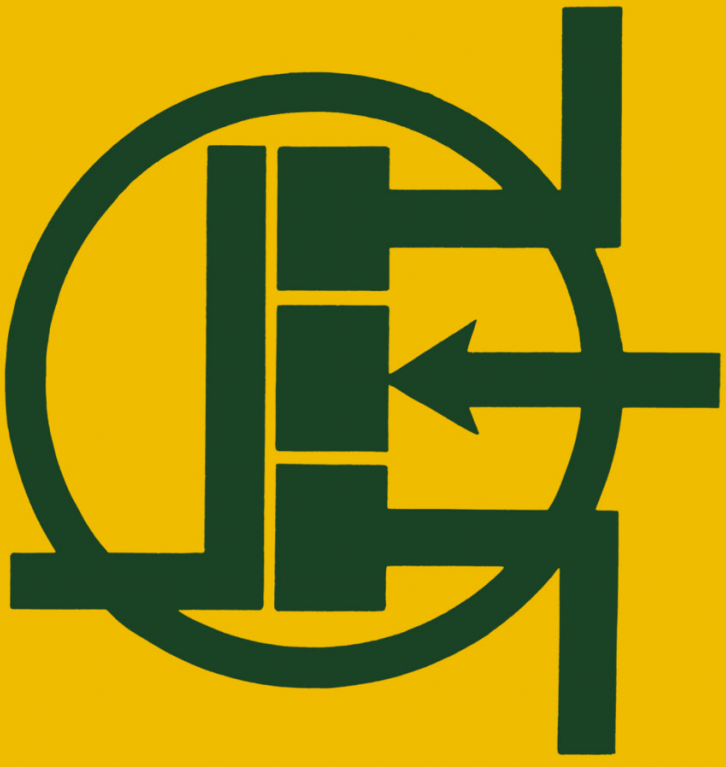
NOEL M. MORRIS

SEMICONDUCTOR DEVICES



MACMILLAN BASIS BOOKS IN ELECTRONICS

Semiconductor Devices

Macmillan Basis Books in Electronics
Series editor Noel M. Morris

Digital Electronic Circuits and Systems Noel M. Morris
Linear Electronic Circuits and Systems G. D. Bishop
Electrical Circuits and Systems Noel M. Morris

Other Related Books

Essential Formulae for Electronic and Electrical Engineers
Noel M. Morris
Semiconductor Electronics by Worked Example F. Brogan
The Electrical Principles of Telecommunications R. Lowe and D. Nave

Semiconductor Devices

Noel M. Morris

Principal Lecturer,
North Staffordshire Polytechnic

M

Softcover reprint of the hardcover 1st edition 1976

First published 1976 by
THE MACMILLAN PRESS LTD
London and Basingstoke
Associated companies in New York Dublin
Melbourne Johannesburg and Madras

SBN 333 185 35 8 (hard cover)
333 185 36 6 (limp cover)

ISBN 978-0-333-18536-0 ISBN 978-1-349-15671-9 (eBook)
DOI 10.1007/978-1-349-15671-9

Set in IBM Press Roman by
PREFACE LTD

Contents

Foreword

Technological progress has nowhere been more rapid than in the fields of electronics, electrical, and control engineering. The *Macmillan Basis Books in Electronics Series* of books have been written by authors who are specialists in these fields, and whose work enables them to bring technological developments sharply into focus.

Each book in the series deals with a single subject so that undergraduates, technicians and mechanics alike will find information within the scope of their courses. The books have been carefully written and edited to allow each to be used for self-study; this feature makes them particularly attractive not only to readers approaching the subject for the first time, but also to mature readers wishing to update and revise their knowledge.

Noel M. Morris

Preface

The boundaries of electronics are continually expanding, and are subject only to the limitations imposed by production technology and our knowledge of device principles. The aim of this book, which is one in the *Macmillan Basis Series*, is to introduce the principles of electronic devices to students in the widest sense of the word. An attempt has been made in this book to avoid leaning too heavily on the physics of devices. As soon as the operation of a particular device has been explained, circuits and applications are introduced to illustrate its use.

The physical basis of semiconductor elements is given in a non-mathematical manner in the first chapter, and provides readers with a broad and progressive approach to the topic. The second chapter deals with basic non-active semiconductor devices such as thermistors, voltage-dependent resistors, Hall-effect devices, and many others that are used as transducing or sensing elements. An important element, the p–n junction diode, is discussed in chapter 3, together with many variants such as Zener diodes, varactor diodes, p–i–n diodes and others. The book logically proceeds to deal with bipolar junction transistors in chapter 4; in this chapter the use of the BJT as an amplifying element and as a switch is explained, the discussions being supported by appropriate circuits. A similar coverage is given to field-effect transistors in chapter 5; both junction-gate and metal-oxide-semiconductor types are described.

The principles of the manufacture of integrated circuits are outlined in chapter 6, including techniques used in BJT and FET circuits. The work in chapter 7 on charge-coupled devices follows logically the coverage of field-effect devices; these devices are of considerable interest in the fields of digital storage, analog shift registers and optical imaging. The rapid development of digital electronics, as exemplified by the ubiquitous pocket calculator, is illustrated in chapter 8 in which a variety of semiconductor memories and memory systems are discussed. Aspects of power electronics are not ignored, and in chapter 9 multilayer devices such as the thyristor, the triac, the diac, the silicon controlled switch and the programmable unijunction transistor are described. The triac is a complex device, and its operation is fully described in chapter 9. Finally chapter 10 deals with a wide range of optoelectronic devices such as photodiodes, solar cells, phototransistors, light-emitting diodes and optically coupled isolators. Liquid crystal displays, although not semiconductor devices, are employed in increasing numbers, and are discussed in the final chapter.

I would like to thank the electronics industry at large for supplying information not only about devices currently in use, but also about devices in the development stage. I am also indebted to my wife for her not inconsiderable efforts in the preparation of the manuscript, and for her patience with my preoccupation while it was being written.

Meir Heath Noel M. Morris

1 Semiconductors

1.1 Introduction

Electrically speaking, a semiconductor is a material whose value of electrical conductivity lies between that of a good conductor and that of a good insulator (strictly, a semiconductor could also be described as a semi-insulator). To understand the operation of semiconductor devices, it is necessary to grasp the principal electronic features of the structure of atoms and of crystals.

1.2 Atomic Structure

The structure of all atoms is broadly the same in that a number of positive charges, or *protons*, concentrated at the centre of the atom in the *nucleus*, are surrounded by orbiting negative charges, or *electrons*. The nucleus also contains other particles known as *neutrons*, having no electrical charge; these particles have no effect on the electrical properties of the material.

Each proton carries a positive charge of 1.6×10^{-19} coulomb and has a mass of 1.67×10^{-27} kg. Although the magnitudes of these quantities are very small, they can be measured and have significance in an engineering context. An electron carries an electrical charge which is opposite to that of the proton, that is, -1.6×10^{-19} coulomb, and its mass is 1840 times smaller than that of the proton. In an isolated atom the electrons orbit in layers or *shells*, rather like planets orbiting around the sun.

Electrons are attracted towards the nucleus by virtue of the force of electrostatic attraction between the two opposite types of charge, and are subjected to a mechanical force away from the nucleus by virtue of their motion; the two forces balance when the electron is in orbit. The total electrical charge on an atom is zero, since it has the same number of protons and electrons.

The nature of atoms is such that each shell contains up to a maximum number of electrons. This fact was first suggested by the physicist Wolfgang Pauli, and is known as *Pauli's exclusion principle*. A simple analogy of this principle can be drawn from that of a multistorey car park that can accommodate different numbers of cars at each level. The parked cars are analogous to electrons, and the parking levels to electronic orbits. Human nature being what it is, the lower levels fill up before the higher levels. So it is in nature that the lower shells of atoms (those closest to the nucleus) are filled up first, followed by the higher levels. Because of

the close proximity of the inner layer of electrons to the nucleus, they are strongly attracted to it and are said to be *tightly bound* to the atom. The electrons orbiting further away from the nucleus do so by virtue of having a higher electronic energy than those in the inner shells. Since these electrons are further away from the nucleus, they are *loosely bound* and can be more easily detached from the atom than can electrons in lower orbits.

The shells themselves have been identified by scientists by means of alphabetical characters, commencing at the letter K; the maximum number of electrons that a shell can contain is

K 2 electrons
L 8 electrons
M 18 electrons
N 32 electrons

The maximum number of electrons is calculated from the expression $2n^2$, where n is the number of the shell (K = 1, L = 2, M = 3, etc.), so that the M shell may contain up to $2 \times 3^2 = 2 \times 9 = 18$ electrons. A simplified atomic structure of a neon atom is shown in figure 1.1, which contains two electrons in the K shell and eight electrons in the L shell, that is, both shells are 'full' of electrons.

The electronic properties of certain materials are explained by the way in which their shells are subdivided. For example, the M shell consists of three subshells containing up to two, six and ten electrons, respectively. In some cases, not all of the subshells are complete. For example, in *silicon* only the first two subshells of the M shell contain electrons (that is, those which may contain up to two and six electrons, respectively). This is of particular significance, as will be shown later.

The shells of an isolated atom can also be described in terms of the amount of energy contained by the orbiting electrons and, in this case, are called *energy levels*. Since a given electron may only have certain values of energy, corresponding to the shells described above, all other energy levels are forbidden. The gaps between those levels are known as *forbidden energy gaps*. More is said about this in section 1.3.

At absolute zero temperature (0 K or $-273\ ^\circ C$), all electrons are at their lowest energy level. The outermost level which contains electrons is known as the *valence level*, and it is the number of electrons in the valence level which dictates the chemical and electrical properties of the substance.

Atoms combine in a number of ways. They either *lose electrons* to other atoms in order to empty a partially filled outer shell, or they *gain electrons* to completely fill a nearly filled shell, or they *share electrons* with other atoms It is those materials which combine using the latter process that are of most interest to electronic engineers. In this process each electron in the valence energy level in each atom orbits not only around the parent atom, but also around another atom; this effectively binds the atoms together. This bond is known as a *covalent bond*. Silicon, a material widely used in semiconductor devices, is of this type, the number

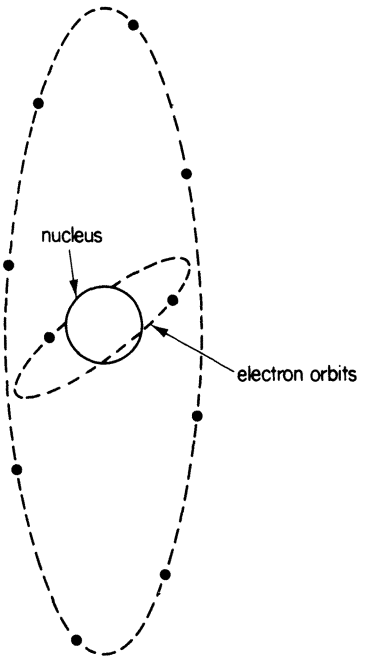


Figure 1.1 Atomic structure of a neon atom

of electrons in the shells of an isolated atom being

K 2 electrons
L 8 electrons
M 4 electrons

Comparing these values with the maximum values given earlier, we see that the K and L shells are full while the M shell (the valence shell in this case) contains only four electrons. As stated earlier the M shell consists of three subshells, only two of which may contain electrons in the silicon atom. Thus the maximum number of electrons in the valence energy shell of the silicon atom is 2 + 6 = 8. In order to 'fill' the valence shell of a silicon atom, it is necessary for each atom to share four electrons with its neighbours. The way in which this is done is illustrated in figure 1.2. In this case the four valence electrons from atom A orbit between atom A and atoms B, C, D and E. Similarly, electrons from the latter atoms orbit around atom A, making eight atoms in all orbiting atom A. This process applies throughout the crystal.

1.3 Energy Bands

In a solid such as silicon, the atoms are close to one another, and the valence electrons are attracted by other nuclei so that the electron orbits are modified in

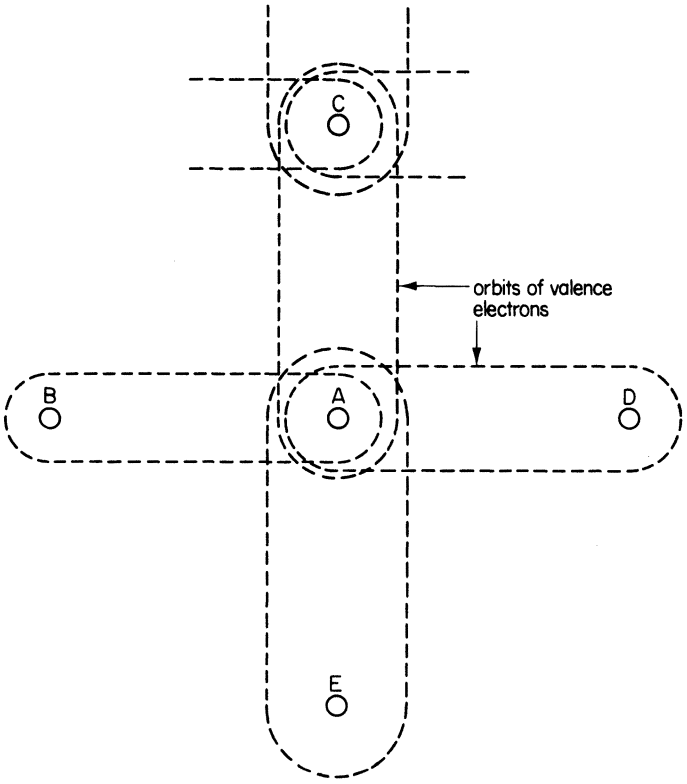


Figure 1.2 Orbits of the valence electrons in silicon

the manner shown in figure 1.2. Consequently the outermost electrons are equally likely to be in orbit around any one of a number of atoms. Thus, in solids, electrons are likely to be found in any one of a number of energy levels. In a semiconductor crystal where the atomic nuclei are close to one another, the allowed energy levels increase in numbers and merge into *energy bands*. An energy-band model of a semiconductor is shown in figure 1.3, in which the valence energy band is full at absolute zero of temperature. The reason for introducing energy level W_F on figure 1.3 is given later.

It is paradoxical that even though the valence energy band is full of electrons at this temperature, the semiconductor is a perfect insulator. In order that the electrical conduction may occur, electrons must be free to move in response to an applied electric field. When the valence energy band is fully occupied by electrons, none can move freely since there is no room to do so and, under this condition, the semiconductor has the characteristics of an insulator.

For an electron to become available for electrical conduction purposes, it must gain sufficient energy to transfer across the forbidden energy gap and appear in the next available empty energy band, which is known as the *conduction band*. Once in

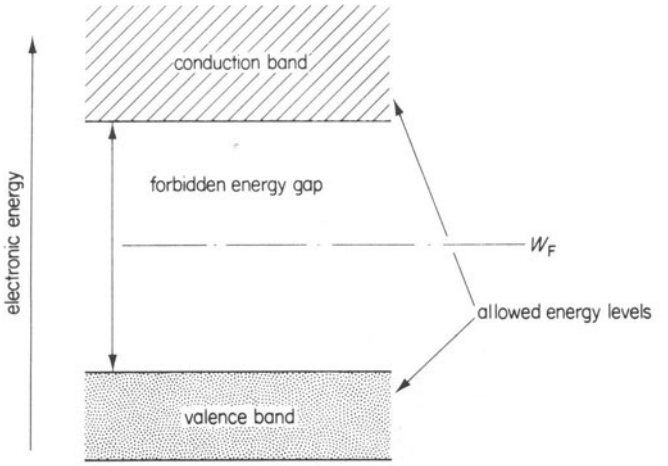


Figure 1.3 Energy-band model for a semiconductor

this band, it is free to move under the influence of an electric field and can take part in the flow of electric current. Electrons may gain the energy required to transfer across the forbidden energy gap in any one of a number of ways; this transfer most commonly occurs when the crystal absorbs energy from heat, light and other electromagnetic and atomic radiation, the former source being the most common. The electrical conductivity due to this cause is known as *intrinsic conductivity* or *i-type* conductivity. Quite clearly, at increased values of temperature the number of electrons which cross into the conduction band increases; a consequence is that the electrical conductivity, σ, of most semiconductors increases with increase in temperature. Since resistivity, ρ, is given by $\rho = 1/\sigma$, the resistivity of semiconductors decreases with increase of temperature. This fact is utilised in several types of semiconductor device, including *negative temperature-coefficient thermistors* (see chapter 2). Clearly if the forbidden energy gap is smaller in one semiconductor than in another, then the intrinsic conductivity is greater in the former than in the latter. Semiconductors in which conduction is entirely due to intrinsic conductivity are known as *i-type semiconductors.*

The amount of energy required to cross the forbidden band is usually expressed in *electron volts*, eV, where

$$1 \text{ eV} = 1 \text{ volt} \times \text{the charge on an electron in coulomb}$$

$$= 1.6 \times 10^{-19} \text{ joule or watt second}$$

This seemingly insignificant amount of energy (there are 6 250 000 000 000 million electron volts in a watt second!) is quite large on the atomic scale. In a silicon crystal at room temperature, the energy required to transfer across the forbidden energy gap is about 1.1 eV. A list of the forbidden energy gaps of a range of popular semiconductors is given in table 1.1.

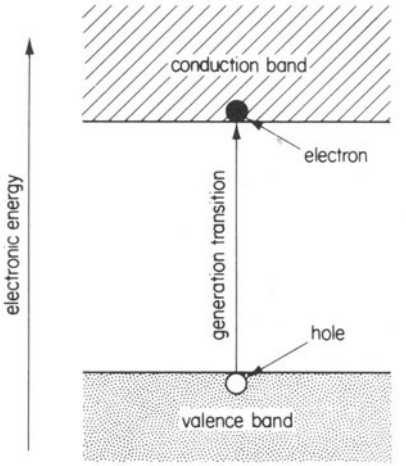


Figure 1.4 Transition of an electron into the conduction band causing the generation of an electron–hole pair

1.4 The Fermi Energy Level, W_F

At absolute zero temperature the valence band is full of electrons, with none in the conduction band. Under these conditions there is a 100 per cent probability of electrons being in the valence band and zero probability of electrons being in the conduction band. There is, theoretically at any rate, an energy level between the valence and conduction bands which has a 50 per cent probability of being filled. This is known as the *Fermi energy level*, and is shown as W_F in figure 1.3. It is named after the physicist Enrico Fermi, who first suggested an explanation for the distribution of electrons in the band structure.

In a pure semiconductor material with the energy-band structure in figure 1.3, the Fermi level is situated half-way between the conduction and valence bands. This is within the forbidden energy gap, so that individual electrons may not have an energy value equal to the Fermi energy. Although the Fermi energy level may not physically exist, as is the case in figure 1.3, it is a useful aid in explaining the operation of semiconductor devices, much as the concept of 'sea level' is a convenient reference for the measurement of altitudes.

1.5 Holes and Electrons

When electrons cross the band gap into the conduction band (more correctly known as a charge carrier *generation transition*), they become free to take part in the process of electric current flow. When the transition takes place, it leaves behind it an empty location in the valence band. The empty location is described as a *hole* (see figure 1.4), and may be regarded as the absence of an electron in a location where one would normally be found. Much as an electron is a negative charge carrier, a hole is regarded as a positive charge carrier.

Table 1.1 Forbidden energy gaps of popular semiconductor materials

Substance	Chemical symbol	Forbidden energy gap (eV)
Germanium	Ge	0.72
Silicon	Si	1.10
Diamond (carbon)	C	6.70
Indium antimonide	InSb	0.20
Indium phosphide	InP	1.25
Gallium arsenide	GaAs	1.40
Gallium phosphide	GaP	2.30
Cadmium telluride	CdTe	1.50
Cadmium selenide	CdSe	1.70
Cadmium sulphide	CdS	2.50
Zinc oxide	ZnO	3.30
Zinc sulphide	ZnS	3.60
Lead selenide	PbSe	0.30
Lead telluride	PbTe	0.30
Lead sulphide	PbS	0.40

The transition of an electron into the conduction band spontaneously generates a hole in the valence band, and the process is said to produce an *electron–hole pair*; it is this mechanism that gives rise to intrinsic conductivity in semiconductors. An atom in which some of its electrons have been temporarily raised above the lowest available level is described as being in an *excited* state.

1.6 Mobility

The charge carrier mobility is an important parameter in semiconductors, and is a measure of the ease with which the carrier can move within the material under the influence of an electric field. The charge carrier mobility, μ, is given by the relationship

$$\mu = \frac{\text{average value of drift velocity (m/s)}}{\text{electric field strength (V/m)}}$$

$$= \frac{v}{E} \quad \frac{\text{m/s}}{\text{V/m}} \text{ or } \frac{\text{m}^2}{\text{V s}}$$

In semiconductors, a high value of mobility is usually required.

Table 1.2 Energy gaps and mobilities of some semiconductors

Element	Band gap (eV)	Electron mobility ($m^2/V\,s$)	Hole mobility ($m^2/V\,s$)
Germanium	0.72	0.39	0.19
Silicon	1.10	0.15	0.05
Indium antimonide	0.20	>6	0.30
Gallium arsenide	1.40	>0.50	0.03
Gallium phosphide	2.25	0.01	0.002

Factors that affect carrier mobility include defects in the crystal structure and also changes in temperature. Defects or irregularities in the structure impede the flow of charge carriers and therefore reduce the mobility. An increase in temperature not only results in an increase in the generation of electron–hole pairs, but also causes the atomic nuclei to vibrate. At elevated temperature the atomic vibration becomes sufficient to impede the flow of charge carriers, and reduces their mobility; this effect is utilised in some *positive temperature-coefficient thermistors* (see chapter 2).

Readers will note in figure 1.4 that electrons are at a higher energy level than are holes. It follows that electrons have higher mobility values than do holes. Typical mobility values of some semiconductors at room temperature are given in table 1.2.

The case of indium antimonide in table 1.2 raises an interesting point in connection with the usefulness of semiconductor materials. Although the charge carriers associated with this substance have high mobilities when compared with other semiconductors, indium antimonide has limited use as a semiconductor due to the low value of band gap. As will be shown later, a relatively high value of band gap is desirable in materials used in diodes and transistors. Unfortunately compounds with high values of band gap usually have relatively low values of mobility. Consequently many materials with the correct type of atomic structure are unsuitable for use in semiconductor manufacture.

1.7 Recombination and Lifetime

An electron in the conduction band moves in a random fashion through the crystal structure and, at some stage, it encounters a hole, that is, it returns to the valence band of one of the atoms. When this occurs, the hole and electron are no longer free; this process is known as a *recombination transition* or simply as a *recombination.*

The process of generation and recombination of charge carriers occurs all the time, and the situation is continually changing. In other words, thermal agitation continually generates new electron–hole pairs, and others disappear as a result of recombinations. The *average time* that a hole or an electron exists between

generation and recombination is known as the *mean lifetime* or simply as the *lifetime* of the charge carrier. The value of the lifetime is important in many semiconductor devices, its value ranging from nanoseconds (1 ns = 10^{-9} s) to hundreds of microseconds (1 μs = 10^{-6} s).

The most important recombination mechanism in silicon and germanium involves what are known as *traps* and *recombination centres*, which are discussed in section 1.11.

1.8 Drift Current and Diffusion Current

Two principal mechanisms exist that cause the movement of charge carriers in a semiconductor material, and are known as drift current and diffusion current.

In the absence of an applied voltage, the density of thermally generated charge carriers at any point in the crystal differs from that at any other point. The charge carriers generated in this way tend to move or diffuse from a region of high concentration of mobile charge carriers to one of a low concentration. The charge carriers consequently appear to move about in a random fashion as the charge density changes its centre. Thus an electron starting at A in figure 1.5 moves about in a random fashion and, some time later, appears at B. Such movement of charge carriers is known as *diffusion current*.

When a voltage is applied to a semiconductor, mobile charge carriers tend to drift in the direction of the electric field – electrons towards the positive pole and holes towards the negative pole. This movement of charge carriers is known as *drift current*.

Both of the above mechanisms are important in describing the operation of semiconductor devices.

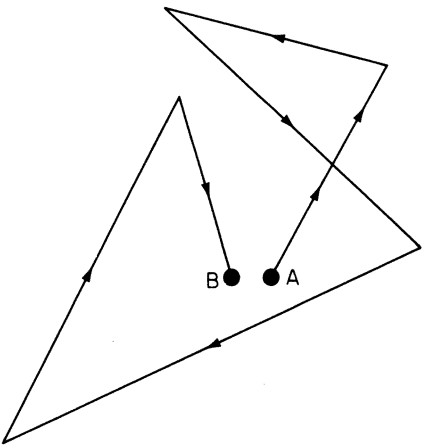


Figure 1.5 Illustrating the mechanism of diffusion current

1.9 Impurity Doping in Semiconductors

In the intrinsic semiconductors so far discussed, electrical conductivity is due to the spontaneous generation of electron–hole pairs. This process, as described above, is a function of thermal and other effects. By adding impurity atoms, that is, by *doping* the pure material in the order of one part in 10^8, the characteristics of the material are modified so that the conductivity is, very largely, independent of temperature over a wide temperature range.

Dopant materials are slected from a knowledge of the periodic table of the elements as follows. The most popular types of semiconductor material – silicon and germanium – contain four valence electrons and are described as *group IV elements* in the periodic table. Elements having three valence electrons are described as group III elements, and those with five valence elements are group V elements.

If a group V element, say arsenic, is introduced into a silicon crystal (a group IV element), only four of the five valence electrons in the arsenic atom are required for atomic bonding purposes. The remaining electron in the valence shell of the arsenic atom is not subject to the same forces binding it to the atomic structure, and is readily available for conduction purposes. An impurity of this kind is known as a *donor impurity* atom, since it donates a mobile electron to the structure. This has the effect of increasing the conductivity of the material, and is independent of temperature. Semiconductors to which impurities have been added are referred to as *extrinsic semiconductors*. Also, as in the case considered above, materials having an excess of mobile electrons are known as *n-type semiconductors* (*n* for *n*egative charge carriers). Since group V materials have five valence electrons, they are known as *pentavelent atoms*. Group V materials such as antimony, arsenic and phosphorus function as donor atoms when used with group IV materials such as silicon and germanium.

If a *trivalent* element having three valence electrons (from group III in the periodic table) is introduced into a pure crystal of silicon or germanium, the atomic bonding is incomplete, since the impurity element has only three electrons in its valence shell. That is, the structure has a deficiency of one electron and, when connected in an electrical circuit, it appears that the new extrinsic semiconductor has an excess of holes or positive charge carriers. The trivalent impurity atom is known as an *acceptor* impurity, since it causes the atomic structure to accept electrons readily in order to make good the bonding deficiency. Semiconductor materials having an excess of mobile holes are known as *p-type semiconductors* (*p* for *p*ositive charge carriers).

The majority of semiconductor devices are manufactured from a group IV material, usually silicon or germanium, doped either with a group V material to give an *n*-type semiconductor, or with a group III material to give a *p*-type semiconductor. A selected list of substances from groups II to VI of the periodic table is given in table 1.3.

Compounds containing only two elements – known as *binary compounds* – can

Table 1.3 A selection from the periodic table of elements

Group II	Group III	Group IV	Group V	Group VI
Be beryllium	Al aluminium	C carbon	As arsenic	O oxygen
Cd cadmium	B boron	Ge germanium	Bi bismuth	S sulphur
Hg mercury	Ga gallium	Pb lead	N nitrogen	SE selenium
Mg magnesium	In indium	Si silicon	P phosphorus	Te tellurium
Zn zinc	Tl thallium	Sn tin	Sb antimony	

be manufactured with an average number of four valence electrons. An example is *gallium arsenide* (GaAs), which is a compound of group III material (Ga) and a group V material (As). If the atomic ratio is 1:1, that is, for each Ga atom there is one As atom, then the result is an intrinsic semiconductor of four valence electrons per atom. Other compounds of this type include gallium anitmonide, aluminium antimonide and indium antimonide (see table 1.1). These compounds are known as III-V compounds, and can be converted into extrinsic materials in a number of ways. For example, if the compound contains more group III atoms than group V atoms it becomes a *p*-type material, and if there is an excess of group V atoms it becomes an *n*-type material. Alternatively an *n*-type material results if a group VI element is used as a donor impurity to replace a group V atom, tellurium and sulphur being examples of donor materials. If a group II material such as zinc is introduced into a pure III-V compound, then it becomes a *p*-type semiconductor.

Binary compounds of materials from groups II and VI can also be used to form a material with an average of four valence electrons. Examples of II-VI compounds include zinc sulphide and cadmium sulphide (see table 1.1).

However, readers should not think that all group IV substances are suitable for use as semiconductors. For example, the group IV substance in the form of a diamond crystal has a very large energy gap of 6.7 eV (see table 1.1) and low electron mobility (0.18 m^2/V s) and is, at normal temperatures, an insulator. At the other end of the group IV table is grey tin with an energy gap of 0.1 eV, which is a conductor at room temperature.

Also, not all intrinsic semiconductors have an average number of four bonding electrons per atom. For example, binary compounds from groups IV and VI with four and six valence electrons, respectively, produce narrow forbidden energy gaps in the range of interest. Compounds of this kind include lead sulphide, lead selenide and lead telluride (see also table 1.1). These materials are useful in infrared detectors.

1.10 Impurity Levels on the Energy-band Diagram

The effect on the energy-band diagram of silicon of introducing impurity atoms is shown in figure 1.6. If silicon is doped with an acceptor impurity to form a *p*-type

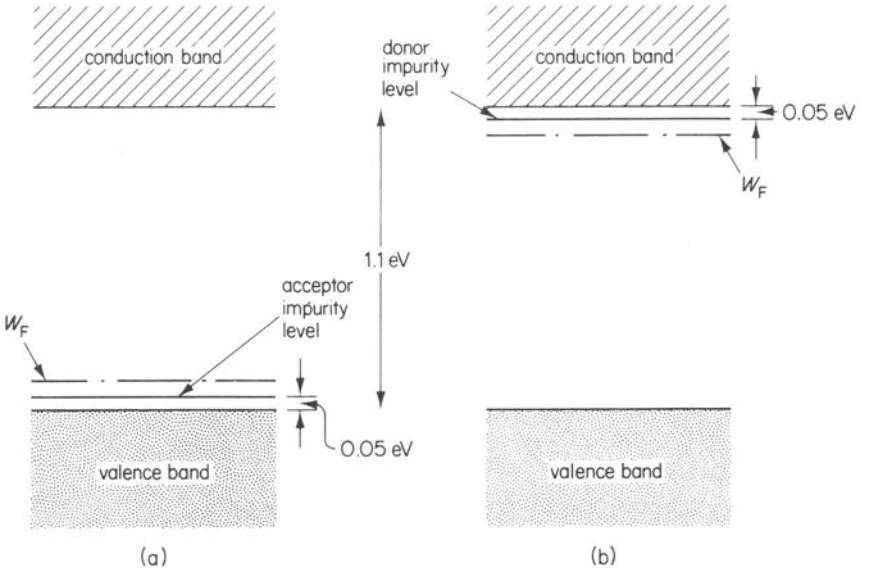


Figure 1.6 Energy-band structures for (a) *p*-type and (b) *n*-type silicon semiconductors

material, it has the effect of creating an empty energy level close to the valence band, shown in figure 1.6a. This band is readily filled by electrons from the valence band, thus creating mobile holes there.

In an *n*-type material the donor atoms create a full energy level close to the conduction band, shown in figure 1.6b. Electrons readily leave this energy level to create mobile negative charge carriers in the conduction band.

A result of introducing the acceptor impurity level in *p*-type semiconductors, figure 1.6a, is the probability of holes appearing in the valence band; this causes the Fermi level to be reduced in value when compared with an intrinsic semiconductor. Similarly, the donor level in the *n*-type semiconductor, figure 1.6b, increases the probability of electrons appearing in the conduction band; in this case, the Fermi level is raised when compared with an intrinsic semiconductor.

1.11 Trapping Levels and Recombination Centres in Semiconductors

As mentioned earlier the lifetime of a charge carrier is the time interval between its generation and its recombination with a charge carrier of the opposite type. Recombination may occur in one of a number of ways. One possibility is that the charge carriers undergo a *direct transition* from the conduction band to the valence band, shown in figure 1.7; only in a very few types of semiconductor does this occur. Transitions of this kind occur in gallium arsenide, which is a material used in light-emitting diodes; this type of material is referred to as a *direct-gap* material.

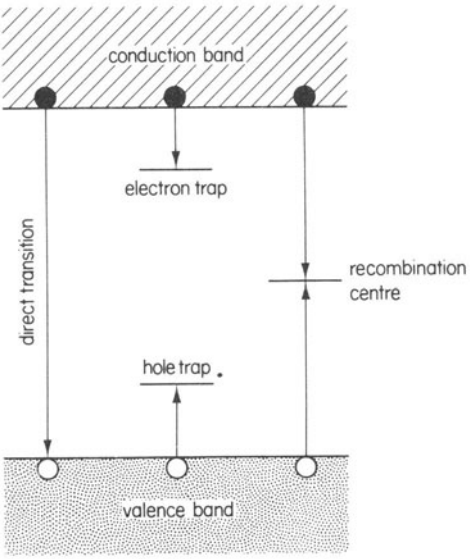


Figure 1.7 Trapping levels and recombination centres in semiconductors

In silicon and germanium, electrons return to the valence band via one or more intermediate energy levels, known as *traps* or *trapping levels*. Trapping levels near to the conduction band are generally referred to as *electron traps*, and trapping levels near to the valence band are known as *hole traps*. Where a level exists near the centre of the energy gap, it acts as a *recombination centre*, which has a high probability of capturing both a hole and an electron. Materials in which recombination takes place via one or more trapping levels are known as *indirect-gap* materials.

Trapping levels result from local discontinuities in the crystal structure, and from the presence of impurity atoms and also from other crystal defects.

In devices used in switching circuits, that is, in some transistors and in thyristors, rapid turn-off times are achieved by using a technique known as *gold-doping*; the improved switching performance is obtained by deliberately introducing recombination centres into the structure of the semiconductors used, by doping the semiconductor with gold ions in order to reduce the lifetime of the charge carriers.

1.12 Majority and Minority Charge Carriers

In *n*-type semiconductors, where there is an excess of mobile electrons, current flow is predominantly due to electron flow from the negative pole of the supply to the positive pole. Consequently, in *n*-type materials, electrons are known as the *majority charge carriers*. Also, at normal operating temperatures, electron–hole pairs are thermally generated, and the holes contribute to current flow. Since the

holes are fewer in number than are electrons in *n*-type material, they are known as *minority charge carriers*. The lifetime of minority carriers is short since they quickly combine with majority carriers; none the less, their lifetime is finite, and it is sometimes desirable to reduce it by various methods including gold doping (see section 1.11).

In *p*-type semiconductors, holes are majority charge carriers and electrons are minority charge carriers.

1.13 Energy-band Diagrams for Insulators and Conductors

Figure 1.8a illustrates the energy-band diagram for an insulator. The forbidden gap in an insulator is wider than that in a semiconductor, consequently the intrinsic conductivity is very low indeed.

At the other extreme, a good conductor is one in which the conduction and valence bands either touch or overlap slightly (see figure 1.8b), and there is an ample supply of mobile charge carriers at even very low temperatures.

1.14 Work Function and Contact Potential

When different types of metal are brought into contact with one another, it is frequently found that a small value of voltage exists between them. This is known as *contact potential* and can be explained in terms of energy-band diagrams. This

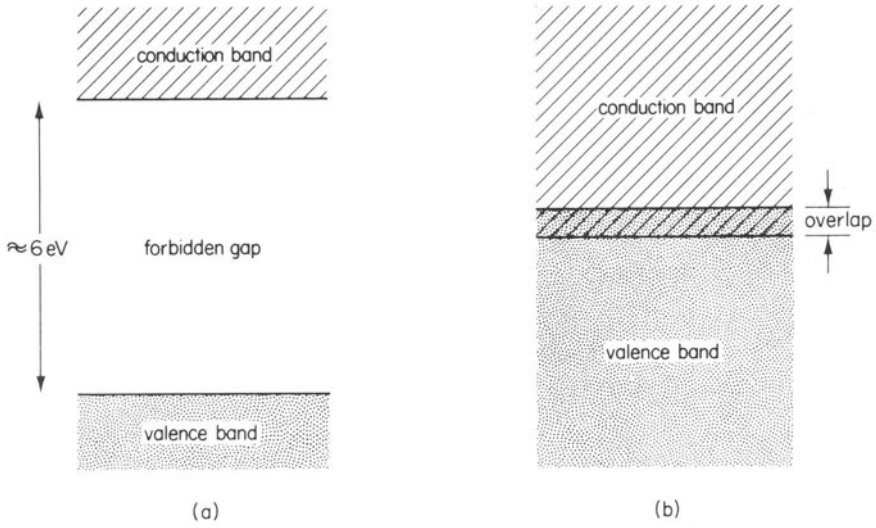


Figure 1.8 Energy-band diagrams for (a) an insulator and (b) a good conductor

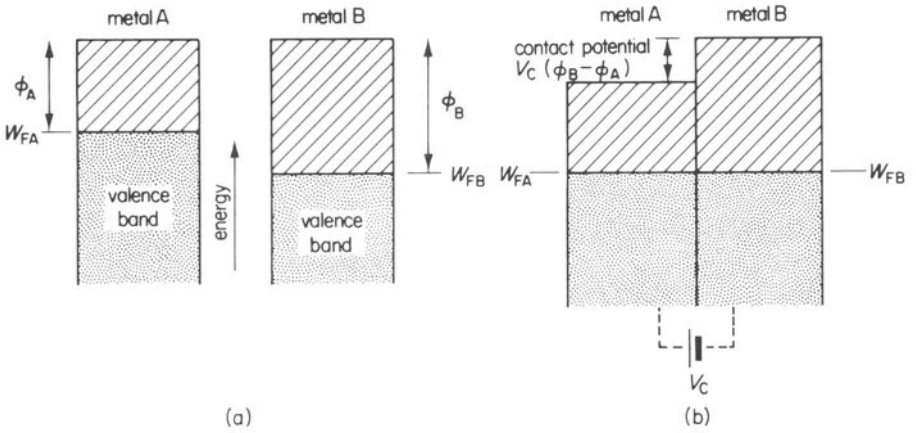


Figure 1.9 Energy-band diagrams (a) for isolated metals of different work functions and (b) for the metals in contact

concept is extended in section 1.15 to discuss contact between metals and semiconductors; contact effects of this type are of practical significance in several semiconductor devices, including *thermoelectric devices* (chapter 2) and *Schottky barrier diodes* (chapter 3).

Consider two metals A and B having the energy-band structures in figure 1.9a. The Fermi levels in metals A and B are W_{FA} and W_{FB}, respectively. In metal A, the amount of energy ϕ_A is the energy required by an electron to enable it to escape from the surface of the metal; this energy difference is known as the *work function* of the metal. Similarly, the work function of metal B is ϕ_B.

When the two metals are brought into contact with one another, the two upper energy levels of the diagrams initially have the same value. However, the electrons in the full valence band in metal A can flow into metal B. Consequently, metal A loses electrons to metal B and metal B gains electrons, hence metal A becomes positive with respect to B. This gives rise to a *contact potential* between the two metals, and is shown as V_C in figure 1.9b. In general **electrons flow from the material with the smaller work function to the material with the higher work function**. The value of the contact potential is directly related to the *energy barrier* or work function difference, $\phi_B - \phi_A$. Another, and important consequence is that the Fermi energy levels in the two metals assume the same values after contact is made between them.

1.15 Metal-to-semiconductor Junctions

Electrical junctions between metals and semiconductors fall into one of two groups, namely *ohmic contacts* and *rectifying contacts*. To illustrate the above categories, we consider the junction between a metal and an *n*-type semiconductor.

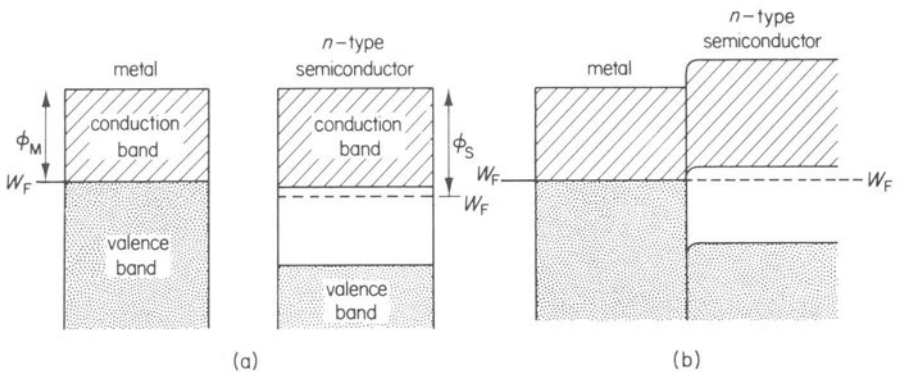


Figure 1.10 Energy-band diagrams for a metal-to-semiconductor ohmic contact (a) before contact and (b) after contact

n-type semiconductor, ohmic contact, $\phi_M < \phi_S$

An ohmic contact is said to be established between the two when the work function ϕ_M of the metal is less than the work function ϕ_S of the semiconductor. Figure 1.10a illustrates the energy-band diagrams before contact is made, W_F indicating the Fermi energy levels in the two materials. When the two are brought into contact, electrons flow from the metal into the semiconductor (which has the largest value of work function) until the Fermi levels coincide, illustrated in figure 1.10b. In this type of contact, the conduction bands overlap and the energy difference $\phi_M - \phi_S$ is small so that electron flow is easy in either direction.

An ohmic connection occurs between a *metal and a p-type* semiconductor when $\phi_M > \phi_S$, that is, when the work function of the metal is *greater than* that of the semiconductor.

n-type semiconductor, rectifying contact, $\phi_M > \phi_S$

In this case the junction allows easy flow of current in one direction only. The energy-band diagrams before contact in this case are shown in figure 1.11a. As before, electrons flow from the lower work-function material, that is, from the semiconductor, until the Fermi levels align. The resulting energy-band structure after contact is shown in figure 1.11b. Under equilibrium conditions a potential barrier, known as the *diffusion potential*, of $\phi_M - \chi$ exists between the two materials, where χ is the depth of the conduction band of the semiconductor. There is a region, known as the *depletion layer*, between the two materials in which there are no free charge carriers since electrons in that region have transferred to the metal. The resistivity of this region is therefore much greater than that of the semiconductor itself. Thus current flow can only take place if the voltage applied to the junction is in the correct direction to overcome the diffusion potential. That is, a rectifying junction is formed which permits current to flow easily in one

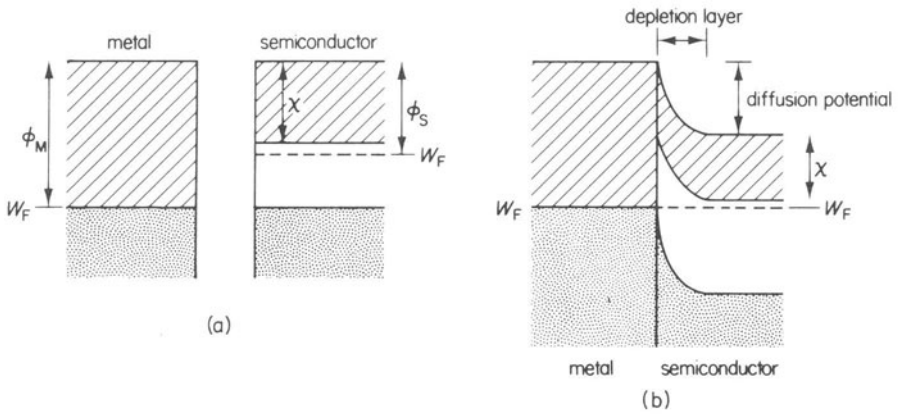


Figure 1.11 Energy-band diagrams for a metal-to-semiconductor rectifying contact (a) before contact and (b) after contact

direction, but not in the reverse direction. This topic is pursued in greater depth in chapter 3. A practical application of the metal to *n*-type rectifying contact is to the so-called *hot carrier diode* or *Schottky barrier diode*.

A rectifying connection occurs between a *metal and a p-type* semiconductor when $\phi_M < \phi_S$, that is, when the work function of the metal is *less than* the work function of the semiconductor. This type of rectifying contact is employed in *copper oxide* and *selenium rectifiers*.

1.16 Some Applications of Semiconductor Materials

The variety of applications of semiconductor materials is increasing all the time, and the following list illustrates applications of some of the more important semiconductors.

Barium titinate thermistors having a positive resistance–temperature coefficient
Bismuth telluride thermoelectric conversion
Cadmium sulphide electroluminescent devices, photoconductive cells
Gallium arsenide diodes, transistors, light-emitting diodes, lasers, microwave generators
Germanium diodes, transistors
Indium antimonide magnetoresistors, piezoresistors, infrared-radiation detectors
Indium arsenide piezoresistors
Lead sulphide infrared-radiation detectors
Lead telluride infrared-radiation detectors
Metallic oxides thermistors having a negative resistance–temperature coefficient
Silicon diodes, transistors, integrated circuits
Silicon carbide voltage-sensitive resistors (varistors)
Zinc sulphide electroluminescent devices

2 Basic Semiconductor Devices

To some degree, all semiconductor devices are affected by changes in conditions such as temperature, pressure, illumination, magnetic and electric fields. In many cases the effect of the change on one of the above quantities has a greater effect on the parameters of the device than do changes in other quantities. In the following, a number of materials used for the measurement of changes in physical quantities are discussed.

2.1 Thermistors

The name 'thermistor' is a contraction of THERMally sensitive resISTOR, the resistance changing over a wide range of values over the operating temperature range of the device.

There are two types of thermistor in popular use, namely those having a *negative resistance–temperature coefficient* (n.t.c.) whose resistance reduces with increasing temperature, and those with a *positive resistance–temperature coefficient* (p.t.c.) whose resistance increases with increasing temperature. Types with p.t.c. characteristics are further divided into two categories, namely those with *elemental* or *linear* characteristics and those with the so-called *switching* characteristics. The latter is the more widely used of the two p.t.c. types.

Typical characteristics for an n.t.c. device and for a switching type of p.t.c. device are shown in figure 2.1, together with their circuit symbols. In both cases, *n*-type semiconductor materials are used.

n.t.c. thermistors

These devices are manufactured from metallic oxides of cobalt, manganese and nickel, whose resistance reduces by a value in the range 1 to 6 per cent/°C. They are used for measurement of temperature in the range from about −60 °C to about +400 °C; in general, high resistance units (100 to 500 kΩ) are used for the measurement of high values of temperature, and low resistance units (100 Ω to 1 kΩ) are used at the lowest temperatures.

Thermistors are manufactured in rod, bead and disc forms, illustrated in figure

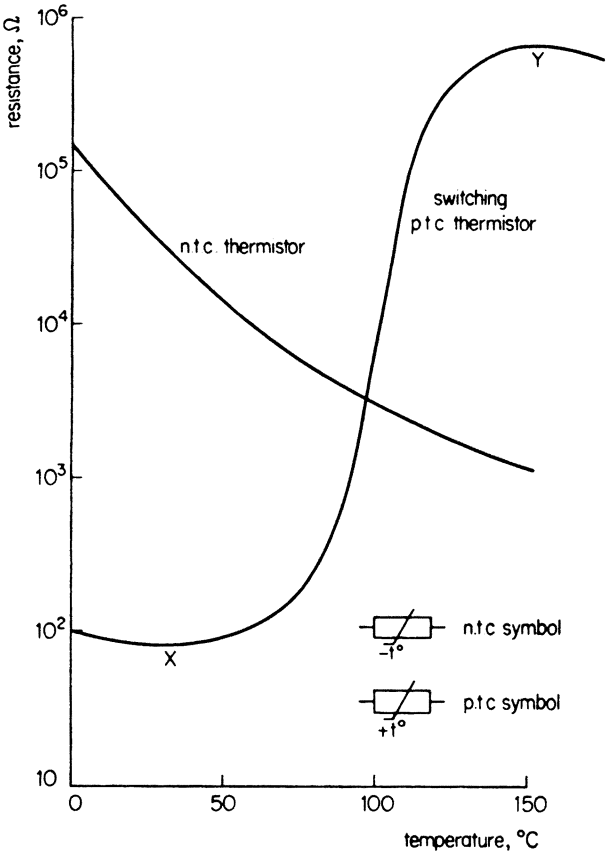


Figure 2.1 Thermistor characteristics

2.2, as well as a number of less usual shapes. In some applications thermistors are deliberately heated either directly by their own current, or indirectly by a heating element; in these cases the thermistor is sensitive to the rate of heat loss to its surroundings. This fact is utilised in instruments used to measure fluid flow rates and liquid levels.

p.t.c. thermistors

The substance barium titinate, when doped with small quantities of such elements as antimony, bismuth, cerium, lanthanum or niobium, exhibits an n.t.c. characteristic in the range between −200 °C and about +40 °C. This corresponds to temperatures below point X on the p.t.c. characteristic in figure 2.1. The latter temperature is the *ferroelectric Curie temperature* of the material, at which point electric polarisation is lost. As the temperature is increased through and beyond this

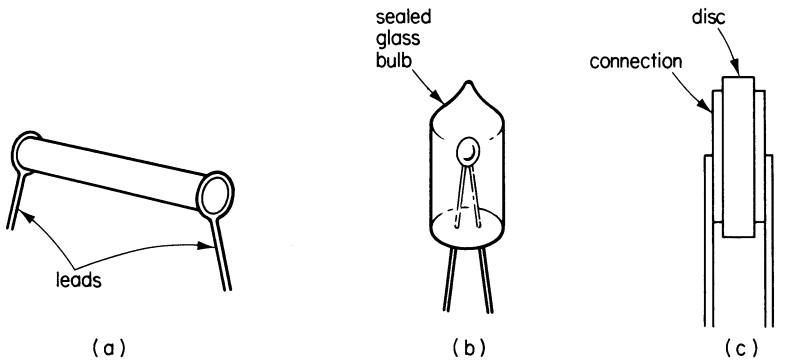


Figure 2.2 Thermistors: (a) rod type, (b) encapsulated bead type and (c) side elevation of a disc type

value, the resistance of the material increases very rapidly with temperature, and resistance changes of 70 per cent/°C over a very limited temperature range have been recorded. At a higher value of temperature, the thermistor reverts to an n.t.c.-type characteristic, but at a much higher value of resistance. This corresponds to temperatures above point Y on the p.t.c. characteristic in figure 2.1.

Devices with the p.t.c. characteristic of the type in figure 2.1 are known as *switching p.t.c. thermistors*, and are characterised by the sudden increase in resistance at the Curie temperature. By adding various substances the Curie point is shifted, allowing the switching temperature to lie in a range from about +60 °C to +180 °C.

Another type of p.t.c. thermistor, known as the *elemental type* or *linear type*, uses heavily doped silicon or germanium, and has a p.t.c. characteristic in the temperature range −65 to +150 °C. These devices are heavily doped with impurities, so that the mobile charge carrier concentration is nearly independent of temperature. The increase in resistance with temperature results from the reduction in charge carrier mobility with increase in temperature (see also section 1.6). The increase in resistance with temperature is relatively small, being about 0.7 per cent/°C.

Applications of thermistors

There are four main areas of application, namely

Temperature sensing
Environmental sensing (other than temperature)
Current control
Power control

Thermistors used for temperature sensing are either of the n.t.c. type or of the elemental p.t.c. type. The sudden change in resistance of the switching p.t.c. type makes them unsuitable for temperature measurement over a wide range of temperature. The most common application of the elemental p.t.c. type is for temperature compensation of semiconductor circuitry.

Environmental conditions such as flow, pressure, level, etc., can be measured by heated thermistors, in the manner described earlier.

