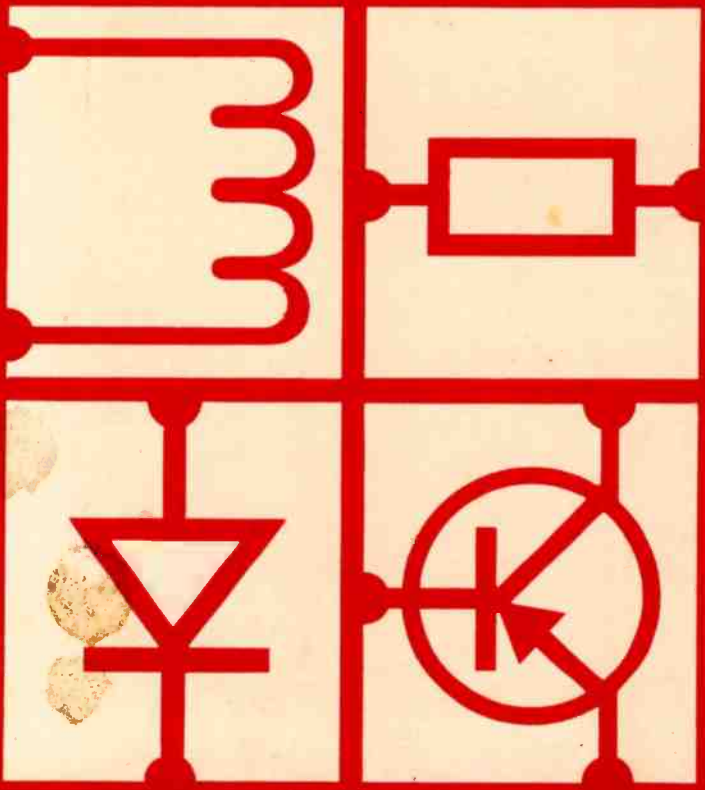
Elements of Electronics

3

Semiconductor Technology



F.A. WILSON

ELEMENTS OF ELECTRONICS
BOOK 3

Semiconductor Technology

ALSO BY THE SAME AUTHOR

BP53 PRACTICAL ELECTRONIC CALCULATIONS AND FORMULAE
BP54 YOUR ELECTRONIC CALCULATOR AND YOUR MONEY
BP62 ELEMENTS OF ELECTRONICS – BOOK 1
BP63 ELEMENTS OF ELECTRONICS – BOOK 2

ELEMENTS OF ELECTRONICS
BOOK 3

Semiconductor Technology

by
F. A. WILSON
C.G.I.A., C.Eng., F.I.E.E., F.I.E.R.E., M.B.I.M.

BERNARD BABANI (publishing) LTD
THE GRAMPIANS
SHEPHERDS BUSH ROAD
LONDON W6 7NF
ENGLAND

First Published – November 1979

British Library Cataloguing in Publication Data
Wilson, F A
Elements of electronics
Book 3
1. Electronic apparatus and appliances
I. Title
621.381 TK7870

ISBN 0 900162 84 8

Printed and Manufactured in Great Britain by C. Nicholls & Co. Ltd.

PREFACE

Knowledge advances by steps and not by leaps.
Macaulay

This is the third book in a series of three written to provide a complete but inexpensive basic electronic theory course especially for readers with very little or no experience of mathematics. The aim of the series is not so much to skim the surface of all electronics to give a rough idea of what it is all about but rather to select the most important features and principles and teach these thoroughly so that the reader *understands* and feels better for it. It is through an intimacy with the basic principles that more advanced ones fall into place with greater ease. It must be emphasized that the series caters mainly for the serious student although the not so serious will gain plenty from it. The tools for the job are also developed and used, for example, graphs have much more to tell those who understand their shapes and slopes than those who have seen but not worked with them. Examples are given where appropriate and the steps explained as necessary in arriving at the answer. The reader may simply accept the calculations shown or may prefer to work through them, possession of an inexpensive book of logarithmic tables only is assumed. However, as one progresses through the book it will soon become obvious that the availability of a calculator, especially a scientific one, is a great advantage.

This, Book 3, covers the elements of semiconductor technology, from the physics, through an understanding of diodes and transistors from their characteristic curves, to rectifiers,

amplifiers, oscillators, switching and integrated circuits. Some elementary ideas of computer technology are also here so that the reader does not feel left out of this fast expanding branch of electronic engineering as so many people are.

For those whose interest in mathematics has grown from their earlier studies, some more advanced mathematical techniques which are especially useful in electronics (equations to graphs etc.) have been deliberately introduced.

Although normally rigorously avoided in text books, some examples of repetition will be found. This is intentional, many small important points need to be encountered more than once to drive them home.

The title of Book 1 is "The Simple Electronic Circuit and Components", of Book 2, "Alternating Current Theory", the two together starting from scratch, that is, the reader should be moderately competent in arithmetic only, all else being taught as the subject progresses. So, for the reader who is already conversant with the principles and mathematics proper to the above two titles, Book 3 stands on its own, there is no need to have studied Books 1 and 2.

Some of the studies covering ancillary material to the main theme are contained within Appendices (which may be indicated by a bracketed, raised reference). The reader may perhaps wish to glance over at least the titles and sub-titles of these first for an indication of what they contain.

Other books are expected to follow based on this basic series covering the elements of more specialized subjects so that readers who wish to expand their electronic wisdom may do so at moderate cost.

F.A. WILSON, C.G.I.A., C.Eng., F.I.E.E., F.I.E.R.E., M.B.I.M.

CONTENTS

	Page
1. THE PHYSICS OF SEMICONDUCTORS	1
1.1 Atomic Structure	1
1.1.1 Composition of the atom	1
1.1.2 Atomic number	2
1.1.3 Shells	2
1.1.4 Valence electrons	3
1.1.5 Positive and negative charges	5
1.1.6 Semiconductor crystals	5
1.1.7 Electrons and holes	7
1.2 Adding Conduction	8
1.2.1 n-Type materials	8
1.2.2 p-Type materials	9
1.3 The p-n Junction	10
1.4 Semiconductors at Work	12
1.4.1 Rectification	12
1.4.2 Amplification	13
1.4.3 Oscillation	13
1.4.4 Switching	14
1.5 Diodes	14
1.5.1 Reverse bias	15
1.5.2 Forward bias	17
1.6 Transistors	18
1.6.1 Construction	21
1.6.2 Shapes and sizes	25
2. SEMICONDUCTOR CHARACTERISTICS	29
2.1 Diodes	29
2.1.1 Measurement of working characteristics	29
2.1.2 Rectifier diodes	30
2.1.3 Voltage regulator diodes	33
2.1.4 The static load line	35
2.2 Transistors	40
2.2.1 Basic circuit configuration	40
2.2.2 The basic equation	43
2.2.3 Common-base characteristic	44
2.2.4 Common-emitter characteristic	50
2.2.5 The relationship between the current amplification factors	53

Page
2.2.6 The common-collector circuit 54
2.3 Effects of Temperature 55
2.4 Semiconductor Capacitance 60
2.5 Field-Effect Transistors (F.E.T.s) 62
2.5.1 The metal-oxide-semiconductor transistor (MOST) 64
3. BASIC SYSTEMS 67
3.1 Rectifiers 67
3.1.1 Power rectification 67
3.1.1.1 The half-wave circuit 68
3.1.1.2 The full-wave circuit 72
3.1.1.3 The bridge circuit 73
3.1.1.4 Voltage doubling 74
3.1.1.5 Filters 75
3.1.1.6 Voltage regulation 78
3.1.2 Demodulation 80
3.2 Amplifiers 83
3.2.1 Expectations 84
3.2.2 Limitations 85
3.2.2.1 Distortion 85
3.2.2.2 Noise 87
3.2.3 h-Parameter analysis 89
3.2.4 The two-stage amplifier 98
3.2.4.1 D.C. bias 99
3.2.4.2 The dynamic load line 101
3.2.4.3 Interstage coupling 103
3.2.5 Power amplifiers 105
3.2.5.1 Push-pull amplifiers 106
3.2.6 Direct coupled amplifiers 110
3.2.7 Negative feedback 113
3.2.7.1 Gain stability 115
3.2.7.2 Reduction of distortion 116
3.2.7.3 Frequency response 118
3.2.7.4 Practical circuits 118
3.3 Oscillators 121
3.3.1 Resonant-circuit oscillators 122
3.3.2 Resistance-capacitance oscillators . . . 125
3.3.3 Crystal-control 128

Page

3.4 Switching 129
3.4.1 The diode as a switch 131
3.4.1.1 Switching resistances 131
3.4.1.2 Switching times 132
3.4.2 The transistor as a switch 134
3.4.2.1 Switching resistances 135
3.4.2.2 Switching times 138
3.4.2.3 The bistable multivibrator 140
3.4.3 Switching logic 142
3.4.3.1 Mathematical logic 143
3.4.3.2 Circuit logic 145
3.4.4 Electronic gates 149
3.4.4.1 The AND gate 150
3.4.4.2 The OR gate 153
3.4.4.3 The NOTgate 155
3.4.5 Computers 155
3.4.5.1 Computer organization 156
3.4.5.2 Computer arithmetic 158
3.4.5.3 Memories 164

4. MICROMINIATURE TECHNOLOGY 167
4.1 Film Techniques 167
4.2 Integrated Circuit Techniques 168
4.2.1 Transistors and diodes 169
4.2.2 Passive components 171
4.3 Integrated Circuits 172
4.3.1 Analogue circuits 172
4.3.1.1 Operational amplifiers 172
4.3.1.2 Special purpose circuits 175
4.3.2 Digital circuits 175
4.3.2.1 Logic gates 176

5. A PAUSE FOR BREATH 179

APPENDICES

1. ABBREVIATIONS 181

2. CIRCUIT SYMBOLS 183

3. BINARY ARITHMETIC AND ITS APPLICATION TO COMPUTERS 185

Page

A3.1 Number Systems 185
A3.2 Binary Arithmetic 189
A3.2.1 Addition 189
A3.2.2 Multiplication 190
A3.2.3 Subtraction 191
A3.2.4 Division 192

4. MATHEMATICS 193
A4.1 The Straight-Line Graph 193
A4.2 Calculation of A.C. Resistance 196
A4.2.1 Gradient of a curve 196
A4.2.2 Graphical method 197
A4.2.3 By calculation 199

1. THE PHYSICS OF SEMICONDUCTORS

In only one book we are going to range over as many important electronic principles governing semiconductors and their offspring as possible. However, if we pause to consider the enormous field of semiconductor use in computers, television, radio, telephony and automation together with the fact that they are expanding further into almost all walks of life, it is clearly of utmost importance that we are not sidetracked into examining features in depth when it is not absolutely necessary. Our minds must be kept on the goal and time never wasted in probing into things which may be of great interest yet are irrelevant to the main project. So it will be throughout this book, we cover all that is necessary for a good *basic understanding* of the subject, delving as deeply as appropriate but never too superficially so that the inquisitive reader is left uncertain. This is especially important when considering the physical aspects of semiconductors for the greater part of each atom from the centre outwards undergoes no change when the atom is engaged in semiconductor duties, so after a brief resumé of atomic structure we pass on to the more active and vital boundary conditions. Just as on a race course, "they're off" means that all the activity will be on the perimeter, none in the vast area it encloses.

1.1 ATOMIC STRUCTURE

Modern physics, that is, the scientific study of the properties and nature of matter, is sufficiently advanced to help explain most of the phenomena we shall meet. In a simplified way we can base our semiconductor reasoning on the following concepts.

1.1.1 Composition of the atom

Except for size, an atom resembles the solar system with its *nucleus* (Latin, kernel) around which electrons are in orbit similar to the Sun with its orbiting Earth and other planets. Whereas it is difficult but possible for us to appreciate the

enormous size of the solar system; it is perhaps more difficult to come to terms with the minuteness of the atom. Like the solar system, most of the atom is space.

1.1.2 Atomic number

There are about one hundred different *elements* (substances which cannot be resolved into anything simpler) each comprising atoms having a specific number of electrons, for example, pure iron is an element, it is not made up of anything but iron. The number of electrons per iron atom is 26. Other examples are the main constituents of air, nitrogen (7) and oxygen (8), also copper (29), gold (79). Two less common but of paramount importance here are silicon (14) and germanium (32). These numbers are known as the *atomic numbers* and each number determines the element to which the atom belongs.

1.1.3 Shells

The whole atom may be considered to be of spherical shape. The orbiting electrons are constrained to move within the confines of concentric *shells* as shown in Fig. 1.1 for the silicon atom as an example. Physicists and chemists use bewildering displays of bars and coloured balls for this purpose to show the arrangement in three dimensions, on the printed page we have to manage with a flat diagram, the depth being left to the imagination.

Each shell of any atom has a definite maximum number of electrons which it can contain. With the shells numbered outwards from the nucleus as shown in Fig. 1.1, this maximum number is equal to $2n^2$ where n is the shell number. It would be most convenient if the arrangement of electrons in the shells followed the simple rule of filling completely each shell before commencing the next. This does apply up to atomic number 18 but thereafter the rules are slightly modified. Determination of the number of electrons in each shell is then more complicated and involves examination of the way in which shells are divided into *sub-shells*. The distribution of electrons in the atoms of a few well-known elements plus those of elements with which we will be concerned later, is

given in Table 1.1 in atomic number order.

1.1.4 Valence electrons

Only the electrons in the outer shell take part in the conduction process, that is, they may be forced away from their parent atoms when sufficient energy is supplied to them. They are known as *valence* electrons (those with vigour, from the Latin. valere – to be strong). From Table 1.1 the outermost shell of silicon is No. 3 and it contains four valence electrons (also see Fig. 1.1), for germanium the outermost shell is No. 4 and again it contains four valence electrons.

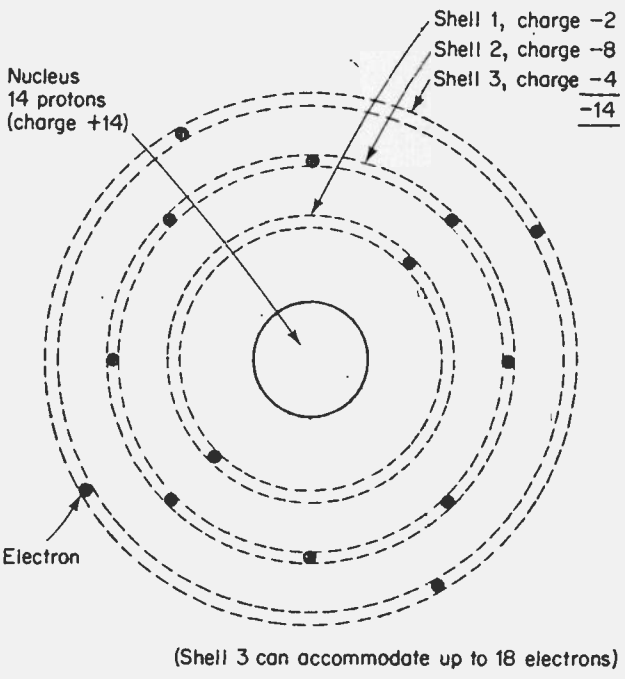


Fig. 1.1 Diagrammatic view of silicon atom.

Shell Number	Element											
	Boron	Silicon	Phosphorus	Iron	Copper	Gallium	Germanium	Arsenic	Silver	Indium	Antimony	Gold
1	2	2	2	2	2	2	2	2	2	2	2	2
2	3	8	8	8	8	8	8	8	8	8	8	8
3		4	5	14	18	18	18	18	18	18	18	18
4				2	1	3	4	5	18	18	18	32
5									1	3	5	18
6												1
Total (Atomic Number)	5	14	15	26	29	31	32	33	47	49	51	79

Table 1.1 **Distribution of Electrons in Atom Shells**

1.1.5 Positive and negative charges

We refer to the charge of an electron as *negative* electricity but an atom with its full complement of electrons is electrically neutral because the nucleus contains the same number of *positive* protons as there are electrons outside. The total charge on the protons exactly balances that of the electrons (see Fig. 1.1). However, under the condition that an electron has broken away, the atom is then positive and is known as an *ion* (Greek – go or wander).

