain.

The inclusion of the base resistance does not significantly alter the gain formulae (5.17), (5.18), (5.19) for the grounded base connection. If the collector resistance r_c is assumed much greater than R_L, then the circuit given in Fig. 5.10 (e) may be used to calculate expressions for gain. Under these conditions it is readily seen that:

$$\text{Voltage Gain} = \frac{\alpha R_L}{r_e + r_b(1-\alpha)}, \tag{5.20}$$

$$\text{Current Gain} = \alpha, \tag{5.21}$$

$$\text{Power Gain} = \frac{\alpha^2 R_L}{r_e + r_b(1-\alpha)}. \tag{5.22}$$

Assuming $\alpha = 0{\cdot}98$, $r_e = 25\ \Omega$, $r_b = 500\ \Omega$, the reader can compare equations (5.17), (5.18), (5.19) with (5.20), (5.21), (5.22). The differences are not very great but the effect of including the base resistance is seen to reduce the voltage and power gain slightly.

5.2. THE GROUNDED EMITTER CIRCUIT

The grounded emitter circuit is shown in Fig. 5.1 (b) and also in Fig. 5.11 with the various currents and voltages labelled. Just as the static input and output characteristics were plotted for grounded base, so they can be

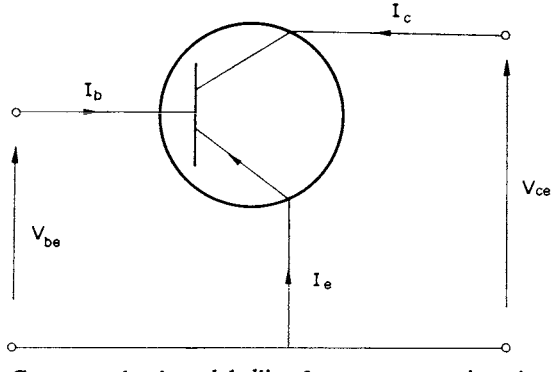


FIG. 5.11. Current and voltage labelling for a *p–n–p* transistor in grounded emitter.

obtained for grounded emitter. The output characteristic for a typical transistor is shown in Fig. 5.12.

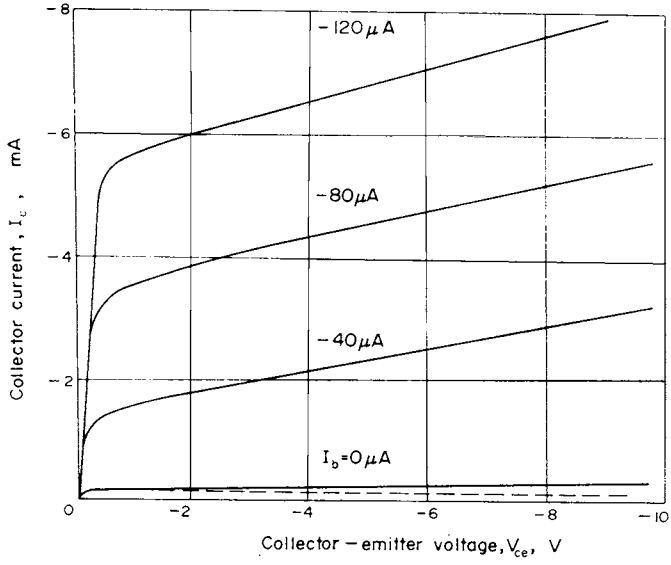


FIG. 5.12. Output characteristic for a *p–n–p* transistor with grounded emitter.

5.2.1. *Characteristics of a Grounded Emitter Amplifier*

The collector current I_c is plotted as a function of collector-emitter voltage V_{ce} for various fixed values of base current. To see how this characteristic is obtained the currents are separated into various components as shown in Fig. 5.13. The current I_e flows into the emitter circuit. The fraction of this reaching the collector is αI_e. The remainder $(1-\alpha)I_e$ flows out from the base. The base–collector junction is a reverse biased diode and the current flowing across this junction is shown as I_{co}, where I_{co} is the base–collector leakage current. Comparing Figs. 5.11 and 5.13

$$I_c = -I_{co} - \alpha I_e, \tag{5.23}$$

$$I_b = [I_{co} - (1-\alpha)I_e]. \tag{5.24}$$

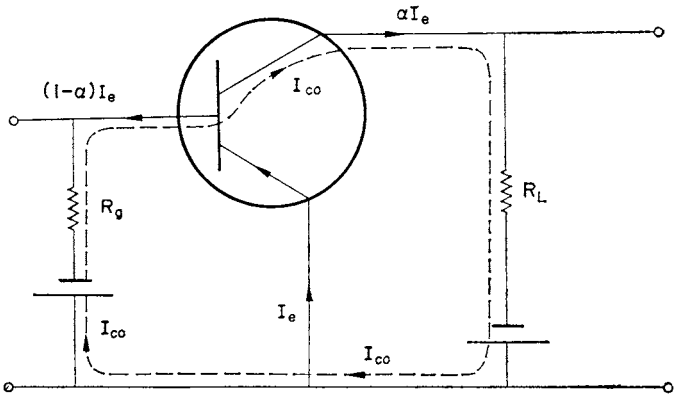


FIG. 5.13. Currents flowing in a *p–n–p* transistor with grounded emitter. The collector leakage current I_{eo} is shown by the broken line.

Eliminating I_e from (5.23) and (5.24)

$$I_c = \frac{\alpha I_b}{1-\alpha} - \frac{I_{co}}{1-\alpha}. \tag{5.25}$$

Now consider the characteristic $I_b = 0$.

In this case:
$$I_c = \frac{-I_{co}}{1-\alpha}. \tag{5.26}$$

As V_{ce} is made negative the base–collector leakage current I_{co} soon saturates. A representative value for the saturated value of I_{co} may be about 5 μA. However, it is noted from equation (5.26) that I_{co} has to be multiplied by $1/(1-\alpha)$ to obtain I_c. Putting in a value of $\alpha = 0{\cdot}95$

$$I_c = \frac{-I_{co}}{1-0{\cdot}95} = -20I_{co} = -\frac{1}{10}\text{ mA}$$

if $I_{co} = 5$ μA.

The transistor connected with grounded emitter is seen therefore to amplify its leakage current $1/(1-\alpha)$ times. For this reason, the collector leakage current is much more important in this circuit than in the grounded base connection. Using the above reasoning, the $I_b = 0$ collector characteristic

might be as shown by the dotted curve in Fig. 5.12. However, the actual characteristic is seen to rise as V_{ce} is made more negative. The explanation is given by the "Early" effect discussed previously in this chapter. As V_{ce} is made more negative the base–collector depletion layer widens and thus reduces the width of the base. This causes α to increase since there is less hole recombination in a thinner base, and from (5.26) it is seen that I_c depends very critically on the value of α when it is near unity. A small increase in α produces a large change in I_c.

The characteristics for constant values of I_b greater than zero are seen from the equation (5.25) to be the $I_b = 0$ characteristic plus a term $\alpha I_b/(1-\alpha)$. This is shown in the curves of Fig. 5.12. However, it should be noticed that the curves have a very pronounced knee near $V_{ce} = 0$ and all tend to pass through the origin eventually. As V_{ce} approaches zero the collector potential is approaching the ground. The base–collector junction becomes forward biased by the emitter–base voltage and the collector current therefore increases rapidly in the opposite direction to normal operating conditions. In Fig. 5.12, therefore, for values of V_{ce} below the knee, the values of I_c are changing direction rapidly since the base–collector junction is becoming forward biased.

Grounded emitter current gain α'. This is defined as

$$\alpha' = \left| \frac{\delta I_c}{\delta I_b} \right|_{V_{ce} = \text{constant}}$$

Again if the transistor is operated above the knee, then using equation (5.25) and remembering that I_{co} is constant if V_{ce} is held constant,

$$\alpha' = \frac{\alpha}{1-\alpha}. \tag{5.27}$$

If $\alpha = 0{\cdot}98$ say, then

$$\alpha' = \frac{0{\cdot}98}{0{\cdot}02} \approx 50.$$

Thus, whilst the current gain in grounded base is just less than unity, the current gain in grounded emitter is much greater than unity.

Input characteristic. For these characteristics I_b is plotted as a function of V_{be} for various values of V_{ce}. From equation (5.24)

$$I_b = [I_{co} - (1-\alpha)I_e].$$

Remembering that I_e varies as

$$\left[\exp\left(\frac{-eV_{be}}{kT}\right) - 1\right],$$

the characteristics shown in Fig. 5.14 are almost self-explanatory. As V_{ce} becomes more negative, α increases towards unity because of the "Early" effect. This is the principal reason for the characteristics being slightly dependent on V_{ce}.

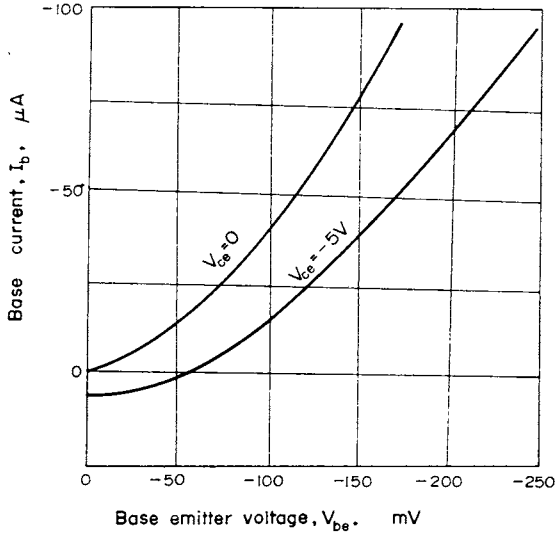


FIG. 5.14. Input characteristic for a *p–n–p* transistor in grounded emitter.

5.2.2. *Equivalent Circuit and Gain Calculations for Grounded Emitter*

In the grounded base connection shown in Fig. 5.1 (a) the values of the resistors R_g and R_L are chosen so that the transistor is biased suitably when no signal is applied. The procedure for choosing R_g and R_L was outlined

briefly. Similar considerations apply to the transistor in grounded emitter shown in Fig. 5.11. The load line is plotted on the collector characteristic Fig. 5.12 and a suitable value of base bias current is chosen. Since the base–emitter voltage is small the value of R_g is obtained by dividing the base–emitter bias voltage by the desired base current. This again is a similar procedure to that outlined earlier for the grounded base connection.

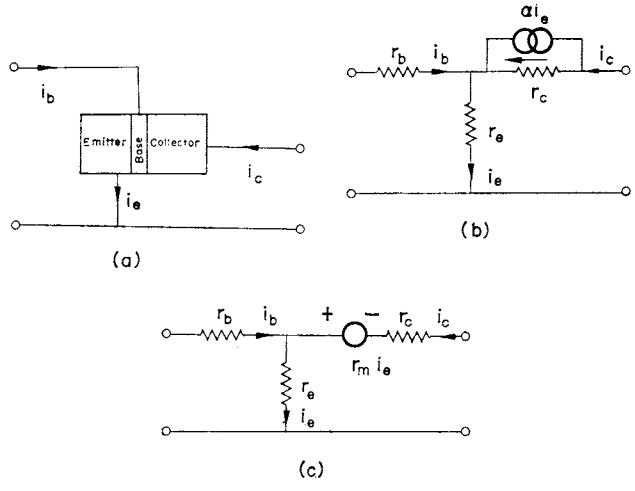


FIG. 5.15. (a) Schematic layout of a transistor in grounded emitter. (b) Equivalent circuit with current generator. (c) Equivalent circuit with voltage generator.

It is desired now to obtain an equivalent circuit for the transistor in grounded emitter. Figure 5.15 shows a schematic layout for the transistor and beside it a possible equivalent circuit. Notice that this is almost the same equivalent circuit as Fig. 5.10 except that r_e and r_b have changed places since the emitter is now grounded and the input signal is applied to the base. There is also one further important difference. The direction of the current generator αi_e has been changed. The reason for this is understood by carefully studying Fig. 5.1 (b). Suppose the base becomes more positive with respect to ground (i.e. the emitter), then the current flowing into the emitter will decrease since the emitter forward bias has been reduced. The collector current, flowing out from the collector, will decrease because the emitter

current has been reduced. Since the load resistance R_L is connected across the terminals B_1B_2 and the current through this load is reduced, the output voltage must decrease. Thus, an increase in input voltage causes the output voltage to decrease, i.e. there is 180° phase shift through a transistor in grounded emitter. To account for this the direction of the current generator has to be reversed.

The equivalent circuit with a voltage source is shown in Fig. 5.15 (c). Note again that the polarity of the voltage source is opposite to that in Fig. 5.10.

In order to see the significant features of the grounded emitter stage, imagine a generator voltage v_{be} connected across the input terminals and a load resistance R_L connected across the output. The circuit is shown in Fig. 5.16 (a). It will further be assumed that $R_L \ll r_c$. In this case r_c may be neglected and the circuit simplifies to Fig. 5.16 (b).

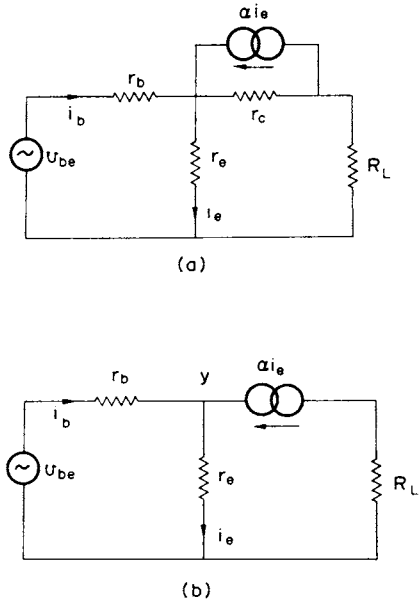


FIG. 5.16. (a) Equivalent current generator circuit with load resistance R_L.
(b) Simplified version of (a) when $r_c \gg R_L$.

Matching currents at the junction Y,

$$i_b+\alpha i_e = i_e,$$
$$\therefore\ i_b = i_e(1-\alpha). \tag{5.28}$$

Applying Kirchhoff's law around the base–emitter circuit

$$v_{be} = i_b r_b+i_e r_e. \tag{5.29}$$

Substituting for i_e from (5.28)

$$v_{be} = i_b r_b+\frac{i_b r_e}{(1-\alpha)}$$
$$= i_b\left(r_b+\frac{r_e}{1-\alpha}\right). \tag{5.30}$$

Thus, the input impedance r_{in} is given by $v_{be}/i_b = r_b+r_e/(1-\alpha)$. Putting $\alpha = 0{\cdot}98$, $r_b = 500\ \Omega$, $r_e = 25\ \Omega$, the input impedance is

$$r_{\text{in}} = 500+25\times 50 = 1750\ \Omega.$$

This is much higher than the value for a grounded base circuit where the input resistance is around $30\ \Omega$.

Current gain. The current gain A_i' is seen from Fig. 5.16 to be

$$A_i' = \frac{\alpha i_e}{i_b}.$$

Using equation 5.28 this becomes

$$A_i' = \alpha' = \frac{\alpha}{1-\alpha}.$$

For $\alpha = 0{\cdot}98$, $\alpha' \approx 50$.

This is in agreement with the results obtained earlier from the characteristic curves.

Voltage gain. The voltage gain A_v' is:

$$A_v' = \frac{-\alpha i_e R_L}{v_{be}}.$$

Using equations (5.28) and (5.30)

$$A_v' = \frac{-\alpha R_L}{(1-\alpha)\left(r_b + \frac{r_e}{(1-\alpha)}\right)} = \frac{-\alpha R_L}{r_e + (1-\alpha) r_b}. \tag{5.31}$$

Note that in the approximation $r_c \gg R_L$, the voltage gain in grounded base. (equation (5.20)) and voltage gain in grounded emitter are equal in magnitude.

Again supposing $R_L = 1\ \text{k}\Omega$, and using the previous values for r_e and r_b of 25 Ω and 500 Ω

$$A_v' = \frac{-(0{\cdot}98)(1000)}{25 + (0{\cdot}02)(500)}$$

$$= \frac{-(0{\cdot}98)(1000)}{35} = -28.$$

The minus sign signifies a 180° phase reversal.

Power gain.

The power gain $$A_p' = \frac{(\alpha i_e)^2 R_L}{v_i i_b}.$$

Using equations (5.26) and (5.28) again

$$A_p' = \frac{\alpha i_b^2 R_L}{(1-\alpha)^2 i_b^2 \left(r_b + \frac{r_e}{1-\alpha}\right)} = \frac{\alpha^2 R_L}{(1-\alpha) r_e + (1-\alpha)^2 r_b}. \tag{5.32}$$

Since the power gain is the product of voltage gain and current gain, it follows that the power gain of a grounded emitter state is $1/(1-\alpha)$ times greater than a grounded base stage.

Thus, the characteristics of a grounded emitter stage are:

Current gain	High
Voltage gain	High
Input impedance	Fairly high
Output impedance	Fairly high
Power gain	High

In the equivalent current generator circuit given in Fig. 5.15 (b), the emitter current i_e has to be used. This is sometimes inconvenient since it is i_b that really represents the input variable. The reader should convince himself that the circuit shown in Fig. 5.17 is equivalent to that in Fig. 5.15 (b). This circuit has the advantage that the generator current is given in terms of i_b and not i_e.

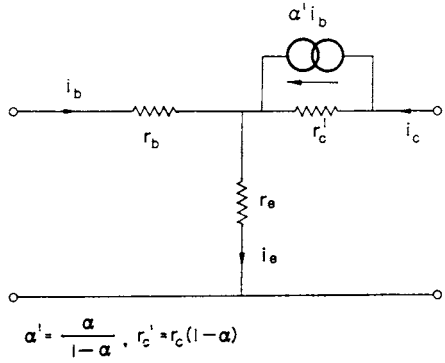


FIG. 5.17. Equivalent current generator circuit for a transistor in grounded emitter. This is equivalent to the circuit of Fig. 5.15 (b).

5.3. HIGH-FREQUENCY EFFECTS

Consider a grounded base *p–n–p* stage suitably biased. The density of minority carriers varies across the base region in the manner shown in Fig. 5.18. This figure is a section of Fig. 4.21. The carriers (holes in this case) diffuse across the base solely under the action of the concentration gradient. The collector hole current is proportional to the gradient of the density of minority carriers at the base collector junction X.

Now suppose that the emitter voltage is increased suddenly as shown in Fig. 5.19 (a). The number of holes injected into the base will suddenly rise and the density will eventually change to the new dotted line in Fig. 5.18. This will not happen immediately of course since it takes a finite time for holes to diffuse across the base. An intermediate distribution is shown by the line of crosses. The new steady state is reached in a time comparable to the transit time of holes across the base, τ say. The concentration gradient of

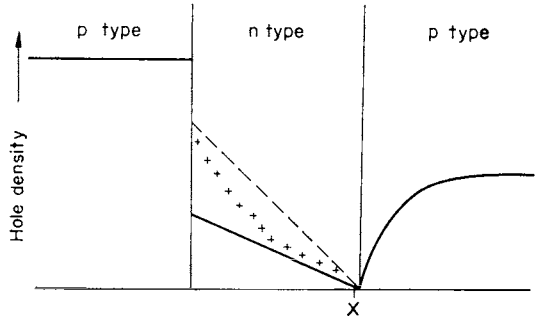


FIG. 5.18. Variation of hole density in the base of a *p–n–p* transistor when the emitter voltage is suddenly increased. Solid line represents initial density; broken line represents the new density after the increase in emitter voltage. The line of crosses represents an intermediate non-equilibrium condition.

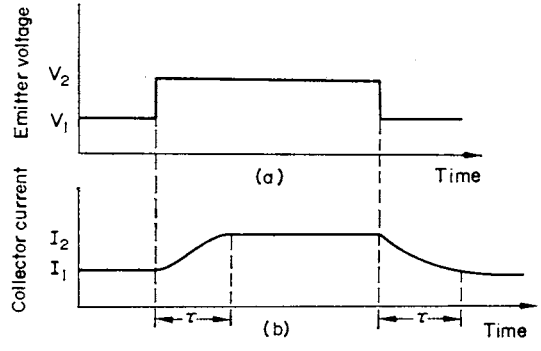


FIG. 5.19. Variation of collector current when the emitter voltage is suddenly increased or decreased.

holes at X is rising during this period so the collector current must rise during this time to its new steady state value. Figure 5.19 (b) shows the collector current increasing with time. The reverse occurs when the base voltage is reduced to its original value at some later time as shown also in Fig. 5.19 (a). If now the base voltage is increased and decreased with a period less than τ then the situation shown in Fig. 5.20 will arise. The collector current will not have time to reach its new equilibrium value before the voltage has

changed. The result then is distortion and a decrease in the collector gain α as the frequency of the applied signal is increased.

The current gain α will start to decrease rapidly when the frequency f of the applied signal approaches $1/\tau$. Thus, for good high-frequency response it is essential to make τ as small as possible. This is most readily achieved by making the base region as thin as possible. The frequency at which α has fallen to $1/\sqrt{2}$ of its low-frequency value is termed the alpha cut-off fre-

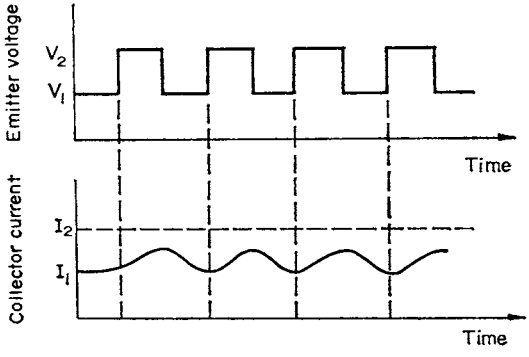


FIG. 5.20. Collector current when the emitter voltage changes more quickly than the hole transit time across the base.

quency. By utilizing special techniques, transistors can be made with alpha cut-off frequencies higher than 1000 MHz.

The high-frequency performance of a transistor is further degraded because of the capacitive effects associated with the depletion layers. Suppose a junction between *n*- and *p*-type material is considered. It will be recalled from Chapter 4 that there is a thin region at such a junction where electrons from the *n*-type material have diffused into the *p*-type and similarly holes from the *p*-type material have diffused into the *n*-type. This leaves positively charged donor atoms behind in the *n*-type material and negatively charged acceptor atoms in the *p*-type. The *n* side becomes positively charged and the *p* side becomes negatively charged. The dipole charge layer thus formed, and shown in Fig. 4.5 of Chapter 4, resembles a charged capacitor, the *n* side of the junction storing positive charge and the *p* side storing negative charge.

This capacitor is effectively in parallel with the junction. Figure 5.21 shows the modified current source equivalent circuit for a transistor in grounded base with depletion layer capacitances C_1 and C_2 included. The emitter–base capacitance C_1 is effectively shorted by the low emitter resistance r_e and is not therefore of much consequence. It is usually omitted from the equivalent circuit. The depletion layer capacitances vary with applied bias voltage but a typical mean value for the collector base capacitance may be around 10 pF.

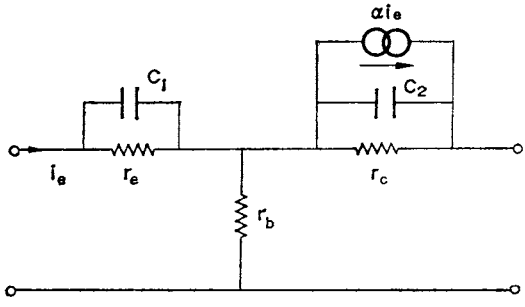


FIG. 5.21. Equivalent current source circuit for a grounded base transistor with depletion layer capacitances C_1 and C_2 included.

Figure 5.22 shows the equivalent circuit in grounded base with a load resistance R_L connected across the output terminals. The base resistance r_b has been left out for simplicity. If, as has been assumed previously, $r_c \gg R_L$, then it is obvious that the collector capacitance C_2 will start to short-circuit

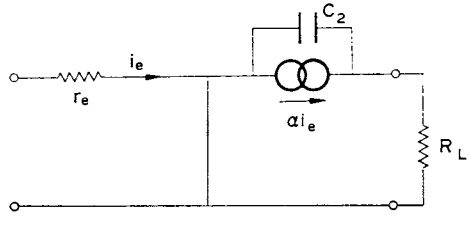


FIG. 5.22. Simplified equivalent circuit with load resistance R_L. If $fC_2/2\pi \prec R_L$, most of the injected hole current αi_e flows through C_2 rather than the load resistance R_L.

the load resistance seriously when its reactance $1/2\pi fC_2$ becomes comparable with R_L. The frequency at which these become equal is given by:

$$f = \frac{1}{2\pi C_2 R_L}. \tag{5.33}$$

Putting $C_2 = 10$ pF, $R_L = 1000\ \Omega$,

$$f = 16 \text{ MHz}.$$

Thus, with fairly low load resistances, the effect of collector capacitance is not great in the audio-frequency range. It can, however, become a serious limitation at higher frequencies in the same manner as the inter-electrode capacitances of triodes and pentodes.

5.4. ALTERNATIVE METHODS OF PRESENTING TRANSISTOR PARAMETERS

There are many other methods of presenting the small signal characteristics of a transistor rather than the equivalent circuits just derived. Suppose the transistor is represented by a "black box" so that it only presents two input and two output terminals to the outside world. This is shown in Fig. 5.23 with the appropriate currents and voltages labelled. The equations relating a.c. input and output voltages and currents may be written:

$$\begin{aligned} v_1 &= r_{11}i_1 + r_{12}i_2 \\ v_2 &= r_{21}i_1 + r_{22}i_2. \end{aligned} \tag{5.34}$$

The quantities r_{11}, etc., are called the resistance parameters (sometimes z's are used and they are then called impedance parameters). This is because the elements of the equivalent circuit become reactive as well as resistive

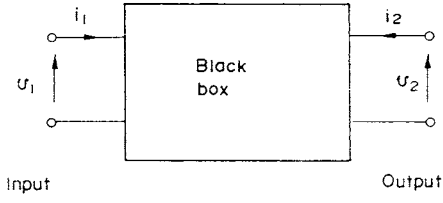


FIG. 5.23. Black box representation of a transistor.

at higher frequencies as noted in Section 5.3. They will depend on the manner in which the transistor is connected within the black box and also on the bias conditions. They are usually quoted by the manufacturer at suitable bias conditions, for grounded base and grounded emitter connections.

In theory all the resistance parameters at low frequency are obtainable from the input and output characteristics. It is difficult to measure them accurately, however, and the values quoted by the manufacturers are usually obtained by direct a.c. measurements at a representative frequency. This may be around 1 kHz for a low frequency transistor. The equivalent circuit parameters r_e, r_b, r_c and α are also expressible in terms of the resistance parameters.

An alternative set of equations to (5.34) is often used. These are:

$$\begin{aligned} v_1 &= h_{11}i_1 + h_{12}v_2 \\ i_2 &= h_{21}i_1 + h_{22}v. \end{aligned} \tag{5.35}$$

The quantities h_{11}, etc., are called hybrid or h parameters. The term hybrid arises since, as reference to (5.35) shows, h_{12} and h_{21} are pure ratios with no dimensions whilst h_{11} has the dimensions of ohms and h_{22} the dimensions of $(\text{ohms})^{-1}$ or S.

In order to help in remembering the significance of the h parameters, equations (5.35) are often written:

$$\begin{aligned} v_1 &= h_{ix}i_1 + h_{rx}v_2, \\ i_2 &= h_{fx}i_1 + h_{ox}v_2 \end{aligned} \tag{5.36}$$

where the subscript x is used to denote the circuit configuration; $x \equiv e$ normally being written for a common emitter configuration and $x \equiv b$ for common base. We can see from equation (5.36) that

h_{ix} = input impedance with the output short-circuited, so that $v_2 = 0$;

h_{ox} = output admittance with the input open-circuited, so that $i_1 = 0$;

h_{fx} = forward short-circuit current-transfer ratio, i.e. i_2/i_1 when $v_2 = 0$;

h_{rx} = reverse open-circuit voltage-transfer ratio, i.e. v_1/v_2 when $i_1 = 0$.

The h parameters can be obtained by performing the experiments suggested above; for example h_{fe} could be obtained by connecting the circuit in common emitter, short circuiting the output voltage at the frequency of measurement (this is most readily done by connecting a large value of capacitance across the collector–emitter terminal, so that the d.c. collector bias voltage is not shorted out) and then measuring the ratio of the a.c. collector current i_2 to the a.c. base current i_1. By measuring the a.c. input voltage on the base v_1 during this experiment, v_1 can be found and hence h_{ie} is readily obtained as the ratio v_1/i_1. If next the base is open-circuited to a.c., h_{oe} and h_{re} can similarly be measured.

Figure 5.24 shows the small signal equivalent circuit for a common emitter amplifier stage which obeys equation (5.36). By comparing equation (5.36) or Fig. 5.24 with the equivalent circuit given in Fig. 5.15 (b), the reader can

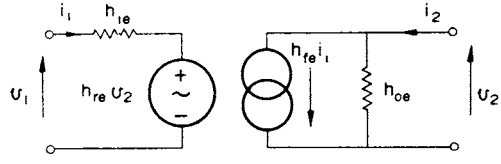


FIG. 5.24. Small-signal equivalent circuit for a common emitter amplifier using the hybrid parameters.

very readily show that the following identities exist, for example, between the physically identifiable parameters r_b, r_e, r_c and α and the h parameters for the common emitter configuration.

$$h_{ie} \approx r_b + \frac{r_e}{1-\alpha}, \tag{5.37a}$$

$$h_{fe} = \frac{\alpha}{1-\alpha} = \alpha', \tag{5.37b}$$

$$h_{re} = \frac{r_e}{r_e+(1-\alpha)r_c}, \tag{5.37c}$$

$$h_{oe} = \frac{1}{r_e+(1-\alpha)r_c}. \tag{5.37d}$$

Using the same values of α, r_e, r_b and r_c as taken in Section 5.1, we note that in this case:

$$h_{ie} = 500 + \frac{25}{0{\cdot}02} = 1750\,\Omega,$$

$$h_{fe} = \frac{0{\cdot}98}{0{\cdot}02} \approx 50,$$

$$h_{re} = \frac{25}{25 + (0{\cdot}02)\,(10^6)} = 1{\cdot}25 \times 10^{-3}$$

$$h_{oe} = 5 \times 10^{-5}\,\text{S}.$$

Very frequently the values of h_{re} and h_{oe} are sufficiently small that they can be neglected, although it is necessary to guard against assuming this statement is always valid for all circuits. For example, neglect of h_{oe} is only valid if the collector load resistor R_L is much smaller than h_{oe}^{-1}. However, as a general rule the small signal behaviour of a stage can be calculated with a fair degree of accuracy if h_{ie} and h_{fe} alone are known.

6. *The Transistor as a Switch*

6.1. THE TRANSISTOR AS A SWITCH

In the previous chapter the response of a transistor to small changes in base or emitter bias has been considered. Such analyses are referred to as small-signal and the equivalent circuit derived under these conditions is very useful in determining the behaviour of the transistor as a linear amplifier. It is, however, often advantageous to use the transistor as a switch where the collector current can rapidly be changed from a very small value (switch open) to a high value (switch closed).

The situation is depicted in Fig. 6.1 where a *p–n–p* transistor is operated in grounded-emitter configuration. Suppose the output characteristic given

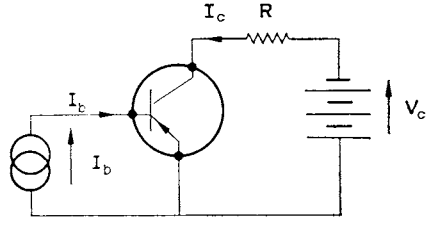


FIG. 6.1. *p–n–p* transistor in common-emitter configuration with base current drive.

in Fig. 5.12 holds. If the base current is zero then reference to Fig. 5.12 shows that the collector current is also very low and the transistor has a high resistance to current flow. If now the base current is suddenly increased to, say, 120 μA, Fig. 5.12 would suggest that the collector current would increase to 7 mA (if $V_{ce} = -6$ volts say). The transistor now has a much lower resistance to current flow.

This mode of operation can no longer be classed as small signal since large fractional changes of I_b and I_c take place. Moreover, our considerations

so far would tend to suggest that the step increase in collector current would follow instantaneously upon the step increase in base current. This is not observed to be so in practice; rather the collector current rises to its new steady state value in a finite time. The reason for this has been alluded to early in Section 5.3, and is because it takes a finite amount of time to establish the new steady state minority carrier density in the base needed to carry the increased collector current across the base.

The transient large signal response of transistors is important to understand since many circuits, particularly in electronic computers, are operated in modes where the transistors are required to switch from one state to another, having widely differing values of collector current, in as short a time as possible. The so-called "charge storage model" of the transistor which will be developed in this chapter is able to predict, with surprising accuracy, the transient response of transistors.

6.2. CHARGE STORAGE MODEL

Consider the common base configuration shown in Fig. 5.2 for a p–n–p transistor. The base region is assumed narrow in order that there be little opportunity for minority carrier recombination. Thus it is a good approximation to assume that the minority carrier distribution varies linearly across the base, as described in Section 4.3.1 and shown in Fig. 4.21. The hole density can thus be written:

$$\frac{dp}{dx} = p_{no} \cdot \left(\frac{w-x}{w}\right) \tag{6.1}$$

where w is the base width, p_{no} is the minority carrier density (holes in this case) at the emitter side of the base and x is distance measured from the emitter–base boundary. The value of p_{no} depends, as equation (4.7) shows, on the applied emitter–base voltage. At the base–collector junction $x = w$, equation (6.1) shows that the minority carrier density falls to zero (see Fig. 4.21) since the base–collector junction is strongly reverse biased.

The minority current across the base is carried solely by diffusion. From equation (6.1) the hole gradient is:

$$\frac{dp}{dx} = -\frac{p_{no}}{w} \tag{6.2}$$

and the corresponding hole current, using equation (3.21), is

$$I_p = \frac{eD_p p_{no} A}{w} \tag{6.3}$$

since holes move only under the action of the concentration gradient. A is the junction cross-sectional area and D_p is the diffusion coefficient for holes. This current I_p can be written as $-I_c$ where I_c is the collector current since it is normally a good approximation to neglect the reverse saturation leakage current I_{co} in comparison with I_c. The minus sign is consistent with the sign convention adopted in Fig. 5.2.

In the absence of any minority current injected into the base from the emitter, the hole carrier concentration is the equilibrium value p_n. Thus the excess concentration under injection is, from equation (6.1),

$$p_{\text{excess}} = p - p_n = p_{no}\left(1 - \frac{x}{w}\right) - p_n \tag{6.4}$$

and the total excess hole charge stored in the base is found by integrating (6.4) across the base. Let this charge be written as Q_F. Then from equation (6.4)

$$Q_F = Ae \int_0^w p_{\text{excess}}\, dx = Aep_{no} \int_0^w \left(1 - \frac{x}{w}\right) dx \simeq \frac{Aep_{no}w}{2} \tag{6.5}$$

where the normally valid assumption $p_{no} \gg p_n$ has been made and in this way p_n can be neglected.