Popular applications of p.t.c. switching thermistors are in the field of electric current control. Electric motor overheating protection is provided by embedding p.t.c. switching thermistors in the stationary part of the motor, and they are connected in series with a relay whose contacts are in the contactor coil circuit which supplies power to the machine. When the motor becomes overheated, the resistances of the thermistors switch to the high value, so causing the motor power supply to be disconnected. Another popular use of p.t.c. switching thermistors is in the degaussing or demagnetisation of colour television tubes. In order to maintain correct colour registration, the tube must be degaussed at frequent intervals; a convenient time to do this is each time the set is switched on. To this end, a p.t.c. thermistor is connected in series with the degaussing winding on the tube, and when the set is swtiched on a high value of alternating current flows through both the coils and the thermistor. When the thermistor heats up to its Curie point, it reduces the coil current to a very low value.

Thermistors are also used in circuits for the measurement of low values of power at radio frequencies up to about 10 GHz (1 GHz = 10^9 Hz). Other applications in electronics include automatic gain control, communications applications, modulators and phase shifters.

2.2 Voltage-dependent Resistors

Voltage-dependent resistors, known as *varistors* and *v.d.r.s*, are devices whose resistance reduces as the applied voltage is increased.

Varistors are manufactured from silicon carbide, and it is thought that their voltage dependence is due to the contact resistance between the crystals which forms a complex series–parallel path. A typical voltage–current characteristic is shown in figure 2.3, the relationship between the quantities being given by

$$V = aI^n$$

where V and I are the voltage across and the current through the varistor, respectively. The values of parameters a and n depend on the composition of the varistor and on its construction. The value of a is usually defined as the voltage across the device at a current of 1 A; it has a value in the range 10 V to 1000 V for disc devices and up to 4500 V for rod devices. Values of n lie in the range 0.12 to 0.25.

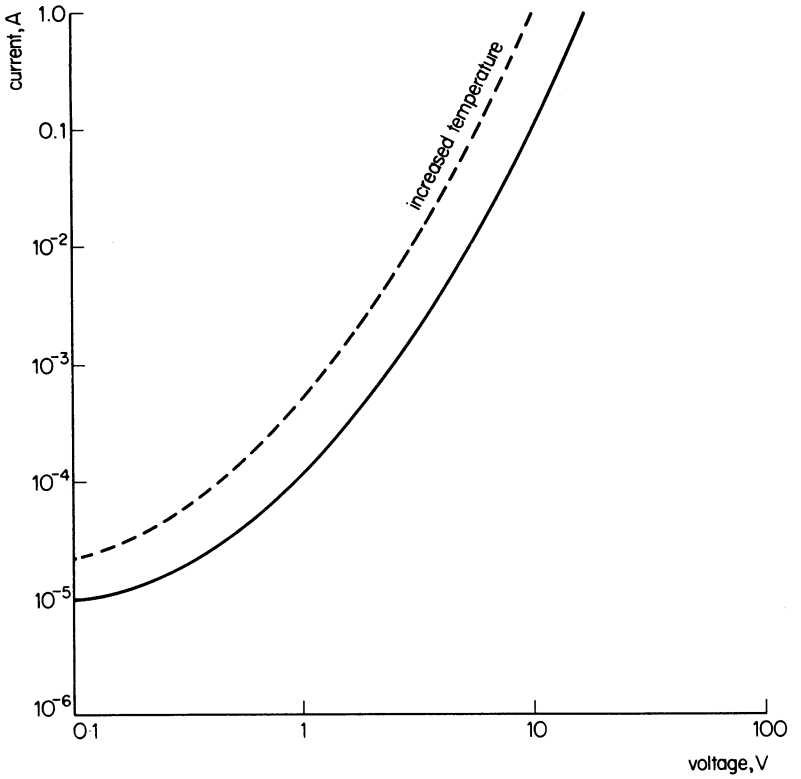


Figure 2.3 Voltage–current characteristic of a varistor

Applications of varistors

Applications include transient voltage suppression, voltage stabilisation and switch contact protection. The former is a popular application, and is achieved by connecting the varistor across the apparatus to be protected; the varistor is selected so that it draws only a small value of current at the normal operating voltage. When a voltage surge develops across the apparatus, the varistor resistance reduces in value and absorbs some of the energy in the transient. This has the effect of reducing the amplitude of the voltage transient. Varistors are used in this way to provide switch contact protection.

2.3 Hall-effect Devices

In 1879, when experimenting with current flowing in a strip of gold leaf, E. M. Hall discovered that charge carriers could be deflected by a magnetic field, in much the same way as electrons are magnetically deflected in a cathode-ray tube.

The principle is illustrated in figure 2.4a for an *n*-type semiconductor material.

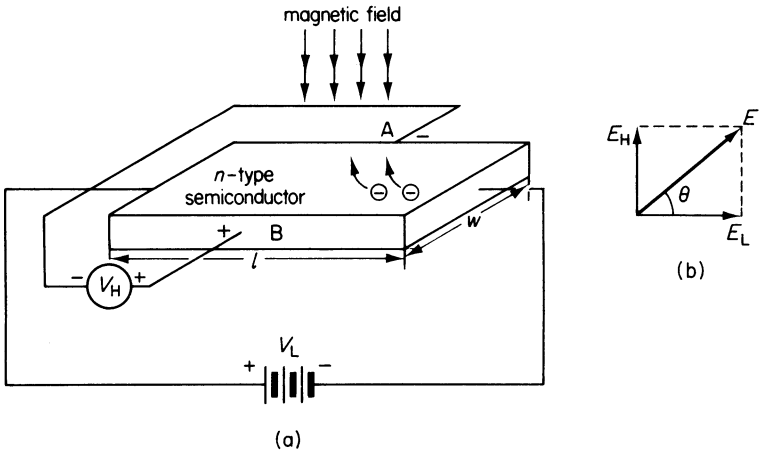


Figure 2.4 (a) The Hall effect in an *n*-type semiconductor and (b) the electric fields in the semiconductor

In *n*-type material, electrons are the majority charge carriers, and when they are subjected to electrical and magnetic fields in the manner shown in the figure, the electrons are swept towards electrode A on one side-face and away from electrode B. Consequently the polarity of electrode A is negative with respect to electrode B. In the case of a *p*-type material, with the same directions of electric and magnetic fields shown in figure 2.4a, the majority charge carriers (holes) are swept towards electrode A, so that it has a positive potential with respect to electrode B.

The angle θ through which the charge carriers are twisted is known as the *Hall angle*, and its value can be calculated from figure 2.4b in which E_L is the longitudinal electric field strength and is equal to V_L/l, and E_H is the Hall-effect electric field strength and is given by V_H/w, where V_H is known as the *Hall voltage*. Two forces act on the electron at right angles to its direction of motion. These are the forces due (1) to the magnetic field, and (2) to the electrons that are already crowded together on face A. The force on an electron due to the magnetic field *towards* electrode A is Bev, where B is the magnetic flux density, e is the charge on the electron, and v is the drift velocity of the electron. From the work in section 1.6 the drift velocity is given by $v = \mu E_L$, where μ is the mobility of the charge carrier. The force on the electron due to the electric charge at electrode A is eE_H, and is in a direction *away* from electrode A. In equilibrium, the two forces just balance each other, or

$$eE_H = Bev$$

hence

$$E_H = Bv = B\mu E_L \tag{2.1}$$

Since $E_H = V_H/w$, then

$$V_H = B\mu w E_L \tag{2.2}$$

Equation 2.2 indicates that the value of the Hall voltage is a function of the mobility μ of the charge carriers. The value of the Hall voltage can be quite large, being up to 0.5 V in semiconductors compared with about 1 μV in metals.

Applications

Hall-effect devices can, from equation 2.1, be used in sensitive *magnetic field meters* or *magnetometers*, the flux density being given by

$$B = \frac{E_H}{\mu E_L}$$

In this application, a high value of Hall voltage is obtained by using a semiconductor with a high mobility. Consequently, indium antimonide (see table 1.2) is often the first choice for this application.

The Hall effect is also used in *the determination of charge carrier mobility* since, from equation 2.1

$$\mu = \frac{E_H}{BE_L}$$

Yet another application is in the field of analog computing. In an instrument known as a *Hall-effect multiplier*, the Hall voltage is proportional to the product of two other voltages; one of these voltages is applied longitudinally to produce the electric field E_L, and the other is used to produce the magnetic field B. From equation 2.2, V_H is proportional to the product of the two voltages.

2.4 Magnetoresistors

Magnetoresistors are devices whose resistance is controlled by means of a magnetic field. They are made from the emiconductor indium antimonide with needle-shaped inclusions of nickel santimonide. When a voltage is applied to the magnetoresistor in the absence of a magnetic field, the current takes the shortest path between the two ends. When the magnetoresistor is subjected to a magnetic field, the stream of current in the semiconductor is twisted at an angle to the shortest route in the manner described in the section on the Hall effect. The result is that the length of the conducting path increases with the magnetic field strength, and with it the resistance of the magnetoresistor increases.

2.5 Semiconductor Strain Gauges and Piezoresistors

In the first chapter it was stated that when atoms were close to one another, the forces binding electrons to the parent atoms were reduced, and the conductivity of the material increased when compared with the case when the atoms were far apart. Conversely, as the atomic centres separate, the conductivity of the material decreases. In the former case the size of the forbidden energy gap is decreased, and in the latter it is increased.

This is the simple principle of the semiconductor strain gauge. Strain gauges are small devices which are bonded to mechanical test-pieces and, when a mechanical force is applied to the test piece, the resulting change in resistance of the strain gauge is used to determine the strain in the test piece.

Strain gauges can be used to measure a wide range of quantities including strain, pressure and flow. A feature of silicon strain gauges is that, for a given value of mechanical strain, the change of resistance is some 50 to 100 times greater than in a conventional strain gauge.

A drawback of semiconductor strain gauges is that they are temperature-sensitive but, provided that the temperature is measured (using, say, a thermistor), then the thermal effects can be compensated for. In some circuits it is possible to use additional silicon strain gauges to provide automatic thermal compensation.

The feature of change of conductivity with mechanical strain is capitalised in some types of *strain-measuring transistor* (see chapter 4). In this case, the mechanical force is applied to the *base region* or control region of the transistor, and it causes the current in the *collector region* of the transistor to change. The change of current is calibrated in terms of change of the strain applied to the base region.

2.6 Thermoelectric Effect

It is found that when current flows across the junction of two dissimilar metals or across the junction of a metal and semiconductor, heat is either absorbed or liberated, depending on the junction. This is known as the *Peltier effect*. The reason for this effect can be explained in terms of the energy diagrams of the junction.

The energy-band diagram for an ohmic contact between a metal and *n*-type semiconductor in the absence of an electric field was shown in figure 1.10. It is found that when the *n*-type semiconductor is mounted as shown in figure 2.5a, the temperature of metal 1 is reduced and that of metal 2 is increased. The reason is as follows.

When metal 1 is connected to the negative pole of the battery, and metal 2 to the positive pole, the energy-band diagram for the semiconductor is twisted or tilted as shown in figure 2.5b. This is because the electrons at the left-hand end of the semiconductor are connected to the negative (high-energy) source, while the right-hand end is connected to the positive (low-energy) source. Due to the high conductivity of the two metals, the 'tilt' given to their energy-band diagrams is insignificant.

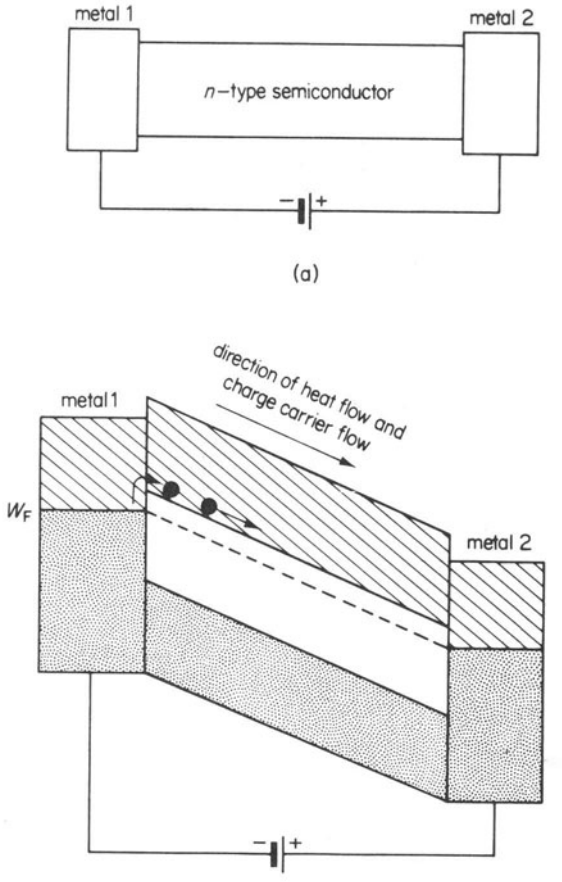


Figure 2.5 Energy-band model illustrating thermoelectric heat transfer

Focusing attention on the junction between metal 1 and the semiconductor, it is observed the when the Fermi levels are aligned, a small energy gap exists between the semiconductor conduction band and the top of the conduction band in the conductor. Electrons in metal 1 that have the greatest energy, that is, those that have absorbed thermal energy, can overcome this energy gap and enter the semiconductor. At this point they are subjected to the electric force towards the positive voltage connected to metal 2. Thus, the most energetic (or hottest) electrons are removed from metal 1 and are attracted to metal 2. In this way, metal 1 is cooled and metal 2 is heated.

In general, **heat flow is in the same direction as the flow of majority charge carriers in that material.** If the semiconductor in figure 2.5a were replaced by a

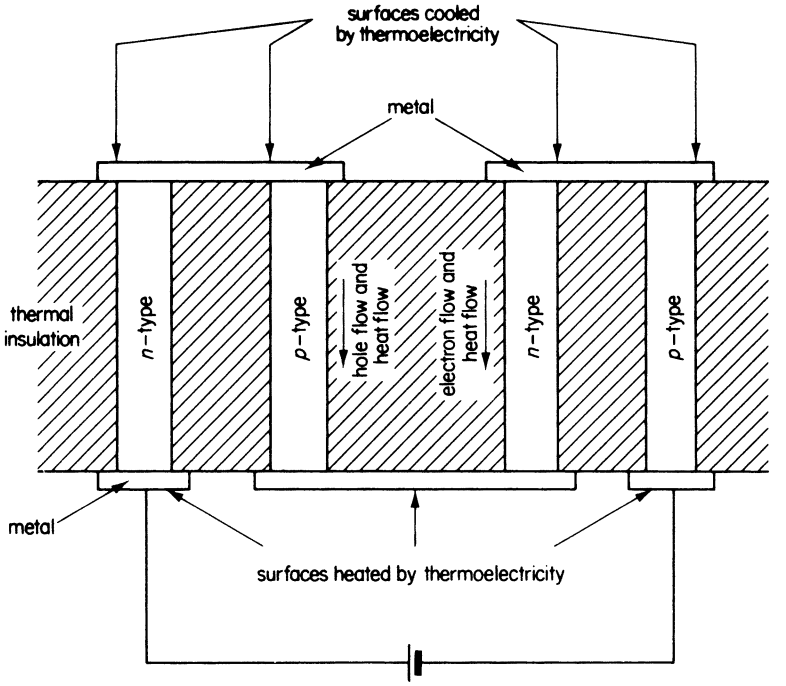


Figure 2.6 The basis of a thermoelectric heat-transfer device

p-type material, then the direction of heat flow would be from the positive terminal (metal 2) to the negative terminal (metal 1).

At low values of current, the cooling effect increases with current, but at some higher value of current, the self-heating effect of the current, that is, the I^2R heating effect, becomes greater than the cooling effect. Thus, there is an optimum value of current for each type of thermoelectric device.

The basis of one form of heat-transfer device is illustrated in figure 2.6. It may, for instance, be used to cool a refrigerator. Using alternate *p*- and *n*-types of semiconductor causes the direction of heat flow to be in a direction away from the upper surfaces.

Semiconductors used in thermoelectric applications must have a high electrical conductivity to allow the 'energetic' electrons to be conducted easily between the metal surfaces, yet must have low thermal conductivity in order to prevent heat 'leaking' back from the hot surface to the cold surface. Bismuth telluride is the most satisfactory semiconductor for this application. Temperature differences of 30 to 80 °C have been obtained in multicouple BiTe devices.

The thermoelectric effect can, alternatively, be used in reverse in *thermoelectric generators*. If, in figure 2.5a, the external battery is removed and metal 2 is heated, the more energetic electrons will move away from metal 2 towards metal 1,

establishing an e.m.f. between the two. Small generators of this kind have been used for operating radios and other portable equipment, and have been used on open fires and on fuel burners in many remote parts of the world.

2.7 The Gunn Effect

In 1963 J. B. Gunn discovered that microwaves were generated in *n*-type gallium arsenide when an electric field strength greater than a threshold value of about 3 kV/cm was set up in it. A similar effect is also produced in indium phosphide and in cadmium telluride. The length of the specimen used is small, so that this order of field strength is achieved using only a small value of voltage.

The operation of these devices depends on the fact that in the above-mentioned materials, electrons at levels above the valence energy band can lie in the conduction band or in a satellite band, the two levels differing in energy by only a fraction of an electron volt. A feature of the two energy levels is that electrons in the lower energy level have a higher mobility than those in the upper level. In a state of thermal equilibrium the majority of the electrons lie in the lower energy level, and the material has a high conductivity due to the high electron mobility. At high values of electric field strength in the material, electrons transfer to the higher-energy, lower-mobility level. This results in a reduction in the conductivity of the material, giving a negative resistance during the transition. At this time, small *domains* of high field strength travel through the material with a velocity of about 10^5 m/s. The frequency of oscillations is dependent on the transit time of the domains, and is typically 10 GHz for a specimen of length 10 μm. The Gunn effect is employed in IMPATT diodes (IMPact Avalanche and Transit Time).

2.8 Photoconductors

When light and other electromagnetic radiation falls on a semiconductor, its intrinsic conductivity increases; devices utilising this effect are known as *photoconductors, photoresistors* or *light-dependent resistors* (l.d.r.s). The increase in conductivity is due to the fact that some of the radiant energy is absorbed by the atomic structure, causing spontaneous generation of electron–hole pairs. The electrons so produced may either be generated in the valence band and transfer either to the conduction band (known as *intrinsic excitation*), or may transfer to an acceptor level (known as an *impurity excitation*), or may leave a donor level and transfer to the conduction band (also an impurity excitation). Photoconductivity, in fact, is principally due to intrinsic excitation from the valence band to the conduction band.

The *minimum value* of a 'bundle' or *quantum* of light energy that will cause an electron to be excited into the conduction band is equal to the value of the forbidden energy gap, W_G, of the semiconductor. The wavelength, λ_c, of a photon

Table 2.1 Values of E_G and λ_c of selected semiconductors

Material	Energy gap (eV)	λ_c (μm)	Spectral responses
Germanium	0.72	1.72	Infrared
Silicon	1.1	1.13	Infrared
Cadmium telluride	1.5	0.83	Infrared
Cadmium selenide	1.7	0.73	Infrared
Cadmium sulphide	2.5	0.49	Visible

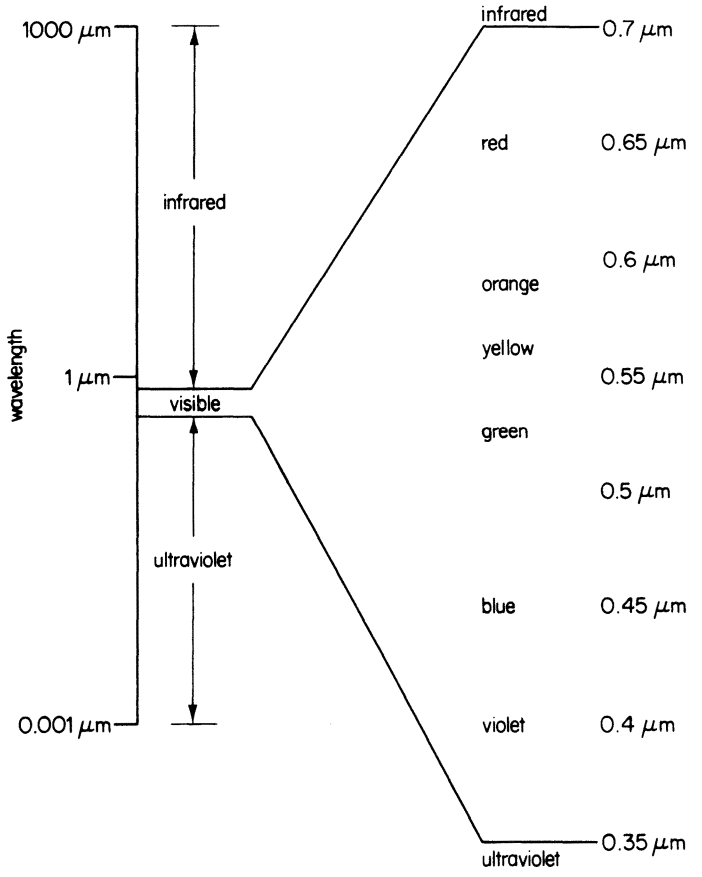


Figure 2.7 A section of the electromagnetic spectrum

of energy which corresponds to the forbidden energy gap is

$$\lambda_c = \frac{1.24}{W_G} \text{ micrometre } (\mu\text{m})$$

where $1\ \mu\text{m} = 10^{-6}$ m. Listed in table 2.1 are the forbidden energy gaps and the appropriate values of λ_c of a number of semiconductors.

A part of the electromagnetic spectrum is illustrated in figure 2.7, and it is seen that while the first four materials in table 2.1 are most sensitive to various

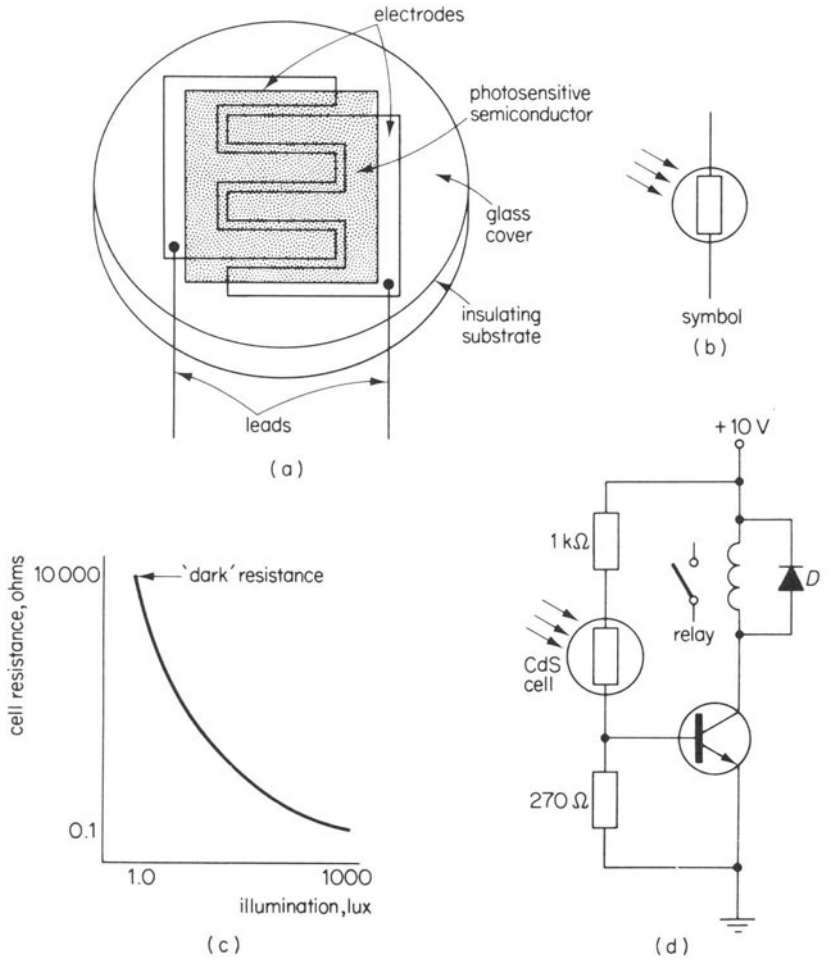


Figure 2.8 Cadmium sulphide cell: (a) one form of construction for end-on illumination, (b) circuit symbol, (c) a typical resistance–illumination curve and (d) a transistor circuit incorporating a CdS cell

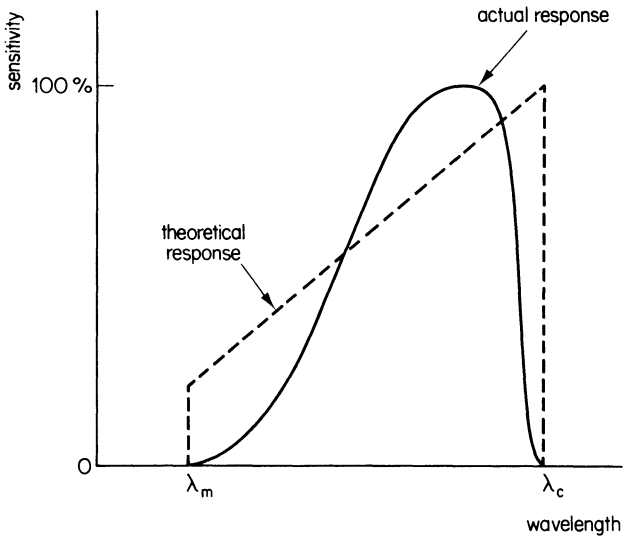


Figure 2.9 Spectral response curve for a photoconductor

wavelengths in the infrared part of the spectrum, only cadmium sulphide is responsive to visible radiation. The wavelength λ_c is the upper limit of the waveband to which the material is sensitive. The material is sensitive to wavelengths shorter than λ_c; its response to longer wavelengths rapidly diminishes to zero. Thus, silicon is much more responsive to a wavelength of 1.13 μm than it is to radiation in the visible spectrum.

For a particular material, frequencies with wavelengths greater than λ_c do not have sufficient energy to excite electrons into the conduction band, while wavelengths less than λ_c can excite electrons into it. There is a minimum wavelength, λ_m, to which the material is sensitive; the energy content of radiation of this wavelength causes electrons to be excited across the energy gap and through the conduction band, so that they escape from the surface of the material (*photoemission*).

The construction of one form of cadmium sulphide (CdS) cell is shown in figure 2.8a, the diameter being typically 1 to 2.5 cm. The cells are manufactured by sintering photoconductive cadmium sulphide into the required shape on to an insulating surface or *substrate*. Connections are made to the CdS by depositing metal electrodes, sometimes in the form of the comb configuration shown, on to it. The device is then protected by a glass or plastic cover. The circuit symbol is shown in figure 2.8b, and a typical resistance–illumination curve is shown in figure 2.8c. The ratio of 'dark' to 'light' resistance of the cell is about 1000:1, and cells may have a 'dark' resistance in the range 10^4 to 10^6 Ω.

A circuit using a CdS cell in a relay control amplifier is shown in figure 2.8d. When the CdS cell is illuminated, the transistor passes a current sufficiently large to

energise the relay. When the illumination level falls to a lower value, the transistor current falls and the relay is de-energised and its contacts open, Diode D is included in the circuit for protection purposes (see also chapter 4).

In the case of a photoconductive cell the ideal spectral response curve extends from zero wavelength up to the limiting value λ_c, the sensitivity to radiation increasing with wavelength. The theoretical response has a lower limit λ_m, described earlier, the theoretical response being illustrated in figure 2.9. The spectral response of an actual device fits between these limits and has the typical lop-sided shape shown. In the case of CdS cells, their spectral response closely matches the response of the human eye. Consequently CdS cells are frequently used in situations where humans can also judge illumination levels, for example, in street lighting schemes, camera exposure settings, etc. Photoconductive devices with frequency spectra in the infrared region, that is, CdSe and CdTe can be used where this feature is advantageous, for example in boiler flame-failure systems, in burglar alarms and in aircraft and missile tracking systems.

3 Semiconductor Diodes and the Unijunction Transistor

3.1 Basic Features of Diodes

A diode is a two-terminal device, one terminal being known as the *anode* and the other as the *cathode* (see figure 3.1a). The characteristic curve of an 'ideal' diode is shown in figure 3.1b. The characteristic of the diode shows that when the anode is positive with respect to the cathode, it functions as though it were a switch whose contacts are closed; in this state the p.d. across the ideal diode is zero for all values of current, and the diode is said either to be in its *forward biased mode* or in its *forward conducting mode*. When the anode is negative with respect to the cathode, the diode functions as though it were a switch whose contacts are open; in this state no current flows through the diode for all values of voltage, and it is said either to be in its *reverse biased mode* or in its *reverse blocking mode*.

The diode may thus be regarded as a *voltage-sensitive electronic switch*, which is closed or is ON when the anode polarity is positive with respect to the cathode, and is open or is OFF for the reverse anode-to-cathode potentials.

Practical diodes come in many forms, the most popular types having a single p–n junction. The characteristic of a practical device has the same general form as that in figure 3.1b, but with the difference that the p.d. across the device when conducting is finite and, when reverse biased, a leakage current flows through it.

3.2 The p–n Junction Diode

The operation of the p–n junction diode can be explained in terms of the energy-band structure of semiconductors. It was shown in section 1.10 that the energy-band structure of isolated p- and n-type semiconductors is as illustrated in figure 3.2a.

During the manufacture of a semiconductor diode, the impurity dopant is changed to form both p- and n-type materials within the same crystal. This results in a p–n junction being formed at the interface between the two types of semiconductor. It was also shown in section 1.15 that, at the interface between two

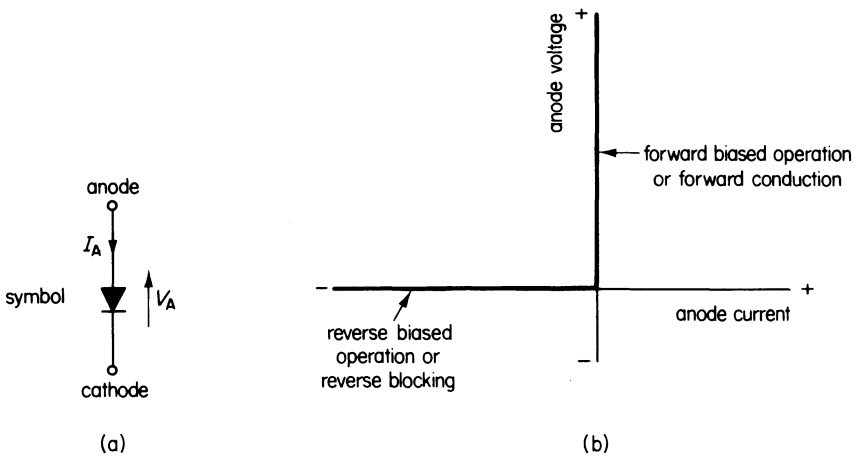


Figure 3.1 (a) Diode circuit symbol and (b) the static anode characteristic of an ideal diode

materials, the Fermi energy levels align. For this to occur, there is a net transfer of charge between the two regions. Initially there is a transfer of mobile electrons from the conduction band of the *n*-type material into the *p*-type and, at the same time, a transfer of mobile holes from the valence band of the *p*-type material into the *n*-type material. Thus the *p*-type material acquires a negative potential with respect to the *n*-type. Since the mobile charge carriers in the junction region have transferred to adjacent regions of the device, the junction is depleted of charge carriers; the junction region is therefore described as a *depletion region* or *depletion layer*. Other names used to describe this region are *transition region* and *space-charge region*. The thickness of this region is about that of one wavelength of visible light, or about 0.5 μm, and within this region there are no mobile charge carriers.

When equilibrium conditions are reached, the Fermi energy levels of the two materials have a common value W_F (see also section 1.15), as shown in figure 3.2b. As readers will recall, a negatively charged material has a higher electronic energy than one which is positively charged; hence the energy bands of the *p*-type material (having acquired a more negative potential than those of the *n*-region) are raised to a higher level on the diagram. Equilibrium is reached when the negative potential acquired by the *p*-region is sufficiently great to prevent further migration of electrons to it; at the same time, the positive potential acquired by the *n*-region prevents further migration of holes from the *p*-region to the *n*-region. Hence, under equilibrium conditions a 'potential hill' is formed between the two regions. The potential hill described above is known as the *contact potential* or *barrier potential*, E_0, and has a value of approximately 0.3 V in germanium and 0.7 V in silicon. Clearly the width of the junction required to build up E_0 depends on the doping

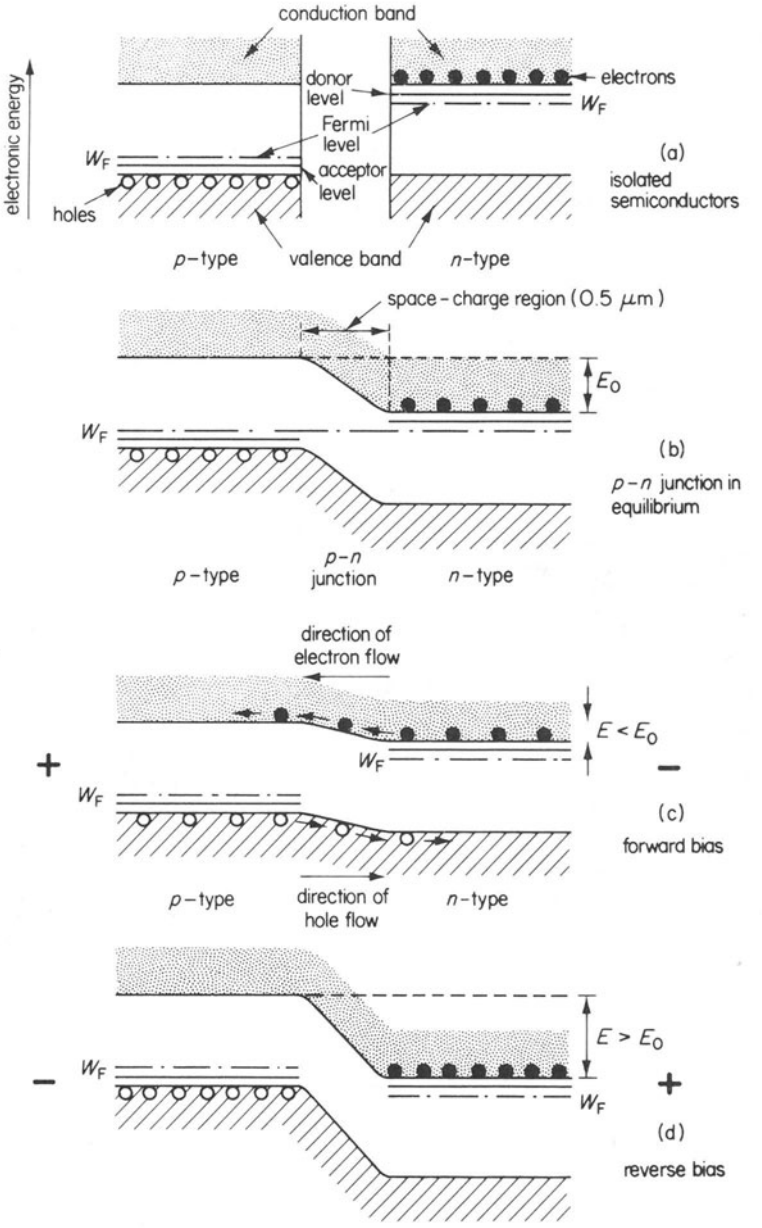


Figure 3.2 Energy-band model of a p–n junction: (a) isolated semiconductors, (b) p–n junction in equilibrium, (c) forward bias and (d) reverse bias

concentration: a heavily doped semiconductor builds up the contact potential by depleting only a very thin layer of charge carriers, resulting in a narrow depletion region. This fact is of importance in the study of the reverse breakdown mechanism of p–n junction diodes (see section 3.5).

A diode becomes *forward biased* when the p-region (the anode) is connected to the positive pole of a power supply, and when the n-region is connected to the negative pole. This externally applied potential alters the energy-band diagram of the diode in the manner shown in figure 3.2c, the energy bands of the p-region being depressed and the bands of the n-region being raised, so reducing the height of the potential hill. This allows energetic charge carriers to move through the diode as shown.

The diode is *reverse biased* when the p-region of the device is connected to the negative pole of the supply, and the n-region is connected to the positive pole. This has the effect of increasing the height of the potential hill, as shown in figure 3.2d. Under reverse bias conditions, even the most energetic mobile charge carriers cannot climb the potential hill, and no current flows through the diode.

3.3 Characteristics of a Practical p–n Junction Diode

A simple test circuit that may be used to determine the static characteristic of a diode is shown in figure 3.3. With the connections shown the diode is forward biased, having its anode connected to the positive pole of the supply. In this mode, the anode current increases rapidly as the 'forward' p.d. across the diode is increased. Typical forward biased characteristics for silicon and germanium diodes are shown in the first quadrant of figure 3.4.

A rise in the ambient temperature results in an increase in the intrinsic conductivity of semiconductor materials. The net result of a change in ambient temperature is illustrated in the case of a silicon diode in figure 3.4. The anode current at point A on the characteristic for a temperature of 25 °C is 24 mA, but at the same value of anode voltage and at a temperature of 75 °C (point B) the forward current is 44 mA. Hence the net result of an increase in temperature is an

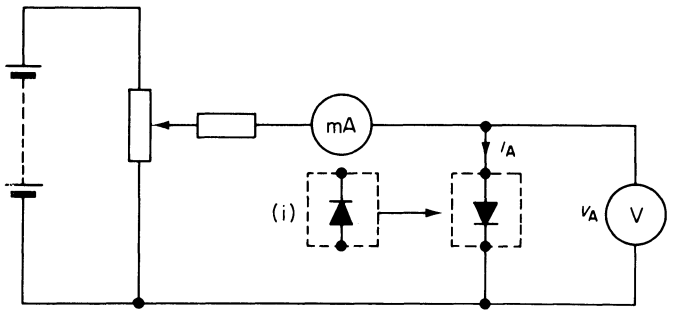


Figure 3.3 Typical test circuit used to determine diode parameters

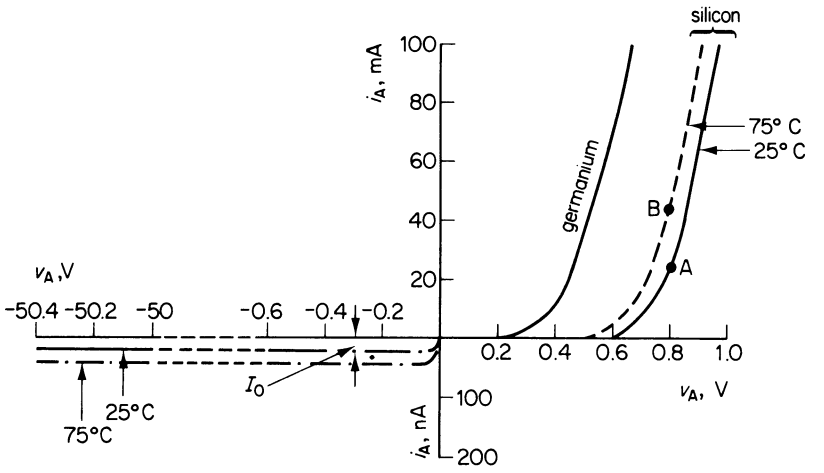


Figure 3.4 Diode characteristics; note the change in scale for the reverse current

upward shift of the forward conducting characteristic. At a constant value of anode current, this shift is equivalent to a reduction of the forward p.d. across the diode; for both silicon and germanium the forward p.d. falls at the rate of about 2.5 mV/°C rise.

When using the circuit in figure 3.3, readers should note that the anode voltage of the diode should preferably be measured using an electronic voltmeter. The reason for this is that the milliameter reads the sum of the diode current and the current drawn by the meter. When the diode is passing a very low value of current, the current drawn by a conventional moving-coil voltmeter may be comparable with (or even greater than) the diode current.

The reverse biased characteristic of the diode is obtained by reversing the connections to the diode, as indicated in inset (i) of figure 3.3. A typical characteristic for a silicon diode obtained from this test circuit is shown in the third quadrant of figure 3.4. The *reverse leakage current*, I_O, of a silicon diode may typically lie in the range 2 to 20 nA (1 nA = 10^{-9} A) and, for most practical applications represents an open circuit. This leakage current is also known as the *reverse saturation current*; it is due not only to thermally generated electron–hole pairs in the junction region, but also to leakage current across the surface of the diode. Due to the lower value of the forbidden energy gap in germanium when compared with that of silicon, the leakage current in germanium diodes has a higher value than that quoted above, and is typically in the range 2 to 20 μA. An increase in the operating temperature of the diode results in increased generation of electron–hole pairs in the junction region and, with it, the leakage current increases. As a guide, **the leakage current doubles in value for each 10 °C rise in temperature.**

3.4 Diode Resistance

When forward biased to point X in figure 3.5, the *instantaneous anode resistance* or the *static resistance*, r_A of the diode is defined as

$$r_A = \frac{\delta v_{A1}}{\delta i_{A1}}$$

This is the resistance 'seen' by the d.c. signal source applied to the diode. Its value varies with the value of the anode current, a typical range of values being 10 to 80 Ω.

In many applications the *slope resistance*, r_a (also known as the *dynamic resistance* or *incremental resistance*) is of importance. Its value is given by the reciprocal of the slope of the characteristic curve at its *operating point* or *quiescent point*, that is, at point X in figure 3.5. Hence

$$r_a = \frac{\delta v_{A2}}{\delta i_{A2}}$$

The value of r_a is generally lower than that of r_A, a typical range of values being 1 to 3 Ω. Although the value of R_a varies a little with anode current, it is convenient to assume in small-signal models of the diode that the value of this parameter is constant.

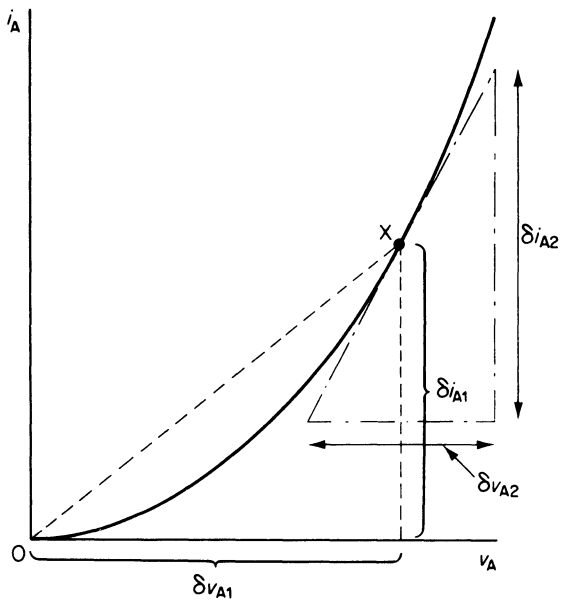


Figure 3.5 Evaluating r_A and r_a

3.5 Reverse Breakdown and Zener Diodes

If the reverse bias applied to a p–n junction diode is increased, a point will be reached at which the junction breaks down and the current flowing in the reverse direction rapidly increases in value (see figure 3.6). The value of reverse voltage at which this occurs, and the breakdown mechanism involved, depends on the construction of the diode. In a conventional rectifier diode, reverse breakdown should not occur within the voltage rating of the diode (which may be several hundred volts); when reverse breakdown occurs in rectifier diodes it is usually catastrophic, and the diode is destroyed.

Diodes which have a 'normal' impurity doping concentration, also have a relatively wide depletion layer between the p- and n-regions. In these devices (which include rectifier diodes), reverse breakdown is due to what is known as *avalanche breakdown*, and is described in the following. Under reverse bias conditions, thermally generated electrons and holes in the junction region are accelerated towards the cathode and the anode, respectively, by the applied voltage. These charge carriers represent the normal leakage current, and as they move through the junction they collide with other atoms in the structure. As the value of the reverse bias voltage is increased, the energy acquired by these charge carriers increases. At a reverse bias equal to the breakdown voltage, a point is reached where their energy is sufficient to dislodge electrons from atoms within the depletion region. This results in a spontaneous generation of electron–hole pairs, and the process escalates to a point at which there is a rapid transition from a reverse blocking state to reverse conduction, as shown in figure 3.6. The avalanche breakdown mechanism predominates in diodes having reverse breakdown voltages above about 8 V. As

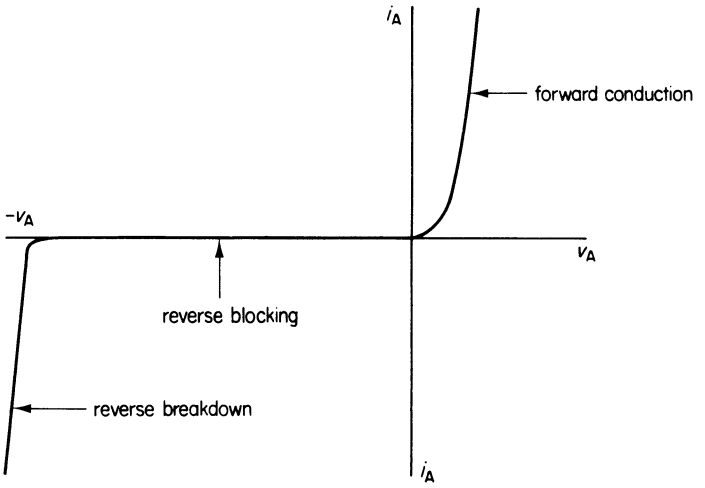


Figure 3.6 Reverse breakdown on a diode characteristic

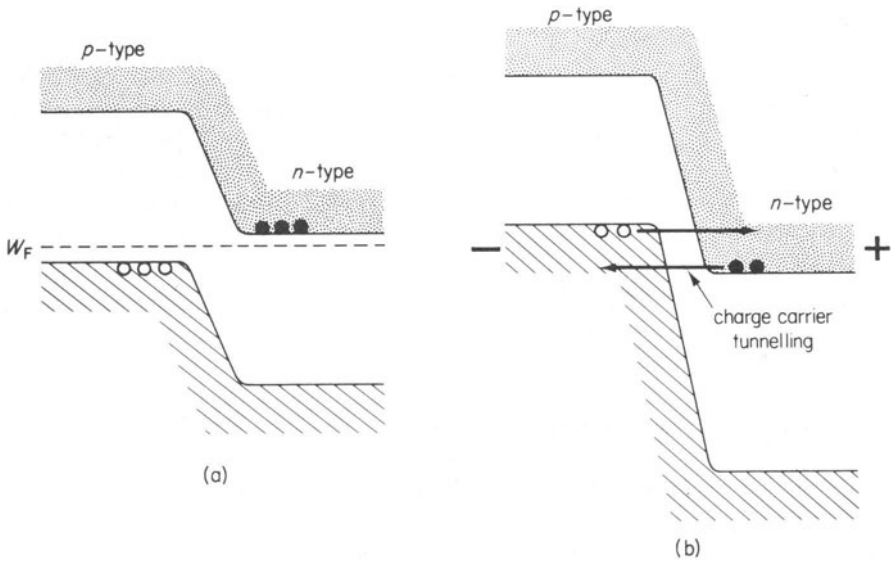


Figure 3.7 The tunnelling phenomenon in Zener diodes: (a) zero bias and (b) reverse bias.

illustrated in figure 3.6, this type of breakdown is typified by a sharp 'knee' on the characteristic (see also figure 3.8).

Where the impurity doping is above that which is 'normal' in diodes, the width of the depletion region necessary to build up the contact potential (see also figure 3.2) between the two regions is very small. In devices having a high impurity concentration, reverse breakdown occurs below about 3 V, and is known as *Zener breakdown* (after the scientist who first described the mechanism). The principle of this breakdown mechanism is best explained in terms of the energy-band diagram in figure 3.7. In 1934 C. Zener predicted that if the transition from one type of impurity to another were sufficiently abrupt then, under reverse bias conditions, charge carriers could be drawn through the narrow depletion region without collisions occurring between the charge carriers and molecules. This mechanism is illustrated in figure 3.7b, in which the charge carriers effectively *tunnel* through the depletion layer. Zener breakdown is typified by a smooth 'knee' on the characteristic curve (see figure 3.8).

When reverse breakdown occurs at voltages in the range 3 to 8 V, then both Zener and avalanche breakdown mechanisms are involved.

It is customary to describe diodes that are continuously operated in the reverse breakdown mode as *Zener diodes*, even though the actual breakdown mechanism may be of the avalanche type. Breakdown voltages of commercially available diodes range from about 1 to 1000 V.

Thermal effects on the reverse breakdown voltage

As with all semiconductor devices, the parameters and characteristics of Zener diodes change with temperature. In general the breakdown voltage of the Zener mechanism *reduces* with increasing temperature (see figure 3.8), and the breakdown voltage of the avalanche mechanism *increases* with temperature. The voltage–temperature coefficient of reverse breakdown diodes is generally about 0.1 per cent/°C, or about 1 mV for each volt of nominal breakdown voltage per °C change in temperature.

Between the 'pure' Zener and the 'pure' avalanche types of characteristic, and at a particular value of current, there is a characteristic that has a voltage–temperature coefficient of zero. Diodes with reverse breakdown voltages in the range 5 to 6 V exhibit this type of characteristic.

A qualitative explanation of the reason for the effects of thermal change is now given. An increase in ambient temperature causes an increased number of charge carriers to be generated. In diodes having a Zener breakdown mechanism the depletion layer is very narrow, and a lower p.d. is required to cause a given number of charge carriers to tunnel through it. This results in the breakdown voltage reducing with increasing temperature.

In avalanche breakdown devices an increase in temperature increases the amplitude of the vibrations of the atoms in the structure. At elevated temperatures, there is thus an increase in the probability of collision between the charge carriers

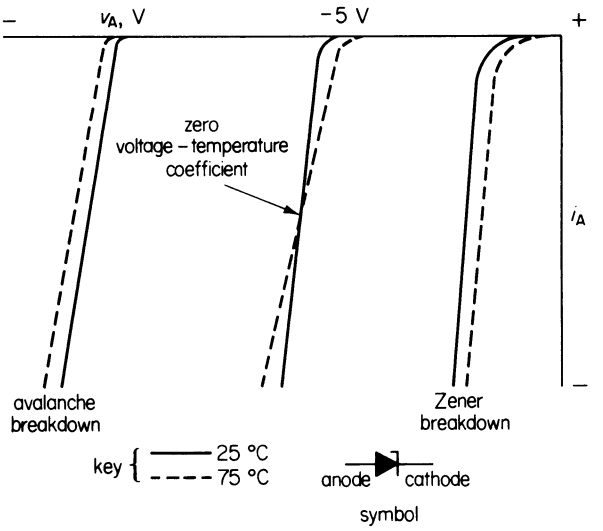


Figure 3.8 The effect of temperature change on the breakdown characteristics of diodes

and the atoms in the depletion region. Arising from this, these charge carriers need to be accelerated by a higher value of voltage before they can acquire sufficient energy to cause reverse breakdown. Thus the breakdown voltage of avalanche breakdown devices increases with increasing temperature.

Slope resistance or dynamic resistance of Zener diodes

A parameter of interest in Zener diodes is their slope resistance (or dynamic reverse conducting resistance) in the reverse breakdown mode. For any given family of Zener diodes, the dynamic resistance is usually a minimum for breakdown voltages in the range 5 to 10 V. This minimum value of resistance varies between types of diodes, and may be in the range 1 to 15 Ω; for values of breakdown voltage which are either greater than 10 V or less than 5 V, the dynamic resistance increases up to a value of several hundred ohms. The dynamic resistance becomes very large at low values of current (at less than, say, 1 mA).

Reference voltage sources

A reference voltage source is a stable voltage source, usually having a low output resistance. A simple form of reference voltage source which uses a reverse biased Zener diode is shown in figure 3.9.

A brief description of the circuit, together with a simplified analysis is given below; a fuller analysis follows. Assuming that the diode is a 'perfect' device exhibiting no leakage current below a reverse bias of V_Z, above which it conducts with zero dynamic resistance, then the output voltage, V_O, is equal to the breakdown voltage of the Zener diode. That is

$$V_O = V_Z$$

Moreover, since the dynamic resistance of the Zener diode is zero, then the output resistance, R_O, of the circuit is also zero.