1.1.6 Semiconductor crystals

Silicon is most commonly used in semiconductors, it is a grey crystalline substance, germanium follows which is silvery grey with a metallic appearance. Other materials are also used, usually for special purposes but we are concentrating on silicon and germanium first. From Sect. 1.1.4 both are seen to have four valence electrons, hence are known as *tetravalent* (Greek, tetra – four). A *crystal* (from Greek – clear ice) is a form some materials take, each having its own particular shape, for example, that of common salt is a cube while the crystalline form of carbon, a common material with usually no shape or beauty whatsoever, is the diamond. The crystalline forms of silicon and germanium have their atoms disposed in a regular, three-dimensional cluster tightly bound into a *diamond lattice.* The close bonding arises because each atom shares its four valence electrons with four other atoms, one with each, in what are known as *covalent bonds.* There are many ways of illustrating such an arrangement, a useful one is shown in Fig. 1.2, again not forgetting that in the atomic world nothing is flat as we have to show it on paper. In the figure the cubical structure of the lattice for atom 1 is shown with this atom at the centre. Atoms 2, 3, 4 and 5 are positioned at those corners of the cube which result in the atoms being diagonally opposite each other. Each of atoms 2, 3, 4 and 5 is equally at the centre of another similar cubical structure bonding it to four other atoms, so continuing to the surface of the material where such bonding must become incomplete. The dotted electron orbits in the figure give us diagrammatically a little more insight into the actual disposition of the valency electrons, showing how they are shared between two atoms so

that all four valency electrons per atom take part in the bonding process. Each electron is labelled, the first number indicating the parent atom. Thus each atom has eight electrons in four shared orbits (four of its own plus one from each of the linked atoms). Although this pictorial representation would not satisfy a physicist, it is all that we need for getting to grips with the basic concept.

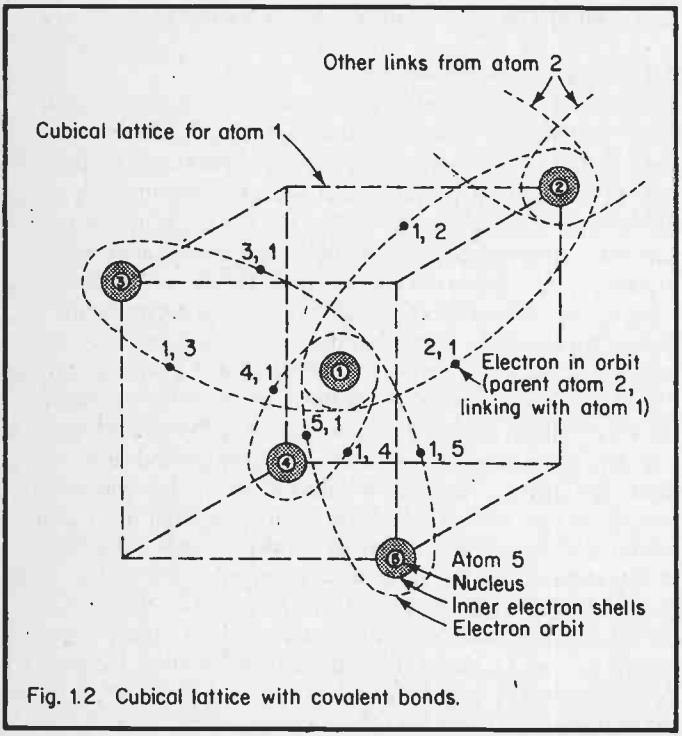


Fig. 1.2 Cubical lattice with covalent bonds.

It would appear from Fig. 1.2 that because of the tight bonds, no electrons are available for current conduction and this is true at absolute zero temperature (–273°C, i.e. 273° below the freezing point of water). When any heat is added to the crystal, energy is supplied. This allows some electrons to become disengaged from their orbits and move freely within

the comparatively enormous spaces between the atoms (as with steam escaping from water). The effect increases with temperature, greater numbers of electrons are released as temperature rises, thus the crystal changes from being a perfect insulator at absolute zero temperature to becoming a poor one at normal temperatures, in fact it is then a *semi-conductor* (Latin, semi – half), being neither a good insulator nor a good conductor. With semiconducting materials, therefore, as temperature rises, the resistance falls.

1.1.7 Electrons and holes

When an electron leaves its parent atom, that atom becomes positive and semiconductor theory has a slightly different way of looking at it. Free electrons constitute negative charges, the concept is that each leaves a *hole*, a convenient and descriptive way of labelling the vacancy created. The hole is positive and is considered to carry the same amount of charge as that on the electron. Thus for each free electron there is a hole and when any electron escapes we say that an *electron-hole pair* is created. We also talk of electrons and holes combining whereas what actually happens is that an electron falls into a vacancy in the outer shell of an atom, the vacancy having arisen from the earlier release of another (or sometimes the same) electron. Holes may also be considered to move because each electron which fills one must have left a hole somewhere else. An odd way of looking at things perhaps, but holes do occur in everyday life, and they move! Consider a queue at a post office or bank counter. The person at the head of the queue is served and leaves, a hole is created. Customer 2 moves up, fills the hole and leaves one further down the queue. Customer 3 moves into this and creates another. What happens is that the hole at the front end of the queue moves steadily to the rear end where it is available to be filled by a new customer. Consider also a crowded party where electrons are represented by people. When people move about they fill and create holes so not only are people moving but the holes are as well. So it is with electrons and holes and just as in the case of the queue where each hole coerces somebody into filling it, so a free electron encountering a hole finds its attraction irresistible.

1.2 ADDING CONDUCTION

Because the close bonding within a crystal arises through its covalent bonds it is clear that sufficient bonds must be effective at any time otherwise the character of the crystal would change. There is therefore a limit to the number of electrons escaping from the valence shells (the outermost – shells nearer the nucleus can be disregarded since they play no part in the conduction process) and conductivity is low. Since this small conductivity is entirely due to electrons belonging to the pure crystal it is known as *intrinsic conduction* (belonging naturally).

For development of semiconductor devices higher conductivities are required and these are obtained by adding an *impurity* to the crystal through a process known as *doping*. The term "impurity" does not infer that what is added is in any way unclean, it is in fact very pure itself but it detracts from the purity of the material to which it is added. Special diffusion techniques are used, the amount of impurity added is extremely small, about one impurity atom to 10^7–10^8 atoms of the base material. Impurity atoms add electrons or create holes and conduction via these is known as *extrinsic* (i.e. not belonging to the main crystal).

1.2.1 n-Type materials

If a *pentavalent* impurity (five electrons in the outermost shell) is added to a tetravalent crystal, its atoms will form bonds with the crystal atoms just as the latter do in the pure crystal and the lattice structure remains unchanged. There is now one electron per impurity atom thrown spare because it cannot enter into a covalent bond. Little energy is required for this electron to escape from its parent atom and become available for conduction. Table 1.1 shows three pentavalent elements as an example, phosphorus (15), arsenic (33) and antimony (51). Each has five electrons in the outermost shell and is therefore suitable as an impurity for doping a tetravalent element such as silicon (14) or germanium (32). Because impurity atoms give up electrons when they form bonds, they are known as *donors* and the doped crystal is

called *n-type* because of the added *n*egative electrons. There will be as many additional free electrons as there are donor atoms but the doped crystal is electrically neutral because each impurity atom added is itself neutral.

The main material now has a lower resistance and if a potential difference (p.d.) is connected across it in either direction most of the current flowing will be due to the added electrons, these are therefore known as *majority carriers*. However a current also flows by virtue of the electron-hole pairs created by heat (Sect. 1.1.6), this current in modern components can usually be kept low. These electrons and holes are known as *minority carriers*.

1.2.2 p-Type materials

Table 1.1 also gives examples of trivalent elements (3 valency electrons), boron (5), gallium (31) and indium (49). Again any atoms of these introduced into a silicon or germanium crystal will form bonds with the main material atoms but in this case one bond of each four per atom is incomplete so giving rise to a hole (i.e. a vacancy has been created for an electron to fill and hence complete the bond). The impurity atoms acquire electrons to complete their bonds, thus are known as *acceptors* and the doped material is called *p-type* because conduction takes place by virtue of *p*ositive holes.

Lest we mislead ourselves on the concept of conduction by holes we recall that although a hole is in reality a vacancy for an electron in the outer shell of an atom, because the atom is thereby positive, holes may be considered instead as representing positively charged particles just as the electron is a negative one. Current flow which so far has without exception been considered as electrons moving away from a negative charge towards a positive one, may thus also be seen as a flow of holes from positive to negative. Thus if a battery is connected across a p-type semiconductor, each hole which reaches the negative end of the semiconductor is filled (i.e. the covalent bond is completed) by an electron from the negative pole of the battery. Simultaneously the positive pole of the battery causes an electron to break away from a covalent bond, leaving a hole.

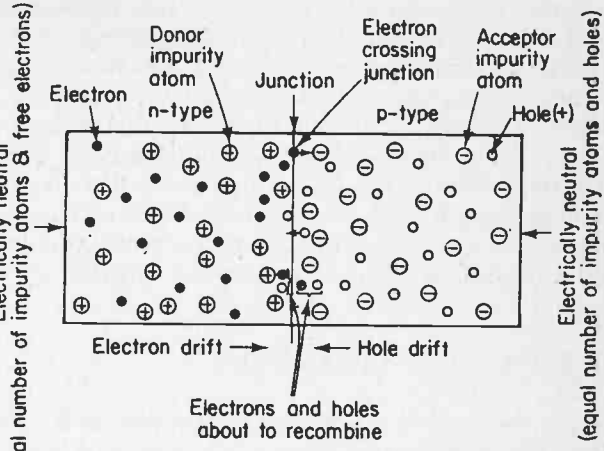
1.3 THE p-n JUNCTION

It is when the two different types of semiconductor, n-type and p-type are in close contact that things begin to happen. The contact is not simply effected by pressing the two separate materials together but by growing one crystal and doping separate but adjacent layers of it with the appropriate impurities. In this way the lattice structure remains continuous but it is where the two layers come together that most of the action takes place

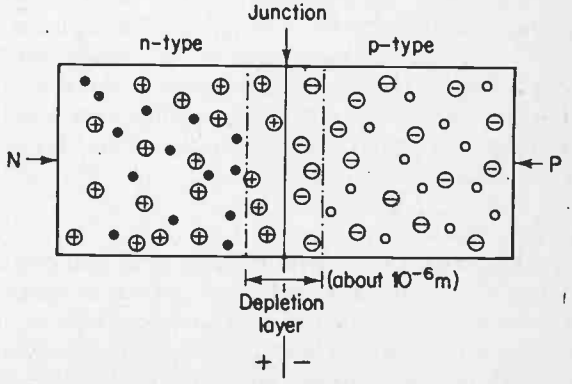
Fig. 1.3 shows diagrammatically such a junction, in practice the dividing line is not as clear-cut as shown but drawing one in this way simplifies explanation. For the same reason electrons and holes released by thermal action are omitted since their effect is minor. The impurity atoms are shown in random positions because this is how they link into the semiconductor material lattice by the diffusion technique employed.

Free electrons and holes, being negative and positive respectively, attract each other, accordingly at the junction the attraction of the holes in the p-type material causes electrons to flow across and equally there is a flow of holes from p-type to n-type. Many of these migrating charges combine with their opposites on the journey through the junction and cancel out. However, since most of the atom is space others are able to penetrate further before encountering other particles and recombining. The time a charge remains free (i.e. until recombination) is on average less than one microsecond.

The semiconductor materials are electrically neutral when the impurity atoms have their correct complement of valency electrons whether these are tied into covalent bonds or free, hence when, for example, electrons from the n-type material cross the junction, it is equivalent to leaving behind an impurity atom one electron short and therefore positive. In the opposite direction as holes leave the p-type material it becomes negative. This is shown in Fig. 1.3(ii). Because on either side of the junction there is a loss of mobile carriers



(i) Electron and hole drift commencing.



(ii) Final build-up of charge.

Fig. 1.3 n-type and p-type semiconductor materials in contact.

this region is known as the *depletion layer.*

Eventually the charges built up across the junction become sufficient to repel further electron or hole crossings. The process is a *dynamic* (active) one because even though the doping concentration is extremely small there are still millions of carriers (the free electrons and holes) milling around, continually recombining and breaking away. The total effect of drift across the junction is such that the potential arising from the charges remaining on each side holds at a level of a few tenths of a volt, depending on the crystal material. It is known as the *barrier potential.*

1.4 SEMICONDUCTORS AT WORK

The discussion, which is still limited to the physics of semiconductors normally moves next into how semiconductor *diodes* and *transistors* function, but it may well be beneficial to digress slightly to be sure first of what we want from them. It may thereby be easier to appreciate their technicalities by having some idea of what they are capable of doing in practice – to see what we are driving at, so to speak. In electronics a *black box* is often used as a way of representing a *system* of connected components having some special purpose when we are not considering the components or their connexions but only the input and/or output. The box is black because at present we do not wish to see inside it, the method is used in the examples which follow.

1.4.1 Rectification

This is the term used to describe the process of changing an alternating current into a related direct current, a typical everyday example being that of a car battery charger in which the a.c. mains supply is *rectified* to a direct current having distinguishable positive and negative polarities for charging the battery. In black box form this is shown in Fig. 1.4(i). Within the box one essential component is the semiconductor *diode* and the box might be conveniently labelled with the general symbol for a rectifier to show its purpose.

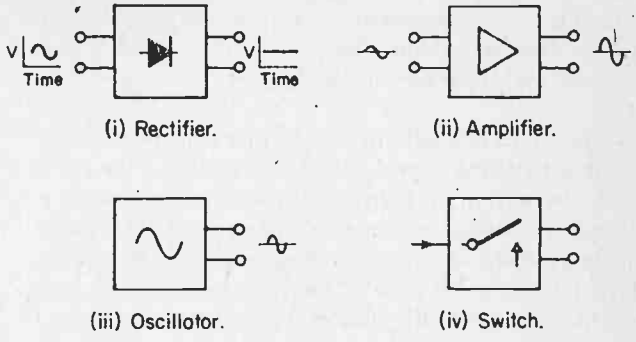


Fig. 1.4 Generol (block box) symbols for basic systems.

1.4.2 Amplification

To *amplify* is to enlarge, a treatment afforded to small *signals* (usually alternating currents carrying information, e.g. radio signals) to make their presence felt more strongly. As an example, the power output of a microphone can do nothing on its own but it may be amplified to a level sufficient to move the cone of a loudspeaker, typically from a fraction of a microwatt to several watts. The information content of the signal is unchanged in the process, thus the speech or music issuing from the loudspeaker is similar to that impinging on the microphone, although it may be of much greater loudness. Fig. 1.4(ii) shows one way of representing an amplifying system, the triangle pointing to the output. Within the box the essential component is the transistor for one of its many capabilities is that of receiving a small signal and delivering an amplified one.

1.4.3 Oscillation

An *oscillator* is a generator of alternating current. Although the turbo-alternator of a power station generates a.c., this is usually referred to as a power generator rather than as an oscillator which has an output at a much lower level. Oscil-

lators are devices which can provide the notes of an electronic organ or the *carrier wave* of a radio transmission; they are to be found in radio and television receivers and even electronic watches. They can produce frequencies from a few to many millions of cycles per second (Hz).

A common type is similar to an amplifier which has a small amount of its output signal fed back to reinforce the input, causing the output to increase further as then does the *feedback*, the effect continues to grow until the system oscillates, rather like keeping a swing oscillating by giving it a slight but sufficient push at the right time on each cycle. Transistors are especially suitable for oscillator circuits and a representation of an oscillator is shown in Fig. 1.4(iii). As the device is a generator, it has no input terminals.

1.4.4 Switching

The mechanical switch which makes or breaks a metallic electrical connexion is everywhere. Most are used in the home or workplace and as with electric light switches, are operated by hand. Electromechanical switches are more complicated, they have similar functions but are operated by an electrical input, often breaking circuits carrying thousands of amperes. For very much smaller currents, semiconductors can carry out switching functions from an electrical input and their use for this purpose has grown to an incredible degree in, for example, the *gate* (i.e. open or closed) circuits of microprocessors and computers; so much so that consideration of semiconductor switching is of equal importance to use of the devices in other fields. The simple black-box representation might be as in Fig. 1.4(iv).

We now lift the lids of our boxes and study first the basic components, the diode and the transistor, which make these facilities possible.

1.5 DIODES

A diode, that is, a device having two electrodes, is in its basic

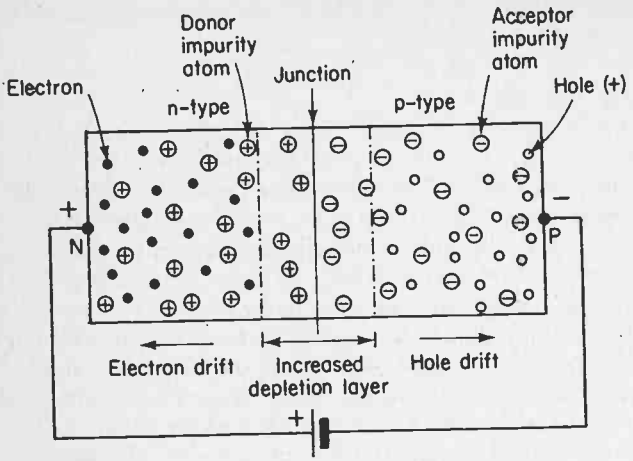
form a single p-n junction. Consider first the simplified sketch of Fig. 1.3(ii), "simplified" because no electrons or holes are shown in the depletion layer; there will of course be many but less than elsewhere.