It is convenient to write p_{no} in terms of the collector current I_c using equation (6.3). Thus from equations (6.5) and (6.3)

$$Q_F = \frac{-w^2 I_c}{2D_p}. \tag{6.6}$$

The quantity $w^2/2D_p$ has the dimensions of time; it is normally written as τ_F.

$$\tau_F = \frac{w^2}{2D_p}, \tag{6.7}$$

$\therefore$ from equations (6.6) and (6.7):

$$Q_F = -\tau_F I_c. \tag{6.8}$$

Equation (6.8) serves to relate the collector current to the total charge stored in the base by minority carriers.

It is instructive at this stage to obtain some estimation of the magnitude of τ_F. A typical base width might be $w = 10$ microns $= 10\times10^{-6}$ m, and $D_p = 10^{-3}$ m^{2}s^{-1} for holes in silicon.

$$\therefore \quad \tau_F = \frac{w^2}{2D_p} = \frac{10^{-10}}{2\times10^{-3}} = 5\times10^{-8}\,\text{sec.}$$

The base current I_b has now to be calculated. The base current is caused, as we have noted in Section 4.3.1, by electrons flowing into the base to recombine with some of the minority carriers that diffuse across the base from emitter to collector. If we denote the lifetime of holes in the base as τ_p (see, for example, problem 4.6, p. 254) then the number of electron–hole pairs recombining in the base per second is given by:

$$A\int_0^w \left(\frac{p-p_n}{\tau_p}\right)\mathrm{d}x.$$

The base current I_b is equal to the above quantity multiplied by the electronic charge $-e$, i.e. the total amount of charge per second flowing into the base to make good the losses due to recombination. Thus,

$$I_b = -Ae\int_0^w \frac{(p-p_n)}{\tau_p}\mathrm{d}x = \frac{-Aep_{no}w}{2\tau_p} = \frac{-w^2I_c}{2D_p\tau_p} \tag{6.9}$$

where again p_n has been neglected in comparison with p_{no} and equation (6.3) has been used. The collector current I_c in equation (6.9) can be replaced by the stored charge Q_F, using equation (6.6). Thus as an alternative to equation (6.9) we have

$$I_b = \frac{-Q_F}{\tau_p}. \tag{6.10}$$

6.3. TRANSIENT CONDITIONS

Under transient conditions equation (6.8) always correctly relates the instantaneous collector current and the instantaneous minority carrier charge stored in the base if the minority carrier distribution has the linear variation given by equation (6.1). We have seen in Section 5.3 that if the minority carrier density at the emitter–base junction is suddenly increased it takes a certain time for the collector current to respond. This time is that needed for the new linear variation in minority carrier concentration to establish itself (see Fig. 5.18) across the base and is obviously of a very similar magnitude to the time taken for minority carriers to diffuse across the base from emitter to collector. It is left as an exercise for the reader to satisfy himself that this time is of order τ_F. Thus equation (6.8) is always correct for transient effects which occur with characteristic times longer than τ_F. We have seen that a typical value for τ_F was about 50 nsec for a 10-micron base width. Thus equation (6.8) would be valid for analysis in this particular instance if one did not wish to study transients taking place in a time shorter than, say, a hundred nanoseconds.

Under transient conditions the base current, however, is not given by equation (6.10) alone. Suppose that at a certain time the value of the minority carrier density at the emitter–base junction is P_{no1}, then from equation (6.5) the stored minority charge is Q_{F1} where

$$Q_{F1} = \frac{Aew}{2} p_{no1}.$$

If now the emitter–base voltage is suddenly increased so that p_{no1} increases to P_{no2}, then the new charge stored in the base will eventually settle down at the new value

$$Q_{F2} = \frac{Aew}{2} p_{no2}.$$

But this will take a finite time of order τ_F as we have already seen. During this time the instantaneous minority carrier charge stored in the base will be changing. This in turn will cause the number of electrons drawn into the base in order to neutralize the increased minority charge, to change also.

Thus for every extra hole injected by the emitter into the base, an electron will be drawn into the base also, through the base terminal. The rate of change of minority carrier stored charge in the base is $\mathrm{d}Q_F/\mathrm{d}t$ and thus the rate of change of electronic charge flowing into the base to neutralize this is $-\mathrm{d}Q_F/\mathrm{d}t$. This then is the extra component of base current that has to be added to equation (6.10) to obtain the total base current. The resultant equation is:

$$-I_b = \frac{Q_F}{\tau_p} + \frac{\mathrm{d}Q_F}{\mathrm{d}t} \tag{6.11}$$

and from equation (6.8)

$$-I_c = \frac{Q_F}{\tau_F}. \tag{6.12}$$

The emitter current I_e is seen from Fig. 5.2 to be equal to $-(I_c+I_b)$ because of current continuity. Thus

$$I_c = -(I_c+I_b) = \frac{Q_F}{\tau_p} + \frac{Q_F}{\tau_F} + \frac{\mathrm{d}Q_F}{\mathrm{d}t}. \tag{6.13}$$

Equations (6.11) to (6.13) are the charge control equations relating the terminal currents to the total hole charge stored in the base Q_F. A further equation, however, is sometimes needed to relate the emitter–base voltage V_{eb} (see Fig. 5.2) to the total base charge Q_F, before we can proceed further. The required relationship can be obtained by noting, from Section 4.2.2, equation (4.7), that

$$p_{no} = p_n \exp\left(\frac{eV_{eb}}{kT}\right).$$

Using this in conjunction with equation (6.5)

$$Q_F \simeq \frac{eAwp_n}{2}\left[\exp\left(\frac{eV_{eb}}{kT}\right) - 1\right]. \tag{6.14}$$

Thus the stored charge in the base increases approximately exponentially with the emitter–base voltage.

Some care must be taken using these simplified charge control equations, since it has been assumed in their derivation that only minority carriers are

injected into the base at the emitter–base junction. This is only correct if the base–collector junction is at all times reverse biased, otherwise minority carriers might be injected from this junction also. In switching circuits this is not always the case; however, the charge control equations can be modified to take care of these conditions but they become rather more complex then. We will not at this stage consider such a situation.

6.4. AN EXAMPLE OF THE USE OF THE CHARGE CONTROL MODEL. STEP FUNCTION RESPONSE OF A COMMON-EMITTER CURRENT AMPLIFIER

In Fig. 6.1, a single common-emitter current amplifier was shown connected so that the base current I_b is produced by a current generator. We

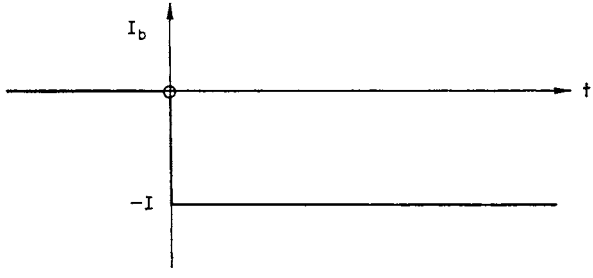


FIG. 6.2. Step function change of base current at $t = 0$ from zero to $-I$.

will assume that I_b suddenly changes from zero value to a constant value $-I$ at time $t = 0$, as shown in Fig. 6.2. Then, from equation (6.11)

$$\frac{Q_F}{\tau_p} + \frac{dQ_F}{dt} = I \quad (t > 0) \tag{6.15a}$$

$$Q_F = 0 \quad (t < 0) \tag{6.15b}$$

and the appropriate solution to these equations is:

$$Q_F = I\tau_p \left[1 - \exp \frac{-t}{\tau_p}\right] \tag{6.16}$$

so that from equation (6.12)

$$I_c = \frac{-I\tau_p}{\tau_F}\left[1-\exp\frac{-t}{\tau_p}\right]. \tag{6.17}$$

The collector current variation with time is shown in Fig. 6.3. The collector current is seen to change exponentially to the steady state value $-I\tau_p/\tau_F$. Thus the ratio τ_p/τ_F is just the current gain of the transistor in grounded emitter which we have previously denoted by α' or alternatively as h_{fe} in Chapter 5. A typical value for α' was seen to be $\simeq 50$.

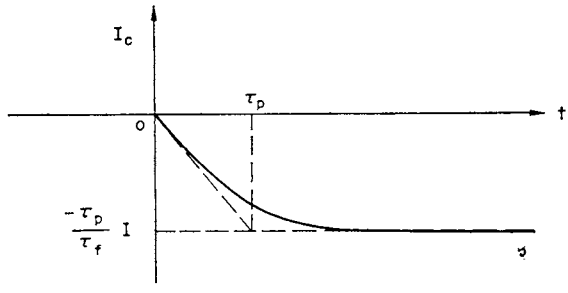


FIG. 6.3. Transient response of collector current to step function change of base current.

The collector current does not follow the sudden change in base current. After a time $t = \tau_p$ it has reached a value equal to $(1-1/e) = 63$ per cent of its steady state value. Thus the turn-on time of the transistor is approximately equal to τ_p. We have just noted that $\tau_p = \alpha'\tau_F$. Using the previously quoted values of $\alpha' \simeq 50$, $\tau_F \simeq 5\times10^{-8}$ sec, then $\tau_p = 50\times5\times10^{-8} = 2{\cdot}5\times10^{-6}$ sec. The transistor therefore takes at least 2·5 μsec in this case to switch on.

This turn-on time is, by present-day standards, comparatively long. Modern logic-circuits require turn-on times of perhaps a thousand times faster than this. It is obviously possible to reduce the turn-on time by reducing τ_F. Study of equation (6.7) shows that this is possible by reducing the base-width w, since $\tau_F \propto w^2$. However, ultimately there are technological limitations which prevent the base width being reduced beyond approximately 1 micron.

Another way of increasing the turn-on time is to use a speed-up capacitor in the base-current. This technique is described in the solution to problem 6.2.

A further technique is to over-drive the transistor. Reference to Fig. 6.4 shows that, if the collector load resistor is R, then the maximum collector current can never exceed the supply voltage V_c divided by R. Under these

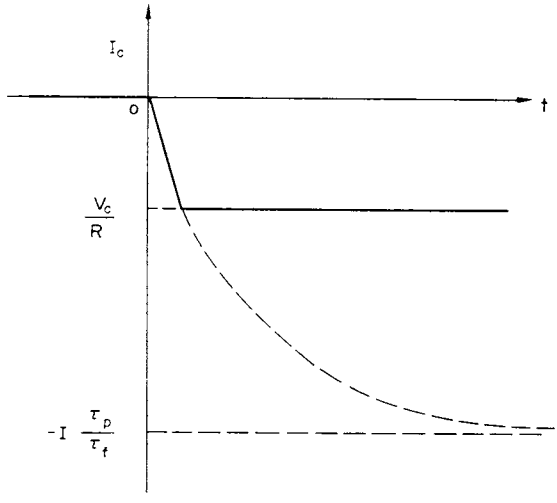


FIG. 6.4. Fast turn-on of a transistor collector current by base over-drive.

conditions the transistor is behaving like a short-circuit. If the base drive current I is such that

$$\frac{I\tau_p}{\tau_F} \gg \frac{V_c}{R}$$

then the collector current increases initially with time as if to reach the asymptotic value $I\tau_p/\tau_F$, but is soon limited to the maximum current that can be drawn from the supply, namely, V_c/R. The situation is shown in Fig. 6.4 and it is obvious now that turn-on is a lot faster than in the case shown in Fig. 6.3. In this case the transistor is not acting as a linear amplifier since

the collector current is no longer proportional to the base current. The transistor is behaving like a switch, being open circuited when $I_b = 0$, (OFF state) and then is rapidly switched to a short-circuit state by the base drive current (ON state). Under these latter conditions the transistor is said to be saturated.

7. *Field Effect Transistors*

THE field effect transistor, or FET, is in some respects more akin to the triode thermionic valve than to the transistor. It is a semiconductor device in which charge transport occurs through the lattice of a semiconductor crystal, but whereas the transistor is basically a current-driven device in which the output current is dependent upon the input current, the FET is a voltage-controlled device with negligible input current.

The FET can be considered to be a variable resistor in which the resistance can be varied by change of electrode voltage, with a consequent change of output current. The possibility of constructing such a solid state device was appreciated many years ago, but the actual device had to await progress in materials technology. Two main types of FET are available, the junction FET and the insulated gate FET. It will be seen that operation of these devices involves motion of charge carriers of one polarity, i.e. majority carriers, only. They are called "unipolar" transistors for this reason, and are thus unlike normal junction transistors which are "bipolar" and depend upon motion of both majority and minority charge carriers.

7.1. JUNCTION FET

The general form of a junction FET is shown in Fig. 7.1. It consists of a channel of doped semiconductor with ohmic contacts at each end, and with semiconductor junction electrodes known as gates formed along the edge of the channel. The gates are doped to give majority carriers of opposite polarity to those of the channel. The figure depicts the main body of *n*-type material and the gates of *p*-type. Silicon is usually used as the semiconductor. One ohmic contact is known as the source and the other as the drain. The device and its voltage supplies are shown schematically in Fig. 7.2

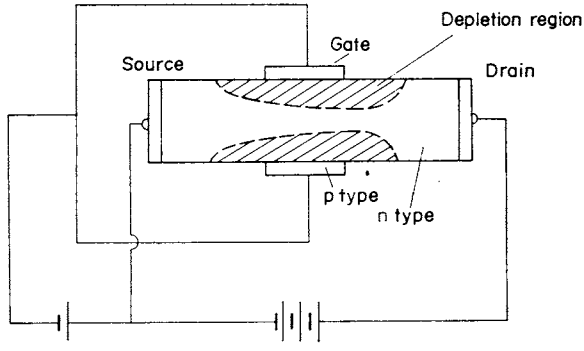


FIG. 7.1. Diagrammatic form of junction FET.

for an *n*-type channel. The arrow shows the direction of gate current if the gate-source diode were forward biased. For a *p*-type channel the arrow direction is reversed.

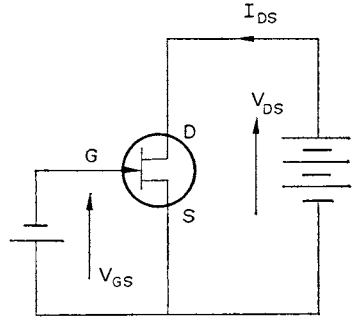


FIG. 7.2. Schematic diagram of a FET.

The *p–n* junction between gate and channel is reverse biased forming a depletion layer at the boundary as shown in Fig. 7.1. As discussed in Section 4.2.4, conditions of low free-charge carrier density exist in the depletion layer, which is therefore a low conductivity region. The action of the gate is thus to vary the width of the depletion layer when the gate voltage is changed, and therefore to reduce the channel conduction cross-section when

the gate reverse bias is increased. The gate voltage therefore effectively controls the main current through the device. An important feature is the low current drain from the gate, and corresponding high input impedance, which follows from the high resistance of a reverse-biased p–n junction. This is in contrast to the relatively low input impedance of a transistor.

It will be seen in Fig. 7.1 that the conducting channel is much narrower near the drain than at the source end of the device. This is a consequence of the voltage difference between source and drain, giving a larger reverse bias voltage at the drain than near the source. Even if the gate should be connected to the source, i.e $V_{GS} = 0$, there may be a considerable reverse bias on the p–n junction close to the drain, and this bias will increase as the drain-source voltage is increased. This effectively limits the current that can pass through the device for a given gate-source voltage as the drain-source voltage V_{DS} increases. The "saturation" current for the gate short-circuited to the source is known as I_{DSS}, and the voltage V_{DS} at which current limitation occurs for $V_{GS} = 0$ is known as the pinch-off voltage. Typical charac-

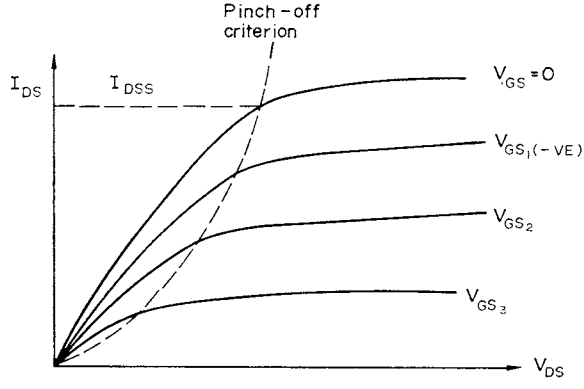


FIG. 7.3. FET characteristics.

teristics for a junction FET are shown in Fig. 7.3. For voltage V_{DS} values greater than the pinch-off level, the current through the device is dependent upon the gate voltage V_{GS}, but does not change significantly with the drain-source voltage. This is a desirable feature for voltage amplification since a

load resistor can be included in the drain circuit without seriously affecting the performance of the FET. It is then said to be operating in the pinch-off region.

Similar characteristics are obtained for a FET with a *p*-type main channel and *n*-type gate regions. Both gate and source are then positive in potential relative to the drain. Since electrons have greater mobility than holes, *n*-type channels tend to have higher conductivities than *p*-type.

Manufacture of the FET is by a process of deposition and diffusion of impurities on a *p*- or *n*-type silicon substrate to form the device on a single chip, e.g. as shown idealized in Fig. 7.4.

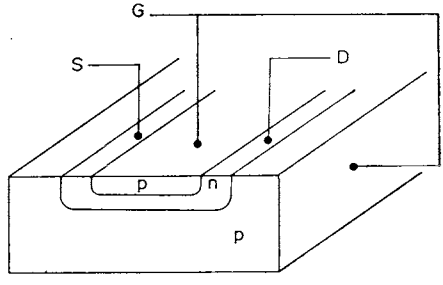


FIG. 7.4. Diffused form of FET.

7.2. THE INSULATED GATE FET

This device, known as an IGFET or MOSFET (metal–oxide–silicon FET), also operates by changing the resistance of a conducting channel. In this case a metallic gate electrode, insulated from the channel by a thin layer of silicon oxide, is used to vary the channel cross-section, instead of the reverse-biased *p*–*n* junction of the junction FET.

Consider successive layers of silicon oxide and metallic conductor on a *p*-type silicon substrate as in Fig. 7.5. When a positive voltage is applied to the metal conductor relative to the silicon, electrons are attracted to the interface region and holes are repelled back into the substrate. This redistribution of free-charge carriers may result in an excess of electrons over holes near the insulation layer, if the applied voltage is sufficiently large, i.e. a layer has been produced in the silicon which has been inverted to become

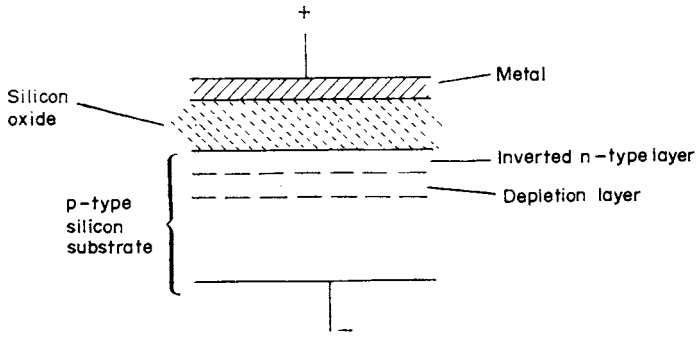


FIG. 7.5. Formation of MOSFET inversion layer.

n-type material. A depletion layer will then form between the *n*-type layer and the main body of the *p*-type silicon. The voltage required for inversion to take place is known as the threshold voltage. The larger the positive voltage applied to the metal contact, the deeper the *n*-type inversion layer. It is possible to make structures in which inversion of the surface layer occurs with zero applied voltage, i.e. the threshold voltage is equal to, or less than, zero. This phenomenon is due to a complex surface effect in which the lattice of the semiconductor is affected by the nearby dielectric layer. In this case an applied positive voltage to the metal increases the depth of the *n*-type region, and an applied negative voltage will decrease it.

The characteristics outlined enable an amplifying device to be constructed. Consider the device shown diagrammatically in Fig 7.6. A metal con-

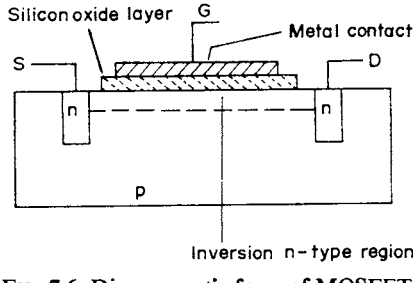


FIG. 7.6. Diagrammatic form of MOSFET.

tact layer is insulated from the *p*-type semiconductor by a silicon oxide layer, and two regions of *n*-type material are formed by diffusion near the ends of the metal/silicon oxide gate region. In the absence of an inversion region no current will result from the application of a voltage between the two *n*-type regions since one of the *p–n* junctions between them will be reverse biased. The formation of an *n*-type layer between them, however, will allow a current to result when a voltage between drain and source is applied. The greater the positive voltage to the gate, the greater the *n*-type layer depth, and therefore the greater the current. The characteristics of the inversion layer, and therefore the source-drain current magnitude, can be simply controlled by the gate voltage.

In a similar way a *p*-type inversion channel can be produced or changed in an *n*-type silicon substrate.

The representative diagram for the MOSFET shows the absence of a conducting link between the gate and the rest of the device, as in Fig. 7.7 (a). The presence of the substrate, which is normally connected to the source, is shown in the schematic diagram of 7.7 (b). Typical characteristic curves

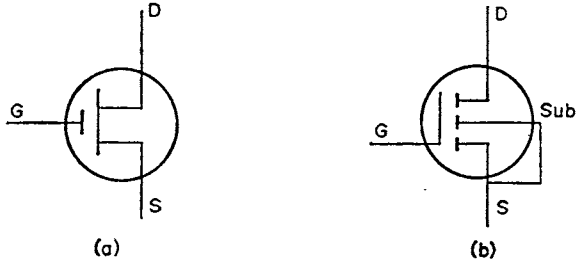


FIG. 7.7. Schematic diagrams of MOSFET.

are given in Fig. 7.8. One difference between these characteristics and those for the junction FET is that the gate may be operated with either polarity over a limited range. The junction FET is normally operated with the gate reverse biased. The insulated gate FET has the gate isolated physically by an insulation layer, and the inversion layer may exist with zero gate-source voltage. In this case, for a *p*-type material device (*n*-type conducting channel) a positive voltage on the gate will enhance the *n*-channel and the device

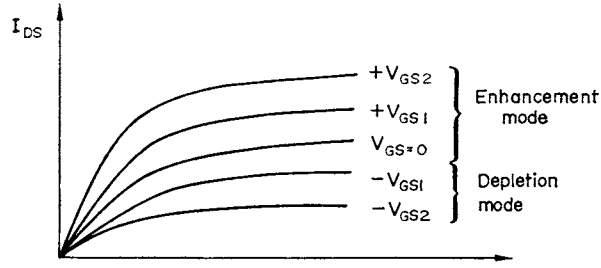


FIG. 7.8. Characteristics of a MOSFET (*p*-type material).

is said to be operating in the enhancement mode, and a negative voltage on the gate will reduce the *n*-channel and the device is said to operate in the depletion mode. This is shown on Fig. 7.8.

7.3. PARAMETERS OF FIELD EFFECT DEVICES

For the majority of small signal applications the junction FET is used in the saturation or "pinched-off" condition. Limited changes of voltage between source and drain have little effect on the drain current. This is true of the IGFET also under normal operating conditions. The drain current is then controlled almost entirely by the gate voltage, and there is little interaction back from the drain circuit to the gate circuit.

The devices can be characterized by parameters in a manner similar to that outlined for the junction transistor in Section 5.1.2, and for thermionic valves in Section 9.3. The absence of interaction between output and input for the voltage-controlled devices leads to only two parameters being required, as compared with four for the bipolar junction transistor, although a third is often included which can be obtained from the two basic parameters.

The three variables associated with field effect transistors are the drain-source voltage V_{DS}, the gate-source voltage V_{GS}, and the drain-source current I_{DS}. A three-dimensional plot would be required to exhibit the relationship between them, but a two-dimensional display is convenient, a family of curves being shown for a series of fixed values for one of the parameters as

in Figs. 7.3 and 7.8 in which fixed values of gate-source voltage are used. In an amplifier it is normally the change of variable that is important, e.g. the change of drain current δI_{DS} when the gate voltage is changed by δV_{GS}. As for the characteristic curves, the parameters show the relationship between change of two variables when the other variable is kept constant. Thus the following two parameters can be defined:

(i) Incremental channel (or drain) resistance r_d or r_{DS}.

This parameter defines the rate of change of drain current with change of drain-source voltage, the gate-source voltage being kept at a fixed value. It is expressed as

$$r_d = \left| \frac{\delta V_{DS}}{\delta I_{DS}} \right|_{V_{GS} = \text{constant}} . \tag{7.1}$$

The reciprocal of the channel resistance is sometimes quoted as an incremental channel conductance, or drain conductance g_d.

The incremental channel resistance is given by the reciprocal of the gradient of the FET characteristics of Figs. 7.3 and 7.8, and typical values range from 100 kΩ to 1 MΩ for a junction FET and from 10 kΩ to 100 kΩ for an IGFET.

(ii) Mutual conductance (or transconductance) g_m or g_{fs}.

This parameter defines the rate of change of drain current with change of gate-source voltage, the drain-source voltage being kept at a fixed value. It is expressed as

$$g_m = \left| \frac{\delta I_{DS}}{\delta V_{GS}} \right|_{V_{DS} = \text{constant}} . \tag{7.2}$$

Typical values of this parameter range from 0·1 to 20 mA/V (or mS) for both junction and insulated gate FETs.

The third parameter that may be quoted is the amplification factor μ. This parameter relates the effectiveness of the gate voltage to the drain voltage in controlling the current through the device. If a change of drain voltage δV_{DS} at constant gate voltage produces a certain small change of drain current, and then a change of gate voltage δV_{GS} at constant drain voltage is required to produce an equal but opposite change of drain current so as to bring it back to its original value, then the amplification factor

can be expressed by

$$\mu = \left|\frac{\delta V_{DS}}{\delta V_{GS}}\right|_{I_{DS} = \text{constant}}. \tag{7.3}$$

The ratio $\delta V_{DS}/\delta V_{GS}$ is a negative one. The negative sign is introduced since the changes in V_{DS} and V_{GS} must be of opposite sign to keep I_{DS} constant. This can be seen, for example, from Fig. 7.3. If V_{DS} is made more positive, V_{GS} must be made more negative to keep I_{DS} at its original value. Normal convention is to express μ as a positive quantity. It follows from the definition that the gate is μ times more effective than the drain in controlling the current through the device. The amplification factor is a ratio of two voltages and there are no units associated with μ, which is dimensionless.

There must be a relationship between the values of the three parameters since there are only three variables V_{DS}, V_{GS} and I_{DS}. This relationship can be seen by writing

$$\mu = \left|\frac{\delta V_{DS}}{\delta V_{GS}}\right| = \left|\frac{\delta V_{DS}}{\delta I_{DS}}\right| \times \left|\frac{\delta I_{DS}}{\delta V_{GS}}\right| = r_d g_m. \tag{7.4}$$

Only two of the parameters are therefore required in order to define the behaviour of the device.

The parameters vary with operating condition since the characteristics are non-linear. They are, however, approximately constant over the working range, and it is these values which are normally tabulated. There is also a considerable spread in parameter values, since devices cannot be manufactured with the precision of thermionic valves, and they are also temperature dependent. Thus calculations cannot be made with the same accuracy as valve predictions.

One of the major advantages of field effect devices is the high input impedance compared with the transistor. The input gate connection is to a reverse biased p–n junction or to an insulated gate, and the input or leakage current is very small. For a junction FET the resistance is typically several hundred megohms, and this will be in parallel with the gate-channel capacitance, usually of about 10 to 100 pF. The input resistance of an insu-

lated gate FET will be even higher, usually in the range 10^9 to 10^{15} ohms, with a parallel capacitance of a few pF. The input gate current is therefore exceedingly small, e.g. 10^{-12} amps.

7.4. AMPLIFICATION AND VOLTAGE GAIN OF THE FIELD EFFECT TRANSISTOR

The FET has been seen to be a device in which a change of drain current may be obtained when a change of voltage occurs between gate and source. The relation between this current and voltage change is given by the mutual conductance, and if there is an input a.c. voltage v_i between source and gate, the a.c. current through the FET will be $g_m v_i$ provided that there is no change in source–drain voltage. If the load resistor R_L is small compared with the incremental channel resistance r_d, where R_L is connected in series with the drain, the output voltage across R_L will be $g_m v_i R_L$. The gain of the stage is thus $g_m R_L$. Because of the high impedance of the FET, R_L is likely to be small compared with r_d, but if this is not the case, the voltage gain magnitude $|A|$ is given by

$$|A| = \frac{g_m r_d R_L}{r_d + R_L} = \mu \frac{R_L}{r_d + R_L}. \tag{7.5}$$

This expression can be obtained by considering the simultaneous change of both drain and gate voltages. A change of drain voltage by δV_{DS} with V_{GS} constant produces a change of drain current δI_{DS}^{1} equal to $\delta V_{DS}/r_d$. A change of gate voltage by δV_{GS} with V_{DS} constant produces a change of drain current δI_{DS}^{11} equal to $g_m \delta V_{GS}$.

The total change of drain current for a small change of both V_{DS} and V_{GS} is then the sum of the two,

i.e. $$\delta I_{DS} = \delta I_{DS}^{1} + \delta I_{DS}^{11}.$$

The term $g_m \delta V_{GS}$ can be written as $(\mu/r_d).\delta V_{GS}$ and therefore

$$\delta I_{DS} = \frac{\delta V_{DS}}{r_d} + \frac{\mu \delta V_{GS}}{r_d} = \frac{1}{r_d}(\delta V_{DS} + \mu \delta V_{GS}). \tag{7.6}$$

This expression is of the form expected since the gate voltage is μ times more effective in controlling the drain current than the drain voltage.

If the FET is used as an amplifier with a resistor in the drain circuit and a constant d.c. supply voltage, the change δV_{DS} operates in opposition to the change δV_{GS}, i.e. the total change of current is less than that for the constant drain-voltage condition, and the gain is less than otherwise expected.

Consider the FET amplifier circuit of Fig. 7.9. The d.c. supply E is assumed to have a very low impedance to a.c. and there is therefore no a.c. voltage across it. A high-value capacitance may be connected across the

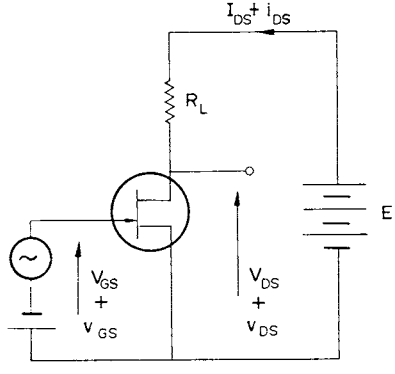


FIG. 7.9. Basic FET amplifier circuit.

supply to ensure that this is so. Each loop of the circuit must satisfy Kirchhoff's laws, i.e. the algebraic sum of the potential differences across all the circuit elements in a closed loop must be equal to zero. There are both a.c. and d.c. voltages, and since the equality to zero must hold at every instant of time, the law must be satisfied by both a.c. and d.c. quantities simultaneously. We can thus write

$$E = V_{DS}+I_{DS}R_L$$

and

$$0 = v_{DS}+i_{DS}R_L. \tag{7.7}$$

The latter equation expresses the fact that the a.c. voltages across the resistor and the FET device are equal in magnitude but opposite in phase, i.e. that there is 180 degrees phase difference between them for sinusoidal

signals. Equation (7.6) can be rewritten putting $\delta V_{DS} = v_{DS}$, $\delta V_{GS} = v_{GS}$, and $\delta I_{DS} = i_{DS}$,

i.e. $$i_{DS} = \frac{1}{r_d}(v_{DS} + \mu v_{GS}). \tag{7.8}$$

Elimination of i_{DS} from (7.7) and (7.8) leads to the expression for voltage gain:

$$\text{Voltage gain} = \frac{v_{DS}}{v_{GS}} = \frac{-\mu R_L}{R_L + r_d}.$$

The negative sign indicates the phase difference of 180 degrees between gate and drain voltages. The voltage gain magnitude $|A|$ is that given in equation (7.5), and is always less than the amplification factor of the device.

Elimination of the voltage v_{DS} from (7.7) and (7.8) gives the expression for the drain current

$$i_{DS} = \frac{\mu v_{GS}}{R_L + r_d}. \tag{7.9}$$

The output circuit is thus equivalent to an a.c. voltage source of magnitude μv_{GS} in series with a total resistance of $R_L + r_d$ of which R_L is external to the device. The device itself is thus equivalent to a voltage source μv_{GS} in series with a resistance equal to the incremental channel resistance r_d, as far as the drain-source circuit is concerned. The voltage equivalent circuit is shown in Fig. 7.10 (a), the values of μ and r_d used being those for the particular operating conditions of the device. The concept of equating the device to a voltage or current source in an appropriate network is a valuable one, as seen in the earlier section on bipolar transistor equivalent circuits. The absence of coupling between output and input for field effect devices leads to a simpler equivalent circuit than for the bipolar transistor. The current source equivalent circuit is given in Fig. 7.10 (b), with a current generator $g_m v_{GS}$ of infinite internal impedance shunted by a resistance equal to the incremental channel resistance r_d. The validity of this form of representation can be checked by adding a resistor R_L across the terminals

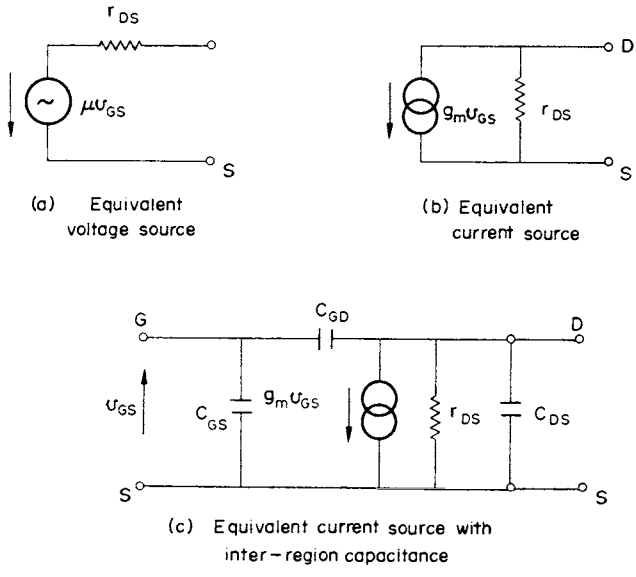


FIG. 7.10. Equivalent circuits for a FET.