The circuit is supplied by an unstabilised power source of voltage V_S, which may vary in value. The function of the voltage reference source is to provide a steady

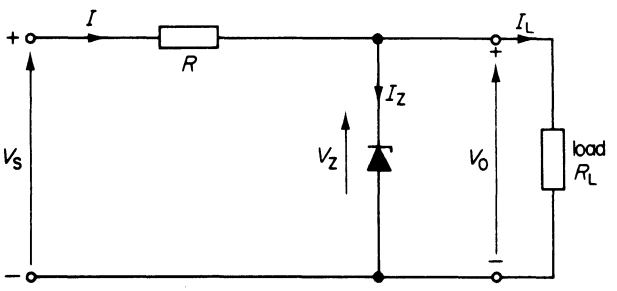


Figure 3.9 A simple voltage reference source

value of output voltage, even though V_S varies; in the circuit shown, any variations in the value of the unstabilised supply voltage are absorbed by the current-limiting resistor R, and are not transmitted to the output terminals. Another function of resistor R is to limit the value of current I flowing through it to the maximum allowable value of Zener diode current, $I_{Z(max)}$; the reason for this restriction in value is explained below. When a load resistor is connected, current I divides between the Zener diode and the load; should the load become disconnected, then current I must flow into the diode and it is for this reason that it is inadvisable to allow I to exceed the diode current rating. If the power rating of the diode is P_Z watts, then

$$V_Z I_{Z(max)} = P_Z$$

or

$$I_{Z(max)} = \frac{P_Z}{V_Z}$$

The value of I is calculated from the expression

$$I = \frac{V_S - V_Z}{R}$$

Combining the above equations gives

$$\frac{P_Z}{V_Z} = \frac{V_S - V_Z}{R}$$

or

$$R = \frac{V_Z(V_S - V_Z)}{P_Z}$$

and the rating of resistor R is

$$I^2R = \frac{P_Z(V_S - V_Z)}{V_Z}$$

For example, if V_S = 10 V, V_Z = 4.7 V and P_Z = 0.4 W, then

$$I_{Z(max)} = \frac{P_Z}{V_Z} = \frac{0.4}{4.7} = 0.085 \text{ A} = 85 \text{ mA}$$

Assuming that I is limited to this value, then

$$R = \frac{V_S - V_Z}{I_{Z(max)}} = \frac{10 - 4.7}{0.085} = 62.4\ \Omega$$

and the power rating of the resistor is

$$I^2R = 0.085^2 \times 62.4 = 0.45 \text{ W}$$

A 68 Ω, 10 per cent tolerance preferred resistor value with a 1 W rating would be suitable for this application. It is advisable to use a 1 W rated resistor rather than a 0.5 W type, since it may have to dissipate 0.45 W continuously. Moreover, should the value of the supply voltage rise above about 10.3 V, then the power dissipated by R would exceed 0.5 W.

The minimum value of load resistance, R_L, that may be connected to the circuit, is computed from the fact that the circuit ceases to function as a regulator when the current through the diode is reduced to zero. In practice it is inadvisable to allow the current through the diode to fall below about 5 mA. In the case of the regulator designed above, the maximum value of load current should be limited to about (85 – 5) mA = 80 mA. Hence the minimum value of load resistance, $R_{L(min)}$, which may be connected to the circuit is

$$R_{L(min)} = \frac{V_Z}{I_{L(max)}} = 4.7/80 \times 10^{-3} = 59\ \Omega$$

Full analysis A detailed analysis is carried out using the circuit in figure 3.10, in which the Zener diode is replaced by an equivalent electrical circuit; in this circuit, r_Z is the dynamic resistance of the diode and V_Z is the breakdown voltage of the diode. The input voltage applied to the circuit is related to the p.d.s and e.m.f.s in the circuit by the following expression

$$V_S = \left(\frac{r_Z + R}{r_Z}\right) V_O - \frac{R}{r_Z} V_Z + I_L R$$

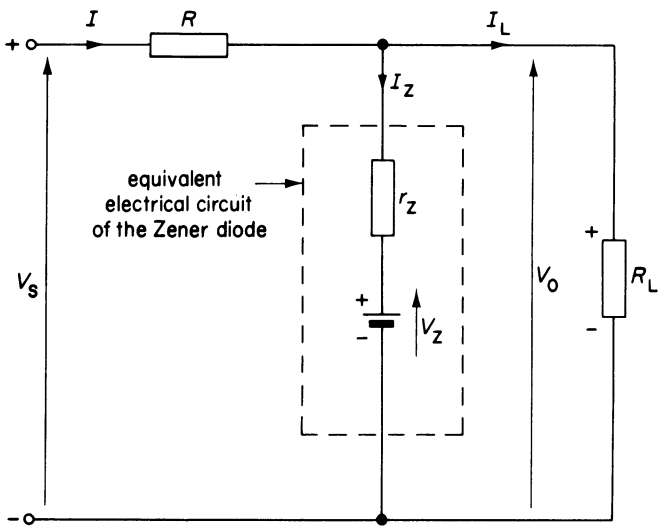


Figure 3.10 Analysis of the voltage reference source in figure 3.9

The *output resistance*, R_O, of the circuit is obtained from the expression

$$R_O = -\frac{\partial V_O}{\partial I_L} = \frac{r_Z R}{r_Z + R}$$

A useful parameter of the circuit is the *stabilisation factor*, S, which indicates the degree of stabilisation of output voltage offered against supply voltage variation. An expression for this factor is derived below.

$$S = \frac{\text{change in output voltage}}{\text{change in supply voltage}} = \frac{\partial V_O}{\partial V_S}$$

$$= 1 \Big/ \left(\frac{r_Z + R}{r_Z} + \frac{R}{R_L} \right)$$

If $R = 68\ \Omega$, $R_L = 1\ \text{k}\Omega$ and $r_Z = 50\ \Omega$, then the output resistance of the circuit is

$$R_O = \frac{r_Z R}{r_Z + R} = 50 \times \frac{68}{50 + 68} = 28.8\ \Omega$$

and the stabilisation factor of the circuit is

$$S = 1 \Big/ \left(\frac{50 + 68}{50} + \frac{68}{1000} \right) = 0.41$$

The value of the stabilisation factor indicates that the output voltage changes by 0.41 V for each volt change at the input of the circuit. Its value can be reduced by using a Zener diode with a lower value of dynamic resistance. For example, if $r_Z = 1\ \Omega$ in the above case, then the value of S would be 0.015, and the output voltage would only change by 15 mV per volt change at the input. This clearly demonstrates the need for Zener diodes with low values of dynamic resistance in this type of application. Typical values of R_O and S for a typical high-quality stabilised power supply are 0.002 Ω and 0.0002, respectively.

Maximum allowable power dissipation, P_{tot}, of the Zener diode

Another important parameter of the Zener diode is its maximum allowable power dissipation, P_{tot}. This parameter can be shown on the reverse characteristic of diodes in the form of a curve (see figure 3.11), where

$$P_{tot} = v_A i_A \text{ watts}$$

where v_A and i_A are the total instantaneous values of anode voltage and current, respectively. Thus if the value of P_{tot} is 1 W, then a curve for the expression $v_A i_A = 1$ can be drawn. For example, when $v_A = -7.5$ V (point L on the P_{tot} curve in figure 3.11), then $i_A = -1/7.5\ \text{A} = -133$ mA. The curve also passes through points M and N at voltages of −10 and −12.5 V, giving respective currents of −100 and −80 mA. The P_{tot} curve is a dividing line between the normal or 'safe' operating region and the overload region. The intersection of the P_{tot} curve with the diode

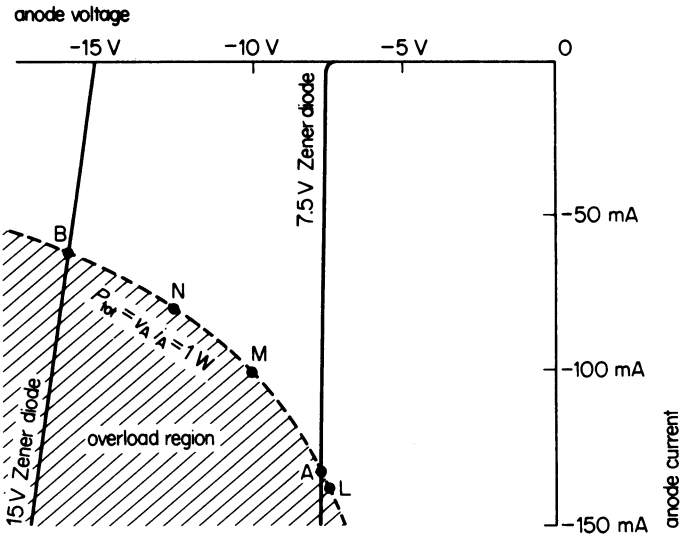


Figure 3.11 P_{tot} dissipation curve

characteristic gives the maximum steady operating current of the diode; thus the 7.5 V Zener diode with the characteristic shown in figure 3.11 should not be allowed to handle a mean current greater than 130 mA.

3.6 Tunnel Diodes

If the impurity doping level is increased beyond that used in Zener diodes, the reverse breakdown voltage progressively reduces until it reaches zero. The impurity doping level can be increased further until it reaches the limit of solubility of impurities in the semiconductor material. This occurs when about 1 atom in every 10 000 has been replaced by an impurity atom, when the breakdown voltage occurs at a small value of forward voltage.

The general shape of the characteristic is shown in figure 3.12a in which

V_P = peak-point voltage

I_P = peak-point current

V_V = valley-point voltage

I_V = valley-point current

V_F = peak forward voltage

The high doping levels cause the width of the depletion layer to be reduced to less than 0.1 μm (or about one-fifth of the wavelength of visible light), so that charge carrier tunnelling can take place at zero bias (see the energy-band diagram in figure

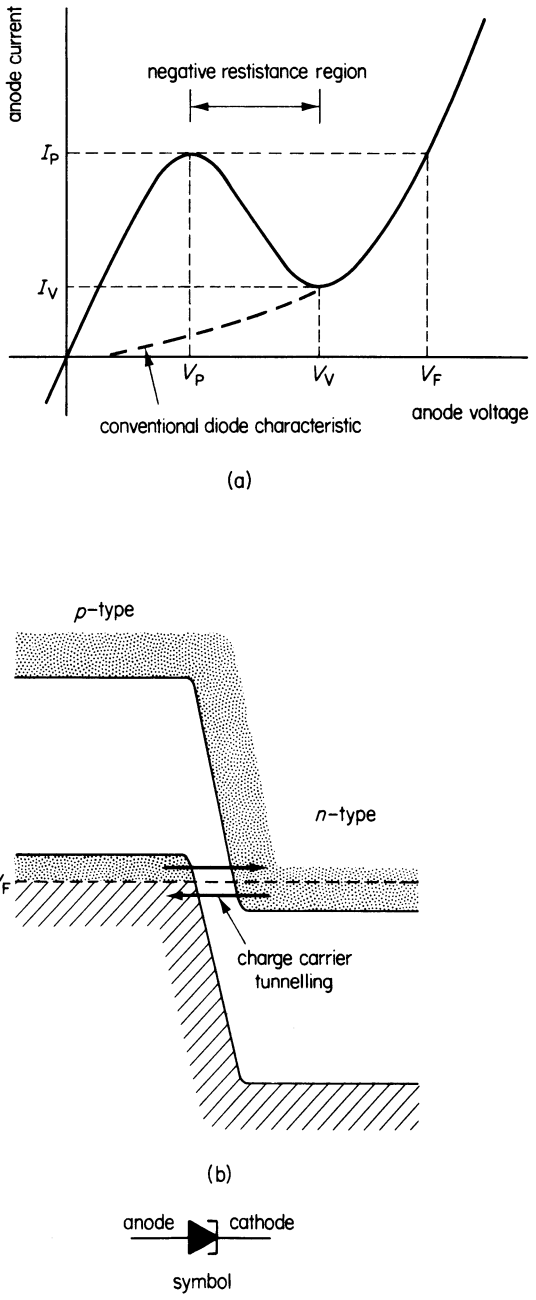


Figure 3.12 (a) Characteristic of a tunnel diode and (b) its energy-band diagram at zero bias

Table 3.1

Material	V_P(mV)	V_V(mV)	V_F(mV)	I_P/I_F
Gallium arsenide	150	500	1000	15
Germanium	55	330	500	8
Silicon	65	420	720	3.5

3.12b). At forward bias voltages up to V_P, current flow is by means of the tunnelling process. In the region between V_P and V_V, the conduction process changes from one of tunnelling to the conventional current flow process in a $p-n$ junction diode. Finally, at higher values of forward bias than V_V, the conduction process is entirely one associated with the forward biased operation of conventional $p-n$ junction diodes.

Arising from the change in operating mode between V_P and V_V, the characteristic exhibits negative resistance in this region; that is, the anode current reduces with increasing anode voltage.

Most commercially available tunnel diodes are manufactured either from germanium or gallium arsenide. They can be manufactured from silicon, but the ratio of peak-to-valley current (I_P/I_V) in this material is small. Typical values of the more important parameters are given in table 3.1.

The propagation of charge carriers by the tunnelling process proceeds at a velocity approaching that of light, so that the speed of operation of tunnel diodes is extremely fast. When used as a switching device, switching times of 1 ns and less are obtained. Also the operating temperature range of these devices is very wide, ranging from about −265 °C to several hundred °C.

The small-signal equivalent circuit for a tunnel diode in the negative resistance region of its characteristics is shown in figure 3.13, in which R and L represent the resistance and inductance, respectively, of the leads and contacts, C_j is the capacitance of the $p-n$ junction (see also section 3.7), and $-R_T$ is the negative resistance of the diode. Typical values are $R = 1\ \Omega$, $L = 8$ nH, $C_j = 15$ pF and $-R_T = -100\ \Omega$. The self-resonant frequency of the circuit in figure 3.13 is approximately $1/[2\pi\sqrt{(LC)}]$ Hz, and with the above values is about 460 MHz. Below this value of frequency, the equivalent circuit may simply be regarded as resistor R in series with the parallel combination of C_j and $-R_T$.

The tunnel diode can be used in a tuned amplifier by connecting it in parallel with a tuned circuit and load. Provided that the 'positive' resistance of the tuned circuit and load are greater than the 'negative' resistance of the tunnel diode, then the voltage across the load is an amplified version of the applied signal.

The amplifier described above can be used as an oscillator by ensuring that the combined 'positive' resistance of the tuned circuit and load is equal to the 'negative' resistance of the diode. The combined circuit then becomes an ideal 'resistanceless' tuned circuit, which causes oscillations to continue indefinitely.

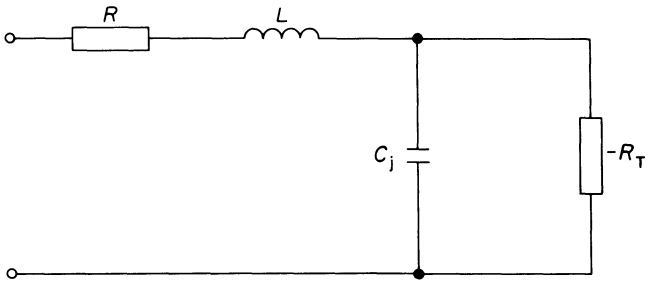


Figure 3.13 Small-signal equivalent circuit of the tunnel diode

3.7 Variable-capacitance Diodes

Under reverse bias conditions, a junction diode can be regarded as a parallel-plate capacitor having two plates (the *p*- and *n*-regions) which are separated by a dielectric (the reverse biased junction or depletion layer). The capacitance, C_T, of such a capacitor is related inversely to the width of the depletion layer. Capacitance C_T is known as the *transition capacitance* or *depletion-region capacitance*. Hence, a narrow depletion-layer width gives rise to a larger value of capacitance than does a larger value of depletion-layer width.

We shall now consider the effect on this capacitance of the value of the reverse bias voltage applied to the diode. As shown earlier the application of a reverse bias

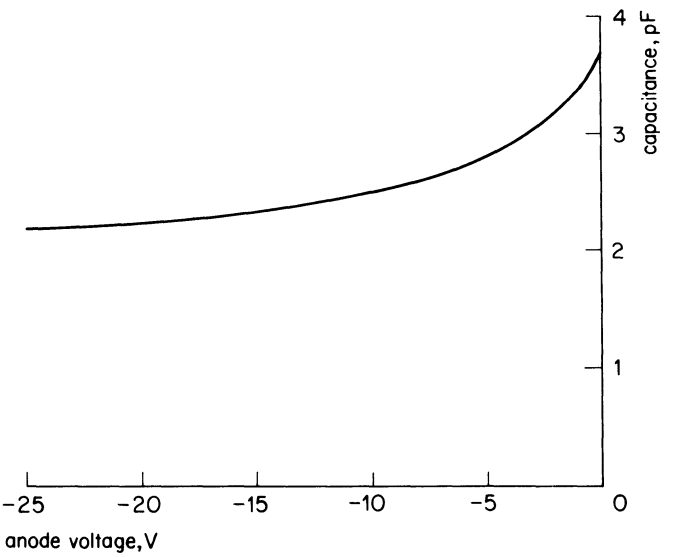


Figure 3.14 Capacitance–voltage characteristic for a varactor diode at a constant temperature

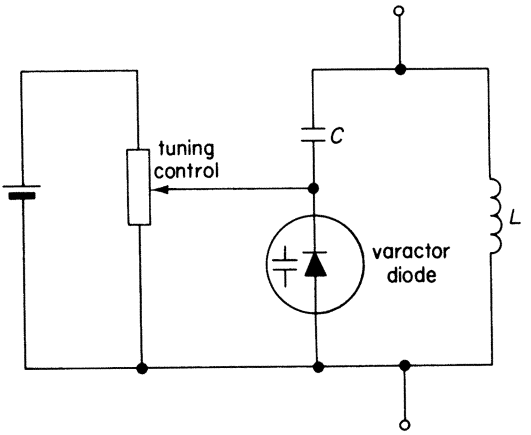


Figure 3.15 A basic voltage-controlled tuning circuit

to a p–n junction increases the height of the potential barrier between the two regions (see figure 3.2d). To cause this increase, a greater width of the semiconductor in the region of the junction must be depleted of charge carriers. Consequently, increasing the reverse bias increases the width of the depletion region and reduces the value of capacitance C_T.

A typical capacitance–voltage characteristic for a diode especially manufactured for use as a voltage-dependent capacitor is shown in figure 3.14. This type of diode is known under various names including *varactor diode, varicap diode* and *tuner diode.*

A popular application of varactor diodes is in voltage-controlled tuning circuits, as may be found in radio and TV receivers. A typical circuit is shown in figure 3.15; as the reverse bias applied to the varactor diode is increased, so its capacitance is reduced. Since the diode is in series with fixed capacitor C, an increase in the reverse bias voltage reduces the net value of the circuit capacitance, and increases the resonant frequency of the circuit.

3.8 *p–i–n* Diodes

A p–i–n diode (see figure 3.16) has an intrinsic semiconductor region between the p- and n-regions of the device. The additional i-region is used to fulfil one of several functions.

When reverse biased, depletion regions develop at the junctions of both p- and n-regions; the effective width of the diode depletion region is also increased by the width of the i-region. As a result, the capacitance of the p–i–n diode is lower for a given value of reverse bias voltage than is a conventional varactor diode.

The i-region also serves to increase the reverse breakdown voltage of the diode as follows. Since the i-region increases the separation between the p- and n-regions, it

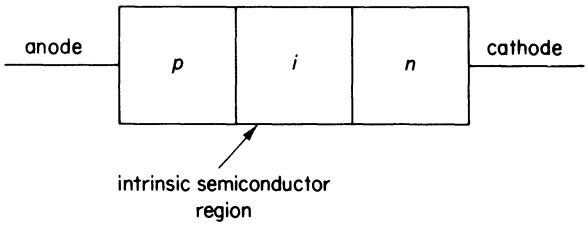


Figure 3.16 A p–i–n diode

also reduces the potential gradient between the two regions when reverse biased. The i-region may therefore be used to provide a higher value of reverse breakdown voltage than is the case with a p–n diode.

These devices are also used in microwave circuits as voltage-controlled attenuators. At the very high frequencies involved in this work, rectification action in the diodes ceases and the impedance of the diode is effectively that of the i-region. When the applied voltage is either zero or it reverse biases the diode, the impedance of the device is high. When forward biased the impedance is reduced. By altering the value of the applied bias voltage, the p–i–n diode can be used at microwave frequencies as part of a voltage-controlled attenuator.

3.9 Diffusion Capacitance or Storage Capacitance of Diodes

A capacitive effect known as *storage capacitance* or *diffusion capacitance*, C_D, is present under forward biased conditions. In this state, current flow takes place in the form of holes passing into the n-region and electrons into the p-region. In conventional p–n diodes, the n-region has a higher resistivity than does the p-region, so that current flow is predominantly due to hole flow from the p-region to the n-region. These holes travel a short distance into the n-region before they recombine with electrons and vanish. If the applied voltage is suddenly reversed, the mobile holes that have not recombined in the n-region are suddenly attracted back to the p-region.

This effect is illustrated in figure 3.17. Diagram (a) shows the basic circuit, and diagrams (b) and (c) show waveforms in the circuit. At time t_1 in diagram (b) the applied voltage is suddenly reversed from V_F to V_R; the reverse voltage instantaneously causes the mobile holes in the region of the junction to return to the p-region, giving rise to a transient reverse current, I_R (see figure 3.17c). The time interval t_s taken for the minority carriers in the n-region to be returned to the p-region is known as the *storage time* of the device. The duration of the storage time may be as small as a fraction of a nanosecond in a high-speed switching diode. Clearly the larger the value of the forward current the greater is the time required to remove the excess charge when reverse bias is applied; in high-current diodes the value of t_s can be up to one millisecond.

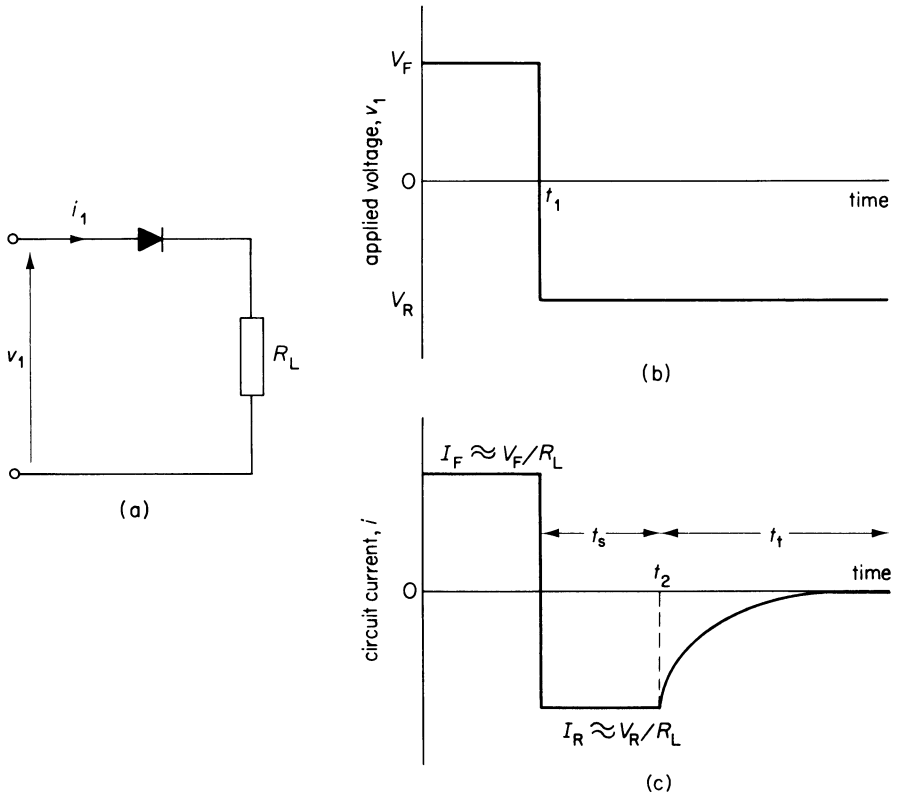


Figure 3.17 Hole-storage effect in a p–n junction diode

The phenomenon described above is known as *hole storage*, and occurs in many types of p–n junction device. The reverse current surge is equivalent to the discharge of capacitance C_D during the hole-storage time interval.

It takes an additional time interval, t_t, known as the *transition time* for the charge carriers to withdraw sufficiently from the p–n junction for blocking conditions to be established, and for reverse biased depletion-region capacitance (see section 3.7) to become fully charged.

3.10 Schottky Barrier Diodes

A Schottky barrier diode or *hot carrier diode* is a rectifying metal-to-semiconductor junction (see also section 1.15). The semiconductor may either be n-type or p-type silicon but, since the mobility of electrons is greater than that of holes, a metal-to-n-type junction is usually used in order to give a higher switching speed.

The current flow mechanism through a Schottky barrier diode differs from that in a p–n diode in that only majority charge carriers are involved in Schottky

diodes. This means that charge storage effects do not occur in a Schottky diode, resulting in switching speeds of less than 0.1 ns being achieved, Schottky diodes are used in conjunction with TTL logic gates (see section 6.6) in order to reduce the switching time of the logic elements.

3.11 Unijunction Transistors (UJTs)

The unijunction transistor is, strictly speaking, not a transistor but is more accurately described as a *double-based diode*. The UJT consists of a bar of *n*-type material to which two ohmic connections are made; *p*-channel UJTs are also manufactured, but are less popular than *n*-channel devices. The connections made to the ends of the channel are known as *base 1* and *base 2* (see figure 3.18a). A *p–n* junction is formed between the bar and a *p*-type emitter. The base 2 termination is

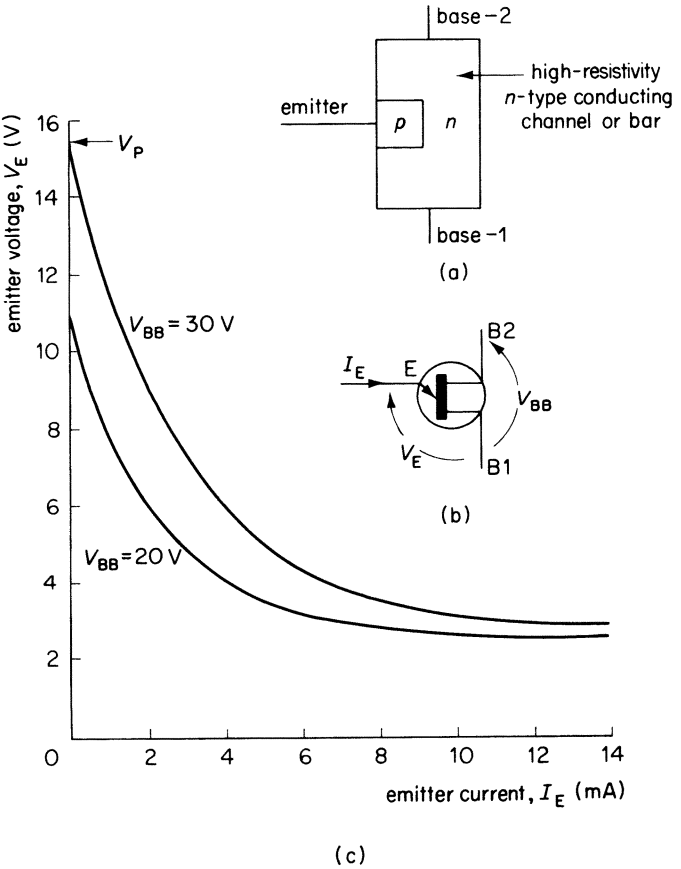


Figure 3.18 The unijunction transistor: (a) construction, (b) circuit symbol and (c) typical characteristics

connected to the positive pole of the *interbase supply*, V_{BB}; in the absence of a signal being applied to the emitter electrode, the value of the *interbase resistance*, R_{BB}, lies in the range 4 to 12 kΩ. The *n*-type bar therefore acts as a potential divider, and the emitter is positioned so that a voltage in the range $0.4V_{BB}$ to $0.8V_{BB}$ appears between base 1 and the emitter junction. The coefficient of V_{BB} above is known as the *intrinsic stand-off ratio*, η, of the UJT. Provided that the potential applied to the emitter is less positive than ηV_{BB}, then the emitter-to-bar $p-n$ junction is reverse biased and no current can flow in the emitter circuit.

Increasing the emitter voltage to a value slightly greater than ηV_{BB} causes the emitter $p-n$ junction to be forward biased, when current carriers are emitted into the bar from the emitter circuit. This occurs at the *peak-point voltage*, V_P, of the UJT (see figure 3.18c); the value of V_P is given by the expression

$$V_P = V_j + \eta V_{BB}$$

where V_j is the forward biased p.d. across the $p-n$ junction and is of the order of a few hundred millivolts. From the above equation, readers will note that the value of the peak-point voltage is very nearly proportional to V_{BB}.

When the triggering action described above occurs, the resistance between the emitter and base 1 falls to a low value and the value of the emitter current, I_E, rises rapidly. Due to its characteristic, the UJT is frequently used as a capacitor discharge device, illustrated below.

A UJT relaxation oscillator

A relaxation oscillator is a circuit which generates non-sinusoidal waveforms by a process of gradually charging a capacitor through a large value of resistance, and then quickly discharging it through a lower value of resistance. One form of relaxation oscillator is shown in figure 3.19, in which capacitor C is charged via resistor R from V_{BB} and, at regular intervals of time, is quickly discharged through R_{B1} via the UJT.

When the supply is first switched on, the capacitor is discharged; the current through R begins to charge the capacitor, and the potential across C (which is the voltage at the gate of the UJT) rises exponentially. So long as the gate potential is less than the peak-point voltage of the UJT, the device is not triggered into conduction and the capacitor voltage continues to rise. However, when the gate voltage rises to V_P, the UJT is triggered into conduction and the capacitor is rapidly discharged through R_{B1}. Resistor R_{B1} has a relatively low value, and the discharge period is very short; during this period of time a large value of current flows through R_{B1}. Once the discharge has been completed, the gate $p-n$ junction of the UJT reverts to its blocking state, and the charging cycle is recommenced.

The circuit shown in figure 3.19 therefore generates a sawtooth waveform at the gate terminal of the UJT; the discharge pulses of the capacitor also generate a sequence of pulses across R_{B1}, each of 5 to 10 μs duration.

The periodic time of the waveform can be estimated from the following.

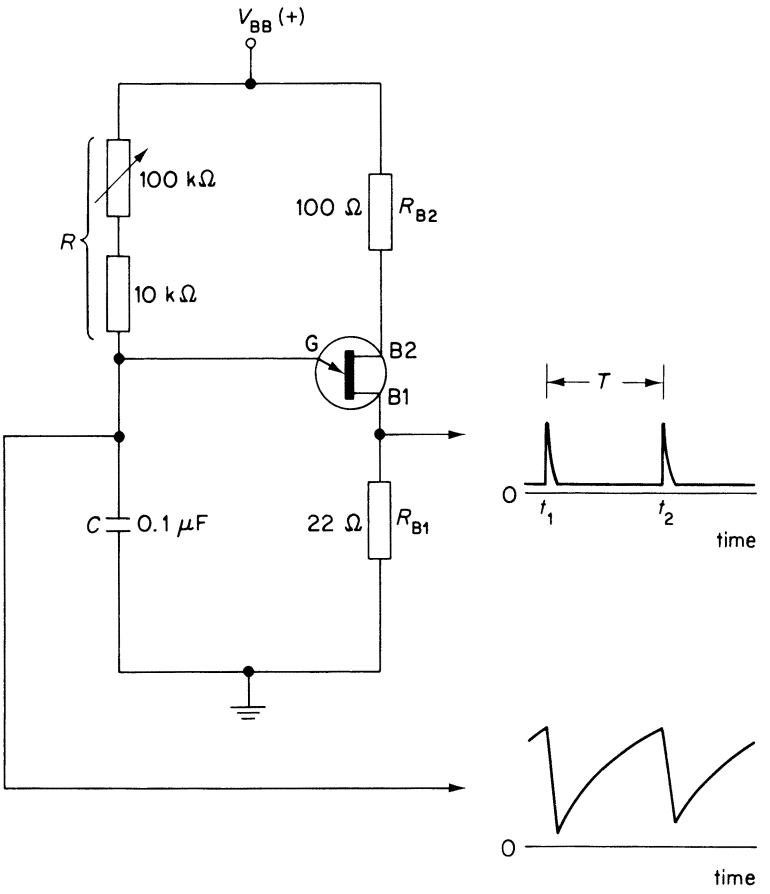


Figure 3.19 A UJT relaxation oscillator

Assuming that the capacitor is fully discharged after each cycle, the equation for the voltage, V_G, at the gate of the UJT during the charging period is

$$V_G = V_{BB}(1 - e^{-t/RC})$$

After time interval T, the value of V_G is equal to ηV_{BB}; that is

$$\eta V_{BB} = V_{BB}(1 - e^{-T/RC})$$

Solving for the periodic time, T, of the cycle yields

$$T = RC \ln\left(\frac{1}{1-\eta}\right) = 2.3\, RC \log\left(\frac{1}{1-\eta}\right)$$

A typical value of η is 0.55, giving $T = 0.8RC$ seconds (R in ohms and C in farads). The pulse period of the oscillator can be controlled by altering the value of either R

or C. In figure 3.19 a variable resistor is used to control the pulse period and, with the values shown would give a pulse repetition rate in the range between 0.8 ms and 9 ms.

In section 3.3 it was shown that the p.d. across a forward biased p–n diode reduces by 2.5 mV for each °C rise in temperature. This also applies to the gate junction of the UJT. Additionally the resistance–temperature coefficient of the n-type bar indicates that the voltage at the gate junction within the bar increases with temperature. By including a suitable value of resistance, R_{B2}, in series with base 2, the two thermal effects can be arranged to cancel each other out. The net result of including R_{B2} in the circuit is to stabilise the value of the peak-point voltage over a wide range of temperature change. The following empirical formula can be used to estimate the value of R_{B2}.

$$R_{B2} \approx \frac{0.7R_{BB}}{\eta V_{BB}}$$

The circuit shown in figure 3.19 is widely used as the basis of many thyristor trigger circuits in power electronics (see also chapter 9), and is also used as the basis of many electronic timing circuits.

4 *Bipolar Junction Transistors, Amplifiers and Logic Gates*

4.1 Introduction to Bipolar Junction Transistors (BJTs)

The name 'transistor' is a contraction of TRANSfer resISTOR, and the bipolar junction transistor was the earliest type of semiconducting amplifying device to come into use. The term 'bipolar' is coined from the fact that both polarities of charge carrier (that is, holes and electrons) play a part in the operation of the transistor.

The BJT is manufactured in a single crystal of semiconductor material, having three regions known as the *emitter*, the *base* and the *collector* regions. As shown below, the emitter region emits charge carriers into the transistor, the majority of them (typically 98 per cent or greater) being collected by the collector region. The base region, which lies between the collector and emitter regions, provides a means of controlling the magnitude of the current flow through the transistor. The name of the base region is derived from the fact that, in early transistors, this region formed the physical foundation or base on which the transistor was constructed.

The physical size of a modern low-power transistor is very small, a typical size being about $6 \times 30\,\mu m$ ($1\,\mu m = 10^{-6}$ m). The thickness of the base region, mentioned above, may be as small as $0.5\,\mu m$; for comparison, readers are reminded that the wavelength of green light is about $0.5\,\mu m$. Details of the method of manufacture of transistors are outlined in chapter 6. So that transistors can be handled, they are encapsulated in physically 'large' canisters. Two types of encapsulation are shown in figure 4.1; diagram (a) illustrates a 'top hat' TO5 metal canister that is hermetically sealed, and diagram (b) shows one form of plastic encapsulation used in a wide range of domestic and industrial equipment.

Two types of BJT in common use are illustrated in figure 4.2. Early devices were manufactured from germanium and were of the *p–n–p* type in these devices the two *p*-regions formed the emitter and the collector (see figure 4.2a) the *n*-region being the base region. Technological developments in silicon devices rapidly overtook those of germanium devices, and the majority of BJTs are now silicon *n–p–n* devices (see figure 4.2b).

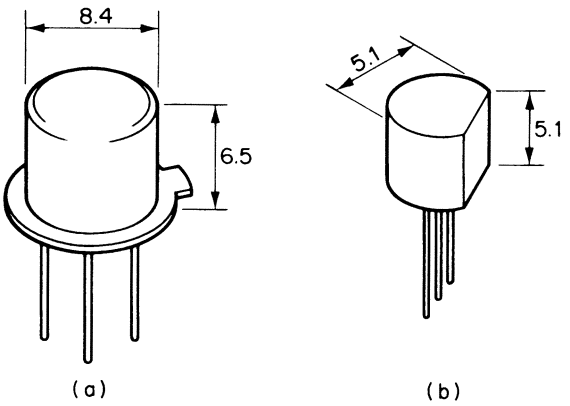


Figure 4.1 Transistor packaging: (a) TO5 canister and (b) one form of plastic encapsulation. All dimensions in mm

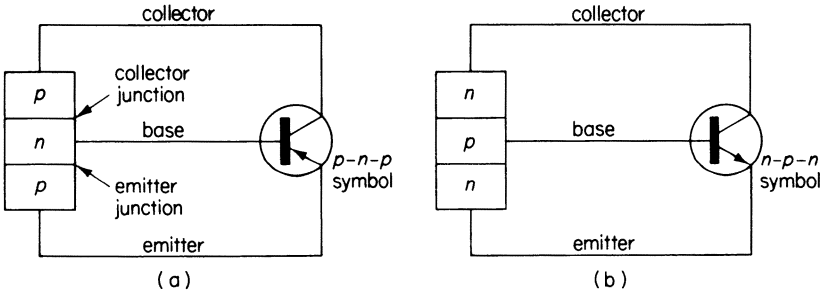


Figure 4.2 Transistor symbols: (a) p–n–p transistor and (b) n–p–n transistor

The direction of *conventional* current flow (that is, hole flow) through the device is given by the direction of the arrow on the emitter electrode in the circuit symbol. In p–n–p transistors (figure 4.2a), the direction of current flow is from the emitter to the collector, so that they operate with the collector negative with respect to the emitter. The direction of current flow in an n–p–n transistor is from the collector to the emitter, so that this type of device operates with its collector positive with respect to its emitter. A description of the operation of an n–p–n transistor is given in section 4.3.

4.2 Transistor Circuit Configurations

In electronic circuits, one of the regions of the transistor is connected to the 'input' signal, another is connected to the 'output' line, and the third region is usually

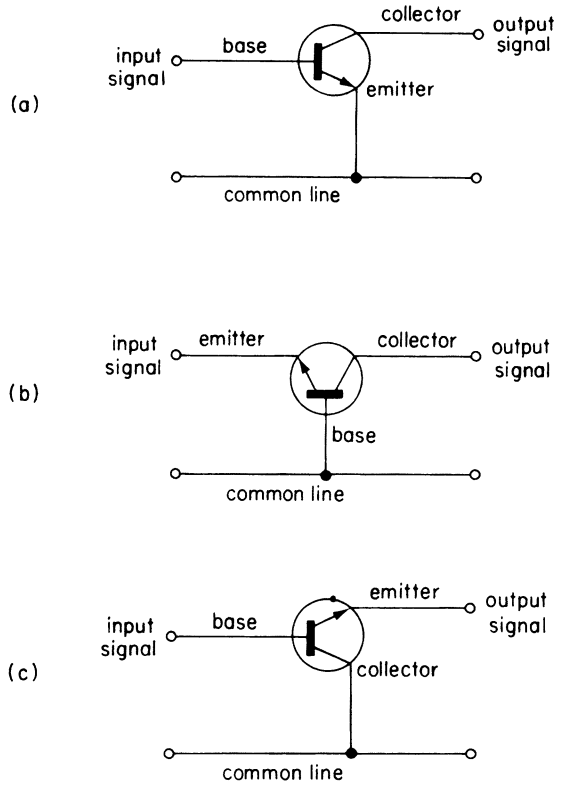


Figure 4.3 Configurations of the *n–p–n* transistor: (a) common-emitter, (b) common-base and (c) common-collector

connected to the line which is 'common' to both 'input' and 'output' signals. The circuit configurations are described in terms of the transistor region that is connected to the 'common' line. Thus a circuit may be described as a *common-emitter circuit*, or as a *common-base circuit*, or as a *common-collector circuit*. Simplified versions of these circuits using *n–p–n* transistors are shown in figure 4.3.

The majority of electronic circuits employ the common-emitter configuration, since it provides a range of operating parameters that are well suited to the requirements of electronic applications. The more important of these parameters are discussed later in the chapter.

4.3 Operation of an *n–p–n* Transistor

A simple circuit which may be used to test the operation of an *n–p–n* transistor is shown in figure 4.4a. When the transistor is used in amplifier circuits, *the emitter*

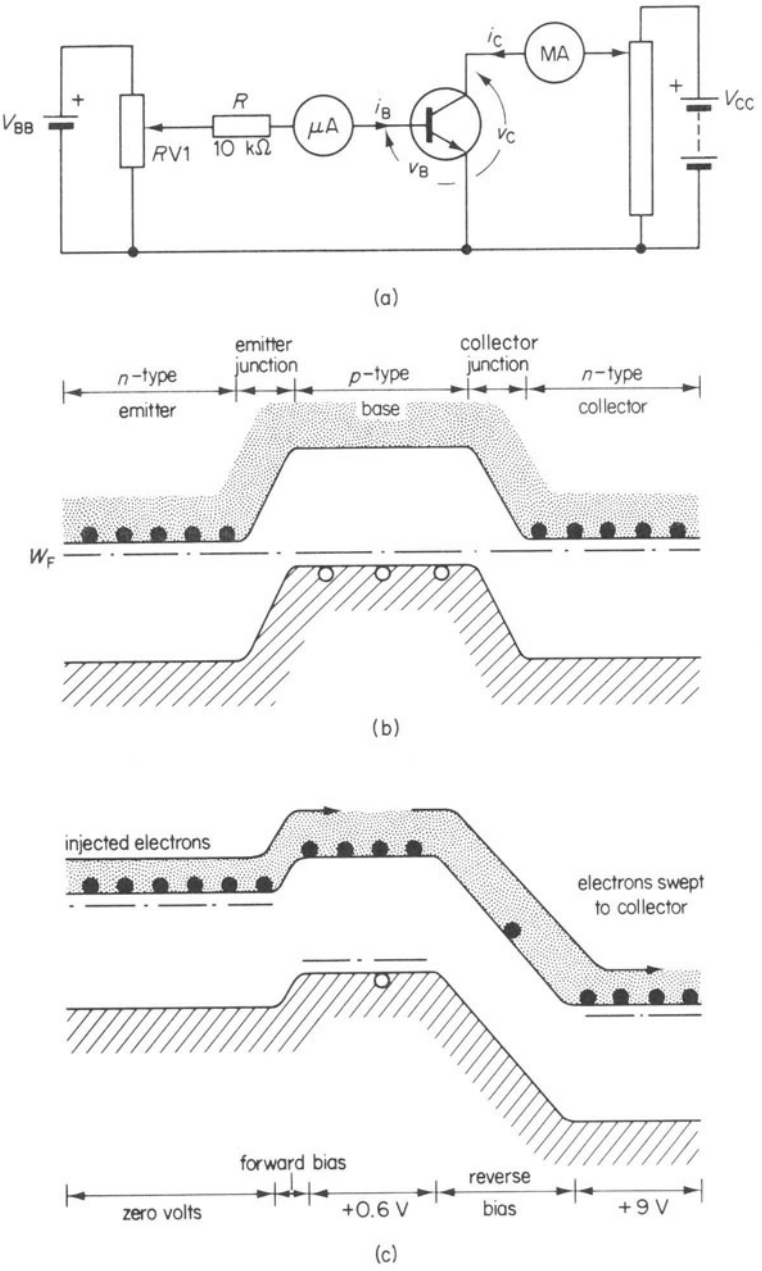


Figure 4.4 (a) A simple test circuit for an n–p–n transistor, (b) the energy-band model for a transistor in equilibrium and (c) the energy-band model with typical bias levels

junction is forward biased by means of the base bias supply, V_{BB}. At the same time, the collector region (an *n*-region) is connected to a more positive potential than that of the base region, so that *the collector junction of the transistor is reverse biased*.

Let us first consider the conditions existing inside the transistor before it is connected to the circuit. As in the case of other semiconductor devices, there is an interchange of charge carriers between the three regions until the Fermi energy levels align. Consequently depletion regions exist at each junction, as shown in figure 4.4b.

When in the electrical circuit in figure 4.4a, the base region is connected to a positive potential so that the emitter junction is forward biased. Consequently charge carriers flow across the junction between the emitter and the base. To understand what happens next, it is necessary to know more about the impurity doping levels of the regions of the transistor.

The principal requirements of bipolar transistors are that most of the current carriers leaving the emitter should reach the collector, and that the value of the current, i_B, flowing into the base region should be as small as possible; these impose certain constraints on the transistor as follows. Firstly, since the collector is constructed from *n*-type material, then the current flowing across both the emitter and collector junctions should primarily be electron flow. This is accomplished by reducing the concentration of impurities in the *p*-type base to a value which is much lower than that in the *n*-type emitter. Also the width of the base region should be very small, so that the electrons have only a short distance to drift before they come under the influence of the collector voltage. Transistors with base regions of about 0.5 μm thickness can readily be manufactured.

Hence, under forward bias conditions, a large concentration of electrons is emitted into the base region, where they become minority charge carriers. A small number of these electrons combine with holes in the base region, and the replenishment of these holes constitutes base current in the external circuit. Since there are relatively few mobile holes in the near-intrinsic base region, the base current has a relatively low value. Once the electrons have been injected into the base region, they drift towards the collector at a slow rate. The time taken for these charge carriers to cross the base region has a considerable influence on the high-frequency performance; to obtain a good high-frequency performance, the time taken for the electrons to cross the base region should be as short as possible. For this reason the base region should be as thin as possible.

When the electrons reach the collector junction, they are rapidly swept into the collector by the positive potential applied to that region. The impurity-doping rules of the collector region have two conflicting requirements. Firstly, it is desirable to keep the doping concentration as small as possible in order to increase the width of the collector junction. This increases the value of the voltage at which avalanche breakdown occurs between the collector and the base, and also reduces the value of the junction transition capacitance (see also section 3.7). Secondly, it is desirable to have a high doping level in order to reduce the collector-to-emitter p.d. to a low

value. Using modern production techniques (see chapter 6), it is possible to obtain a compromise which satisfies both requirements.

In operation it is found that the collector current has a value which lies between about fifty and several hundred times the value of the base current.

4.4 Common-emitter Characteristics

The static characteristics of a transistor, determined using the test circuit in figure 4.4a, are of particular importance since they allow the more important parameters to be determined.

The *output characteristic* or *collector characteristic* in figure 4.5a is valuable to circuit designers, since it shows how the collector current, i_C, varies both with the collector voltage, v_C, and with the base current, i_B, over the normal range of values for that device. When the base current is zero, no charge carriers are injected into the base region from the emitter, and the collector current is zero. In this mode of operation, the transistor is said to be *cut off*.

When the emitter junction is forward biased, that is, $i_B > 0$, electrons are emitted into the base region in the manner described above; the majority of these charge carriers are swept into the collector region, giving rise to collector current.

The relationship between the magnitude of the direct current in the collector

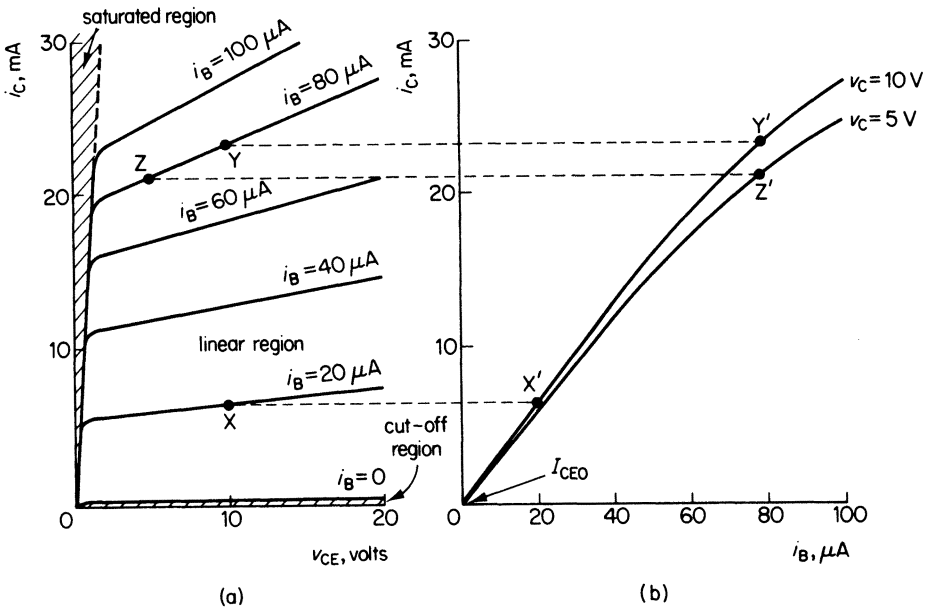


Figure 4.5 Characteristics of an *n–p–n* transistor in the common-emitter mode: (a) output characteristics and (b) transfer characteristics

and that in the base is known as the *large-signal forward current gain in the common-emitter mode*, and is given the symbol h_{FE}, where

$$h_{FE} = \frac{\text{collector current}}{\text{base current}} \text{ at a constant value of collector voltage}$$

At a collector voltage of 10 V on the output characteristics in figure 4.5a, the large-signal current gain at a base current of 20 μA (point X on the diagram) is

$$h_{FE} = \frac{i_C}{i_B} = \frac{6.75 \text{ mA}}{20 \,\mu\text{A}} = 337.5$$

and at a base current of 80 μA (point Y on the figure) the current gain is

$$h_{FE} = \frac{i_C}{i_B} = \frac{23 \text{ mA}}{80 \,\mu\text{A}} = 287.5$$

Readers will note that as the collector voltage increases in value, the output characteristics separate from one another. A reason for this phenomenon was put forward by J. M. Early in 1952, when he showed that as the collector voltage is increased, the collector depletion region extends further into the base region and reduces the 'electronic' width of the base region. Thus there is less chance at the higher values of voltage for charge carrier recombinations to take place in the base region, so that a greater number of charge carriers reach the collector. This effect is known as the *Early effect* or *base-width modulation*.

The region on the output characteristics in which the collector current increases in a more or less linear manner with base current, is known as the *linear region* of the characteristics. Transistors used in *linear amplifiers* operate in this region, and they provide an output signal which is an amplified and undistorted version of the input signal.

When the collector voltage has a very low value (ideally zero), the transistor is said to be saturated, and it operates in the saturated region of the characteristics (see figure 4.5a). In this operating mode both the emitter and the collector junctions are forward biased. The transistor is operated in this region in saturated switching circuits, such as in many forms of electronic logic gates.

The static transfer characteristics (figure 4.5b) are useful in many applications, and show the relationship between i_C and i_B at constant values of v_C. The transfer characteristic for a collector voltage of 10 V is obtained by projecting values of collector current at v_C = 10 V from the output characteristics. Thus, corresponding to point X on the output characteristic, there is a point X′ on the v_C = 10 V transfer characteristic. There is also a point Y′ corresponding to point Y on the output characteristics. The transfer characteristics for other values of v_C are plotted in this way; for example, there is a point Z′ on the transfer characteristic for v_C = 5 V corresponding to point Z on the output characteristics.