1.5.1 Reverse bias

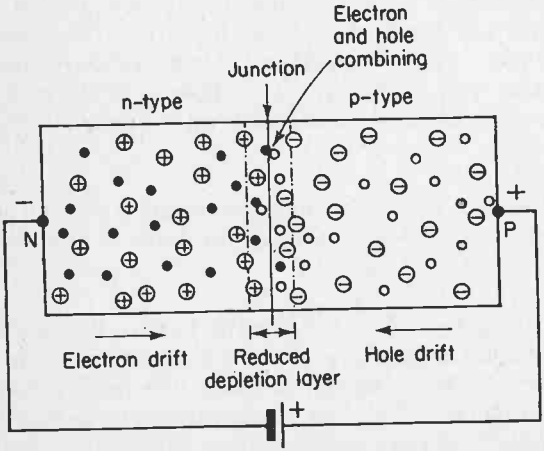
If a potential such as supplied by a battery is connected to the points N and P so that N is made +ve and P –ve, then electrons in the n-type material will be attracted towards the +ve battery terminal and similarly holes in the p-type will flow towards the –ve terminal. But this cannot give rise to a continuous current because the depletion layer loses its majority carriers and therefore creates a discontinuity in the circuit. Fig. 1.5 (which is Fig. 1.3(ii) redrawn to show this feature and note that the impurity atoms remain locked within the lattice structure) shows that both electrons and holes, having moved outwards from the junction because of the effect of the battery potentials have in effect widened the depletion layer. Looking at this in another way, since the normal charges in the depletion layer act against majority carrier movement across the junction, the battery potential is in fact aiding the junction potential and therefore dissuading the carriers from crossing even more. Connecting a battery as in the figure to apply potentials to the diode is known as *biassing* (= influencing). In this case the bias potential is known as a *reverse voltage* or *reverse bias.*

In summary therefore, biassing a p-n junction +ve to n and –ve to p inhibits current flow through the device, that is, its resistance becomes high.

We can no longer ignore the minority carriers, that is, the electrons and holes generated thermally as distinct from those added by doping. Although, as their names implies, they are of lower number, their effect cannot always be disregarded as in this case, for they provide a normal but relatively small current flow when a battery potential is applied, such current increasing with temperature. We must get things into perspective here, the minority carriers are few because the basic material is of high resistance, it is the doping which gives rise



(i) Reverse bias.



(ii) Forward bias.

Fig. 1.5 Effects of biassing an n-p junction.

to the multitude of majority carriers. With a reverse bias applied some minority carriers flow across the junction as a *reverse saturation current*, also known as the *leakage current* because it represents a small leak operating against what otherwise would be a complete obstruction to current by the depletion layer.

1.5.2 Forward bias

Again referring to Fig. 1.3(ii), with the battery connexions changed over, i.e. –ve to N, +ve to P, it is evident that electrons in the n-type and holes in the p-type will be driven towards the junction and will enter the depletion layer since the battery potential opposes and therefore effectively reduces the barrier potential. Recombination occurs in the vicinity of the junction, the consequent reduction in majority carriers allowing more to flow from the battery into the device, or using our semiconductor way of looking at current flow, electrons enter the n-region at N while holes enter at P. This is shown in Fig. 1.5(ii). It is not in conflict with our belief that current flow is a movement of electrons in one direction round a complete circuit because holes entering at P are exactly the same as electrons leaving, for an electron deserting an atom is equivalent to a hole being created.

The diode with forward bias has a low resistance and minority carriers need no consideration because their effect is swamped by the main flow of majority carriers.

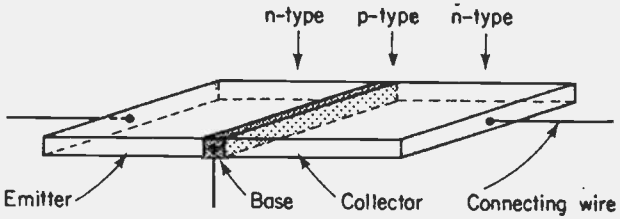
The semiconductor diode is therefore a device ideally suited to give the unidirectional current flow characteristic required of a rectifier as explained in Sect. 1.4.1. (we use the term "characteristic" here for any graph which demonstrates a particular quality or attribute of a component). The diode also has switching capabilities (Sect. 1.4.4) because it can approach *short-circuit* (zero resistance) or *open-circuit* (infinite resistance) conditions, depending on the polarity applied. The current/voltage relationships will be examined in the next chapter.

1.6 TRANSISTORS

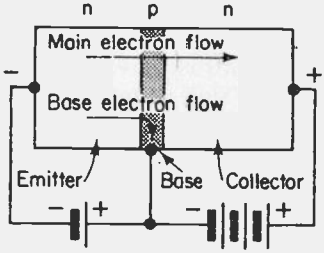
Explanation of transistor action follows logically on that of the diode because the transistor consists essentially of two n-p junction diodes back-to-back (in series with one reversed), that is n-p–p-n or p-n–n-p. Because the centre sections are of the same doping, they merge into one, giving types npn and pnp (the dashes are superfluous, so we now drop them). In the manufacture all three regions are grown in one crystal. The central region is very thin (a few μm) so the drawing in Fig. 1.6(i) is not to scale nor are transistors necessarily constructed in this way, there are many variations. The centre region is normally doped less than the outer ones. In the figure which shows an pnp transistor as an example, names are given to each of the regions, the derivations of these will be evident after the physical aspects of the transistor have been discussed. A pnp transistor is similarly labelled. The three connecting wires are bonded to the regions as shown.

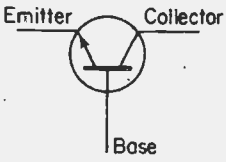
Fig. 1.6(ii) shows diagrammatically the transistor and its bias potentials. With the batteries so connected, –ve polarity is applied to the left-hand (l.h.) n region (the emitter) and +ve to the r.h. n region (the collector), the established way of producing electron flow from left to right in the diagram. However, the impurities and junctions have yet to be considered for they modify this current flow considerably.

The emitter-base junction is a forward biassed diode, hence electrons flow from emitter to base. The base-collector junction is reversed biassed, on its own therefore practically no current flows. However (and this is the secret of success of the transistor) owing to the thinness of the base layer and the relatively low number of impurity atoms in it, most of the electrons injected into the base by the emitter shoot straight through into the collector leaving only a small number (some 2% or less) combining with holes in the base region, there giving rise to a small base current. The relatively high +ve potential on the collector while not normally producing a base-collector current because it aids the junction barrier potential (which as shown in Fig. 1.5(i) prevents *holes* from



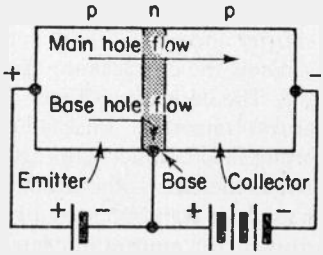
(i) npn junction transistor.

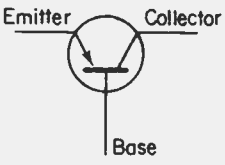




Symbol.

(ii) Biassing the npn transistor.





Symbol.

(iii) Biassing the pnp transistor.

Fig. 1.6 Junction transistors.

crossing from p to n) now finds itself able to attract a copious supply of *electrons* unaffected by the barrier. The electrons therefore flow into the collector and constitute a reverse current which flows on to the battery +ve terminal.

What is of the utmost importance in this action necessary to give the facility required for amplification (Sect. 1.4.2) is that

(i) variation in the potential applied between the emitter and base changes the effect of the barrier potential, hence the number of electrons injected into the collector, thus the (input) emitter-base voltage controls the (output) collector current. Since the emitter-base voltage also controls the base current then the important relationship that the collector current is controlled by the very much smaller base current applies, this is amplification;

(ii) the above can be looked at in another way. The emitter-base resistance is low because the junction is forward-biassed. The base-collector resistance is high because of reverse biassing. The current is approximately the same from emitter right through to the collector therefore the (output) power at the collector is greater than the (input) power to the emitter, this is another characteristic of amplification. We now discern how the word *transistor* was coined, from *transfer resistor.*

The derivation of the terms *emitter* and *collector* are perhaps obvious, the emitter emits or injects the charges into the base, and the collector collects them. The derivation of *base* is not immediately apparent, the earliest transistors consisted of two fine wires with their ends sprung into contact with a thin wafer of semiconductor known as the base – this arrangement is also evident in the symbols used for transistors as in Fig. 1.6 where the direction of the arrow on the emitter indicates *hole* current. This is in line with the symbol already in use for a diode in that the arrow points in the direction of conventional current, not electron flow. The symbols are very frequently drawn with the base vertical.(A2)

Explanation so far has centred on the npn transistor. The

action within a pnp transistor follows similar principles. Again the emitter-base junction is forward-biassed while the base-collector junction is reverse-biassed, therefore the external bias polarities are the opposite to those for the npn. The arrangement is shown in Fig. 1.6(iii) and the symbol shows that for a pnp transistor hole current flows into the base from the emitter.

The use of the transistor in an oscillator circuit (Sect. 1.4.3) follows from the fact that it has amplifying properties. Furthermore its use as a switch (Sect. 1.4.4) arises from the control the base current has over the collector current where a small current change at the base can turn on or off a much larger collector current. These facilities will become more apparent on examination of current/voltage relationships in the next chapter.

1.6.1 Construction

Germanium, the original semiconductor material still has some uses but most transistors are now silicon and come under the heading *NPN Silicon Planar Epitaxial*, we therefore concentrate on the construction of this type as an example, especially since the techniques have many similarities with those used in integrated circuit manufacture, as discussed in Chapter 4. This is not to say that pnp transistors of this type or other methods do not exist or are not important, there are many variations and in fact the one considered is not suitable in every case, for example, for high voltage working.

We understand the meaning of npn and silicon but the words "planar" and "epitaxial" are met for the first time and therefore need explanation. At the same time those other features which combine to make the whole process possible and convenient must be considered.

(i) Silicon Dioxide

Apart from its suitable valency for a basic semiconductor material, silicon has another advantage which puts it ahead of germanium for transistor mass production. It has a great affinity for oxygen, so it is quite easy to transform silicon

into its very useful dioxide (SiO_2). One method is to pass steam over heated silicon when the chemical reaction $Si + 2H_2O = SiO_2 + 2H_2$ takes place, in plain language, silicon plus water (in the form of steam) produces silicon dioxide by combining with the oxygen in water, leaving hydrogen to be given off as a gas. Equally oxygen gas may be used. Now silicon dioxide is very conveniently an insulator and also can be used as a barrier to other chemicals, (just as a coat of varnish might) and it is this facility of being able to produce an insulating film and barrier on a silicon slice which is so useful in building up the doped layers of a transistor.

(ii) Diffusion

It is evident that doping an element to the correct degree is at the heart of transistor action. Several ways of doing this have been developed but the one which has found most favour is by vapour diffusion. Slices of n-type silicon are placed in an oven and heated to such a temperature that the silicon is hot but not melting. Dry hydrogen gas carrying, for example, boron vapour (Sect. 1.2.2) passes over the slices, on each of which is thus diffused a p-type coating. The slice-coating junction is therefore n-p.

(iii) Epitaxy

In (ii) above a slice of n-type silicon is known as a *substrate* (a lower layer or foundation). When crystals are grown on a substrate by the diffusion process the deposited material has its atoms orientated to match those of the substrate and the crystal structures are in-line and continuous, a condition which is essential to transistor operation. From this arises the description *epitaxial* (Greek, epi – upon, the same axis).

(iv) Planar construction

The *planar* (meaning related to a plane, i.e. flat surface) technique became possible because of the development of the diffusion method. It differs from other methods (such as growing crystals while adjusting the amount of doping or fusing pellets of opposite type onto a base slice) by using a plane surface of one type and masking it so that diffusion

only takes place over a predetermined area.

(v) Photo-etching

Photo is derived from a Greek word meaning "light" while *etching* means eating away by acid, these two derivations describe the basic principles of the method. In more detail the process is as follows and is shown diagrammatically in Fig. 1.7. Consider a substrate of silicon (the doping does not matter in this explanation) shown at (1).

(2) a layer of silicon dioxide is formed on its plane surface. The dimensions in the figure are of course out of proportion, the layer being very thin.

(3) a *photoresist* lacquer coating is applied over the silicon dioxide film. The lacquer is one which reacts to ultraviolet light by changing its chemical composition, thereafter being insoluble in certain solvents whereas the unexposed lacquer is.

(4) a photographic mask is placed over the photoresist lacquer. Where eventually diffusion is to be effective on the substrate, the mask is black, elsewhere transparent. The combination of substrate and mask is then exposed to ultraviolet light.

(5) the mask is removed and by applying the special solvent, the photoresist lacquer is dissolved away where the light has not reached it because of the shielding by the mask. See Fig. 1.7(ii).

(6) immersion in hydrofluoric acid etches away the silicon dioxide where the photoresist has been removed, leaving the desired window (circular in our example) in the silicon dioxide, the remainder of the latter protects the rest of the substrate so that the ensuing diffusion process affects the window part only. The remainder of the photoresist is then removed [Fig. 1.7(iii)].

This may seem a tedious process but a large number of components is manufactured at the same time on one larger substrate. When it is realised that for example a low power

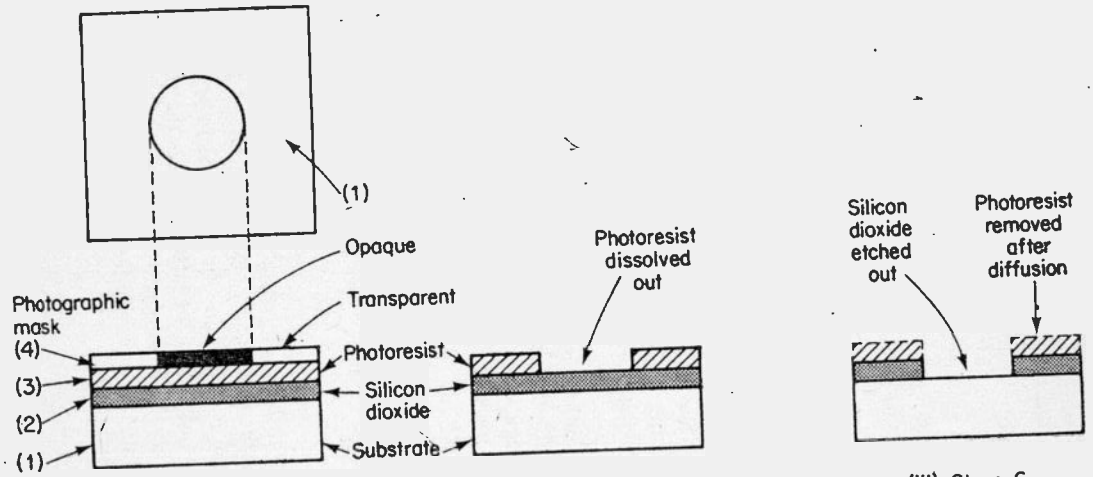


Fig. 1.7 Photo-etching.

silicon transistor may be mounted inside a can only some 4 mm across, its tiny size can be appreciated. Thus all stages above cater not just for one transistor, but for very many on each main slice and the photo-etching mask for one circular window per transistor would appear as a larger sheet covered in rows of black dots.

Putting all the above together we can now follow through the construction of a typical transistor of this type. Fig. 1.8 refers:

Stage 1.1: through the epitaxial process an n-type layer is formed on a lower resistance substrate to provide the collector;

Stage 1.2: a silicon dioxide coating is formed on the substrate and the appropriate window made in the dioxide;

Stage 1.3: a p-type base is diffused into the n-type material through the window only since the remainder of the slice is protected by the silicon dioxide film.

Stage 2: a second silicon dioxide coating is added (2.1), a window is formed in it by etching and by diffusion an n-type emitter is inserted (2.2). This completes the n-p-n system but connexions must be added. The substrate is in contact with the collector and a wire can therefore be bonded to this to provide the collector connexion. Base and emitter connexions are made as follows:

Stage 3: a ring-shaped window is formed above the p-region and aluminium is evaporated into it to make contact with the base (3.1). At the same time an aluminium contact is made on the emitter region in the centre (3.2). Wires are then bonded to the aluminium to give the base and emitter connexions.