DS. The current i_{DS} in R_L is then

$$i_{DS} = \frac{g_m v_{GS} r_d}{R_L + r_d} = \frac{\mu v_{GS}}{R_L + r_d}$$

which corresponds with equation (7.9).

These simple equivalent circuits are first-order approximations that are adequate for many applications. They are not accurate, for example, for very high frequencies when the intercapacitance between gate, source and drain must also be included in the equivalent circuit. A more complete equivalent circuit is given in Fig. 7.10 (c).

It is possible to use the FET as an amplifier in all three main configurations, i.e. as a common source, a common gate or a common drain amplifier. The common source arrangement of Fig. 7.9, in which the input is applied between gate and source and the output is effectively between drain and source, is the most frequently used configuration. It has a high input impedance of several hundred megohms, an output impedance usually of many

kilohms, and a voltage gain of perhaps 10–100. Special purpose FET devices have been used at frequencies in excess of 1 GHz (1000 MHz). The high input impedance makes this amplifier especially valuable where the signal source to be amplified also has a high impedance, e.g. the output from a ceramic record player pick-up, and for oscilloscope input amplifiers or electrometer input stages. The common drain configuration shown in Fig. 7.11 has the output load resistor in the source circuit rather than connected to the drain. The input impedance is even higher than for the common

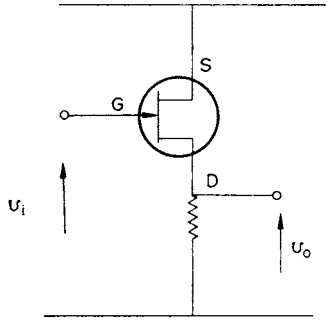


FIG. 7.11. Common drain configuration.

source connection, and it has a low output impedance. The voltage gain is less than one, but the power gain can be large as the output impedance is low compared with the input impedance. The common gate configuration in which the gate is common to both input and output circuits features a low input impedance and a high output impedance. It finds application as an impedance transformer in some high-frequency circuits, but its use is limited.

7.5. THE FET AS SWITCH AND VARIABLE RESISTOR

The source-drain resistance of the FET can be changed from a low to high value by change of source-gate voltage, as can be seen from the FET characteristic, and the process is reversible. It is thus a device which can approximate to an open-circuit or a short-circuit depending upon the gate

voltage, and can be used like the bipolar transistor as a switch for digital applications, as discussed in Chapter 6, as well as an amplifier.

For intermediate source-gate voltages the FET behaves as a voltage-variable resistor between source and drain, with the resistance value controlled by the source-gate voltage, and it can be used in integrated circuits in this way, particularly in its MOSFET form. The MOSFET construction also allows capacitors to be formed in a similar manner, using the silicon oxide layer as a dielectric between the metal contact layer and the semiconductor substrate as capacitor electrodes. Thus the MOSFET arrangement is particularly valuable for the integrated circuit approach.

As for the bipolar transistor, the maximum voltage for a FET is limited to about 20 volts. If the voltage across a reverse-biased p–n junction becomes too large, impact ionization and avalanche processes somewhat akin to those in a gas-discharge take place, and breakdown occurs.

8. *Manufacture of p-n Junctions, Transistors, and Integrated Circuits*

8.1. ZONE REFINING AND CRYSTAL GROWING

The first requirement for transistor manufacture is very pure germanium or more commonly silicon. In order to obtain transistor action the pure basic semiconductor material must be doped in a controlled manner with impurity atoms to the extent of about one impurity atom per million parent atoms. In some cases even less doping than this is required. It is usual therefore to begin with basic material that contains an impurity atom concentration of less than one part in 10^{10}.

Impurities are removed from the unrefined ingots of silicon or germanium by a technique known as zone refining. The method is shown in Fig. 8.1. The material to be refined is placed in a quartz or graphite boat which is drawn slowly through a quartz tube. Around this tube several radio-frequency heater coils are wound and the eddy currents induced in the ingot by these coils cause sufficient heat to melt the ingot locally. The whole of the quartz tube is filled with an inert gas in order that the ingot is not oxidized at all. In Fig. 8.1 only the material inside a heating coil is molten. The boat is supposed to be moving from left to right and thus the molten region moves along the ingot from right to left. Suppose the impurity concentration in a region of the ingot that is just about to be melted is B atoms/unit volume; then when it becomes molten the impurity concentration is still B. As this molten region passes away from the heater coil and cools, the concentration of impurity atoms is found to be considerably reduced in the newly solidified material, say to A atoms/unit volume. The excess impurity atoms have moved into the region that is now molten. The ratio A/B is called the segregation constant. It may range from around 0·1 to 0·001 depending on

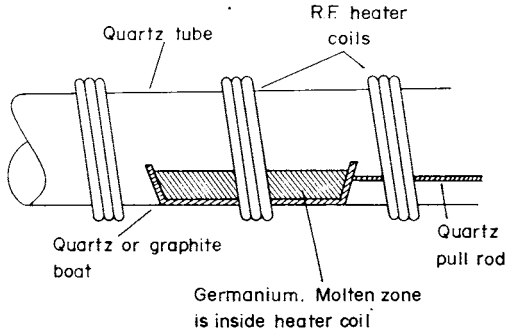


FIG. 8.1. Zone refining. The boat containing the material to be refined is pulled slowly past the heating coils.

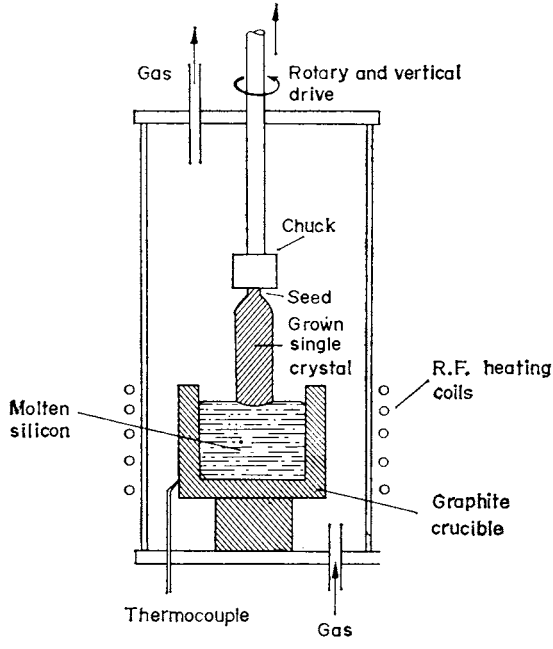


FIG. 8.2. A crystal-growing furnace.

the type of impurity. Thus, as the bar is pulled through a heater coil, the impurities concentrate at the left-hand end. By drawing the ingot through several heater coils the impurity level is successively reduced. The left-hand end of the ingot which now contains most of the impurity atoms is subsequently cut off and discarded.

The purified metal must now be grown into a single crystal. This technique is shown in Fig. 8.2. The crystal-growing furnace allows the base material, silicon for example, to be melted by a radio-frequency heating coil in an inert atmosphere again, to prevent oxidation. A small seed crystal of silicon held in a rotating chuck is dipped into the liquid silicon and then slowly withdrawn. If the temperature of the melt and the pulling rate are correct a single crystal is continuously drawn from the melt. The chuck is rotated to agitate the melt and also to ensure the symmetrical growth of the crystal.

If a small controlled amount of *n*- or *p*-type impurity is added to the melt, then either *n*- or *p*-type single crystals may be grown in this manner.

8.2. EARLY METHODS FOR THE PRODUCTION OF JUNCTION TRANSISTORS

The earliest techniques for producing transistors were the *grown-junction* and the *alloy-junction* methods. The former method will be described as it would be applied to grow a *n–p–n* transistor. The junction is grown directly as the crystal is pulled from the melt in the crystal-growing furnace. First of all an *n*-type impurity is added to the melt and the withdrawn crystal is therefore *n*-type. This is the transistor emitter region. After a length of crystal has been grown, *p*-type impurity is added to the melt in sufficient quantity to outweigh the effect of the *n*-type. The crystal now becomes *p*-type and a small length is grown to make the base region of the transistor. Then more *n*-type impurity is added to the melt to cause the crystal to revert to *n*-type. This region becomes the collector. After the single crystal has been grown it is cut lengthwise into several *n–p–n* junctions measuring possibly $1 \times 1 \times 3$ mm long. Each junction is then etched to remove surface contamination; the lead wires are fastened on, and the whole assembly is hermetically sealed into a glass or metal container. Without sealing, moisture ingression would cause surface oxidation of the transistor. Surface roughness

and oxidation both provide mechanisms for increased minority carrier recombination in the base region. This will obviously degrade the transistor performance.

An example of the construction of a *p–n–p* alloy-junction transistor is shown in Fig. 8.3. The base material is a thin wafer of *n*-type germanium say, perhaps 3 mm square by 0·3 mm thick. On either side of the wafer are

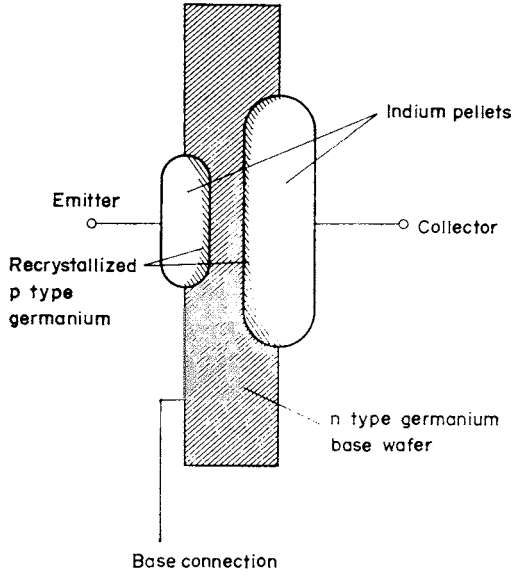


FIG. 8.3. Construction of a fused junction *p–n–p* transistor.

placed pellets of a *p*-type impurity metal, say indium. The collector pellet is about three times greater in area than the emitter pellet. The assembly is heated in an inert atmosphere above the melting point of indium but below that of germanium. The indium diffuses into the germanium to make it *p*-type. The depth of diffusion and thus the thickness of the *n*-type base layer is controlled by time and temperature. Lead wires are connected to each region and the transistor is encapsulated.

The collector surface area is made greater than the emitter area in order

that as large a proportion as possible of the holes leaving the emitter reach the collector. Since holes move across the base region by diffusion only there is a much greater tendency for holes to be lost by transverse diffusion and surface recombination if the emitter and collector have the same cross-sectional area than if the collector is bigger than the emitter.

8.3. THE PLANAR TECHNOLOGY

The previous two methods of transistor and diode manufacture were very important in the early years following the invention of the transistor and the alloy junction technique is still used for the mass production of certain germanium devices. However, as we have noted on several occasions earlier in this text, an important factor in determining the speed of response of a transistor is the base width. In the alloy process, the location of the junctions are determined by the depth of penetration of the recrystallized regions into the semiconductor. This depth is hard to control accurately and therefore alloy junction transistors with very narrow base regions are difficult to make.

A technique which gives much superior control is that of the diffused junction. Diffused junctions are produced by exposing the surface of the semiconductor to a high concentration of opposite type impurities, carried normally in the gaseous phase. By doing this at an appropriate temperature the impurities penetrate the semiconductor by solid state diffusion. The rate of diffusion, and thus the depth of penetration of the diffusion, can be controlled very precisely.

A further significant discovery was made when it was realized that a thin layer of silicon dioxide, which can be grown readily on the surface of a single crystal silicon slice by placing it in an oxidizing atmosphere, can readily mask and protect against diffusion of the important acceptor and donor impurities. By opening up areas in the SiO_2 protective layer and diffusing impurities through these exposed regions, very close control on the lateral dimensions of the device can also be obtained.

The situation is shown schematically in Fig. 8.4. In Fig. 8.4 (a) an *n*-type silicon layer is first exposed to an oxidizing ambient of oxygen or steam at approximately 1000°C. A SiO_2 layer perhaps 500 Å thick is grown in this

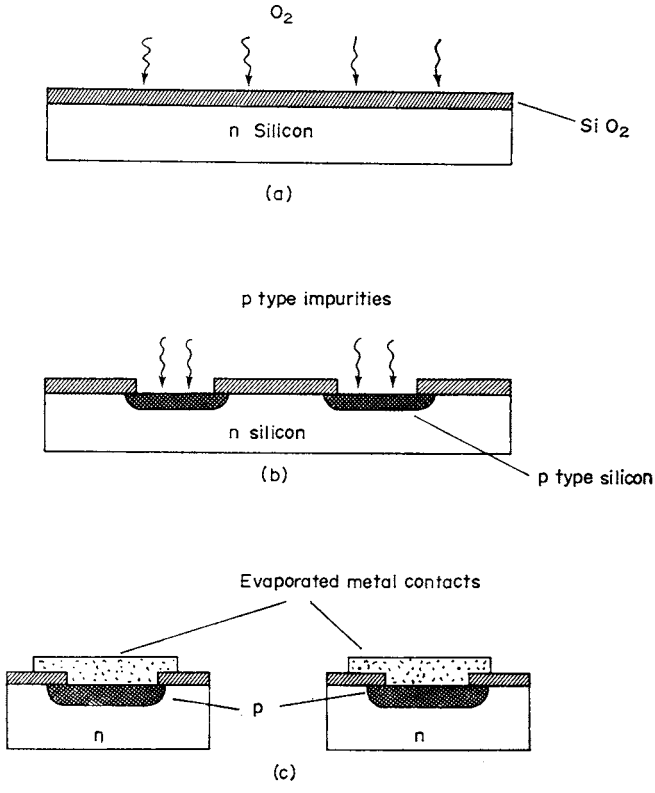


FIG. 8.4. Stages in producing a diffused junction: (a) thermal oxidation of silicon, (b) *p*-type diffusion, (c) top contact evaporated, dice sliced into discrete devices.

manner. Figure 8.4 (b) shows the next stage. Holes are etched in the oxide as described in Section 8.6 to expose the *n*-type material. The slice is then placed in a reactor where it is subjected to an atmosphere of possibly boron trioxide (B_2O_3) if a *p*-type diffusion is required. The temperature is again around 1000°C and the reaction

$$2B_2O_3 + 3Si \rightleftharpoons 4B + 3SiO_2$$

takes place, leaving a very thin surface layer of boron atoms a few tenths of a micron deep. The boron impurities are then diffused deeper during the so-called "drive-in diffusion" stage. The ultimate depth of the diffusion is controlled by time and temperature. Figure 8.4 (c) shows the final stage where a metallic top contact to the *p*-region has been evaporated and the slice scribed and broken-up to make several discrete *p–n* junctions.

This is an example of the planar technology which started in 1960 to revolutionize the fabrication of solid state devices. The very close control that can be held on doping profiles and dimensions by this method and the high packing density of devices it allows on a single slice led rapidly to the development of integrated circuits. The SiO_2 protective or passivating layer which, as Fig. 8.4 (c) shows, completely protects the *p–n* junction interface, also gives rise to devices with much better characteristics (less leakage current for example) than the earlier methods of manufacture.

8.4. EPITAXIAL LAYER GROWTH

The growth of a SiO_2 layer as described in Section 8.3 is one example of a vapour-phase growth technique. The most important vapour-phase technique for device fabrication is the growth of a single crystal film upon a single crystal substrate, normally of the same material but with a different doping. This method of growth is called epitaxy, and the film so grown is called an epitaxial layer (epitaxial—from the Greek for "arranged upon"). It is possible in this manner to grow a thin epitaxial layer of *n*-type silicon on a *p*-type substrate of silicon, or vice-versa, *p*-type on an *n*-layer.

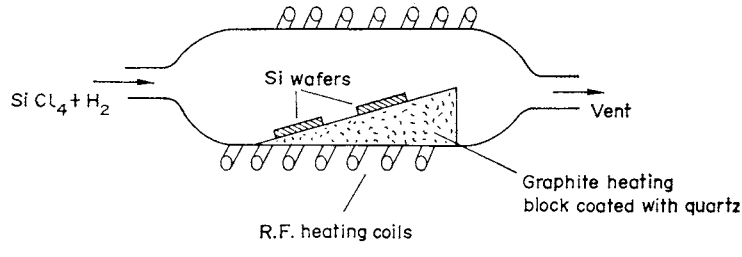


FIG. 8.5. Epitaxial reactor for silicon.

The most widely used method of growing silicon epitaxial layers is the vapour-phase reduction of silicon tetrachloride. A typical quartz reactor system is shown in Fig. 8.5. Here $SiCl_4$ and hydrogen gases are passed over silicon single crystal wafers. Dopant gases are also added, normally diborane (B_2H_6) for *p*-type doping, and phosphine (PH_3) for *n*-type. The silicon wafers are heated by r.f. induction heating to approximately 1000°C and the reaction

$$SICl_4 + 2H_2 \rightleftharpoons Si(solid) + 4HCl$$

takes place. The silicon grows as a doped single crystal on the host crystal (or substrate) at a rate of about 1 micron per minute. The substrate may well be produced in the first instance by the crystal-pulling technique described in Section 8.1. An example of the use of an epitaxial layer in device manufacture is described in Section 8.6.

8.5. THE INTEGRATED CIRCUIT CONCEPT

As we have just noted the planar technology provides a method for allowing whole circuits, or large parts of circuits, to be constructed on a single small piece, or chip, of semiconductor material a few millimetres in overall dimensions.

For specialized purposes, and when a very large number of identical circuits is required, the whole circuit will be included on the single chip if possible and economic. When a complete circuit is included on a single chip it is called a monolithic integrated circuit. The cost of setting up to produce a sophisticated chip is very high, however, and only becomes economic if sufficient numbers are required or if the saving in space and weight is essential. For other types of application greater flexibility is achieved by leaving some determining elements from the circuit, to be connected externally. A single chip can then serve a variety of purposes by connecting a small number of external circuit elements appropriate to the particular requirement, i.e. to meet the overall circuit function needed.

Amongst the advantages offered by integrated circuits are a great reduction in size and weight, and a reduction in cost in some cases. The reduction

in dimensions may be as great as a factor of 100 or more, enabling the most sophisticated electronic equipment, e.g. computers, to be manufactured with a reasonable overall size and a weight reduction of perhaps 50 to 1.

Integrated circuits can be produced using either bipolar or unipolar transistors. Field effect transistors have advantages over bipolar transistors in many cases however. They are relatively high-impedance devices, with a corresponding reduction in current and power dissipation, and yet have high power gain. The reduction in power dissipation is particularly important where a complex circuit is to be concentrated into a small space. The problem of extracting the heat generated in the circuit may then be a difficult one. The form of construction of the FET, and particularly the MOSFET, also lends itself well to integrated circuit fabrication, and enables resistors and capacitors to be included readily in the integrated circuit.

Most electronic circuits are composed of active devices, e.g. transistors and diodes, together with resistors (for bias, collector load, impedance transformation, etc.) and capacitors (e.g. for coupling a.c. signals while blocking d.c. supplies). Each of these elements can be produced in a form suitable for integrated circuit inclusion within limitations, e.g. capacitance values must not be too large. Some elements are difficult to produce in a suitable form, e.g. inductive elements, or large capacitors. Usually some alternative circuit form can be devised that dispenses with the requirement. Otherwise they must be included as an external lumped element.

8.6. INTEGRATED CIRCUIT REALIZATION

Some indication of the way in which whole circuits may be included on a single chip will be given at this stage.

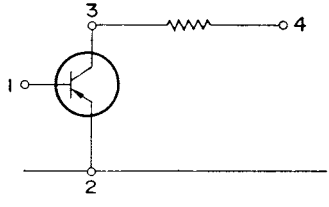
There are basically four types of material that may be required in the integrated circuit, *p*- and *n*-type semiconductor, insulating layers and conducting layers. The starting point is a piece of single-crystal semiconductor. It is not usual that intrinsic material is required, so the normal starting crystal is a doped substrate, e.g. *p*-type silicon, with a plane exposed surface. Selected areas of the surface can be operated upon in the following manner:

(i) The layer near the surface can be made a *p*- or *n*-type semiconductor by allowing impurities to diffuse in through a mask defining the areas to be changed, as described in Section 8.3. The depth of the layer so affected can be controlled, depending upon the time, temperature, etc., of the operation.

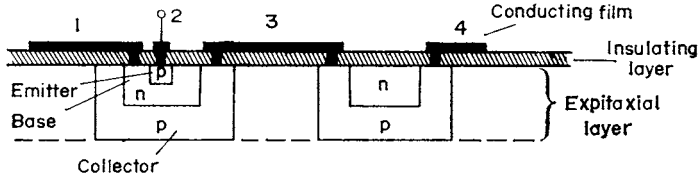
(ii) An insulating layer, exceedingly thin (e.g. 10^{-6} m), of silicon oxide can be formed by heating the substrate to about 1000°C in an atmosphere of oxygen.

(iii) Interconnection can be made by vacuum deposition of aluminium in selected areas using a masking technique.

Selection of areas for treatment is usually by a photolithographic technique. A film of photosensitive emulsion is coated on to the surface, and



(a) Circuit form



(b) Monolithic form of integrated arrangement

FIG. 8.6. Simple integrated circuit arrangement.

exposed to light (usually U.V.) through an appropriate negative. Treatment with a suitable chemical dissolves the unexposed areas, leaving the required surface pattern of emulsion which protects the underlying area from the process which follows.

In many cases an oxide layer is first formed with the required pattern using the emulsion mask, and this oxide layer is then used as a more robust mask pattern for the process to follow.

The technique can be illustrated by considering the simple circuit of Fig. 8.6, consisting of a transistor and collector resistor. The starting point in this case is a substrate of *n*-type material. An epitaxial layer of *p*-type material is first grown on it by the method described in Section 8.4 and the surface is oxidized. Patterns are etched in the oxide layer by first laying down the appropriate mask by the photolithographic technique, and then etching away the unprotected areas with hydrofluoric acid to expose the silicon surface. The first pattern exposes the areas between the proposed devices to *n*-type impurities which are allowed to diffuse down to the substrate, thus forming isolated islands of *p*-type material. Whilst these islands are formed in conducting material they are electrically isolated since there are two *p–n* junctions back-to-back between each. The surface is then reoxidized and a new pattern laid down to allow *n*-type impurities to diffuse into the centre of the *p*-type islands, and the process repeated to diffuse *p*-type impurities into the centre of the transistor area to form the emitter. A final oxidized layer has holes etched into it to allow electrical contact with the various regions. A thin film of aluminium is then deposited on the device and etched away using a photoresist mask to leave connecting strips as required. Clearly the technique can be extended to give circuits of great complexity. Capacitors can be included by using the silicon oxide as dielectric between two conducting film areas, or by using the capacitance between a reverse biased *p–n* junction.

In production a large number of identical circuits can be manufactured on a single slice, or slices, of semiconductor, and cut up into separate circuits after completion. Connections to the integrated circuit are then made by bonding very fine gold wire to the appropriate part of the circuit and connecting it to a pin on the holder—or header—of the circuit.

Only a glimpse of the potentialities and concepts of integrated circuits

has been given, and the reader is directed to one of the modern specialist books on the subject for futher details. The additional advantage that the reliability of the overall integrated circuit is similar to that of an individual component leads to the possible construction of systems with a complexity that cannot be envisaged for the discrete component if a good reliability is required.

9. *Thermionic Valves*

MANY thermionic valves are highly complex devices, containing a large number of electrodes, but the majority in common use belong to one of the types classified as diodes, triodes, tetrodes or pentodes depending upon whether they have two, three, four or five electrodes respectively. Discussion will be limited mainly to these four basic types of valve. Diodes and triodes are also available with a low pressure gas—usually an inert rare gas such as neon or argon—introduced into the vacuum envelope of the valve. The small amount of gas introduced brings about a significant change in the characteristic of the valve, and its effect will also be considered in this chapter. In each case a thermionic cathode is used to produce free electrons in the region between cathode and anode, and the subsequent motion of these electrons determines the behaviour of the valve in an electric circuit.

9.1. THE HIGH VACUUM DIODE

The geometry of the diode is very simple, consisting of a thermionic cathode surrounded by an anode, these often being arranged as a pair of concentric cylinders. The electrodes are mounted in an evacuated glass envelope. The diode is symbolized in Fig. 9.1 (a) and (b), where valves with either a directly or indirectly heated cathode are shown. Whilst the number of electrons that can be emitted by the cathode depends upon the cathode temperature, this is not necessarily the number that is collected by the anode. Indeed if the anode is made negative with respect to the cathode there is no current through the valve. It is this property of allowing current when the anode is at a positive potential with respect to the cathode, and no current when the anode is at a negative potential, that is responsible for most of the applications of the thermionic diode.

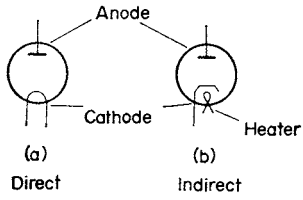


FIG. 9.1. Schematic diode representation.

The relationship between the voltage applied to the valve and the current through it is shown in Fig. 9.2. Three distinct regions can be distinguished, region *A* where the anode is negative and there is virtually no current, region *B* where the current is increasing as the anode potential is raised, and region *C* where the current is almost independent of the voltage provided that the cathode temperature remains constant.

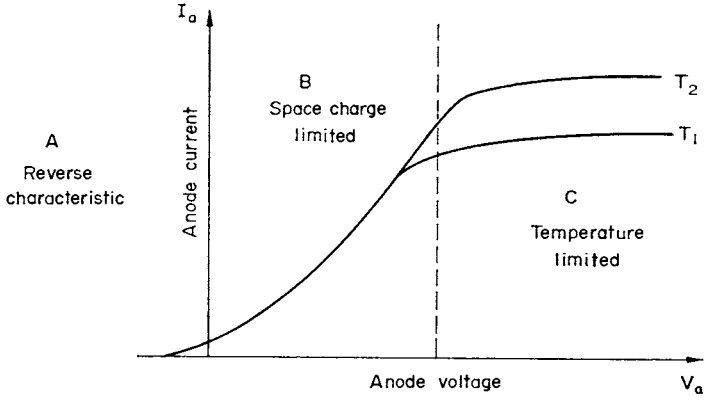


FIG. 9.2. Diode characteristic.

In region *A*, the negatively charged electrons emitted by the cathode are repelled by the negative anode potential back to the cathode. No electrons can therefore cross from cathode to anode and no current flows through the valve. When the anode and cathode are at the same potential, i.e. there is zero voltage across the valve, no applied electric field exists in the valve, and electrons emitted from the cathode tend to drift relatively slowly

across the valve. This drift occurs since electrons are emitted with a small but finite velocity. A cloud of electrons, known as a space charge cloud, is then built up in the interelectrode region. This negatively charged cloud exerts a repelling force on any further electrons emitted by the cathode. A dynamic equilibrium is set up whereby the number of electrons that leave the cathode and join the electron cloud is equal to the number collected by the anode from the cloud. A small current flows through the valve. If now the anode is made positive in potential, the equilibrium is upset since electrons are attracted from the cloud to the anode at a greater rate and therefore more are allowed to leave the cathode and join the electron cloud. Under these conditions the diode is said to be space charge limited and it can be shown both theoretically and experimentally that the current through the valve is proportional to the anode voltage raised to the power of three halves. The relationship is known as the Child–Langmuir law and can be expressed as

$$I_a = k(V_a + V_e)^{3/2} \tag{9.1}$$

where k is a constant depending upon the geometry and dimensions of the valve, I_a and V_a are the current through and the voltage across the valve, and V_e is a voltage introduced to allow for the fact that a small current flows even when the voltage across the valve is zero. The value of this voltage is related to the energy with which electrons are emitted from the surface of the cathode. When the anode voltage becomes sufficiently great, electrons are attracted from the electron cloud at a rate equal to the maximum possible emission from the cathode at its particular temperature and the electron cloud disappears. The current through the valve is then limited to the cathode emission as given by Richardson's equation. This current, called the saturation current, is of course dependent upon the temperature of the cathode as can be seen in Fig. 9.2. Even in region C there is a small increase of current when V_a is increased (Schottky effect) since the electric field at the cathode due to the anode voltage has the effect of reducing the apparent work function of the cathode. This effect can normally be neglected unless very high anode voltages are used.

9.1.1. *Space-charge Limited Conditions*

The variation of electric potential between anode and cathode for a diode with plane-parallel electrodes is shown in Fig. 9.3. Three distributions are given, distribution *A* corresponding to the case of zero or negligible space-charge, *B* to the space-charge limited condition assuming that the electrons are emitted with a velocity that is close to zero, and *C* to the space-charge

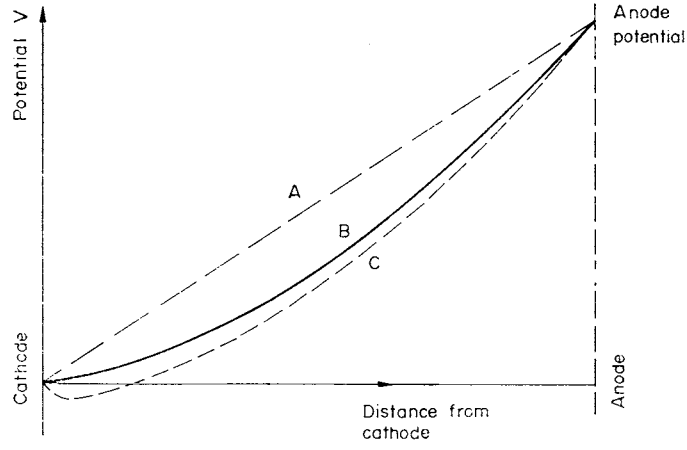


FIG. 9.3. Potential distribution for planar diode.

limited condition with a significant electron emission velocity. The linear distribution is obtained when there is little or no space-charge cloud; for example if the cathode temperature is so low that relatively few electrons can be emitted, or if the anode voltage is high and the diode is temperature limited. When an electron space-charge cloud is formed, the accumulation of negative electric charge reduces the potential close to the cathode until the electric field (given by the slope of the potential distribution curve) at the cathode surface approaches zero. There is then little force at the cathode surface to accelerate electrons away from the surface. It is this reduction of electric field at the cathode surface that is responsible for the self-regulating character of the space-charge limited condition. When electrons are

emitted with a significant velocity the space-charge cloud builds up to give a higher density, and the potential a short distance away from the cathode surface becomes negative with respect to the cathode as in curve *C* of Fig. 9.3. The maximum negative value of the potential is almost equal to the voltage V_e of the Child–Langmuir equation (9.1). A repelling electric field then exists at the cathode and emitted electrons experience a decelerating force as they leave the cathode surface. The electron velocity is almost zero at the position of maximum negative potential. For a given voltage between anode and cathode the velocity with which electrons reach the anode is independent of whether or not a space-charge cloud is present, as required by the energy equation (2.8) for the electrons. The effect of the space-charge cloud is to reduce the electron velocity in the vicinity of the cathode below that obtained for the space-charge-free condition, but to increase the acceleration closer to the anode so that the final velocity is the same for both cases.

The electric current (rate of transfer of electric charge) equivalent to a single electron in motion is given by ev, where v is the electron velocity (Section 2.4.2), and therefore the current density at a point of electron density n is given by

$$J = nev. \tag{9.2}$$

For a diode with plane-parallel electrodes the current density must be the same at all points between the electrodes. It follows from equation (9.2) that since the product nv is the same throughout the inter-electrode region, the electron density is proportional to $1/v$ and will be greatest where the velocity is least, i.e. at the position of minimum potential in the region of the cathode. As the electron velocity increases towards the anode the electron density will be reduced to a relatively low value.

9.1.2. *Diode Operation*

In region *B* of Fig. 9.2 the diode is space-charge limited and the valve current is dependent upon the anode voltage but not upon the cathode temperature, provided that this temperature is sufficiently high to ensure that saturation current is not being drawn from the cathode. In region *C* the

current is dependent upon the cathode temperature but is approximately independent of the anode voltage. Thermionic valves are normally operated under space-charge limited conditions. The current through the valve is then insensitive to changes of cathode temperature.

Most applications of the diode, such as the rectification of alternating voltages or the detection of radio signals, make use of it as a switch, the valve acting as an open-circuit or a closed-circuit depending upon whether the anode is at a negative or a positive polarity with respect to the cathode. Whilst the open-circuit condition can be closely achieved in practice, the effective resistance of the diode is not zero in the conducting condition. For example, if a diode passes a current of 100 mA when the anode voltage is 100 V, the resistance of the valve is 1000 Ω. A comparison with other diode types is given in Chapter 10.

9.2. THE HIGH-VACUUM TRIODE VALVE

The triode is similar in construction to a diode with the addition of a third electrode placed between the anode and cathode. This third electrode is an open wire structure through which electrons can pass, and is called the control grid. It often takes the form of a fine wire helix as shown in Fig. 9.4, and the corresponding schematic representation is given in Fig. 9.5. Electrons can pass through the control grid unimpeded, but the electric field at the cathode surface can be changed by varying the electric potential of the control grid. There are two important features of such a grid that enable the triode valve to be used as a voltage amplifier. If the grid is placed physically very close to the cathode small changes of the voltage between the cathode and grid will produce a large change of electric field in the cathode-grid region and therefore a large change of current through the valve. Moreover, if the grid has a potential which is slightly negative relative to the cathode no electrons can travel from the cathode to the grid, and hence no conduction current flows in the grid circuit. If no current is taken from the grid supply no power is required to change the grid voltage and therefore to change the anode current. This is why very small signals from a low power source, e.g. from a radio aerial, can produce a significant effect in the first stage of a radio receiver. In practice a very small current does

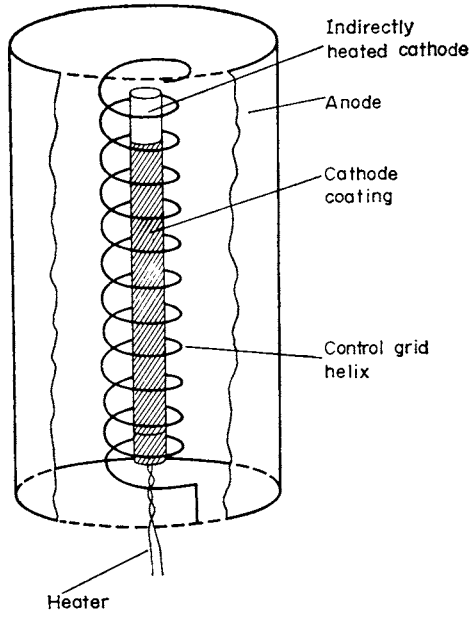


FIG. 9.4. Cut-away view of triode electrode system.