A typical set of BJT input characteristics is shown in figure 4.6. The input characteristics relate the total current, i_B, flowing into the base to the value of

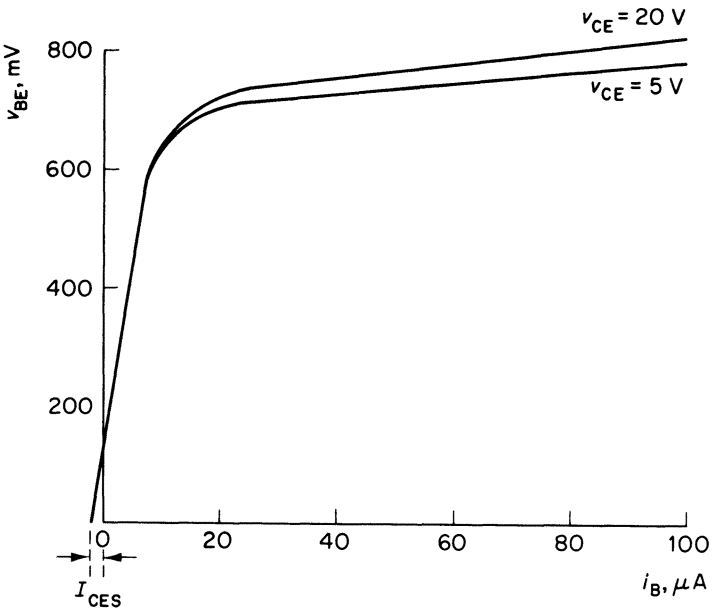


Figure 4.6 Typical common-emitter input characteristics for a silicon *n–p–n* transistor

voltage, v_{BE}, between the base and the emitter. Since the emitter junction is basically a diode operated in its forward biased mode, the resulting characteristics are diode-like in nature. When the input characteristics are plotted, the ordinates are reversed compared with those of a diode, that is, the current is plotted horizontally and the voltage is plotted on the vertical axis. This is adopted simply for convenience in determining the parameters of the device.

Readers will note that when the base voltage is reduced to zero, a current flows *out* of the base region; the value of this current is given the symbol I_{CES}. The value of I_{CES} in a low-power silicon transistor may be as low as 10 nA, but in some germanium power transistors it may be several hundred microamperes.

A value of leakage current which is frequently quoted for transistors is I_{CEO} (see figure 4.5b), and is the leakage current flowing into the transistor collector when the base current is zero; that is when the base circuit is open, and $i_B = 0$. In low-power silicon transistors, I_{CEO} and I_{CES} have much the same value.

4.5 Common-base Characteristics

In certain instances the transistor offers better performance when operating in the common-base mode than it does in the common-emitter mode; for example, this

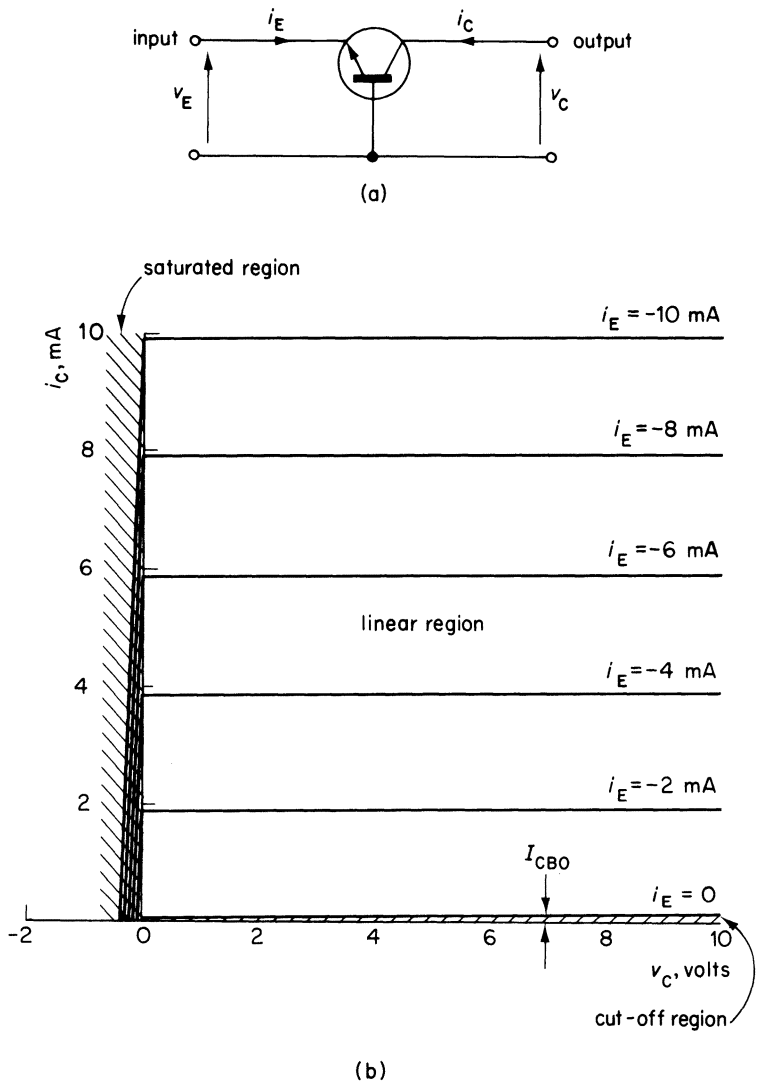


Figure 4.7 (a) The common-base connection and (b) typical output characteristics

mode provides a better high-frequency response than is obtained from the common-emitter configuration.

A typical set of common-base output characteristics is shown in figure 4.7b. In the common-base configuration the input signal is applied to the emitter, and the output signal is taken from the collector. Since the emitter current is the sum of the collector and base currents, its value is slightly greater than that of the collector

current. This fact is recorded on the output characteristics; for example, corresponding to an emitter current (i_E) of 10 mA, the collector current (i_C) is about 9.95 mA. Readers will also note that i_E is given a negative sign; this is due to the fact that, in transistor circuit theory, current is assumed to flow *into* all the regions of the transistor. In the case of an n–p–n transistor, current actually flows *out* of the emitter and is therefore given a negative sign.

The ratio i_C/i_E at any point on the characteristics is known as the static value of the *forward current gain* in the common-base mode, and is given the symbol h_{FB}. If the instantaneous value of the collector current is 9.95 mA when the emitter current is −10 mA, then $h_{FB} = 9.95/(-10) = -0.995$. The value of h_{FB} usually lies in the range −0.98 to −0.998. The relationship between values of the common-base and common-emitter current gains is

$$h_{FB} = -\frac{h_{FE}}{1 + h_{FE}}$$

In the common-base mode, the transistor is said to be *cut off* when the emitter current is zero, that is, when the emitter is open-circuited. Under this condition, the leakage current flowing into the collector is designated the symbol I_{CBO}, the value of I_{CBO} being much less than I_{CEO}, the leakage current in the common-emitter mode. The relationship between the two values is given by

$$I_{CBO} = I_{CEO}(1 + h_{FB})$$

If $h_{FB} = -0.99$, then $I_{CBO} = 0.01\, I_{CEO}$.

As in the case of the common-emitter configuration, the common-base characteristic exhibits a *saturated region*, a *linear region* and a *cut-off region* (see figure 4.7). In the majority of applications, the transistor operates in the linear region of the characteristics. Readers will note that the output characteristics are practically parallel to one another, and are uniformly spaced. This fact indicates that the common-base amplifiers are free from the Early effect distortion which exists in common-emitter amplifiers.

4.6 Thermal Effects on Transistor Characteristics

The principal effects of temperature changes on the common-emitter and common-base characteristics are discussed below.

Effects on common-emitter characteristics

In all bipolar transistors the leakage current, the current gain and the base–emitter voltage are temperature-dependent factors. The net effect of an increase of temperature on the output characteristics is illustrated in figure 4.8 for the case when the base current is 40 μA; it shows an increase in the separation between the output characteristics.

The effect on the input characteristics is twofold. Firstly, the base–emitter

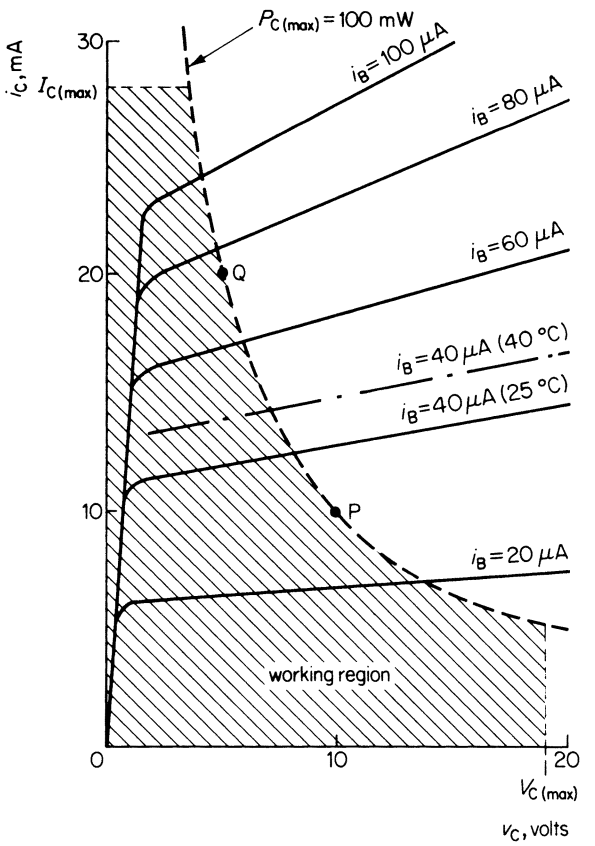


Figure 4.8 The working region on a family of common-emitter output characteristics

voltage decreases at a rate of about 2 mV for each °C rise. Secondly, the leakage current increases in value. In the case of silicon transistors the value of the leakage current is very small, and any change in it has an insignificant effect on the input characteristics.

Effects on common-base characteristics

The changes due to thermal effects on the common-base output characteristics are generally quite small. The change in the base–emitter voltage with temperature is much the same as in the common-emitter case.

General comments

In general, all types of amplifiers are so designed that the effects of temperature change on their performance are negligible. The greatest problems arise in the case

of power amplifiers in which the mean power dissipated in the transistor may approach its power rating, P_{tot}. In such a case, an increase in ambient temperature causes the collector current to rise which, in turn, may result in the mean power generated by the device exceeding its rating. Should this happen, the power consumed by the transistor exceeds its capability to dissipate the heat generated – with a consequent further increase in temperature. If this phenomenon is progressive, it leads to an operating temperature at which the transistor fails catastrophically. This is known as *thermal runaway*. Thermal runaway should not occur in well-designed amplifiers. Circuits which provide stabilisation against thermal effect are described in section 4.12.

4.7 Maximum Power Dissipation

One limitation on the allowable working area on the output characteristics of a transistor is the maximum power which may be continuously dissipated by the collector. Suppose that the *maximum allowable continuous collector power dissipation*, $P_{C(max)}$, by the collector is 100 mW (see figure 4.8); in this case the area on the characteristics that may safely be used lies below a curve described by the equation

$$v_C i_C = 100 \text{ mW}$$

Points P and Q in figure 4.8 both lie on this curve, having co-ordinates of 20 mA and 5 V, and 10 mA and 10 V, respectively. Also it is usual for manufacturers to quote the value of the *maximum permissible continuous power dissipation*, P_{tot}, of the transistor. This parameter is calculated from the equation

$$P_{tot} = v_C i_C + v_B i_B = P_{C(max)} + v_B i_B$$

where v_B and i_B are the total instantaneous values of the base voltage and base current, respectively. Since the product $v_B i_B$ has a value which is less than about 0.1 per cent of $P_{C(max)}$, it is reasonable to assume that $P_{C(max)}$ and P_{tot} have about the same value.

A second limitation on the working area of the output characteristics is the *maximum allowable collector current*, $I_{C(max)}$. The limiting value of this current is generally determined by the value of collector current which causes the current gain of the transistor to reduce significantly. A further limitation on the working area is the *maximum collector voltage*, $V_{C(max)}$.

4.8 Transistor Derating Curves

The value of P_{tot} is usually specified at a temperature of 25 °C. Above this temperature, the transistor dissipation is usually derated in order to limit the junction temperature to a safe value. The derating factor is given by the relationship

$$\text{derating factor} = \frac{P_{tot}}{\Delta T}$$

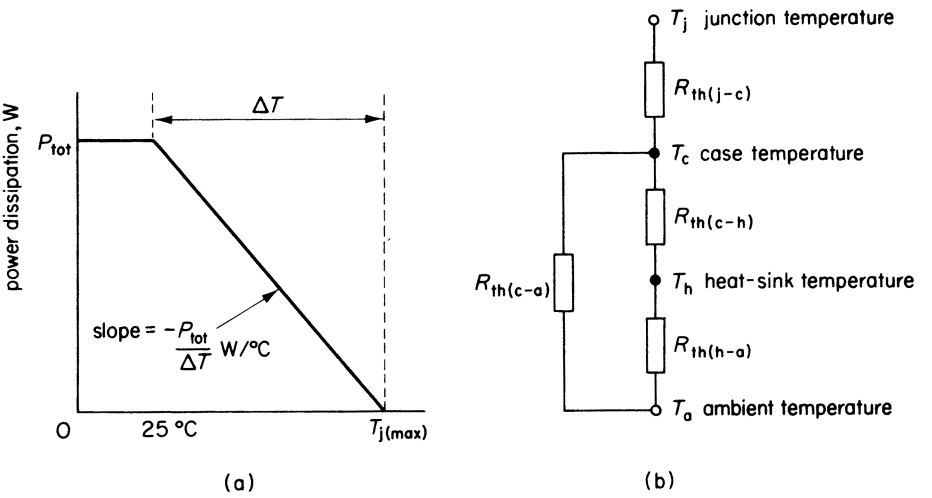


Figure 4.9 (a) Thermal derating curve for a transistor and (b) an equivalent heat-flow circuit

where ΔT is as defined in figure 4.9a. In this figure the maximum allowable temperature of the junction is $T_{j(max)}$, and has a value in the range 75 to 100 °C for germanium transistors and 150 to 200 °C for silicon transistors. When the junction temperature is equal to $T_{j(max)}$, the transistor power rating is derated to zero. If P_{tot} = 100 mW and $T_{j(max)}$ = 175 °C then, from figure 4.9a

$$\text{derating factor} = \frac{100 \times 10^{-3}}{175-25} = 0.00067 \text{ W/°C}$$

Typical values for the derating factor range from about 0.0001 W/°C for a low-power transistor in free air to about 5 W/°C for a high-power transistor mounted on an efficient heat sink.

A parameter known as the *thermal resistance* is used in transistor heat-flow problems, where

$$\text{thermal resistance} = R_{th} = \frac{1}{\text{derating factor}} \text{ °C/W}$$

Derating factors of 0.0001 and 5 W/°C correspond to thermal resistances of 10 000 °C/W and 0.2 °C/W, respectively.

To reduce the thermal resistance between the junction and atmosphere, the transistor may be secured in a cooling clip or to a heat sink, both increasing the effective heat radiating area. For example, a transistor whose thermal resistance between its junction and atmosphere is 290 °C/W, may have its thermal resistance reduced to 140 °C/W when secured to a cooling clip. If the cooling clip is secured to a blackened-aluminium heat sink, the thermal resistance may be reduced to about 80 °C/W. If a fan-cooled finned heat sink is used (sometimes known as an 'infinite'

heat sink), the thermal resistance between the transistor and atmosphere could fall to, say, 55 °C/W.

An equivalent heat-flow diagram for a transistor is shown in figure 4.9b. In this diagram, temperatures T_j, T_c, T_h and T_a respectively represent the temperature at the p–n junction, the case of the transistor, the heat sink and the atmosphere. The effective thermal resistance between the junction and the case is $R_{th(j-c)}$, and between the case and the heat sink is $R_{th(c-h)}$, etc. The effective value of the thermal resistance of a parallel circuit is dealt with in the same manner as an electrical parallel resistive network. When a transistor is mounted on a heat sink, the majority of the heat flow takes place via the heat sink rather than from the case of the transistor directly to the atmosphere. Thermal resistance values for transistors and heat sinks can be obtained from manufacturers' literature.

Example

A transistor has a P_{tot} rating of 5 W, and the maximum allowable junction temperature is 100 °C. If $R_{th(j-c)} = 2$ °C/W, $R_{th(c-a)} = 25$ °C/W, $R_{th(c-h)} = 0.5$ °C/W and $R_{th(h-a)} = 2$ °C/W, estimate the maximum allowable ambient temperature for safe operation.

Solution

The effective thermal resistance, $R_{th(j-a)}$, between the transistor junction and atmosphere is

$$R_{th(j-a)} = R_{th(j-c)} + \frac{R_{th(c-a)}(R_{th(c-h)} + R_{th(h-a)})}{R_{th(c-a)} + R_{th(c-h)} + R_{th(h-a)}}$$

$$= 2 + \frac{25(0.5 + 2)}{25 + 0.5 + 2} = 4.27 \text{ °C/W}$$

Now

$$T_j - T_a = R_{th(j-a)} P_{tot}$$

$$= 4.27 \times 5 = 21.35$$

Therefore

$$T_a = T_j - 21.35 = 100 - 21.35 = 78.65 \text{ °C}$$

4.9 *h*-parameters

To quantify the behaviour of transistors, a range of parameters and equivalent circuits is used. Of particular interest here is a group of parameters known as *hybrid parameters* or *h-parameters*, which describe the low- and medium-frequency operation of bipolar transistors. These parameters, described below, are widely used since they are relatively easy to measure.

Consider the 'black box' transistor circuit in figure 4.10, in which the transistor may be connected in any one of its three basic configurations. The *h*-parameters of

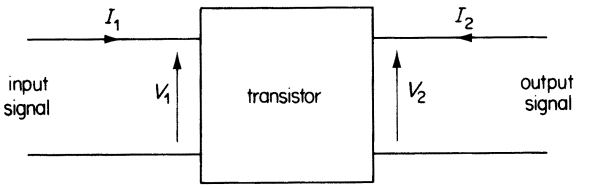


Figure 4.10 *h*-parameters of a transistor

the transistor can be determined either for large-signal (or d.c.) operation, or for small-signal operation. The latter are of great interest to electronic engineers since, in most cases, the signal which is to be amplified is usually an alternating signal of small amplitude. The small-signal *h*-parameters are defined by the equations

$$V_1 = h_i I_1 + h_r V_2 \tag{4.1}$$

$$I_2 = h_f I_1 + h_o V_2 \tag{4.2}$$

where V_1 and I_1 are the r.m.s. values of the sinusoidal input voltage and current, respectively, and V_2 and I_2 are the r.m.s. values of the respective output quantities. The *h*-parameters in equations 4.1 and 4.2 are defined in the following form.

$h_i = \left(\frac{V_1}{I_1}\right)_{\delta V_2 = 0}$ = input resistance (with output short-circuit to a.c.), having dimensions of resistance

$h_r = \left(\frac{V_1}{V_2}\right)_{\delta I_1 = 0}$ = reverse voltage feedback factor (with input open-circuit to a.c.), and is dimensionless

$h_f = \left(\frac{I_2}{I_1}\right)_{\delta V_2 = 0}$ = forward current gain (with output short-circuit to a.c.), and is dimensionless

$h_o = \left(\frac{I_2}{V_2}\right)_{\delta I_1 = 0}$ = output conductance (with input open-circuit to a.c.), having dimensions of conductance

The above parameters are called hybrid parameters since the dimensions of the parameters are not consistent, one having the dimensions of resistance, one being a conductance, and two being dimensionless.

Depending on the circuit configuration, that is, common-emitter, common-base or common-collector, one of the following three subscripts is added to each of the above parameters.

e = common-emitter configuration

b = common-base configuration

c = common-collector configuration

Thus the input resistance parameters, h_i, in the common-emitter, common-base and common-collector configurations are h_{ie}, h_{ib}, and h_{ic}, respectively. The voltages and currents in the general circuit, figure 4.10, are converted to specific values, that is, V_b, I_b, V_c, I_c, etc., in the particular application. Hence the small-signal h-parameter equations for the common-emitter configuration are written as follows

$$V_b = h_{ie}I_b + h_{re}V_c \tag{4.3}$$

$$I_c = h_{fe}I_b + h_{oe}V_c \tag{4.4}$$

Equations 4.3 and 4.4 define the type of equivalent circuit used to replace the BJT for the purposes of circuit analysis. The input circuit (see figure 4.11) is described mathematically by equation 4.3. This equates the applied voltage, V_b, to the sum of two other voltages, one being the p.d. in a resistor of value h_{ie} and the other being a voltage source whose value depends on the fraction of reverse feedback, h_r, within the transistor. Normally the value of the voltage $h_{re}V_c$ is small compared with the value of $h_{ie}I_b$.

Equation 4.4 mathematically describes the form of the output circuit. In this circuit (see figure 4.11), the collector current, I_c, is shared between a current source of value $h_{fe}I_b$ and a path of conductance h_{oe}.

The transistor parameters can be obtained by means of dynamic tests (that is, a.c. tests) in the required configuration. Alternatively the values of the parameters may be estimated from the input and output characteristics, shown for the common-emitter configuration in figure 4.12. From diagram (a)

$$h_{ie} = \frac{\delta v_{B1}}{\delta i_{B1}} \quad \text{ohms}$$

$$h_{re} = \frac{\delta v_{B2}}{\delta v_{C1}} \quad \text{dimensionless}$$

and from figure 4.12b

$$h_{fe} = \frac{\delta i_{C1}}{\delta i_{B2}} \quad \text{dimensionless}$$

$$h_{oe} = \frac{\delta i_{C2}}{\delta v_{C2}} \quad \text{siemens}$$

For example if, at the operating points on the input and output characteristics, the following measurements are taken

$$\delta i_{B1} = 5\ \mu\text{A}, \delta v_{B1} = 9\ \text{mV}, \delta v_{B2} = 3\ \text{mV}, \delta v_{C1} = 5\ \text{V}$$

$$\delta i_{B2} = 10\ \mu\text{A}, \delta i_{C1} = 1.05\ \text{mA}, \delta v_{C2} = 8\ \text{V}, \delta i_{C2} = 0.8\ \text{mA}$$

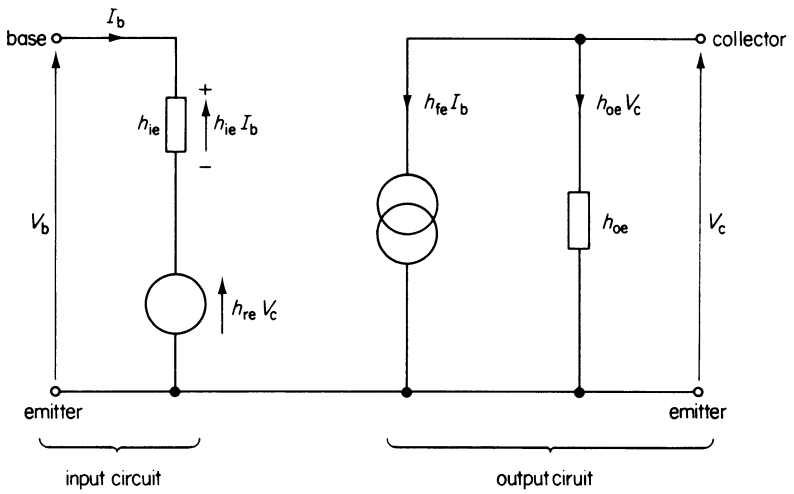


Figure 4.11 The small-signal *h*-parameter equivalent circuit of the bipolar junction transistor

then

$$h_{ie} = \frac{\delta v_{B1}}{\delta i_{B1}} = \frac{9 \times 10^{-3}}{5 \times 10^{-6}} = 1800\ \Omega$$

$$h_{re} = \frac{\delta v_{B2}}{\delta v_{C1}} = \frac{3 \times 10^{-3}}{5} = 0.0006$$

$$h_{fe} = \frac{\delta i_{C1}}{\delta i_{B2}} = \frac{1.05 \times 10^{-3}}{10 \times 10^{-6}} = 105$$

$$h_{oe} = \frac{\delta i_{C2}}{\delta v_{C2}} = \frac{0.8 \times 10^{-3}}{8} = 0.0001\ \text{S}$$

The above figures may be regarded as being typical of a low-power silicon transistor.

Even though great strides have been taken in the manufacture of transistors, devices of a given type show a wide spread of parameter values. For example, the parameters of the BC108 transistor, which is a general purpose silicon *n–p–n* type, show the spread depicted in table 4.1. Thus the actual value of any of the above parameters in a particular case lies in a range from about 50 per cent below the typical value to about 100 per cent above it, that is, a parameter spread of 3:1.

Other parameters of a transistor family also vary; an example is the voltage v_{BE}, which in the BC108 family may have a value between 0.55 and 0.7 V, a typical value being 0.62 V. The latter figures are quoted at a junction temperature of 25 °C and at a collector voltage of 5 V.

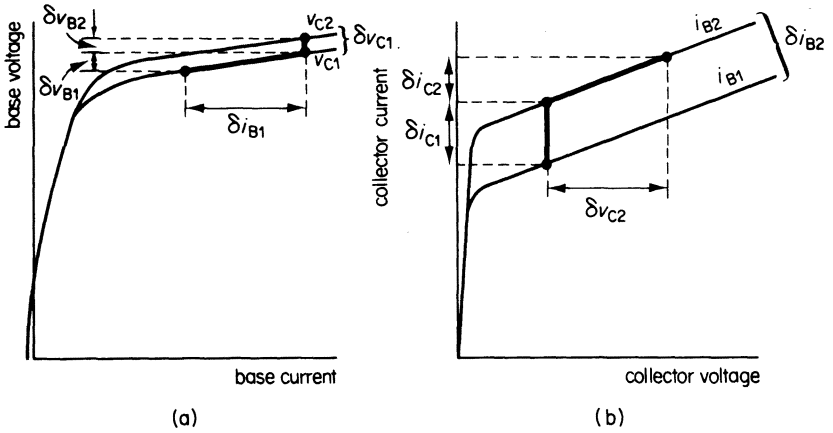


Figure 4.12 Estimating the common-emitter *h*-parameters

Table 4.1

Parameter	Minimum value	Typical value	Maximum value
h_{ie}(kΩ)	2.7	4.5	8.7
h_{re}(x 10^{-4})	1.5	2.0	3.0
h_{fe}	220	330	600
h_{oe}(μS)	18	30	60

When the transistor is operated as a switching device, engineers are primarily interested in the large-signal *h*-parameters which are defined in the common-emitter mode by the equations

$$v_{BE} = h_{IE} i_{BE} + h_{RE} v_{CE}$$

$$i_{CE} = h_{FE} i_{BE} + h_{OE} v_{CE}$$

4.10 Relationships between the Common-emitter, the Common-base and the Common-collector Parameters

The equations for the common-base and common-collector parameters, expressed in terms of the common-emitter parameters are

Common-base

$$h_{ib} = \frac{h_{ie}}{1 + h_{fe}}$$

$$h_{rb} = \frac{h_{ie} h_{oe}}{1 + h_{fe}}$$

Common-collector

$$h_{ic} = h_{ie}$$

$$h_{rc} = \frac{1}{1 + h_{re}}$$

Common-base	*Common-collector*
$h_{fb} = \dfrac{-h_{fe}}{1 + h_{fe}}$	$h_{fc} = -(1 + h_{fe})$
$h_{ob} = \dfrac{h_{oe}}{1 + h_{fe}}$	$h_{oc} = h_{oe}$

Using the values evaluated in the example above, that is, $h_{ie} = 1800\ \Omega, h_{re} = 0.0006$, $h_{fe} = 105$ and $h_{oe} = 0.0001$ S, the corresponding common-base and common-collector values are

$h_{ib} = 17\ \Omega$	$h_{ic} = 1800\ \Omega$
$h_{rb} = 0.0017$	$h_{rc} = 0.9994$
$h_{fb} = -0.991$	$h_{fc} = -106$
$h_{ob} = 0.94 \times 10^{-6}$ S	$h_{oe} = 0.0001$ S

4.11 A Basic Small-signal Linear Amplifier

The transistor amplifier circuit in figure 4.13a forms the basis of many advanced designs. In the following, some of the more important features of the circuit, together with an outline of its operation are given.

The operating point

In order that the amplifier can amplify or magnify the alternating signal, V_1, applied to its input, the transistor must be *biased* so that it operates in the linear region of its characteristics. The *operating point* or *quiescent point*, Q (see figure 4.13b), on the output characteristics indicates the steady value of the collector current (I_C) and of the collector voltage (V_C) when the a.c. input signal is zero, that is $V_1 = 0$; this is sometimes known as the no-signal condition. The quiescent point is established by means of the d.c. bias circuit consisting of the bias resistor, R_B, and the supply voltage V_{CC}. This circuit causes the d.c. bias current, I_B, to flow in the base circuit, and is known as the *quiescent value of base current*. The quiescent point lies on the output characteristic corresponding to a base current of I_B (see figure 4.13b). The value of I_B can be calculated from the following equation.

$$I_B = \frac{V_{CC} - V_B}{R_B}$$

where V_{CC} is the value of the supply voltage, and V_B is the d.c. value of the base–emitter voltage. In the case of a silicon transistor, $V_B \approx 0.6$ V, hence

$$I_B \approx \frac{V_{CC} - 0.6}{R_B}$$

In some cases $V_{CC} \gg 0.6$ V, when $I_B \approx V_{CC}/R_B$.

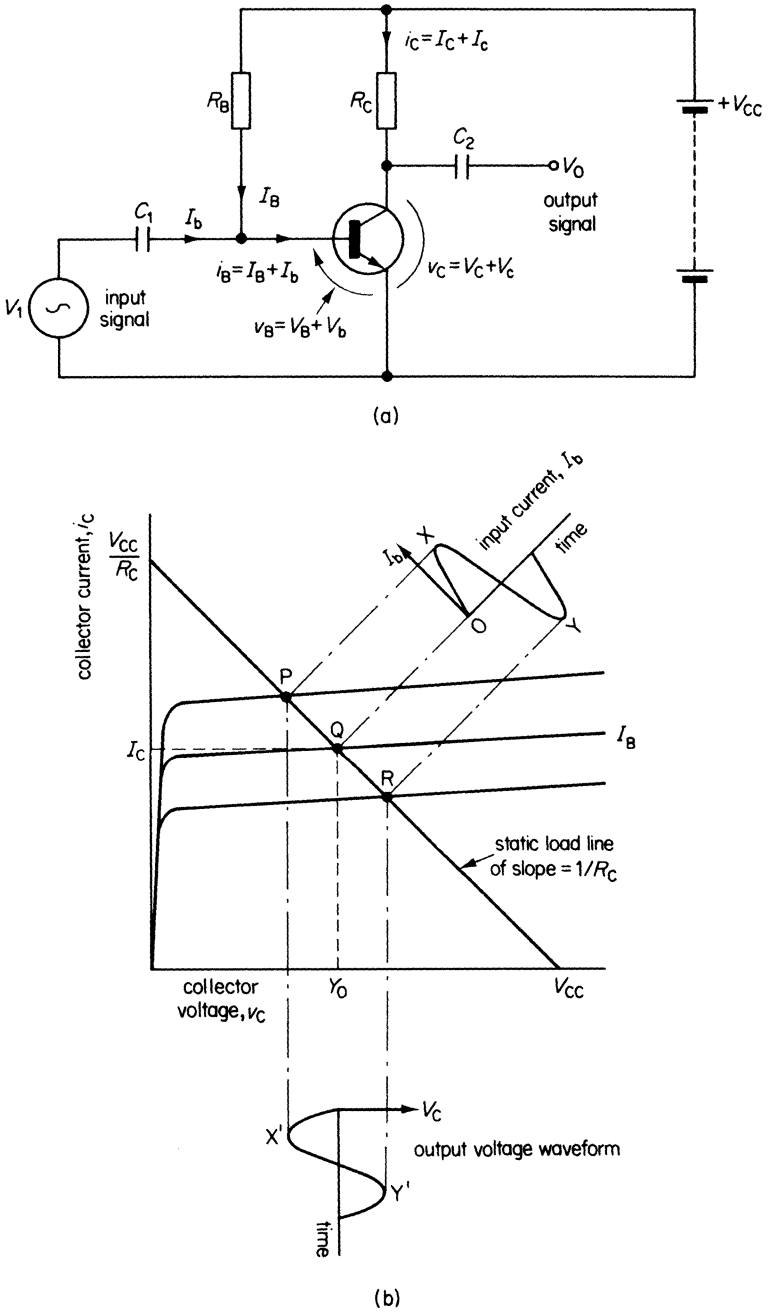


Figure 4.13 (a) A circuit diagram of a simple a.c. coupled amplifier and (b) the load-line construction

Capacitive coupling

To prevent the resistance of the a.c. signal source from upsetting the bias conditions, capacitor C_1 (known as a *blocking capacitor*) is connected between the signal source and the transistor. At zero frequency (d.c.) the reactance of the capacitor is infinitely large, and it effectively blocks or prevents the flow of direct current from the bias circuit into the signal source. Capacitor C_2 at the output of the amplifier also fulfils the function of a blocking capacitor, and prevents the quiescent collector voltage from being applied to the load.

The capacitances of both C_1 and C_2 (typically 25 μF or greater) are chosen so that, even at the lowest frequency of operation, their reactances are small compared with the input impedance of the amplifier and of the load, respectively, and allow a.c. signals to pass unimpeded.

Principle of operation

Under no-signal conditions ($V_1 = 0$) the circuit is in its 'quiet' or quiescent state, and the collector current and voltage have steady values. When an a.c. signal is applied, it causes an *alternating component* of current, I_b, to flow in the base circuit, the total base current being expressed by the relationship

$$\text{total base current} = i_B = I_B + I_b$$

In turn, this base current results in a collector current of

$$i_C = I_C + I_c$$

where I_C is the quiescent value of the collector current, and I_c is the *alternating component* of the collector current and is due to the flow of I_b in the base circuit. That is, an increase in base current (due to an increase in V_1) causes the collector current to increase in value. A direct result of the increase in collector current is an increase in the p.d. across resistor R_C in the collector circuit. Now the value of the voltage at the collector of the transistor is given by the equation

$$v_C = V_{CC} - i_C R_C$$

Hence, since the p.d. $i_C R_C$ increases with i_C, the collector voltage reduces as the value of i_C increases. That is, when the signal voltage V_1 instantaneously has a positive potential, then the instantaneous value of the collector voltage decreases. Consequently if V_1 has a sinusoidal waveform, then the output a.c. waveform has a (– sine) curve, that is, the amplifier is *phase inverting*. This feature is illustrated in figure 4.13b; an input signal that causes the alternating component, I_b, of the base current to *increase* from O to X results in the instantaneous operating point shifting to point P, when the collector voltage *falls* to X′. An input signal which causes I_b to *reduce* to Y results in the instantaneous operating point moving to R, and the collector voltage *rising* to Y′.

As mentioned earlier capacitor C_2 functions as a blocking capacitor which

prevents the quiescent component, V_C, of the collector voltage from being transmitted to the output. In this way only the a.c. component of the collector signal appears at the output terminals. For this reason the circuit is described as an *a.c. coupled amplifier* since, due to the action of the blocking capacitors, it amplifies only a.c. signals.

The static load line

Applying Kirchhoff's second law to the collector circuit yields

$$V_{CC} = v_C + i_C R_C$$

or

$$v_C = V_{CC} - i_C R_C$$

The equation for v_C is that of a straight line terminating at the lower end (when $i_C = 0$) at $v_C = V_{CC}$, and at the upper end (when $v_C = 0$) at $i_C = V_{CC}/R_C$. This line is known as the *static load line* of the circuit, and is illustrated in figure 4.13b. The slope of this line is $(-1/R_C)$ siemen.

The operating point of the amplifier lies at the intersection of this line with the output characteristic for the operating value of base current. The quiescent point, Q, is given by the intersection of the load line with the output characteristic corresponding to the quiescent base current, I_B. In this type of circuit the quiescent point is generally selected so that the value of the quiescent collector voltage is equal to about one-half of the supply voltage.

The dynamic load line or a.c. load line

In practice the output signal from the amplifier in figure 4.13a is used to drive current into an electrical load, symbolised by R_L in figure 4.14a.

In the absence of an input signal, the operating point lies at the quiescent point, Q, on the static load line. When an a.c. signal is applied to the amplifier, capacitor C_2 (figure 4.14a) has a very low reactance to this signal, and allows it to be transmitted to the load R_L. Hence the value of the 'a.c. load' resistance or *dynamic load* resistance, R'_L, is equivalent to R_C in parallel with R_L, or

$$R'_L = \frac{R_C R_L}{R_C + R_L}$$

A line known as the *a.c. load line* or *dynamic load line*, of slope $(-1/R'_L)$ and passing through point Q on the static load line (see figure 4.14b), is used to determine the actual value of the output voltage from the amplifier. Since $R'_L < R_C$, the slope of the dynamic load line is steeper than that of the static load line, so that for a given change in base current, δi_B, the change in output voltage, δv_{C2} (figure 4.14b), is less than the change in output voltage (δv_{C1}) which occurs in the unloaded amplifier. That is, the *voltage gain* of the amplifier with R_L

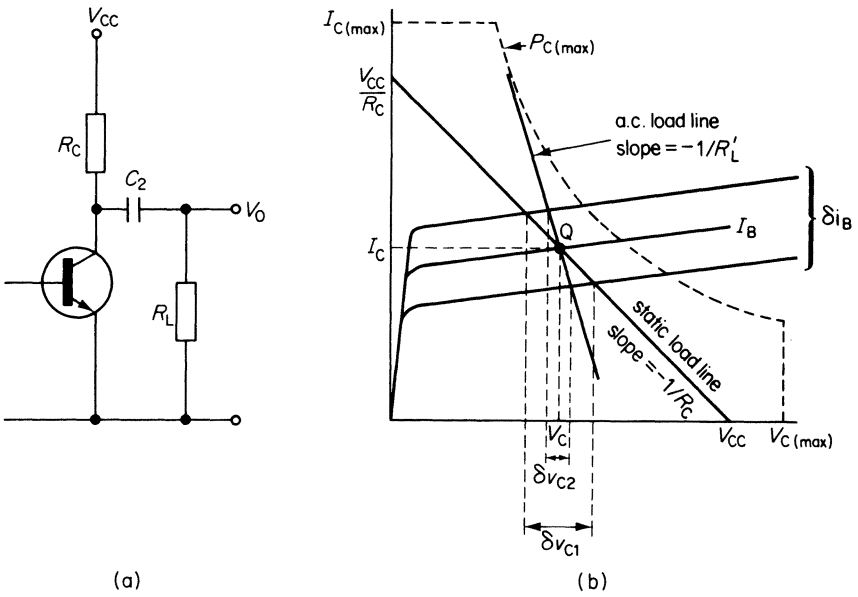


Figure 4.14 Construction of the dynamic load line

connected is less than in the case when R_L is not connected. If K_{v0} is the voltage gain of the unloaded amplifier and K_{vL} is the voltage gain of the loaded amplifier, it can be shown that the following approximate relationship holds good.

$$K_{vL} = \frac{K_{v0}R_L}{R_C + R_L}$$

To prevent the possibility of damage from thermal runaway, or from overvoltage or overcurrent, the load line should lie within the safe working area bounded by the $P_{C(max)}$ curve, the $V_{C(max)}$ line and the $I_{C(max)}$ line (see figure 4.14b).

The r.m.s. value of the output voltage for a sinusoidal input signal under no-load conditions ($R_L = \infty$) from figure 4.14b is $\delta v_{C1}/(2\sqrt{2})$, and when loaded is $\delta v_{C2}/(2\sqrt{2})$. The voltage gain in the two cases is

$$\text{unloaded voltage gain} = K_{v0} = \frac{\delta v_{C1}}{2V_1\sqrt{2}}$$

$$\text{loaded voltage gain} = K_{vL} = \frac{\delta v_{C2}}{2V_1\sqrt{2}}$$

4.12 Thermal Stability

A feature of the fixed bias circuit in figure 4.13a is that if the ambient temperature changes, then changes occur in v_{BE}, h_{FE} and in the leakage current. The net result

is an upward drift of the individual characteristics with increasing temperature. The implication is that the quiescent point shifts along the load line towards saturated working. The practical consequences are a change in the quiescent working conditions and a change in the voltage gain of the circuit. It may also lead to the output signal becoming distorted at high values of temperature due to the shift towards a region on the characteristics where the curves are more cramped together. A number of circuits have been designed that limit the shift in the Q-point with temperature change. Two circuits are considered here.

Collector-voltage bias circuit

The circuit in figure 4.15 is used in many low-cost amplifiers, and provides improved thermal stability when compared with the fixed-bias circuit. In this circuit, capacitors C_1 and C_2 are blocking capacitors. In figure 4.15 the base bias current is obtained from the collector voltage via the resistor chain R_B. The reason for dividing the resistor into two parts, and for C_3 is given later. The reason for the improvement in thermal stability over the fixed-bias circuit is now given.

Assume for the moment that the ambient temperature is increasing. This causes the quiescent value of the collector current to increase, and with it the collector voltage reduces. Since the base current is obtained from the collector voltage, the above reduction in collector voltage results in a reduced value of d.c. base current. This reduction in base current reduces the rise in collector current with temperature to a value less than that which occurs in the fixed-bias circuit.

The values of the components used in figure 4.15 can be estimated as follows. Suppose that V_{CC} = +9 V, that the value of the transistor parameter h_{FE} is 91, and that the quiescent value of the base–emitter voltage is 0.6 V. the quiescent collector voltage should also be about $V_{CC}/2$, that is, about 4.5 V. If the quiescent

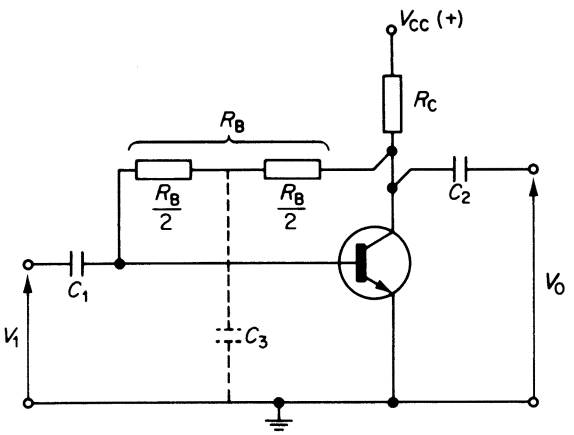


Figure 4.15 A common-emitter amplifier using collector-voltage bias

collector current is 1 mA, then

$$R_C = \frac{V_{CC} - V_C}{I_C} = \frac{9 - 4.5}{10^{-3}} = 4500\ \Omega$$

Resistor values are selected from a *preferred value* range (the preferred range of 10 per cent tolerance resistors includes decade multiples of 1, 1.2, 1.5, 1.8, 2.2, 2.7, 3.3, 3.9, 4.7, 5.6, 6.8 and 8.2), and a 4.7 kΩ resistor is selected. The quiescent base current is about $I_C/h_{FE} = 1/91 = 0.011$ mA. The value of R_B is computed from the equation

$$R_B = \frac{V_C - V_B}{I_B} = \frac{4.5 - 0.6}{0.011 \times 10^{-3}} = 354.5\ \text{k}\Omega$$

For reasons discussed below, R_B is constructed from two series-connected resistors each of value 180 kΩ. Resistor R_B not only provides the base bias current but, when an alternating input signal is applied to the amplifier, it also feeds back a current that is proportional to the alternating component of the collector voltage. This latter signal has the effect of reducing the voltage gain of the amplifier. The unwanted a.c. feedback signal can be prevented from reaching the base circuit by connecting the centre-point of R_B to the chassis by means of capacitor C_3. The value of C_3 is chosen so that its reactance at the lowest frequency of operations, f_L, is much less than the resistance of $R_B/2$. That is

$$\frac{1}{2\pi f_L C_3} \approx \frac{R_B/2}{10}$$

or

$$C_3 \approx \frac{10}{\pi f_L R_B}$$

If $f_L = 32$ Hz, then $C_3 \approx 0.28$ μF. In practice a capacitance value greater than this would probably be chosen. Capacitor C_3 in figure 4.15 is known as a *decoupling capacitor*, since it reduces the electrical coupling between the two sections of the circuit, and prevents the unwanted feedback signal from reaching the input terminals.

Self-bias or emitter bias circuit

The circuit in figure 4.16 offers improved thermal stability when compared with that in figure 4.15. In this circuit the value of the quiescent base voltage, V_B, is fixed at a constant value by the resistor chain R_1, R_2. The circuit operation is now described.

As the ambient temperature increases, so the collector current tends to increase, and with it the p.d. across R_E increases. The base–emitter d.c. voltage is given by

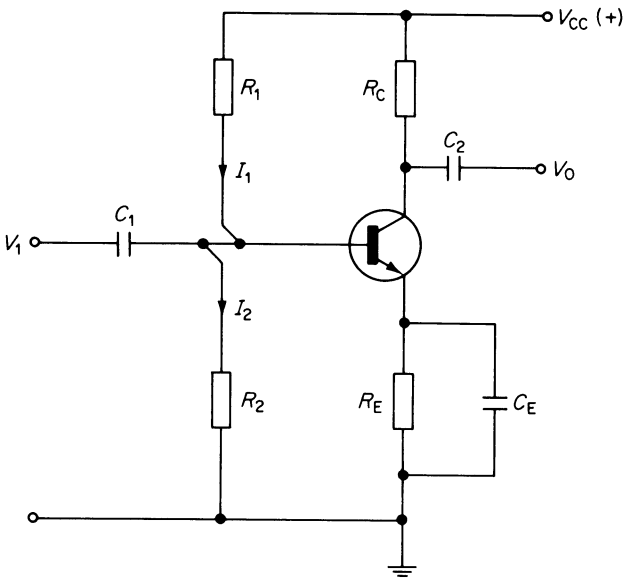


Figure 4.16 A self-biasing circuit

the expression

$$V_{BE} = V_B - I_E R_E$$

Consequently as I_E increases so V_{BE} reduces. In turn this reduces the value of the base current; the net result is that the change of collector current with temperature variation is reduced to a low value.

In figure 4.16 capacitor C_E is a decoupling capacitor and provides a low reactance path for the flow of the alternating component of the emitter current. Consequently we may say that the d.c. component of the emitter current flows in resistor R_E, and the a.c. component flows through C_E. The reactance of C_E at the lowest frequency of operation should be much less than the resistance of R_E, or

$$\frac{1}{2\pi f_L C_E} \approx \frac{R_E}{10}$$

or

$$C_E \approx \frac{10}{2\pi f_L R_E}$$

If f_L = 32 Hz, then $C_E \approx 0.05/R_E$ F (R_E in Ω) $\approx 50/R_E$ μF (R_E in kΩ).

The values of the circuit elements can be estimated using relatively simple methods, as shown in the following example. Suppose that V_{CC} = +9 V, and that the value of the transistor parameter h_{FE} is 75 and that V_{BE} is 0.6 V. To provide adequate thermal stability, the p.d. across R_E should be at least 10 per cent of the

value of V_{CC}; also, from a practical viewpoint, the value of R_2 should be about ten times that of R_E. Hence, with a supply voltage of 9 V, the p.d. across R_E should be greater than about 0.9 V. Let us assume that the quiescent collector current is to be 1 mA. Hence

$$R_E \geqslant \frac{0.9\text{ V}}{1\text{ mA}} \geqslant 0.9\text{ k}\Omega$$

A preferred value of 1 kΩ is used for R_E, and the mean value of the p.d. across R_E is about 1 V. From the above work on C_E

$$C_E \approx \frac{50}{1}\,\mu\text{F} = 50\,\mu\text{F}$$

Since V_{BE} is 0.6 V, the quiescent value of base voltage is 1.6 V. Also, since $R_2 \approx 10\,R_E = 10\text{ k}\Omega$, then the value of current I_2 that flows in R_2 is

$$I_2 = \frac{V_B}{R_2} = 0.16\text{ mA}$$

and

$$I_B = \frac{I_C}{h_{FE}} = \frac{1}{75} = 0.013\text{ mA}$$

Hence

$$I_1 = I_2 + I_B = 0.173\text{ mA}$$

The estimated value of R_1 is

$$R_1 = \frac{V_{CC} - V_B}{I_1} = \frac{9 - 1.6}{0.173 \times 10^{-3}}\,\Omega = 42.8\text{ k}\Omega$$

The value computed for R_1 above is at the upper limit of a 39 kΩ, 10 per cent tolerance resistance, and at the lower end of a 47 kΩ, 10 per cent tolerance resistance. Either of the two values given above may be chosen.

Since the supply voltage is 9 V and the emitter voltage is 1 V, a p.d. of 8 V appears across the series combination consisting of R_C and the transistor. In the quiescent state, a p.d. of about one-half of this value should appear across R_C, hence

$$R_C \approx \frac{4\text{ V}}{1\text{ mA}} = 4\text{ k}\Omega$$

A preferred value of 3.9 kΩ would be selected for R_C.

The no-load voltage gain of the circuit described above is about 330, and is computed using the relationships developed in section 4.14.

4.13 Stability Factors

The transistor current is a function of a number of factors including I_{CEO}, v_{BE} and h_{FE}, which are temperature-dependent; the total change ΔI_C in the collector current over a given temperature range is given by the expression

$$\Delta I_C = S\Delta I_{CBO} + S'\Delta v_{BE} + S''\Delta h_{FE}$$

where S, S' and S'' are *stability factors* (or, more correctly, instability factors) corresponding to changes ΔI_{CBO}, Δv_{BE} and Δh_{FE} in the leakage current, base–emitter voltage and current gain, respectively. The stability factors are described by the partial derivatives below.

$$S = \frac{\partial I_C}{\partial I_{CBO}}$$

$$S' = \frac{\partial I_C}{\partial v_{BE}}$$

$$S'' = \frac{\partial I_C}{\partial h_{FE}}$$

For the circuit in figure 4.16

$$S = \frac{R_E + R_b}{[R_E + R_b/(1 + h_{FE})]}$$

where $R_b = R_1R_2/(R_1 + R_2)$

$$S' \approx -1/R_E$$

$$S' \approx \frac{I_{C1}S}{h_{FE1}h_{FE2}}$$

where h_{FE1} is the current gain at collector current I_{C1} and h_{FE2} is the current gain at some other value of collector current. The value of S usually lies between about $0.1h_{FE}$ and $0.2h_{FE}$.

4.14 *h*-parameter Analysis of a Linear Amplifier

The analysis of a linear amplifier using the hybrid parameters is best carried out using the block diagram in figure 4.17. The appropriate equations are

$$V_1 = h_iI_1 + h_rV_2$$

$$I_2 = h_fI_1 + h_oV_2$$

$$V_2 = -I_2R'_L$$

The resistance R_S is the internal resistance of the signal source. The general *h*-parameters are used in the above equations, and the resulting equations are valid

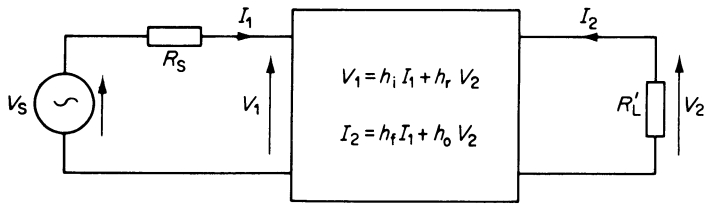


Figure 4.17 *h*-parameter analysis of a linear circuit

for any of the three configurations, provided the parameters for that configuration are inserted into the appropriate equations. Resistor R'_L is the effective load or a.c. load connected to the amplifier.