1.6.2 Shapes and sizes

Even with some standardization, there are many different forms of both diode and transistor. The physical bulk depends mainly on the maximum power likely to be dissipated within the device but apart from size, the method of mounting may

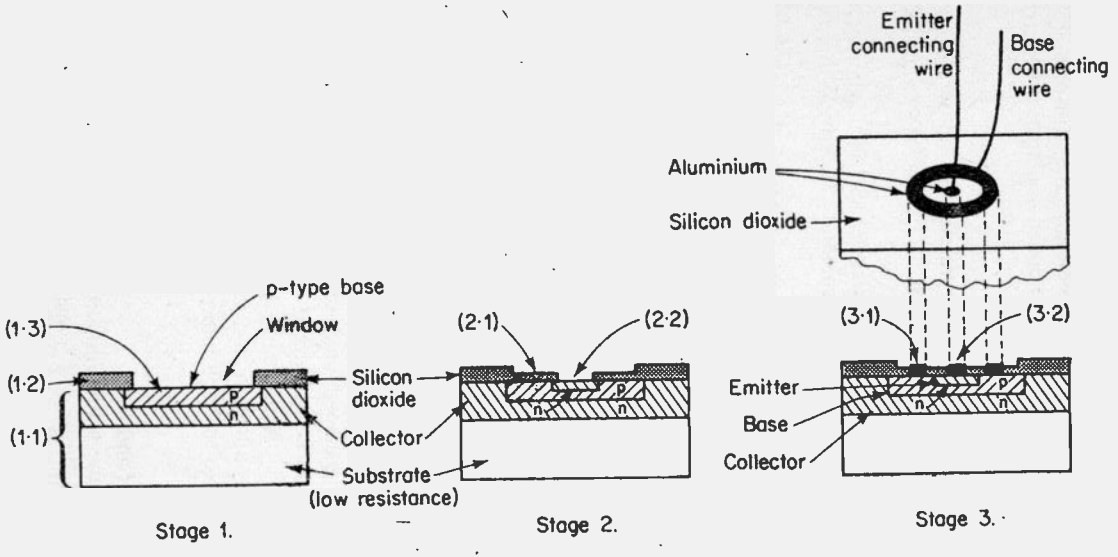
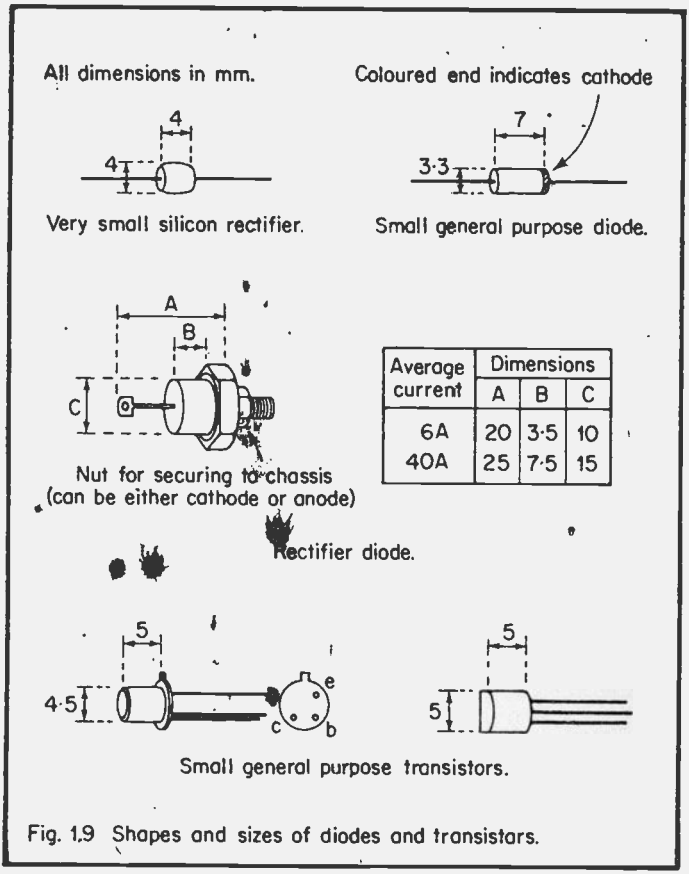


Fig. 1.8 Construction of npn silicon planar epitaxial transistor.

be a major factor, for example, a component may be supported on its own wires, fixed into printed wiring (a copper laminate on an insulating board) or bolted to a chassis. Fig. 1.9 shows a few typical forms.

Undoubtedly most diodes and transistors are part of integrated circuits where the size is so very tiny that drawings are not appropriate, the construction of the latter is discussed later. Clearly from what has been said above, if the size is very small, so too is the power-handling capacity.



Average current	Dimensions		
	A	B	C
6A	20	3·5	10
40A	25	7·5	15

Fig. 1.9 Shapes and sizes of diodes and transistors.

2. SEMICONDUCTOR CHARACTERISTICS

Chapter I concerns itself entirely with promoting an understanding of *how* semiconductors work. From the typical characteristics discussed in this chapter we shall begin to understand their function in practical circuits, changing as it were from considering the motivation of a few atomic particles to the combined effect of the many millions of them which constitute a measurable electronic current.

The reader may perhaps be wondering how we can reconcile random (haphazard or by mere chance) movements of electrons with steady currents which can be measured. Actually the number moving past a point in a given time measured on a meter in, say, μA or mA is not constant even though the meter says it is. The number of electrons per second constituting one ampere of current (6.25×10^{18}) is the *mean* or *average*, the actual number at any instant varies above and below this value but the meter does not notice quite a few either way in a total number of this magnitude. Nevertheless the random nature of electron movement cannot always be ignored, its effects are therefore discussed in more detail in Section 3.2.2.2.

2.1 DIODES

Although the diode is the simplest form of semiconductor device, it is available with a surprisingly wide range of characteristics. Special purpose diodes are available for high-frequency working, high-speed switching and for many other applications. We consider, however, mainly the "general purpose diode" and one special variation for voltage regulation.

2.1.1 Measurement of working characteristics

The current-voltage graph of a diode is obtained from straightforward measurements of voltage across the diode, simultaneously with the current flowing through it. A typical

circuit is given in Fig. 2.1. The diode D under investigation is connected in series with an ammeter A to measure the current while the voltmeter V shows the p.d. applied. A tiny inaccuracy arises since V actually indicates the p.d. across both diode and ammeter in series, hence reads slightly high. Connecting the r.h. terminal of V to point X only results in the ammeter indicating the sum of the diode and voltmeter currents. Fortunately in practice the error using modern meters is small enough to be neglected.

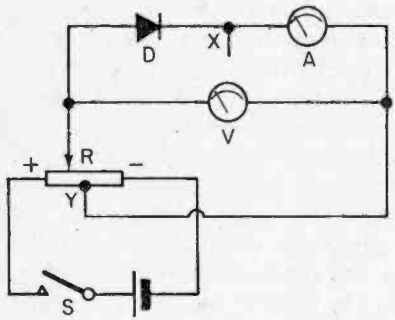


Fig. 2.1 Circuit for measurement of diode current/voltage characteristic.

When switch S is operated the battery voltage (say, 3 V) appears across resistor R. Y is the centre-tap of R and relative to Y the l.h. end of R is therefore +1.5 V and the r.h. end −1.5 V. Movement of the slider from left to right varies the potential across the diode from 1.5 V forward to 1.5 V reverse. Current readings are taken over a range of voltages enabling the characteristic to be drawn on graph paper (e.g. Fig. 2.2).

2.1.2 Rectifier diodes

A good diode characteristic for most rectification applications is that of very low forward resistance but infinite in reverse. The latter requirement cannot quite be met because of the

reverse saturation current (Sect. 1.5.1). Although very small, this must be taken into account especially as it is part of the basic diode equation from which we can predict the theoretical current through a pn junction knowing only the reverse saturation current, i_s, junction voltage, V and its temperature T. T is the *absolute temperature* measured in degrees Kelvin (°K, after Lord Kelvin, a British physicist). The scale starts at 0°K, which is −273°C, the *absolute zero*, therefore, for example, melting ice at 0°C has an absolute temperature of 273°K and water boils at 100°C or 373°K.

We cannot derive the equation for this would require a whole chapter, but by using it we can see what the *shape* of any pn junction current/voltage characteristic is like. The diode equation modified for our purpose is

$$I = I_s \left(e^{\frac{11600V}{T}} - 1 \right)$$

where I is the current at voltage V, I_s is the reverse saturation current and T is the junction temperature in °K.

By choosing a temperature of 305°K which is a convenient figure to make the exponent of e a whole number:

$$I = I_s (e^{38V} - 1)$$

For a given voltage therefore the ratio of the diode current to the reverse saturation current can be calculated:

$$\frac{I}{I_s} = e^{38V} - 1$$

As an example the calculation follows for V = 0.05 volts:

$38\ V = 1.9$ $\qquad$ $[\log_{10} e^{1.9} = 1.9 \log_{10} e$

$e^{1.9} = 6.686$ $\qquad$ $= 1.9 \times 0.4343 = 0.8252$

$\therefore \quad e^{1.9} - 1 = 5.686$ antilog $0.8252 = 6.686$]

i.e. $I = I_s \times 5.69$

Since I_s is constant for any particular diode, the shape of the current/voltage curve is obtained as in Fig. 2.2. The curve for any other diode will approximate to this shape and it is described as *exponential* because the independent variable V appears in the exponent.

Since V is the junction voltage which is not measurable by ordinary meters, practical values, where the voltage is measured across the whole diode, will not be fully in agreement. Also there will be a further discrepancy at a junction temperature different from that chosen above. Nevertheless the shape will be similar and what we learn from Fig. 2.2 is that:

(i) the reverse saturation current which arises from electron-hole pairs created by thermal agitation and not through the action of the majority carriers, is very nearly constant at reverse voltages greater than about 0.1 V. This is also clear from the formula because e raised to any negative power approaches 0 as the exponent increases (e.g. when $V = -0.1$, $e^{38V} = 0.022$, when $V = -0.2$, $e^{38V} = 0.0005$), hence I/I_s approaches -1, i.e. $I \approx -I_s$. We also now see why it is called a saturation current, it does not increase at higher reverse voltages (an exception is considered later).

In a modern silicon diode, reverse saturation current may be as low as 10–100 nA, giving a reverse resistance of the order of megohms.

(ii) the rise in forward current is very rapid as forward voltage increases and forward resistances fall to values as low as a few ohms. In practice the steep rise in forward current occurs at a lower voltage for germanium than for silicon.

Discussion of the use of diodes of this type in rectifier circuits continues in Section 3.1.

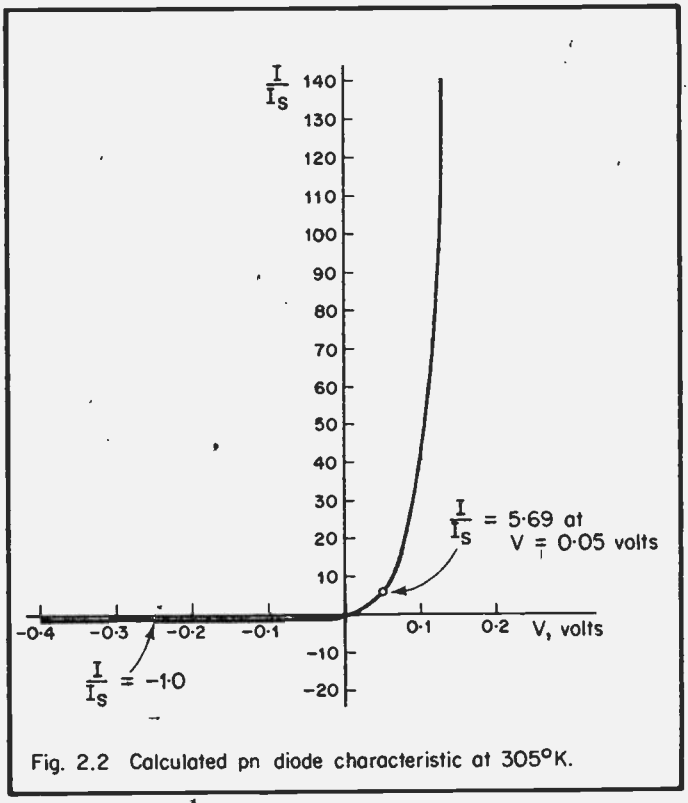


Fig. 2.2 Calculated pn diode characteristic at 305°K.

2.1.3 Voltage regulator diodes

A reservation is made in the previous section about reverse saturation current not increasing at higher reverse voltages. Within a certain range of voltage this is true and the general diode is worked within this range. However, if the reverse voltage is increased sufficiently an exceptional condition arises within the diode where *voltage breakdown* occurs.

An electron of the reverse saturation current moving at great speed because of the high reverse potential collides with a fixed atom and breaks up one of its valence bonds, thereby releasing another electron and creating a hole. The second

electron is itself accelerated by the high potential and the energy it gains allows it to create a further electron-hole pair in the same way. Since the first electron may still be active the effect is cumulative, resulting in the release of electrons on a rapidly increasing scale, known as an *avalanche* – just as one small rock, gaining energy as it falls down a mountain-side releases others on its way and these each release more until an avalanche, consisting of a huge mass of tumbling rocks, crashes down.

Diodes are manufactured with breakdown voltages from just below one up to about 200.

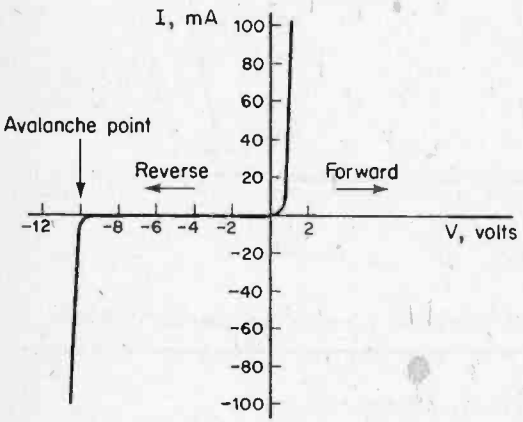


Fig. 2.3 Junction diode characteristic showing reverse voltage breakdown.

The characteristic of a typical avalanche diode (at 10 V in this case) is shown in Fig. 2.3. The interesting feature is that at the breakdown or avalanche point the reverse current changes considerably for little change in reverse voltage thus the diode may be used to regulate (keep constant) the voltage of a supply. Precautions in circuit design are then needed to

avoid destruction of the diode by excessive current causing overheating.

Such diodes are also known as *Zener diodes* (after the American physicist C. M. Zener, who was one of the first to study the avalanche effect) or *reference diodes*, the latter because under avalanche conditions a known, stable voltage is available from the device.

2.1.4 The static load line

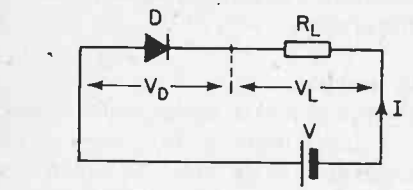
Ohm's Law has served us well so far in circuit analysis where the conditions are linear, meaning that graphs of current/voltage relationships are straight lines. We have now met the diode characteristic which is not linear, although at any point on it Ohm's Law applies. Thus given a voltage and the current it produces we can calculate the resistance from V/I at that point. But this value of resistance does not apply on other parts of the characteristic, so V/I does not give the same result. Hence calculation of circuit quantities when the diode is coupled with other components needs some artifice and this is where the *load line* is useful. Consider the diode-resistance circuit of Fig. 2.4(i), the diode D having a typical (silicon) characteristic as at (ii). Suppose V = 1.5 V and the *load* resistance R_L = 5 Ω. A simple circuit indeed, but how can the current be calculated when the voltage drop across D depends on it? We could, of course, assume various values of current and for each value calculate:

(i) the voltage V_L across R_L; and
(ii) from the characteristic read off the voltage V_D across D.

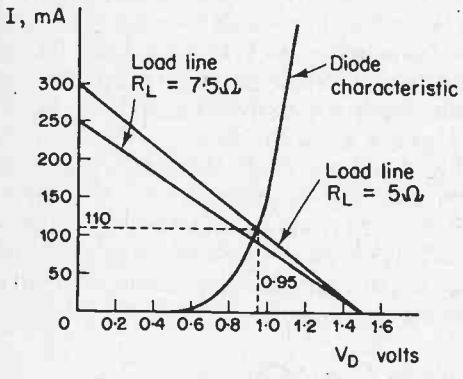
Then, $V = V_L + V_D$ and at only one current will V = 1.5 V. This can be determined by drawing a graph or by a *cut and try* process as shown below:

I	50	100	150	110	mA
V_D	0.90	0.94	0.99	0.95	V
V_L	0.25	0.50	0.75	0.55	V
V	1.15	1.44	1.74	1.50	V

which shows that the correct current for V = 1.5 V is 110 mA. The *static load line* ("static" because it refers to the d.c. condition) obviates such work which is not always as simple as in this example.