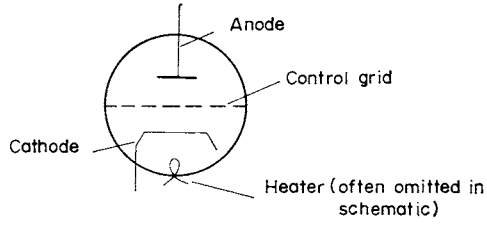


FIG. 9.5. Triode schematic diagram.

flow in the grid circuit, and the ratio of the grid voltage to the grid current is expressed as the input impedance of the valve. It is typically many megohms under normal operating conditions, and thus resembles the field effect transistor rather than the bipolar transistor.

The triode is normally operated under space-charge limited conditions, an electron cloud forming close to the cathode surface. As in the case of the diode, the current leaving the cloud depends upon the electric field

near the cathode. In the triode the electric field is the sum of two components, produced by the anode and control grid potentials. The former aids and the latter, because of its negative potential, suppresses emission from the cloud. The situation is made somewhat more complex due to the close proximity of the grid wires to the cathode, the effect of the grid being greatest at those parts of the cathode close to the grid wires. There is thus a non-uniform flow of current from the cathode surface. As the grid potential is made more negative there is a reduction in the area of the space-charge cloud from which electrons are drawn into the region between anode and control grid. When the negative potential of the control grid is sufficiently great, no electrons are able to leave any part of the cloud, and the current to the anode is reduced to zero. The valve is then said to be "cut-off".

9.2.1. *Triode Characteristics and Parameters*

There are three variables associated with a triode valve, the anode voltage V_a, the grid voltage V_g, and the anode current I_a. All voltages are expressed relative to the cathode. The direction of current-flow corresponds by convention to the direction of transfer of positive charge, and the current therefore flows in the opposite direction to the electron motion, i.e. from anode to cathode. A three-dimensional plot is required to exhibit the relationship between three variables, but a family of curves is obtained for a series of different values of one parameter kept constant along each individual curve as for the FET. Figure 9.6 shows the variation of the anode current as the anode voltage is changed for several discrete values of grid voltage. These are called the anode characteristics of the valve. When the grid is at a slightly positive potential, approximately equal to that which would exist there if the grid were removed, we would expect the characteristic to be similar to that of the diode in the space-charge limited condition. As the grid is made negative the current is reduced for a given anode potential but the overall shape of the characteristic is little changed, the characteristic curve being displaced along the anode voltage axis. This is as expected, since an increase of repelling field at the cathode due to a larger negative voltage on the grid can be overcome by an increase in the positive potential of the anode. A set of characteristic curves is obtained which are approximately parallel and equi-spaced for equal increments of grid voltage, over much of

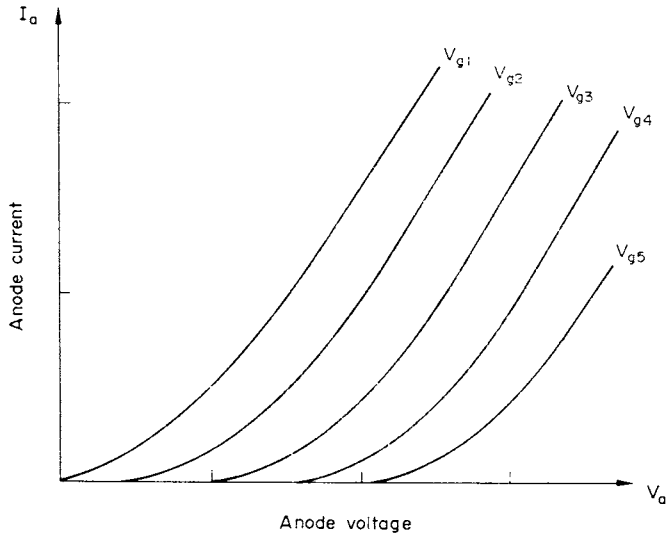


FIG. 9.6. Triode anode characteristics.

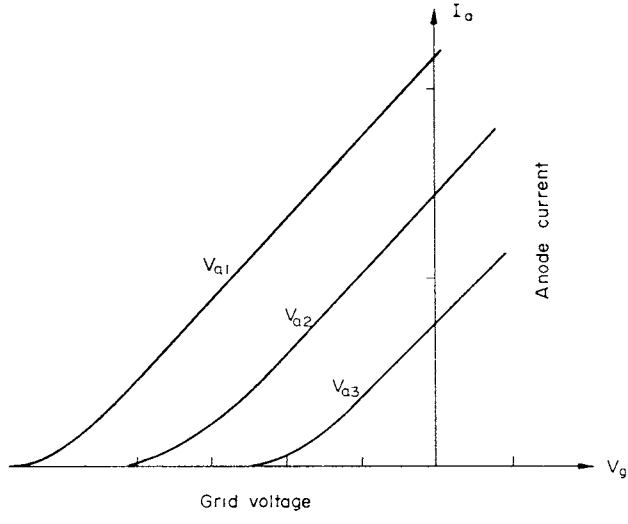


FIG. 9.7. Triode mutual characteristics.

their range. From this characteristic the change of anode current for a given change of anode voltage at constant grid voltage can be obtained by following a single curve. The change of anode current for a given change of grid voltage at constant anode voltage is obtained by erecting a vertical line at the appropriate value of V_a. Precisely the same information can be replotted in a different form as in Fig. 9.7. Here it is the anode voltage which is kept constant during each particular characteristic curve, and the graph is known as the mutual characteristic. The cut-off condition can be clearly seen, the cut-off grid voltage becoming more negative as the anode voltage is increased.

When used as a small signal voltage amplifier, change of voltage and current is important and the shapes of the characteristics define the amplifier behaviour. The parameters of the triode are similar to those of the FET, and can be defined as follows:

(i) Anode slope resistance r_a

This parameter defines the rate of change of anode current with anode voltage, the grid voltage being kept at a constant value, i.e.

$$r_a = \left| \frac{\delta V_a}{\delta I_a} \right|_{V_g = \text{constant}} . \tag{9.3}$$

The parameter magnitude is equal to the reciprocal of the gradient of the anode characteristic, and typical values for a triode valve range from 1 kΩ to 50 kΩ.

(ii) Mutual conductance (or transconductance) g_m

This parameter defines the rate of change of anode current with grid voltage, the anode voltage being kept at a constant value, i.e.

$$g_m = \left| \frac{\delta I_a}{\delta V_g} \right|_{V_a = \text{constant}} . \tag{9.4}$$

Typical values for the mutual conductance range from 0·1 to 10 mA/V (or m Siemens).

(iii) Amplification factor μ

As for the FET, this parameter can be obtained from the other two, and expresses the effectiveness of the grid (or gate) to that of

the anode (or drain) in controlling the current through the device, i.e.

$$\mu = \left|\frac{\delta V_a}{\delta V_g}\right|_{I_a = \text{constant}}, \tag{9.5}$$

and the relationships between the parameters is given by

$$\mu = \left|\frac{\delta V_a}{\delta V_g}\right| = \left|\frac{\delta V_a}{\delta I_a}\right| \left|\frac{\delta I_a}{\delta V_g}\right| = r_a g_m. \tag{9.6}$$

Typical values of μ for a triode valve range from 2 to 200.

9.3. THE TRIODE AS A VOLTAGE AMPLIFIER

When the triode is used, in conjunction with an appropriate circuit, to give magnification of a small time-varying signal, e.g. a small a.c. voltage, the d.c. grid voltage—known as the grid bias voltage—and the d.c. anode voltage

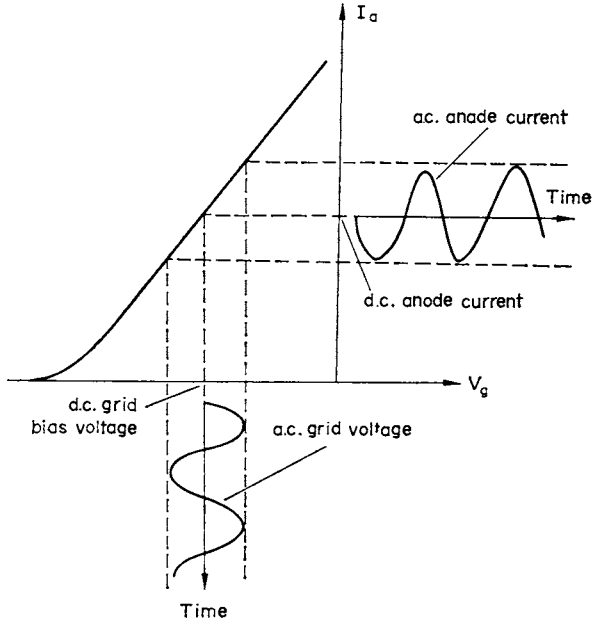


FIG. 9.8. Triode anode current with a.c. and d.c. grid voltage components.

are chosen so that the valve is operating on the linear part of the characteristic. When a time varying voltage is then also applied to the grid, a time varying component of anode current of the same form as the grid voltage will pass through the valve. This is illustrated in Fig. 9.8. An a.c. voltage will be produced across a resistor connected in series with the valve as in Fig. 9.9. If the resultant a.c. voltage is greater than that initially applied to the grid, voltage amplification has been achieved. The voltage gain, given by

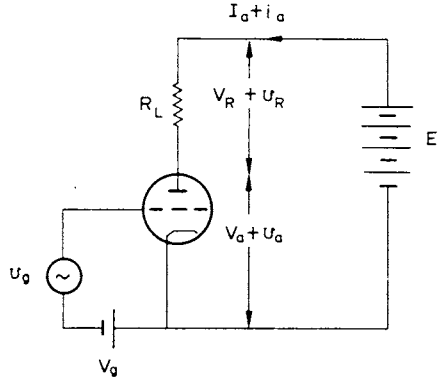


FIG. 9.9. Simple triode voltage amplifier.

the ratio of output to input a.c. voltages, cannot be simply calculated however since the anode voltage of the valve will change when the grid voltage is changed, due to the variation of voltage across the resistor. The actual anode current through a valve connected in series with a fixed resistor and a constant voltage d.c. supply, due to a change of grid voltage, is given by the dynamic characteristic.

9.3.1. *Dynamic Characteristics*

The dynamic characteristic of a valve and its circuit can be obtained by a graphical method as follows. There are two relationships between the anode voltage and anode current which must be satisfied, one being the anode characteristic of the valve itself and the other relating to the circuit in which the valve is used. This second relationship is obtained by applying

the law that the algebraic sum of the potential difference across each element in a closed circuit is equal to zero (Kirchhoff's law). Thus for the anode circuit of Fig. 9.9, considering that only d.c. components (represented by capitals) are present, $E = V_a + V_R = V_a + I_a R_L$. In this equation E and R are constants for a given circuit condition. The graph of this equation, which is a straight line cutting the I_a and V_a axes at the points E/R_L and E respectively, is known as the load line for the circuit, and is shown together with the anode characteristic in Fig. 9.10. The intersection of these graphs gives the values which satisfy both relationships and therefore the actual values of I_a and V_a, for different grid voltages. It can be seen that the change

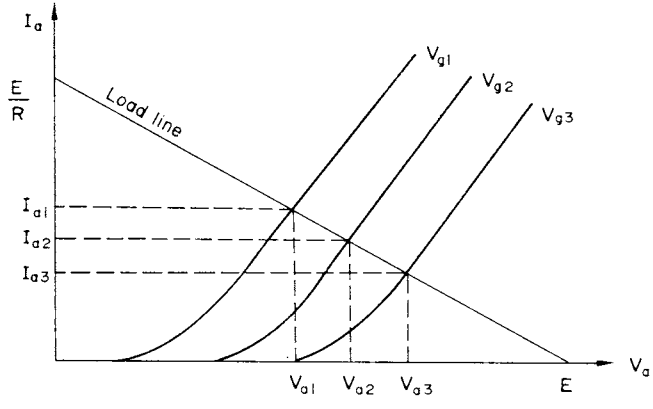


FIG. 9.10. Internal and external characteristics for resistive load amplifier.

of I_a for a given change of V_g is now less than for the case of operation at constant anode voltage. The relationship between I_a and V_g under these conditions is shown in Fig. 9.11 and compared with the static mutual characteristic of the valve. It is seen that the effective mutual conductance is lower than that normally quoted for the valve. The voltage gain (or amplification ratio as it is sometimes known) is seen from Fig. 9.10 to be

$$(V_{a2} - V_{a1})/(V_{g2} - V_{g1}).$$

Alternatively, the approach used in Section 7.4 can be taken, in which the effect of both grid and anode voltage on the anode current can be taken

into account. Following the same argument it can be shown that

$$\delta I_a = \frac{1}{r_a}(\delta V_a + \mu \delta V_g).$$

As for the FET case, when a resistor is included in series with the anode to give gain, the resultant δV_a change is in the sense to reduce the change of δI_a produced by δV_g, i.e. the effective gain is reduced. This corresponds with the dynamic characteristic of Fig. 9.11.

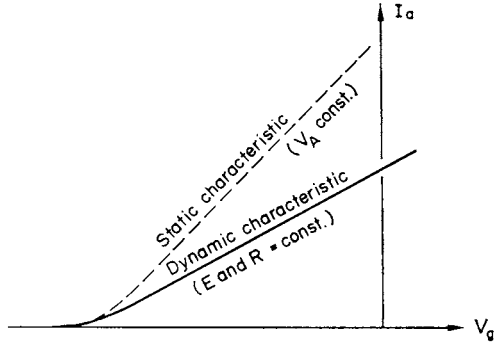
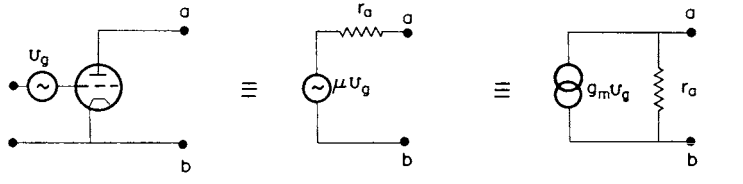


FIG. 9.11. Comparison of static and dynamic characteristics.

The argument of Section 7.4 also leads to the expression for the voltage gain of the amplifier

$$A = \frac{v_a}{v_g} = \frac{-\mu R_L}{R_L + r_a}$$

and the equivalent circuits of Fig. 9.12.



(a) Equivalent voltage source (b) Equivalent current source

FIG. 9.12. Equivalent anode circuits for a triode (considering a.c. components only).

9.3.2. *Triode Limitations*

The triode valve has serious limitations which become apparent when the frequency of the signal to be amplified is high (many MHz). These difficulties are mainly associated with the capacitance that exists between the various electrodes of the valve. The capacitance values, though very small (typical values are in the range 2–10 pF) are sufficient to affect the performance. Alternating currents exist through the grid-anode and grid-cathode inter-electrode capacitors, the former being generally the greater since the a.c. voltage between the anode and grid is usually greater than that between the cathode and grid. Two effects result from these currents—the input impedance between grid and cathode which is otherwise many megohms falls to a relatively low value, and an a.c. voltage is induced in the grid circuit by the current through the impedance between grid and cathode. The reduction of input impedance can seriously affect the behaviour of any circuit connected to the grid, whilst the latter effect, known as internal feedback, can either reduce the voltage gain of the amplifier to a low value or enhance the gain to such an extent that the amplifier becomes self-oscillating. The choice between these two possibilities depends upon the phase of the voltage induced in the grid circuit, this in turn depending upon the impedance of the anode and grid circuits. The effect of feedback is considered further in Chapter 10.

9.4. MULTI-ELECTRODE VALVES

Whilst valves such as the heptode (seven electrodes) and the hexode (six electrodes) are used for specialized purposes, multi-electrode valves are predominantly tetrodes or pentodes. These latter were developed in the search for valves with low interelectrode capacitance and therefore capable of operation at high frequencies. Initial attempts were made to reduce the interelectrode capacitance of a triode by reducing electrode dimensions, but there is a limit to what can be done in this direction. A great improvement was achieved by the introduction of a second grid to form the tetrode valve.

9.4.1. *The Tetrode Valve*

The capacitance between two conductors can be effectively reduced to zero by enclosing one of the conductors in a conducting screen held at a constant electric potential. Electric flux leaving one of the conductors does not then terminate on the other, but is terminated rather on the enclosing screen. Whilst a continuous screen cannot be placed between the anode and control grid of a triode as electrons have to pass from one to the other, a great reduction in capacitance between them is obtained if an open wire structure such as a fine wire helix is placed so as to enclose the cathode and control grid. Thus a tetrode is similar in construction to a triode except that a further grid, called the screen grid, is placed between the anode and control grid. The screen grid is maintained at a constant potential slightly less than that of the anode, and a capacitor is usually connected externally from the screen grid to the cathode so as to provide a low impedance path for the current through the capacitor formed by the anode and screen grid. A schematic diagram is shown in Fig. 9.13. The introduction of the screen grid can reduce the anode-control grid capacitance by a factor of a thousand. This effect is accompanied by other changes of valve characteristic. Not only is the control grid screened from the effect of the anode, but there is also additional screening of the cathode from the anode. Thus a given change of anode voltage produces a much smaller change of current from the cathode than before, and the ratio $\delta V_a/\delta I_a$, for V_g = constant, is much increased.

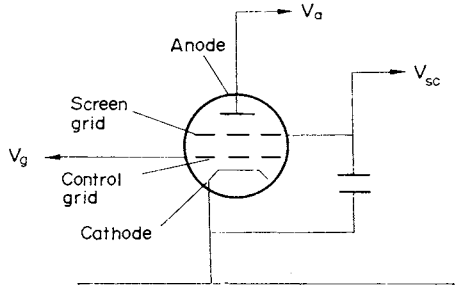


FIG. 9.13. Tetrode schematic with screen-cathode capacitor.

The anode slope resistance for the tetrode is of the order of several hundred thousand ohms. The mutual conductance, being governed by conditions in the cathode-control grid region, is virtually unchanged by the addition of the extra grid. The amplification factor, given by the product $r_a g_m$, is therefore greater than that of the triode, being typically in the range 100–200. The tetrode amplifier thus has the possibility of a greater gain than one using a triode valve. The introduction of the screen grid has an unfortunate effect on the anode characteristic of the valve however. There are now four variable parameters to be taken into consideration, and two have to be kept at a constant value for a two-dimensional plot to be made. Figure 9.14 shows a typical anode characteristic for a fixed screen grid potential and several

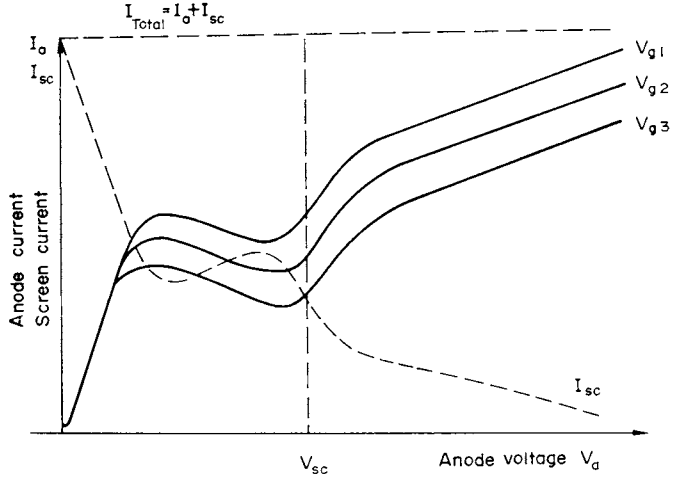


FIG. 9.14. Tetrode anode and screen characteristics for given screen voltage.

different but constant values of control grid potential. As before all potentials are relative to the cathode as zero. The useful range of values of anode potential for linear amplification is seen to be limited. The distortion of the characteristic is due to the secondary emission of electrons from the anode, produced by the impact of electrons that have crossed the valve from the cathode. Such secondary emission invariably takes place at the anode of a valve when a current flows through it, but in the case of a diode or triode

the secondary emitted electrons are released in an electric field whose direction is such that they are attracted back to the anode and there is no net effect in the external circuit. When the tetrode is operated under the condition that the potential of the anode is less than that of the screen grid, the secondary emitted electrons are attracted away from the anode and collected by the screen grid. Over the portion of the characteristic having a negative slope there is a greater loss of electrons by secondary emission than gain of electrons due to the increase of beam current, when the anode voltage is increased. If the anode potential becomes greater than that of the screen grid secondary emitted electrons are recollected by the anode and the screen grid current falls to a low value. This current is that due to interception of the electrons by the screen grid. As expected the sum of anode and screen grid currents is approximately constant and equal to the cathode emission current. The negative slope of the anode characteristic, corresponding to a negative anode slope resistance has applications; a valve operating in this region can be used to make an oscillator, but it is generally an undesirable feature. The distortion due to secondary emission can be removed by introducing a further electrode, producing the pentode valve.

9.4.2. *The Pentode Valve*

The secondary emission of electrons from the anode can be suppressed by imposing at the anode surface an electric field to repel the electrons back to the anode. This is accomplished by the introduction of a third fine wire helix, known as the suppressor grid, between the screen grid and the anode. The suppressor grid is connected, usually internally, to the cathode of the valve. It is thus always at a negative potential with respect to the anode. Electrons are still attracted away from the cathode since there is an electrode —the screen grid—at a high positive potential closer to the cathode than the suppressor grid. The electrons are slowed down in the region between the screen and suppressor grids, but experience additional acceleration between the suppressor grid and anode to make up for this.

Pentode valves are similar in construction to triode and tetrode valves. The schematic representation of a pentode valve is shown in Fig. 9.15.

The mutual conductance of the pentode valve is little different from that of the triode or tetrode valve as conditions close to the cathode are unchanged.

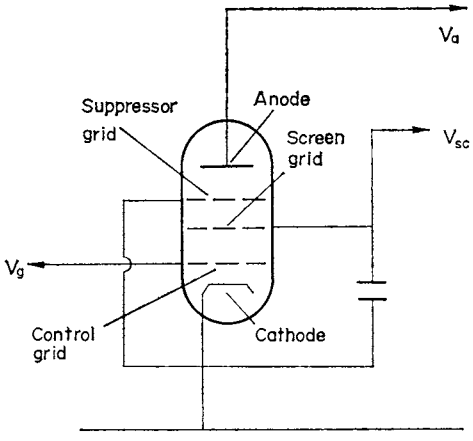


FIG. 9.15. Schematic of pentode and connections.

There is even more screening between the anode and the cathode region and the values of anode slope resistance and amplification factor are therefore greater than those for the tetrode, being of the order of a megohm (a million ohms) and a thousand respectively. A typical anode characteristic for a pentode, as shown in Fig. 9.16, has a wide range of permissible anode voltage without the characteristic becoming non-linear.

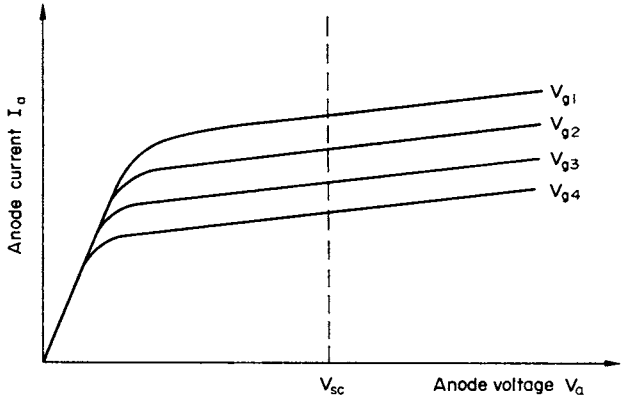


FIG. 9.16. Pentode anode characteristics.

Tetrodes and pentodes can be used in conjunction with anode load resistors or other impedances to give voltage amplification as already considered for the case of the triode. The equivalent circuit of all three amplifier valves is the same, the appropriate values of μ and r_a being used for each valve. Since r_a for a pentode is very high nearly all the current from the source in Fig. 9.12 (b) must flow through the circuit connected to *ab*. The pentode can act therefore as a constant current device.

Because of its large amplification factor, large permissible anode voltage variation, and very low anode-control grid capacitance, the pentode valve is the main thermionic valve used in high-gain low-power amplifiers. The high anode-slope resistance limits the alternating current that can pass through the valve to a very small value. If large a.c. currents are needed, for example to work an electromechanical device such as a loudspeaker, a low impedance valve must be used. Thus for thermionic valve radio or television receivers it is usual to first obtain amplification of the minute signal received at the aerial by a number of amplifier stages using pentode valves, and then when the voltage of the signal is sufficiently large, to use it to produce a large alternating current in a low impedance valve for operating the loudspeaker.

9.4.3. *The Beam-power Tetrode*

An alternative approach for overcoming the problem of electron emission from the anode is to increase the electron density between the screen grid and the anode to such an extent that electrons leaving the anode are repelled back to the anode by the space-charge force of the high density electron cloud. As for the space-charge limited diode (Section 9.1.1), the presence of the electrons depresses the electric potential. It is necessary for the decrease of potential in the screen grid–anode region to be sufficient to suppress electron emission even if the screen grid is at a positive potential with respect to the anode.

The increased electron density is achieved by forcing the electron trajectories into a constricted area of the valve cross-section, using beam forming electrodes at cathode potential as in Fig. 9.17. The control-grid and screen-grid helices are made with the same wire spacing and are accurately aligned to ensure that the screen-grid wire does not intercept the electron beam.

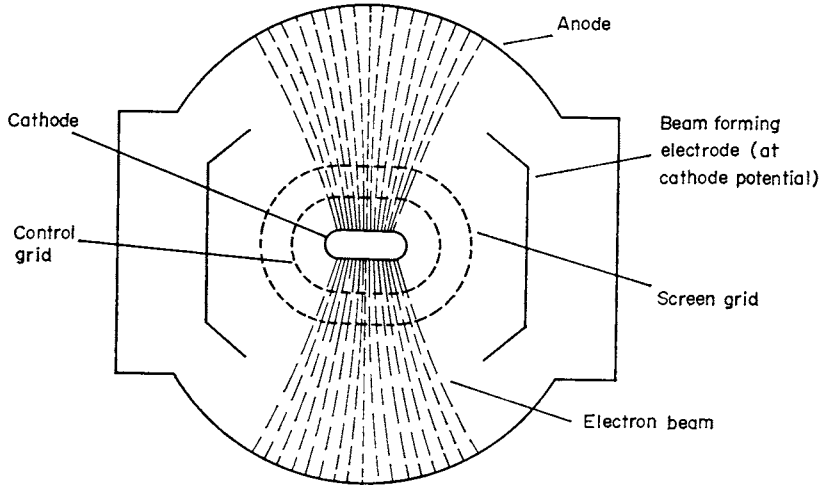


FIG. 9.17. Plan view of beam power tetrode.

Beam power tetrodes have a lower impedance, and are therefore higher power valves, than the pentode. Moreover, they are capable of greater amplification than the triode. They are used as medium-power amplifier valves.

9.5. GAS-FILLED VALVES

The characteristics of high vacuum diodes and triodes are greatly changed when a small quantity of gas is introduced into the envelope, and the resultant valves have many important applications. The gas diode is widely used as a low voltage rectifier and the gas triode, or thyratron as it is called, can be used as a fast-acting electronic switch. Schematically the presence of gas in the valve is shown by shading the diagram or placing a dot within it as in Fig. 9.18.

FIG. 9.18. Gas valve schematics.

9.5.1. *The Gas Diode*

The main disadvantage of the high vacuum diode is that the current through the valve is limited by space-charge effects to a relatively low value (usually tens of milliamps) and this limit can only be raised by an increase of voltage across the valve. The resistance introduced into the circuit by the valve is high and the valve dissipates appreciable energy, e.g. a valve passing a current of 50 mA, with an anode voltage of 100 V, is equivalent to a resistance of 2000 Ω and dissipates 5 W. A quite different characteristic is obtained when a rare gas such as neon or argon, or mercury vapour, is added at a fraction of a mm Hg pressure (atmospheric pressure is equivalent to 760 mm Hg pressure). The voltage–current characteristics of the high vacuum and gas diode are shown for comparison in Fig. 9.19. When the voltage across the valve is less than the ionization potential V_i for the gas

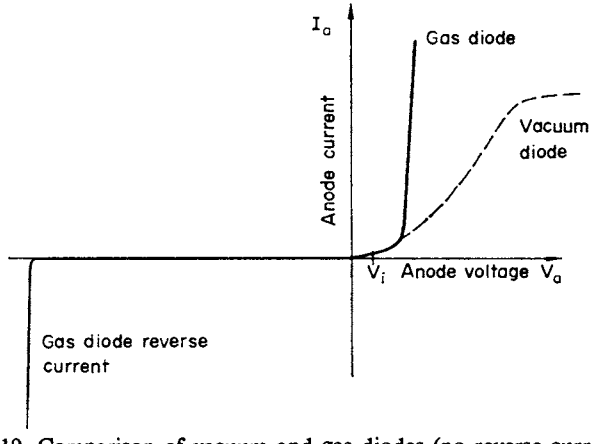


FIG. 9.19. Comparison of vacuum and gas diodes (no reverse current for vacuum diode).

the current through the gas diode is a little less than that for the vacuum diode due to the effect of collisions of the electrons with gas molecules. For positive anode voltages a little greater than the ionization potential the anode current in the gas diode rises to a large value limited mainly by the resistance of the external circuit. The difference in behaviour of the two types of

diode can be explained in terms of the electron–gas molecule collisions described in Section 2.5. Elastic collisions are the only ones that take place when the anode voltage is less than the ionization potential since no electron can then acquire sufficient energy to ionize the molecule with which it collides. The resistance of the gas diode is then a little greater than that of the vacuum diode since a fraction of the energy of the electron is lost to a molecule on each collision. As soon as the anode voltage becomes appreciably greater than the ionization potential of the gas, ionization takes place, and electron avalanches occur in the anode–cathode region. These avalanches produce large numbers of positive ions and electrons which can carry charge across the valve, and result in a greatly enhanced valve current. Perhaps the most important role is played by the positive ions, which because of their large mass compared with that of the electron move relatively slowly towards the cathode and intermingle with the electrons of the space-charge cloud. This results in neutralization of the space-charge when the positive ion density is equal to that of the electrons, and there is no longer a field at the surface of the cathode repelling emitted electrons back to the cathode surface. The cathode can then freely emit the full electron current corresponding to its temperature, as given by the Richardson–Dushman equation. The combined effect of the removal of the space-charge limitation and the production of copious additional electrons allows the current to rise to such a value that the valve would be damaged by bombardment of the cathode with positive ions unless a resistor is included in series with the valve to limit its current to a safe value. A gas diode passing a current of 200 mA with an anode voltage of 20 V corresponds to a resistance of only 100 Ω and a power dissipation of 4 W. The great disadvantage of the gas diode is seen from the reverse part of its characteristic, i.e. when the anode is negative with respect to the cathode. It is desirable in most applications of the diode that no current should flow under this condition, and this is found to be true for anode voltages up to a certain value, usually of the order of a few hundred volts. At higher negative voltages than this the valve again becomes conducting and a current passes through the valve in the opposite direction to that in which the current normally passes.

For the negative anode potential condition no electrons can be emitted by the cathode and pass to the anode. Suppose it is imagined that an elec-

tron is introduced in some way into the valve. The electron will be accelerated by the field towards the cathode and if the voltage is sufficient an electron avalanche will result. The positive ions produced in the avalanche are accelerated towards the negative anode and upon collision with it can liberate electrons from its surface by secondary emission. Each liberated electron can itself produce an electron avalanche across the valve. If the effect at the negative anode surface of all the positive ions resulting from a single avalanche is to cause on the average more than one electron to be emitted the effect is cumulative, a chain reaction is set off, and the number of electrons and positive ions grows very rapidly. Electrical breakdown is said to take place, the valve becomes electrically conducting, and can no longer be used for rectification of alternating currents. The gas diode must therefore be operated with the anode voltage less than this breakdown voltage.

The choice of gas is governed by the desire for a gas that is easily ionized, and one that is chemically inactive and that does not therefore react with tube materials such as the cathode coating. One of the rare gases such as xenon, or low pressure mercury, is normally used.

Because of its desirable characteristic of being able to pass a large current at low anode voltages, with corresponding low values of resistance and power dissipation, and subsequent high efficiency, the gas diode is generally to be preferred to the vacuum diode provided that the operating voltages are not too high. It is however more easily damaged if the anode current becomes too high. In high voltage applications the high vacuum diode has to be used.

9.5.2. *The Thyratron Valve*

The thyratron is an electronic switching tube which behaves as an effective open or closed circuit between anode and cathode, depending upon whether the voltage applied between the control electrode and cathode is less than, or greater than a certain critical value. Hence the derivation of the name, *thyra* being the Greek for door.

In a high vacuum triode no anode current can flow if the control electrode is made sufficiently negative, i.e. if the valve is beyond cut-off. If a low pressure inert gas were to be introduced into such a valve when beyond cut-

off, no electron avalanches would be formed even if the anode voltage is considerably greater than the gas ionization potential, since no electrons pass through the grid to initiate the avalanches in the region between the grid and anode. A reduction of the negative voltage at the control grid so that the valve is no longer beyond cut-off would allow electrons to pass through the grid, ionization and electron avalanches to occur, and a large current to pass through the valve. The resistance of the thyratron between anode and cathode thus falls from a very large to a very small value simply by a change of control grid voltage. Once the valve becomes conducting positive ions move to the negative control grid where their positive charge neutralizes the effect of the negative grid voltage and the valve cannot be made to cease conduction simply by making the grid more negative. In order to stop conduction the anode voltage must be reduced to a low value so that ionization can no longer take place. It is therefore possible to make the valve conducting by "triggering" it, i.e. by applying a positive pulse of voltage to the control electrode whose d.c. voltage is such that the valve is normally just beyond cut-off.

The relationship between the anode and grid voltages for the onset of conduction is shown in Fig. 9.20. The onset is governed by the cut-off

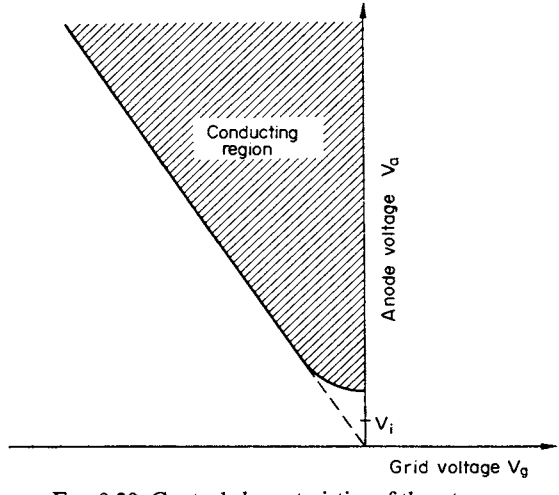


FIG. 9.20. Control characteristics of thyratron.

conditions, and since an increase of grid voltage will require a proportional increase of anode voltage to keep the valve just at cut-off, it is not surprising that the relationship in Fig. 9.20 is a linear one over most of the range. There is a deviation from linearity at the low voltage end of the characteristic since no ionization can take place if the anode voltage is too small. The slope of the linear region, V_a/V_g, is known as the control ratio of the thyratron, being a measure of the relative effectiveness of the anode and grid at the cathode surface. It is somewhat akin to the amplification factor of the triode.