Solving the above equations for the current gain (K_i), the input resistance (R_{in}), the voltage gain (K_v), the power gain (K_p) and the output resistance (R_{out}) yields

$$\text{current gain} = K_i = \frac{I_2}{I_1} = \frac{h_f}{1 + h_o R'_L}$$

$$\text{input resistance} = R_{in} = \frac{V_1}{I_1} = h_i - h_r R'_L K_i$$

$$\text{voltage gain} = K_v = \frac{V_2}{V_1} = -\frac{K_i R'_L}{R_{in}}$$

$$\text{power gain} = K_p = \frac{K_i^2 R'_L}{R_{in}} = \left| K_i K_v \right|$$

$$\text{output resistance} = R_{out} = \frac{1}{[h_o - h_f h_r/(h_i + R_S)]}$$

4.15 The Transistor as a Switch – the NOT Gate

An elementary form of transistor switching circuit, known as a *resistor–transistor logic* (RTL) NOT gate, is shown in figure 4.18a. A circuit symbol for the NOT gate is shown in figure 4.18b. The input signal, A, is provided by a switch that could, for example, be mounted in a sensor or transducer in a vending machine or on a piece of machinery or on a conveyor. The electrical voltage applied to resistor R_B could either be zero or could be V_{CC}.

Since the input voltage can have one of two possible values, it is described as a *binary* signal; for convenience, signal A is said to have a binary '0' value or logic '0' value when the applied voltage is zero, and is said to have binary '1' or logic '1' value when the applied voltage is V_{CC}. When a logic '0' signal is applied to R_B, the base current is zero and the transistor is cut off, that is, the collector current is zero. In this condition no current flows in R_C, and the collector voltage is

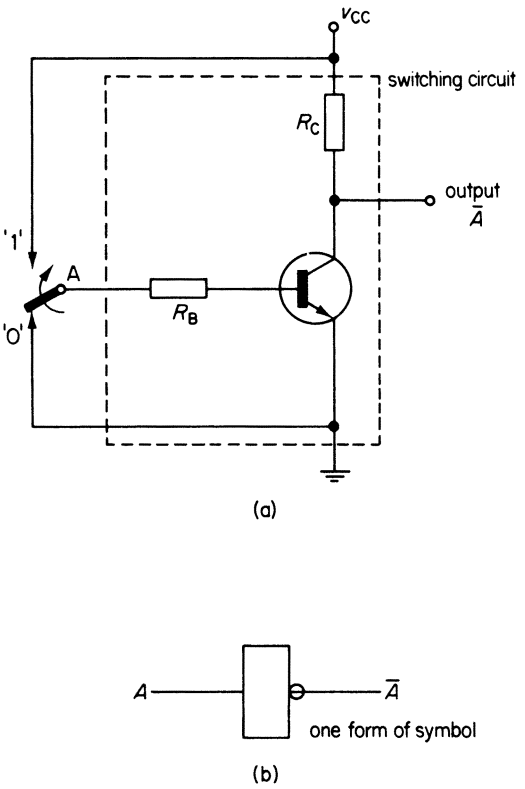


Figure 4.18 A basic resistor–transistor logic (RTL) switching circuit

practically equal to V_{CC}. Hence, when the input signal is logic '0', then the output signal is logic '1'. This type of circuit therefore provides an output signal which is the *logical complement* or *logical inverse* of the input signal. That is to say, **the logic signal at the output of the gate is NOT equal to the logic signal applied to the input.** The output signal, f, is given by the expression

$$f = \text{NOT } A = \overline{A}$$

The NOT function or negation function is represented by placing a bar over the top of the function, illustrated above.

The circuit shown (and other logic circuits for that matter) is described as a *gate* since it can either be open and allow information to pass, or it can be closed to the flow of information.

When the input signal is switched to the logic '1' state, current flows into the base of the transistor. The value of R_B is chosen so that the base current is large enough to saturate the transistor. In this operating state, the p.d. across the

transistor falls to a low value, typically 100 to 300 mV. Since this value is low when compared with the supply voltage, V_{CC}, the collector current, $I_{C(sat)}$, in the saturated state is approximately

$$I_{C(sat)} \approx \frac{V_{CC}}{R_C} \tag{4.5}$$

The operational current gain of the transistor in the saturated state is designated the symbol $h_{FE(sat)}$, and

$$I_{C(sat)} = h_{FE(sat)} I_{B(sat)} \tag{4.6}$$

where $I_{B(sat)}$ is the base current in the saturated state. The value of the current gain in the saturated state may be as low as 10 or 20. Also, when saturated, the base–emitter p.d., $V_{BE(sat)}$, of a silicon transistor is typically 0.7 V, hence

$$I_{B(sat)} = \frac{V_{CC} - V_{BE(sat)}}{R_B}$$

If it can be assumed that $V_{BE(sat)} \ll V_{CC}$, then

$$I_{B(sat)} \approx \frac{V_{CC}}{R_B} \tag{4.7}$$

Substituting equations 4.5 and 4.7 into equation 4.6 yields

$$R_B \approx h_{FE(sat)} R_C$$

Hence if $V_{CC} = 9$ V and $h_{FE(sat)} = 33$, then if the maximum collector current is 9 mA, suitable values for R_C and R_B are 1 kΩ and 33 kΩ, respectively. In practice, to allow for variations in supply voltage, component tolerance and for external loading, a value of about one-quarter of the theoretical value of R_B is used. That is, a resistor of preferred value 8.2 kΩ would be used for R_B.

The operation of a logic gate or logic network is described by its *truth table*, which gives the logical output from the circuit for all possible input conditions. The truth table for the NOT gate is given in table 4.2.

Table 4.2 Truth table for a NOT gate

Input A	Output $f = \bar{A}$
0	1
1	0

Table 4.3 Truth tables for three-input OR and AND gates

Inputs			OR gate output	AND gate output
A	B	C	$f = A + B + C$	$f = A.B.C.$
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	1	0
1	0	0	1	0
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

4.16 OR and AND Logic Functions

A *logic-OR gate*† is an electronic circuit having a number of input lines that are independently energised by logic signals, and if *any* input line is energised by a logic '1', then the output is logic '1'. Since each input line may have either logic '0' or logic '1' applied to it then, for a gate with N input lines, there are 2^N possible combinations of input signal in the truth table. The truth table for a three-input OR gate is shown in table 4.3; since there are three input lines, there are $2^3 = 8$ possible combinations of input signal.

The truth table indicates that the output from the OR gate is logic '1' whenever any input line has a logic '1' applied to it. The logic-OR function is symbolised by the plus (+) sign; thus

$$f = A \text{ OR } B \text{ OR } C = A + B + C$$

A *logic-AND gate* is a circuit having a number of input lines that are independently energised by logic signals; the AND gate provides an output of logic '1' only when *all* the input lines are simultaneously energised by logic '1' signals. The truth table for a three-input AND gate is also given in table 4.3. The AND function is symbolised by a dot (.) symbol, as follows

$$f = A \text{ AND } B \text{ AND } C = A.B.C.$$

4.17 NOR and NAND Gates

The majority of practical logic gates generate what are known as NOR and NAND functions, described below. The NOR function is expressed by the statement

$$\text{NOR} = \text{NOT OR}$$

†The topic of electronic logic is dealt with in depth in Noel M. Morris, *Digital Electronic Circuits and Systems*, Macmillan, 1974.

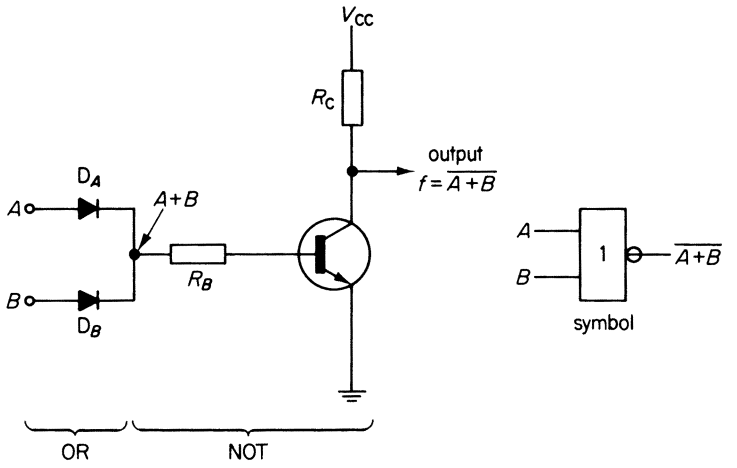


Figure 4.19 A basic form of DTL NOR gate

That is, the output from a NOR gate is the logical complement of the output from an OR gate which has the same number of input lines. A two-input *diode–transistor logic* (DTL) NOR gate is illustrated in figure 4.19. The diodes D_A and D_B form the logic OR function of signals A and B at their common cathode connection. The remainder of the circuit is that of a NOT gate, which inverts the OR function, to give the over-all NOR function at the output. The truth table for the circuit is given in table 4.4. In the case of figure 4.19, the logical output, f, is expressed as follows

$$f = \text{NOT}\,(A + B) = \overline{A + B}$$

The NAND function is expressed by the statement

$$\text{NAND} = \text{NOT AND}$$

In the case of a two-input NAND gate (see figure 4.20), the logical output, f, is

Table 4.4 Truth table for two-input NOR and NAND gates

Inputs		NOR gate output	NAND gate output
A	B	$f = \overline{A + B}$	$f = \overline{A.B}$
0	0	1	1
0	1	0	1
1	0	0	1
1	1	0	0

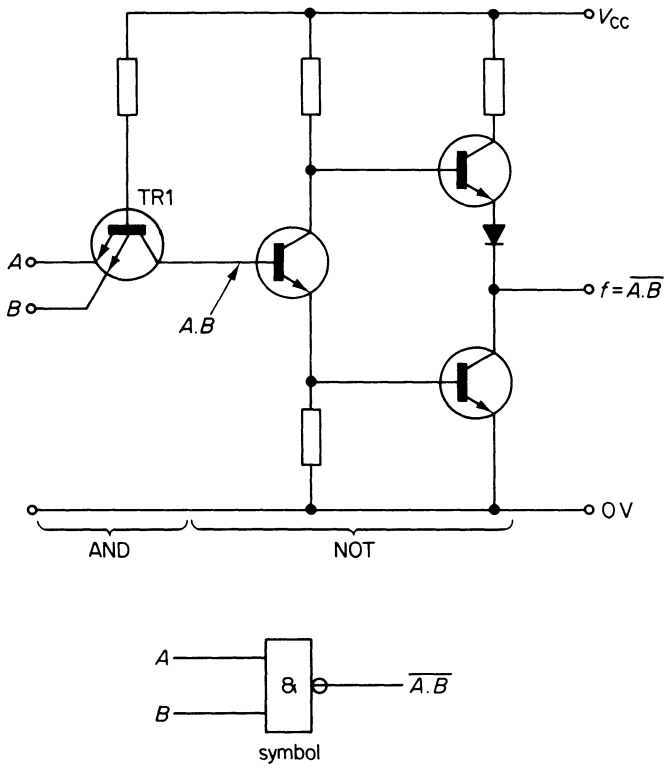


Figure 4.20 A two-input TTL NAND gate

expressed in the form

$$f = \text{NOT } (A \text{ AND } B) = \overline{A.B}$$

A two-input *transistor–transistor logic* (TTL) NAND gate is shown in figure 4.20. In this circuit, the AND function of the input signals is formed at the collector of the multiple-emitter transistor TR1. The remainder of the circuit is a high-speed logic signal invertor, which gives the over-all NAND function at the output. The truth table of the circuit is also given in table 4.4.

5 *Field-effect Transistors, Amplifiers and Logic Gates*

5.1 Types of Field-effect Transistor

A field-effect transistor (FET) is a device that depends for its operation on the control of a current by means of an electric field in a semiconductor. There are two principal types of FET, namely (a) the *junction-gate FET* (JUGFET, JFET or simply FET), and (b) the *insulated-gate FET* (IGFET, MOSFET or MOST). The operating principles of the two types differ from one another, and are examined in this chapter. The most popular application of JUGFETs is in input stages of high-input impedance linear amplifiers, while IGFETs are widely used in digital integrated circuits.

A family tree showing branches of the FET family is given in figure 5.1. Natural forms of JUGFET are depletion-mode devices (this term is explained in section 5.3), while natural forms of IGFET are enhancement-mode devices (see section 5.9).

Field-effect transistors differ from bipolar junction transistors in several respects, including the following.

(1) Current flow in FETs consists only of majority charge carriers, and is therefore *unipolar* (that is, only one type of charge carrier is involved).
(2) Control of current flow is by means of an electrical field and, theoretically, no current is drawn from the input signal source. The input resistance is therefore very high (typically many tens or hundreds of megohms).
(3) FETs are not only simpler to fabricate than BJTs in IC form (see chapter 6), but also occupy less space.
(4) The gain–bandwidth product of FETs is generally smaller than that of BJTs.

5.2 The Junction-gate Field-effect Transistor

The basic structure of an *n*-channel JUGFET is shown in figure 5.2a, and consists of an *n*-type conducting bar or channel which has two p^+-type regions diffused into it

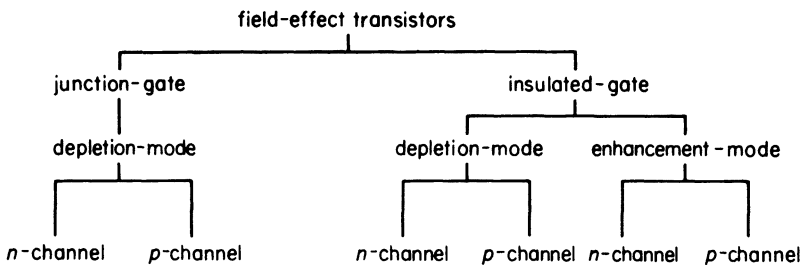


Figure 5.1 FET family tree

on opposite sides of the bar. A p.d. v_{DS}, is applied between the ends of the bar, and current flows longitudinally through it. Since the bar is made of *n*-type material, current flow through it is due to the movement of electrons from the negative pole of the supply to the positive pole. For this reason the end of the conducting channel connected to the negative pole of the supply is known as the *source* electrode; the end connected to the positive pole is known as the *drain* electrode.

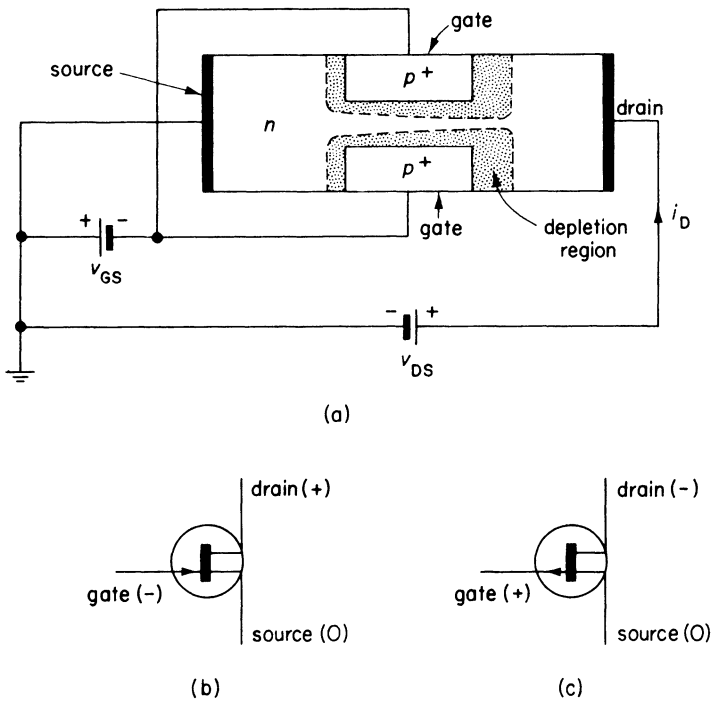


Figure 5.2 (a) The basic structure of an *n*-channel JUGFET showing the supply polarities, (b) symbol for an *n*-channel JUGFET and (c) a symbol for a *p*-channel JUGFET

The heavily doped (p^+) *gate regions* form p–n junctions with the conducting channel. The gate-to-source p–n junctions are operated in a reverse biased mode, and depletion regions form around the junctions in the manner shown in figure 5.2a. The magnitude of the reverse bias voltage at any point under the gate region is produced by two voltages which are, firstly, the reverse bias v_{GS} applied between the gate and the source and, secondly, the p.d. due to the potential divider effect of the conducting channel when connected to the drain-to-source voltage, v_{DS}. The former produces a depletion region whose depth is proportional to the gate bias voltage, v_{GS}. The magnitude of the latter p.d. increases in value from zero at the source electrode to a maximum at the drain, and results in a depletion region which increases in depth from the left-hand end of the gate region (see figure 5.2a).

A circuit symbol for an n-channel JUGFET is shown in figure 5.2b. The drain and source electrodes in the symbol are linked by a line that symbolises the conducting channel existing between them. The arrow on the gate electrode points from the p-region to the n-region, that is, the convention for p–n junction diodes is followed. The polarities of the voltages applied to the electrodes of an n-channel device are also illustrated in figure 5.2b.

p-channel devices with n-type gate regions are also manufactured, but are less popular than n-channel FETs since the lower mobility of the charge carriers (holes) results in an inferior high-frequency performance to that of n-channel devices. The circuit symbol for an n-channel JUGFET is shown in figure 5.2c.

5.3 Characteristics of an n-channel JUGFET

A typical set of common-source output characteristics for an n-channel JUGFET is shown in figure 5.3a. The characteristics show how the drain current, i_D, varies

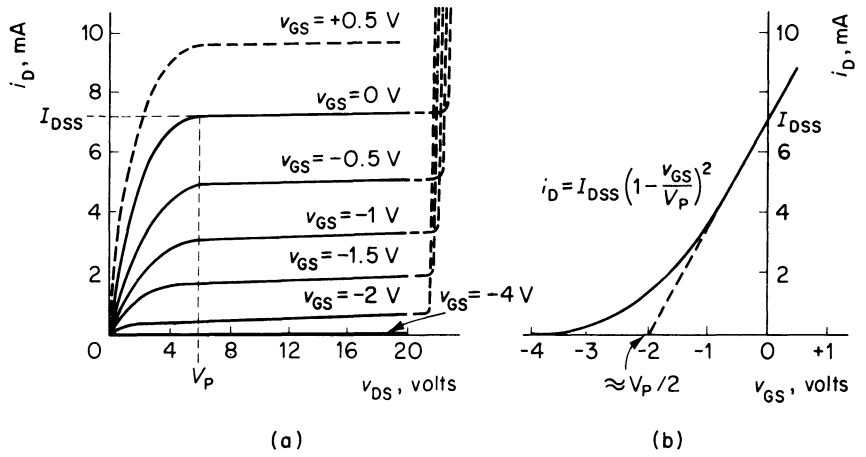


Figure 5.3 (a) Drain or output characteristics for an n-channel JUGFET and (b) the transfer, mutual or transconductance curve for v_{DS} = 14 V

with the drain voltage, v_{DS}, for various values of gate voltage, v_{GS}. A qualitative explanation of the shape of the characteristics follows.

Each output characteristic can broadly be divided into four regions, three of which are defined in terms of a voltage known as the *pinch-off voltage*, V_P, which is defined below. The four regions are

(1) $v_{DS} < V_P$ (resistive region)
(2) $v_{DS} = V_P$ (onset of pinch-off)
(3) $v_{DS} > V_P$ (pinch-off region)
(4) breakdown region

For convenience, the descriptions which follow refer to the $v_{GS} = 0$ characteristic, that is, the output curve corresponding to the condition when the gate is directly connected to the source electrode.

$v_{DS} < V_P$ (*resistive region*)

When operating with a low value of drain voltage, the gate depletion region does not extend very far into the conducting channel, as indicated in figure 5.4a. The value of the drain current is limited only by the resistance of the source-to-drain conducting channel, so that an increase in the drain voltage produces a proportional increase (or nearly so) in drain current. Thus, for low values of v_{DS}, the FET acts as though it were a linear resistance. In the case of the output characteristic for $v_{GS} = 0$ in figure 5.3a, the value of this resistance is about 300 Ω.

$v_{DS} = V_P$ *(onset of pinch-off)*

When the drain voltage reaches a value known as the *pinch-off voltage*, V_P, the depletion regions associated with the reverse biased p–n gate–channel junctions has extended into the channel until the current is 'pinched off' into a very thin film (see figure 5.4b). From this argument it would seem that it is possible to completely cut off the current flow simply by increasing the value of the source voltage. This is, of course, an impractical proposition since if the current were cut off, the p.d. in the channel would be zero and the depletion regions (which have caused the current to be pinched off) would vanish. Consequently the drain current cannot be cut off simply by increasing the value of the drain voltage.

$v_{DS} > V_P$ *(pinch-off region)*

For drain voltages greater than V_P, the value of the drain current remains fairly constant. The value of the drain current for $v_{GS} = 0$ is known as the *drain–source saturation current*, I_{DSS} (see figure 5.3a).

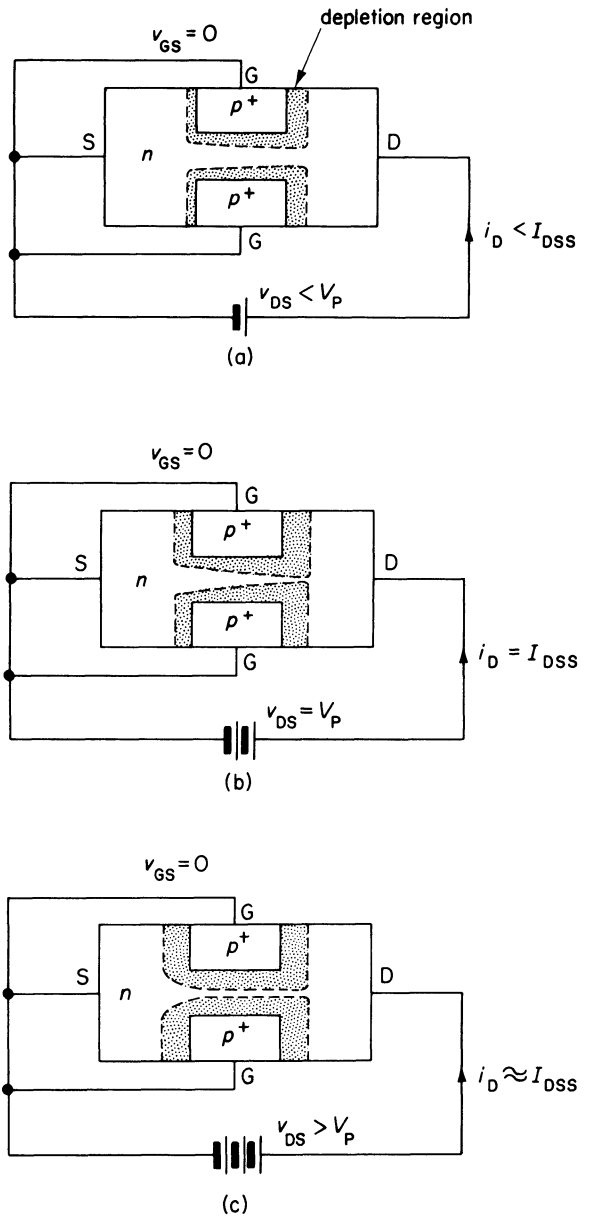


Figure 5.4 The effect of v_{DS} on the depth of the depletion region

A reason for the drain current assuming a constant value under this operating condition is found in the basic relationships of semiconductor devices, as follows. The current in a semiconductor is dependent on the drift velocity, v, of electrons, where

$$v = \mu E$$

in which μ is the mobility of the electrons (see also section 1.6), and E is the electric field strength. For field strengths greater than about 10^5 V/m it is found that, in n-type silicon, electron mobility is inversely proportional to E. This value of field strength is encountered in FETs at pinch-off; consequently the drift velocity (and current) in the channel remain constant.

In this region of operation, the width of the conducting channel remains constant but, as v_{DS} is increased, the length of the channel that is 'pinched off' increases (see figure 5.4c).

Breakdown region

At a high value of v_{DS}, the drain current suffers a dramatic increase (see figure 5.3a) and, if not limited in value, results in catastrophic failure of the device. The mechanism involved is a reverse breakdown of the p–n junction.

The effect of varying v_{GS} – depletion-mode operation

Increasing the value of v_{GS}, that is, making it more negative, causes the depletion-region depth to increase and the width of the conducting channel to reduce, with a consequent reduction in the value of the drain current. Thus increasing the value of v_{GS} has the effect of reducing or depleting the magnitude of the drain current; as a result, a JUGFET operating with a reverse biased gate junction is said to operate in *depletion mode*. JUGFETs are normally biased to operate in this mode.

It is also possible to operate JUGFETs with a small value of forward bias applied to the gate – up to about +0.5 V in the case of an n-channel device – to give an increase in drain current above I_{DSS} (see figure 5.3a). The effect of the forward bias is to reduce the depth of the depletion region at the gate junction, and to reduce the resistance of the channel. A voltage greater than about 0.5 V would cause the p–n junction to be forward biased, when the gate would draw current from the signal source; JUGFETs are not normally operated in this mode.

JUGFET transfer characteristics (figure 5.3b)

A transfer characteristic is a curve relating the drain current to the gate current at a particular value of v_{DS}. There is a family of such curves for each FET. The curve in figure 5.3b is for v_{DS} = 14 V. The transfer characteristics are also called the *mutual characteristics* or the *transconductance characteristics*. For drain voltages greater

than V_P, the equation of the transconductance curve is

$$i_D = I_{DSS}\left(1 - \frac{v_{GS}}{V_P}\right)^2 \tag{5.1}$$

Inserting $v_{GS} = 0$ in the above equation gives the vertical intercept of the curve as $i_D = I_{DSS}$ (see figure 5.3b).

A parameter known as the *mutual conductance*, g_m, of the FET at any point on the characteristic is given by the slope at that point, hence

$$\text{mutual conductance} = g_m = \frac{\partial i_D}{\partial v_{GS}} = -\frac{2I_{DSS}}{V_P}\left(1 - \frac{v_{GS}}{V_P}\right)$$

In particular, the mutual conductance at the point where the transfer characteristic meets the y-axis (when $v_{GS} = 0$) is

$$g_{m(0)} = -\frac{2I_{DSS}}{V_P} \tag{5.2}$$

Hence

$$g_m = g_{m(0)}\left(1 - \frac{v_{GS}}{V_P}\right) \tag{5.3}$$

Typical values of $g_{m(0)}$ lie in the range 0.05 to 10 mS. The mutual conductance is also known as the *transconductance* (g_{fs}) of the FET.

Rearranging equation 5.2 allows the pinch-off voltage to be predicted from the transfer characteristics as follows

$$-\frac{V_P}{2} = \frac{I_{DSS}}{g_{m(0)}}$$

The above expression indicates that a straight line of slope $g_{m(0)}$, which cuts the y-axis at I_{DSS}, also cuts the x-axis at $v_{GS} = -V_P/2$ (see figure 5.3b). From the foregoing, a FET having a high value of I_{DSS} usually has a high value of V_p and vice versa.

Manufacturers of FETs usually measure V_P by a method based on equation 5.1, as follows. At a value of v_{DS} which is greater than V_P, the gate bias is increased until $i_D \approx 0$. Since $I_{DSS} \neq 0$ then, from equation 5.1, $v_{GS}/V_P = 1$. That is, $v_{GS} = V_P$ when the value of the drain current is just equal to zero; the value of V_P is usually specified at a drain current of about 0.5 to 1 nA.

5.4 Common-source Small-signal Equivalent Circuit of the FET

The common-source small-signal equivalent circuit for the FET at low frequencies is shown in figure 5.5a. In this equivalent circuit it is assumed that the signal source provides a sinusoidal waveform, and that the voltages and currents shown are r.m.s. values of the respective quantities. The input circuit has a high value of input impedance, and can be regarded as a current source comprising resistor r_g in parallel

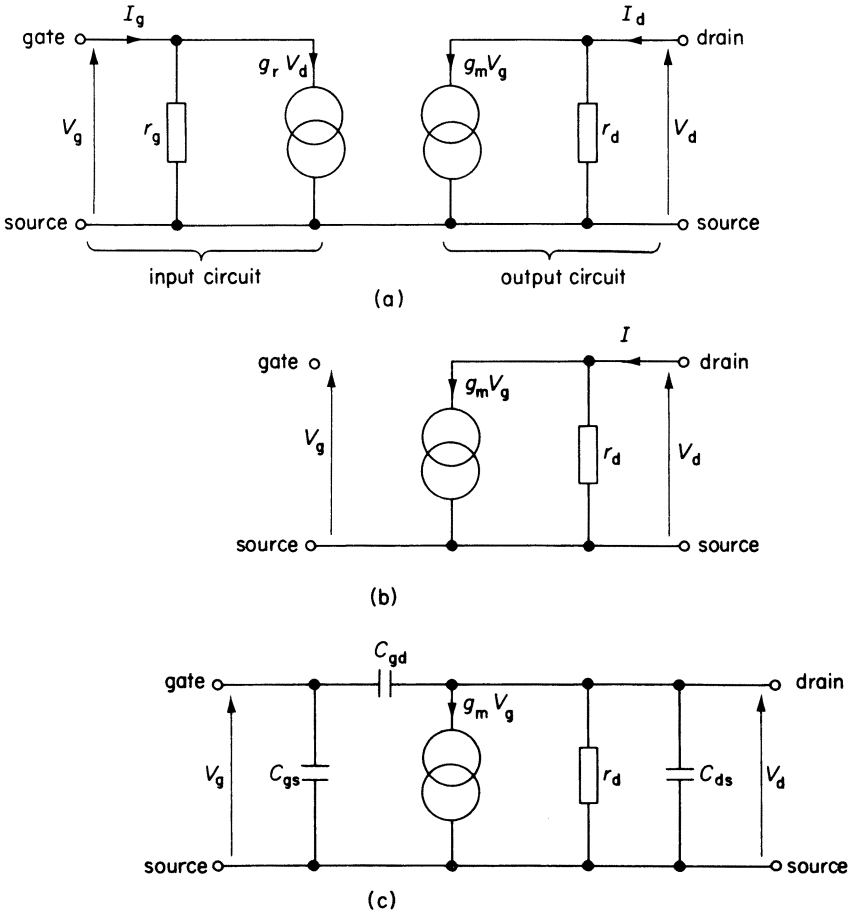


Figure 5.5 Small-signal equivalent circuits for the FET: (a) low-frequency equivalent circuit, (b) simplified low-frequency equivalent circuit and (c) a simplified high-frequency equivalent circuit

with an 'ideal' current source $g_r V_d$, where g_r is the reverse conductance parameter of the FET, and V_d is the r.m.s. value of the drain voltage.

The output circuit is represented by another current-generator circuit comprising an 'ideal' current source $g_m V_g$ in parallel with resistance r_d. The value of r_d is equal to the value of the reciprocal of the slope of the output characteristic at the operating point. The equations satisfying figure 5.5a are

$$I_g = \frac{V_g}{r_g} + g_r V_d \tag{5.4}$$

$$I_d = g_m V_d + \frac{V_d}{r_d} \tag{5.5}$$

When in normal operation, and at low frequency, the gate-to-channel junction is reverse biased, and $I_g \approx 0$, that is, $r_g = \infty$ and $g_r = 0$. The low-frequency small-signal equivalent circuit may therefore be simplified to that in figure 5.5b, which is represented in mathematical form by equation 5.5.

Since the gate junction is reverse biased, depletion-layer capacitances C_{gs} and C_{gd} exist between the gate and the source and between the gate and the drain, respectively (see figure 5.5c). At frequencies beyond the audio-frequency range these capacitances must be taken into account, and are shown in the high-frequency equivalent circuit, figure 5.5c. Capacitance C_{ds} in the figure represents the drain-to-source capacitance of the channel; this capacitance is largely the header capacitance of the FET. Since C_{gs} and C_{gd} are due to the depletion layer in the channel, their values vary inversely with the square root of the reverse bias voltage at the gate. The capacitive reactances of these capacitors not only shunt both the input and output circuits, but also introduce feedback between the input and output circuits. The net result at high frequencies is a reduction in voltage gain as the signal frequency increases. Typical values of the parameters in the equivalent circuit are listed below.

Parameter	*Typical value*
g_m	0.1–10 mS
r_d	0.1–1 MΩ
r_g	> 100 MΩ
C_{gs}	1–10 pF
C_{gd}	1–10 pF
C_{ds}	0.1–1 pF

5.5 Thermal Effects on JUGFET Characteristics and Parameters

A change in ambient temperature from, say, 25 °C to 100 °C causes the transfer characteristic of an *n*-channel JUGFET to alter in the manner shown in figure 5.6. The reason for the change in shape is as follows.

An increase in temperature results in a reduction of the mobility of the charge carriers in the channel, leading to an increase in r_d and a reduction in i_D. This factor has its primary effect in the region between points X and Y on the characteristics. This change effectively reduces the value of g_m in this region of the curve. Since, in this region, the value of the drain current reduces with increasing temperature, the phenomenon of thermal runaway (see section 4.6) associated with some BJT amplifiers is not normally encountered in FET circuits.

A second effect of an increase in temperature is that the barrier potential of the reverse biased *p–n* gate junction has a negative temperature coefficient of value about –2.2 mV/°C. The latter effect results in an increase in i_D with increasing temperature. This has its major effect on the region of the characteristic to the left of point X in figure 5.6.

At point X the two effects mentioned above cancel each other out, and the drain current has a constant value over a wide temperature range. Point X on the

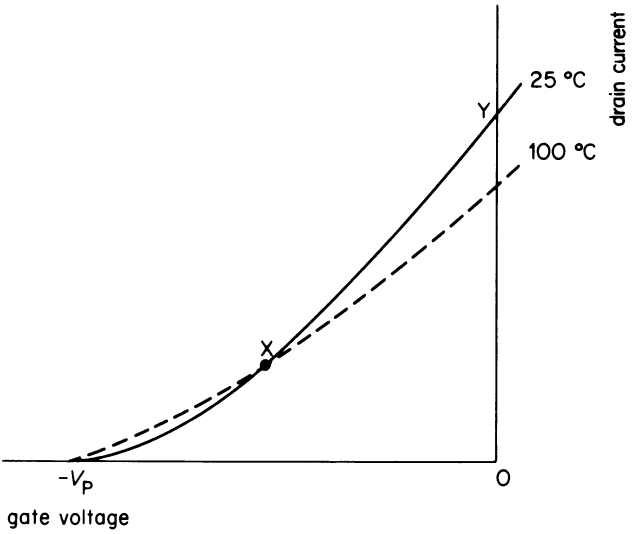


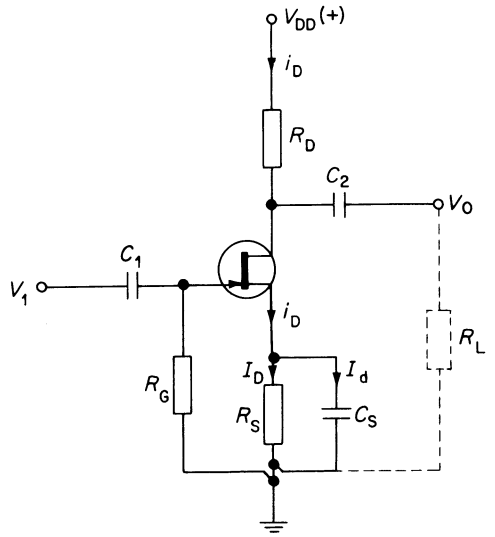
Figure 5.6 The effect of temperature change on the transfer characteristics of an *n*-channel JUGFET

characteristic is the operating point aimed at in FET amplifiers which are to have low-drift performance.

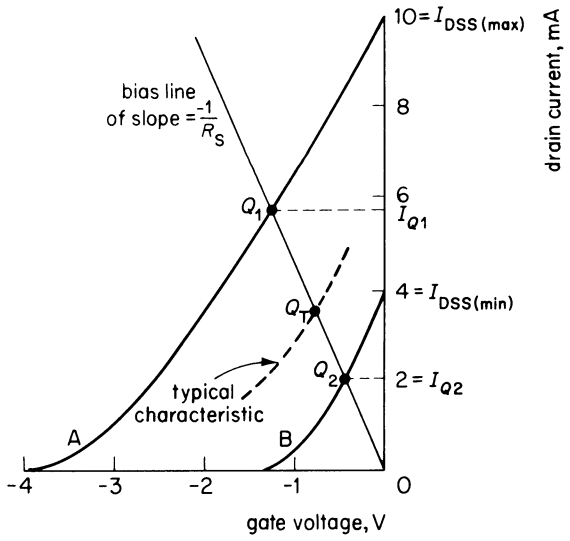
If a FET amplifier is biased to a point between X and Y on the transfer characteristics, then it offers low distortion since the mutual characteristic is fairly linear. If biased to point X, it offers low drift. If biased to the left of point X, it offers the highest voltage gain. The reason for the latter feature is not apparent at first sight, but can be understood from the following. If the reverse bias voltage applied to the gate is increased to the point where the quiescent drain current is reduced by a factor of, say, ten times when compared to the 'normal' condition, then a load resistor of ten times the previous value can be employed in order to give the 'normal' value of quiescent drain voltage. The voltage gain, K_v, of the FET is given approximately by the expression $K_v = -g_m R_L$ (see section 5.8), where R_L is the load resistance; if g_m is reduced by, say, one-third by the increased bias voltage (see also equation 5.3) then, since R_L has been increased by a factor of 10, the over-all voltage gain when operating at low values of i_D can be larger than that obtained at other operating points. However, since the characteristic is curved in this region, the input signal should have a small magnitude if distortion is to be avoided.

5.6 A Simple Common-source Amplifier

A basic type of common-source amplifier using an *n*-channel JUGFET is illustrated in figure 5.7a. The gate bias voltage, V_G, is equal to the mean value of the voltage



(a)



(b)

Figure 5.7 (a) A basic common-source amplifier using self-bias and (b) determination of the working point

developed across R_S, that is, $V_G = I_D R_S$, where I_D is the quiescent value of the drain current. Since the lower end of R_S is earthed, and the potential of its upper end is equal to $I_D R_S$ then, since no current flows in R_G, the mean potential of the gate electrode relative to the source is $-I_D R_S$. In practice some leakage current flows out of the gate electrode; this current causes a p.d. to develop across R_G and, to prevent this p.d. becoming too large and affecting the bias voltage, the value of R_G is usually limited to a value not greater than about 1 MΩ.

Each family of FETs has a 'maximum' and a 'minimum' transfer characteristic, typified by curves A and B in figure 5.7b. When used in the amplifier in diagram (a) of the figure, the FET would operate at point Q_1 on curve A, this point being determined by the intersection of the *bias line* of slope $(-1/R_S)$ with the 'maximum' transfer characteristic. Operating point Q_2 on the 'minimum' transfer characteristic is determined in a similar manner. Thus, with a device having the range of characteristics shown and with a value of R_S of 220 Ω, the quiescent drain current can have a value in the range between I_{Q1} and I_{Q2}, which is about 5.7 and 2.0 mA, respectively. Manufacturers also supply a statistically 'typical' transfer characteristic which, in the case considered, gives an operating point Q_T.

In the case of a 'one off' design, the correct value of quiescent current can be obtained by inserting a suitable value of R_S into the circuit. This arrangement is unsatisfactory in the case of mass-produced circuits, and a bias circuit must be used, which limits the range of possible values of quiescent current. Such a circuit is described in section 5.7.

A figure of merit of the circuit discussed here is its *thermal stability factor*, S_F, which is defined as

$$S_F = \frac{\text{change in } I_Q \text{ with temperature with } R_S \text{ in circuit}}{\text{change in } I_Q \text{ with temperature when } R_S = 0}$$

An analysis of the circuit in figure 5.7a yields

$$S_F = \frac{1}{1 + g_m R_S} \tag{5.6}$$

Thus a circuit with a high value of $g_m R_S$ has a better thermal stability than one with a lower value for this product. The circuit design is often a compromise between several factors including good thermal stability, low power consumption and good linearity.

Capacitor C_S is a bypass capacitor, and the procedure for selecting a suitable value is generally similar to that outlined in the case of BJT amplifiers. A suitable value is given by the expression $C_S \geqslant 10/2\pi f_{min} R_S = 5/\pi f_{min} R_S$, where f_{min} is the minimum operating frequency of the amplifier. Assuming that the value of f_{min} is 32 Hz, then $C_S \geqslant 0.05/R_S$ F (R_S in Ω) or $C_S \geqslant 50/R_S$ μF (R_S in kΩ).

Capacitors C_1 and C_2 in figure 5.7a are blocking capacitors and, using the usual

methods of approximation

$$C_1 \geqslant \frac{50}{R_G}\,\mu\text{F}(R_G \text{ in k}\Omega)$$

$$C_2 \geqslant \frac{50}{R_G}\,\mu\text{F}(R_L \text{ in k}\Omega)$$

As will be shown in section 5.8, the voltage gain, K_v, of the amplifier is

$$K_v = -g_m R'_L$$

where R'_L is the a.c. load resistance, which comprises R_D in parallel with R_L. Since a typical value of g_m is 1 to 4 mS, the value of R'_L needs to be fairly high in order to obtain a reasonable value of voltage gain. *Note*: The value of g_m at point Q_T in figure 5.7b is about 3 mS.

5.7 A Practical Form of JUGFET Amplifier

The circuit in figure 5.8a is one of the most popular of many possible circuits used to stabilise the working point of the FET against variations in device characteristics. It differs from figure 5.7a in that the gate is connected to potential V_{GG}, which is obtained from the potential divider chain R_1R_2. The value of V_{GG} is

$$V_{GG} = \frac{R_2 V_{DD}}{R_1 + R_2}$$

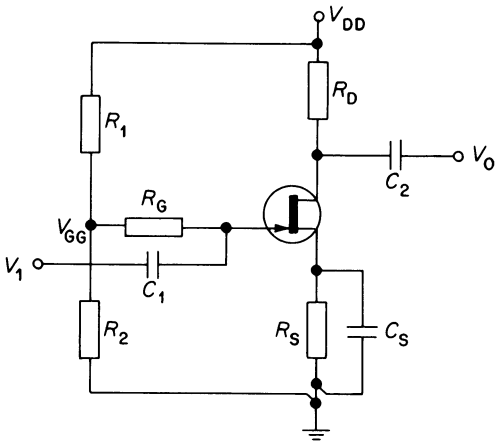
The bias line terminates at one end at $+V_{GG}$ (see figure 5.8b), and has a slope of $-1/R_S$. The slope of the bias line in figure 5.8b is less steep than that in figure 5.7b, which implies that the value of R_S has a larger value in the circuit in figure 5.8 than it has in figure 5.7.

An examination of equation 5.6 indicates that increasing the value of R_S increases the thermal stability of the circuit, so that the thermal stability of the circuit in figure 5.8a is better than that in figure 5.7a. However, a larger value of R_S means that a higher value of supply voltage may be required.

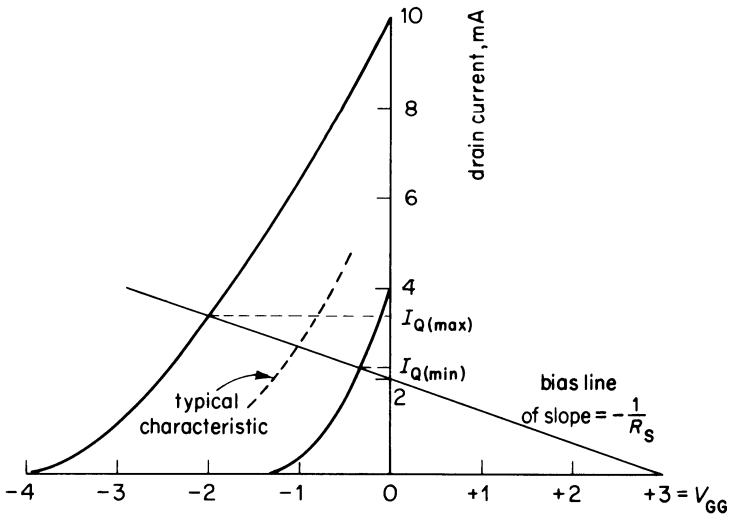
If V_{GG} = +3 V and R_S = 1.5 kΩ (see figure 5.8b), the intercept of the bias line with the minimum and maximum transfer characteristics gives $I_{Q(min)}$ = 2.3 mA and $I_{Q(max)}$ = 3.4 mA. Comparing these with the values 2.0 and 5.7 mA, respectively, for figure 5.7b, shows that a considerable improvement in stabilising I_Q has been obtained by using the modified circuit. If V_{DD} = 15 V, the bias voltage could be obtained using R_1 = 10 kΩ and R_2 = 39 kΩ (or any other suitable values in the same ratio).

5.8 Analysis of FET Amplifiers

The analysis of a FET amplifier is carried out using the block diagram in figure 5.9. In this circuit, R_S is the internal resistance of the signal source, and R'_L is the



(a)



(b)

Figure 5.8 (a) A practical form of JUGFET linear amplifier and (b) the bias line construction

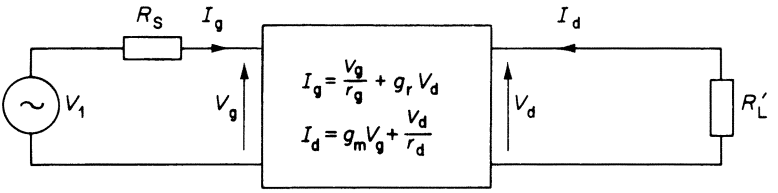


Figure 5.9 Small-signal analysis of a FET amplifier

effective value (the a.c. value) of the load resistance. The equations of the circuit are

$$I_g = \frac{V_g}{r_g} + g_r V_d$$

$$I_d = g_m V_g + \frac{V_d}{r_d}$$

$$V_d = -I_d R'_L$$

Solving the above equations for the voltage gain ($K_v = V_d/V_g$), the current gain ($K_i = I_d/I_g$), the input resistance ($R_{in} = V_g/I_g$) and the output resistance (R_{out}) yields the relationships shown in table 5.1.

The approximate relationships given in the table are deduced from an analysis that assumes that $r_g = \infty$ and $g_r = 0$; these relationships are adequate for many audio-frequency applications.

Table 5.1

Parameter	Exact relationship	Approximate relationship
K_v	$-g_m \Big/ \left(\frac{1}{r_d} + \frac{1}{R'_L}\right)$	$-g_m R'_L$
K_i	$\frac{g_m + K_v/r_d}{K_v g_r + 1/r_g}$	infinity
R_{in}	$1 \Big/ \left(K_v g_r + \frac{1}{r_g}\right)$	infinity
R_{out}	$1 \Big/ \left(\frac{1}{r_d} - \frac{g_m g_r}{(1/R_S + 1/r_g)}\right)$	r_d

Relationships between the h-parameters and the FET parameters

$$h_i = r_g \qquad r_g = h_i$$

$$h_r = -g_r r_g \qquad g_r = -\frac{h_r}{h_i}$$

$$h_f = g_m r_g \qquad g_m = \frac{h_f}{h_i}$$

$$h_o = \frac{1}{r_d} - g_r g_m r_g \qquad r_d = \frac{1}{h_o - h_r h_f / h_i}$$

5.9 Insulated-gate Field-effect Transistors (IGFETs)

In the IGFET the gate is insulated from the semiconductor by an oxide layer of thickness about 0.15 μm. A section through a *p*-channel IGFET is shown in figure 5.10. This device is fabricated in a small volume or chip of *n*-type silicon, into which two heavily doped p^+-regions are diffused; the diffused regions act as the source and the drain of the FET. The spacing between the source and drain electrodes is typically 10 to 20 μm. Connections are made to these regions through holes or 'windows' in the oxide layer which covers the surface of the FET. The gate electrode in figure 5.10 is simply a thin layer of aluminium, which is deposited on the surface of the oxide. By virtue of their construction, IGFETs are frequently referred to as *metal-oxide-semiconductors*, (MOS) FETs or simply as MOSTs.

For reasons given below, the MOSFET in figure 5.10a is a *p*-channel device, and it has its source electrode connected to the positive pole of the supply, and its drain electrode to the negative pole. Since the *p*-type source and drain regions are separated by an *n*-type semiconductor then, assuming the gate voltage to be zero, the *p*–*n* junction at the drain is reverse biased and the drain current is zero. This fact is implied in the circuit symbol for the *p*-channel MOSFET (figure 5.10b) by the broken link between the drain and source. The separation between the gate and the semiconductor is also illustrated on the circuit symbol by the physical separation between the gate electrode and the 'broken' channel. The type of device, that is, *p*-channel, is symbolised by the direction of the arrow on the substrate link; when the arrow points away from the drain-to-source channel it is a *p*-channel device, and when pointing towards it the device is described as an *n*-channel device (see figure 5.10c).

At room temperature, electron–hole pairs are continuously generated in the *n*-type substrate material. This material is lightly doped (in fact it is nearly intrinsic), and the application of a negative potential to the gate electrode causes some of the holes generated in this way to be attracted to the underside of the oxide-semiconductor interface. As the negative potential applied to the gate electrode is progressively increased in value, more and more of the minority charge

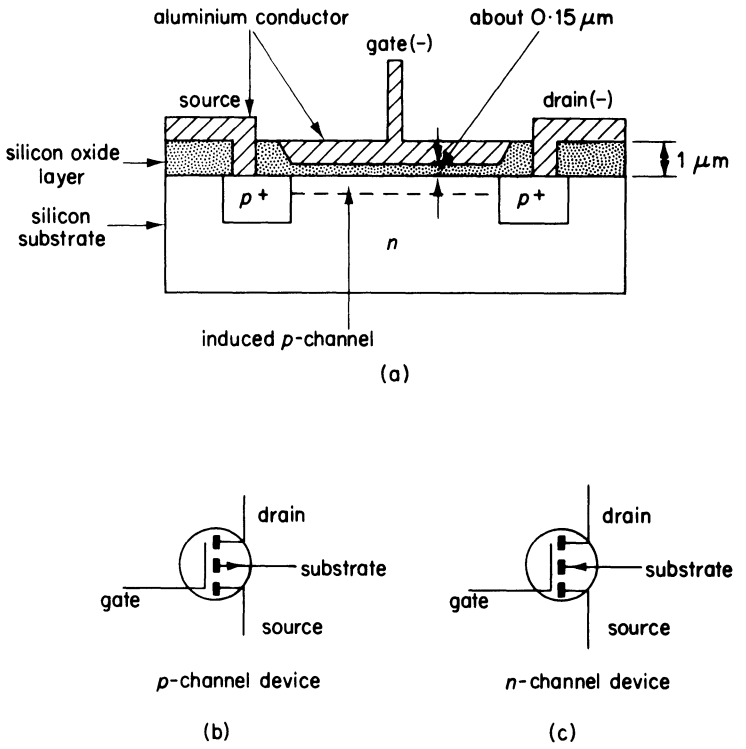


Figure 5.10 (a) A section through a *p*-channel MOS transistor and (b) its circuit symbol; (c) the symbol for an *n*-channel MOST

carriers (holes) in the *n*-region are attracted to the interface until, at a voltage known as the *threshold voltage*, V_T, a conducting channel or *inversion channel* of *p*-type material is formed between the source and the drain (see figure 5.10a). The value of V_T is typically about -4 V in standard types of MOSFETs, but is reduced to about -2 V by using special construction techniques.

The application of a higher gate voltage than V_T attracts more holes to the conducting channel, so increasing its conductivity. Consequently increasing the value of the gate voltage increases or enhances the value of the drain current. This type of FET is therefore described as an *enhancement-mode* device. Moreover, since the conducting channel is an inversion layer of *p-type semiconductor, the device in* figure 5.10a is described as a *p*-channel MOSFET.

A typical set of common-source output characteristics for a *p*-channel MOSFET is shown in figure 5.11a. The transfer characteristic, figure 5.11b, is deduced from the output characteristic in the manner outlined in section 5.3. This curve cuts the $i_D = 0$ axis at V_T.