(i) Diode/resistance circuit.



(ii) Characteristic and load lines.

Fig. 2.4 Static load lines.

The current through the diode at various voltages is given by the characteristic (as measured or as published by the manufacturer). To this we add a second characteristic calculated from the current/voltage relationship of the load. Except at one point only, the two characteristics cannot agree because the diode has its own different ideas on what the current

should be. It is, however, this one point in which we are interested, where the two characteristics meet, for here both tell the same story about the current. This is explained in more practical terms by determining the load line for Fig. 2.4(i).

A table might be constructed for calculating the graph points as follows ($R_L = 5\ \Omega$):

V_D	0	0.5	1.0	1.5	V
$V_L (= 1.5 - V_D)$	1.5	1.0	0.5	0	V
$I (= (V_L \times 1000)/5)$	300	200	100	0	mA

Only a few points are shown because it soon becomes obvious that the graph is a straight line. We examine the straight line graph a little more closely in Appendix 4, Section 1, from which it is clear that the equation for I is of the general form $y = mx + c$ since

$$I = \frac{V - V_D}{5} \text{ A}$$

$$\therefore \quad I = \frac{-V_D}{5} + \frac{V}{5}$$

that is, $m = -1/5$ and $c = 1.5/5 = 0.3$, giving $I = -1/5.\ V_D + 0.3$. the negative sign indicating negative slope.

For a straight line, two points only are necessary to place it, the most suitable choices being on the axes at $V_D = 0$, i.e. $I = V/R_L$ and at $V_D = 1.5$ (the full supply voltage) where $I = 0$.

In Fig. 2.4(ii) the load line for $R_L = 5\ \Omega$ is drawn and the point at which this line cuts the diode characteristic gives the current and voltage for the circuit in (i). These values of

$I = 110$ mA, $V_D = 0.95$ V are seen to agree with the "cut and try" method used first. Summing up, the load line is drawn between two points:

(i) V/R_L on the current axis;
(ii) V on the voltage axis.

A second load line for $R_L = 7.5\ \Omega$ is also shown, it is drawn between 200 mA and 1.5 V. The circuit quantities are read off as $V_D = 0.9$ V, $I = 80$ mA.

Dexterity with graphs and load lines is of paramount importance in semiconductor engineering. The following example is designed to give a little more practice with both so that we are well prepared for *dynamic* (a.c. conditions) load lines when we study transistor amplifiers.

EXAMPLE:
A rectifier conforms to the equation $i = 1.5\ V^4$ over the range for V of 0–0.8 V. It is connected in series with a load of 3.5 Ω.

(a) What current flows when the applied voltage is 1.4; and
(b) when the load resistance is changed to 2 Ω and the applied voltage reduced to 1.2?

The circuit is shown in Fig. 2.5(i). First calculate the graph points, one example of which, using logarithms follows:

when $V = 0.5$ V,

$I = 1.5\ V^4 = 0.09378$ A

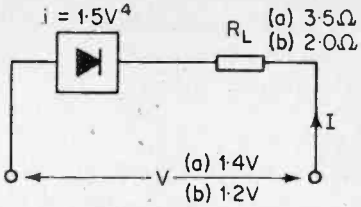
$= 93.78$ mA

[log V = $\bar{1}.6990$

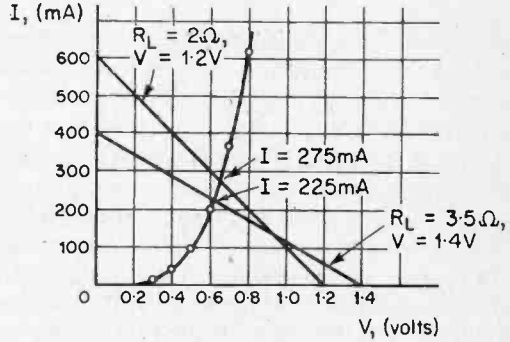
log V^4 = $\bar{2}.7960$
log 1.5 = 0.1761

log $1.5\ V^4$ = $\bar{2}.9721$
antilog = 0.09378]

The calculations are then assembled in a table which could include columns or lines for the logarithm steps if desired.



(i) Circuit.



(ii) Characteristic with load lines.

Fig. 2.5 Current through rectifier and load.

V(volts)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
I(mA)	0.15	2.40	12.2	38.4	93.8	194.4	360.2	614.4

and the graph is then drawn on suitable squared paper as shown in Fig. 2.5(ii).

The simplicity of the load line calculation now becomes evident:

(a) $R_L = 3.5\ \Omega$, $V = 1.4$ V

Point on I axis at $V/R_L = 1.4/3.5 \times 1000 = 400$ mA
Point on V axis at 1.4 V.
The load line is then drawn between these two points, it crosses the diode characteristic at I = 225 mA, this is therefore the circuit current.

(b) $R_L = 2\ \Omega$, V = 1.2 V
Points for load line at 600 mA, 1.2 V.
This line indicates a circuit current of 275 mA.

2.2 TRANSISTORS

The single current/voltage characteristic of the diode now gives way to several for the transistor with its three connexions instead of two. Hence we must first understand the generally accepted method of identifying through which of the three regions a current is flowing or across which two a voltage is developed. Base, collector and emitter are easily denoted by B, C and E respectively, so if a voltage exists across base and emitter it is shown as V_{BE}, the two-letter subscript indicating the pair of regions. As circuits are developed there will usually be found a common line to which many of the components and one pole of the battery are connected, this line may also be connected to *chassis*, that is, the case of the equipment, sometimes even to *earth*. When a common line or *rail* is used it is frequently marked 0 V and voltages are quoted relative to it. Under this condition a single subscript may be used, for example, the collector voltage on a transistor which has its emitter connected directly to the common need not be designated V_{CE} but simply V_C. Currents flow into or out of each region and therefore usually need only one subscript, for example, I_B represents the current flowing into (or out of) the base. At times it is important to show when two regions are being considered that the third is on open-circuit (nothing connected to it), thus I_{CBO} represents the current flowing between collector and emitter with the base on open-circuit.

2.2.1 Basic circuit configurations

Referring to Fig. 1.4(ii) shows that an amplifier has two input

and two output terminals, four altogether, yet the transistor, already shown to be a suitable device to be put into the box, has only three. Hence as an amplifier one region must be made common to both input and output and Fig. 2.6 shows the three choices for an npn transistor. A pnp can be connected similarly but in this case with battery potentials reversed. Note the common or 0 V rail to which one pole of each battery is connected.

Fig. 2.6(i) shows the elements of the *common-base* circuit, described first because of its close relationship with the drawing of an npn transistor in Fig. 1.6(ii). The base is connected to the common rail as is one terminal of the output circuit and effectively one of the input circuit. In such an arrangement it must be presumed that there is a d.c. path through the circuit connected to the input terminals, otherwise the bias potential V_{EB} will not be applied. Note that the biassing potentials agree with Fig. 1.6(ii) in that the emitter-base junction is forward biassed while the base-collector junction is reverse biassed. Comparison with Fig. 1.6(iii) checks the statement above that a pnp common-base circuit would be similar except that the battery connexions are reversed. In Fig. 2.6(i) voltages have been marked in accordance with the system described above.

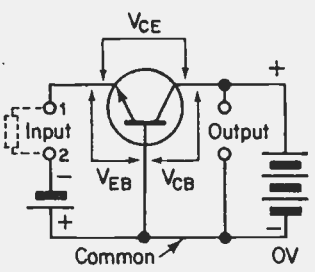
Fig. 2.6(ii) shows the more practical common-base circuit which has a *load resistance* R_L added. R_L is necessary to avoid the output circuit being *shunted* (i.e. another resistance connected in parallel) by the low resistance of the battery so losing most of the output signal in this resistance instead of producing a useful current in the output circuit. Calculation of the optimum value for R_L features prominently in circuit design.

Fig. 2.6(iii) shows the *common-emitter* circuit, the input being applied to emitter and base as before but in this case the output is derived from collector and emitter since the emitter is connected to the common. Again note the biassing, remembering that the arrow direction of the transistor symbol indicates hole current, not electron flow.

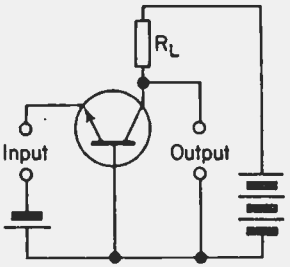
For current to flow the arrow must therefore point towards a negative battery terminal or away from a positive one.

The remaining method of connexion is *common-collector* and is illustrated in Fig. 2.6(iv). The input is between base and collector and the output derived from emitter and collector, the collector being connected to the common.

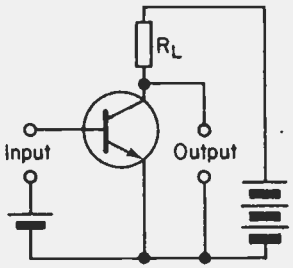
We next study some of the many static characteristics (i.e. currents and voltages remaining constant while being



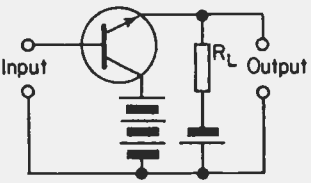
(i) Elements of common-base circuit.



(ii) Common-base circuit with load resistance.



(iii) Common-emitter circuit.



(iv) Common-collector circuit.

Fig. 2.6 Bosic npn transistor connexions.

measured). They tell us much about the tiny device and how it will function in practical circuits. Such measurements are made with high-resistance voltmeters and low-resistance ammeters connected as shown typically for the diode in Fig. 2.1.

In moving towards the practical aspects of diode and transistor circuits, there is no need to be too pedantic about directions of currents and voltages even to the extent of drawing graphs in different quadrants because this adds little towards achieving an understanding of what happens in the output circuit when conditions change at the input. Thus we draw graphs entirely in the first quadrant, labelling currents or voltages as negative only when this seems appropriate.

The common-emitter circuit is probably the one most frequently used, the common collector least. However, we look at common-base first because as mentioned previously, this follows automatically from the drawings in Fig. 1.6.

2.2.2 The basic equations

Voltage : Consider Fig. 2.6(i). The two batteries which supply V_{CB} and V_{EB} are effectively in series and therefore their voltages are additive, i.e. $(V_{CB} + V_{EB})$. This is seen to be equal to V_{CE} which must have a negative sign because it operates in the opposite direction :

$$\therefore \quad V_{CB} + V_{EB} = -V_{CE}$$

$$\therefore \quad \underline{V_{CB} + V_{EB} + V_{CE} = 0.}$$

Current: The current I_E flowing into the emitter divides between the base and the collector, that is, it flows out as $(I_B + I_C)$:

$$\therefore \quad I_E = -(I_B + I_C)$$

$$\therefore \quad \underline{I_E + I_B + I_C = 0.}$$

2.2.3 Common-base characteristics

Again from Fig. 2.6(i), I_E flows in the input circuit and I_C in the output circuit. The ratio I_C/I_E is known as the *current gain* or *current amplification factor*, denoted by α_B (B for common-base and a capital letter to indicate the d.c. condition). For the common-base circuit I_C is slightly less than I_E because of recombination in the base region, resulting in I_B, hence α_B is less than but usually very nearly 1.

We seem to be far removed from amplification when the output current is actually less than the input current but these two currents flow in different resistances, low at the input, high at the output, hence as mentioned in Section 1.6, there is a voltage gain. Our interest in any transistor is therefore mainly in (i) the input characteristic to determine the input resistance, similarly (ii) the output characteristic and (iii) the transfer characteristic relating input and output currents, from which the current gain follows.

(i) Input characteristics

A typical input characteristic is given in Fig. 2.7(i). This shows the relationship between the voltage applied across emitter and base, V_{EB} and the emitter current I_E. As might be expected, the curve is the exponential one of a forward-biassed junction diode. The measurement work in this particular case simply involves adjusting V_{EB} at, say, 25 mV intervals and reading I_E at each. The collector-base voltage V_{CB} is quoted because it does affect the curve slightly by modifying the depletion layer.

Ohm's Law gives the static (d.c.) resistance at any point on the curve simply by dividing the voltage at that point by the current it produces, for example at point X in the figure, 675 mV/6.05 mA = 111.6 Ω, at point Y, 750 mV/75 mA = 10 Ω, showing that as the curve becomes steeper the resistance is falling.

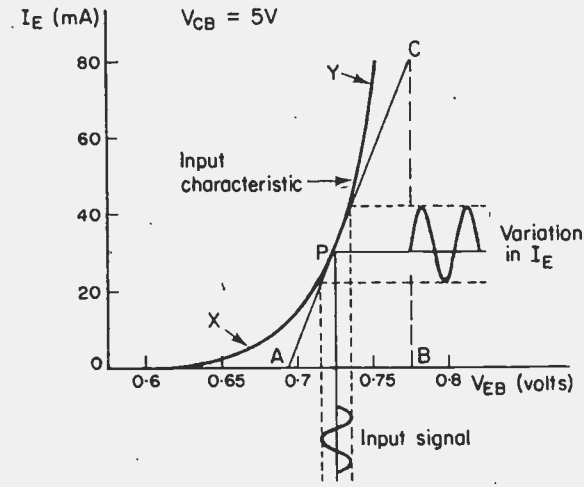
For amplifiers we are more interested in the *a.c. resistance* (strictly it is an impedance because the junction has some

capacitance) which is in effect, the resistance to small alternating currents. This is quite different from the value of resistance as calculated from the static values of voltage and current at any point. It is calculated from how much the current changes for a given small change in voltage. This occurs when a small waveform is applied to the input terminals and therefore adds and subtracts its peak value of voltage to and from that of the battery providing the forward bias. Because the a.c. resistance occurs under working conditions it is also known as the *dynamic resistance* or impedance.

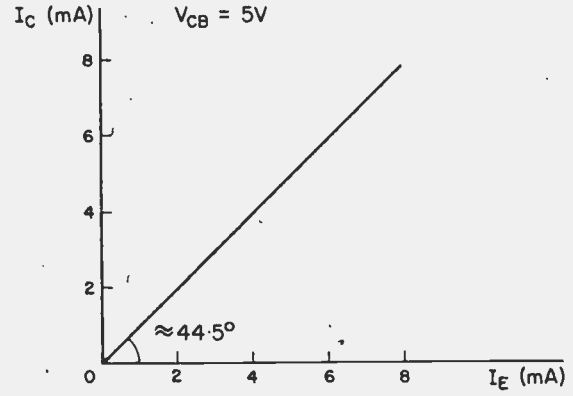
A useful way of illustrating the resultant emitter current from a small sine-wave input voltage has been added to Fig. 2.7(i). Suppose the bias battery to have a voltage of 725 mV, producing a steady emitter current of 30 mA as shown. If an input waveform has a peak voltage of 10 mV, then when it is in such a direction that input terminal 2 is positive and terminal 1 negative [Fig. 2.6(i)], it will add this value to V_{EB} and alternatively subtract it on reversing. Thus V_{EB} swings between 715 and 735 mV and from the curve we see that I_E swings from 22 to 42 mA.

Two points immediately arise, (i) as the graph is continually changing its slope the two half-cycles of the input waveform may not produce the same change in current and therefore (ii) the a.c. resistance varies over the portion of the characteristic used. The first consideration gives rise to *distortion* because the change in I_E does not follow exactly the change in V_{EB} and the larger the input signal, the more apparent this becomes. Equally the straighter the characteristic, the less the distortion, the inference from this being that a higher bias voltage should have been used.

We calculate the a.c. resistance at the operating point as though the curve there had changed to a straight line. Graphically the a.c. resistance is measured from the slope of this line and because measurement of the slope of a characteristic is a general requirement in electronics, not one limited to this particular type of characteristic only, the technique is set out

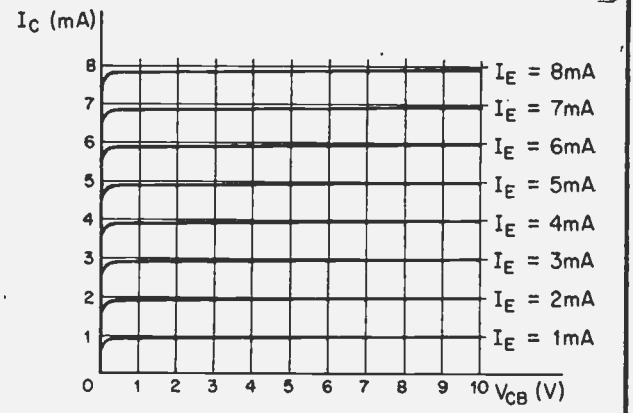


(i) Input characteristic.