The physical construction of the thyratron is very different from that of the high vacuum triode. Bombardment of the control electrode by positive ions results in considerable dissipation of energy and a fine wire grid would soon become red hot and disintegrate. A disc with a central hole takes the

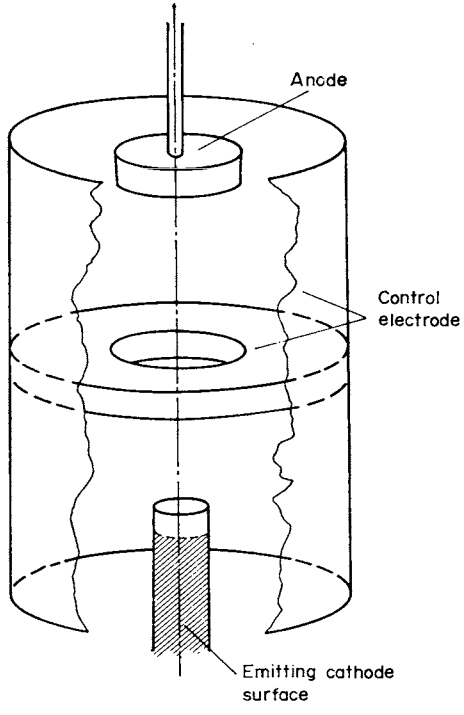


FIG. 9.21. Cut-away view of thyratron electrodes.

place of the wire helix as in Fig. 9.21. An attempt is made to shield the cathode surface from positive ion bombardment by putting the cathode coating on the curved surface of an axial cylinder. Even though the physical construction is so different from a helix grid tube, the principle of operation is still as previously outlined.

9.5.3. *The Neon Stabilizer Valve*

Use is made of the phenomenon responsible for conduction in the gas diode with negative anode voltage, in the voltage stabilizer valve. The schematic representation, physical construction, and characteristics of this valve are shown in Fig. 9.22. It consists of two metal electrodes mounted in a glass envelope containing a low pressure gas, usually neon. When the voltage

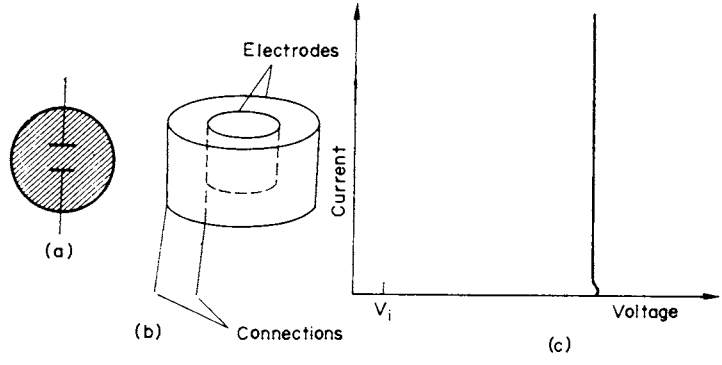


FIG. 9.22. Schematic diagram, electrode system, and *V–I* characteristic of voltage stabilizer valve.

across the valve is sufficiently high, e.g. 80 volts, the valve becomes conducting. The magnitude of the voltage required depends upon the gas type and pressure, and the electrode geometry, dimensions and material. The voltage across the valve varies very little even when the current through the valve is varied over a wide range, and it can therefore be used as a constant voltage device. It has now been largely superseded by solid-state devices, e.g. the Zener diode, especially for lower voltage applications. These devices use the characteristic of approximately constant voltage across the diode under the reverse-bias electrical breakdown condition.

10. *Electronic Circuits*

10.1. APPLICATIONS OF ELECTRON DEVICES

Most electron devices fall into one of two categories. The first category includes those whose application depends upon their ability to allow current in one direction only, as in a thermionic or solid-state diode. The second category consists of devices in which the current can be controlled, and amplification or switching of an electric signal can be obtained when the device is connected into an appropriate circuit. This category includes voltage-controlled devices such as the field effect transistor and multi-electrode thermionic valves, and bipolar transistors.

The unidirectional property of diodes is used for the rectification of alternating voltages, and the detection of amplitude modulated radio waves. In both cases a unidirectional voltage is produced from an alternating supply voltage. Rectification is used for example to produce a d.c. supply from an a.c. supply at power frequencies (e.g. 50 Hz). Detection is a process in which a high-frequency signal whose amplitude is varying with time is made to give a unidirectional voltage varying in the same way as the envelope variations of the h.f. signal. The conversion is usually at low power levels as in a radio receiver.

Amplification enables alternating voltages to be increased in magnitude without changing the shape of the wave. Such amplification is required for example if the signal is to work an electromechanical device such as a loudspeaker where powers of watts are required, and the available signal received from a microphone or gramophone pick-up is very small, e.g. a few microwatts. It is possible by the use of an appropriate electrical circuit, to return a fraction of the amplified signal to the input of the amplifier. This process is known as feedback, and if this feedback is sufficiently strong and of the right phase, the system can become unstable and break into electrical

oscillation. The electron device operated in this way acts as an a.c. generator, converting energy from the d.c. supply to energy in the form of a high-frequency signal.

In recent years there has been an increased emphasis on digital rather than analogue methods. The analogue approach uses electrical signals which at all times are proportional to the intensity of the non-electrical parameter (e.g. sound intensity) of interest. The digital approach is to sample the signal at very short time intervals, and use a series of pulses that represent the magnitude of the sample. A basic requirement is then switching circuits, which are either on or off and must be fast acting.

There are an almost infinite number of different ways and different circuits in which electron devices can be used. They can be used for frequency changing, voltage stabilization, wave forming and shaping, pulse generation, pulse counting, triggering and for many other purposes, in addition to the more straightforward applications of rectification, amplification and switching already mentioned. Discussion here will be limited to some of the basic modes of operation.

10.2. RECTIFICATION

Electric power distribution is by alternating voltage supplies at low frequency. There are many reasons for this. An alternating voltage is the natural output from a coil rotating in a magnetic field: it is possible to transform a.c., e.g. a low-voltage high current supply can be converted to one of high voltage and low current, and distribution can take place at high voltage and low current with a subsequent reduction in conductor dimensions: there are many ʌɹəʌ convenient a.c. electrical machines such as the induction motor. However, many pieces of electrical apparatus require the use of d.c. voltages, especially those employing thermionic and semiconductor devices, and some means of producing a d.c. supply from an alternating one is needed.

Three main types of rectifying devices are available, the vacuum and gas thermionic diodes, and the semiconductor diode. A comparison of their characteristics is shown in Fig. 10.1. The good conductivity of the semiconductor diode at very low voltages has resulted in its almost universal use

for relatively low-voltage rectification. For applications in which the voltage is many hundreds of volts or more, the maximum reverse voltage allowable in the gas diode and semiconductor diode is not sufficient and the high vacuum diode must be used.

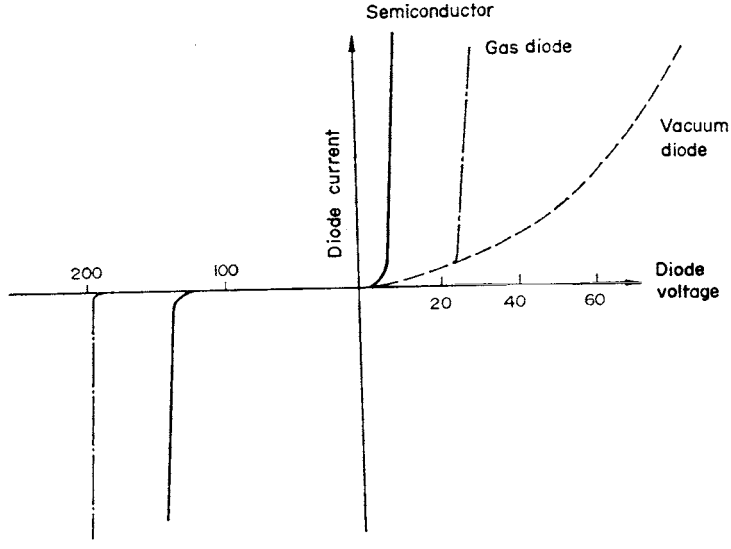


FIG. 10.1. Comparison of *V–I* characteristics of diodes. (Note change of scale.)

10.2.1. *Rectifier Circuits*

Consider the circuit of Fig. 10.2, in which a diode (thermionic valve or semiconductor) is connected in series with a resistor R to an a.c. supply. Current can only flow when the diode voltage has the right polarity; the current through the resistor and the voltage across it are then as shown. This voltage is unidirectional but pulsating, and can be converted to an almost constant voltage by connecting a large capacitor across the resistor as in Fig. 10.3. During the conducting part of the cycle for the diode, the capacitor will be charged up to the peak value of the a.c. voltage. The capacitor cannot discharge back through the diode since the polarity is

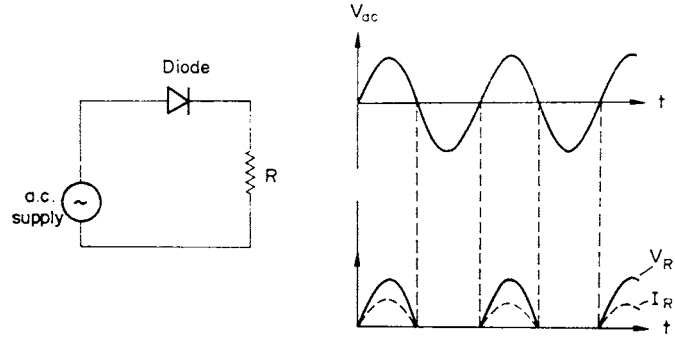


FIG. 10.2. Voltage and current waveforms for diode and resistor connected to an a.c. supply.

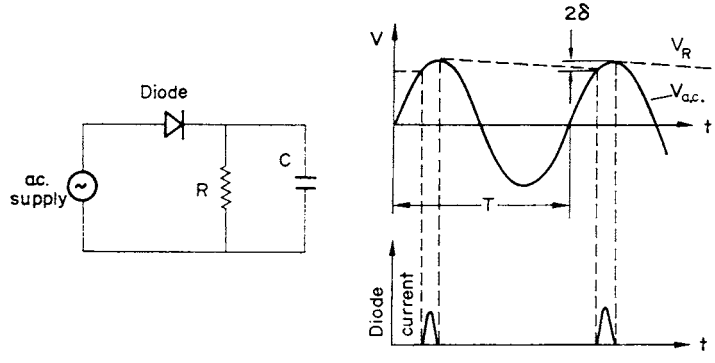


FIG. 10.3. Voltage and current waveforms for half-wave rectifier circuit.

wrong, and it will simply discharge through the load resistor R. The variation of voltage across R is thus as shown. Current will flow through the rectifier only during the very short period when the polarity is right and when the instantaneous value of the applied sinusoidal voltage is greater than the instantaneous voltage across the capacitor. The current flowing through the diode is then just enough to recharge the capacitor to the peak voltage of the a.c. This current flow is also shown in the figure.

The magnitude of the ripple accompanying the d.c. output voltage can be calculated from a knowledge of the way in which the voltage across a ca-

pacitor C, initially charged to a voltage V_o, varies when discharging through a resistor R. This voltage is given by

$$v = V_o \exp\left(-\frac{1}{RC}t\right). \tag{10.1}$$

If the peak value of the a.c. voltage is V_m, after a time T which is the period of the a.c. waveform and very closely the discharge time for the capacitor, the voltage across the capacitor has fallen to

$$V_m \exp\left(\frac{-1}{RC}T\right),$$

and the total decrease of voltage 2δ is given by

$$2\delta = V_m\left[1-\exp\left(-\frac{1}{RC}T\right)\right]. \tag{10.2}$$

The exponential $\exp(ax)$ can be expanded in the series

$$\exp(ax) = 1+ax+a^2x^2/2+\ldots$$

and if $(1/RC)T$ is small compared with unity,

$$2\delta \doteq V_m\frac{T}{RC}.$$

The ripple is normally expressed as half the total drop of voltage, the ripple being considered as an alternating voltage imposed on the d.c. output voltage. The peak value of this alternating component is then δ, given by

$$\delta = V_m\frac{T}{2RC}. \tag{10.3}$$

The ripple magnitude thus depends upon the values of R, C, and the period T (and hence the frequency f) of the a.c. supply, becoming small if R, C, and f are large. However, the supply frequency and the load resistance are usually beyond control, being governed by other considerations, and C is the only parameter that can be changed at will. The value of C is normally limited in practice to tens of microfarads since the larger the value of C the greater the peak charging current through the diode.

Apart from increasing the value of C indefinitely, two methods are available for decreasing the amplitude of the ripple voltage. A low pass filter can be inserted between the capacitor and the load to attenuate the ripple, or the time interval between successive charging of the capacitor can be reduced by a factor of two by the use of both halves of the a.c. waveform.

The filter consists of an inductor L and a further capacitor C' connected in series as in Fig. 10.4. V_o is the d.c. voltage, and v_1 and v_2 the input and output ripple voltages. The output d.c. voltage is taken across the second

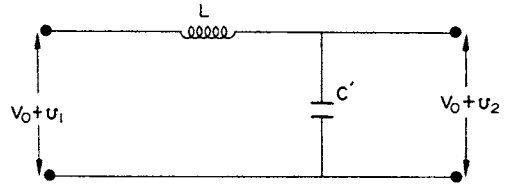


FIG. 10.4. Ripple filter.

capacitor. The filter can be considered as a potential divider, and if the ripple frequency is f_r, ($\omega_r = 2\pi f_r$), we obtain

$$\frac{v_2}{v_1} = \frac{1/\omega_r C'}{\omega_r L - 1/\omega_r C'} = \frac{1}{\omega_r^2 LC' - 1} \doteq \frac{1}{\omega_r^2 LC'}. \qquad (10.4)$$

It is assumed that $\omega_r^2 LC'$ is much greater than unity. v_2 is thus much less than v_1, and the ripple is greatly reduced. For a half-wave rectifier f_r is equal to the a.c. supply frequency. An increase in the ripple frequency makes this filter more effective.

Schemes using both half-periods of the wave require two rectifiers, and are known as full-wave rectifier circuits in contrast to the half-wave rectifier circuit already considered. Figure 10.5 shows a full-wave circuit with filter, together with the variation of voltage across the first capacitor C_1. When A and B are positive and negative respectively relative to O, rectifier A provides the charging current, and rectifier B is effectively an open circuit. On the reverse half-cycle it is rectifier B that passes the current, and A that is open circuit. The magnitude of the ripple across the first capacitor has

been reduced by a factor of about two by the addition of the second rectifier, and the ripple frequency has increased by a factor of two. Since the filter effectiveness is proportional to the square of the frequency of the ripple, a total

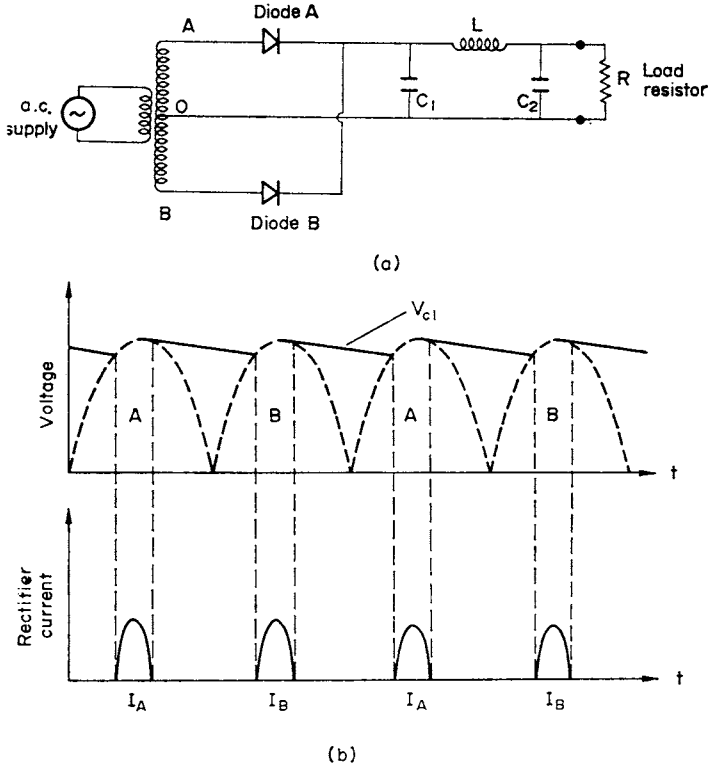


FIG. 10.5. (a) Full-wave rectifier circuit; (b) voltage and current waveforms. Full-wave rectification.

reduction by a factor of eight is achieved compared with the half-wave rectifier circuit.

It should be noted that the voltage across the diode is equal to the sum of the d.c. voltage across the capacitor and the a.c. voltage of the supply. Since the d.c. voltage is approximately equal to the peak value of the a.c. voltage, the maximum voltage across the diode is equal to twice the d.c.

output voltage and occurs when the diode is not conducting. The diode must therefore have a permissible reverse voltage at least twice that of the output d.c.

An alternative scheme, known as a bridge rectifier circuit, is given in Fig. 10.6. This circuit requires four diodes, but has the merit that the output

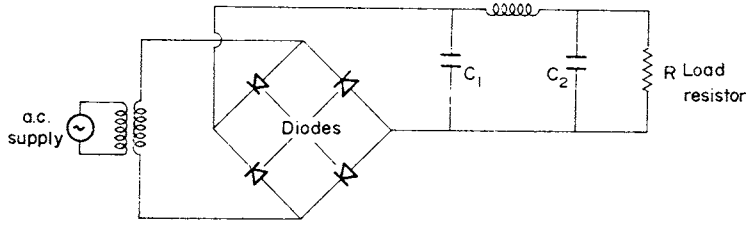


FIG. 10.6. Bridge rectifier circuit.

d.c. voltage corresponds to the full secondary voltage of the transformer, and each diode is subject to a maximum reverse voltage equal to the output d.c. voltage rather than twice its value.

10.3. DETECTION OF RADIO WAVES

Amongst the earliest electron devices were the vacuum diode and "cats-whisker" solid-state detectors used by Fleming and Bose respectively near the beginning of this century to detect radio waves. Audio signals, i.e. signals to which the human ear is sensitive, have frequencies up to about 20 kHz. However, frequencies much higher than this are required if electromagnetic radio waves are to be successfully transmitted over long distances. One way of reconciling this frequency difference is to vary the amplitude of a very high-frequency electric wave (known as the carrier wave) at the audio frequency. This is shown in Fig. 10.7, and is called amplitude modulation. If a current of this form were passed through a loudspeaker, in which the instantaneous force on the moving system is proportional to the instantaneous current passing through it, there would be no movement of the speaker cone. The mechanical system of the loudspeaker cannot respond to the high frequency of the carrier wave, and since the mean current is zero the mean

force on the cone will also be zero, i.e. there will be no response. One way of producing a net force is to remove the negative half of the signal by the use of a unidirectional device such as a diode. The process is called demodulation, and the circuit used is similar to that of Fig. 10.3, with the output voltage being taken as before across resistor R. The values of C and R in this case must be such that the voltage across the capacitor decays little in the period τ_c of the carrier wave, but has a large variation over the period τ_a

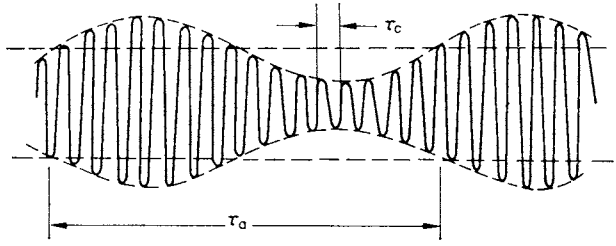


FIG. 10.7. Amplitude modulated carrier wave.

of the audio frequency. The output voltage cannot then follow the carrier wave variations, but will tend to follow the envelope of the wave. An alternative method of demodulation uses the fact that some diodes have a voltage–current characteristic in which the current is approximately proportional to the square of the voltage. If the modulated signal is connected to such a diode the output current is always positive. This square law detection method is essentially a low-power device, and since it introduces distortion, it is only used when no other method is available, which is the case at ultra-high frequencies.

10.4. AMPLIFICATION CIRCUITS

The basic processes of amplification using thermionic valves and transistors were discussed in Chapters 5, 7 and 9. It was shown that the amplitude of an alternating voltage wave could be increased without change of the wave-shape, when the electronic device was incorporated in an appropriate circuit. In order to achieve this voltage amplification, the alternating com-

ponent of the current through the device must be made to flow through some electrical impedance. In the previous chapters discussion was limited to examples in which this impedance was purely resistive.

10.4.1. *Comparison of Field Effect Devices and Bipolar Transistors as Amplifiers*

Circuits used to produce amplification with field effect devices, e.g. thermionic valves, and transistors are similar as shown in Fig. 10.8. The transistor shown is a *p–n–p* type, and is connected in the grounded emitter

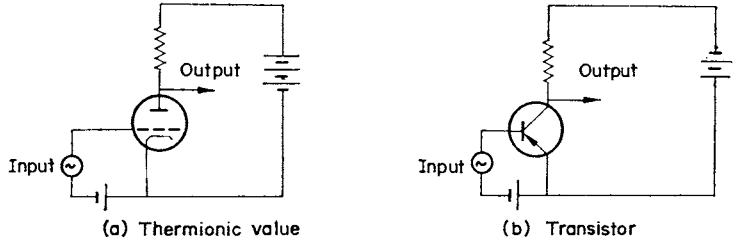


FIG. 10.8. Basic amplifier circuits.

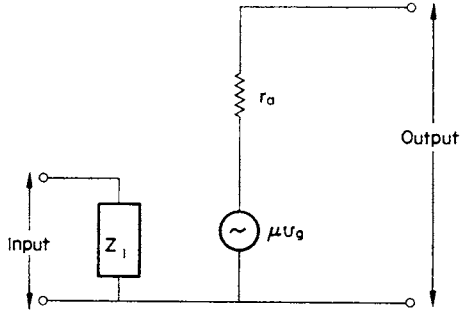
state—by far the most common form of connection. In spite of the similarity between these basic circuits, there are however several important differences between the two amplifying systems. These include:

(i) the triode or FET are essentially *voltage* controlled, whilst the transistor is base *current* controlled. This is exemplified by the different definitions of amplification factor. For a thermionic valve, for example, the amplification factor μ was defined as $|\delta V_a/\delta V_g|$ for a constant value of anode current, whilst the amplification factor α' for a grounded-emitter transistor was defined as $|\delta I_c/|\delta I_b|$ for a constant value of collector-emitter voltage. In practice α' for the transistor is of the same order of magnitude as μ for a triode.

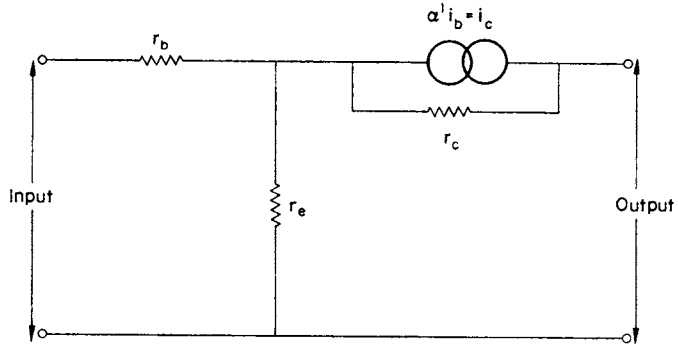
(ii) the input impedance of the transistor is very small compared with that of a field effect device. In the simplest description of triode operation, the control grid has a negative voltage relative to the cathode, no electrons can therefore be collected by the grid, and therefore no grid conduction current

can flow. The input impedance is thus very high and little or no power is required to control the current flow. In practice the input impedance has a value of many megohms. By contrast, the base current of a transistor is significant, the input impedance is of the order of only thousands of ohms, and this will affect the behaviour of circuits connected to the input of the transistor unless they are low impedance circuits.

(iii) the input and output circuits of the thermionic valve or FET can be considered to be quite separate and not interconnected unless a circuit has been deliberately introduced for this purpose. This isolation does not occur in the transistor, since a fraction of the emitter current passes into the base



(a) Thermionic valve or FET



(b) Bipolar transistor

FIG. 10.9. Equivalent circuits.

circuit, and there is resultant coupling between the input and output circuits.

(iv) there are important differences between the equivalent circuits of the two devices, which are contrasted in Fig. 10.9. The thermionic valve or FET are most conveniently represented by a voltage source of magnitude equal to the product of amplification factor and input a.c. signal, in series with the anode slope or incremental channel. The transistor is most usefully represented by a current source and three resistors, of which r_c has a large value and can be neglected in many cases as it is in parallel with the current source. This current source has a value equal to the product of grounded emitter amplification factor α' and the a.c. base current. This latter equivalent circuit also indicates the interconnection between input and output, the resistor r_e being common to both circuits. Typical values for r_a are 10,000 Ω for a triode, 100,000 Ω for a tetrode, and 500,000 Ω for a pentode and FET; dipolar transistor resistances r_b, r_e and r_c are typically 200 Ω, 20 Ω, and 1 MΩ respectively.

The field-effect transistor was discussed in Chapter 7, and shown to be similar in characteristics to the thermionic valve as a small signal amplifier with a high input impedance, negligible coupling between output and input, and a similar equivalent circuit. The FET has the added advantage that it may operate without bias of the gate.

10.4.2. *Single-stage Amplifier Circuits*

The gain of a thermionic valve or FET amplifier is always easier to calculate than that of the bipolar transistor amplifier because of its simpler equivalent circuit. The anode load impedance has simply to be connected across the terminals of the equivalent circuit and the voltage across the load impedance calculated. The problem thus becomes a straightforward a.c. circuit problem. For example, if an impedance Z_L is used as the anode load, consisting of a resistance R and a reactance X in series, the equivalent circuit is as in Fig. 10.9 (a) with Z_L across the output terminals, and the magnitude of the voltage gain of the amplifier is given by

$$\text{voltage gain} = \mu \sqrt{\left(\frac{R^2+X^2}{(R+r_a)^2+X^2}\right)}.$$

A similar process can be used for any other load impedance. The corresponding equivalent circuit when an impedance Z_L is connected into the collector circuit of a transistor is as in Fig. 10.9 (b) with Z_L across the output terminals. In this case the output a.c. voltage is $i_c Z_L$, whilst the input a.c. voltage is $i_b Z_i$, where Z_i is the input impedance. The voltage gain is then given by

$$\text{voltage gain} = \frac{i_c Z_L}{i_b Z_i} = \alpha' \frac{Z_L}{Z_i}.$$

The value of Z_i has to be calculated taking into account r_b, r_e and Z_L.

The load impedance can take one of many different forms. A particularly important case is that previously considered, of a purely resistive load. As the impedance of the load does not change with frequency the resultant amplifier has a gain which is independent of frequency. This is important if no frequency distortion is to be introduced, e.g. as in a high fidelity audio amplifier for sound reproduction. The parallel resonance circuit is another important example. The impedance of a parallel combination of a capacitor and an inductor has a very high value at the resonant frequency, and a low value at frequencies well removed from this value, as in Fig. 10.10. An amplifier with such an impedance for its load will have a large voltage gain only at or near the resonant frequency and hence the property of frequency discrimination. A particular frequency can be selected for

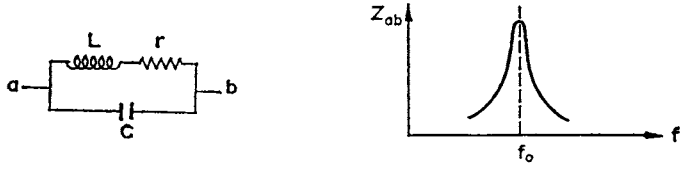


FIG. 10.10. Parallel resonance circuit and frequency response.

amplification in the presence of many others. This has obvious applications in broadcast receivers. Provided that the resistance in the circuit is small compared with the reactance values, the condition for resonance is that $\omega^2 LC = 1$. It is sometimes required to design an amplifier having a particular gain-frequency characteristic, and this can be achieved if an a.c.

circuit can be devised having an impedance–frequency characteristic of the same form. As a simple illustration, if it is required that the gain should increase with frequency a series circuit consisting of inductance and resistance can be used, or a resistor and capacitor in parallel could be used if the gain is to decrease with frequency.

10.4.3. *Grid or Emitter Bias Circuits*

In order to have a high input impedance and give linear amplification, the control grid of a thermionic valve must be provided with a certain d.c. negative voltage relative to the cathode. Likewise the transistor must have an energy source in the input circuit to provide the base d.c. current bias. In previous discussion these bias supplies have been shown as small batteries, but this is obviously an inconvenient method in practice. Alternative schemes are available for dispensing with these batteries. An automatic grid bias circuit for a triode is shown in Fig. 10.11. The capacitor value is suffi-

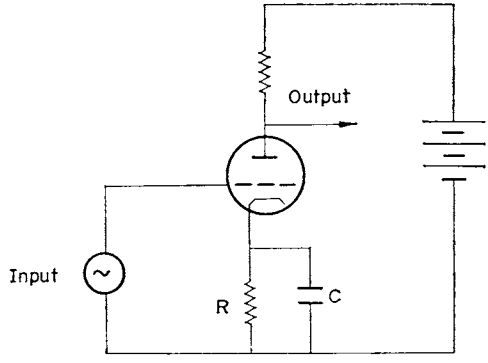


FIG. 10.11. Grid bias circuit for thermionic valve.

ciently high so as to provide a shunt path to the resistor at the operating frequency of the amplifier. If this shunt path has a low impedance relative to the resistor, the d.c. current component will pass through the resistor setting up a d.c. voltage across it, but the a.c. current component will pass through the capacitor giving rise to only a very small a.c. voltage. Resistor

R is chosen so that the voltage developed across it, when the specified anode current flows through it, is equal to the specified grid bias voltage.

A similar scheme may be used to bias a FET when required.

The potential of the base of the bipolar transistor is intermediate between that of the emitter and the collector, i.e. for a p–n–p transistor both the base and the collector are negative relative to the emitter. The same battery can therefore be used to provide both potentials if a potential divider is used. Such a scheme is shown in Fig. 10.12. R_1 and R_2 should have values such that the current through the divider chain is large compared with the base

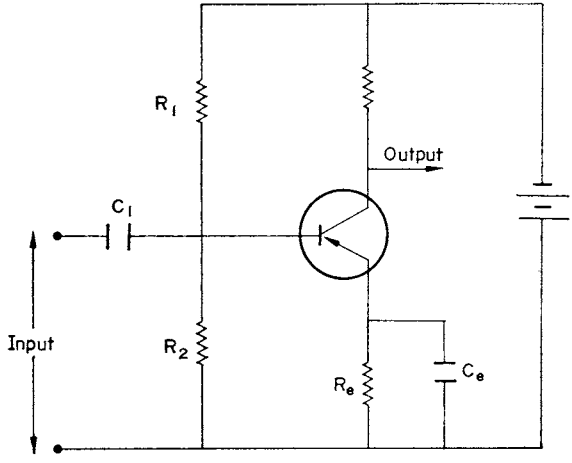


FIG. 10.12. Bias and stabilizing circuits for a transistor amplifier.

current. The base potential is then not critically dependent upon the base current. The capacitor C_1 is included to prohibit flow of d.c. currents into the input signal source. The circuit comprising R_e and C_e is included to stabilize the current through the transistor if the ambient temperature changes. Transistors tend to be temperature dependent, and the d.c. current through them varies as the temperature changes. However with the introduction of the stabilizing circuit, if the d.c. current should increase, the voltage between emitter and base decreases, and the emitter current is restored to its former value. Capacitor C_e is included to provide a low impedance path for the a.c. current.

10.4.4. *Coupling between Amplifier Stages*

The gain obtainable from a single stage may not be sufficient to satisfy the amplifier requirement, and a number of amplifier stages must be connected together to increase the overall gain. Two main methods of achieving this are available, capacitor coupling and transformer coupling.

The output of one device cannot normally be connected directly to the input of another because of their possible different potentials. The anode of a valve for example will normally be a few hundred volts positive with respect to the cathode, whilst the grid will be a few volts negative relative to the cathode. This difficulty can be overcome by connecting the devices via a coupling capacitor C_c, as in Fig. 10.13. Ideally this capacitor behaves as an open circuit to d.c. and as a short-circuit to a.c. The amplifier shown

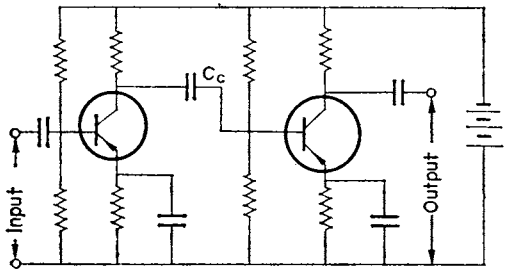


FIG. 10.13. *RC* coupled amplifier.

is known as an *RC* coupled amplifier. The value of C_c is chosen so that its impedance at the operating frequency is small compared with the input impedance of the following stage. Coupling is carried out in the same way for a multi-stage valve amplifier. The value of the coupling capacitor must be much larger for a transistor amplifier than that in a thermionic valve amplifier since the impedances in the circuit are so much smaller. However the much lower operating voltages allow these capacitors still to be physically small.

Since transistors are low voltage devices they are often directly connected, with an appropriate circuit readjustment to ensure each transistor has acceptable d.c. operating conditions.

Coupling via transformers is carried out by connecting the transformer primary winding into the anode or collector circuit, and the secondary winding into the grid-cathode or base-emitter circuit of the following valve or transistor. Since there is no d.c. connection between the primary and secondary windings there is no problem associated with the difference in d.c. potentials. In order to get maximum energy transfer from one stage and the next it is necessary that the transformer should be matched. In the thermionic valve, the anode circuit has an impedance which is usually much less than the input impedance at the grid of the next valve. The primary winding therefore should have less turns than the secondary. The transformer thus gives a step-up of voltage, which is helpful for a device which is basically a voltage amplifier. In the transistor however it is the input base–emitter impedance that is considerably less than the emitter–collector output impedance, and so a transformer with more turns on the primary than on the secondary is required. This would seem to be appropriate for a transistor which is basically a current amplifier, since such a transformer will give a current step-up from primary to secondary. One of the advantages of transformer coupling is that the d.c. resistance of the windings is small, and so nearly all the battery voltage appears across the valve or transistor.