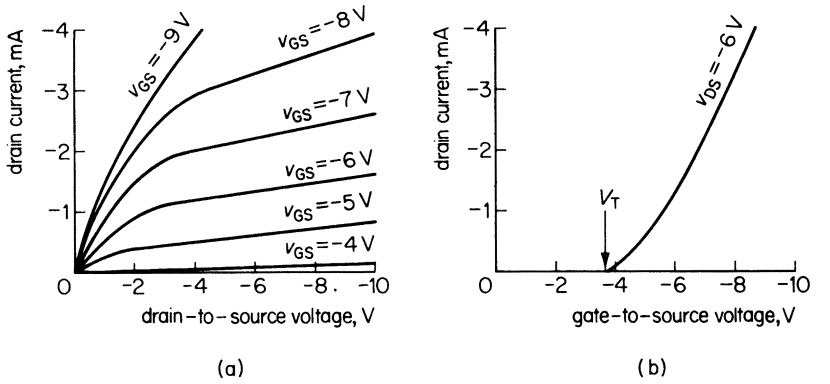


Figure 5.11 (a) Typical set of output characteristics for a *p*-channel enhancement mode MOSFET and (b) a transfer characteristic for the device

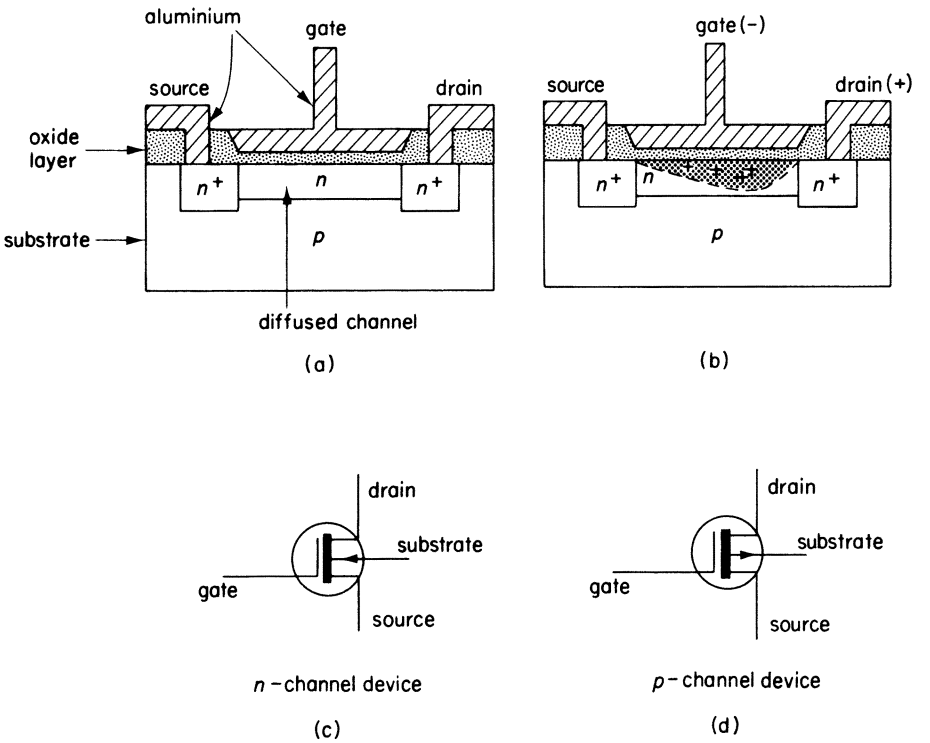


Figure 5.12 (a) The basic structure of a depletion-mode *n*-channel MOSFET and (b) channel depletion with a negative gate potential; (c) and (d) circuit symbols for *n*-channel and *p*-channel MOSFETS, respectively

5.10 Depletion-mode MOSFETs

A second type of MOSFET is manufactured that has a similar basic structure to that in figure 5.10a, but has a conducting channel (known as the *initial channel*) diffused between the source and the drain during manufacture.

Consider the operation of an *n*-channel depletion-mode MOSFET. A section through a typical device is shown in figure 5.12a, and the reader will note that the source and the drain are electrically connected by the diffused initial *n*-channel. Consequently when $v_{GS} = 0$ current can flow between the drain and the source. This fact is accounted for in the circuit symbol, figure 5.12c, by the continuous line that links the drain and the source. A set of output characteristics for the depletion-mode FET is shown in figure 5.13a. In the case of an *n*-channel depletion-mode MOSFET, the application of a positive potential to the gate causes the minority charge carriers (electrons) in the *p*-region to be attracted to the underside of the gate electrode; these increase the channel conductivity and result in an increased value of drain current. In this mode the drain current is enhanced. Applying a negative potential to the gate electrode induces positive charge carriers to appear in the *n*-channel, reducing its conductivity. Thus, a negative gate potential depletes the value of the drain current; this type of device is known as a *depletion-mode* MOSFET, since it can operate as a depletion-mode device. A mutual characteristic of the device is illustrated in figure 5.13b.

5.11 Equivalent Circuits of the MOSFET

For all practical purposes the equivalent circuits in figure 5.5, developed for the JUGFET, can be used for MOS devices. Typical values of the parameters are

g_m 0.1–25 mS
r_d 1–50 kΩ
r_g $> 10^{10}$ Ω
C_{ds} 0.1–1 pF
C_{gs} 1–10 pF
C_{gd} 1–10 pF

5.12 The FET as a Voltage-dependent Resistor (VDR)

Both IGFETs and JUGFETs exhibit an 'ohmic' region at low values of drain voltage, the value of resistance being controlled by the gate bias voltage, v_{GS}. When JUGFETs are operated in this mode they are known as *pinch-effect resistors*.

5.13 Comparison of *p*-channel and *n*-channel MOSFETs

It is found that the *p*-channel MOSFETs have more reliable characteristics than do *n*-channel devices, but this is a picture that is changing continuously. In the

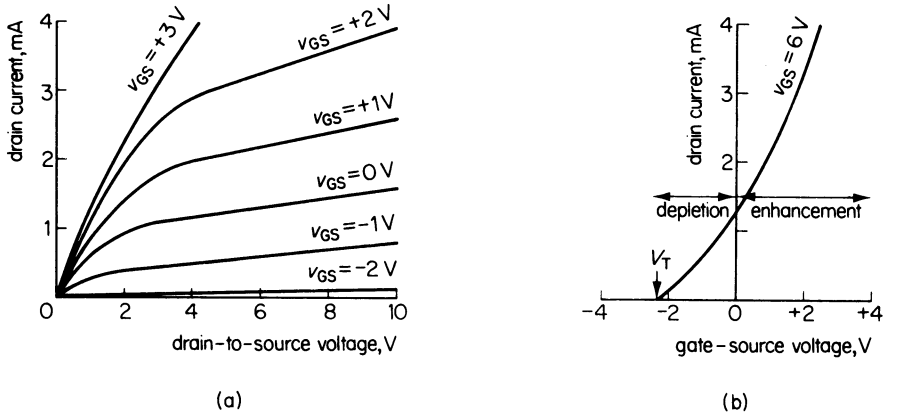


Figure 5.13 (a) Output characteristics of an *n*-channel depletion-mode MOSFET and (b) a transfer characteristic

production of MOSFETs, the majority of the contaminants in the oxide layer are positively charged silicon ions. Since the gate voltage applied to an *n*-channel enhancement MOSFET has a positive polarity, it drives these mobile positive ions into the interface between the oxide and the silicon. In turn these charges attract electrons to the interface, causing the inversion channel to form prematurely. The net result is that the threshold voltage in an *n*-channel device may vary erratically. In *p*-channel enhancement devices, which use a negative gate voltage, the positive ions are attracted to the gate electrode and do not affect the performance of the FET.

However, *n*-channel devices are not without their advantages. Since the mobility of electrons in the material silicon is about twice that of holes, the physical size of an *n*-channel MOSFET for a given ON resistance when operating as an electronic switch, is about one-half that of an equivalent *p*-channel device. This means that *n*-channel devices are very attractive as switching elements. Moreover, their small size reduces their self-capacitance when compared with *p*-channel devices, so that their switching speed is higher.

In certain types of logic device, both *p*- and *n*-channel FETs are employed in the same element (see section 5.15).

5.14 MOS Linear Amplifier

MOSFETs exhibit higher noise figures than JUGFETs, and are rarely used in discrete form in linear amplifiers. Moreover, since most MOSFETs are *p*-channel devices, their upper frequency limit is lower than that of *n*-channel FETs owing to the lower mobility of the charge carriers.

There are a few exceptions to the above, such as their use in electrometer amplifiers where their high input resistance is of value. A circuit which is used in

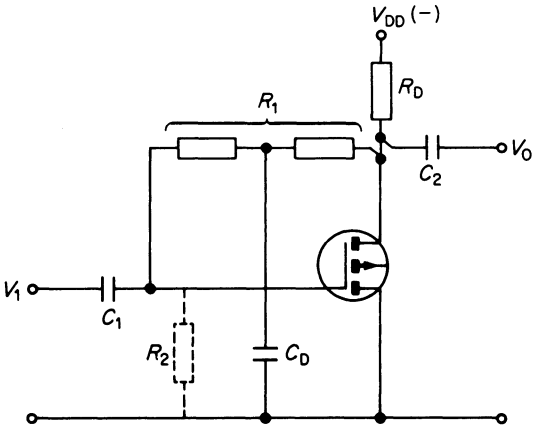


Figure 5.14 A MOSFET linear amplifier

some applications is shown in figure 5.14. In this circuit, capacitors C_1 and C_2 are blocking capacitors, and C_D is a decoupling capacitor that prevents the feedback of a.c. signals from the drain to the gate. The d.c. bias voltage is obtained via R_1, and since the gate does not draw current, the quiescent values of the gate and drain voltages are equal to one another, that is, $V_{GS} = V_D$. The operating point on the output characteristics is determined by the intersection of the load line and a curve corresponding to $V_{GS} = V_D$.

If it is necessary to operate with a bias voltage that is less than V_D, then resistor R_2 (shown dotted in figure 5.14) can be connected between the gate and the common line. The gate voltage is then $V_{GS} = R_2 V_D/(R_1 + R_2)$.

5.15 MOS Logic Circuits

MOS devices are widely used in digital electronic circuits. Owing to the high values of the time constants associated with the input circuits of MOS elements, they are somewhat slower in operation than are bipolar devices. Their principal assets are that their power consumption is lower than that of bipolar circuits, and that they can be fabricated at a higher density in integrated circuit form.

NOT gates

A *p*-channel MOS NOT gate is shown in figure 5.15. In this circuit, TR2 acts as a resistor whose characteristic is not quite linear. Since the gate of TR2 is connected to its drain electrode, TR2 is always in a conducting state and, for the purpose of the circuit shown, it may simply be regarded as a resistor.

When $A = 0$, TR1 is cut off and no current flows through it; as a result the output potential is 'high', and is regarded as a logic '1' signal. Typically, with

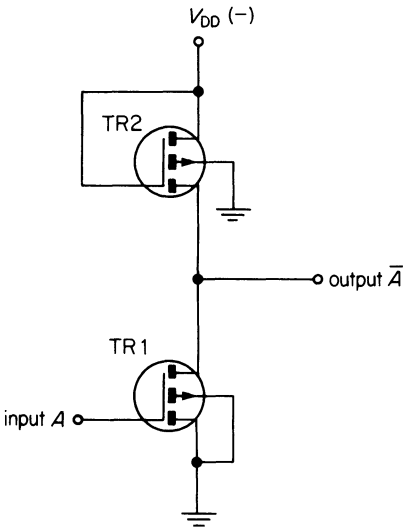


Figure 5.15 A *p*-MOS NOT gate

$V_{DD} = -20$ V, the logic '1' output corresponds to a voltage in the range −11 to −14 V.

When $A = 1$, TR1 conducts and the p.d. across it falls to a 'low' value, that is, the output is logic '0'. A typical logic '0' output voltage lies in the range −2 to −3 V. The above is known as *negative logic notation*, since the greater negative potential represents logic '1'.

By using *complementary MOS* (CMOS) FETs in the manner shown in figure 5.16a, the mean power consumption can be reduced to about 10 nW per gate (1 nW = 10^{-9} W), and the delay time involved in propagating a signal through the gate is roughly halved when compared with the *p*-MOS circuit in figure 5.15. The former is brought about by the fact that both TR1 and TR2 are switched by the input signal, A, and when TR1 is ON then TR2 is OFF and vice versa. Since one of the FETs is OFF at all times, no current (other than leakage current) flows through them. The reduction in propagation delay mentioned above, results from the use of an *n*-channel device, which has improved switching speed when compared with the *p*-channel device it replaces.

A feature of CMOS logic is that it tolerates a wide range of supply voltage, and may be supplied by a voltage in the range 3 to 15 V, and is therefore compatible with bipolar logic families. The 'high' output voltage from the gate is almost equal to the supply voltage used, and the 'low' output voltage is practically zero (typically 10^{-2} V).

CMOS NOR and NAND gates

NOR and NAND gates are logical extensions of the CMOS NOT gate described above. Typical two-input CMOS NOR and NAND gates are illustrated in diagrams

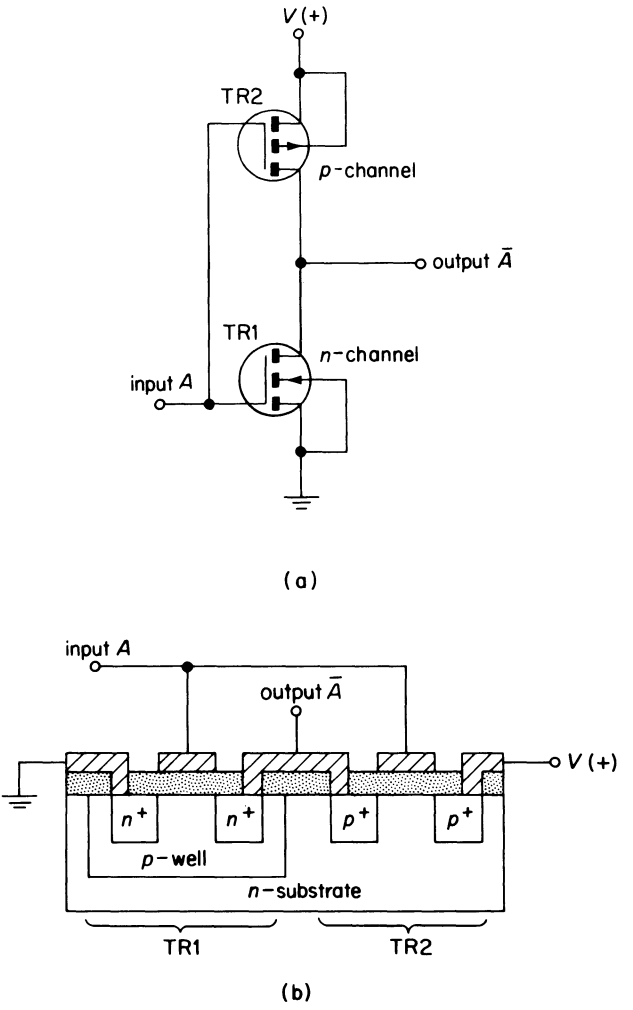
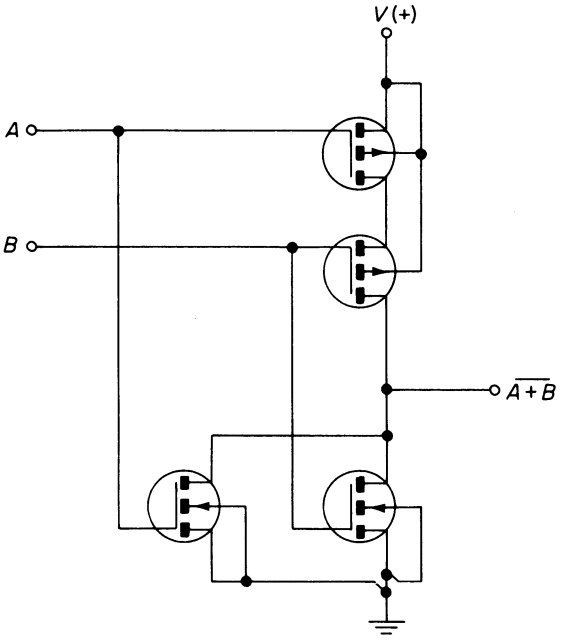


Figure 5.16 (a) A CMOS NOT gate and (b) a section through the gate

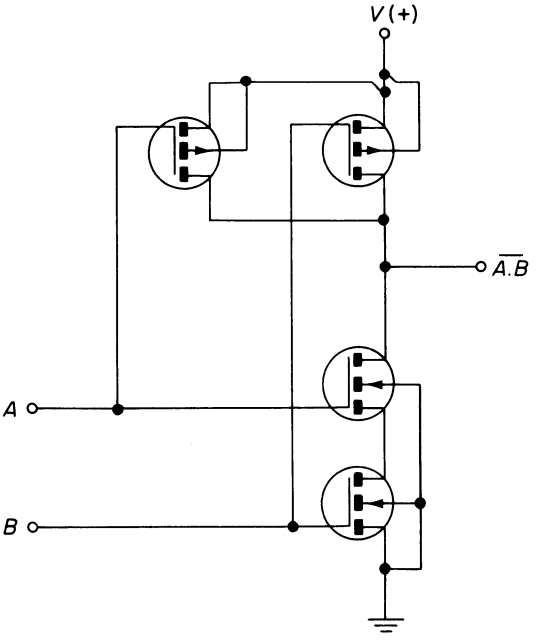
(a) and (b), respectively, of figure 5.17. In common with other CMOS gates, the circuits shown operate over a wide supply voltage range, and consume very little power.

5.16 MOS Gate Insulation Protection

Since the oxide layer between the gate and the semiconductor in MOS devices is very thin, it is easily punctured by relatively low values of transient voltage. For example, it is not uncommon for humans to become charged to voltages in excess of 20 kV; this potential may be generated by friction between the skin and many types of material such as clothing, plastics, etc. To prevent MOSFETs being



(a) CMOS NOR gate



(b) CMOS NAND gate

Figure 5.17 CMOS gates

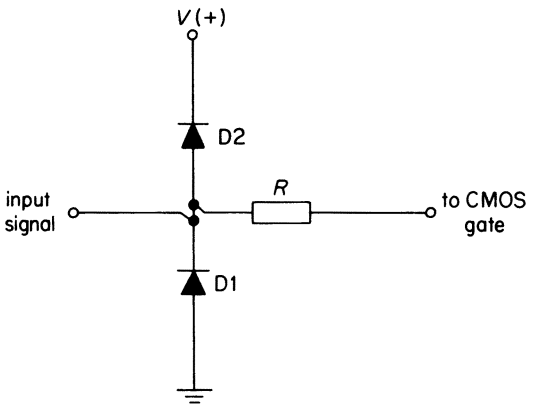


Figure 5.18 CMOS gate input protection

accidentally damaged from this cause during handling and shipping, they are normally packed in a conductive plastic material. The conductive packing is not normally removed until the devices are mounted in the circuit.

Most MOS logic circuits have in-built protection against voltage transients; one form of circuit used in the input lines of CMOS gates is shown in figure 5.18. In this circuit, diodes D1 and D2 protect the gate against negative and positive transients, respectively. The value of resistor R is typically in the range 200 to 2000 Ω.

5.17 Silicon-on-sapphire (SOS) Structure

All the MOS devices so far described use silicon as the substrate material. The operation of the transistor itself takes place in the upper 1 μm of the substrate, the remaining part having no significant function. In fact the remaining silicon

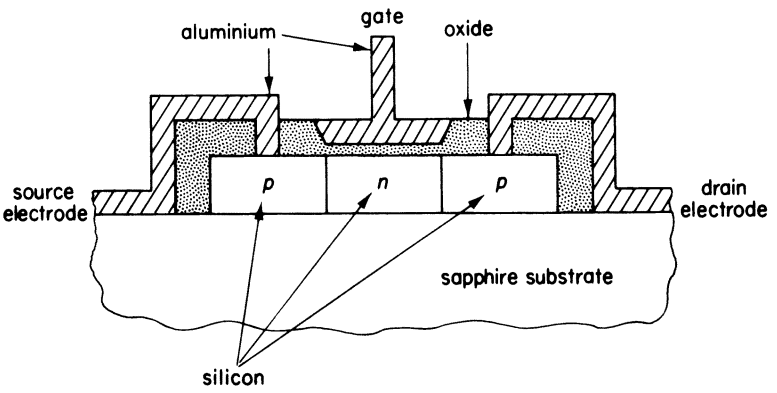


Figure 5.19 SOS MOS structure

handicaps the operation of the device, since it contributes significantly to the leakage current and to the device capacitance.

One method of overcoming the disadvantages of the silicon substrate is to replace it with an insulating material whose crystal lattice is sufficiently similar to the silicon. Sapphire is such a material. The basis of the SOS MOS is shown in figure 5.19. The process offers MOS-circuit packing density and bipolar speeds.

5.18 The Silicon-gate Structure

In this structure, a heavily doped polycrystalline silicon is used as the gate electrode (see figure 5.20). In its manufacture, the gate oxide and the silicon gate are formed before the source and drain diffusions are introduced. The gate acts as a diffusion mask during manufacture, and gives accurate alignment between the gate and the source and drain electrodes. The net result is a small structure, having low values of interelectrode capacitance; the switching speed of such a device is higher than that of a conventional FET.

Moreover, with a p^+ silicon gate, the positive ions in the silicon are equivalent to an electrical voltage on the gate. This has the effect of reducing the threshold voltage by about 1 V when compared with a conventional MOSFET.

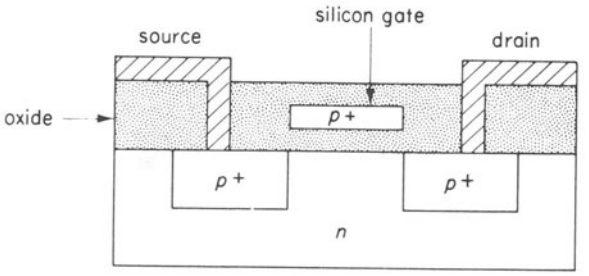


Figure 5.20 The silicon-gate structure

5.19 Fetrons

FETRONs are high-voltage junction-gate depletion-mode FETs which are used as pin-for-pin replacements for thermionic values. More than one FET may be used in each device in order to obtain the required characteristics. The advantages claimed for these devices are that they are more reliable than their valve equivalents, and that there is a considerable reduction in power consumption. They are not suitable for replacement for all valve types, but can be used to replace popular pentode types – such as the EF95 and the EF91 – and triodes, including ECC81-83.

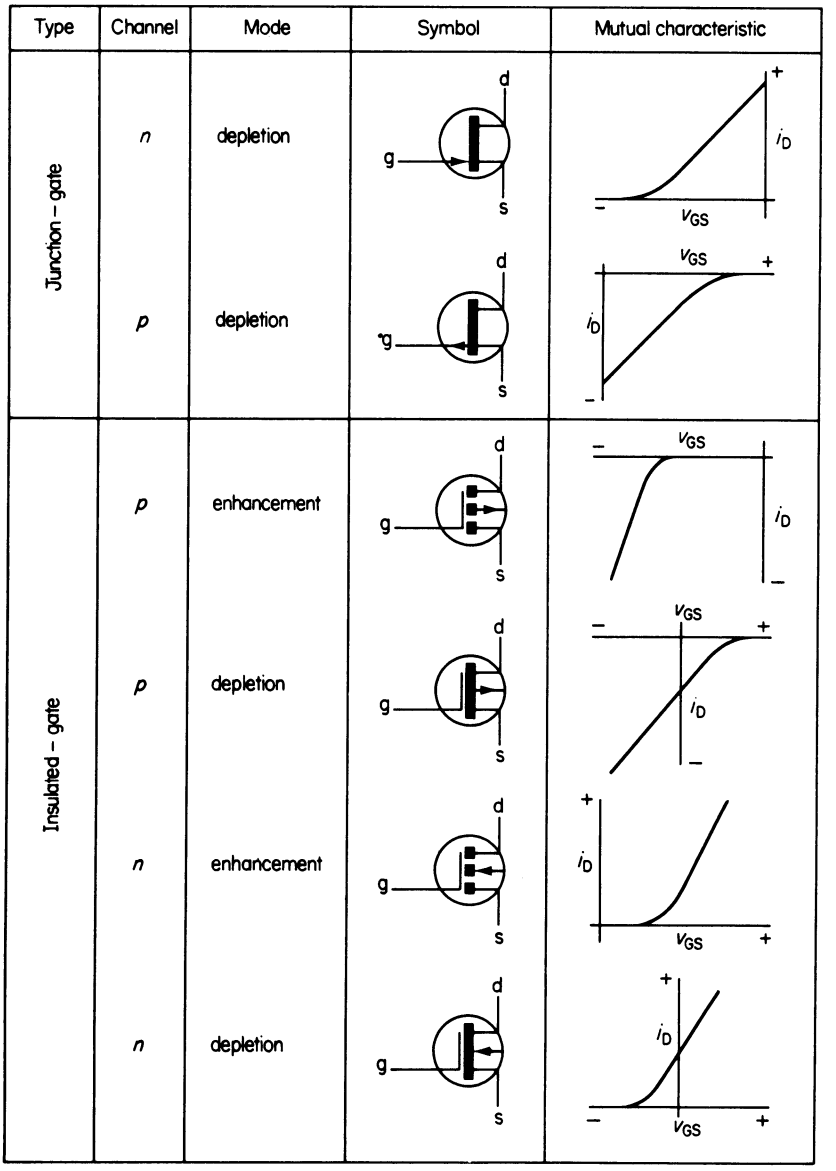


Figure 5.21 Summary of FET symbols and transfer characteristics

5.20 Summary of FETs and FET Characteristics

The foregoing work is summarised in the following and in figure 5.21.

JUGFETs: Depletion-mode devices; *n*-channel devices are the most popular; principal applications are in linear amplifiers

MOSFETs: Basic operation in enhancement mode; most popular type is *p*-channel; principal applications are in digital electronics

In the transfer characteristics in figure 5.21, the actual polarities of the quantities involved are shown. In the case of the drain current, i_D, the 'positive' direction of flow is *into* the drain electrode.

6 Monolithic Integrated Circuits

A monolithic *integrated circuit* (IC) is a complete circuit or group of circuits manufactured in a single piece of silicon, a typical physical size being 1.25 mm square (or about fifty thousandths of an inch square). Such a circuit may contain fifty or more components such as transistors or resistors. The word monolothic is derived from the two Greek words 'monos' and 'lithos', meaning single and stone, respectively. The word 'monolithic' implies that the circuit is manufactured within a single crystal. This type of integrated circuit is sometimes described as a *planar* IC, since it takes the form of a flat surface.

Semiconductor devices must be protected from atmospheric pollution, and one method of doing this is to cover much of the surface area with an oxide layer. For this reason, silicon is invariably used for the production of ICs, since a chemically stable oxide layer, SiO_2 (glass), can be produced on the surface of the silicon. The oxide layer provides a relatively simple means of controlling the diffusion of impurities into the surface of the silicon during its production. Oxides of germanium are not stable, and cannot be used as a diffusion barrier.

6.1 Silicon Crystal Production

Zone refining

Common sand (silica) is first reduced to an impure form of silicon, the impurities are situated at the edges or grain boundaries of the crystals. In order to produce high resistivity pure silicon, these impurities must be removed.

One technique used to remove these impurities is known as *zone refining*, and is illustrated in figure 6.1. The impure rod of silicon is slowly moved through a radio-frequency heating coil, where it becomes molten. The impure substances concentrate in the molten zone, and are gradually swept to one end of the rod as it is passed through the coil. In practice several heating coils are used, the rod being allowed to 'freeze' or to solidify between zones. The bar is passed through the system a number of times.

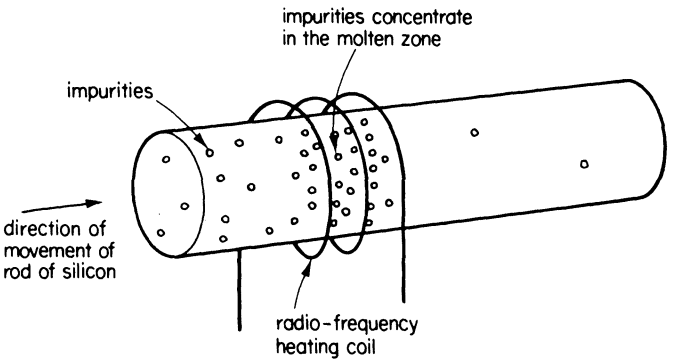


Figure 6.1 Zone-refining method

Crystal growing

The pure silicon produced by the zone-refining method is made up of random-orientated crystals. The resulting 'jagged' edges at the grain boundaries give rise to variations in resistivity throughout the rod. Silicon used in the production of ICs must not have grain boundaries, and its crystal structure must be arranged in a regular pattern to form a 'perfect' crystal.

One method of producing such a crystal is the *Czochralski* (pronounced 'chockralski') process, illustrated in figure 6.2. The pure silicon is placed in a quartz vessel and is melted by r.f. heating; the steady operating temperature is about 20 °C above the melting temperature of silicon, the oven being meanwhile filled with a hydrogen–nitrogen gas. Minute quantities of the appropriate dopants are added to

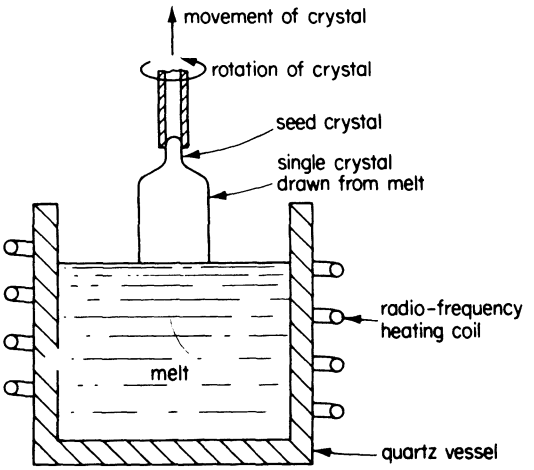


Figure 6.2 The Czochralski method of crystal growing

the melt to produce the required type of semiconductor, that is, *p*- or *n*-type. A suitable 'seed' crystal is then lowered into the melt, and is rotated and slowly withdrawn. The seed crystal does not melt at the temperature of the oven, since it has a higher resistivity than the melt itself.

Great care is taken to ensure that the correct plane of the seed is exposed in order to produce the correct crystal growth. The planes of the crystal are represented by a crystallographic notation. For example, if silicon using the (100) orientation is used in the manufacture of MOSFETs, it results in devices having threshold voltages of the order of about one-half that of devices manufactured from a crystal with the (111) orientation. The withdrawal rate of the seed is about 5 cm/h, and a typical crystal has a diameter in the range 2.5 to 7.5 cm (1 to 3 in.), and has a length of about 30 cm (12 in.).

Wafer preparation

The silicon crystal is cut into discs or wafers of about 500 μm (0.02 in.) thickness by means of a diamond saw. Any damage to the wafer resulting from the cutting process is removed by grinding and polishing, when the thickness of the wafer is finally reduced to about 200 μm (0.008 in.). The wafer, which is the *substrate* on which the IC is constructed, is now in a state to be used in the next phase of production.

6.2 Manufacture of a Simple Bipolar IC

Significant manufacturing advantages arise from the production of complete electronic circuits in IC form. Included in these are the following.

(1) The parameters and characteristics of transistors are closely matched.
(2) The manufacturing *yield* (the percentage of fault-free circuits per wafer) is very large.
(3) Very complex circuits can readily be manufactured in a very compact form.
(4) The cost is very low when compared with the equivalent discrete component circuit.
(5) They are very reliable.

Consider now the production in IC form of the circuit in figure 6.3.

The buried layer

The complete IC is constructed in what is known as an epitaxial layer; the manufacture of this layer is described later. The collector region of the transistor TR is part of the epitaxial layer, and the resistivity of this layer is fairly high (typically 0.5 Ω cm); two disadvantages of transistors with a high collector

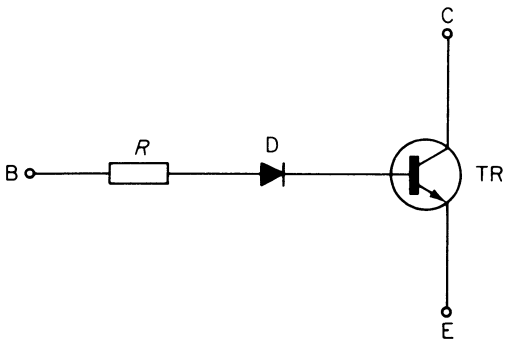


Figure 6.3 A simple transistor circuit which can be converted into a monolithic IC

resistance are

(1) a high value of saturated resistance or ON resistance
(2) a high value of $V_{CE(sat)}$.

These effects can largely be overcome by diffusing a low resistivity n^+-region into the substrate immediately below the point where the transistor is to be constructed. This region is known as a *buried layer*, since it will ultimately be buried below the epitaxial layer. The process of diffusion described below is generally similar to that used in fabricating other parts of the circuit.

Oxide growth and photomasking

The surface of the wafer is first oxidised by passing steam over it, to form an oxide layer about 1 μm thick (see figure 6.4a). The upper surface of the oxide is then coated with a uniform of photosensitive emulsion known as *photoresist*. A photographic negative is then used to *mask* the photoresist, the whole then being exposed to ultraviolet radiation (see figure 6.4a). The radiation passes through the transparent regions of the mask, causing the exposed regions of the photoresist to be polymerised. The mask is removed and the surface washed in a 'developer' which hardens the exposed regions, and dissolves the unexposed regions. The wafer is next immersed in an etching solution, which removes the oxide from the exposed areas. This leaves a *window* or aperture in the oxide film through which diffusion can take place. The remaining area of photoresist in then removed.

Diffusion

The wafer is placed in a diffusion furnace, heated to 1200 °C and subjected to gases containing *n*-type dopants (see figure 6.4b). The dopants cause the exposed area of the substrate to be converted to an n^+ semiconductor (figure 6.4c). The final depth of the buried layer is about 8 μm. The oxide layer is then removed by means of a

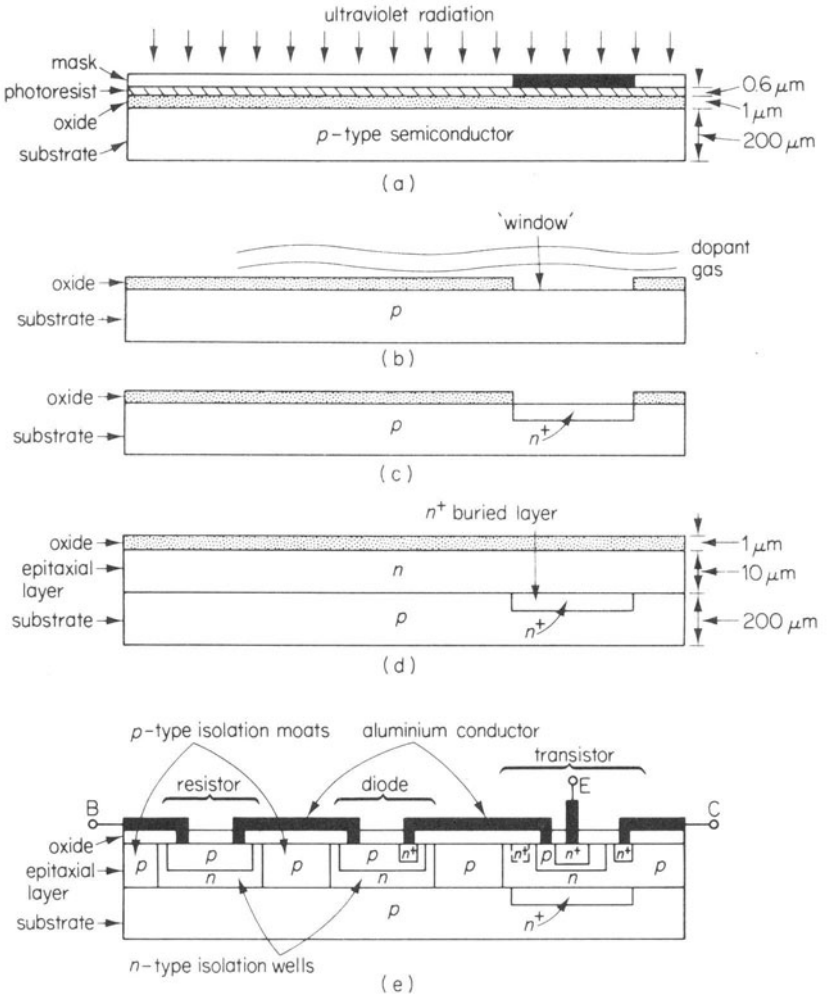


Figure 6.4 Illustrating some of the processes involved in the production of an IC

chemical solvent and an abrasive process, leaving the upper surface of the substrate and the buried layer ready for the next step.

Epitaxial layer

The wafer is once more heated in a furnace to 1200 °C and is exposed to a number of gases. The gases hydrogen and silicon tetrachloride are admitted, to yield silicon atoms which cause the *epitaxial layer* to 'grow' on the upper surface of the substrate; as it grows, it is subjected to the gases phosphene and hydrogen to give the phosphorus impurity essential for the production of an *n*-type semiconductor.

In this way an *n*-type epitaxial layer is grown on the surface of the substrate; the thickness of this layer is about 10 to 15 μm (see figure 6.4d). The rate of growth is about 1 μm/m; the process of epitaxy proceeds at a much faster rate than that of diffusion, the former being about sixty times the rate of the latter.

Circuit components

The surface of the wafer is once more covered with an oxide layer in the manner described above, and the process of photomasking, etching and diffusion is repeated many times until the circuit is complete (see figure 6.4e). A summary of the principal processes involved in the production of the components is given below.

(1) *p*-type isolation moats are diffused around the sites of the components.
(2) *p*-type diffusions are introduced to form the resistor, the diode anode and the transistor base regions.
(3) n^+-type regions are diffused for the diode cathode, the transistor emitter and for the connection to the collector of the transistor.

Aluminium metallisation

To provide an electrical connection between the 'outside' circuit and the IC, yet another set of windows is opened in the oxide layer at the points where the contacts are to be made. A thin layer (typically 1.5 μm thick) of aluminium is deposited over the whole surface of the oxide and, using a photoetching technique similar to that described above, the unwanted areas of aluminium are removed.

A feature of aluminium is that it is one of the group III materials in the periodic table of elements, and forms a *p*-type impurity with silicon. Thus direct contact may be made between aluminium and *p*-type silicon. However, unless steps are taken, it forms a *p*–*n* rectifying junction with *n*-type silicon (see also the Schottky diode in section 1.15); to prevent this happening, silicon–aluminium contacts are made with n^+-type silicon. The n^+ silicon contains a large concentration of phosphorus atoms, which ensures an ohmic contact. Thus the cathode of the diode must be of n^+-type material, as must be the connection between terminal C and the collector of the transistor.

The rectifying effect between aluminium and *n*-type silicon is used to advantage in some applications (see section 6.6).

Completing the IC

In all, about 80 to 100 individual processes are involved in the production of a bipolar IC. The wafer may contain several hundred individual circuits and, after the metallisation process is complete the wafer is scribed with a diamond-tipped tool and the individual circuits are separated. Each small section is known as a ***chip*** or ***die***. Each chip is mounted on a suitable header, and electrical connections are

made between the IC and the external connecting 'pins' by aluminium or gold wires, which may be only 0.025 mm (0.001 in.) in diameter. Finally the IC is encapsulated.

6.3 Encapsulation of ICs

The three most popular types of IC packages (often abbreviated to *packs*) are

(1) TO5 canisters (or cans), or reduced-height versions of these
(2) flatpacks
(3) plastic encapsulated dual-in-line (DIL) packs.

Outline diagrams of the three types above are shown in figures 6.5a, b and c respectively.

TO5 cans and flatpacks are hermetically sealed, and can be used over a temperature range from –55 to 126 °C, and are used wherever space and weight are at a premium. Plastic DIL packs are cheap to manufacture and are widely used in domestic, commercial and industrial applications; the operating temperature range is usually 0 to 70 °C. The most popular form of IC package is the fourteen-pin DIL pack (see figure 6.5c), with seven connecting pins on each of two opposing faces. The spacing of the pins is 2.5 mm (0.1 in.), so that they fit easily into printed circuit boards; the pins on the opposite faces are in line with one another.

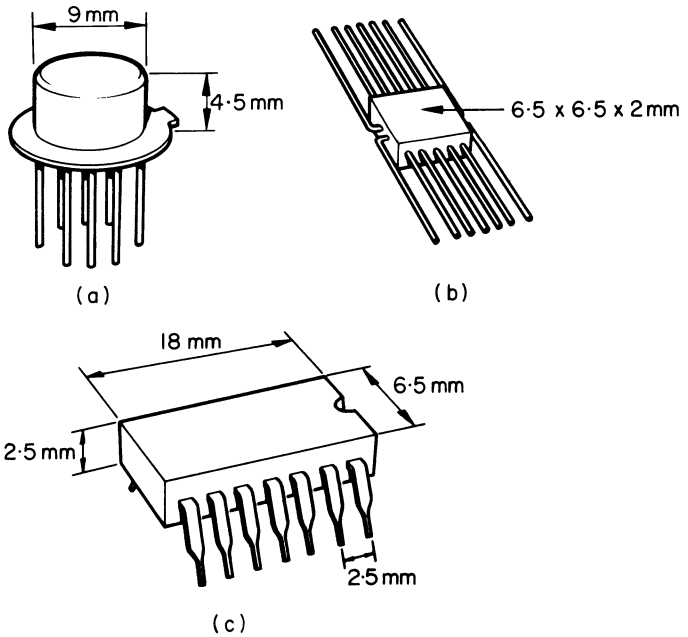


Figure 6.5 Encapsulation of integrated circuits

6.4 Parasitic Components in ICs

The components in each IC are isolated from one another by reverse biasing the p–n diode formed between the components and the substrate. In the case of the IC in figure 6.4e, the p-type substrate is connected to the point in the circuit with the lowest potential (frequently this is the chassis of the equipment).

While providing electrical isolation between the components, this technique has the unfortunate effect of introducing a number of *parasitic diodes* into the IC. Each parasitic diode has two undesirable effects; firstly, each provides a leakage path for the flow of current and, secondly, since the diodes are reverse biased, each introduces a parasitic capacitance. These parasitic components contribute towards the degradation of the high-frequency performance of the IC.

6.5 Monolithic Diodes

Diodes used in ICs are frequently constructed from bipolar transistor arrays. For example, the diode in figure 6.4e uses the equivalent base and emitter of an n–p–n transistor as the anode and cathode, respectively, of the diode.

Three basic diode configurations are shown in figure 6.6; in each case the anode is marked A, and the cathode K. The diode in figure 6.6a has the lowest forward p.d. of the configurations shown, but its reverse breakdown voltage is limited to about 7 to 9 V. The latter comment also applies to the diode in figure 6.6b, which also uses the base–emitter junction of the transistor; the forward p.d. of this diode has an intermediate value between the diodes in diagrams (a) and (c). The diode in diagram (c) uses the collector–base p–n junction, and has the highest breakdown voltage rating of those illustrated, together with the highest forward p.d.

6.6 Schottky Diodes and Schottky Transistors

It was shown in chapter 1 that the junction between a metal and a semiconductor can produce either an ohmic connection or a rectifying connection. Also, in section 6.2 it was mentioned that aluminium acts as though it were a p-type material when in contact with silicon. The latter fact is utilised in monolithic Schottky diodes, an example of which is shown in figure 6.7a; in this diagram the junction between the aluminium and the n-type epitaxial layer produces a Schottky

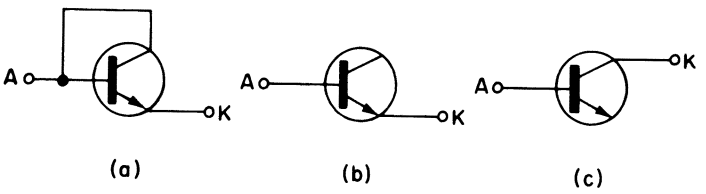


Figure 6.6 IC diodes

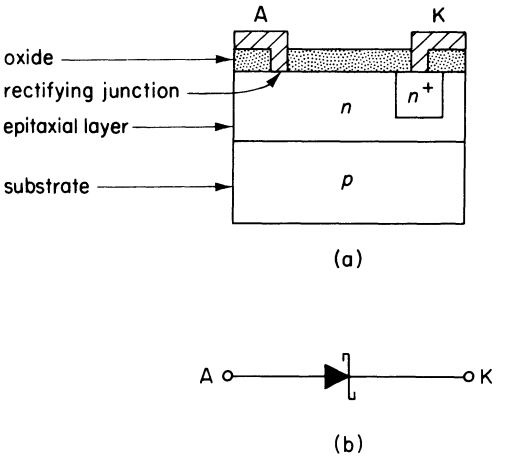


Figure 6.7 (a) A cross-section through a monolithic Schottky diode and (b) a circuit symbol for this type of diode

rectifying junction. The contact between the aluminium and the n^+ diffusion at the right-hand side of the figure forms an ohmic junction; the reason for this was described in section 6.2.

The principal advantages of the Schottky diode over conventional silicon p–n junction diodes are

(1) the forward p.d. at which conduction commences is about 0.3 V, compared with about 0.5 to 0.6 V for p–n diodes;
(2) Schottky barrier diodes have negligible storage time (see section 3.10).

One limitation of Schottky diodes when compared with p–n junction diodes is their relatively low values of reverse breakdown voltage.

In high-speed TTL gates (see also section 4.17), the propagation delay can be reduced by using what are known as *Schottky transistors* or *Schottky diode clamped transistors* instead of conventional bipolar transistors. A section through a transistor of this kind is shown in figure 6.8a; an important feature to note is that the base metallising connection (B) bridges the p-type base region and the n-type collector region. This results in the equivalent electrical circuit in figure 6.8b, in which a Schottky diode is shown connected between the base and the collector of the transistor.

The function of the diode is to prevent the transistor from being driven into saturation, as follows. When an attempt is made to saturate the transistor by increasing the value of the base current, the collector voltage of the transistor begins to fall. When a conventional transistor is driven into saturation, the collector voltage falls below that of the base by about 0.5 V. In the Schottky transistor, the diode limits the value of the forward p.d. between the base and the collector to

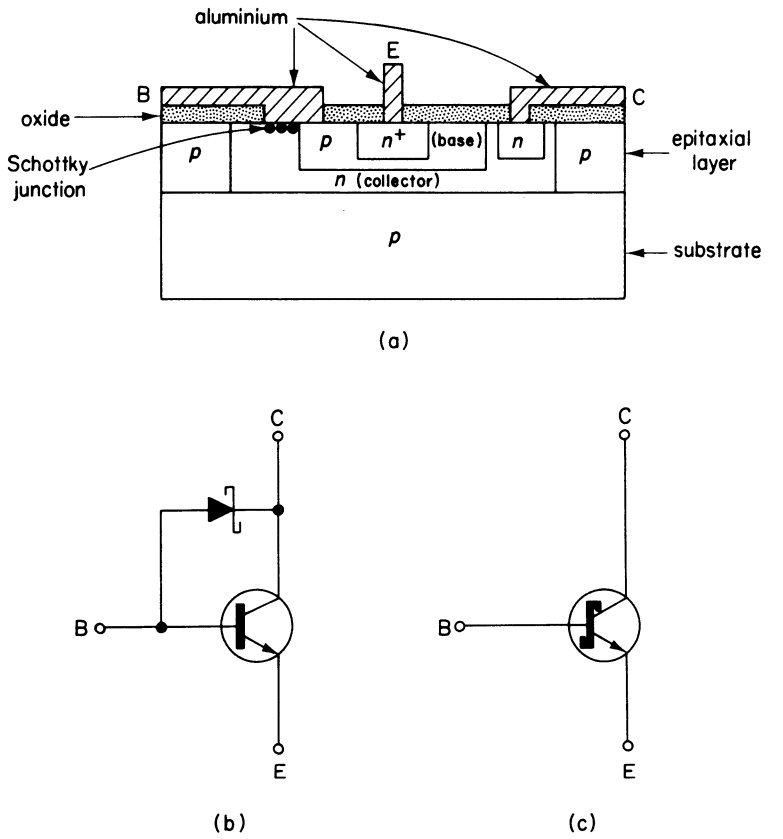


Figure 6.8 (a) A cross-section through a monolithic Schottky transistor, (b) the equivalent electrical circuit and (c) a circuit symbol for the transistor

about 0.3 V (that is, to the forward p.d. of the diode), thereby preventing the transistor from becoming fully saturated. The net effect is to reduce the stored charge in the transistor, so that when the transistor base current is suddenly reduced to zero, the storage-delay component of the turn-off time is very small. The over-all result is a significant reduction in turn-off time when compared with that of a conventional BJT.

6.7 Monolithic Capacitors

Small values of capacitance can be obtained by utilising the transition capacitance of a reverse biased p–n junction diode, illustrated in figure 6.9a. The capacitance obtained by this means is, of course, voltage-dependent.

Alternatively, metal-oxide-semiconductor capacitors can be introduced in the manner shown in figure 6.9b. Here the upper electrode is the aluminium conductor,

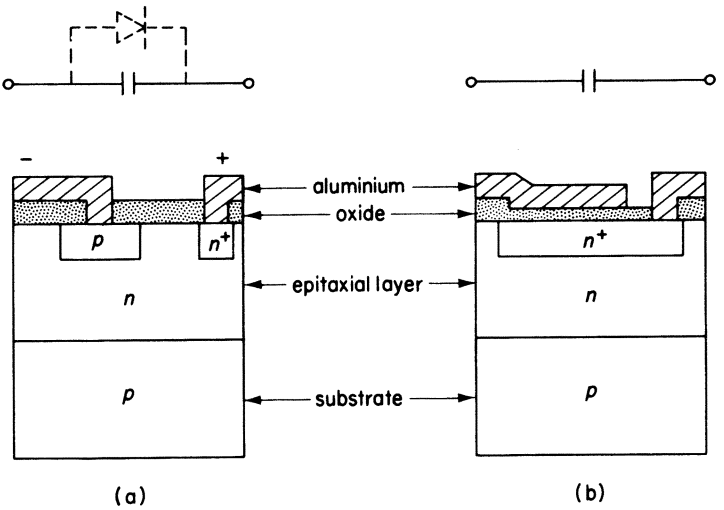


Figure 6.9 (a) A capacitance can be formed using a reverse biased p–n junction diode and (b) a MOS capacitor

and the lower electrode is an n^+-region introduced into the epitaxial layer of the IC. The oxide acts as the dielectric.

6.8 Inductors

Inductors cannot be satisfactorily manufactured in monolithic form. If possible, circuits should be redesigned to avoid the use of inductors. If this is not possible, then inductors must be connected as external discrete components to the IC.

6.9 MOS Transistors

MOS transistors are relatively easy to manufacture in monolithic form – sections through two types of structure were illustrated in figures 5.10 and 5.12. The majority of MOS logic applications employ either p-MOS or CMOS logic, the latter having been discussed in section 5.15 (see also figure 5.16).

An advantage of MOS devices over bipolar elements is that the complex isolation methods required in the production of the latter are not necessary in MOS devices. Moreover, MOS devices occupy only a fraction of the area required by bipolar elements so that, for a given chip area, either more circuits can be manufactured or a more complex logic function can be generated than is the case with bipolar elements.

The threshold voltage of MOSFETs can be reduced in monolithic circuits by employing silicon nitride in the gate insulation. If the dielectric strength of the gate insulation can be increased, the thickness of the gate insulation can be reduced. A

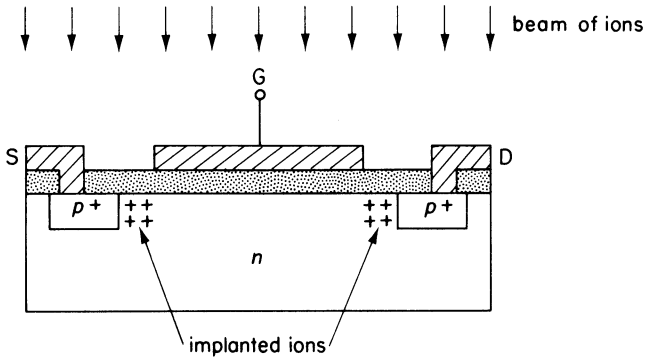


Figure 6.10 Ion implantation in MOS devices

dielectric material which may be used for this purpose is silicon nitride but, unfortunately, it results in unwanted changes in the characteristics of the transistor. A compromise solution consists of a gate insulation comprising a layer of silicon oxide (about 0.06 μm thick) and a layer of nitride (about 0.04 μm thick). The gate electrode is deposited on the surface of the nitride. The resulting FET is known as a *metal-nitride-oxide-semiconductor* (MNOS) transistor. The threshold voltage is reduced to about 2 V by this means.