(iii) Transfer characteristic.

Fig. 2.7 Typical common-base characteristics.



(ii) Output characteristics.

in Appendix 4, Section 2, both graphical and by calculation.

The graphical method in this case gives a tangent AC as shown. BC gives the change in I_E for a change in V_{EB} given by AB. The a.c. resistance follows from AB/BC. In the figure the a.c. resistance at point P is approximately 0.08 V/0.08 A = 1.0 Ω

By calculation, following the technique of Appendix 4.2, the a.c. resistance is 1.091 Ω, the two answers are in reasonable agreement.

The graphical method may suffer from some inaccuracy of measurement or from incorrectly drawing the tangent. For readers who may be interested in the mathematical alternative this has been added to the appendix mainly to emphasize the indispensability of mathematics in electronic calculations especially when the alternative is an answer subject to human choice. The mathematics in Appendix 4.2 have already been covered except for one brief encounter with the art of calculus.

(ii) Output characteristics

The output characteristics are again current/voltage relationships but for the collector, i.e. how I_C changes with V_{CB}. The collector receives most of the current that the emitter supplies according to its bias voltage, thus a *family* of characteristics is necessary, each relating to a different value of I_E. Fig. 2.7(ii) shows such a family. The curves are obtained by maintaining I_E at each chosen value, say at 1 mA intervals, varying V_{CB} at 1 or 2 volt steps and noting the value of I_C at each.

Evidently, irrespective of the value of V_{CB} (except at very low ones), there is practically no change in I_C simply because the maximum current allowed through the device is controlled by the conditions on the emitter. I_C falls slightly short of I_E especially at the higher values by the value of the base current.

When $V_{CB} = 0$ it is seen that collector current is flowing. This is because the base itself has a resistance through which the base current flows creating a tiny voltage in such a direction as to allow a small collector current to flow (polarities depend, of course on whether the transistor is npn or pnp). To reduce the collector current to zero it is necessary to reverse the polarity of V_{CB} slightly, in fact making the base-collector junction just forward-biassed.

We would have difficulty in finding the a.c. resistance by measuring any of the curve gradients graphically because they are almost horizontal, showing that a very large increase in V_{CB} is necessary to cause even a small increase in I_C. But from this the a.c. resistance is seen to be very high indeed.

(iii) Transfer characteristic

From the family of output characteristics can be produced the *transfer characteristic* which links input current I_E to output current I_C. At any particular value of V_{CB} on the output characteristics, values of I_C are read off for each value of I_E, for example at $V_{CB} = 5$ V, $I_C \approx 5.9$ mA when $I_E = 6$ mA (unfortunately such small values are not evident in Fig. 2.7(ii)). Plotting I_C against I_E gives the current transfer characteristic as shown in Fig. 2.7(iii).

One adjustment is necessary before current amplification factors can be calculated, it is that I_C includes a very small current, the collector-base leakage current which plays no part in overall amplification because it is not injected via the emitter. This current, which is usually of the order of nA is represented by I_{CBO} (see Sect. 2.2, the O in the subscript shows that the emitter is on open-circuit, hence collector and base are being considered as a diode). Thus the effective output current is $I_C - I_{CBO}$ and

$$\alpha_B \doteq \frac{I_C - I_{CBO}}{I_E} ,$$

typically α_B might be around 0.98 or higher. This is the static or d.c. value of the common-base amplification factor.

From Appendix 4.2 it can be appreciated that the d.c. value may not apply to changing currents so when a small sine-wave or other alternating waveform is applied to the emitter-base circuit we shall be concerned with the changes it causes in both I_C and I_E thus,

$$\alpha_b = \frac{\delta I_C}{\delta I_E}$$

at a given value of V_{CB} (δ = a small change in). The lower case subscript to α (the b) is used to indicate small signals, that is, those which swing the input and output currents only a small proportion of the total possible. It is also clear that for the small signal we are concerned with the gradient of the curve but in this particular case the "curve" is almost a straight line so no tangent has to be drawn and α_b has the same value as α_B.

This leads to the general conclusion that transistors connected in common-base have:

low to very low input impedance

high to very high output impedance
a current gain or amplification factor slightly less than 1,

(the term "impedance" is used since we recognize that there is a small reactive element).

2.2.4 Common-emitter characteristics

Of the three configurations the common-emitter is probably the one most frequently used and because of this much of the data published by manufacturers is for this mode. Thus having examined the common-base circuit in some detail in the previous section we can now be more brief having had some experience with the types of characteristic and their purpose.

(i) Input characteristics

The input is applied between base and emitter (Fig. 2.6(iii)) while the collector voltage is held constant at some predetermined value. The base current is measured over a range of input voltages, a typical silicon characteristic is shown in Fig. 2.8(i). One of our interests is in the input resistance of the transistor as measured by the reciprocal of the slope of the curve at a particular point.(A4.2) An example is shown using the graphical method at 1.0V V_{BE}, point P on the curve. The tangent AC is drawn, the reciprocal of the slope, i.e. AB/BC giving the a.c. resistance, in this instance 0.55 V/2.0 mA = 275 Ω. For interest, following the technique of calculation outlined in Appendix 4.2 the calculated value at 1.0 V V_{BE} is 273 Ω, a good correlation in this case, but we need not be too dismayed over greater discrepancies (up to a few per cent) considering the relatively large spread of values between transistors of the same type.

The input resistance of a transistor in the common-emitter mode is thus appreciably higher than for common-base.

(ii) Output characteristics

The curves for the same type of transistor as in (i) are given in Fig. 2.8(ii), each being measured at a constant value of I_B. The effect of I_B on I_C is shown by the different levels of the curves at any given value of V_{CE}. Compared with common-

base the curve slopes are greater, therefore their reciprocals which represent output resistances indicate only moderately high values, for example for the curve for I_B = 0.2 mA, the a.c. resistance above about V_{CE} = 10 V is some 4000 Ω, less for the curves above.

(iii) Transfer characteristic

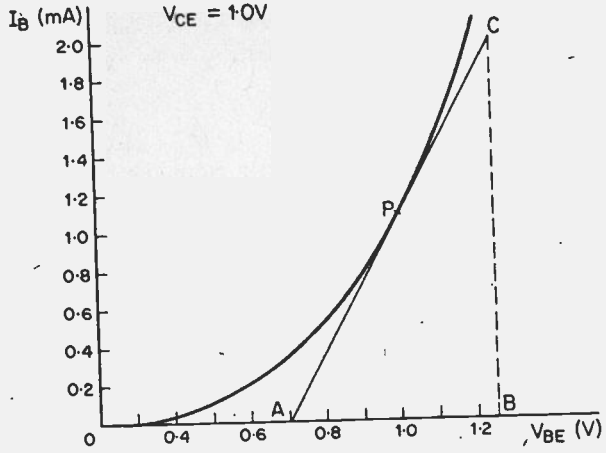
This is quite easily obtained directly from the output characteristics. V_{CE} is chosen and working for example at 15 V, we read that at I_B = 0.2 mA, I_C = 53 mA, the other values of I_C moving up the curves are 118, 182 and 245 mA. These are plotted in Fig. 2.8(iii). The characteristic is more curved than for common-base, hence α_E will not be the same as α_e and both vary with the value of I_B. Note the very different scales when compared with Fig. 2.7(iii), at I_B = 0.5 mA for example, α_E is I_C/I_B = 118/0.5 = 236, a very different current amplification factor compared with common-base. Considering that input and output resistances are both in the medium category it is easier to see for common-emitter how amplification arises because a tiny signal applied to base and emitter will give rise to a signal of similar shape but much greater amplitude between collector and emitter. The calculation of α_E is slightly approximate because no account has been taken of leakage currents.

By graphical construction at I_B = 0.5 mA, α_e = 185, somewhat less than the value for α_E.

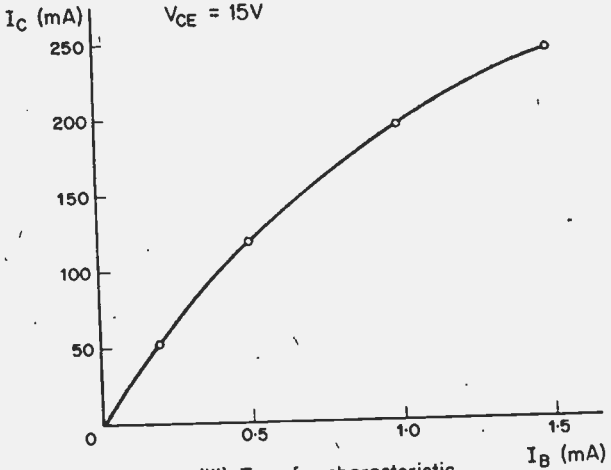
These figures are, of course, for demonstration only. They refer to one particular type of npn silicon transistor out of the very many different types available. Thus we need take no note of the figures themselves, only how they instruct us in reading important facts about a transistor from its static characteristics. These are available from the manufacturer or can be measured.

For the common-emitter mode therefore the figures show:

medium values of input impedance
medium values of output impedance

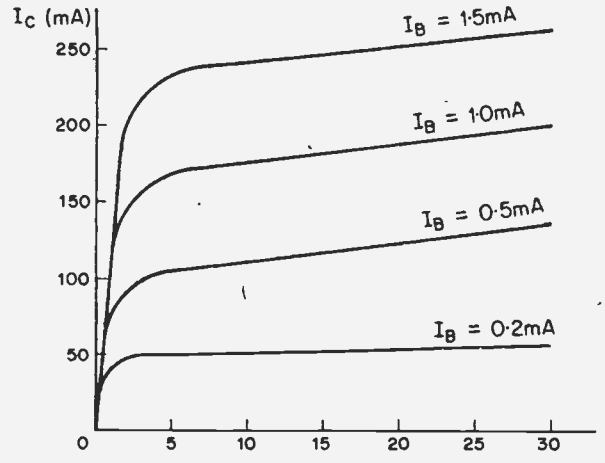


(i) Input characteristic.



(iii) Transfer characteristic.

Fig. 2.8 Typicol common-emitter characteristics.



(ii) Output characteristics.

a current gain or amplification factor very much greater than 1.

2.2.5 The relation between the current amplification factors

In this section an approximate but useful relationship is developed. Ignoring the small leakage currents, for a given transistor:

$$\alpha_E = \frac{I_C}{I_B}$$

at a given value of V_{CE}

$$\therefore \quad \frac{1}{\alpha_E} = \frac{I_B}{I_C} \quad \text{and since} \quad I_B = I_E - I_C$$

[The transistor current equation gives $I_B = -(I_E + I_C)$

but we ignore current directions otherwise we end up with an unnecessary complication that α_B is negative. Thus we simply say, quite correctly, that I_B is the difference between I_E and I_C.]

$$\frac{1}{\alpha_E} = \frac{I_E - I_C}{I_C} = \frac{I_E}{I_C} - \frac{I_C}{I_C} = \frac{I_E}{I_C} - 1 .$$

But

$$\frac{I_E}{I_C} = \frac{1}{\alpha_B}$$

for the same collector voltage

$$\therefore \quad \frac{1}{\alpha_E} = \frac{1}{\alpha_B} - 1 = \frac{(1 - \alpha_B)}{\alpha_B}$$

$$\therefore \quad \alpha_E = \frac{\alpha_B}{1 - \alpha_B}$$

and if the same procedure is repeated for the small signal case, i.e. using

$$\alpha_e = \frac{\delta I_C}{\delta I_B} \quad \text{etc.,} \quad \alpha_e = \frac{\alpha_b}{1 - \alpha_b}$$

Typically, if in common-base the current amplification factor α_B or α_b is 0.99, then α_E or α_e when the transistor is reconnected for use in the common-emitter mode is

$$\frac{0.99}{1 - 0.99} = 99 .$$

2.2.6 The common-collector circuit

This is shown in Fig. 2.6(iv) and although the collector is not

connected directly to the common, it is effectively so as far as an alternating signal is concerned since arrangements are made to keep the battery or power supply impedance very low. The battery in series with R_L forward biasses the emitter-base junction.

Because the characteristics of the common-collector circuit are not as suitable in most amplifiers as are those of common-base and common-emitter, this analysis is in summary form only, the principles outlined previously apply.

The output voltage generated across the load resistance R_L tends to follow the input voltage so closely that the voltage gain is approximately unity. For this reason the circuit is also known as an *emitter follower*. There is, however, a current gain as with common-emitter because a small base current is controlling a much larger emitter (output) current. The collector-base junction is reverse-biassed, therefore the input resistance is high; the base-emitter junction is forward-biassed so the output resistance is low. This confirms the absence of voltage gain because although there is a current gain, it operates in a low-resistance output circuit and can only give rise to a low voltage.

The common-collector circuit is usually used for impedance matching, for example from a high-impedance source to a low-impedance load, when so used it avoids the power loss of resistive or transformer matching.

2.3 EFFECTS OF TEMPERATURE

So as not to complicate matters in the preceding sections the effect of junction temperature on a transistor has been mainly bypassed. Temperature is of great importance however because it does place restrictions on the operation of semi-conductors. Both diodes and transistors are affected but we examine the latter in greater depth because having two junctions the effects of temperature are slightly more involved.

All the static characteristics change with temperature because of thermally generated minority carriers, some change considerably so the circuit designer must take into account both internal and external temperature conditions. Heat is also generated within the junctions through power dissipation and although the power in many cases is very small, so too is the physical size of the junction, accordingly temperature rises may be quite significant. In small components heat is removed from the junction by conduction to the case, especially along the connecting wires. Where this proves inadequate, *heat sinks* are employed, of various forms but usually a fin or blade of metal fixed to the case so that heat is conducted out of the device and dissipated in the surrounding air. Frequently heat is conducted away by the metal chassis on which the component is mounted.

Considering the power dissipated first, it is totalled from voltage times current for both the junctions, i.e.

Total power dissipated =

$$P_{tot} = (V_{BE} \times I_B) + (V_{CE} \times I_C)$$

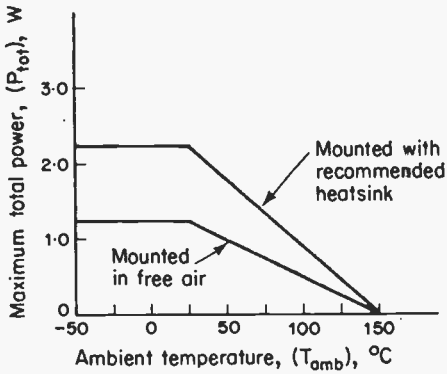


Fig. 2.9 Transistor total dissipation rating curves.

The limited power dissipation may be shown by a manufacturer by *total dissipation rating curves*, typically as shown in Fig. 2.9 from which it is seen that above 25°C the total dissipation allowed falls as ambient (surrounding) temperature increases until a limit is reached at 150°C. The upper curve demonstrates how a much greater power may be generated within the device when a heat sink is employed.

The curves are obtained through measurement of temperature gradients across heat conduction paths and substituting these in the formula:

$$P_{tot} = \frac{T_j - T_{amb}}{\theta}$$

where T_j is the semiconductor junction temperature in °C, T_{amb} is the ambient temperature in °C and θ is the *thermal resistance* of the heat loss path from junction to surrounding air.

θ needs more explanation. It is called thermal *resistance* because it has the nature of resistance in the Ohm's Law type of relationship where power is analogous to current and temperature to voltage.

We understand the generation of heat by expending power, this is the opposite, the removal of heat by subtraction of power either by conduction along some metal or material or by warming the surrounding air. Heat flows from a body at a higher temperature to one at a lower temperature and the greater the temperature difference, the greater the *rate* of flow. Thus the temperature difference needed to cause a heat flow rate of one watt is a measure of the thermal resistance of the path. It is measured in the unit θ, °C/W and evidently a path needing 10°C temperature difference between its two ends to cause a heat flow rate of 1 W has ten times the thermal resistance of a path needing only 1°C difference. In the reverse direction θ could be defined as the temperature rise in °C per watt of input power. Visualized in more practical

terms it becomes painfully obvious to the user of a soldering iron that the heat flow rate from the working end is greater than that from the handle, in terms of the thermal resistance θ, the handle has the much higher value.

Evidently if thermal resistance is of such a form that it can be substituted in an Ohm's Law type of equation, then separate paths may be considered in series and parallel in the normal resistance manner.

Considering a transistor such as shown in Fig. 1.9, θ consists of two paths in series, θ_{j-c} from the junction to the transistor case and θ_{c-amb} from the case to the air. A heat sink surrounding the case places a path in parallel with θ_{c-amb}.

From Fig. 2.9 it is evident that the maximum junction temperature allowed in this particular case is 150°C for above that temperature no power at all may be dissipated within the transistor.