10.5. POSITIVE AND NEGATIVE FEEDBACK

Suppose that a proportion of the output signal from an amplifier is taken and combined, by addition or subtraction, with the input signal. Under these conditions feedback is said to be applied around the amplifier and

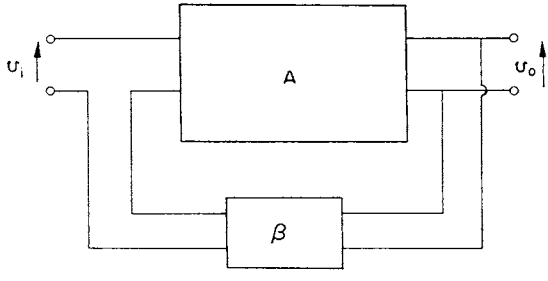


FIG. 10.14. An amplifier with feedback loop.

certain very important and useful effects can be produced in this way. If the fed-back signal is arranged so as to increase the input signal then the feedback is considered to be positive. Negative feedback occurs when the fed-back signal reduces the total input signal.

Figure 10.14 shows schematically an amplifier with a voltage gain of A in the absence of any feedback. The output voltage v_o is also applied to an attenuating network from which a fraction β of this voltage is obtained and then fed back in series with the input voltage v_i. The voltage at the input terminals of the amplifier is therefore $v_i+\beta v_o$. This is amplified A times to give the output voltage v_o.

$$\therefore \quad v_o = A(v_i+\beta v_o), \tag{10.5}$$

$$\therefore \quad \frac{v_o}{v_i} = G = \frac{A}{1-A\beta}. \tag{10.6}$$

G is the gain of the system with feedback applied and is sometimes called the closed-loop gain. If the feedback loop is removed then the gain is A and in consequence this quantity is called the open-loop gain; β is called the feedback fraction.

Consider the voltage v_i to be applied to the input circuit whilst the feedback loop is open. The output voltage is thus Av_i. Now suppose the feedback loop is suddenly closed; the input voltage to the amplifier increases to the value $v_i+A\beta v_i$. This in turn will be amplified, a fraction β of it fed back, and the input voltage becomes

$$v_i+(v_i+A\beta v_i)\,A\beta.$$

Repeating this argument for several cycles around the loop shows that the input voltage to the amplifier builds up as:

$$v_i[1+A\beta+A^2\beta^2+A^3\beta^3+\ldots]. \tag{10.7}$$

The term in square brackets in equation (10.7) is a geometric progression and its sum to n terms can be shown to be:

$$\frac{v_i[1-(A\beta)^n]}{1-A\beta}. \tag{10.8}$$

Positive Feedback

If $A\beta \geqslant 1$ it is obvious that the series in equation (10.7) just goes on increasing and the input voltage to the amplifier will never reach a steady value. The amplifier, however, cannot handle an ever-increasing input signal; some stage in it will probably saturate and the open circuit loop gain A will drop to a very low value or perhaps to zero. The output voltage will then fall to a low value (or zero) and the input voltage will return to the applied voltage v_i. Conditions are now ready for the whole cycle of events to start again. The amplifier has in fact become an oscillator. The waveform and frequency of the output voltage depend very much on the circuit details in the amplifier. The waveform may be far from sinusoidal.

If $0 \leqslant A\beta < 1$, reference to equation (10.8) shows that since $(A\beta)^n$ tends to zero as n tends to infinity, the input voltage to the amplifier settles down at the steady value of $v_i/(1-A\beta)$. The output voltage is thus $Av_i/(1-A\beta)$ and the closed-loop gain is therefore $A/(1-A\beta)$. This is the same value as given by the more direct but perhaps less illuminating method leading to equation (10.6).

For the condition $0 \leqslant A\beta < 1$, the closed-loop gain is seen to be finite and greater than or equal to A. Applying feedback to an amplifier, so that the condition above is satisfied, provides a method for increasing its gain above the value without feedback. Great care, however, has to be taken to ensure that $A\beta$ does not increase beyond unity otherwise oscillations may occur. This technique, sometimes called "regenerative feedback" is not often used because it results in an amplifier that is inherently unstable.

Negative Feedback

Suppose now $A\beta$ is negative. This could easily be done by reversing the connections to the feedback loop. The closed-loop gain is seen from equation (10.6) to be

$$G = \frac{A}{1+A\beta} \tag{10.9}$$

and is always less than the open-loop gain. If $A\beta$ is much greater than unity

then

$$G \approx \frac{A}{A\beta} = \frac{1}{\beta}.$$

Under these conditions the gain is just the reciprocal of the feedback fraction and is independent of the open-loop gain A. Negative feedback is thus often used in order to maintain the gain of an amplifier constant even though the open-loop gain may vary.

As an example, suppose it is desired to make a transistorized amplifier with a gain of 100 that must be held as constant as possible over the working life of the batteries. The designer has at his disposal transistors that may be connected to give a voltage gain of 10 per stage when the supply batteries are new but only 8 when the supply batteries are nearly exhausted. Suppose the following designs are considered:

(a) Three stages with a feedback fraction $\beta = 9/1000$.

At start of life: $A = 10\times10\times10 = 1000,$

$$\therefore\ \beta A = 9,$$

$$\therefore\ G = \frac{1000}{1+9} = 100.$$

End of life: $A = 8\times8\times8 = 512,$

$$\therefore\ \beta A = 4{\cdot}60,$$

$$\therefore\ G = \frac{512}{5{\cdot}60} = 91.$$

(b) Two stages no feedback:

Start of life. Gain $= 10\times10 = 100.$

End of life. Gain $= 8\times8 = 64.$

A gain variation from 100 to 64 with no feedback has to be compared with a variation from 100 to 91 with feedback.

Negative feedback may also be used to eliminate gain variation due to ageing components or production spreads in transistor and valve parameters.

The open-loop gain A may change with frequency if the coupling networks used between stages are frequency dependent or if the valve or transistor

parameters change with frequency (for example α decreases with increase in frequency—see Section 5.3). Negative feedback applied in this case will reduce the overall gain but make it less frequency dependent. There is one factor, however, that must always be considered in design; the open-circuit loop gain which at low frequencies is positive, say, may become negative at higher frequencies. The change results usually from the phase shift introduced by the interstage coupling networks and spurious electrode capacitances. Under such conditions the feedback is no longer negative but becomes positive and oscillation may occur if $A\beta > 1$.

10.6. CATHODE FOLLOWER-TYPE CIRCUITS

The cathode follower circuit has been widely used as a means of increasing the input impedance, and decreasing the output impedance, of electron devices. It is essentially a negative feedback circuit. It will be illustrated by the cathode follower using a triode valve, as shown in Fig. 10.15. A similar

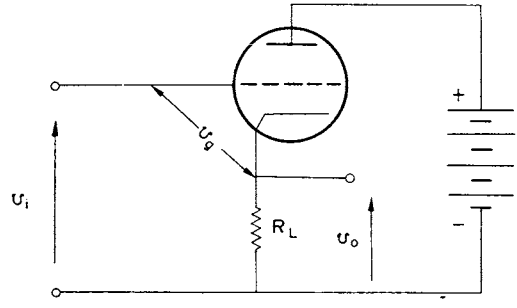


FIG. 10.15. A cathode follower stage.

circuit can be used and analysed in an identical manner for a FET source follower, and it can be extended for a bipolar transistor emitter follower circuit.

The load resistance R_L is seen to be connected in the cathode lead rather than the anode circuit. The input signal voltage to the valve v_g (i.e. the grid cathode voltage) is equal to the difference between the input voltage to the stage v_i and the output voltage v_o.

Thus, $$v_g = v_i - v_o. \tag{10.10}$$

There is therefore 100 per cent negative feedback (i.e. $\beta = 1$) applied between the output and input circuits.

From equation (10.6), when $\beta = 1$,

$$\text{Gain} = \frac{A}{1+A} \quad \text{if} \quad A \gg 1.$$

Thus, the cathode follower might be expected to have a gain just less than unity. In this case

$$v_o \approx v_i.$$

The cathode voltage must therefore follow the grid voltage very closely; hence the name cathode follower.

To be more specific; for the circuit in Fig 10.15

$$v_o = \left[\frac{\mu v_g}{r_a + R_L}\right] \cdot R_L \tag{10.11}$$

where r_a is the anode slope resistance of the triode and μ the amplification factor. From equations (10.10) and (10.11) eliminating v_g:

$$\frac{v_o}{v_i} = G = \frac{\mu R_L}{r_a + R_L + \mu R_L}.$$

Suppose $r_a \gg R_L$ and $\mu > 1$. These conditions are usually true in practice. Then,

$$\frac{v_o}{v_i} = G \approx \frac{1}{1 + 1/g_m R_L} \tag{10.12}$$

where $g_m = \mu / r_a$.

If $g_m R_L \gg 1$, then

$$\frac{v_o}{v_i} \approx 1.$$

A simple equivalent circuit can be deduced from equation (10.12) and this is shown in Fig. 10.16. It should be checked that this circuit gives the same result as equation (10.12). One of the important roles of the cathode

follower is clearly seen from this circuit. It acts as a voltage source equal to the input voltage v_i with an output impedance equal to $1/g_m$. The output impedance $1/g_m$ is typically around 300Ω for a triode and is thus small. The output voltage v_o is practically equal to the input voltage v_i for all values of load resistance R_L greater than several times $(1/g_m)$.

It is relatively easy to show, although the proof is not included here, that the input impedance of a cathode follower stage is approximately $1/(1-G)$

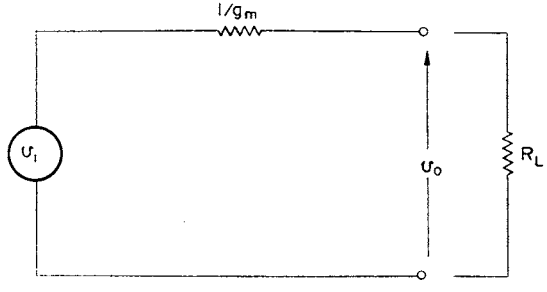


FIG. 10.16. An equivalent voltage generator circuit for the cathode follower.

times the input impedance of the same valve used as a conventional voltage amplifier. Since G is nearly unity for a cathode follower its input impedance is extremely high.

Thus, although the cathode follower has a gain less than unity, it has a very high input impedance together with a very low output impedance. It

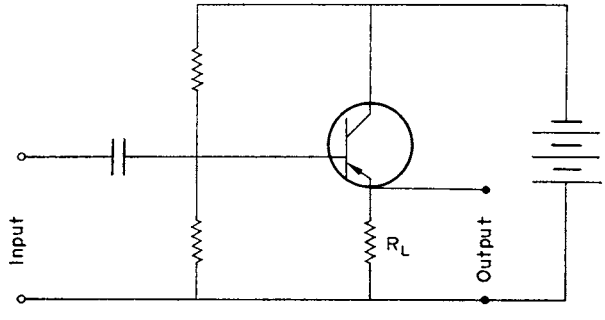


FIG. 10.17. Transistor emitter follower.

can be used therefore to supply a low impedance load from a high impedance source.

A similar approach leads to the emitter follower circuit of Fig. 10.17 for a bipolar transistor.

The emitter follower circuit can be analysed using the rather more complex equivalent circuit for the transistor, and its general characteristics are similar to those of the cathode follower. The two main differences are that the gain is closer to unity than the cathode follower, and both the input and output impedances are lower for the transistor circuit than for the valve circuit.

10.7. OSCILLATORS

Intentional positive feedback is used to produces elf-sustained oscillators. It is usually desirable that the waveform and frequency of an oscillator be controlled carefully. A sinusoidal waveform is most commonly required but in some cases square waveforms, sawtooth waveforms and so on are required. Consideration will be limited here to sinusoidal oscillations only

10.7.1. *The Tuned Drain FET Oscillator*

Consider the circuit shown in Fig. 10.18. A FET with a mutual conductance g_m is shown having a parallel tuned circuit as its drain load. The shunt resistance R represents the combined effect of losses in the tuned circuit and the external load resistance. Feedback occurs since a portion of the output voltage is fed back to the gate circuit via the mutual inductance M.

The equivalent current generator circuit is shown in Fig. 10.19. It is assumed that the incremental channel resistance r_d is much greater than R and can thus be neglected.

Equating currents around the circuit:

$$g_m v_g = i_C + i_L + i_R. \tag{10.13}$$

Equating the voltage drops across the shunt elements:

$$v = i_R R = L \frac{\mathrm{d}i_L}{\mathrm{d}t}, \tag{10.14}$$

$$i_C = C\frac{dv}{dt}. \tag{10.15}$$

Also
$$v_g = M\frac{di_L}{dt}. \tag{10.16}$$

Suppose the required solution is now guessed; this is likely to be a voltage v of fixed angular frequency ω and constant amplitude V_o say,

$$v = V_o \sin \omega t. \tag{10.17}$$

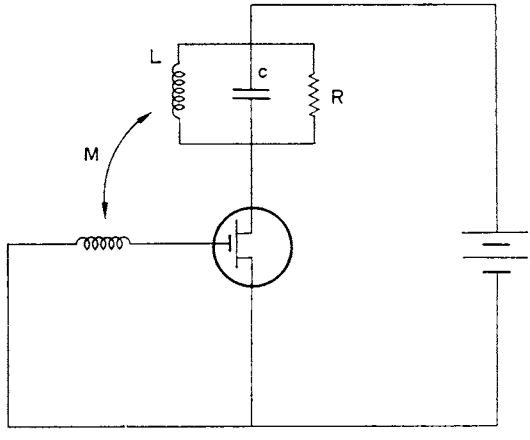


FIG. 10.18. Tuned drain oscillator. The feedback is via the mutual coupling between drain and source circuits.

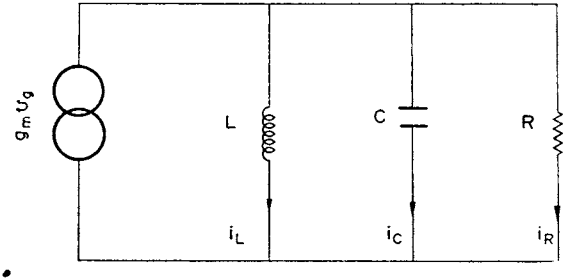


FIG. 10.19. Equivalent current generator circuit for the tuned drain oscillator.

From (10.14) and (10.17)

$$i_R = \frac{V_o \sin \omega t}{R}, \tag{10.18}$$

$$i_L = \frac{V_o}{L} \int \sin \omega t \, \mathrm{d}t = \frac{-V_o}{\omega L} \cos \omega t. \tag{10.19}$$

From (10.15) and (10.17) $i_C = CV_o\omega \cos \omega t.$ (10.20)

From (10.14), (10.16) and (10.17)

$$v_g = \frac{M}{L} V_o \sin \omega t. \tag{10.21}$$

Substituting for i_R, i_L, i_C and v_g in (10.13) from (10.18), (10.19), (10.20) and (10.21)

$$g_m \frac{Mv_o}{L} \sin \omega t = CV_o\omega \cos \omega t - \frac{V_o \cos \omega t}{\omega L} + \frac{V_o \sin \omega t}{R}. \tag{10.22}$$

Equating terms in (10.22) in $\cos \omega t$:

$$\omega^2 LC = 1. \tag{10.23}$$

Equating terms in $\sin \omega t$:

$$\frac{g_m M}{L} = \frac{1}{R}$$

or

$$g_m R \left(\frac{M}{L}\right) = 1. \tag{10.24}$$

Equation (10.23) shows that the frequency of oscillation is the natural resonant frequency of the tuned circuit. Equation (10.24) imposes a condition on the circuit parameters in order that oscillations can be sustained. A more detailed solution would show that if

$$\frac{g_m RM}{L} < 1$$

then oscillations would not occur. The amount of feedback increases as M increases and equation (10.24) gives the condition when there is just sufficient feedback to start oscillations.

At resonance, $\omega^2 LC = 1$ and the positive susceptance of the capacitor $+\omega C$ just balances the negative susceptance $-(1/\omega L)$ of the inductance, leaving the effective drain load as a pure resistance R. The open loop gain A at this frequency therefore is $g_m R$. The feedback fraction β is M/L.

For oscillations to be sustained it has already been shown that $A\beta \geq 1$.

$$\therefore \quad g_m R \left(\frac{M}{L}\right) \geq 1.$$

Thus, equation (10.24) is the limiting condition for oscillations to start.

At frequencies off resonance, the capacitive and inductive susceptances do not balance and there is therefore a finite susceptance in parallel with R. Under these conditions the drain impedance is always less than R and the open circuit gain is not sufficient now to make $A\beta \geq 1$. The system does not oscillate therefore at any frequency other than that given by $\omega^2 LC = 1$.

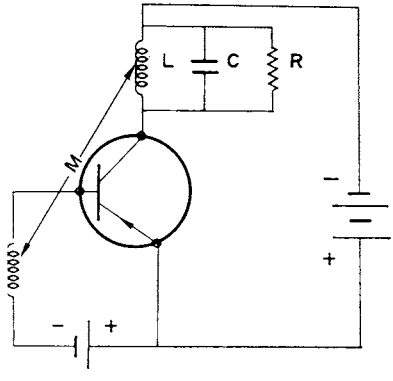


FIG. 10.20. A tuned oscillator using a *p–n–p* transistor in grounded emitter.

There must be some initial gate voltage to trigger off the oscillations. Small fluctuations in the standing drain current through the FET, caused by random motion of charge carriers through the device, are sufficient to induce voltages in the drain circuit and thereby start the oscillations.

There are many other oscillator circuit configurations, involving valves or transistors, but they nearly all rely somewhere on a positive feedback

mechanism. In Fig. 10.20 a tuned collector oscillator circuit using a *p–n–p* transistor with grounded emitter is shown. This is the transistor equivalent of the tuned drain FET oscillator shown in Fig. 10.18.

10.7.2. *Regenerative (or Positive Feedback) Switching Circuits*

In Chapter 6 we have seen how it is possible to operate a transistor in either the OFF or the ON state, and we have noted that with sufficient base current I_b the transistor may become saturated in the ON state. In this section we want to apply positive feedback to a cascaded pair of transistors and observe the resultant behaviour since this is very basic to the operation of many of the circuits used in computers to perform logic functions.

Suppose we connect two identical common emitter transistors T_1 and T_2 in cascade as shown in Fig. 10.21 and then couple the output of the second transistor back to the input of the first so as to apply strong positive feed-

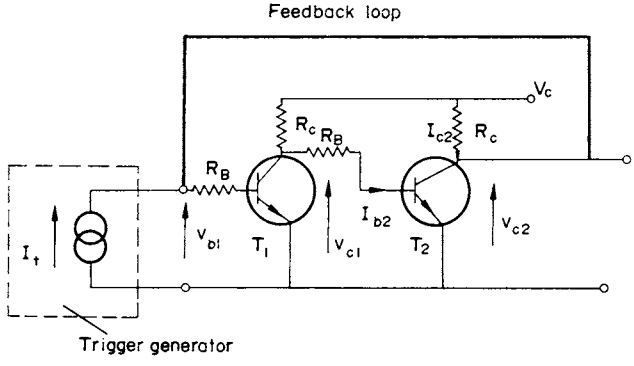


FIG. 10.21. Bistable circuit and trigger generator.

back. The feedback connection is drawn heavily in Fig. 10.21. Let us investigate one possible stable state of this circuit, namely T_1 OFF and T_2 ON. The base current I_{b2} of T_2 is

$$I_{b2} \simeq \frac{V_C}{R_C + R_B} \tag{10.25}$$

if we neglect the small forward bias base–emitter voltage drop across T_2 in comparison with V_C. The collector current of T_2 is

$$I_{c2} \simeq \frac{V_C}{R_C} \tag{10.26}$$

where the collector–emitter drop has been neglected in comparison with V_C since the transistor is assumed to be switched to its low impedance ON state.

We have noted in Chapter 6, Section 6.4 that for saturation to be ensured for transistor T_2 then:

$$I_{b2}\frac{\tau_p}{\tau_F} > I_{c2}$$

or

$$I_{b2} > \frac{I_{c2}}{\alpha'} \tag{10.27}$$

where $\alpha' = \tau_p/\tau_F$, and is the current gain in common-emitter. Using equations (10.25) and (10.26) in equation (10.27), the condition for T_1 to be OFF and T_2 to be ON is therefore

$$\frac{V_C}{R_C+R_B} > \frac{V_C}{\alpha' R_C} \tag{10.28}$$

or

$$\alpha' > \frac{R_C+R_B}{R_C}$$

which is not difficult to arrange since $\alpha' \gg 1$.

By a similar argument, since the circuit is symmetrical with respect to both transistors T_1 and T_2, a second state also exists with T_1 switched ON and T_2 OFF. Thus we are confronted with a circuit with two stable states. For obvious reasons, such a circuit is called a BISTABLE circuit, or more colloquially a flip-flop.

Suppose a situation exists in which T_1 is OFF and T_2 ON. If now an additional current generator I_t is arranged so that it can drive a step function of base current through T_1, in such a direction as to switch T_1 on, then, following the argument used in Chapter 6, the collector current of T_1 will increase, v_{c1} will decrease and the bias voltage on the base of transistor T_2

will also decrease tending to turn it OFF. The collector voltage V_{c2} on T_2 will then increase, causing the base voltage V_{b1} of T_1 to also increase. This in turn will enhance the base current drive just applied from the external current generator in order to turn T_1 ON. If the loop gain around the circuit is large enough the cumulative positive feedback is very great and the circuit quickly switches to a new state with T_1 ON and T_2 OFF. The current generator needed in this example to initiate the transition is called the trigger signal. In practice either a current or voltage source could be used to provide the trigger signal. The trigger signal only need be applied long enough to take the transistor T_2 from its saturated ON state back into the small signal regime where it is biased neither hard on nor hard off. After this the positive feedback cycle just described suffices to drive T_1 on even though the trigger bias is now removed.

The full transient behaviour, which is beyond the detail of this treatment, can be analysed using the charge control equations as outlined in Chapter 6, modified to take account of the behaviour of a transistor when it is saturated. Under such conditions, the base-collector junction is no longer reverse biased and minority carriers are injected from it into the base as well as from the emitter-base junction.

The bistable circuit as shown in Fig. 10.21 is seldom used without further modification. However, it is relatively simple to alter it so that a change of state (T_1 off, T_2 on; to T_1 on, T_2 off and vice versa) can be obtained every time a trigger pulse is applied. In this way the circuit reverts to its original state after every second trigger pulse. A cascade of these "scale of two" circuits can be used to count pulses on the binary number scale.

Further modifications to the circuit shown in Fig. 10.21 can produce a monostable circuit which, although capable of being switched from one state to another by a trigger pulse as the bistable was, will revert after a predetermined time to the original state. Only one single stable state exists in this case. A regenerative switching circuit which has no single stable state but which alternates between two metastable states (i.e. between [T_1 on, T_2 off] and [T_1 off, T_2 on]) is called an astable circuit, or multivibrator. The length of time the circuit spends in each state can be controlled by the circuit elements.

11. *The Cathode-ray Tube*

11.1. BASIC CONCEPTS

When used with appropriate auxiliary apparatus, the cathode-ray tube enables a visible display of electrical phenomena to be obtained. It is amongst the most versatile of electrical devices, allowing the characteristics of electrical signals that are varying with time to be displayed, or presenting in visible form information that has been transmitted electrically as in a television receiver. Many non-electrical quantities, such as pressure, light intensity, velocity and temperature, can be made to produce electrical effects by the use of an appropriate transducer, and can therefore also be investigated. The instrument used for such investigation, consisting of a cathode-ray tube and its associated electrical circuitry, is known as a cathode-ray oscilloscope.

Two very important properties are made use of in the cathode-ray tube. An electron moving through a region of transverse electric or magnetic field is deflected by the field, and if the electron is caused to strike a thin

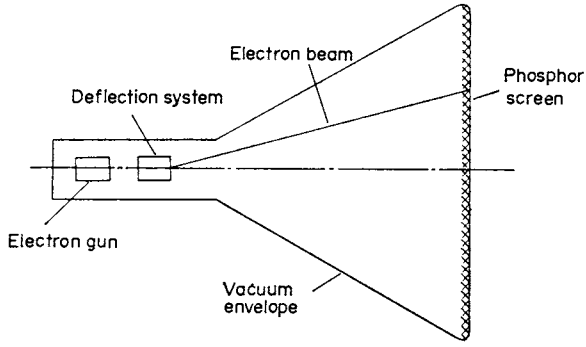


FIG. 11.1. Basic elements of the cathode-ray tube.

film of phosphor coated on a glass screen its position is shown by the presence of a spot of light at the position of impact.

There are three main parts to a cathode-ray tube, as shown in Fig. 11.1, an electron gun which produces a small diameter beam of high speed electrons, a deflection system by which the electrical signal being studied is made to set up a transverse electric or magnetic field to deflect the electron beam, and a detecting phosphor screen to indicate the position of the beam. These elements are enclosed in a glass envelope in which a very low vacuum pressure is maintained. A large range of screen sizes is available, from 1 inch diameter for small portable oscilloscopes to 2 feet in diameter for some modern television receivers. The three constituent parts of the tube will be considered in turn.

11.2. PRODUCTION OF A SUITABLE ELECTRON BEAM

In order to give good definition at the detecting screen it is necessary, not only to obtain emission of electrons from a cathode and accelerate them to a high velocity towards the detecting screen, but also to ensure that the beam has a very small diameter when it collides with the screen. A fine point of light is then produced.

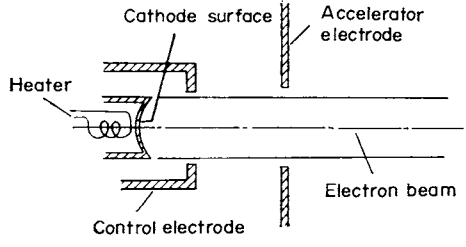


FIG. 11.2. Electron gun.

A typical electron gun is shown in Fig. 11.2. Electron emission is obtained from an oxide-coated cathode, with the emitting surface on the end face of a cylinder enclosing the heater filament. Emitted electrons are attracted away from the cathode surface by an accelerator electrode having a high positive

potential (e.g. a few thousand volts) with respect to the cathode. A central hole in this electrode allows the electron beam to pass along the axis of the tube. An additional electrode, known as the control electrode, is inserted between the cathode and accelerator and has a potential slightly negative (a few volts) relative to the cathode. The function of this electrode, whose potential can be varied, is to vary the electric field at the cathode surface and therefore to control the electron emission rate from the cathode, and hence the brightness of the light spot on the detecting screen.

Since the electrons in the beam all have a negative charge there will be mutual repulsion between them and they will diverge, the beam diameter becoming greater as they do so. This space-charge force was considered in Section 2.6, and unless some focusing scheme is introduced to counteract the effect a large diameter spot is obtained at the screen. Three methods are available for focusing the beam, gas, electrostatic, and magnetic focusing. Of these gas focusing was the earliest historically but is now little used. The technique was deliberately to introduce a small quantity of gas into the vacuum envelope. Subsequent ionizing collisions between electrons and gas molecules result in positive ions being produced along the beam, which tend to neutralize the negative charge of the electrons and remove the cause of spreading. The large mass of the positive ions compared to the electrons results in their being removed only slowly from the beam region and a sufficient density of ions can soon be built up.

In the electrostatic method, electrodes are introduced into the tube to set up electric fields having a radial component. Radial forces are then exerted on electrons passing through the region. Field distributions can be set up so that there is a net inward radial force, producing a net inward component to the electron velocity to counteract the outward space-charge force. Such an electrode system is called an electron lens, and two electrode schemes—an aperture lens and a cylinder lens—are shown in Fig. 11.3. The electric field lines given indicate by convention the force direction on a positive charge, and the force on an electron is in the opposite direction to the field lines. In both cases the radial components of the electric field tend to produce a net reduction of electron beam diameter, as the electrons tend to follow the field lines. The right-hand side of (b) is a divergent lens, but as the electron is accelerated axially whilst passing through the lens it has a

greater velocity in the divergent part and therefore greater inertia, and the divergent effect of the second part of (b) is less than the convergent effect of the first part. A practical scheme for a cathode-ray tube will usually consist of a combination of aperture and cylinder lenses.

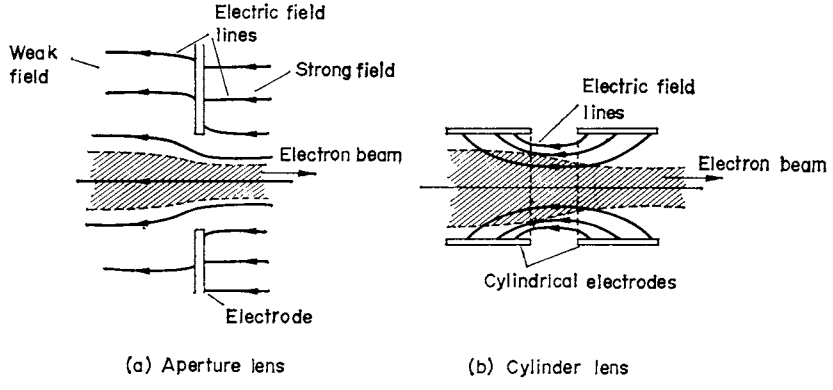


FIG. 11.3. Electron lenses.

Magnetic focusing is achieved by imposing an axial magnetic field along part of the length of the electron beam. In the middle of the region the field is wholly axial, but at the beginning and end of the region there is a large radial component as in Fig. 11.4. Incoming electrons interact with this radial component and a force is exerted on the electrons in a direction mutually perpendicular to the radial field component and to the electron velocity direction, i.e. the force is in a circumferential direction and the electron beam rotates about its axis. There is then an interaction between the electrons as they move circumferentially and the axial magnetic field component, resulting in an inward radial force exerted on the electrons. The radial magnetic field component with which the electron beam interacts on exit slows down the circumferential motion, and the electron beam has no rotation when it is clear of the magnetic field region. A suitable magnetic field can be set up by a current-carrying coil or a permanent magnet magnetized axially as in Fig. 11.4 (b).

The choice between focusing systems is generally an economic one. An electrostatic scheme is built into the tube and is cast away with the tube if

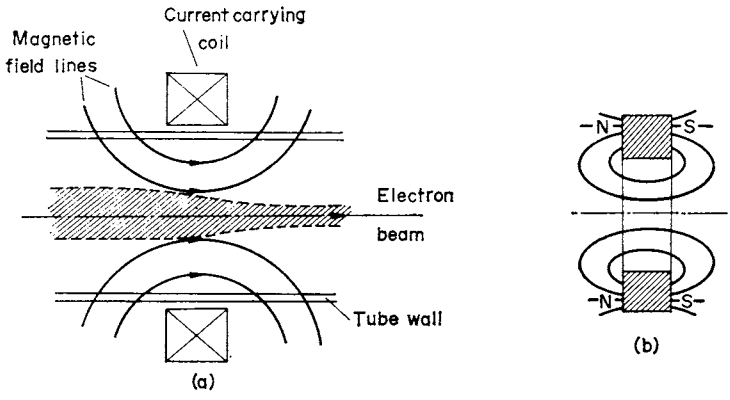


FIG. 11.4. (a) Magnetic focusing scheme. (b) Magnetic field of annular magnet magnetized axially.

this should fail. The cathode-ray tube of a television receiver may be viewed for many hours per week and its lifetime may therefore be relatively short. It is then cheaper to use magnetic focusing which can be transferred from tube to tube. Oscilloscopes invariably use electrostatic focusing.

11.3. DEFLECTING THE ELECTRON BEAM

Electrostatic and magnetic deflection systems differ in that the former is a high impedance method and requires a voltage but no current to set up the electric field, whilst the latter is a low impedance method requiring a current at low voltage to set up the magnetic field. An electric field perpendicular to the electron beam can be produced by placing a pair of electrodes one on each side of the beam, and applying a voltage between them. The beam is then attracted towards the positive electrode as it moves through the deflection system. A magnetic field can be produced by placing a pair of coils one on each side of the beam and passing a current through them. The force on the beam in this case is perpendicular to both magnetic field and beam directions. The transverse acceleration resulting from the deflection force will cause the electrons to have a transverse velocity when they leave the deflection system. Electrons will then continue with this trans-

verse velocity component in addition to the axial velocity until they strike the detecting screen, if no further force is exerted on them.

Formulae for the beam deflection at the screen can be derived as follows, using the fundamental equations discussed in Section 2.4.

11.3.1. *Electrostatic Deflection*

The relevant dimensions are given in Fig. 11.5. The electrons from the electron gun have been accelerated through a potential difference V_o and have a resultant velocity v_o along the axis as they enter the deflection system. Since no axial electric fields exist, the axial velocity is assumed to remain

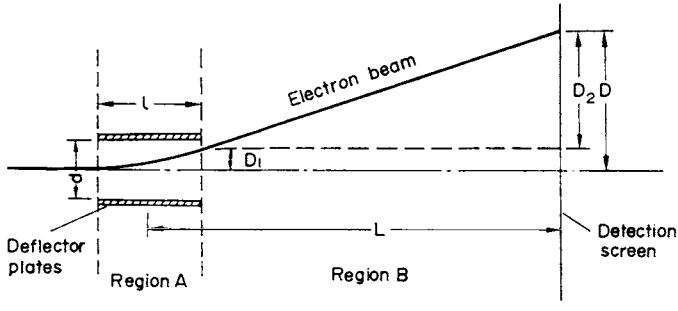


FIG. 11.5. Electrostatic deflection system.

constant through the whole region. A potential difference V_D exists between the deflector plates to deflect the beam. These deflector plates are parallel, and fringing effects are not taken into account, i.e. it is assumed that there is a transverse electric field $E = V_D/d$ between the plates for a distance l, and that no electric field exists outside this region. In practice the electric field lines will bulge out at the ends of the plates and the actual deflection region length will be greater than l. Thus the calculated deflection will be a little less than that obtained experimentally. The deflection calculation can be divided into two parts. Region A in Fig. 11.5 is one of constant acceleration of the electrons resulting in a transverse deflection D_1, and region B is one of constant transverse velocity resulting in a transverse deflection D_2. The total deflection D is then the sum of D_1 and D_2.

In region A: The transverse acceleration $a = -eE/m = eV_D/md$. The transit time for the electron to move through region A is given by $\tau_1 = l/v_o$.

The relationship between the electron velocity and the potential through which it has moved, equation (2.8), is $v_o = \sqrt{(m/2eV_o)}$ and hence $\tau_1 = l\sqrt{(m/2eV_o)}$.

Therefore deflection $D_1 = \frac{1}{2}a\tau_1^2 = \dfrac{V_D l^2}{4dV_o}$.

In region B: The transverse velocity $= a\tau_1$.