The switching speed of FETs can be increased if the interelectrode capacitances of the FET can be reduced. One method of achieving this is to reduce the overlap which occurs between the gate electrode and the diffusions which form the source and the drain. If the MOSFET is fabricated initially with a gate electrode that does not overlap the diffused regions (see figure 6.10), and the whole surface is bombarded with ions that have been accelerated to energies in the range 50 to 100 keV, the ions are implanted in the epitaxial layer in areas not masked by the aluminium. In this way the effective size of the electrodes is increased, but no overlap exists between the gate metallisation and the drain and source electrodes. The implanted ions must be of the correct kind for the type of channel in use: boron for *p*-type and phosphorus for *n*-type (see table 1.3).

6.10 Medium-scale and large scale Integration

The terms *medium-scale integrated* (MSI) circuit and *large-scale integrated* (LSI) circuit are in common usage, and refer to the number of logic gates contained in a single IC package. The terms are not precisely defined, and the following are typical

MSI contain 10 to 100 gates per IC
LSI contain more than about 100 gates per IC

Typical MSI chips include adders, counters, decoders and shift registers. LSI chips include calculators, random-access memories (RAMs) and read-only memories (ROMs).

7 Charge-coupled Devices

A charge-coupled device (CCD) is a multigate MOS device capable of transferring electrical charge from its source electrode to its drain electrode in the form of a series of charge 'packages'. Charge-coupled devices have properties which are particularly useful in the fields of memory devices, in analogue delay lines and in solid-state optical imaging.

7.1 Basic Principles of CCDs

The principle of charge transfer can be understood from figure 7.1. Figure 7.1a shows a basic *p*-channel MOS CCD (*n*-channel devices are also manufactured). This device has an *n*-type substrate, and the application of a negative potential to the electrode on the surface of the CCD repels the mobile negative majority charge carriers away from the underside of the oxide. This results in the formation of a depletion region below the electrode. As the electrode potential is increased, the depletion region extends further into the substrate; at the same time, the negative potential on the gate attracts minority charge carriers (holes) to the depletion region until, at the threshold voltage, these minority carriers form a conducting channel at the oxide–semiconductor interface.

For the purpose of explaining the operation of the device, it is convenient to regard the depletion region as being a *potential well* (see figure 7.1a). Also, even though the 'holes' in the inversion layer are located immediately below the oxide, it is convenient to think of them 'filling' the lower part of the potential well. Thus, the charge 'package' is constrained in the well.

In the basic CCD structure (diagrams (b) to (e) of figure 7.1), three electrodes are required to complete the transfer of one charge packet. In consequence the three electrodes are referred to as one *element* of the CCD. The gap width between the electrodes is kept as small as possible to give a reasonable value of *charge transfer efficiency*, η (which usually has a value in the range 99.9 to 99.99 per cent), between the elements; the value of the voltage applied to the electrodes is usually greater than the threshold voltage of the device (the reason for this is given later). The mechanism of charge transfer is described in the following.

The drive lines supplying the electrodes are energised by a three-phase supply, the basic voltage waveforms applied to ϕ_1, ϕ_2 and ϕ_3 are as shown in the figure. Figure 7.1b illustrates the instant of time when the potential applied to ϕ_1 is

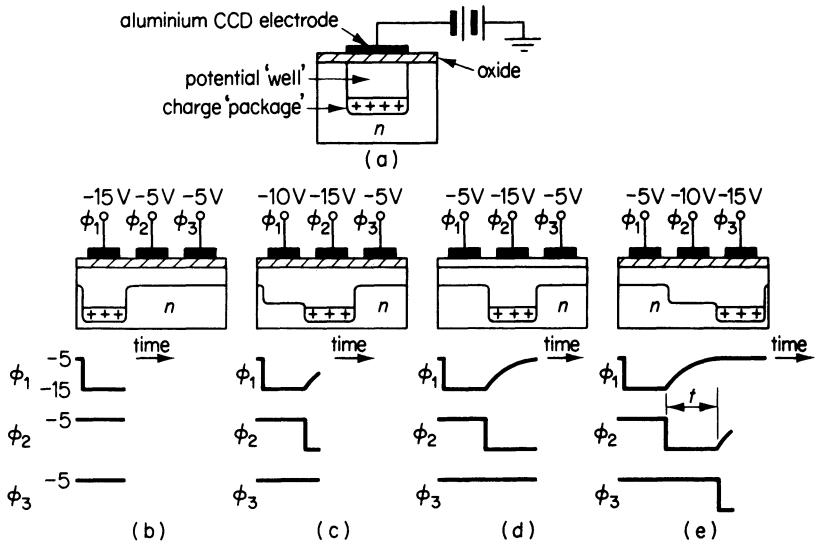


Figure 7.1 The principle of charge-coupled devices

greater than that applied to either ϕ_2 or ϕ_3, so that the potential well is deepest under ϕ_1. In consequence the charge package is located under ϕ_1. A short time later (figure 7.1c), the potential applied to ϕ_2 is increased, and that applied to ϕ_1 is reduced. This results in the charge packet transferring once more to the 'deepest' part of the potential well, this time under ϕ_2. When the potential applied to ϕ_1 has decayed to its smallest value (figure 7.1d), the potential well has transferred completely to ϕ_2. A little time later, the potential applied to ϕ_3 is increased (figure 7.1e), and that applied to ϕ_2 is reduced, and the charge packet is once more moved to the right. Since the charge cannot be transferred in zero time, it is necessary to allow the potential applied to the electrodes to decay slowly; the time, t, of this overlap is known as the *overlap time* (figure 7.1e).

7.2 A Three-phase CCD

A basic form of a two-element three-phase CCD is shown in figure 7.2. It consists essentially of two elements of the type described above, together with a means of injecting charge packets and a means of collecting them.

Charge injection

Let us suppose for the moment that the input signal is in the form of a series of digital signals that are to be transmitted through the CCD. Several modes of signal injection are possible, one is described in the following.

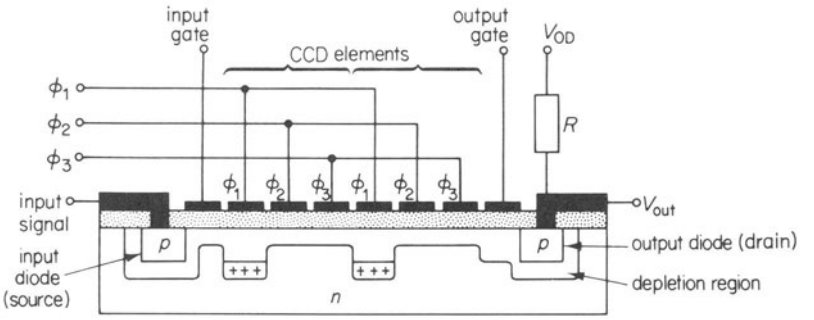


Figure 7.2 A three-phase CCD structure

If the *p*-region of the input diode is earthed, it can be used as an 'infinite' source of holes, which are the required type of charge carriers for the CCD shown. When an impulsive potential of negative polarity is applied to the input gate electrode (see figure 7.3), it creates an inversion channel between the input diode and the potential well under ϕ_1. As a result, a charge package is injected into the potential well.

The CCD shown has six transfer electrodes, and is capable of storing and transferring two packets of charge (or bits of data). Transfer of data between the input diode and the output diode is by the sequential application of the three-phase control potentials. A logic '0' signal is injected into the CCD by maintaining zero gate voltage during the time that ϕ_1 is supplied at a high voltage.

Charge collection

The output diode of the CCD (figure 7.2) is reverse biased, its depletion region coupling with that of the final storage electrode. As a result, when ϕ_3 falls to its smallest value, the charge stored in the final potential well is collected by the output diode, and results in the flow of an impulsive current in resistor *R*.

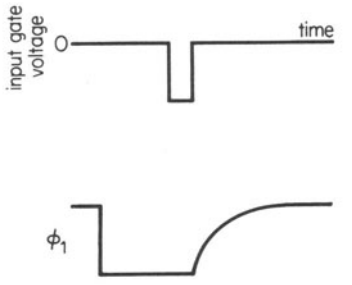


Figure 7.3 One method of charge injection

A typical set of voltages for a *p*-channel CCD are: input gate voltage, –4.5 V; output gate voltage, –5.5 V; clock voltages (ϕ_1, ϕ_2, ϕ_3), –30 V; drain supply voltage, –10 V.

7.3 Charge Transfer Efficiency

During each charge transfer process, a small amount of the charge is 'left behind' in its potential well. This charge manifests itself as a form of distortion of the output signal. The two main reasons for the 'loss' of charge are

(1) incomplete transfer in the time allowed for the process,
(2) charge carriers are trapped in carrier-trapping levels in the silicon–oxide interface.

The lowest frequency of operation is set by the fact that the potential wells may be 'filled' by minority carriers generated in the substrate by thermal effects at very low frequencies. This effectively destroys the information stored in the potential wells. The upper frequency limit is controlled to a great extent by the physical geometry of the device. A typical range of operating frequency is from about 50 kHz to about 10 MHz.

The worst effects of charge-carrier trapping can be overcome in three-phase devices by the so-called *fat zero* technique. In this operating mode, the electrode voltages never fall below a minimum value which allows a constant trickle of current through the device.

7.4 Digital Shift Register

A digital shift register is a sequence of memory elements which store information in digital form. A shift register is one in which the information is moved or is 'shifted' along the register by means of the applied clock pulses. Three-phase CCD described earlier is one form of digital shift register, in which digital information is entered into the register (the CCD) at the input diode and, after two complete sets of clock pulses (ϕ_1, ϕ_2, ϕ_3), the data is 'shifted' to the output of the register. The 'length' of the register depends on the number of CCD elements in the device.

The CCD shift register can also be used as a *digital delay line* or as a *dynamic memory*, simply by continuously recirculating the stored data as follows. As the data appears at the output of the CCD, the waveform is regenerated into a rectangular wave and is reinserted into the input of the register. In this way the information is stored in a dynamic mode, the circulation rate being dependent on the clock frequency.

7.5 Analog Shift Register or Delay Line

In this mode of operation (see figure 7.4), the input signal is injected into the input diode via blocking capacitor C_1; the input gate is energised so that its depletion

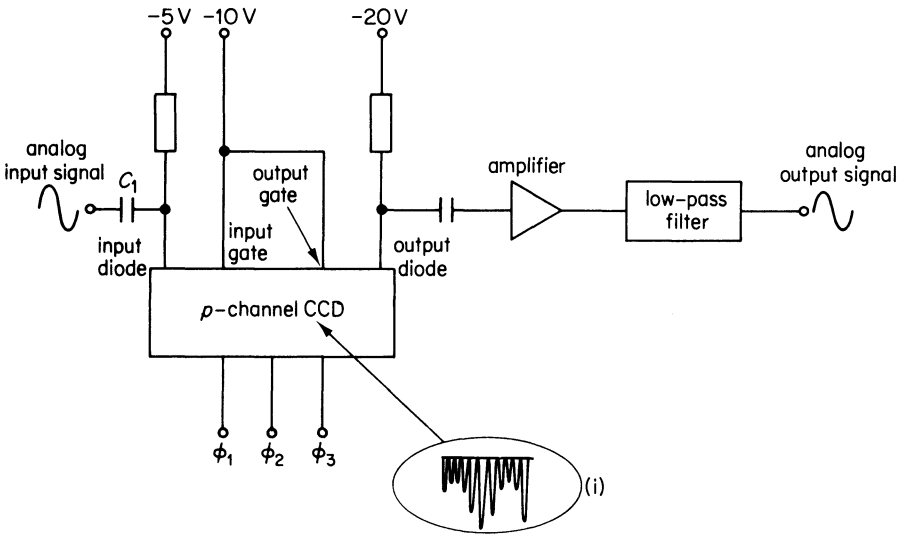


Figure 7.4 A *p*-channel CCD analog shift register

region links with that of the input diode and with that of the first storage electrode. When the potential well under ϕ_1 is deepest, holes flow from the input diode into the potential well under ϕ_1; the charge stored in the well is dependent on the magnitude of the input signal. The three-phase sequence of clock pulses causes this charge package to be transmitted along the register.

Some time later, when the potential well under ϕ_1 is again at its deepest, the input signal is again sampled and its magnitude is converted into the form of a charge package. Thus the analog signal at the input of the CCD is 'sampled' at the clock pulse-rate, and is transmitted along the register in a series of pulses, the amplitude of each pulse being related to the amplitude of the input signal at the instant of sampling. Inset (i) in figure 7.4 illustrates the nature of the signal transmitted along the CCD.

The output signal is finally amplified, the high-frequency components in the signal being removed by a low-pass filter. This leaves an output waveform which has the same shape as the input signal, but is delayed in time by the time taken for the signal to propagate through the CCD and the associated equipment.

7.6 Optical Imaging

When light irradiates the surface of silicon, minority charge carriers are released. In a CCD structure, these charge carriers flow into the potential wells under the electrodes. By allowing the charge to accumulate for about 1 ms (this period of time is known as the *integration time*), a detectable amount of charge which is proportional to the incident illumination is stored in the potential wells. Applying a series of three-phase clock pulses 'shifts' the stored data along the register.

A CCD can store a 'line' of optical information whose output can be used, for example, in conjunction with a cathode-ray tube to give a visual display. This technique can be extended for use as an area imager to give a complete TV picture. Applications include closed-circuit TV, videophones and electronic page readers.

7.7 Two-phase and Four-phase Devices

Devices operating from a two-phase clock supply offer some attractions over the three-phase devices described above. Firstly, they require simplified clocking arrangements and, secondly, a smaller surface area is required. A feature of the basic two-phase system is that data can be moved in one direction only, but this is rarely a disadvantage. A variety of structures can be used to impart directionality to data flow, some requiring fairly complex fabrication processes.

Structures using four-phase clocking systems produce some improvement in clocking performance when compared with other types. Four-phase CCDs employ a two-level conductor arrangement, and can move the stored data in either direction.

7.8 CCD Logic Gates

A variety of logic gates have been fabricated which use the CCD structure. A NOR-type gate can be constructed which uses transfer electrodes which are in series with one another between the input and output diodes. A NAND-type gate structure would have the gates arranged in parallel around the output diode.

8 Semiconductor Memories

8.1 Types of Memory

Many electronic circuits require memory elements to record the state of operation of the circuit at particular instants of time. The most basic form of memory element, known as the *set–reset flip–flop* (or the S–R flip–flop), consists of two cross-connected NOR gates (for details see section 8.2). This circuit is widely used not only as a memory element in its own right, but also as the basis of more sophisticated elements known as *master–slave flip–flops* (see section 8.3). The S–R flip–flop is one of a family of memories known as *static memories*, which retain the stored information indefinitely so long as the power supply is maintained; both bipolar and MOS technologies are used in the manufacture of static memories.

Another family of memories, known as *dynamic memories*, use MOS technology, and depend for their operation on the ability of the gate insulation of MOS devices to retain an electrical charge for a relatively long period of time. Ultimately, the charge stored in the gate dielectric decays in value, and must periodically be 'refreshed'.

In general, the information stored both in static and in dynamic memories is lost when the power supply fails. Such storage systems are known as *volatile stores*.

A semiconductor memory chip contains an array of cells, usually in the form of a matrix of the type in figure 8.1. Each cell can be independently *addressed* by energising appropriate X- and Y-address wires. For example, if row wire B and column wire F are energised, then cell V is addressed; by using additional wires (not shown), information can either be *written* into or *read* from the cell. Addressing a particular *location* in the memory in this way is known as *X–Y selection* or *coincident selection*; this allows one BInary digiT or *bit* either to be written into or to be read from the selected location. If the organisation associated with the addressing logic is altered, it is possible to access several cells simultaneously. For example, if all three X-wires are energised and, if column wire F is also energised, then cells U, V and W are simultaneously addressed. This is known as *word selection* or *linear selection*; a binary *word* is a group of binary digits which form the normal unit in which information is stored. In figure 8.1, a word contains three bits. The store shown may be regarded as storing either nine bits of data or three words each of three bits.

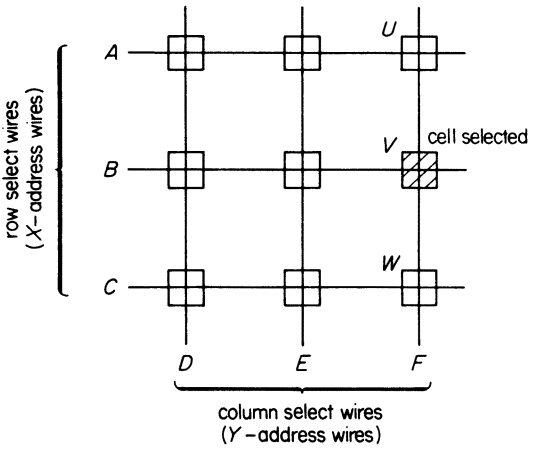


Figure 8.1 One method of addressing an individual cell in a random access memory

The type of memory in figure 8.1 is known as a *random-access memory* (RAM), since the bits (or words) stored may be selected at random. This type of memory is capable of being accessed very rapidly.

Memories known as *read-only memories* (ROM) store data that cannot normally be altered. The data stored in the ROM is frequently specified by the user, and is inserted in the ROM either at the manufacturing stage or, in the case of electrically *programmable ROMs* (PROMs), when it is installed on site. These memories are *non-volatile*, and are used in a number of applications including storing *micro-programs* for computer control, and for storing tables (that is, sine, cosine, etc.); also for storing character patterns for use with optoelectronic display devices. The information stored in *reprogrammable ROMs* (RPROMs) can be altered either by electrical or optical methods; the data stored in these devices decay very slowly, and refreshing is required only at very infrequent intervals.

8.2 The S–R Flip–Flop

The basic circuit is shown in figure 8.2a and has two input lines, which are the set line or S-line and the reset line or R-line, and two output lines. The 'normal' output line is taken from terminal Q, and the second output line designated $\overline{Q}$ provides a signal which is the logical complement of the signal on the Q-line. Table 8.1 is a 'dynamic' truth table of the S–R flip–flop, in which the state of output Q prior to the stated values of S and R being applied is Q_n, and its state after the conditions listed have been applied is Q_{n+1}.

If the initial state of Q is logic '1' ($Q_n = 1$), then when $S = 0$, $R = 0$ the output remains unchanged at '1' (that is, $Q_{n+1} = 1$); the output also remains unchanged for the same input conditions if the initial value of Q was '0'. However, if input

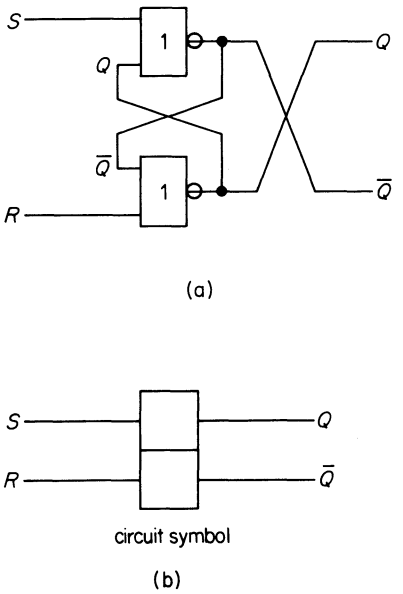


Figure 8.2 The S–R flip–flop

conditions $S = 0$, $R = 1$ are applied, output Q is reset to '0' ($Q_{n+1} = 0$) irrespective of the previous state of Q. Similarly, if input signals $S = 1$, $R = 0$ are applied, then Q is 'set' to logic '1' ($Q_{n+1} = 1$). The X in the final row of the Q_{n+1} column of the truth table is described as a 'don't know' condition; this suggests that we cannot be certain of the resulting state of output Q. Strictly speaking, this statement is not quite correct since – with the circuit shown – the input conditions $S = 1$, $R = 1$ result in outputs of $Q = 0$, $\bar{Q} = 0$, which is a contradictory state of affairs since if $Q = 0$, then $\bar{Q}$ should be '1'. So far as possible, the input state $S = R = 1$ should be avoided.

Table 8.1 Truth table of the S–R flip-flop

Inputs		Output
S	R	Q_{n+1}
0	0	Q_n
0	1	0
1	0	1
1	1	X

8.3 Master–Slave Flip–Flops

As high-speed computer systems were developed, many forms of flip–flop were investigated in order to establish the most suitable type. The most popular family for a wide range of uses was found to be the master–slave family, whose basic principles can be understood from figure 8.3.

This type of flip–flop contains two memory stages known as the master stage and the slave stage. The operation of the circuit is controlled by means of switches S1 and S2, whose functions are regulated by a 'clock' signal. The clock signal is a square waveform which ensures that when S1 is open then S2 is closed and vice versa. When the clock signal is at logic '0' (see figure 8.3b), S1 is open and S2 is closed, and when it is at logic '1' then S1 is closed and S2 is open. The operating sequence is as follows.

Clock signal '0' (point A on figure 8.3b) The master stage is isolated from the input signal, and is connected to the slave. Output Q from the slave stage (which is also the output from the complete master–slave flip–flop) also stores the state of the master stage.

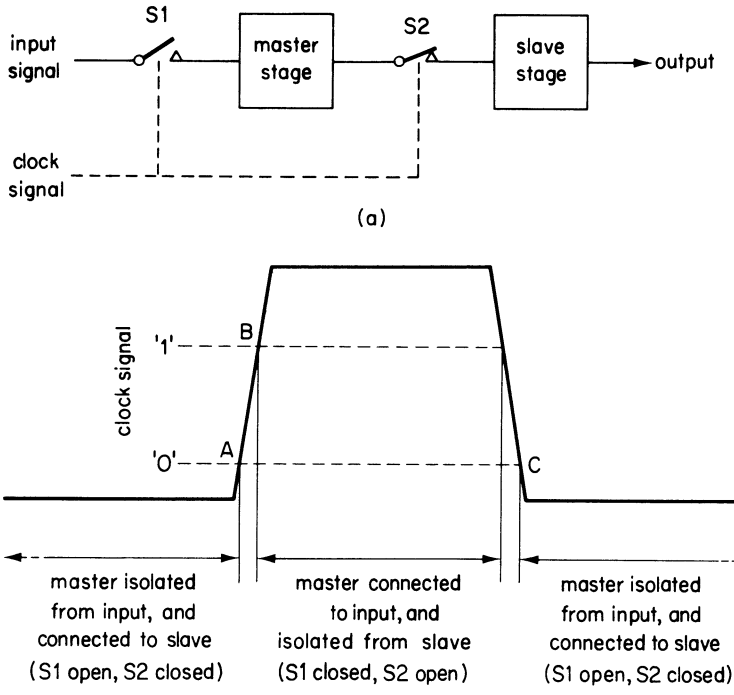


Figure 8.3 The basis of the master–slave flip–flop

Clock signal '1' (point B on figure 8.3b) The master is isolated from the slave, and is connected to the input signal. The master stage therefore stores the new state of the input signal. The slave stage continues to store the previous state of the master stage, so that the output from the slave stage is unchanged.

Clock signal '0' (point C on figure 8.3b) The master stage is isolated from the input signal, and is connected to the slave stage. In this condition, the new state stored by the master is transmitted to the slave, and thence to the output.

Thus data is gated through the flip–flop by the clock pulse, and finally appears at the output terminals when the clock signal falls to logic '0'.

The most popular type of master–slave circuit is the *J–K flip–flop*, illustrated in figure 8.4. In this flip–flop, gates G1A and G1B fulfil the function of S1 in figure 8.3, and gates G2A and G2B are equivalent to S2. The correct operating sequence of the electronic switches is obtained by the use of invertor gate G3. The latter gate ensures that when the clock signal is at logic '0', a logic '1' is applied to G2A and G2B, and allows data to be transferred from the master stage to the slave stage. The truth table for the J–K flip–flop is given in table 8.2.

If, in table 8.2, we consider the *J*-line as being equivalent to the *S*-line of the S–R flip–flop, and the *K*-line and *R*-line are considered to be equivalent, then the first three rows of the J–K flip–flop are equivalent to those of the S–R flip–flop. That is, the J–K element can perform the function of the S–R flip–flop. For this reason, J–K elements have largely supplanted S–R flip–flops. The truth table of the J–K flip–flop differs from that of the S–R element only in the final line. In this case, when $J = K = 1$, the output from the flip–flop *changes state each time the clock pulse falls to zero*. That is, if the initial value of Q is '1', the sequence of

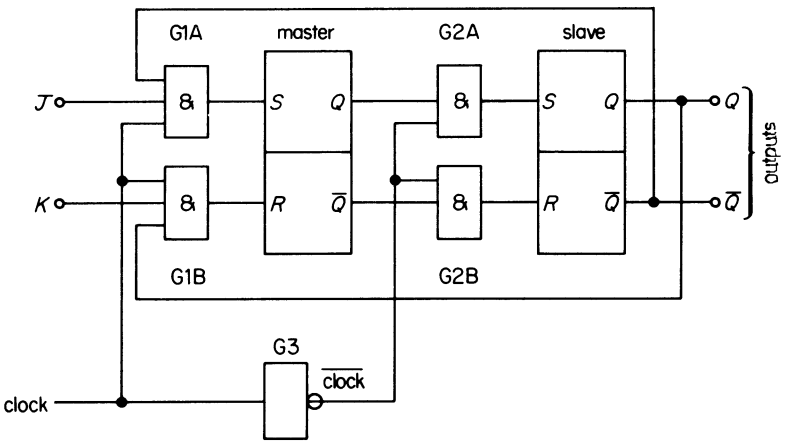


Figure 8.4 One form of master–slave J–K flip–flop

Table 8.2 Truth table for the master–slave J–K flip–flop

Inputs		Output	
J	K	Q_{n+1}	
0	0	Q_n	S–R operation
0	1	0	
1	0	1	
1	1	$\bar{Q}_n$	trigger or toggle operation

logical values of Q following the application of a train of clock pulses is 0, 1, 0, 1, 0, 1, . . . This feature is brought about by the pair of cross-connected feedback links in figure 8.4 between Q and G1B and between $\bar{Q}$ and G1A, respectively.

The J–K element is a most versatile circuit, and is the most widely used individual type of memory element.

8.4 Static RAMs

A basic unit or *cell* of the bipolar static RAM is the S–R flip–flop. One form of RAM cell is illustrated in figure 8.5, in which TR1 and TR2 are the active elements

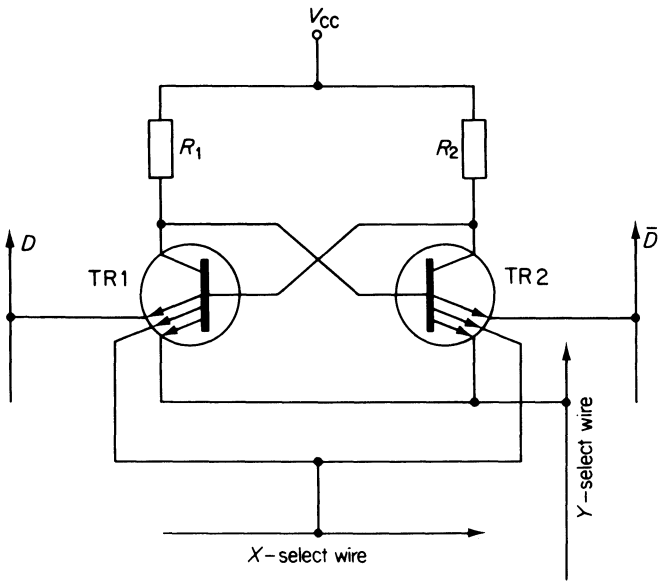


Figure 8.5 One form of bipolar static memory cell

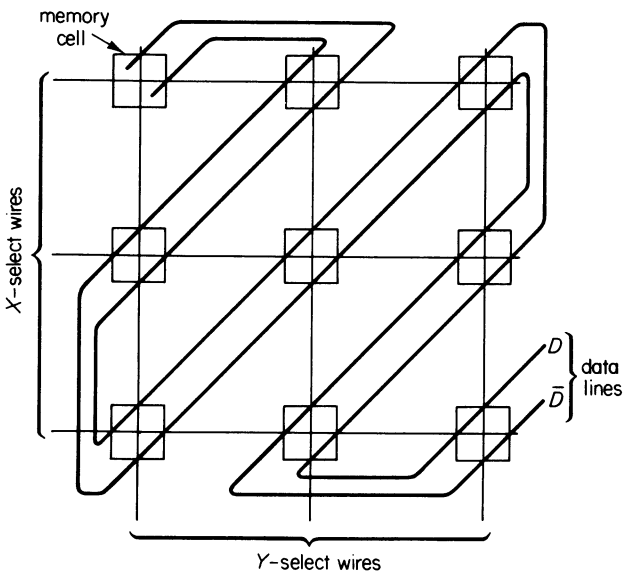


Figure 8.6 A static nine-bit memory matrix

in a cross-connected NOR flip–flop. In operation, when the cell is storing data, the X- and Y-select wires are at a low potential so that the current in the transistor which is ON flows to both of the 'selection' lines.

To *read* the data stored, both the X- and the Y-select wires are raised to a positive potential; this causes the emitter current to flow either into the D-line or the $\bar{D}$-line (depending on which transistor is ON). The state of the stored data is therefore indicated by a current pulse in the appropriate data line. To *write* data into the cell, both the X- and Y-select lines are raised to a positive potential and, if TR1 is to be turned ON, line D is switched to a low voltage and $\bar{D}$ is raised to a positive potential.

A basic method of organising a nine-bit X–Y select static RAM is illustrated in figure 8.6. Each X- and Y-row is addressed by a single wire, and a common pair of data lines is associated with each cell. By addressing one X-line and one Y-line, only one cell is selected; data can then either be read from or written into that cell in the manner described above.

8.5 Dynamic RAMs

A popular form of dynamic memory cell employing three MOSFETs is shown in figure 8.7. The data is stored in the form of an electrical charge on the gate capacitance, C, of transistor TR2. Activating the 'write select' line turns TR1 ON, and connects capacitor C to the 'write data' line; either a '1' or a '0' may be written into the memory element by connecting an appropriate voltage to this line or,

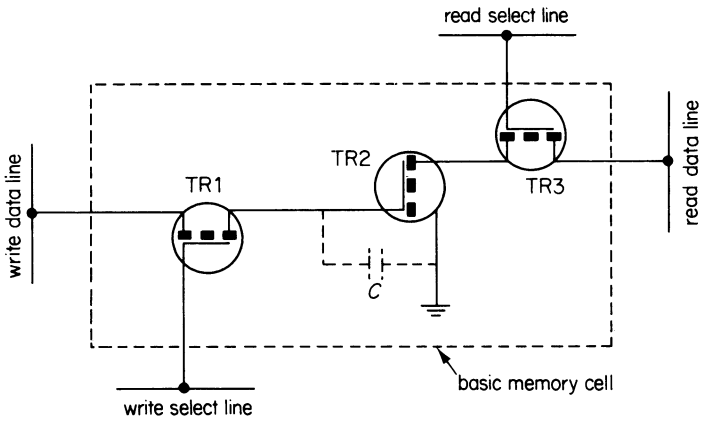


Figure 8.7 A three-transistor dynamic MOS memory cell

alternatively, the stored data may be 'refreshed'. Since the stored charge in the capacitor ultimately decays, it is necessary to carry out the refreshing operation or recharging operation every few milliseconds.

To read data from the cell, the 'read select' line is energised, which turns TR3 ON and connects the 'read data' line to TR2. If a logic '1' is stored, TR2 is ON and current flows in the 'read' line. If a '0' is stored, TR2 is OFF and no current flows in the 'read' line.

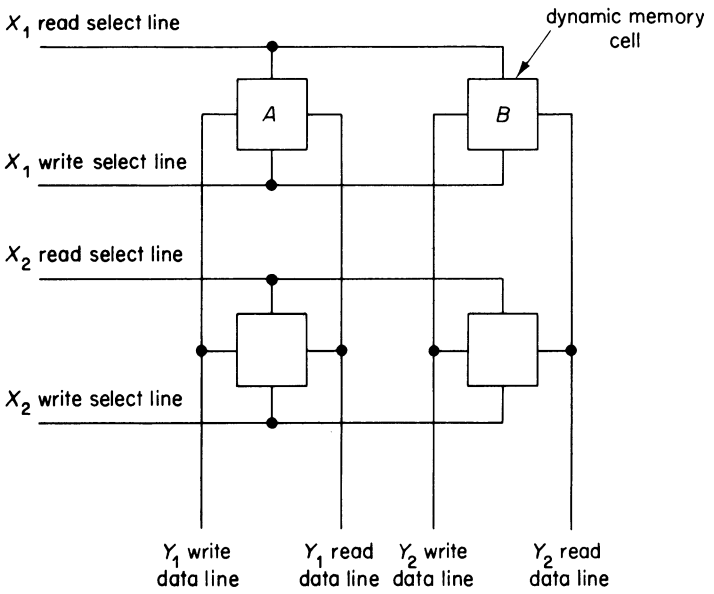


Figure 8.8 Memory organisation for a four-bit dynamic RAM

One form of organisation of a four-bit dynamic memory matrix is shown in figure 8.8. When the X_1 'read select' line is energised, cells A and B are addressed, and data is read from cell A by monitoring the current in the Y_1 'read data' line. Data is written into cell A by energising the X_1 'write select' line and, simultaneously, applying the data to the Y_1 'write data' line. Data is refreshed by simultaneously reading data from the cell and writing it back in again.

8.6 Review of RAMs

Random access memories are manufactured using a wide range of technologies including bipolar, *p*-MOS, *n*-MOS, CMOS and SOS. The capacity of the memories is in the range between about ten bits to several thousand bits. Small high-speed memories (known as *scratch-pad* memories) are usually static memories using bipolar technology. The larger units, usually functioning at a slower speed than scratch-pad memories, are frequently dynamic memories using MOS technology.

8.7 Content Addressable Memories (CAMs)

Content addressable memories are designed with the computer programmer in mind, rather than the designer. These memories are addressed by their contents (or a part of their contents), rather than by the physical location within the memory. For example, an employer may wish to know the names of all his employees under 40 years of age, earning over £5000 and with a mathematics 'A'-level qualification. This data can be used to 'address' the CAM.

8.8 Read-only Memories (ROMs)

In general, ROMs contain information which is accessed frequently, but is changed only infrequently (or never changed in some cases). The name 'ROM' is perhaps misleading, since the information must be 'written' into them at some time. Some ROMs have data written into them only once during their lifetime, while others have the data changed at infrequent intervals. Since ROMs generally store information which is rarely changed, it is possible to organise the 'reading' cycle so that the data can be accessed far more rapidly than is the case with a RAM. On the other hand, when the information is altered in a ROM, the 'writing' time takes considerably longer than is the case with a RAM.

There are three broad categories of ROMs, namely *mask programmable, electrically programmable* and *reprogrammable*. Versions of the three types are described below.

Mask programmable ROMs

The program stored in this type of ROM (which may be specified by the customer) is introduced into the ROM during the manufacture of the IC. Mask programming can be understood from figure 8.9.

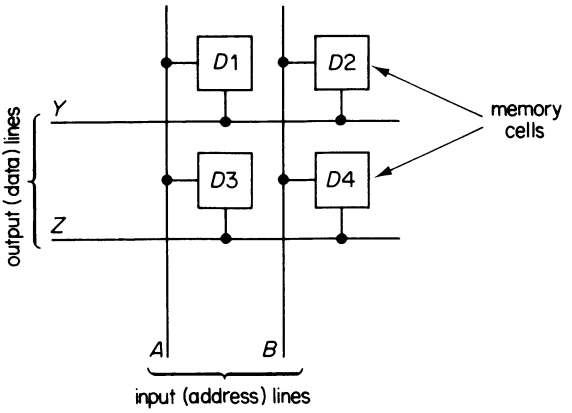


Figure 8.9 The basis of the mask programmable ROM

As outlined earlier a number of photographic masks are used in the manufacture of monolithic ICs. By altering some of the masks it is possible to cause some of the devices (which are either diodes or transistors) either to be permanently ON or to be permanently OFF. In this way the memory cell either stores a '0' or a '1'.

The semiconductor memory array in figure 8.9 contains four such cells ($D1$ to $D4$), and if the ROM had been mask programmed so that $D1$, $D2$ and $D3$ are ON, and $D4$ is OFF, then the application of a logic '1' signal to address line A causes 1s to appear at data lines Y and Z. A logic '1' applied to address line B causes a '1' to appear on output line Y, and a '0' on line Z.

Electrically programmable ROMs (PROMs)

A disadvantage of mask programmable ROMs is that the initial manufacturing cost is high unless, that is, very large production quantities are involved. To overcome this problem, particularly where small production quantities are involved, electrically programmable ROMs can be used.

A simple PROM is shown in figure 8.10. It is electrically similar to the matrix in figure 8.9, but with each cell in the form of a diode in series with a fusible link. This link may, for example, be in the form of either a thin aluminium link or a polysilicon fuse.

In cases where a diode is to be left in the ON state, the fuse is left intact. Where it is to be off, the fuse is blown by the application of a current pulse between appropriate address and data lines. This process can be carried out by the manufacturer or by the user by means of suitable electronic apparatus.

Reprogrammable ROMs (RPROMs)

Reprogrammable memory cells are MOS elements constructed in such a way that the application to the gate region of a voltage above a certain critical value causes

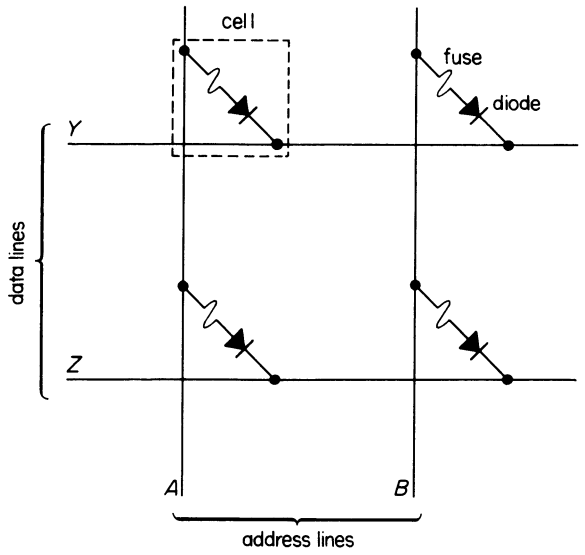


Figure 8.10 One form of PROM

charges to be trapped in the insulation below the gate. This produces a conducting inversion layer which links the source and drain of the MOSFET. The amount of charge stored depends not only on the value of the applied voltage, but also on the time interval for which it is applied. Over a period of time the charge slowly leaks away and, for a low value of applied voltage and for a short time-duration, data may be stored for, say, one day. A higher voltage and a longer time-duration cause the charge to be retained for a longer period – a figure quoted by one manufacturer is that a 24 V, 10 ms pulse results in data being retained for more than 10^{11} read accesses, which may amount to many years in use.

Data may be erased by applying a voltage pulse to the gate, but of reverse polarity to the 'writing' pulse. In other cases, data is erased by exposing the chip to ultraviolet radiation.

Since reprogrammable ROMs are most frequently used as read-only memories, they are also known as *read-mostly memories* (RMM). Another name given to them is *electrically alterable read-only memories* (EAROM).

8.9 Programmable Logic Arrays (PLAs)

A ROM may be considered to be a combinational logic array, having outputs which are logical functions of the inputs. In this way the ROM can be used to replace one or more logic networks. When used in this way the device is known as a programmable logic array. Once the truth table of a particular logic network has been specified, the equivalent logic network can be replaced by a PLA.

8.10 Amorphous Memories

An amorphous substance is one that is non-crystalline, and amorphous semiconductors (such as some types of glass oxide) have a resistivity of the order of 10^7 Ω cm. An amorphous semiconductor can be converted into a crystalline form having a resistivity of about 0.1 Ω cm by applying a critical value of electrical field strength to it. To reset the cell into its amorphous state, a high current pulse is applied to the cell. With the above technique, programmable ROMs can be manufactured using amorphous semiconductors.

9 *Thyristors and Other Multilayer Devices*

A thyristor is a bistable semiconductor switching device that functions as a 'controlled diode', having an OFF state or *blocking state*, and is triggered into its ON state or *conducting state* by the application of a signal to its *gate* electrode. Other methods of triggering the device ON are possible, and are discussed below. There are two main forms of thyristor, namely

(1) the *reverse blocking thyristor* (referred to simply as a *thyristor*)
(2) the *bidirectional thyristor* (referred to as a *triac*).

9.1 The Reverse Blocking Thyristor

The thyristor is a four-layer p–n–p–n device whose operation can be explained in terms of a *two-transistor analogy* (see figure 9.1). By dividing the centre p–n regions of the thyristor in the manner shown in figure 9.1b, two transistors are formed. Transistor TR1 is an n–p–n device and TR2 is a p–n–p device (figure 9.1b); the two transistors are interconnected in the manner shown in figure 9.1c. The extreme p-region acts as the anode of the thyristor, and the extreme n-region as its cathode. Region p_2 in figure 9.1a is normally used as the gate electrode (marked G in the figure); this electrode is sometimes known as the *cathode gate*, this name being due to the proximity of the gate region to the cathode. The n_1 region (marked G2 in figure 9.1a) is also used as a control electrode in some devices (see section 9.9 – the PUT); this gate is sometimes known as the *anode gate* owing to its proximity to the anode of the thyristor.

A symbol frequently used to represent the thyristor is shown in figure 9.1d, while the symbol in figure 9.1e is a general symbol used to represent the thyristor. The operation of the thyristor is explained in terms of the two-transistor analogy below.

Anode negative with respect to the cathode

In this case the junctions p_1–n_1 and p_2–n_2 are reverse biased, and only leakage current flows through the thyristor. In this operating state, the device blocks the flow of current and is said to be in its *reverse blocking mode* (see figure 9.2). In this

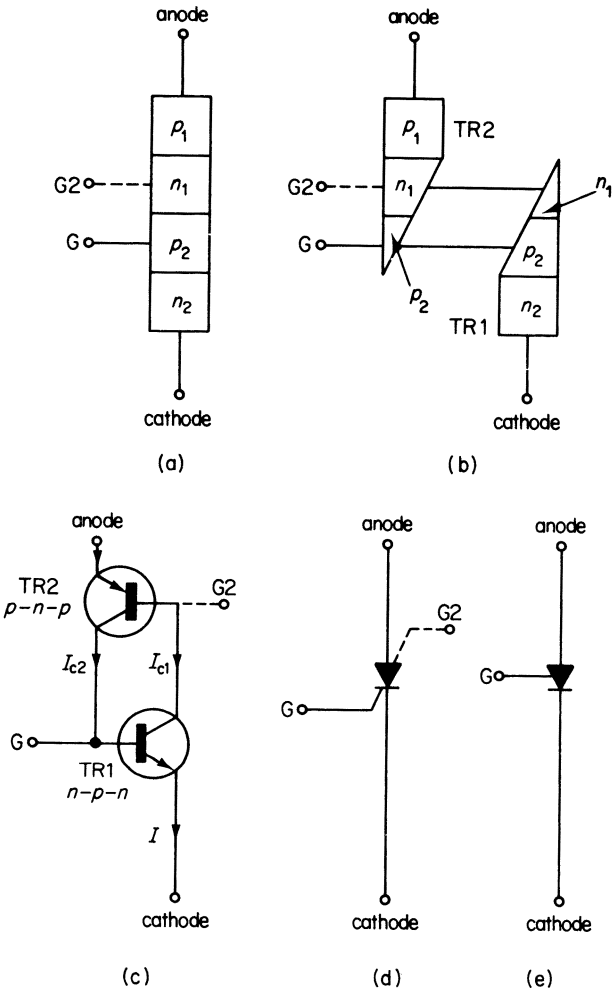


Figure 9.1 A two-transistor analogy of the thyristor

mode, any signal applied to the gate region has no effect on the operation of the thyristor. If the magnitude of the negative (or reverse) anode voltage is increased, a point is finally reached where the anode current increases rapidly. This is known as *reverse breakdown*, and is due to avalanche breakdown of the device. Unless the current is limited to a safe value, the device may be destroyed.

Anode positive with respect to the cathode

In this condition the junctions p_1-n_1 and p_2-n_2 are forward biased but, in the absence of a gate signal, junction p_2-n_1 is reverse biased and blocks the flow of

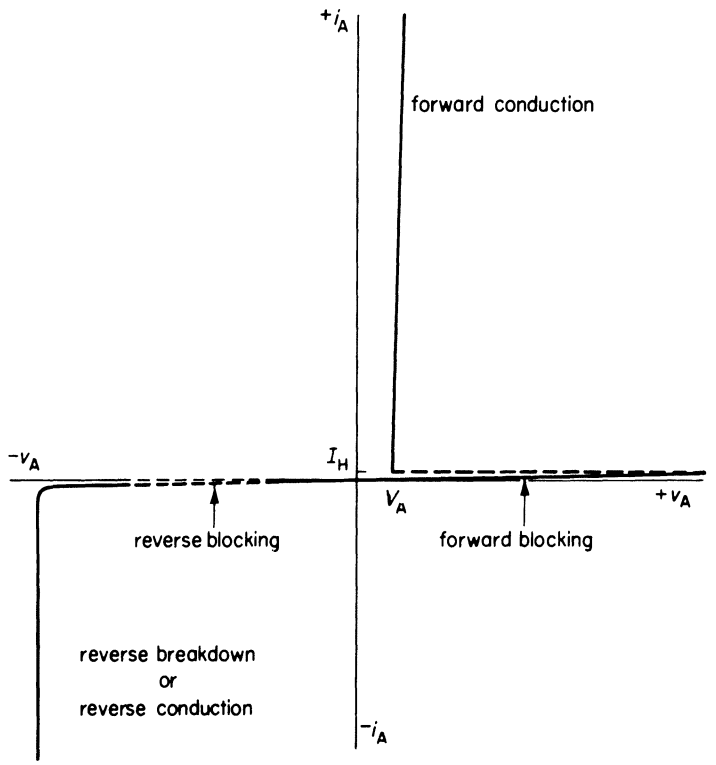


Figure 9.2 The anode characteristic of a thyristor

current. This is known as the *forward blocking mode* of operation. In this mode the application of a negative potential to the gate region reverse biases junction p_2-n_2, and the thyristor continues to block the flow of anode current.

When a positive potential is applied to the gate of the thyristor, it has the effect of injecting charge carriers into the base of TR1 in the two-transistor equivalent circuit (see figure 9.1c). This causes TR1 to turn ON, which effectively connects region n_1 to the cathode of the thyristor. Since the collector current of TR1 also flows in the base of TR2, it causes the latter transistor to turn ON also. Hence the resistance between the anode and the cathode falls to a very low value; the net result is a rapid rise in the anode current together with a reduction in the anode voltage. That is, the thyristor has been triggered into its *forward conducting mode* (see figure 9.2). In this mode, the voltage, V_A, between the anode and cathode is approximately equal to the sum of $v_{BE(sat)}$ and $v_{CE(sat)}$ for silicon transistors; for normal values of load current the value of V_A lies in the range 0.7 to 1.2 V, but under overload conditions it may rise to 2 to 3 V.

Readers will also note that the collector current flowing in TR2 also flows into the base of TR1; thus each transistor supplies base current to the other, and the

two transistors form a 'self-latching' pair. This is known as *regenerative latching*, or simply as *regeneration*. The time taken for this to take place after the application of a gate signal is only a few microseconds; hence an impulsive signal of a few microseconds' duration is all that is necessary to trigger the thyristor into conduction.

A simple analysis of the turn-on mechanism of the thyristor, known as the *alpha analysis*, is given below. For TR1 in figure 9.1c

$$\alpha_1 = \textit{modulus} \text{ of the common-base current gain of TR1}$$

$$= |\, h_{FB1} \,|$$

and for TR2

$$\alpha_2 = |\, h_{FB2} \,|$$

Now

$$I_{c1} = \alpha_1 I$$

and

$$I_{c2} = \alpha_2 I$$

where I is the current flowing through the thyristor. Also

$$I = I_{c1} + I_{c2} + I_{CO} = \alpha_1 I + \alpha_2 I + I_{CO}$$

where I_{CO} is the leakage current, hence

$$I = \frac{I_{CO}}{1 - (\alpha_1 + \alpha_2)}$$

For the thyristor to have two operating states, the value of the sum $(\alpha_1 + \alpha_2)$ must either have a value that is less than unity, or a value that is equal to unity. If the sum is less than unity, then the value of I is very small, and the device is in its OFF or blocking state. If the sum is equal to unity, then the value of I is $I_{CO}/0 = \infty$; that is, the thyristor is ON or is conducting. Quite clearly the latter value of current cannot be allowed, otherwise the thyristor would be destroyed; in practice the value of the load current is limited by the impedance of the load circuit.

The foregoing implies that the value of each of the αs must, in the OFF state, be less than 0.5. From this standpoint silicon is the most suitable material for the manufacture of thyristors since, at normal operating temperatures, the αs obtainable at low current are both stable and less than 0.5. Germanium transistors have αs that are too high, and result in unstable triggering characteristics, while gallium arsenide provides αs that are too low at normal temperatures.

9.2 Thyristor Turn-on Methods

All thyristor turn-on methods cause the sum of the αs of the equivalent transistors to increase in value. This, in turn, increases the current flowing through the device,

which further increases the αs until their sum is unity, when the thyristor turns ON. The principal mechanisms involved are given below.

Applied voltage turn-on

Applying a high voltage between the anode and the cathode in the 'forward' direction, increases the value of the leakage current and results in an avalanche effect. The net increase in current ultimately causes the αs to give rise to a value when the transistors turn ON.

Light-activated thyristors

If a translucent 'window' is left in the encapsulation so that light can fall on to the gate region, the light energy causes mobile charge carriers to be generated in the gate region. The carriers are equivalent to gate current, and cause the thyristor to trigger into its conducting mode.

dv/dt *turn-on*

If a voltage is suddenly applied to the anode of a thyristor that causes it to be forward biased (the gate signal meanwhile being zero), it establishes a depletion region at the junction p_2-n_1 of the thyristor in figure 9.1a. The resulting displacement of charge carriers from this region is equivalent to the flow of charging current in an equivalent capacitor, C, connected between the anode and the gate of the device (see figure 9.3). The instantaneous value of the charging current is given by the expression $i = C\,\mathrm{d}v_A/\mathrm{d}t$, where $\mathrm{d}v_A/\mathrm{d}t$ is the rate of rise of anode voltage. If the value of i is sufficiently great, it causes the thyristor to turn ON. Triacs are particularly prone to this type of failure. One of the most important methods used by manufacturers to increase the dv/dt rating of thyristors is the *shorted-emitter* technique, and is described in section 9.4.

One method that users can adopt to reduce the value of the dv/dt applied to the thyristor is to shunt each device with a series-connected RC circuit (known as a

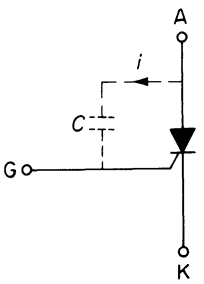


Figure 9.3 dv/dt turn-on

voltage-snubber circuit). The value of R used in this circuit is fairly small, typically 10 Ω. If R_L is the resistance of the load circuit and V_S is the value of the supply voltage, then the value of the time constant $R_L C$, where C is the capacitance of the 'snubber' circuit capacitor, is given by

$$R_L C = 0.63 V_S \Big/ \left(\frac{dv}{dt}\right)$$

where dv/dt is the rated maximum value of dv/dt for that type of thyristor.