At 25°C for the free-air curve and T_j at a maximum of 150°C

$$P_{tot} = \frac{T_j - T_{amb}}{\theta_{j-amb}}$$

$$\therefore \quad 1.25\ W = \frac{150° - 25°}{\theta_{j-amb}}$$

$$\therefore \quad \theta_{j-amb} = \frac{125}{1.25} = 100°C/W$$

meaning that a temperature difference of 100°C is needed across the thermal resistance between the junction and the surrounding air (θ_{j-amb}) to maintain a heat flow rate of 1 watt.

Thus by knowing the maximum temperature to which a transistor junction can be raised without fear of its destruction and the appropriate values of θ from laboratory temperature measurements, such curves as in Fig. 2.9 may be produced. For example if the ambient temperature is 100°C and the heat loss path is unchanged,

$$P_{tot} = \frac{150 - 100}{100} = \frac{50}{100} = 0.5 \text{ W}$$

as shown on the curve. For the heat sink curve, at 25°C

$$2.25 \text{ W} = \frac{150° - 25°}{\theta_{j-amb}}$$

(θ_{j-amb} now contains the resistance of the heat sink path in parallel)

$$\therefore \quad \theta_{j-amb} = \frac{150 - 25}{2.25} \text{ °C/W} = 55.6\text{°C/W}$$

showing how well the heat sink has reduced the total thermal resistance.

EXAMPLE:
A transistor which should not be operated at a junction temperature exceeding 175°C can be used in free air maintained at 20°C when the thermal resistance of the junction-to-air path is 170°C/W. If used with a heat sink the thermal resistance is 90°C/W. It is proposed to use the device at a total power dissipation of 2.0 watts. Is the heat sink necessary?

In free air

$$P_{tot} = \frac{T_j - T_{amb}}{\theta} = \frac{175 - 20}{170} = 0.91 \text{ W}$$

clearly 2 W will produce too high a junction temperature.

With heat sink

$$P_{tot} = \frac{175 - 20}{90} = 1.7 \text{ W}$$

i.e. even with a heat sink the heat transference is too small. In fact at 2 W the junction temperature would become (with heat sink)

$$2 = \frac{T_j - 20}{90}$$

$$\therefore \quad 180 = T_j - 20$$

$$\therefore \quad T_j = 200^\circ\text{C},$$

well above the limit.

It follows that the considerations above of the *derating* of a device (that is, reduction of its operating range) as ambient temperature and/or thermal resistance to heat loss from the junction increase, applies equally to diodes as to transistors. Diodes have only one junction to consider, hence the calculation of junction dissipation is less complicated.

2.4 SEMICONDUCTOR CAPACITANCE

In Fig. 1.3(ii) which shows the conditions within an unbiassed diode, we recall that the depletion layer is a region with few current carriers and has positive and negative charges on the two sides. Because the depletion layer is therefore effectively a partial insulator the impression is given that within the diode there is the equivalent of a charged capacitor since there are two opposing charges separated by a dielectric. Furthermore, by moving on to Fig. 1.5(i) showing the diode with reverse bias it is seen that the width of the depletion layer has

increased and the capacitance would then be expected to fall because effectively the thickness of the dielectric has increased

$$\left(C \propto \frac{\text{area of plates}}{\text{thickness of dielectric}} \right)$$

These effects are confirmed in practice and Fig. 2.10 shows a typical capacitance curve for the collector-base junction of a transistor where the fall in capacitance as reverse-bias (V_{CB}) increases from zero is demonstrated. The capacitance is small (a few picofarads) but the reactance can be quite a disadvantage at the higher frequencies as shown by way of example on the graph.

Alternatively, a capacitance which can be varied electronically for example, by a changing voltage, has many uses. A common use is in automatic tuning of radio and television

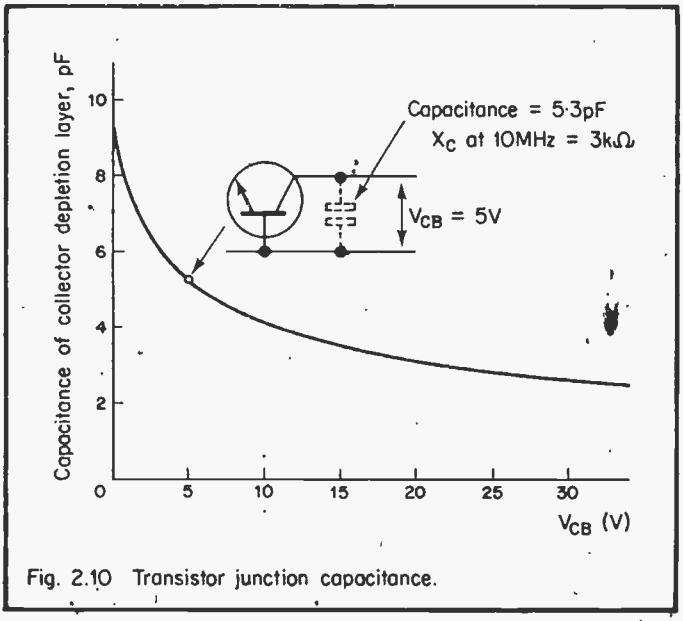


Fig. 2.10 Transistor junction capacitance.

receivers in which arrangements are made that slight mistuning gives rise to a voltage. This is fed to a variable capacitance (tuner) diode (a device deliberately designed to make use of the junction capacitance effect) which changes the tuning of the receiver accordingly.

When relatively high capacitances are required, the bias is changed to forward whereupon the depletion layer is reduced, so increasing the capacitance.

2.5 FIELD-EFFECT TRANSISTORS (F.E.T.s)

So far we have only discussed the standard *bipolar* transistor, so named because it has both positive and negative carriers (holes in p-type, electrons in n-type, see Fig. 1.3). There is another type of transistor with somewhat different operating principles, it is known as a *field-effect transistor* and its current path is not through both p and n materials but wholly in one. N-type material is more commonly used because the mobility of its carriers is higher than for p-type (electrons have considerably less mass than holes and therefore accelerate more quickly). The type is known as *n-channel* and the reason for the term "channel" will be appreciated from consideration of Fig. 2.11 which shows a field-effect transistor diagrammatically.

Current flow is along a bar of n-type semiconductor material from *source* to *drain* (somewhat analogous to emitter and collector in a bipolar transistor) under normal conduction principles. The free electrons within the n-type material (obtained from the donor impurity atoms) are repelled from the source and attracted towards the drain because of the applied potentials. However, the p-type *gate* electrodes form pn junctions as shown and associated with these are depletion layers as we see in Fig. 1.3. Since the d.c. potential along the bar rises positively from source to drain the depletion layers are not of constant width but increase on the drain side because the potential between p and n is greater there.

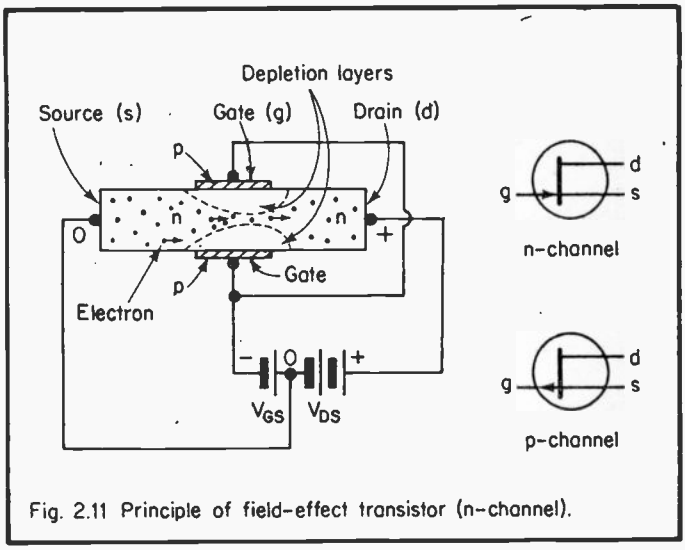


Fig. 2.11 Principle of field-effect transistor (n-channel).

Now we note from Fig. 1.5(i) that the pn junctions are reverse-biassed and because the width of the depletion layer varies with the magnitude of the bias, as V_{GS} becomes more negative the effective width of the channel and therefore electron flow decrease. In Ohm's Law terms, for a given voltage between drain and source (V_{DS}), the resistance of the F.E.T. increases hence the drain current I_D falls as the negative potential V_{GS} increases. Thus as with a bipolar transistor a relatively large current is dominated by a small voltage. But there is one important difference, the F.E.T. control is by application of an electric field, fields are set up by extremely small currents hence the gate input resistance is very high (in the MΩ range), a property often needed when the device must absorb practically no power from the input circuit.

As V_{GS} increases the expansion of the depletion layers ultimately becomes such that the channel is constricted almost to closure and the current is then nearly independent of drain-source voltage. The output characteristics of Fig. 2.12 show this effect where above a certain value of V_{DS}, I_D changes

little, while the adjacent transfer characteristic shows how at $V_{DS} = 10$ V, I_D is reduced to 0 at $V_{GS} \simeq -5.8$ V, the latter is known as the *pinch-off* voltage and occurs when the depletion regions meet. The F.E.T. is therefore a unipolar device with output characteristics very similar to those of the bipolar transistor but with a much higher input impedance.

P-channel F.E.T.s are based on similar principles.

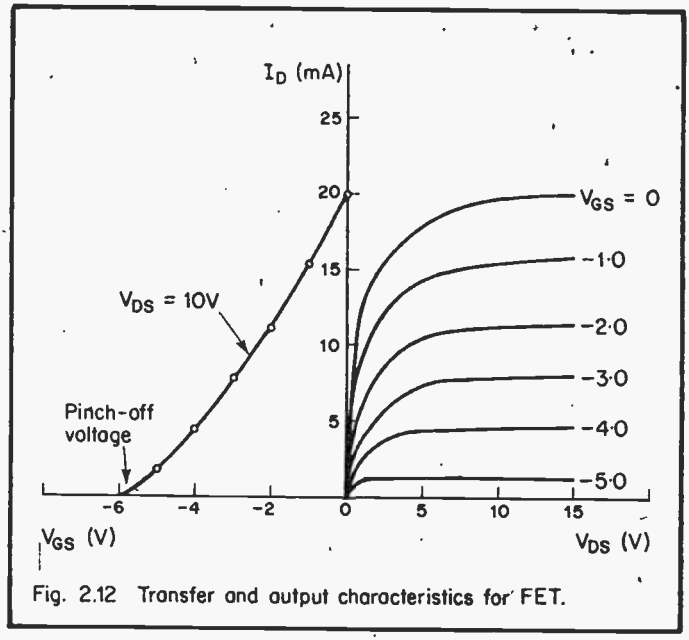


Fig. 2.12 Transfer and output characteristics for FET.

2.5.1 The metal-oxide-semiconductor transistor (MOST) This works on an extension of the F.E.T. principle in that by insulating the gate from the channel, even higher input impedances are obtained, many thousands of megohms in fact. It is known as an *insulated-gate field-effect transistor* (IGFET) but more commonly as a *metal-oxide-semiconductor transistor* (MOST).

By planar and diffusion technology as in Section 1.6.1 a MOS transistor may be constructed as shown in Fig. 2.13 and the principle of operation can be seen from this. The particular drawing is for an induced n-channel type but opposite polarities are equally used, especially in CMOS (Complementary MOS) where one n-channel is paired with one p-channel transistor to provide special facilities in switching circuits.

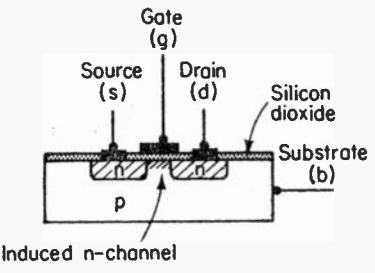


Fig. 2.13 Construction of metal-oxide semiconductor transistor.

A p-type silicon wafer has source and drain electrodes formed by n-type diffusion with the result that between them are two separate pn junctions in opposition, thus irrespective of the polarity across the pair one is always reverse-biassed and practically no current can flow between source and drain. If however a positive potential is connected to the gate, the electric field due to the charge on it attracts electrons to the surface of the p-type channel, changing it more and more to n-type according to the magnitude of the potential. Thus to a certain extent the pn junctions are destroyed, the channel becoming a continuous n-type and therefore conductive. The F.E.T. principle applies in that the voltage on the gate controls the current flow through the channel beneath it. The output characteristics have the general F.E.T. form.

3. BASIC SYSTEMS

In this chapter we examine in more detail the four basic "systems" of Chapter 1, Section 4. Having studied the main semiconductor components, diodes and transistors, we are now in a position to remove the lids of the black boxes to see what they contain.

3.1 RECTIFIERS

The general principles of rectification are already evident from firstly our earlier skirmish with Fourier Analysis of a rectified wave (Book 2) and secondly from what has been seen of semiconductor diodes in Chapter 2 in that they simply do not let current through when biassed in the reverse direction and hence are unidirectional devices. This is the basis of all rectification, which might be defined as the transformation of alternating current into some sort of equivalent direct current and as shown in Fig. 1.4(i) we expect the d.c. output of a system to be a straight horizontal line on the voltage/time graph. The ideal characteristic for the diode is simply zero resistance forward, infinite in reverse, this cannot be realized therefore how closely the ideal system output characteristic is approached depends on the cost of the circuit components other than that of the diodes themselves.

3.1.1 Power rectification

The power mains fed to homes, factories and in fact practically everywhere, are sine-wave, alternating in most countries at a frequency of 50 or 60 Hz. The supply is alternating mainly because of the ease with which transformers can change the voltage compared with direct voltages which are not so conveniently changed. Change of voltage results in economy of transmission for it needs smaller conductors to transmit a certain amount of power at high voltage, low current than at low voltage, high current (it is the magnitude of the current which determines the size of the conductor). It is true that high-voltage transmission brings its own problems but reducing

the cost of cables takes precedence. The consumer generally can accept power in an alternating form, heating equipment for example is indifferent as to the direction of current, it produces heat either way. Filament lighting is the same because most of it is via heat. However there are some systems which need d.c., transistor equipment is an example, telephone exchanges and some traction systems are others and for these semiconductor rectifier diodes are very suitable. They are available with forward currents in excess of 1000 A and by using several units in parallel can cater for most requirements.

3.1.1.1 The half-wave circuit

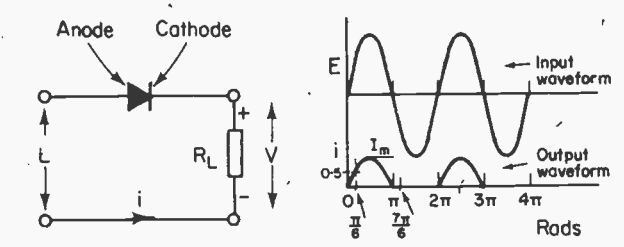
The simplest but least efficient circuit is of a single diode in series with the supply, as in Fig. 3.1(i). The forward resistance changes with applied voltage and there is a small current in the reverse direction, however for simplification the graphs of Fig. 3.1 show rectified forms as though the diodes were ideal.

E represents the r.m.s. value of the power supply sine wave alternating voltage transformed as necessary to such a value that after rectification the required voltage V is obtained. The instantaneous current i is shown on the graph, there is a half-wave of current when E is positive at the diode anode, none when E is negative.

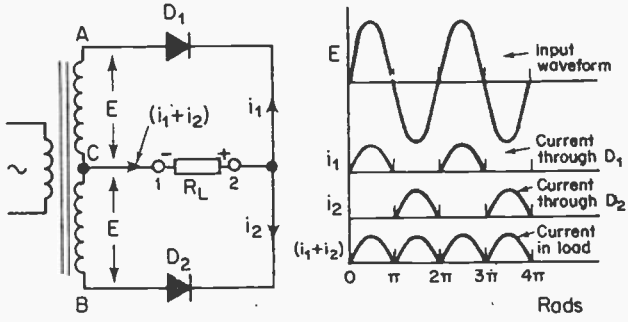
Earlier we had a look at the way in which a wave can be analysed into its components by the technique of Fourier Analysis. The Fourier equation to the half-wave rectified sine wave has been shown as:

$$e = \frac{2E_m}{\pi}\left(0.5 + \frac{\pi}{4}\sin \omega t - \frac{\cos 2\omega t}{3} - \frac{\cos 4\omega t}{15} - \frac{\cos 6\omega t}{35} - \ldots\right)$$

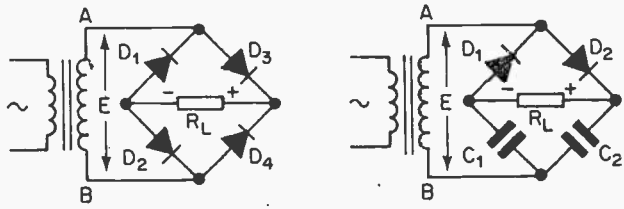
and perhaps it is as well to recall what this awesome looking



(i) Half-wave.