The transit time for the electron to move through region B is given by $\tau_2 = (L - \frac{1}{2}l)/v_o$.

The deflection D_2 is then given by the product of the transit time and transverse velocity in region B,

i.e. $$D_2 = a\tau_1\tau_2 = \frac{V_D l}{2V_o d}\left(L - \frac{1}{2}l\right).$$

Addition of D_1 and D_2 results in $$D = \frac{V_D L l}{V_o 2d}. \qquad (11.1)$$

In making use of this equation the units used must of course be consistent, i.e. the units of deflection will be the same as those used for the dimensions of the tube. As an example, if V_o is 2500 V, V_D is 100 V, L is 0·25 m, d is 10^{-2} m, and l is 4×10^{-2} m, the resultant deflection of the beam at the detecting screen is 2×10^{-2} m from its undeflected position.

It should be noted that the deflection is proportional to the deflection voltage and the device is therefore a linear one. The deflectional sensitivity, defined as the deflection at the screen per unit deflection voltage, is inversely proportional to the beam voltage and proportional to the ratio of length to separation of the deflector plates. Thus an increase of beam voltage to give

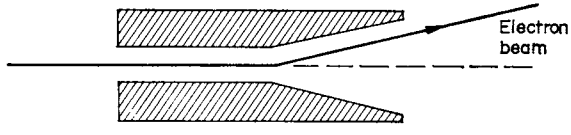


FIG. 11.6. Divergent deflector plates.

more brightness at the screen results in a reduced deflectional sensitivity. There is a limit to the improved sensitivity resulting from increasing the ratio l/d when the beam strikes the end of the deflector plates. Some further improvement can be achieved by the use of plates with divergent ends as in Fig. 11.6, and such plates are often used.

11.3.2. *Magnetic Deflection*

It was shown in Section 2.4.2 that the force exerted on an electron moving in a direction perpendicular to an imposed magnetic field is given by the product of electron charge $-e$, electron velocity v_o, and magnetic flux density B. Comparing this product $-ev_oB$ with the product of electron charge and electric field $-eE$ relevant to the electrostatic case, it can be seen that the electric field is simply replaced by the product v_oB in order to determine the corresponding deflection in a magnetic field. Thus if the magnetic field is uniform over a distance l with no magnetic flux outside this region, the beam deflection calculation has exactly the same form as the electrostatic calculation with E replaced by v_oB. Making this replacement in the final equation, and expressing v_o in terms of V_o as before, the expression for the deflection of the beam at the detecting screen for the magnetic field case becomes

$$D = \frac{lLB}{\sqrt{V_o}} \sqrt{\frac{e}{2m}}. \tag{11.2}$$

It must be remembered that the transverse deflection of the electron beam in the magnetic case is in the direction perpendicular to the magnetic field lines.

The deflection equation (11.2) derived above is an approximation since it assumes that the velocity of the electron remains axial as far as the calculation of transverse force is concerned. This approximation is good provided that the deflection is small. A more general form of the deflection equation can be obtained by using the fact that the electron will move on an arc of a circle when it enters the magnetic field. This was considered in Section 2.4.2 where it was shown that the radius of this circle is given by $r = mv/eB$. After leaving the magnetic field the electron will travel on to

the screen in a straight line as in Fig. 11.7. The angle θ is equal to arc AB divided by r, and therefore very closely equal to l/r. The deflection D is then given by $D = L \tan \theta$.

For the case where the deflection is small, $\tan \theta \doteq \theta$, and hence

$$D = \frac{lL}{r} = \frac{lLBe}{mv} = \frac{lLB}{\sqrt{V_o}} \sqrt{\frac{e}{2m}}$$

as before.

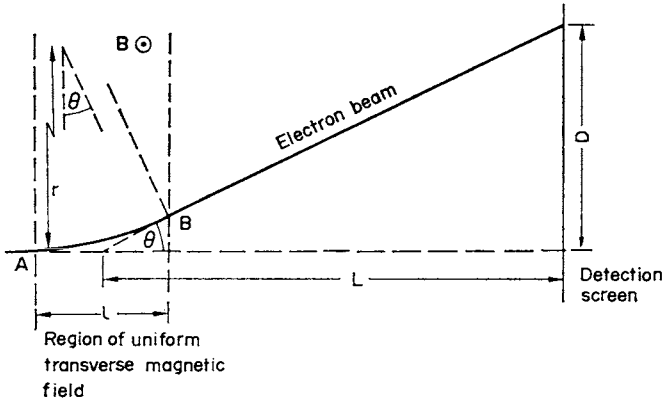


FIG. 11.7. Magnetic deflection system.

The errors introduced by the neglect of fringing of the field are much greater in the case of magnetic fields than in the electrostatic case. Since the coil producing the field is outside the cathode-ray tube the magnetic fields cannot be confined so easily as in the electric field case. The accuracy becomes greater if the product lB is replaced by the integral of B along the whole interaction region. This integral quantity can be measured experimentally by a flux measuring apparatus. For an electron beam of voltage $V_o = 2500$ volts, a distance L of 0·25 m, an interaction length l of 4×10^{-2} m, and a magnetic flux density of 10^{-3} Tesla (or Weber per square metre) uniform over the region, a total deflection of 6×10^{-2} m is obtained at the detecting screen.

The beam deflection is proportional to the flux density, and if air-cored

coils are used to produce the magnetic flux the deflection is proportional to the current in the coils. This system is therefore also a linear one. The tube sensitivity—in this case the deflection per unit coil current—is inversely proportional to the square root of the beam voltage and the tube can be used at higher voltages than in the electrostatic case without such a serious decrease in sensitivity.

An electrostatic deflection system has the advantage of not requiring current from the deflection signal source. The input impedance, defined as the input voltage to the deflection plates divided by the input current is typically many megohms. This enables the deflection plates to be connected between various points in a high impedance electric circuit without affecting the circuit itself. This is a very important feature for normal oscilloscope work. The electric field deflection scheme can also be used at very high frequencies, e.g. at many MHz. The main application of the magnetic deflection scheme is in television receivers. Here the frequency of the incoming signal is relatively low and a high power deflection signal source is available. Television tubes use a high-voltage beam, 20 kV for example. A magnetic field system gives a reasonable deflection sensitivity, and has the additional merit that it is not discarded with the tube.

11.4. THE DETECTING SCREEN

The function of the detecting screen is to give a visible indication of the instantaneous position of the electron beam. There are many compounds, known as phosphors, which will emit visible radiation when bombarded by electrons. These phosphors absorb some of the kinetic energy of the incident electrons and re-emit it as electromagnetic radiation, most of which can be in the visible part of the spectrum. Fluorescence and phosphorescence are the terms used to denote the emission of light during the actual electron bombardment and after the bombardment has ceased, respectively. The term luminescence covers both these phenomena. When a suitable detecting screen phosphor is to be chosen from the many available, to fit in with a particular requirement, there are several characteristics that must be considered. The most important of these are colour, sensitivity, persistence, and stability and life when subjected to electron bombardment.

A wide range of colours is available. For oscilloscope work it is usually required to produce the maximum effect at the eye, i.e. for the screen to be as bright as possible visually. Zinc orthosilicate, which emits mainly in the green part of the spectrum, is widely used because of its high light output. A blue emitting phosphor such as calcium tungstate is often preferable if photographic recording of the oscilloscope trace is required, since photographic films are more sensitive to radiation at the blue end of the spectrum. The white screen for monochrome television receiver cathode-ray tubes are a combination of many different phosphors, each covering a part of the visible spectrum. Screens for colour television use a mosaic of three phosphors to give the three colours from which the whole picture is produced.

Persistence, or "after glow" as it is sometimes called, denotes the degree of phosphorescence associated with the screen, i.e. the amount of light emitted after the bombardment by electrons has ceased. A long persistence tube will continue to give out light from those parts previously bombarded many seconds after the electron beam has been removed from the screen. A short persistence is normally required to avoid interference and multiple displays between successive pictures or traces. In applications like radar and the study of single pulses a long persistence screen is required in order to hold the display on the screen whilst it is being studied.

Very careful control is needed in the preparation of phosphor materials since the addition of a very small quantity of impurity, such as one part in a million, may be sufficient to markedly change the characteristics of the screen. A very thin coating of the phosphor is applied to the inner face of the cathode-ray tube. This is often backed by an exceedingly thin film of aluminium deposited on the phosphor. This serves to reflect emitted light in the forward direction and the tube is said to be "aluminized".

11.5. APPLICATIONS OF THE CATHODE-RAY TUBE

The cathode-ray tube is basically a piece of apparatus in which a light spot on a screen can be deflected a distance which is proportional to the amplitude of an electric signal applied to the tube. The mass and therefore the inertia

of the electron is so small that the system can respond to very rapid changes. There are a vast number of ways in which the tube has been used, often with great ingenuity, and only a few of the more important examples can be given here. The application with the greatest social impact has, of course, been television.

11.5.1. *The Cathode-ray Oscilloscope*

It is invariably the electrostatic deflection tube that is used in the cathode-ray oscilloscope. The spot deflection is then proportional to the instantaneous voltage applied between the deflector plates. An obvious application is to the measurement of voltage. The deflectional sensitivity of the tube can be measured experimentally. If a d.c. voltage is applied to the plates the spot position will be moved, and the magnitude of the voltage can be obtained by measuring this deflection. If an alternating voltage is applied the spot position will oscillate along a straight line whose length is proportional to the maximum variation of voltage, i.e. it is proportional to twice the peak voltage, or to the peak to peak voltage as it is usually called. The r.m.s. voltage can then be obtained from the relationship $V_{\text{peak}} = \sqrt{2}V_{\text{rms}}$. As the oscilloscope is bulky, inconvenient and rather inaccurate (± 5 per cent) as a voltmeter, it is only used in this way if it is the only device available, or when its advantages of high input impedance and high operating frequency are required.

The main applications of the oscilloscope are in the display of voltages which are time varying, and in the display of the relationship between two associated parameters.

In order to display a time-varying voltage on an x–y coordinate system on the screen, the y deflection must be made proportional to the voltage and the x deflection be made proportional to time. This latter can be accomplished by the use of a saw-tooth waveform voltage of the form shown in Fig. 11.8 (b). The voltage and therefore the x deflection increases linearly with time, and the light spot moves with constant velocity across the screen. After a time τ_x the voltage drops rapidly to zero and the spot returns almost instantaneously to the origin of the coordinate system, thereafter continually retracing it spath across the screen. This saw-tooth voltage is

called the time-base of the oscilloscope. If an unknown voltage waveform is connected to the y deflection plates when the time-base is in operation and the x deflection is proportional to time, the variation of the unknown volt-

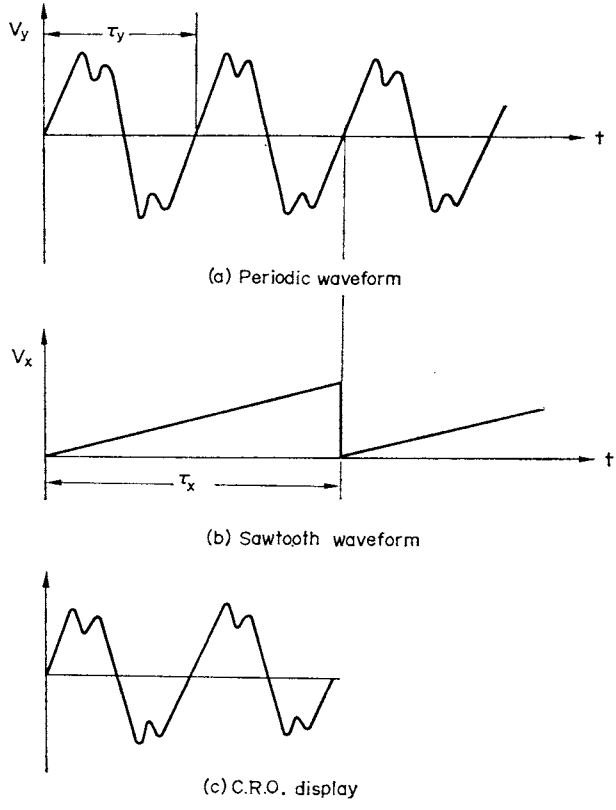


FIG. 11.8. Display of periodic voltage. (a) Periodic waveform. (b) Saw-tooth waveform. (c) C.R.O. display.

age with time will be displayed. Periodic voltage waveforms, such as that shown in Fig. 11.8 (a), are particularly suitable for display. The period τ_x is then made a simple multiple of the period of the waveform, i.e. $\tau_x = n\tau_y$, and the light spot will return to zero at just the right time to enable the waveform to be continually retraced and a stationary display obtained. This

condition is known as synchronization, and there are then n-cycles of the unknown waveform on the screen. There are some requirements in which a single transient, such as a single lightning stroke current, is being investigated. A single tooth of the saw-tooth waveform is then used with facilities in the associated electronic apparatus to start the time-base at the instant it is required. If the signals being investigated are too small in amplitude to give a significant deflection of the spot, they must be amplified before being applied to the deflector plates. Oscilloscopes normally have built-in amplifiers for this purpose.

The relationship between two interdependent variables, such as the voltage across and the current through an electric circuit element, the stress and strain of a mechanical specimen, or the magnetic field strength and flux density in a magnetic material, can be displayed directly. Voltages proportional to the independent and dependent variables are connected to the x and y deflector plates respectively. The light spot then takes up a coordinate position on the screen corresponding to the values of the parameters being investigated. If the amplitude of the independent variable is then changed the spot will move to a new position whose coordinates correspond to the new values of the parameters. A continuous line showing the relationship between the parameters is obtained if the independent variable is made to continuously change in amplitude. A sinusoidal variation is often very convenient, using the normal a.c. power supply to give variations at 50 Hz.

Time intervals can be measured by making use of a calibrated time-base. If electrical signals are produced at the instants that the events to be recorded take place, they can be used to give vertical marker deflections as the light spot moves across the screen. The horizontal distance between these markers can be measured and the corresponding time interval obtained from the calibration. Radar is an example of such a measurement. A marker is made when a short pulse of electromagnetic waves is sent out from the aerial, and a second marker added when a wave is received back at the aerial after reflection by a distant object. The distance between the markers is a measure of the distance away of the reflecting object. Most radar systems use a much more sophisticated method of display than this, of course. The frequency of a periodic waveform can be measured if the time-base frequency is calibrated.

11.5.2. *Application to Television*

A detailed explanation of the way in which a picture is produced on a television screen is outside the scope of this book, but the basic concept is easily visualized. Saw-tooth waveform voltages are generated to give a time-base for both vertical and horizontal deflections. The period of the

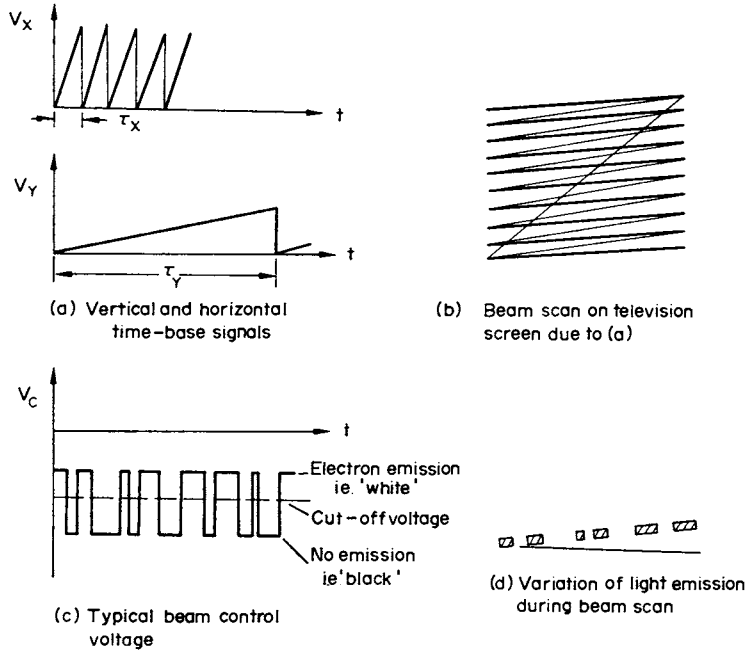


FIG. 11.9. Formation of picture mosaic on a television screen. (a) Vertical and horizontal time-base signals. (b) Beam scan on television screen due to (a). (c) Typical beam control voltage. (d) Variation of light emission during beam scan.

vertical deflection time-base is made many hundred times greater than that of the horizontal time-base, i.e. in Fig. 11.9 (a), $\tau_y \gg \tau_x$. The result is that the electron beam scans a rectangular area as in Fig. 11.9 (b). The intensity of the electron beam, and therefore the intensity of the light spot produced by it, depends upon the voltage of the control electrode in the electron gun.

This voltage is normally held sufficiently negative relative to the cathode to ensure that very few electrons leave the cathode and that there is subsequently little light emitted by the screen. As the beam moves across the screen, the voltage of the control electrode is changed as in Fig. 11.9 (c) by a signal received at the aerial from the television transmitter. If the point on the screen being scanned is required to be white the control electrode is made less negative, more electrons leave the cathode, and more light is emitted at the screen. In this way a mosaic of white and black dots is formed on the screen, giving the appearance of a continuous picture if viewed at a distance from the screen.

In a colour television receiver three pictures are produced in a manner similar to that for a monochrome receiver, in three primary colours, and superimposed. The three colours chosen are red, green and blue, from which any other colour (including white) is obtained by addition, with the appropriate intensity of each. Several schemes enable this to be done. For example,

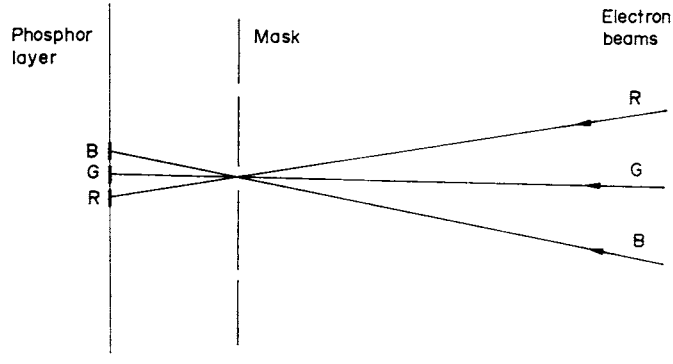


FIG. 11.10. Use of a shadow mask to produce a colour picture.

a screen can be made by laying down a mosaic of small dots of the red, green and blue phosphor, accurately positioned, and then using three electron beams each of which is allowed to fall on dots of one colour only. Each electron beam then produces a picture in one of the three primary colours. This can be achieved by the use of a shadow mask, as shown in Fig. 11.10. The mask is a sheet perforated with small holes through which

the electron beam may pass. The separation of the electron guns gives a different angle of incidence for each electron beam, which is then enabled to fall on the mosaic dot of the corresponding colour. Since the intensity of the separate colours can be independently controlled, and the dot size is very small, the overall picture seen from a distance is one of normal and continuous gradation of colour and intensity.

APPENDIX 1

Approximate Values of some Physical Constants

Electronic charge e	$-1{\cdot}602\times10^{-19}$ coulomb
Electronic mass m	$9{\cdot}108\times10^{-31}$ kilograms
Boltzmann's constant k	$1{\cdot}380\times10^{-23}$ joule degree^{-1}
Planck's constant h	$6{\cdot}624\times10^{-34}$ joule second
Proton rest mass	$1{\cdot}672\times10^{-27}$ kilograms
Permittivity of free space ε_0	$8{\cdot}854\times10^{-12}$ farad metre^{-1}
Permeability of free space μ_0	$1{\cdot}257\times10^{-6}$ henry metre^{-1}
Velocity of light	$2{\cdot}998\times10^{8}$ metres sec^{-1}
Avogadro's number	$6{\cdot}0254\times10^{23}$ per gram molecule
Loschmidt's number	$2{\cdot}687\times10^{25}$ metre^{-3}

(Loschmidt's number is the number of molecules in unit volume of any gas at N.T.P.)

APPENDIX 2

References for Further Reading

ALLISON, J. *Electronic Engineering Materials and Devices* (McGraw-Hill).

GRAY, P. E. and SEARLE, C. L. *Electronic Principles, Physical Models and Circuits* (Wiley).

GROVE, A. S. *Physics and Technology of Semiconductor Devices* (Wiley).

LINDMAYER, J. and WRIGLEY, C. Y. *Fundamentals of Semiconductor Devices* (Van Nostrand).

PRIDHAM, G. J. *Electronic Devices and Circuits* (Pergamon).

SPARKES, J. J. *Junction Transistors* (Pergamon).

SPARKES, J. J. *Transistor Switching and Sequential Circuits* (Pergamon).

APPENDIX 3A

Problems

Chapter 2

2.1. A tungsten wire cathode 1 cm in length is to be operated at 2650 K and provide an emission current of 100 mA. Assuming that the coefficient A in the Richardson–Dushman equation is 6×10^5 A/m^2/K^2, calculate the wire diameter required (ϕ for tungsten is 4·5 V.)

2.2. A tungsten filament cathode, which is temperature limited, is operated at 2650 K. If the temperature is increased by 50 K, by what percentage does the emission current increase?

(Simplify the expression $\frac{\text{final emission current}}{\text{initial emission current}}$ before inserting the operating temperature given.)

2.3. An electron is emitted with an insignificant initial velocity from one of a pair of parallel plane electrodes in an evacuated region. The emitting electrode is at a potential of -1500 V relative to the other plane electrode. The separation between the electrodes is 2 cm.
(a) How long does it take the electron to reach a speed of 10^7 m/sec?
(b) How far does the electron travel before reaching this speed?
(c) With what velocity does the electron hit the far electrode?
(d) How much energy does the electron give the far electrode?

2.4. An electron is injected perpendicularly through a hole in one of a pair of parallel plane electrodes 1 cm apart, with a velocity of 5×10^6 m/sec. The other electrode is at a potential of -100 V relative to the electrode through which the electron is injected. What will happen to the injected electron?

2.5. The potential difference between the electrodes of problem 2.4 is reduced to zero, and a magnetic field is imposed between and parallel to the electrodes. What will be the electron trajectory for a flux density of (a) 10^{-2} T, (b) 3×10^{-3} T, and (c) 10^{-3} T, for the same electron injection velocity as before?

2.6. A plane sheet of electrons, of charge density ϱ and thickness d, occupies the region midway between two parallel plane electrodes, a distance D apart. Both electrodes are at the same potential. Using Gauss' law, derive expressions for the electric field between the sheet of electrons and the electrodes, and for the potential of the outer surface of the sheet of electrons relative to the electrodes.

Chapter 3

3.1. The allowed orbit radii and energy levels of the Bohr atom are given by equations (3.5) and (3.7) respectively. Using the values of physical constants given in Appendix 1, show that these equations may be written as $(0{\cdot}529\ n^2)/Z$ angstrom units and $(-13{\cdot}6\ Z^2)/n^2$ electron volts.

3.2. An electron in a hydrogen atom makes a transition from a quantum state with principal quantum number 2 to the ground state ($n = 1$). Calculate the energy released by this electron and also the frequency of the radiation from the atom when this transition occurs.

3.3. Calculate the velocity of an electron in the ground state of hydrogen. At what fraction of the velocity of light is the electron travelling?

3.4. A shell of principal quantum number n can contain n subshells. A subshell can contain $2(2l+1)$ electrons where l is a positive integer or zero, but must never be greater than $(n-1)$. Using this information, show that a shell cannot contain more than $2n^2$ electrons.

3.5. Use formula 3.8 to calculate the intrinsic electron and hole densities in germanium on a very cold day ($T = 273$ K) and a very hot day ($T = 310$ K). Show that the increase in density due to this temperature variation is almost sevenfold. The energy gap for germanium is 0·72 eV.

3.6. A silver wire has a resistivity of $1{\cdot}5\times10^{-8}$ ohm metres at a specified temperature. If the electric field along the wire is 100 V/m, compute the average drift velocity of electrons, assuming there are 10^{29} free conduction electrons per cubic metre. Calculate also the mobility and mean free time between collisions.

3.7. The resistivity of intrinsic germanium at 300 K is 0·47 ohm metre. The electron and hole mobilities in germanium are 0·36 m²/volt sec and 0·17 m²/volt sec respectively. Calculate from this information the intrinsic density of electrons and holes.

3.8. The material used in question 3.7 is doped with antinomy impurity atoms so that there is one impurity atom per 10^6 germanium atoms. Calculate the electron and hole densities at 300 K. It may be assumed that all antimony atoms are ionized at this temperature. The density of germanium atoms is $4{\cdot}4\times10^{28}$ per m³. What is the resistivity of this doped material?

3.9. From the information given in question 3.7 calculate the diffusion coefficients both for holes and electrons in germanium at 300 K.

Chapter 4

4.1. Using equation (4.1), show that the contact potential V_{pn} developed across a $p-n$ junction is given by:

$$V_{pn} = \frac{kT}{e}\log_e\frac{p_p}{p_n} = \frac{kT}{e}\log_e\frac{n_n}{n_p}$$

where p_p and n_p are the hole and electron densities in the p material and p_n and n_n are the hole and electron densities in the n material.

Hence show that:

$$p_p n_p = p_n n_n.$$

Why must this be so?

4.2. A $p-n$ junction consists of a region of p-type germanium with a conductivity of 10^4 S and a region of n-type germanium with a conductivity of 100 S. The mobilities of electrons and holes in germanium are 0·36 m²/volt sec and 0·17 m²/volt sec respectively. The intrinsic density of holes and electrons in germanium at 300 K is $2{\cdot}5 \times 10^{19}$ m^{-3}. Calculate the contact potential that would appear across the junction at 300 K.

4.3. The following, rather oversimplified, argument may be used to calculate the saturation current density in a $p-n$ junction.

The reverse current in a $p-n$ junction saturates when the minority carriers are extracted through the depletion layer at the maximum possible rate. This is equal to the production rate of electron-hole pairs by the intrinsic process. The maximum possible rate, however, is only realized if there is no chance for minority carriers to recombine. Such conditions can only exist however over a distance from the depletion layer of the order of a diffusion length for the minority carriers concerned.

Show, using these details, that an order of magnitude estimate of the saturation current density is

$$J_s = \frac{D_p e p_n}{L_p} + \frac{D_n e n_p}{L_n} .$$

4 4. The diffusion length for both electrons and holes in germanium is $\frac{1}{10}$ cm. Using this information and that given in problem 4.2, calculate the saturation current density at 300 K. Determine also the ratio of the saturation hole current to electron current.

4.5. Find the voltage that would have to be applied across the $p-n$ junction described in question 4.4 in order to cause a forward current density of 10^5 A/m² to flow.

If both the p- and n-regions of the diode are 3 mm long, calculate the total voltage drop across the terminals of the diode in the above instance.

4.6. A specimen of n-type material has a majority electron carrier density of n_n and a minority hole carrier density of p_n under equilibrium conditions. Suppose that at some instant of time $t = 0$ the minority carrier density throughout the specimen momentarily increased to the value p_{no}. Show that the hole density p would decay back to the original equilibrium value p_n according to the law:

$$p = p_n - (p_n - p_n) \exp(-t/\tau_p)$$

where $\tau_p = L_p^2/D_p$.

This example shows that any small fluctuations in minority hole carrier density decay in a time of order τ_p. The quantity τ_p is called the lifetime of holes. When electrons are the minority carriers a similar expression for the lifetime of electrons may be derived, viz.

$$\tau_n = \frac{L_n^2}{D_n} .$$

4.7. Using values taken from questions 4.2, 4.4, and 4.6, calculate the lifetimes of holes and electrons in germanium.

Chapter 5

5.1. A transistor is connected in grounded base. Show that the emitter resistance r_e is given approximately by the relation:

$$r_e = \frac{kT}{eI_e} \quad \text{where } I_e \text{ is the emitter bias current.}$$

If the emitter bias current is 2 mA, estimate the emitter resistance at room temperature (300 K).

5.2. A transistor is connected in grounded base. The following small signal parameters apply:

$$r_b = 100\ \Omega \qquad r_c = 10^6\ \Omega$$
$$r_e = 25\ \Omega \qquad \alpha = 0{\cdot}95$$

The collector load resistance is 10,000 Ω.

Sketch the equivalent circuit and calculate the voltage gain, current gain, power gain and input resistance.

5.3. An a.c. voltage source of internal resistance 100 Ω is connected to the input of the grounded base transistor described in question 5.2. The open circuit voltage from the source is found to be 1 mV and yet the output voltage from the transistor is measured as 72 mV. Why is this value much less than to be expected from the value of voltage gain calculated in question 5.2?

5.4. The voltage source in question 5.3 is connected now to the input of the transistor via a transformer having a turns ratio of 1 : n. Find the optimum value of n in order that the output signal be as great as possible. Has the transformer a step-up or step-down turns ratio?

5.5. The transistor described in question 5.2 is now connected in grounded emitter. The collector load resistance as before is 10,000 Ω.

Calculate the voltage gain, current gain, power gain and input resistance.

5.6. An a.c. voltage source of internal resistance 100 Ω and open circuit voltage 1 mV is connected to the input of the grounded emitter transistor described in question 5.5. Calculate the output voltage appearing across the load resistance.

Chapter 6

6.1. A transistor is connected in common-emitter. The base current is I_b' and the corresponding collector current is I_c'. The transistor is not saturated. At time $t = t_1$ the base current is suddenly switched to zero. Show that the collector current varies for $t > t_1$ as

$$I_c = I_c' \exp -\frac{(t - t_1)}{\tau_p}.$$

By observing this decay on an oscilloscope, suggest how you might be able to measure the value of τ_p.

6.2. A pure voltage generator is connected in series with a parallel RC network to the base terminal of a grounded-emitter amplifier. The voltage generator suddenly increases by a step ΔV at time $t = 0$. Show that the collector current follows the step change faithfully if $RC = \tau_p$. Show also that the corresponding change ΔI_c in the collector current is given in this case by:

$$\Delta I_c = \frac{C}{\tau_F} \Delta V,$$

where C is the capacitance of the RC network. You may assume that all the applied change of voltage ΔV appears across the RC network since the forward-biased base–emitter junction has a relatively low impedance.

6.3. A transistor connected in common-emitter has charge control parameters τ_p and τ_F. Use the charge control equations to show that the magnitude of the short circuit current gain varies with angular frequency ω for a sinusoidal base input current, as:

$$h_{fe}(\omega) = \frac{\tau_p}{\tau_F} \frac{1}{\sqrt{(1+\omega^2\tau_p^2)}}.$$

Chapter 9

9.1. The anode characteristic of a diode is given by the following values of I_a and V_a:

V_a volts	0	10	20	30	40	50
I_a mA	0·2	2·2	5·6	9·7	14·7	19·7

Plot a graph of $\log_{10} I_a$ as a function of $\log_{10} V_a$, and verify that the relationship between I_a and V_a is of the form $I_a = KV_a^n$. What value for n is indicated?

9.2. The diode of problem 9.1 is connected in series with a resistor of 10 kΩ and a battery of emf 100 volts. A 30 kΩ resistor is connected between the anode and the cathode of the diode. Plot the two relationships between V_a and I_a corresponding to the valve characteristic and to the characteristic of the external circuit (using Kirchhoff's laws), and hence determine the valve anode current and voltage. (The point of intersection of the graphs is the only point where both relationships are satisfied.)

9.3. A diode valve has a characteristic given by $I_a = KV_a^{3/2}$ where K is 2×10^{-3} when I_a is in mA and V_a is in volts. The diode is connected in series with a resistor of 20 kΩ and a battery of emf E. A voltmeter of resistance 50 kΩ is connected across the diode and indicates 100 volts. What is the emf of the battery, and what will be the voltage across the diode when the voltmeter is removed?

9.4. The characteristics of a triode valve can be expressed in the form

$$I_a = 0{\cdot}01(V_a + 24V_g)^{3/2} \text{ mA}.$$

Calculate the amplification factor and the anode slope resistance of the valve for the condition $V_a = 200$ volts and $V_g = -6$ volts.

9.5. A triode valve has the following anode characteristic:

$V_g = -1$ volt					
V_a volts	50	75	100	125	150
I_a mA	1·0	2·5	4·0	6·0	8·0
$V_g = -2$ volts					
V_a volts	50	75	100	125	150
I_a mA	0	0·1	1·5	3·0	5·0

The valve is connected in series with a 20 kΩ load resistor to a d.c. supply of 200 volts. Plot the anode characteristic and the load line on the same graph and hence determine the change of anode voltage when the grid voltage is changed from -1 to -2 volts. Compare the result with that obtained using the normal expression for the voltage gain of a resistive load amplifier, in which the anode slope resistance and the amplification factor are taken from the plotted anode characteristic.

9.6. A triode valve is used as a voltage amplifier with an anode load resistor of 30 kΩ, and a gain of 25 is obtained. When the load resistance is reduced to 20 kΩ the voltage gain falls to 20. Calculate the anode slope resistance, the amplification factor, and the mutual conductance of the triode.

Chapter 10

10.1. A half-wave rectifier circuit operates from a 50 Hz supply of 240 V r.m.s. The load consists of a resistor of 10 kΩ in parallel with a capacitor of 10 μF. Neglecting the resistance of the rectifier element, calculate the mean output voltage and the magnitude of the r.m.s. ripple voltage.

10.2. The peak ripple voltage on the output of the d.c. supply of problem 10.1 is required to be reduced to less than 3 per cent of the output voltage. A further 10 μF capacitor is available. Calculate the minimum value of an inductance to be used in conjunction with the capacitor to satisfy the ripple requirement.

What would be the magnitude of the ripple if full-wave rectification were to be used in place of the half-wave rectifier circuit?

10.3. A FET with a mutual conductance of 1 mA/V and an amplification factor of 20, has a load of 10 mH inductance and negligible resistance in the drain circuit. What is the maximum possible voltage gain, and at what frequency does the voltage gain fall to one-half of this maximum value?

10.4. An alternating voltage amplifier uses a FET of mutual conductance 4 mA/V and amplification factor 40. The drain load consists of a resistor of 20 kΩ in parallel with a capacitor C. It is required that the voltage gain at 10 kHz should be 75 per cent of the maximum possible voltage gain. What value of capacitance should be used?

10.5. A triode is self-biased by means of a parallel R–C bias circuit connected in series with the cathode of the valve. The d.c. grid bias voltage is to be -4 V and the d.c. valve anode current 2 mA. The a.c. current passing through the bias resistor is to be less than 1 per cent of the a.c. current through the capacitor over the operating frequency range of 50–1000 Hz. Calculate the required value of bias resistor, and the minimum value of the parallel capacitor.

10.6. An amplifier for a voltmeter must have an overall gain of 60. Gain variations of only ± 1 per cent can be tolerated. It appears from the manufacturers' catalogues that owing to production spreads in the device characteristics, variations of ± 12 per cent in the open-loop gain may be expected. Determine the minimum value of the feedback fraction and also the open-loop gain needed to satisfy the above condition.