Thermal triggering

At elevated temperatures the leakage current increases to a value that can cause thyristors to trigger into their conducting state.

9.3 Thyristor Turn-off Methods

Once a thyristor is turned ON, current flow is self-sustaining until the anode current is reduced below a value known as the *holding current,* I_H (see figure 9.2). The value of this current may be less than a milliamp in a low-current device, and up to about 50 mA in high-current devices. The *turn-off time* of the thyristor is the time taken for it to achieve its full blocking capability, and is the time taken for the depletion regions to be established at the junctions. Typical values of turn-off time lie between 5 and about 200 μs. Rapid turn-off times can be achieved by introducing recombination centres into the semiconductor material by techniques such as gold doping.

The potential applied to thyristors used in a.c. circuits reverses once during each cycle. Provided that the circuit inductance is not too large, the reversal of the supply voltage polarity ensures that the thyristor turns off when the anode voltage falls to zero or, at worst, at an early point in the negative half-cycle.

To cause thyristors used in d.c. circuits to turn OFF, it is necessary to force-commutate the anode current to zero for a length of time that is either equal to or greater than the thyristor turn-off time. Additional circuitry is required for this purpose. A description of these circuits is available in the literature – see, for example, Noel M. Morris, *Advanced Industrial Electronics* (McGraw-Hill, Maidenhead, 1974).

9.4 The 'Shorted-emitter' Construction

The shorted-emitter structure, illustrated in figure 9.4a is widely used to improve the dv/dt rating of thyristors. In this case the cathode metallisation makes contact both with the cathode region (n_2) and with the gate region (p_2), so providing an ohmic connection between the gate and the cathode. This is represented in the equivalent circuit in figure 9.4b by resistor R.

This ohmic connection provides an alternative path for the flow of the

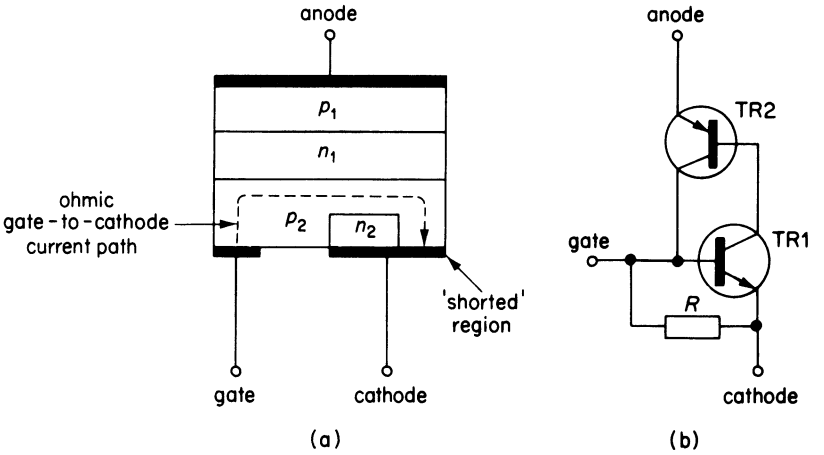


Figure 9.4 The 'shorted-emitter' structure

capacitance displacement current arising from a high value of dv/dt applied to the anode. This reduces the transient current that is available to turn TR1 ON, thereby reducing the possibility of the incidence of dv/dt failure. Typical values of dv/dt rating for thyristors lie in the range 1 to 2 kV/μs.

9.5 Turn-on Behaviour of Thyristors

When a thyristor which has a simple form of gate construction begins to conduct, current flow starts in a localised area and finally spreads to the whole cross-section of the thyristor. The lateral conduction *spreading velocity* in such a device is typically 0.1 mm/μs. Since conduction is initiated in a small area, the allowable rate of rise of current, di/dt, must be limited in value, otherwise localised overheating may cause the device to fail catastrophically; such a failure is known as di/dt *failure*. The rated value of di/dt may lie in a band between about 20 A/μs and several thousand A/μs.

One method of protecting thyristors against this type of failure is to include an inductor in the circuit. The minimum value of circuit inductance, L_{min}, can be estimated from the expression

$$L_{min} = V_S \Big/ \left(\frac{di}{dt}\right)$$

where V_S is the value of the supply voltage, and di/dt is the rated value of di/dt for the device.

The turn-on performance of thyristors can be improved in several ways. One method is to construct thyristors using large gate areas; these devices provide a larger initial conducting area than do conventional thyristors. Another method is to use some form of gate-signal amplifying thyristor, which is built into the geometry

of the main thyristor. The operation of these devices relies on the main anode current acting as a high-current gate signal. Names given to the latter technique include *accelerated cathode excitation, amplified gate construction* and *dynamic gate construction.*

9.6 The Triac or Bidirectional Thyristor

The basic structure of the triac is shown in figure 9.5 and, like the thyristor, it has a gate electrode and two main terminals, T1 and T2, through which the main current flows. It differs from the reverse blocking thyristor in that it can be triggered into conduction either when T1 is positive with respect to T2, or vice versa. Moreover, it may be triggered into conduction by a gate signal which is either positive or negative with respect to T1.

The current and voltage ratings of triacs are generally lower than those of thyristors, and the dv/dt withstand capability is less. An advantage of triacs over thyristors is that, since they may be triggered with either polarity of supply voltage, they can be used in relatively simple circuits to control current flow in a.c. systems.

Figure 9.6 shows the general shape of the static characteristic of a triac, and readers will note that it can be triggered into conduction either in the first quadrant (when T2 is positive with respect to T1), or in the third quadrant (T2 negative). The operating modes of the triac are listed in table 9.1.

The sensitivity of the triac to gate signals can be altered to some extent by the geometry of the device, those in popular usage being most sensitive in the I^+ and III^- modes, and generally least sensitive in the III^+ mode.

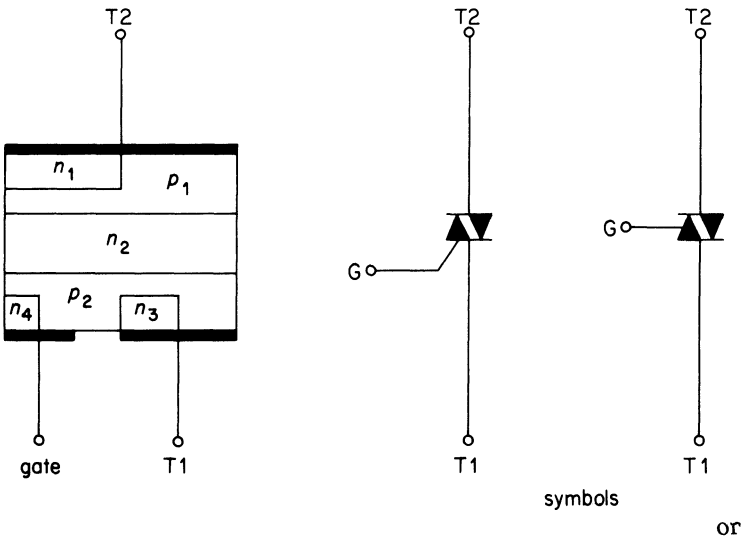


Figure 9.5 The bidirectional thyristor or triac

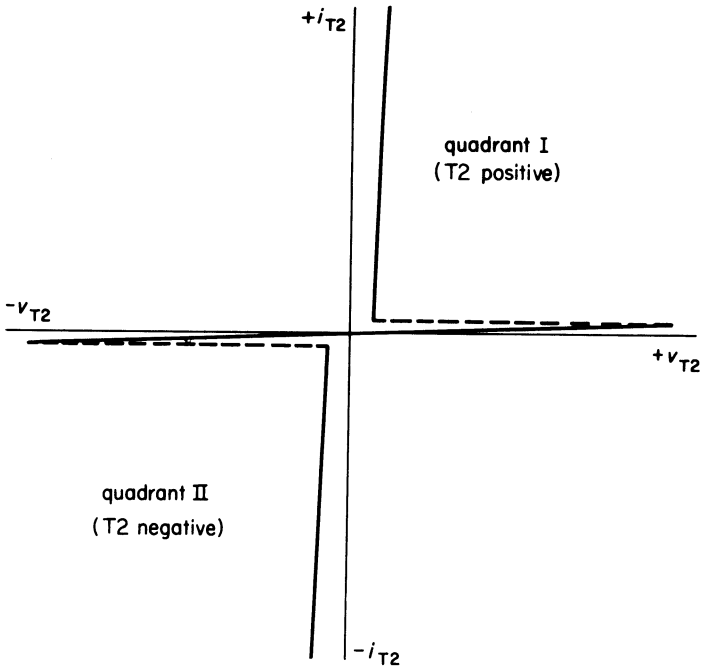


Figure 9.6 Static characteristics of the triac

Operation in the I^+ mode: T2 positive, positive gate voltage

In this mode the triac behaves as though it were a conventional thyristor, with the main current flowing through regions $p_1-n_2-p_2-n_3$ (see figure 9.5).

Operation in the I^- mode: T2 positive, negative gate voltage

The mechanism associated with this mode of operation is illustrated in figure 9.7. Diagram (a) shows the current paths involved, and diagram (b) is used to explain

Table 9.1

Mode	Potential of T2 with respect to T1	Potential of the gate with respect to T1
I^+	positive	positive
I^-	positive	negative
III^+	negative	positive
III^-	negative	negative

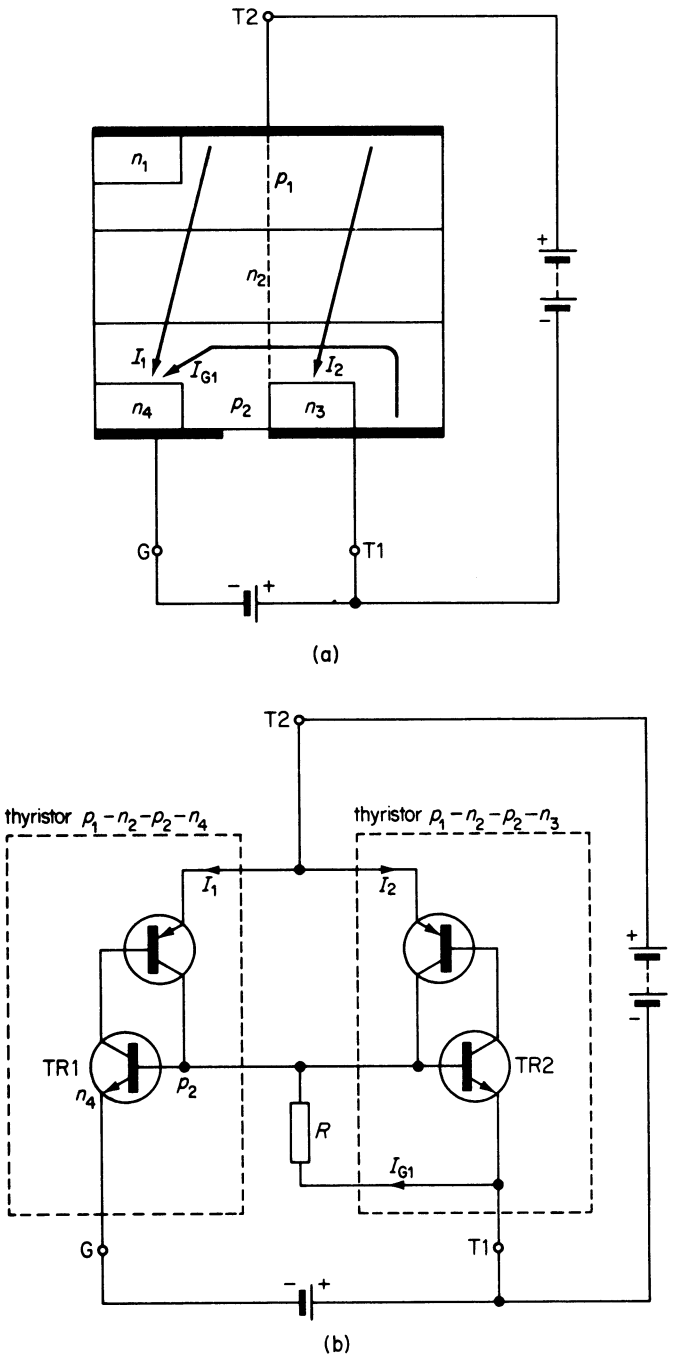


Figure 9.7 Operation of the triac in the I^- mode

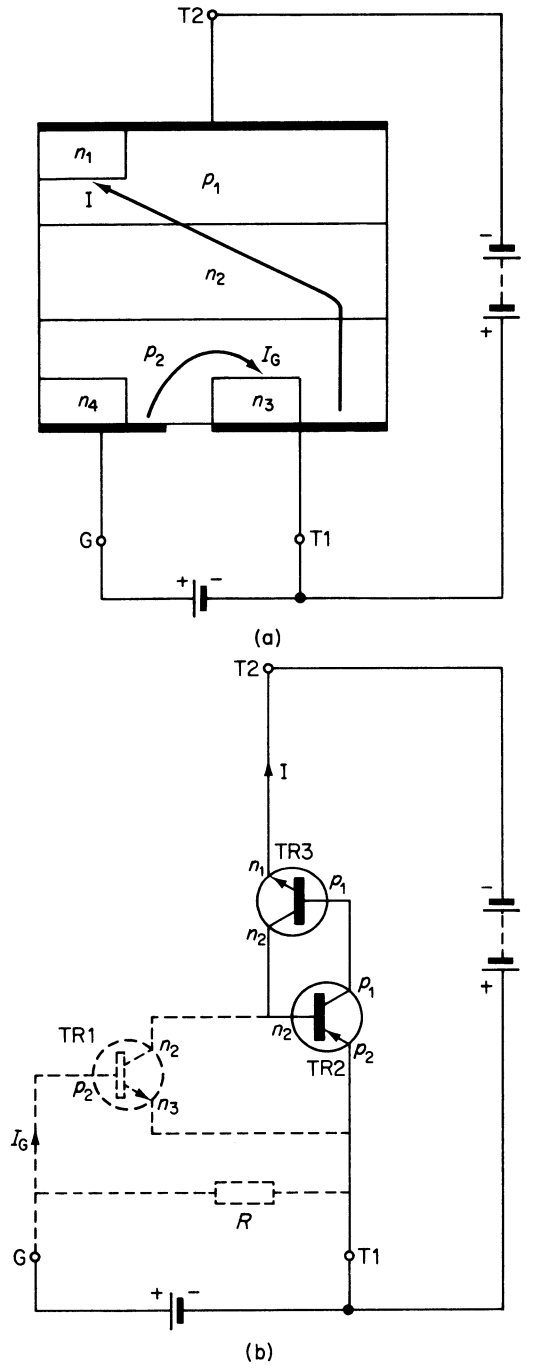


Figure 9.8 Operation of the triac in the III $^+$ mode (remote gate operation)

the triggering mechanism. The triac can be considered in this mode to consist of two thyristors, namely thyristor $p_1-n_2-p_2-n_4$ and thyristor $p_1-n_2-p_2-n_3$. Initially, with the gate negative with respect to T1, current I_{G1} flows from T1 through resistor R (see figure 9.7b) into the forward biased emitter junction p_2-n_4 of TR1. Resistor R represents the lateral resistance of the p_2 region of the triac (see also the shorted-emitter construction in section 9.4). Current I_{G1} triggers the left-hand thyristor ($p_1-n_2-p_2-n_4$) into conduction, and current I_1 flows in it.

As a result of the above sequence of events, part of current I_1 also flows into the base of TR2, which causes the main structure ($p_1-n_2-p_2-n_3$) to turn ON. When fully conducting, the main current, I_2, flows in the main structure as illustrated in figure 9.7.

III$^+$ operating mode: T2 negative, positive gate voltage

In this mode of operation the main current follows the path $p_2-n_2-p_1-n_1$ (see figure 9.8a). Triggering is by means of a *remote-gate method*, a term that arises from the fact that the gate current is 'remote' from the main structure (see below). The operation is best described in terms of figure 9.8b, in which the triac is divided into three equivalent transistors; in this diagram, resistor R once more represents the lateral resistance of the p_2 region.

The application of a positive potential to the gate of the triac causes a current to be injected into the base of TR1, whose collector current flows in the base of TR2. The latter current triggers the main structure (TR2 and TR3) into conduction.

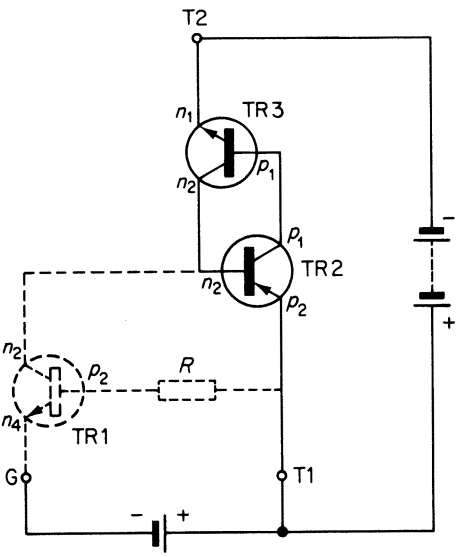


Figure 9.9 Triac operating in the III$^-$ mode

III operating mode: T2 negative, negative gate voltage

The thyristor is triggered once more by a remote-gate method, but this time using a different transistor structure for TR1 (see figure 9.9) than is the case in figure 9.8. Here TR1 is formed from the $n_2-p_2-n_4$ regions. Applying a negative potential to the n_4 region of TR1 (the triac gate) turns this transistor ON, causing electrons to be injected into the n_2 region of the triac. These charge carriers cause the main structure (TR2–TR3) to be triggered into conduction.

9.7 The Diac or Bidirectional Breakdown Diode

The diac is a two-terminal device exhibiting a characteristic generally similar to that in figure 9.10. Its structure is essentially that of a transistor, with the equivalent emitter and collector connections being used. The device blocks the flow of current for both forward and reverse voltages up to its *breakover voltage*, V_{BO}. When the current exceeds the breakover value, I_{BO}, the device exhibits a negative resistance region on its characteristic.

The diac is used in pulse-generator circuits for thyristors and triacs; it is usually employed as a capacitor-discharge device in a relaxation oscillator. A basic form of triac control circuit incorporating a pulse generator of this kind is illustrated in figure 9.11. In the circuit shown, the a.c. supply is connected to the timing circuit comprising resistor R and capacitor C; the supply also energises the load circuit and triac. The p.d. across the capacitor builds up at a rate which depends on the time

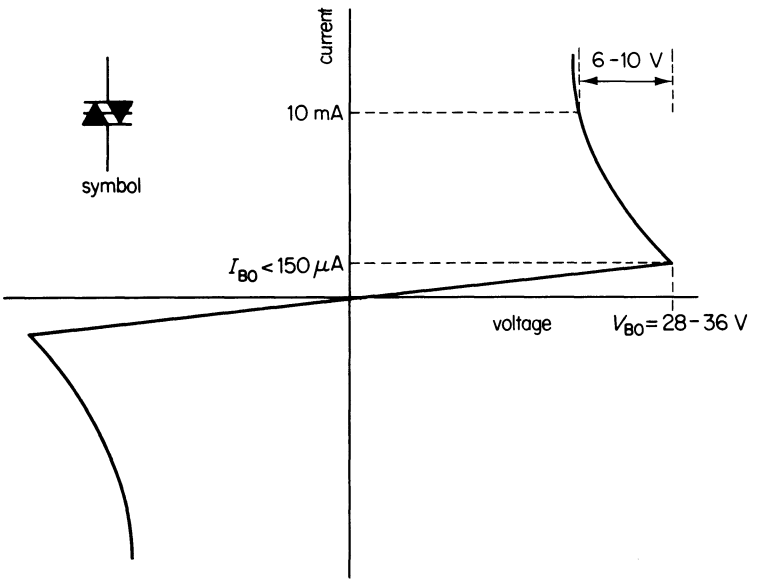


Figure 9.10 Typical diac characteristic

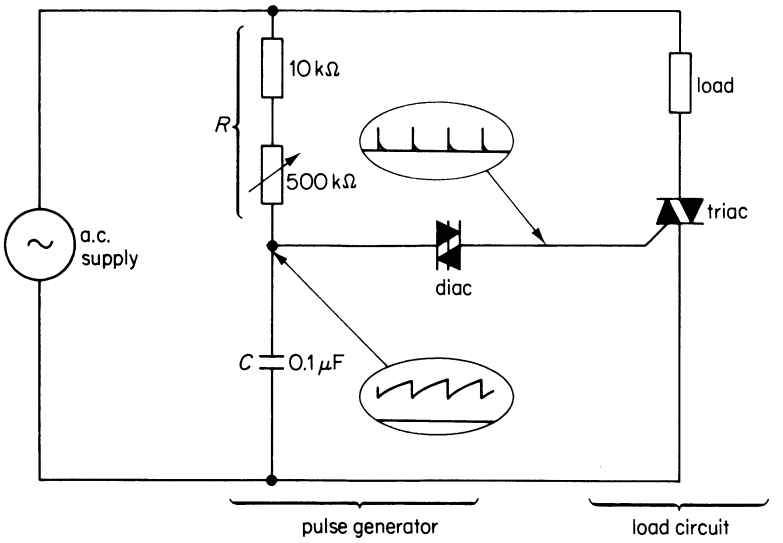


Figure 9.11 A simple pulse-generator circuit

constant, RC, of the timing circuit, and when its value reaches the breakover voltage of the diac, this device is triggered into conduction. The capacitor is partially discharged into the gate of the triac, which is triggered into its conducting state. The waveforms shown in the insets in figure 9.11 refer to the positive half-cycles of the supply voltage waveform. In the negative half-cycles of the supply waveform, a sawtooth of negative polarity appears across the capacitor, and negative-going pulses are applied to the gate of the triac. The triac is then triggered in its III$^-$ mode.

9.8 The Silicon Controlled Switch (SCS)

The silicon controlled switch is a low-power thyristor having all four regions brought out to terminations. A circuit symbol used for the SCS is shown in

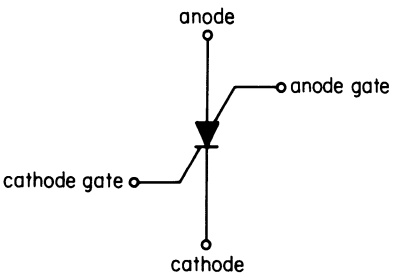


Figure 9.12 The silicon controlled switch

figure 9.12. The additional gate connection increases the versatility of the device when compared with other four-layer devices, allowing it to be used in a wide range of applications including numerical indicator tube drives and in switching applications (including its use as a memory element).

Light-sensitive SCSs are manufactured which are capable of switching a current of several amperes, and having turn-on and turn-off times in the range 50 to 100 μs.

9.9 The Programmable Unijunction Transistor (PUT)

The PUT is, in fact, not a UJT but a version of the SCS whose triggering voltage (the peak-point voltage of the PUT) is controlled by the voltage applied to the anode gate of the PUT. In this way it is possible to control the intrinsic stand-off ratio of the PUT (see also section 3.11).

10 Optoelectronics

Optoelectronic elements include devices that are sensitive to electromagnetic radiation (light) in the visible, infrared and ultraviolet regions, and also devices that emit electromagnetic radiations in those regions. The section of the electromagnetic spectrum covered by these devices ranges from wavelengths of 0.001 to about 1000 μm; ultraviolet radiations cover wavelengths in the range 0.001 to 0.35 μm, visible light covers wavelengths of 0.35 to 0.7 μm, and infrared from 0.7 to 1000 μm. This information was illustrated in diagrammatic form in figure 2.7. Frequently the wavelengths of electromagnetic radiations are quoted in Ångstrom units (symbol Å), where 1 Å = 10^{-10} m.

The effect of illumination on semiconductors was discussed in chapter 2, and a summary is presented here. When radiant energy is absorbed by the atomic structure, it causes the spontaneous generation of electron–hole pairs that increase the conductivity of the material. The minimum value of radiant energy that can cause an electron to be excited from the valence energy band to the conduction band (intrinsic excitation) is equal to the forbidden energy gap, W_G, of the semiconductor. The critical wavelength, λ_c, of a photon of energy that causes this type of excitation is given by the expression

$$\lambda_c = \frac{1.24}{W_G} \mu m \qquad (10.1)$$

10.1 Photodiodes

A photodiode is a *p–n* junction which is reverse biased in normal operation, and whose junction is exposed to the radiant energy source through a 'window' in its encapsulation. The physical size of an encapsulated photodiode is small, a typical size being 5 mm long x 3 mm diameter (0.2 x 0.125 in.). Both germanium and silicon have been used in photodiodes, but germanium has largely been supplanted by silicon since the former material has a higher value of *'dark' current*, I_O – this is the leakage current that flows when the incident radiation is zero. The dark current associated with a germanium device may be as high as 10 mA, and in a silicon device may be as low as 20 nA.

An increase in illumination results in an increase in the number of electron–hole pairs generated in the semiconductor, and the minority charge carriers which are

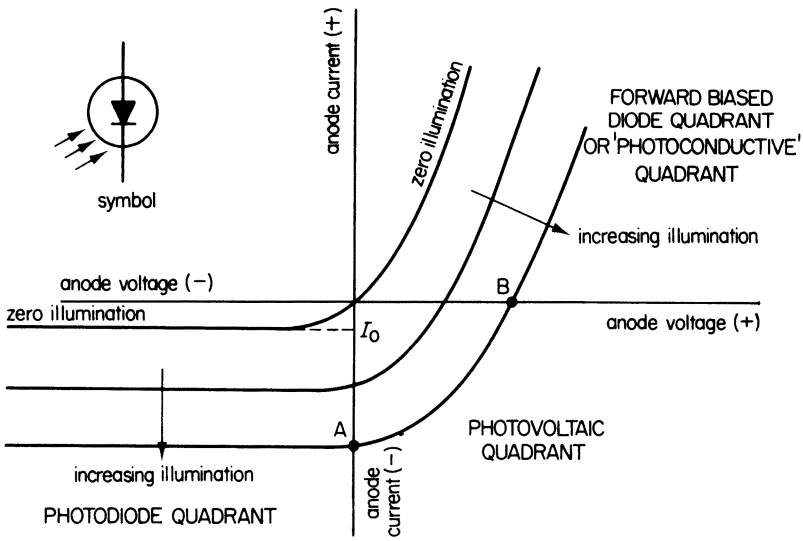


Figure 10.1 Photodiode and photovoltaic cell characteristics

generated close to the junction are swept across it. The carriers that cross the junction in this way constitute flow of *photocurrent*, I_P, through the diode. When illuminated, the diode current under reverse biased conditions becomes $I_O + I_P$.

The electrical characteristics of the photodiode lie in the third quadrant of figure 10.1; the sensitivity of photodiodes range from about 10 mA/lm to about 50 mA/lm, and they are responsive to illumination variations up to a frequency of 10 MHz. The characteristics of these devices can easily be modified during manufacture by altering the resistivity of the semiconductor material; for example, a good high-frequency response can be obtained by reducing the junction capacitance by using the *p–i–n* construction (see also section 3.8). In order to obtain maximum sensitivity to illumination, it is necessary that the incident radiation be focused as nearly as possible on the junction, since the minority charge carriers which are produced at some distance from the junction are likely to vanish due to recombination effects before they reach the junction.

Due to their high operating speed, photodiodes are used in high-speed tape readers, in character recognition equipment, and in laser data links. Some photodiodes have a very low noise current, and are suitable for the detection of low-level optical signals, such as star tracking and photometry applications.

Unlike many other opteolectronic devices, the value of the photocurrent varies in a roughly linear fashion with intensity, which is a useful feature. The principal limitation of photodiodes is the low value of power level involved.

Photodiodes can also be used in a 'photoconductive' mode by applying a forward bias. In this case the device operates in the first quadrant of figure 10.1.

10.2 Photovoltaic Cells or Solar Cells

The photovoltaic cell differs from other types of optoelectronic device in that there is a direct conversion from illumination to electrical energy within it.

Solar cells are large-area silicon devices whose peak spectral response is arranged to match that of the sun's radiation. They consist of a thin layer (1 μm or 0.04×10^{-3} in.) of *n*-type material on a *p*-region. At high illumination levels, an individual silicon cell may generate an open-circuit voltage of up to 0.5 V. Operational efficiencies of about 15 per cent are obtainable, and cells can be connected in series to give an increased output voltage, or in parallel to give a higher current.

Looking now at the characteristics of the photovoltaic cell, shown in the fourth quadrant of figure 10.1, at point A on the curves the anode voltage is zero and the current flow is due entirely to leakage current. This corresponds to the condition when the cell is short-circuited. The application of a small forward bias to the junction reduces the height of the potential barrier, so that a few majority charge carriers can cross the junction. The net result is a reduction in the reverse current. A progressive increase of forward bias gives a progressive reduction in reverse current until, at point B on the characteristic, the anode current is zero. This condition is, in fact, obtained when the cell is open-circuited. Thus the transition from point A to point B on the characteristic curve, is brought about merely by increasing the value of the external load resistance connected to the cell from zero to infinity.

Applications of photovoltaic cells include exposure meters, punched-tape and punched-card readers and aerospace applications.

10.3 Phototransistors

A bipolar phototransistor (also known as a *photoduodiode*) is a three-layer device having its base region exposed to illumination. Both *n–p–n* and *p–n–p* devices are manufactured. They are usually connected in the common-emitter configuration, and may either have the base connection open-circuited or be connected through a resistance to the common line. In some cases it is possible to forward bias the transistor.

A basic circuit incorporating an *n–p–n* phototransistor is shown in figure 10.2. In the absence of illumination, the collector current, I_C, is

$$I_C = (h_{FE} + 1)I_{CBO}$$

When illuminated, electron–hole pairs are generated in the base region and, by transistor action, the new value of collector current is

$$I_C = (h_{FE} + 1)(I_{CBO} + I_P)$$

where I_P is the photocurrent component of the base current.

A typical collector characteristic of a phototransistor is shown in figure 10.3. The characteristics show how the collector current alters with change in

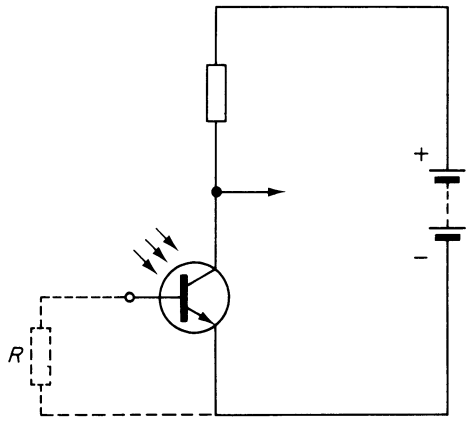


Figure 10.2 An *n–p–n* phototransistor circuit

illumination; the illumination is expressed in units of lux (1 lx = 1 lm/m^2). Some phototransistors have a lens built into the encapsulation to concentrate the illumination on to the base region, giving increased sensitivity.

By manufacturing MOSFETs with transparent gate regions, light energy can be used as a means of releasing charge carriers in the substrate. This is the simple principle of the *photo-FET*.

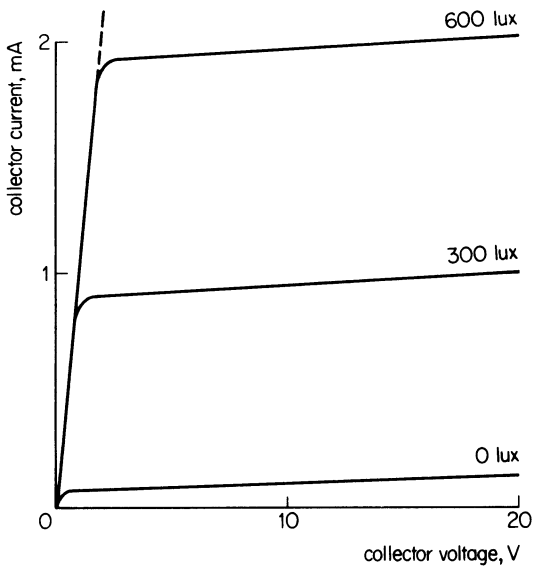


Figure 10.3 Output characteristics of a phototransistor

10.4 Photothyristors

If light is allowed to pass through a 'window' in the canister containing a thyristor, the incident energy falling on the gate region releases charge carriers that trigger the thyristor into conduction. The gate irradiance necessary to trigger a 1.6 A, 200 V thyristor lies typically in the range 0.5 to 10 mW/cm^2. *Note*: Irradiance is the radiant flux density on a surface. Light-activated silicon controlled switches (see also section 9.8) are also manufactured.

10.5 The Light-emitting Diode (LED) or Electroluminescent Diode

Just as light energy can cause an electron to be excited into the conduction band then, when recombination occurs, energy is released. The wavelength of the radiation resulting from a recombination depends on the band gap through which the charge carrier falls. In *direct-gap* semiconductors, such as those based on gallium arsenide (GaAs), direct transitions take place between the conduction band and the valence band. In *indirect-gap* materials, such as silicon and germanium, transitions take place via trapping levels (see section 1.11). Direct-gap semiconductors are of interest to us here, since the radiation emanating from these is in the visible and the infrared regions.

A section through a LED is illustrated in figure 10.4. In order to obtain the highest possible radiation efficiency, the p–n junction must be very close to the surface of the diode, and the anode must be constructed so as to provide very little

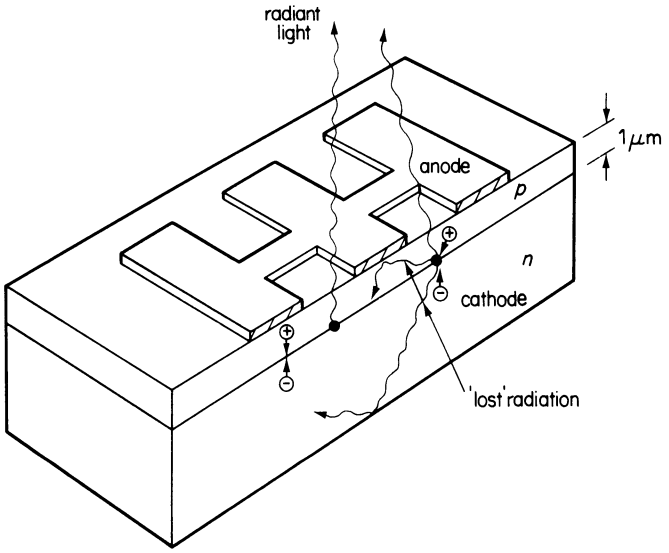


Figure 10.4 A section through one form of LED

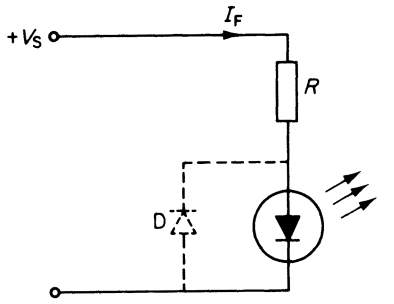


Figure 10.5 A LED circuit

impedance to the radiated light. Even so, only a fraction of the light generated is actually able to leave the surface of the LED. This is due to two factors, namely

(1) the transparency (or lack of it) of the LED material to the emitted wavelength
(2) the high refractive index of the LED material.

The former causes absorption of light. The latter causes light to be 'bent' away from a straight path, and in a material such as gallium arsenide phosphide (GaAsP) any light striking the surface at an angle greater than about 17° is reflected back into the chip. Also some of the light which is perpendicular to the surface is reflected back into the chip. The radiating efficiency is improved by the combined use of anti-reflection coatings and lenses. In addition, colour filters provide improved contrast.

The light output from these sources varies as (current)^n, where n has a value in the range 1.2 to 1.5. The display brightness can therefore be increased by applying current pulses of high value to the diode, yielding an apparently bright display, yet requiring a relatively low mean value of current. Popular display colours are red (GaP or GaAsP), orange (GaAsP), yellow (GaP or GaAsP) and green (GaP).

A popular circuit used to drive a LED is shown in figure 10.5. The reason for diode D is discussed later. Resistance R is used for current-limiting purposes, and its value is computed from the relationship

$$R = \frac{V_S - V_F}{I_F}$$

where V_S and V_F are the values of the supply voltage and the forward voltage across the LED, respectively, and I_F is the forward current flowing through the diode. The values of V_F and I_F depend on the type of diode (that is, on the colour of the emitted light), and are typically 1.6 to 2.5 V and 5 to 25 mA, respectively, for a red colour, and 2 to 3.5 V and 10 to 40 mA, respectively, for orange, green and yellow colours.

A range of LEDs known as *resistor LEDs* include a fixed value of resistance

inside the encapsulation. When operated at the rated voltage, the external current-limiting resistor is not required.

The reverse breakdown voltage of LEDs is typically 3 to 11 V and, when used with an a.c. supply, should be protected against reverse breakdown. A simple method of protecting against this eventuality is to shunt the LED with a conventional diode in the manner shown in figure 10.5. This diode conducts when the reverse voltage across the diode exceeds about 0.4 V. Alternatively, a conventional p–n junction diode can be connected in series with the LED.

10.6 LED Displays

The most popular type of LED *numerical display* is the *seven-segment* type illustrated in figure10.6a. Each of the segments, a to g, may consist either of a group of photodiodes or of a single photodiode in association with an optical magnifying system. By illuminating groups of segments, as shown in figure 10.6b, it is possible to generate the decimal numbers 0 to 9. Certain alphabetical characters can also be generated by the seven-segment display. For example, the letter L in figure 10.6b can be used with a pocket calculator to display the fact that the

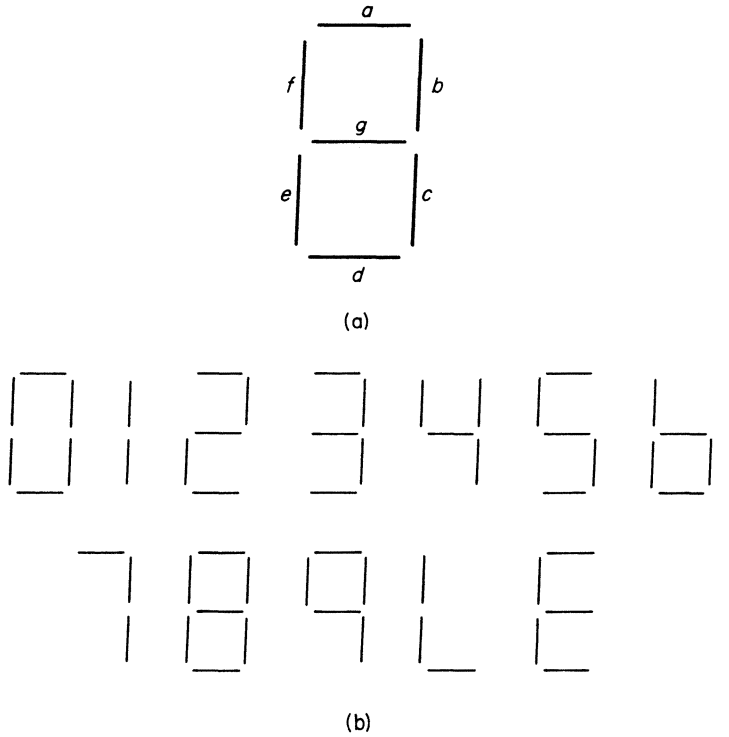


Figure 10.6 A seven-segment LED display

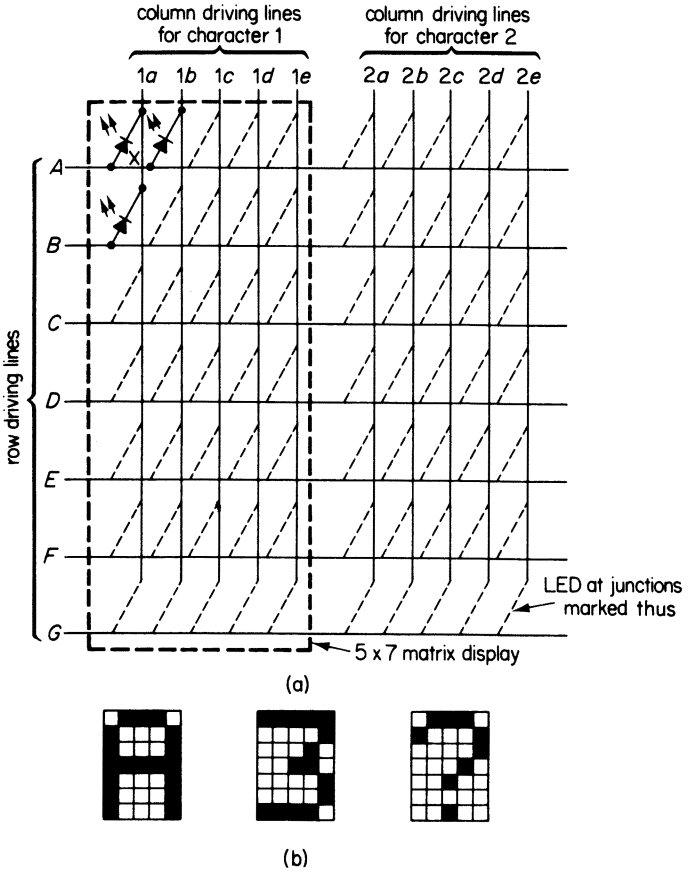


Figure 10.7 (a) Schematic diagram of a 5 x 7 dot matrix display and (b) examples of characters that may be displayed

battery voltage is low. The letter E is sometimes used to indicate that the value computed by a calculator is in excess of the storage capacity of the machine. A mathematical negative sign is generated if segment *g* is illuminated alone.

Another popular type of display is the *5 x 7 dot matrix display*, illustrated in figure 10.7a. Each character contains 35 LEDs in a 5 x 7 array, and each diode in the character is $X-Y$ addressable. When row wire *A* in figure 10.7a is connected to a positive potential, and column wire 1a is connected to zero potential, then diode X radiates light. To display a particular character, the rows and columns in the matrix are addressed sequentially; this procees is known as *multiplexing*, or as *scanning* or *strobing*. To prevent visual 'flicker', each complete character is strobed at about 100 times per second.

The 5 x 7 matrix display can not only produce the full range of alphabetical and

numerical characters (and for this reason it is known as an *alphanumeric display*), but also generate a wide range of other characters. Three typical characters are displayed in figure 10.7b.

10.7 Optically Coupled Isolators

An optically coupled isolator is a photoemissive device which is optically coupled to a photoelectric device, both contained in the same encapsulation. These devices

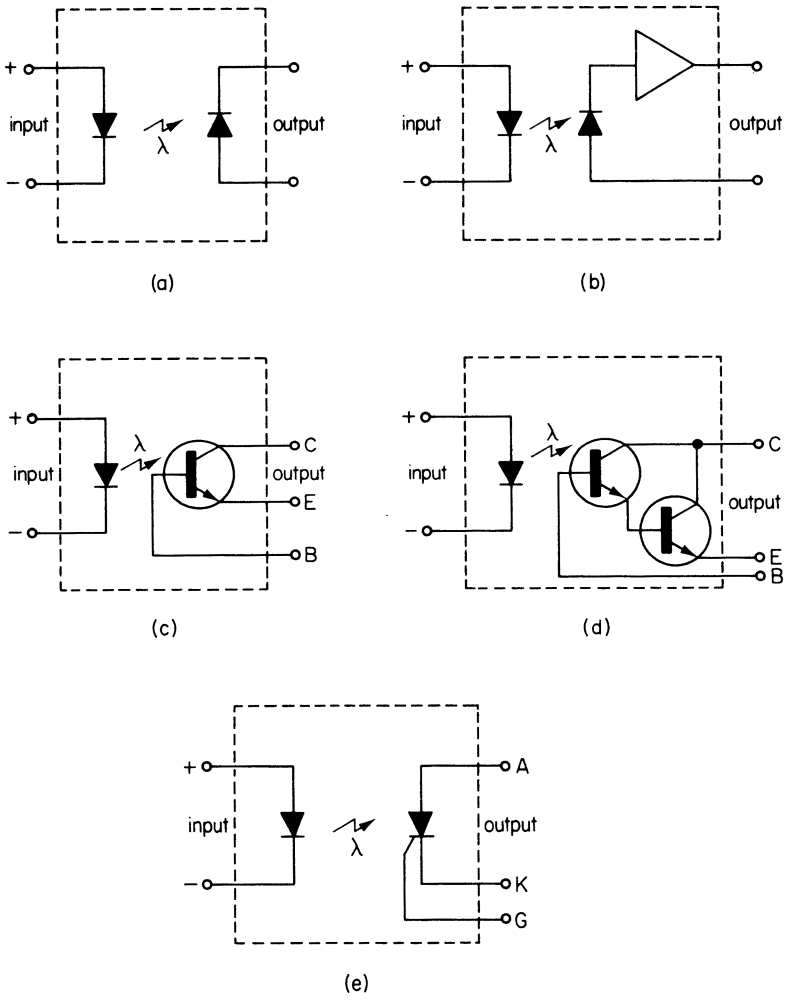


Figure 10.8 A selection of optically coupled isolators: (a) LED and photodiode, (b) LED and a photodiode/amplifier combination, (c) LED and phototransistor, (d) LED and photo-Darlington pair and (e) LED and photothyristor

provide electrical isolation between the two circuits, and typically provide a bandwidth from d.c. to about 100 kHz. Early types contained either a neon lamp or a tungsten filament lamp together with a photoconductive cell. Modern semiconductor optically coupled isolators (also known as *optical couplers* and as *photon couplers*) usually have an LED in combination with either a photodiode or a phototransistor. Some devices also contain an amplifier within the encapsulation. Darlington connected transistor pairs and photothyristors may be used as output devices. Basic circuits of a selection of devices are illustrated in figure 10.8.

The feature of complete electrical isolation between the input and output circuits of the optically coupled isolator, allows a signal from, say, a low-voltage circuit to trigger a high-voltage thyristor via the output from the optical coupler. This feature is also advantageous in other fields such as biomedical engineering.

10.8 Semiconductor Lasers

The *laser* (*L*ight *A*mplification by *S*timulated *E*mission of *R*adiation) is a class of light emitter that depends for its operation on atoms being stimulated by external means so that they emit light. Semiconductor lasers are stimulated by means of current injection, and are frequently known as *injection lasers*. The injection laser is a development of the GaAs diode, in which a reflecting cavity is formed by polishing the ends of the crystal in a plane perpendicular to the p–n junction. These surfaces reflect some of the light back to the junction, which stimulates further emission, and a photon avalanche builds up in the junction.

10.9 Liquid Crystal Displays (LCDs)

Liquid crystals are not semiconductor devices, but since they are used as display devices in conjunction with electronic systems, they are briefly discussed here. Unlike the semiconductor lamps described above, LCDs do not radiate illumination, but either reflect incident illumination or transmit it.

Liquid crystals are organic fluids, the type of interest here being known as *nematic liquid crystals*. The name nematic comes from the Greek, meaning *thread-like*, which describes the thread-like nature of the molecules of the material. The nematic molecules are shaped like minute cigars, and are aligned by their electrical molecular coupling. The liquid crystal is sealed between glass sheets, each having a transparent conducting surface (figure 10.9a); a typical spacing between the back and front plates is 10 μm. The application of a relatively low value of voltage between the conducting coatings causes the crystal molecules to rearrange their orientation to produce the display.

There are two types of LCD in popular usage, namely *dynamic scattering displays* and *field-effect displays* (or *twisted nematic displays*). In the former, the application of a voltage to the device causes the crystals to become very efficient scatterers of white light. In the absence of an applied voltage the display is transparent. Consequently the energised segment of the display is white against a

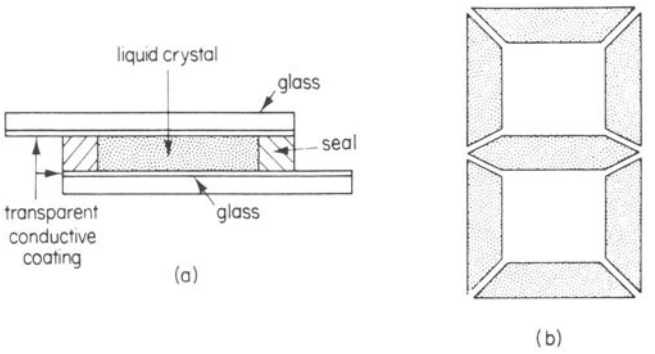


Figure 10.9 (a) A section through a LCD and (b) a typical seven-segment display

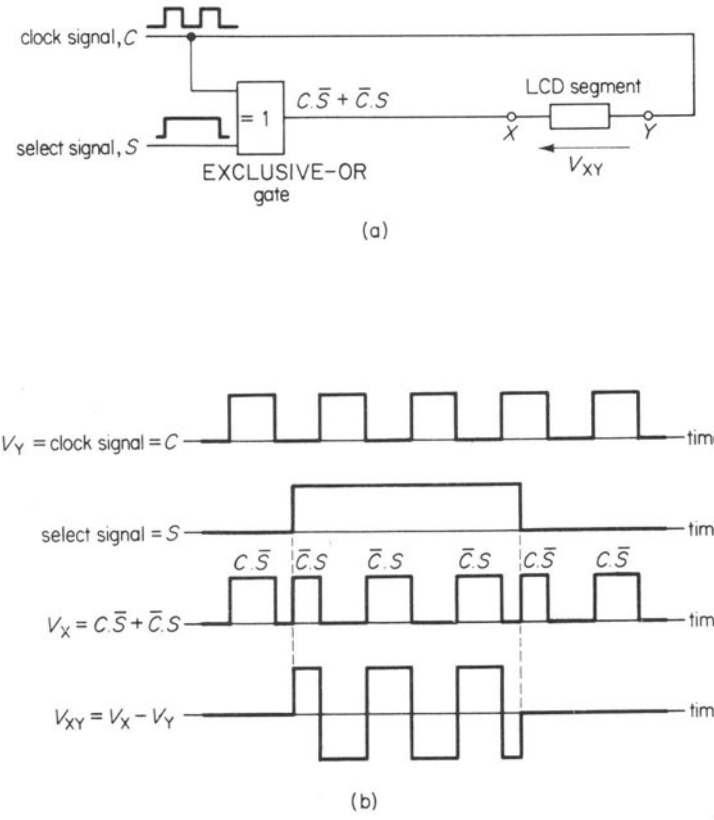


Figure 10.10 (a) One form of LCD segment driver circuit and (b) typical waveforms

transparent background; a high value of incident illumination produces a bright display. Any shape of display is possible with these devices, the seven-segment version in figure 10.9b being very popular.

In unexcited field-effect LCDs the molecular structure of the material causes the plane of polarisation of the incident light to be twisted through 90°. Polarising surfaces are used with these devices and, in the unexcited state, these surfaces result in light being transmitted completely through the device. A mirror may be placed below the device so that the light is returned, and it arrives at the upper polariser in the correct plane to pass through it. In this mode the display is transparent. The application of a potential across the LCD causes the molecular structure to change so that the plane of the polarisation of the light is not twisted, and the light is completely absorbed by the lower polariser. This results in the energised segments appearing as a dark character, which contrasts with the brighter surroundings.

When energised by a d.c. source, both types of LCD suffer from electrolytic dissociation problems, which shorten the life of the display. To overcome this the LCDs are energised by an a.c. signal at a frequency in the range 30 to 100 Hz. A circuit frequently used to generate the a.c. signal is illustrated in figure 10.10a. The EXCLUSIVE-OR gate generates the logic function $C \,.\, \bar{S} + \bar{C} \,.\, S$, which is applied to terminal X of the LCD element; clock signal C is applied to terminal Y. Typical waveforms in the circuit are illustrated in figure 10.10b. When a logic '0' signal is applied to the 'select' line of the EXCLUSIVE-OR gate, the waveforms applied to terminals X and Y are equal in magnitude, with the result that the p.d. across the LCD element is zero. When line S is activated by a logic '1' signal then, when the clock signal is '0', the logic level at X is '1' and that at Y is '0'. When the clock signal is logic '1', the logic levels at X and Y reverse. This causes an alternating voltage to be developed across the LCD element when the 'select' line is activated.

Index