Chapter 11

11.1. A cathode ray tube has an electron beam that has been accelerated through a potential difference of 1000 V. The beam is then passed through a pair of parallel deflector plates 5 cm long and 1 cm apart, with a potential difference of 50 V between them. If the detecting screen is placed 20 cm from the centre of the deflector plates, calculate:
(a) the time taken for the beam to pass through the region between the deflector plates;
(b) the transverse acceleration of the electrons between the plates;
(c) the total deflection of the electron beam at the detecting screen.

11.2. Calculate the magnitude of the magnetic field which, acting over the same distance of 5 cm and replacing the electrostatic deflection system of problem 11.1, will give the same deflection at the detecting screen.

APPENDIX 3B

Solutions

Chapter 2

2.1. Current density J A/m^2 $= AT^2 \exp\left(\dfrac{-11{,}600\phi}{T}\right)$

$= 6{\cdot}0\times10^5\times(2650)^2\times\exp(-19{\cdot}7) = 1{\cdot}2\times10^4.$

Total current $= 0{\cdot}1$ amp $= \pi dJ\times10^{-2}$ for 1 cm length of cathode.

$d = 2{\cdot}7\times10^{-4}$ m.

2.2. $\dfrac{J_2}{J_1} = \dfrac{\left(T_2^2\exp-\dfrac{T_o}{T_2}\right)}{T_2^2\exp\left(-\dfrac{T_o}{T_2}\right)} = \left(\dfrac{T_2}{T_1}\right)^2\exp\left(T_o\dfrac{T_2-T_1}{T_2T_1}\right)$

$= 1{\cdot}04\exp(0{\cdot}365) = 1{\cdot}48.$

2.3. Accelerating field $E = 1500/2\times10^{-2} = 7{\cdot}5\times10^4$ V/m.
Acceleration $a = eE/m = 1{\cdot}32\times10^{16}$ m/sec^2.

(a) $\tau_1 = \dfrac{v}{a} = \dfrac{10^7}{1{\cdot}32\times10^{16}} = 7{\cdot}6\times10^{-10}$ seconds;

(b) $x = \frac{1}{2}a\tau_1^2 = \frac{1}{2}\times1{\cdot}32\times10^{16}\times(7{\cdot}6)^2\times10^{-20} = 3{\cdot}8\times10^{-3}$ m;

(c) $\frac{1}{2}mv^2 = eV$.

Final velocity $= \sqrt{\left(\dfrac{2eV}{m}\right)} = \sqrt{\left(\dfrac{2\times1{\cdot}6\times10^{-19}\times1500}{9{\cdot}11\times10^{-31}}\right)}$

$= 2{\cdot}3\times10^7$ m/sec;

(d) energy per electron $= eV = 1{\cdot}6\times10^{-19}\times1500 = 2{\cdot}4\times10^{-16}$ joule.

2.4. To come to rest, the electron needs to move through a potential

$$V = \frac{mv^2}{2e} = \frac{9{\cdot}11\times10^{-31}\times25\times10^{12}}{2\times1{\cdot}6\times10^{-19}} = 71\text{ V}.$$

Thus the electron cannot reach the far electrode and returns to the electrode through which it was injected.

2.5. Radius of gyration $r = \dfrac{mv}{eB}$

(a) when $B = 10^{-2}$ T, $r = \dfrac{9{\cdot}11\times10^{-31}\times5\times10^{6}}{1{\cdot}6\times10^{-19}\times10^{-2}} = 2{\cdot}9\times10^{-3}$ m

and the electron returns to injection electrode, a distance of 0·58 cm from the point of injection;

(b) when $B = 3\times10^{-3}$ T, $r = 2{\cdot}9\times10^{-3}\times\dfrac{10^{-2}}{3\times10^{-3}} = 9{\cdot}7\times10^{-3}$ m

and the electron returns a distance 1·9 cm from the injection point;

(c) when $B = 10^{-3}$ T, $r = 2{\cdot}9\times10^{-3}\times\dfrac{10^{-2}}{10^{-3}}$

$= 2{\cdot}9\times10^{-2}$ m

and the electron collides with the far electrode.

2.6. Taking a one metre square of the electron sheet,
The charge in the square $= \varrho d$.
The electric flux is perpendicular to the sheet, and emerges over an area of 2 m² (both sides of the sheet).

$$\therefore \text{ The electric flux density } D = \frac{\varrho d}{2}$$

$$\text{and the electric field } E = \frac{D}{\varepsilon_0} = \frac{\varrho d}{2\varepsilon_0}.$$

Distance between the electrodes and the outer edges of electron sheet is $\frac{1}{2}(D-d)$.

$$\therefore \text{ Potential of outer surface of sheet } = \frac{\varrho}{4\varepsilon_0} d(D-d).$$

Chapter 3

3.1. From equation 3.5

$$r_n = \frac{\varepsilon_0 h^2 n^2}{\pi m Z e^2} \text{ m}$$

$$= \frac{(8{\cdot}854\times10^{-12})\,(6{\cdot}624\times10^{-34})^2\times10^{+10}}{\pi(9{\cdot}107\times10^{-31})(1{\cdot}601\times10^{-19})^2}\left(\frac{n^2}{Z}\right) \text{ angstrom units}$$

$$= \frac{0{\cdot}529n^2}{Z} \text{ angstrom units.}$$

From equation (3.7)

$$W_n = -\frac{me^4Z^2}{8\varepsilon_0^2 h^2 n^2}\ \text{J}$$

$$= -\frac{(9{\cdot}107\times10^{-31})\,(1{\cdot}601\times10^{-19})^4\,Z^2}{8(8{\cdot}854\times10^{-12})^2(6{\cdot}624\times10^{-34})^2\,n^2}$$

$$= -\frac{21{\cdot}8\times10^{-19}\,Z^2}{n^2}\ \text{J}$$

$$= -\frac{21{\cdot}8\times10^{-19}\,Z^2}{1{\cdot}601\times10^{-19}\,n^2}\ \text{eV}$$

$$= -\frac{13{\cdot}6\,Z^2}{n^2}\ \text{eV}.$$

3.2. For hydrogen $Z = 1$,

$$n = 2, \quad W_2 = -\frac{13{\cdot}6}{(2)^2}\ \text{eV} = -3{\cdot}4\ \text{eV},$$

$$n = 1, \quad W_1 = -13{\cdot}6\ \text{eV}.$$

$\therefore$ Difference in energy $= 13{\cdot}6 - 3{\cdot}4 = 10{\cdot}2$ eV.

If f is the frequency of the radiation

$$hf = (10{\cdot}2)\times(1{\cdot}601\times10^{-19})\ \text{J}.$$

$$\therefore f = \frac{10{\cdot}2\times1{\cdot}601\times10^{-19}}{6{\cdot}624\times10^{-34}}$$

$$= 2{\cdot}47\times10^{15}\ \text{Hz}.$$

3.3. Bohr postulate:

$$\text{Angular momentum} = \frac{nh}{2\pi}.$$

$$\text{For ground state } mvr = \frac{h}{2\pi}.$$

From problem (3.1) $r = 5{\cdot}29\times10^{-11}$ m.

$$\therefore v = \frac{(6{\cdot}624\times10^{-34})}{2\pi(5{\cdot}29\times10^{-11})\,(9{\cdot}107\times10^{-31})}$$

$$= 2{\cdot}19\times10^{6}\ \text{m/sec}.$$

$$\frac{v}{c} = \frac{2{\cdot}19\times10^{6}}{3\times10^{8}} = 7{\cdot}3\times10^{-3}.$$

3.4. The subshell $l = 0$ (s subshell) contains two electrons, the $l = 1$ (p subshell) subshell contains $2(2+1) = 6$ electrons and so on. Thus in a shell of principal quantum number n with n subshells there are:

$$2+6+\ldots+2(2l+1)\ \text{electrons}$$

where l extends to $(n-1)$.

$$\therefore \text{ Total number of electrons} = 2\sum_{l=0}^{n-1}(2l+1).$$

This series is an arithmetic progression. The sum of an A.P. is given by:

$$(\text{No. of terms}) \times (\text{Average term}) = \frac{2n[1+2(n-1)+1]}{2}$$
$$= 2n^2.$$

3.5. The intrinsic density is

$$N = 5\times 10^{15} T^{3/2} \exp \frac{-Wg}{2kT} \text{ (cm)}^{-3}.$$

At 273 K
$$\frac{Wg}{2kT} = \frac{0{\cdot}72\times 1{\cdot}601\times 10^{-19}}{2\times 273\times 1{\cdot}38\times 10^{-23}}$$
$$= 15{\cdot}3.$$
$$\therefore N = 5\times 10^{15}\times (273)^{3/2} \exp(-15{\cdot}3)$$
$$= 5\times 10^{12} \text{ (cm)}^{-3}.$$

At 310 K
$$\frac{Wg}{2kT} = 13{\cdot}4.$$
$$\therefore N = 5\times 10^{15}\times (310)^{3/2} \exp(-13{\cdot}4)$$
$$= 42\times 10^{12} \text{ (cm)}^{-3}.$$
$$\therefore \% \text{ increase} = \frac{(42-5)}{5}\times 100 = 740 \text{ per cent.}$$

3.6. From equations (3.9) and (3.12)

$$v_D = \frac{\sigma E}{Ne}$$
$$= \frac{100}{(1{\cdot}5\times 10^{-8})\times 10^{29}\times (1{\cdot}6\times 10^{-19})}$$
$$= 0{\cdot}42 \text{ m/sec.}$$

Also $v_D = \mu E.$

$$\therefore \mu = \frac{0{\cdot}42}{100} = 4{\cdot}2\times 10^{-3} \text{ m}^2/\text{V sec},$$
$$\tau = \frac{2m\mu}{e} = \frac{2(4{\cdot}2\times 10^{-3})\times (9{\cdot}1\times 10^{-31})}{1{\cdot}6\times 10^{-19}}$$
$$= 4{\cdot}8\times 10^{-14} \text{ sec.}$$

3.7.
$$\sigma = Ne(\mu_e + \mu_p).$$
$$\therefore N = \frac{1}{(0{\cdot}47)\times (1{\cdot}6\times 10^{-19})(0{\cdot}36+0{\cdot}17)} \text{ (m)}^{-3}$$
$$= 2{\cdot}5\times 10^{19} \text{ electrons or holes per m}^3.$$

3.8. Density of donor impurities $= (4{\cdot}4\times10^{28})\times10^{-6} = 4{\cdot}4\times10^{22}\ (\mathrm{m})^{-3}$.
From problem 3.7 the intrinsic density

$$= 2{\cdot}5\times10^{19}.$$

$$\therefore \text{ Hole density} = \frac{(2{\cdot}5)^2\times10^{38}}{4{\cdot}4\times10^{22}}$$

$$= 1{\cdot}42\times10^{16}\ (\mathrm{m})^{-3}.$$

Thus, Electron density $= 4{\cdot}4\times10^{22}\ (\mathrm{m})^{-3}$.

Hole density $= 1{\cdot}4\times10^{16}\ (\mathrm{m})^{-3}$.

$$\text{Resistivity} \approx \frac{1}{(4{\cdot}4\times10^{22})\,(1{\cdot}6\times10^{-19})\,(0{\cdot}36)}$$

$$= 4\times10^{-4}\ \Omega\mathrm{m}.$$

3.9. The Einstein relation is

$$D = \frac{kT}{e}\,\mu.$$

For electrons

$$D_n = \frac{(1{\cdot}38\times10^{-23})\,(300)\,(0{\cdot}36)}{1{\cdot}601\times10^{-19}}$$

$$= 9{\cdot}3\times10^{-3}\ \mathrm{m^2/sec}.$$

For holes

$$D_p = \frac{(1{\cdot}38\times10^{-23})\,(300)\,(0{\cdot}17)}{1{\cdot}601\times10^{-19}}$$

$$= 4{\cdot}4\times10^{-3}\ \mathrm{m^2/sec}.$$

Chapter 4

4.1. From equation (4.1)

$$E_x\,\mathrm{d}x = \frac{kT}{e}\,\frac{\mathrm{d}p}{p} = \frac{-kT}{e}\,\frac{\mathrm{d}n}{n}.$$

Integrating this expression across the junction:

$$V_{pn} = \frac{kT}{e}\log_e\frac{p_p}{p_n} = \frac{kT}{e}\log_e\frac{n_n}{n_p}.$$

4.2. For the p material, using the formula $\sigma_p = ep_p\mu_p$ where σ_p is the conductivity of the p material and μ_p the mobility of holes:

$$p_p = \frac{10^4}{(0{\cdot}17)\,(1{\cdot}6\times10^{-19})} = 3{\cdot}68\times10^{23}(\mathrm{m})^{-3}.$$

For the n material

$$n_n = \frac{100}{(0{\cdot}36)(1{\cdot}6\times10^{-19})} = 1{\cdot}75\times10^{21}\ (\mathrm{m})^{-3}.$$

Now $p_n n_n$ = (Intrinsic density)^2.

$$\therefore\ p_n = 3{\cdot}57\times 10^{17}(\text{m})^{-3}$$

Thus the result derived in problem 4.1

$$V_{pn} = \frac{1{\cdot}38\times 10^{-23}\times 300}{1{\cdot}6\times 10^{-19}}\log_e\frac{3{\cdot}68\times 10^{23}}{3{\cdot}57\times 10^{17}}$$
$$= 0{\cdot}35\ \text{V}.$$

4.3. The number of hole electron pairs produced per unit volume per second by the intrinsic process = g and

$$g = rn_i p_i = rn_n p_n = rn_p p_p. \tag{a}$$

Consider firstly the p-region near the depletion layer. The minority carriers here are electrons and if they can be swept out of a region of length L_n without chance of recombining, then the number crossing unit area of the depletion layer per second and moving into the n-region is of order:

$$L_n g = Lrn_n p_p. \qquad \text{from (a)}$$

Thus the saturation electron current density crossing the depletion layer is of order

$$eL_n rn_p p_p. \tag{b}$$

But by definition (equation (3.37))

$$L_n^2 = \frac{D_n}{rp_p}. \tag{c}$$

From (c) and (b)

$$\text{Saturation electron current density} = \frac{en_p D_n}{L_n}.$$

A similar argument shows that the saturation hole current density

$$= \frac{ep_n D_p}{L_p},$$

$$\therefore\ \text{Total saturation current} = \frac{D_p e p_n}{L_p} + \frac{D_n e n_p}{L_n}.$$

4.4.
$$J_s = \frac{D_p e p_n}{L_p} + \frac{D_n e n_p}{L_n}.$$

From problem 4.2

$$p_n = 3{\cdot}57\times 10^{17}\ (\text{m})^{-3},$$
$$n_p = 1{\cdot}7\times 10^{15}\ (\text{m})^{-3}.$$

The diffusion coefficients are given by

$$D_p = \frac{kT}{e}\mu_p \quad \text{and} \quad D_n = \frac{kT}{e}\mu_n.$$

$$\therefore J_s = kT\frac{p_n\mu_p}{L_p} + \frac{n_n\mu_p}{L_n}$$

$$= (1{\cdot}38\times10^{-23})(300)\left[\frac{(3{\cdot}57\times10^{17})(0{\cdot}17)+(1{\cdot}7\times10^{15})(0{\cdot}36)}{1/1000}\right]$$

$$= 250 \text{ mA/m}^2.$$

$$\frac{\text{Hole saturation current}}{\text{Electron saturation current}} = 100.$$

4.5.

$$J = J_s\left(\exp\frac{eV}{kT} - 1\right).$$

$$\therefore \exp\left(\frac{eV}{kT}\right) - 1 = \frac{10}{25\times10^{-6}}$$

$$= 4\times10^5.$$

$$\therefore \frac{eV}{kT} = \log_e 4\times10 = 12{\cdot}9.$$

$$\therefore V = \frac{(12{\cdot}9)(1{\cdot}38\times10^{-23})\times300}{1{\cdot}6\times10^{-19}}$$

$$= 0{\cdot}33 \text{ V}.$$

Let cross-sectional area of diode = A m^2.
Total resistance of p and n regions

$$\frac{0{\cdot}3}{A}\left(\frac{1}{100}+\frac{1}{1}\right) = \frac{0{\cdot}3}{A}\,\Omega.$$

Current flowing = $10A$ A.
$\therefore$ Voltage drop due to ohmic resistance of p and n material

$$= 10A\times\frac{0{\cdot}3}{A} = 3 \text{ V}.$$

$\therefore$ Total voltage drop across the diode = $3+0{\cdot}33 = 3{\cdot}33$ V.

Note from this example that at large current densities the diode appears almost as a pure ohmic resistance. This resistance arises from the bulk resistance of the p and n materials of the diode.

4.6. The intrinsic production rate for hole-electron pairs = g. The recombination rate = $rn_n p$. In equilibrium when $p = p_p$, the production and recombination rates are equal.

$$\therefore g = rn_n p_n.$$

When $p \neq p_n$ $\qquad (rn_n p - g) = \dfrac{-\mathrm{d}p}{\mathrm{d}t}$.

The solution to this is

$$p = p_n + (p_{no} - p_n) \exp(-trn_n)$$

remembering that $p = p_{no}$ at $t = 0$ and

$$p = p_n \quad \text{as} \quad t \to \infty.$$

From equation (3.37) $\qquad rn_n = \dfrac{D_p}{L_p^2}$.

4.7. For holes $\qquad \tau_n = \dfrac{L_p^2}{D_p}$ and $D_p = \dfrac{kT}{e}\mu_p$.

$$\therefore \tau_p = \frac{L_p^2 e}{kT\mu_p} = \frac{(1/10)^2 \times (1{\cdot}6 \times 10^{-19})}{(1{\cdot}38 \times 10^{-23})(300)\,1700} = 228\ \mu\text{sec}.$$

$$\tau_n = \frac{L_n^2 e}{kT\mu_n} = \frac{(1/10)^2 \times (1{\cdot}6 \times 10^{-19})}{(1{\cdot}38 \times 10^{-23})(300)\,3600} = 108\ \mu\text{sec}.$$

Chapter 5

5.1. The emitter–base circuit is a forward biased junction. The volt-ampere law for such a junction is:

$$I_e = I_s\left(\exp\frac{eV_{eb}}{kT} - 1\right) \qquad \text{(a)}$$

where I_s is the saturation current.

Now $\qquad r_e = \left|\dfrac{\delta V_{eb}}{\delta I_e}\right|_{V_{cb}\ \text{constant}}$

$$\approx \frac{kT}{e(I_e + I_s)}. \qquad \text{from (a)}$$

But $I_e \gg I_s$. $\qquad \therefore\ r_e \approx \dfrac{kT}{eI_e}$.

Putting $I_e = 2 \times 10^{-3}$ A, $\quad T = 300$ K,

$$r_e \approx \frac{(1{\cdot}38 \times 10^{-23})\,300}{(1{\cdot}6 \times 10^{-19}) \times 2 \times 10^{-3}} = 13\ \Omega.$$

5.2. Voltage gain $= \dfrac{(0{\cdot}95)\,(10{,}000)}{25 + 100\,(0{\cdot}05)} = 317$

Current gain $= 0{\cdot}95$

Power gain $= 317 \times 0{\cdot}95 = 302$

Input resistance $= 25 + 100(0{\cdot}05)$

$= 30\ \Omega.$

5.3. The input resistance is 30 Ω.
∴ Actual voltage appearing across emitter–base terminals

$$= \frac{30}{100+13} = 0{\cdot}23 \text{ mV}.$$

∴ Output voltage = $(0{\cdot}23)\times 317 = 72$ mV.

5.4. Referred to the transistor input terminals, the source voltage is n mV and the source impedance is $100n^2$ Ω.

The voltage appearing across the emitter–base terminals $= \dfrac{30n}{100n^2+30}$.

The output voltage $= \dfrac{317\times 30\times n}{100n^2+30}$.

The maximum value of this voltage is seen by direct differentiation to occur when

$$n = \left(\frac{30}{100}\right)^{1/2} = 0{\cdot}55.$$

∴ Output voltage = 87 mV.
The transformer has a step down turns ratio.

5.5. Voltage gain $= \dfrac{(0{\cdot}95)\,(10{,}000)}{25+(0{\cdot}05)\,(100)} = 317.$

Current gain $= \dfrac{0{\cdot}95}{1-0{\cdot}95} = 19.$

Power gain $= 317\times 19 = 6040.$

Input resistance $= 100+\dfrac{25}{1-0{\cdot}95} = 600\ \Omega.$

5.6. Input resistance = 600 Ω.
∴ Voltage appearing across input terminals of transistor

$$= \frac{600}{100+600} = 0{\cdot}86 \text{ mV}.$$

∴ Output voltage = $(0{\cdot}86)\times 317 = 272$ mV.

Chapter 6

6.1. From equation (6.11), since

$$I_b = 0,$$

$$\frac{Q_F}{\tau_p}+\frac{\mathrm{d}Q_F}{\mathrm{d}t} = 0.$$

$$\therefore\ Q_F = A\exp-\left(\frac{t}{\tau_p}-t_1\right).$$

At $t = t_1$, $\quad Q_F = -I_c\tau_F$, from (6.12).

$$\therefore\ Q_F = -I_c'\tau_F \exp -\frac{(t-t_1)}{\tau_p}$$

and using (6.12) the collector current is seen to be:

$$I_c = -\frac{Q_F}{\tau_F} = +I_c' \exp -\frac{(t-t_1)}{\tau_p}\,.$$

By observing the decay τ_p can be found. Both I_c' and I_b' can be measured and $I_c'/I_b' = \tau_p/\tau_F$. Thus τ_F can be obtained.

6.2. The base current I_b is given by

$$I_b = \frac{\Delta V}{R} + C\frac{\mathrm{d}(\Delta V)}{\mathrm{d}t} \tag{1}$$

becau se of the RC network. The charge control equation (6.11) gives

$$-I_b = \frac{Q_F}{\tau_p} + \frac{\mathrm{d}Q_F}{\mathrm{d}t} \tag{2}$$

or using (6.12) in (2)

$$I_b = \tau_F\left\{\frac{I_c}{\tau_p} + \frac{\mathrm{d}I_c}{\mathrm{d}t}\right\}. \tag{3}$$

If $RC = \tau_p'$, (3) becomes

$$I_b = \frac{\tau_F}{C}\left(\frac{I_c}{R} + C\frac{\mathrm{d}I_c}{\mathrm{d}t}\right). \tag{4}$$

Equations (1) and (4) are identical if we write $\Delta V = (\tau_F/C)I_c$. Thus there is a direct one-to-one correspondence between the change in voltage ΔV and the collector current if $RC = \tau_p$.

We note now that the addition of R and C has provided an amplifier with very rapid response to changes in input voltage. The added capacitor C is called a speed-up capacitor.

6.3. Assume the base current can be written as $I_b \sin \omega t$. Then the stored charge Q_F is the solution of

$$\frac{\mathrm{d}Q_F}{\mathrm{d}t} + \frac{Q_F}{\tau_p} = -I_b \sin \omega t$$

giving as a solution

$$Q_F = -\frac{I_b}{1+\omega^2\tau_p^2}\{\tau_p \sin \omega t - \omega\tau_p^2 \cos \omega t\}.$$

$$\therefore\ I_c = -\frac{Q_F}{\tau_F} = \frac{I_b(\tau_p/\tau_F)}{1+\omega^2\tau_p^2}\{\sin \omega t - \omega\tau_p \cos \omega t\}. \tag{1}$$

Thus from (1)

$$\left|\frac{I_o}{I_b}\right| = |h_{fe}(\omega)| = \frac{\tau_p}{\tau_F} \cdot \frac{1}{\sqrt{(1+\omega^2\tau_p^2)}}.$$

Chapter 9

9.1.

$\log_{10} V_a$		1·000	1·301	1·477	1·602	1·699
$\log_{10} I_a$	$\bar{1}$·301	0·342	0·748	0·987	1·167	1·295

The resulting graph is a straight line.
Slope of graph $= n = 1{\cdot}4$.

9.2. $E = I_1R_1 + V_a$,

$I_1 = I_a + I_2$,

$I_2 = V_a/R_2$.

Hence $E = I_aR_1 + V_a\left(\frac{R_1+R_2}{R_2}\right)$,

$100 = 10I_a + \frac{4}{3}V_a$ if I_a is in milliamps.

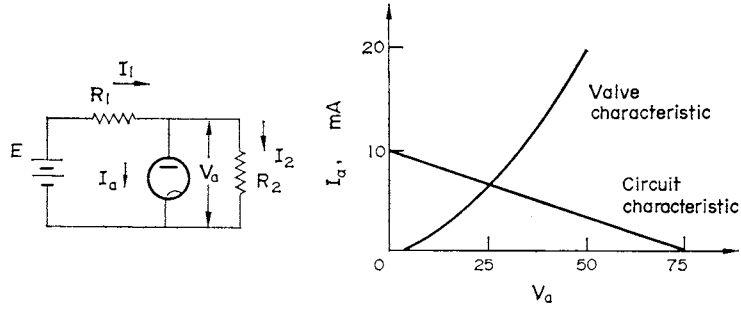


FIG. P. 9.2.

The point of intersection of the graph of this equation with that of the valve characteristic, gives for the operating conditions

$$I_a = 6{\cdot}8 \text{ mA and } V_a = 23 \text{ V}.$$

9.3. The circuit diagram and equation is as for problem 9.2.
Thus $E = 20I_a + \frac{7}{5}V_a$.
For the valve, $I_a = 2\times10^{-3}V_a^{3/2}$
and $\therefore E = 4\times10^{-2}V_a^{3/2} + \frac{7}{5}V_a$.
When $V_a = 100$ V, $E = 180$ V.
On removal of the voltmeter, $150 = 20I_a + V_a = 4\times10^{-2}V_a^{3/2} + V_a$.

This equation can be solved numerically:

V_a	100	110	120	122	124	125
$V_a^{3/2}$	1000	1150	1320	1350	1380	1400
$V_a+V_a^{3/2}$	140	156	173	176	179·2	181

The diode anode voltage is therefore 124 V.

9.4. $I_a = 10^{-5}(V_a+24V_g)^{3/2}$ A.

Differentiating with respect to V_g for the condition I_a = constant,

$$0 = 10^{-5}\times\tfrac{3}{2}(V_a+24V_g)^{1/2}\left(\frac{dV_a}{dV_g}+24\right),$$

$$\therefore \frac{dV_a}{dV_g} = \mu = -24.$$

Differentiating with respect to I_a for the condition V_g = constant,

$$1 = 10^{-5}\times\tfrac{3}{2}(V_a+24V_g)^{1/2}\left(\frac{dV_a}{dI_a}\right),$$

$$\therefore \frac{dV_a}{dI_a} = r_a = \frac{10^5}{\tfrac{3}{2}(V_a+24V_g)^{1/2}}.$$

When $V_a = 200$ V and $V_g = -6$ V,

$r_a = 8{,}900\ \Omega$.

9.5. The load-line is of the form $E = I_aR_a+V_a$,

$$\text{i.e. } 200 = 20I_a+V_a \text{ if } I_a \text{ is in mA.}$$

From the graph of the triode characteristic and the load-line, the anode difference corresponding to the two points of intersection is

$$V_B - V_A = 23 \text{ V.}$$

Thus, since the difference in grid voltage is 1 V,

$$\text{amplifier voltage gain} = 23.$$

From the anode characteristic:

in the operating region, $\dfrac{1}{\text{slope}}$ (= anode slope resistance) $= \dfrac{V_Q-V_P}{I_B-I_Q} = 13{,}400\ \Omega$, the anode voltage difference for constant anode current (= amplification factor) $= V_Q - V_P = 38$.

$$\therefore \text{ expected voltage gain} = \frac{38\times 20}{20+13{\cdot}4} = 23.$$

9.6. Voltage gain $= \dfrac{\mu R_a}{r_a + R_a}$.

Thus $25 = \dfrac{30\mu}{r_a + 30}$ and $20 = \dfrac{20\mu}{r_a + 20}$.

These two equations give $\mu = 50$ and $r_a = 30$ kΩ,

$$g_m = \frac{\mu}{r_a} = \frac{50}{30} 10^{-3} = 1{\cdot}66 \text{ mS}.$$

Chapter 10

10.1. Mean output voltage $= V_m \left(1 - \dfrac{\tau}{2RC}\right)$.

$V_m = \sqrt{2}\, V_{RMS} = 340$ V.

$\tau = \frac{1}{50} = 2 \times 10^{-2}$ sec, $R = 10^4\ \Omega$, $C = 10^{-5}$ F.

$\therefore$ Mean output voltage $= 306$ V.

Peak ripple voltage $= V_m \left(\dfrac{\tau}{2RC}\right) = 34$ V

or R.M.S. ripple voltage $= 24$ V.

10.2. Required peak ripple voltage $= 306 \times 3 \times 10^{-2} = 9{\cdot}2$ V.
$\therefore$ Ripple must be reduced by a factor of $34/9{\cdot}2 = 3{\cdot}7$.
An LC filter gives a reduction by a factor $\omega^2 LC$,

$$\therefore L = \frac{3{\cdot}7}{\omega^2 C} = \frac{3{\cdot}7}{4\pi^2 \times 50^2 \times 10^{-5}} = 3{\cdot}7 \text{ H}.$$

For full-wave rectification, initial ripple reduced by factor of two. LC filter is four times more effective (ω increased by factor of two).

$\therefore$ output peak ripple becomes $9{\cdot}2/8 = 1{\cdot}15$ V.

10.3. The equivalent circuit is shown. $r_a = \dfrac{\mu}{g_m} = \dfrac{20}{10^{-3}} = 20$ kΩ.

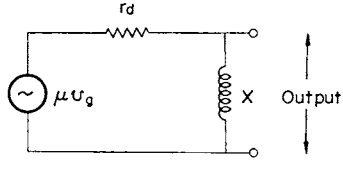


FIG. P. 10.3.

Voltage gain $= \dfrac{\mu X}{\sqrt{(r_a^2 + X^2)}}$ where $X = \omega L$.

This is a maximum when $\omega L \gg r_a$, i.e. at high frequencies, when the gain $\to \mu = 20$. When the gain is 10,

$$10 = \frac{20X}{\sqrt{[(20\times 10^3)^2 + X^2]}} \quad \text{or} \quad X = 1{\cdot}16\times 10^4 = \omega L.$$

$$\text{Frequency } f = \frac{\omega}{2\pi} = \frac{X}{2\pi L} = 184 \text{ kHz}.$$

10.4. The equivalent circuit is shown, in which

FIG. P. 10.4.

$$r_d = \frac{\mu}{g_m} = \frac{40}{4\times 10^{-3}}\ 10 \text{ k}\Omega. \qquad R_L = 120 \text{ k}\Omega.$$

By the use of Thévenin's theorem, the circuit can be reduced to the simple circuit shown, where

$$R_o = \frac{10\times 20}{10+20} \text{ k}\Omega = 6{\cdot}66 \text{ k}\Omega$$

and $$v_o = 40v_g \times \frac{20}{10+20} = 26{\cdot}7v_g.$$

$$\text{The voltage gain} = \frac{26{\cdot}7X}{\sqrt{[(6{\cdot}66\times 10^3)^2 + X^2]}} \quad \text{where} \quad X = 1/\omega C.$$

The gain is a maximum when $X \gg r_d$, i.e. at very low frequencies, when the gain is 26·7.

For the new condition, gain $= 26{\cdot}7\times 0{\cdot}75 = 20$.

$$\text{Then } 20 = \frac{26{\cdot}7X}{\sqrt{[(6{\cdot}66\times 10^3)^2 + X^2]}} \quad \text{or} \quad X = \frac{1}{\omega c} = 7{\cdot}6\times 10^3\ \Omega$$

$$C = \frac{1}{2\pi\times 10^4\times 7{\cdot}6\times 10^3} = 0{\cdot}0021\ \mu\text{F}.$$

10.5. The bias resistor $= \dfrac{4}{2\times 10^{-3}} = 2 \text{ k}\Omega.$

For $< 1\%$ of a.c. current through the resistor, the maximum capacitor impedance must be $2000\times 1/10^2 = 20\,\Omega$, and this will be at the lowest frequency, at 50 Hz,

$$C = \frac{1}{2\pi\times 50\times 20} = 160\ \mu\text{F}.$$

10.6. $$G = \frac{A}{1+A\beta}. \qquad \text{(a)}$$

Suppose a change in open-loop gain dA produces a change dG in closed-loop gain. By differentiating (a)

$$dG = \frac{[(1+A\beta)-A\beta]\,dA}{(1+A\beta)^2} = \frac{dA}{(1+A\beta)^2}$$

$$= \frac{dA}{A^2/G^2}.$$

$$\therefore \left(\frac{dG}{G}\right)\frac{1}{G} = \left(\frac{dA}{A}\right)\frac{1}{A}. \qquad \text{(b)}$$

Now $\frac{dG}{G} = \pm 1$ per cent; $\frac{dA}{A} = \pm 12$ per cent.

$$\therefore A/G = 12 \qquad \text{from (b)}$$

Since $G = 60$, $A = 12\times 60 = 720$.

From (a) $$60 = \frac{720}{1+720\beta}.$$

$$\therefore \beta = \frac{660}{60\times 720} = 0{\cdot}0153.$$

Chapter 11

11.1. Electron velocity $v = \sqrt{\left(\frac{2eV}{m}\right)} = \sqrt{\left(\frac{2\times 1{\cdot}6\times 10^{-19}\times 1000}{9{\cdot}11\times 10^{-31}}\right)}$

$$= 1{\cdot}87\times 10^{7} \text{ m/sec.}$$

$\therefore$ (a) to travel 5 cm, transit time $= \frac{5\times 10^{-2}}{1{\cdot}87\times 10^{7}}$

$$= 2{\cdot}7\times 10^{-9} \text{ sec;}$$

(b) transverse acceleration $= \frac{eE}{m} = \frac{1{\cdot}6\times 10^{-19}\times 50}{9{\cdot}11\times 10^{-31}\times 10^{-2}}$

$$= 8{\cdot}8\times 10^{14} \text{ m/sec}^2;$$

(c) total deflection $= \frac{V_D}{2V}\frac{lL}{d}$

$$= \frac{50}{2\times 10^{3}}\times\frac{5\times 20}{1}\times 10^{-2} \text{ m}$$

$$= 2{\cdot}5 \text{ cm.}$$

11.2. For equivalence of the electric and magnetic field, vB must be equal to E,

i.e. $$\frac{V_D}{d} = vB \quad \text{and} \quad B = \frac{V_D}{vd} = \frac{50}{1{\cdot}87\times 10^{7}\times 10^{-2}}$$

$$= 2{\cdot}7\times 10^{-4} \text{ T.}$$

Index