(ii) Full-wave



(iii) Bridge.

(iv) Voltage doubler.

Fig. 3.1 Rectifier circuits.

expression tells us. We can substitute i and I_m for e and E_m, so expanding the expression and numbering the terms:

$$i = \underset{(1)}{\frac{I_m}{\pi}} + \underset{(2)}{\frac{I_m}{2} \cdot \sin \omega t} - \underset{(3)}{\frac{2I_m}{3\pi} \cdot \cos 2\omega t}$$

$$- \underset{(4)}{\frac{2I_m}{15\pi} \cdot \cos 4\omega t} - \underset{(5)}{\frac{2I_m}{35\pi} \cdot \cos 6\omega t} - \ldots$$

Each of the terms is the equation to the graph of one of the components of the output waveform of Fig. 3.1(i) and we may perhaps revive our confidence in the Fourier equation by calculating the total value of the components at two points to see if they add up as expected, bearing in mind that there is no need to consider all the components to infinity because the higher the harmonic, the less is added to the final result. Let I_m = 1 A.

At $\pi/6$ radians (30°), Fig. 3.1(i) shows the expected result to be 0.5 x I_m = 0.5 mA and at $7\pi/6$ radians (210°) we expect 0.

Equation term	(1)	(2)	(3)	(4)	(5)	SUM (= i amps)
$\omega = \pi/6$ rads	0.318	+0.250	−0.106	+0.021	+0.018	0.501
$\omega = 7\pi/6$ rads	0.318	−0.250	−0.106	+0.021	+0.018	0.001

Considering that we have used no more than five terms, the result of adding together the instantaneous values of these components is very close to what is expected from the output waveform. Perhaps a tedious but nevertheless instructive process.

The second term is at the power frequency, all subsequent terms represent even harmonics, neither the power frequency nor the harmonics adding to the rectified current. This is therefore given by the first term only, that is, I_m/π, the only term with no variation in its value, which therefore represents the direct current, where I_m is the maximum value of i. [It is understood that there must be a d.c. path through the power source, usually a transformer winding, for a current to flow.]

By chopping off half of the waveform it is being changed to a considerable degree, so we realise why the harmonic content of the diode output is high. If the d.c. is used for battery charging, harmonics are ineffective but for certain other systems Fourier analysis proves its worth for the harmonics it predicts are very real and active. With a d.c. supply of this type to transistor amplifiers, the harmonics become amplified and interfere with the legitimate signal. In the output of an electronic instrument harmonics are classed as music but when unwanted as in this case we label them *noise*. Undesirable harmonics can be reduced to an ineffective level by *filter* circuits, to be discussed later.

Because the harmonics are not d.c. they represent the inefficiency in the system so let us consider the efficiency of the half-wave rectifier in more detail:

The d.c. power output is equal to

$$\left(\frac{I_m}{\pi}\right)^2 \cdot R_L$$

Considering the power supply input, if the maximum current when the diode is conducting is I_m, then the r.m.s. value

$$I = \frac{I_m}{\sqrt{2}}$$

and the power is

$$I^2R_L = \frac{I_m^2}{2} \cdot R_L .$$

This occurs for only half the time of a complete cycle, therefore the mean input power is equal to

$$\frac{I_m^2 R_L}{4} .$$

Knowing both input and output powers enables us to calculate the efficiency of a system as

$$\frac{\text{d.c. power output}}{\text{a.c. power input}} \times 100\%$$

$$= \frac{I_m^2}{\pi^2} \cdot R_L \times \frac{4}{I_m^2 . R_L} \times 100\% = \frac{400}{\pi^2} = 40.5\%$$

clearly an inefficient, but of course inexpensive circuit.

Care must be taken with any diode that the maximum peak reverse voltage is not exceeded otherwise avalanche breakdown occurs. This value is usually quoted by the manufacturer as V_{RRM}, the *repetitive peak reverse voltage*, sometimes as PIV, *peak inverse voltage.* In the case of the half-wave circuit the value is equal to E_m.

3.1.1.2 The full-wave circuit

A transformer is shown for this circuit in Fig. 3.1(ii) since by centre-tapping its secondary winding two equal voltages in antiphase are available. Suppose for example that the whole winding AB has 100 V induced in it with A positive to B. From the point of view of the centre-tap C which is connected to terminal 1 of the load, A is +ve 50 V and B is –ve 50 V,

hence D_1 conducts and D_2 is reverse-biassed and therefore not conducting. When the input waveform reverses so that A is −ve to B, D_2 conducts while D_1 does not. Currents i_1 and i_2 flow on alternate half-cycles in the same direction through R_L as shown in the graphs and the net result $(i_1 + i_2)$ is shown at the bottom. Thus the title of the circuit becomes apparent for the full input wave has been rectified as compared with the half-wave previously described.

Fourier analysis shows that there is no component in the rectified output at the supply frequency, the lowest harmonic is the second, that is, at double the supply frequency.

Without going into a mathematical proof again it is perhaps obvious that both the direct current and the a.c. input power are double that of the half-wave circuit, leading to a circuit efficiency of

$$\left(\frac{2I_m}{\pi}\right)^2 R_L \times \frac{2}{I_m{}^2 R_L} \times 100\% = \frac{800}{\pi^2} = 81\%$$

twice that of the half-wave circuit.

V_{RRM} can be deduced from consideration of, say, diode D_2 which is reverse-biassed when C is +ve to B by the maximum voltage E_m. At this instant A is +ve to C by E_m, hence A is +ve to B by $2E_m$. Now because diode D_1 is forward-biassed the +ve potential at A is applied directly to the cathode of D_2, hence the peak reverse voltage across D_2 is $2E_m$.

3.1.1.3 The bridge circuit

This is again a full-wave circuit but one which has the advantage of requiring no centre-tapped transformer winding. It uses four diodes connected as shown in Fig. 3.1(iii). Consider a point in the input wave cycle when A is +ve to B. Current flows from B through D_2, R_L and D_3 back to A. When the input wave reverses, current flows from A, through D_1, R_L and D_4 back to B. The current through R_L is unidirectional. On each half-cycle two diodes in series are forward-biassed

the remaining two are reverse-biassed. The output waveform is similar to that for the full-wave circuit of Fig. 3.1(ii).

For any diode which is reverse-biassed, it is seen to be in series with another in the forward direction, across the supply voltage, E. Since the second diode has low resistance, then the full voltage E_m effectively appears across the reverse-biassed diode at the peak of the wave, i.e. $V_{RRM} = E_m$.

3.1.1.4 Voltage doubling

A circuit which again does not require a centre-tapped transformer, yet can double the output voltage compared with the previous ones is shown in Fig. 3.1(iv). The cost of this facility is in the requirement of the two capacitors, C_1 and C_2. The circuit is particularly suitable where a higher voltage at lower current is required. Voltage multiplication to a greater degree (x3 or x4) is possible but the circuits become more complex and recourse to a greater step-up in the transformer instead may be more economical.

Considering the figure, when A is –ve to B, current flows through D_1 and charges up C_1. D_2 is reverse-biassed and contributes no current. On the alternate half-cycle D_2 and C_2 are in operation. The actual charging process is complicated for although we have met all the basic formulae, there is a changing time constant to contend with (the diode resistance varies with forward voltage) and moreover a changing input voltage. So we leave well alone, but all this actually adds up to the simple fact that after a few cycles both capacitors charge up to a voltage approaching E_m. They are charged in series-aiding, hence the total voltage across the two capacitors approaches $2E_m$. This is the voltage applied across the load, R_L. The amount by which the voltage falls short of $2E_m$ depends on the load current drain for the particular capacitance values employed, these must therefore be high.

A reminder about current direction in Fig. 3.1. Although we affirm that current is the flow of negative electrons away from a –ve charge to a +ve one, the symbol for the diode may

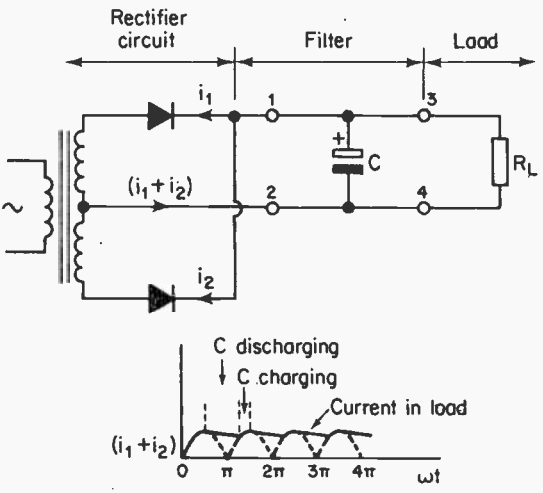
confuse us by telling the opposite. We have to live with this. One thing is worth noting however which is that in all rectifier circuits the diode symbol points to the positive terminal of the load.

3.1.1.5 Filters

It is evident from the foregoing that a.c. components in a rectifier output are an embarrassment both to efficiency and also in many cases to *noise-free* working of semiconductor circuits which the d.c. energizes. These components can be minimized by use of *filter* circuits of varying complexity, the simplest being either a single parallel capacitor which bypasses the alternating components or a series inductance which presents a high impedance to them, neither the capacitor nor the inductor having an appreciable effect on the d.c. Through the use of either of these the *ripple* on the d.c. output is greatly reduced but even more reduction is possible by using combinations of both series inductance and parallel capacitance, known as an LC filter.

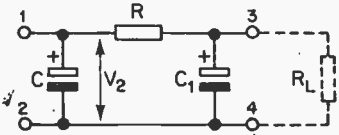
The single capacitor filter

Frequently used because of its simplicity and cheapness this consists of one large-value capacitor connected across the rectifier output. Because the polarity across the load is fixed, advantage can be taken of the lower cost of an electrolytic capacitor which has high capacitance because of a very thin chemical dielectric. The correct polarity across the capacitor is necessary to maintain the dielectric film, reversal of polarity is likely to destroy it with resultant breakdown. Such capacitors are suitably marked to indicate the positive terminal and may have capacities running into many thousands of microfarads. As an example, Fig. 3.2(i) shows the full-wave circuit redrawn to separate rectifier and filter circuits and where C is the filter capacitor. The current in the load now differs appreciably from that in Fig. 3.1(ii) because when the input voltage falls from its peak to a value below the voltage of the capacitor, then the latter supplies current from its charge to the load. Hence the one diode which is conducting at the time only supplies current during that part of the cycle marked "C charging". The result is a waveform across the

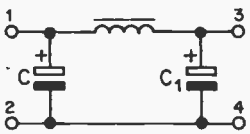


(i) Single capacitor.

(ii) Single inductor.



(iii) Resistance – capacitance.



(iv) Inductance – capacitance.

Fig. 3.2 Rectifier filters.

load as shown, evidently much *smoother* than without the capacitor. C is frequently referred to as a *reservoir* capacitor.

The single inductor filter

This is shown in Fig. 3.2(ii) but is less likely to be used on its own because its cost far exceeds that of a capacitor for equal reduction of ripple. Technically it works as a store as the capacitor does, not storing charge but magnetic energy. When the voltage applied by the diodes exceeds that across the load, energy is stored in the magnetic field of the inductor, when the voltage falls below, some of the field collapses and supplies energy to the load, the load variations are therefore smoothed.

The resistance–capacitance filter

A series resistance and second parallel capacitor following the main reservoir capacitor C provides additional smoothing (Fig. 3.2(iii)). In the example, the full-wave circuit lowest harmonic is the 2nd so the value of R is chosen to be reasonably high compared with the reactance of C_1 to this harmonic

$$\left(X_C = \frac{1}{\omega C_1} = \frac{1}{2\pi \times 2fC_1} = \frac{1}{4\pi fC_1} \right.$$

where f is the supply frequency.)

Considering the 2nd harmonic ripple voltage V_2 across the reservoir capacitor C and assuming that the reactance of C_1 at the 2nd harmonic frequency 2f is considerably lower than the resistance of the load R_L, then clearly the 2nd harmonic current due to V_2 will be bypassed from the load by the low reactance path of C_1 in parallel. As an example, a load current of 20 mA at 10 V implies a load resistance of 500 Ω, at a power supply frequency of 60 Hz, the reactance of C_1 to the 2nd harmonic is only 5 Ω for a value of just over 250 μF, indicating a reduction in 2nd harmonic ripple in the load to about one-hundredth. Greater reductions are possible with larger values of C_1.

Alternatively one might look at the harmonic reduction from the point of view of the series resistor R which feeds into a low impedance at the 2nd harmonic frequency due to the addition of C_1. At this frequency the current is therefore greater and hence the voltage drop across R according to its value. There is a compromise limit to raising the value however since R causes an undesirable d.c. voltage drop when load current flows.

What has been said about the 2nd harmonic applies even more to higher order harmonics for which the reactance of C_1 falls accordingly.

The inductance–capacitance filter
The limitation imposed on making the value of R high as above is overcome by use of an inductor which provides high series reactance to the harmonics yet has low resistance to the d.c. component. Combined with the two capacitors this filter is capable of reducing ripple to a satisfactory level for the most stringent conditions. Introduction of an inductor naturally increases the cost.

3.1.1.6 Voltage regulation
When a stable direct voltage supply is required irrespective of load current or supply voltage changes within certain limits, the Zener diode (Sect. 2.1.3) may be used, usually as a parallel element across the supply V_S as shown in Fig. 3.3. The series resistance R prevents excessive current flowing through the diode through its greater voltage drop when the diode current tries to increase. For circuit efficiency the diode current should be comparatively low compared with the maximum load current, for the diode current is in effect the price paid for regulation. Also the value of R should be as low as possible to minimize its power loss.

A typical practical example follows based on the values shown in brackets in the figure. When the load current I_L is maximum we make the diode current I_z minimum, then if load current falls, the diode current increases to compensate. The total current $I = (I_L + I_z) = 18 + 2 = 20$ mA.

The voltage, V across R $= V_S - V_L = 20 - 9 = 11$ V

$$\therefore R = \frac{11}{20} \times 1000 = 550\ \Omega$$

i.e. for regulation, with V_S at 20 V the current through R must be maintained at 20 mA to provide $V_L = 9$ V (irrespective of variations in I_L).

We have set the conditions so that the diode is not worked below its minimum current. It is also necessary to check that the diode chosen is not taken above its maximum current. This could occur when the load current is minimum, in this case 4 mA.

Then $I_Z = I - I_L = 20 - 4 = 16$ mA.

The maximum power dissipation for the diode is 400 mW, therefore at 9 V

$$I_Z(\text{max}) = \frac{0.4}{9} \times 1000 = 44\ \text{mA}$$

The diode current in the circuit is therefore well below the maximum allowed.

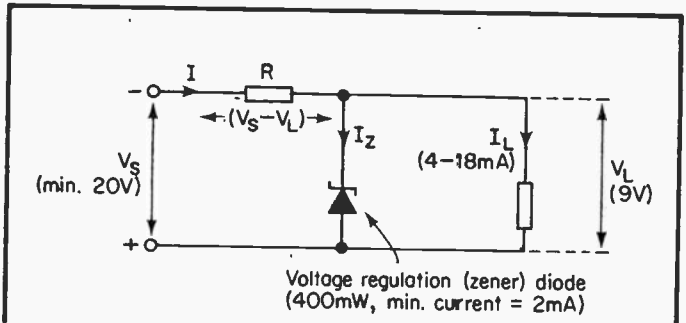


Fig. 3.3 Voltage regulation using zener diode.

V_S is quoted as 20 V minimum, should this voltage rise there will be an increased current through the diode with only a very small increase in its voltage, the voltage drop across R increases accordingly, hence V_L remains almost constant.

Many other more complex voltage regulation circuits are used, depending on the degree of regulation required, the Zener diode method has the great advantage of moderately good regulation for a simple, low cost circuit.

3.1.2 Demodulation

There is a further rectification process which has nothing to do with power supplies yet should not be overlooked for the process itself leads to an insight into the basic principles of radio communication. It is known as *demodulation* or *detection.* However, before we can demodulate a wave, it has to be modulated so we consider the latter term first.

We start with some form of *information*, a general term covering speech, music, television, morse, telegraph or computer codes, in fact anything which informs and which can be converted to an electronic signal whatever shape or form that may take. This has to be transmitted to some distant point.

We are already accustomed to the graceful elegance of the sine wave and we continue to use this for analysis even though most forms of information (such as speech and music) when displayed on an oscilloscope appear as a mass of agitated and entangled frequencies. This is legitimate because our considerations about transmitting a sine wave by radio apply equally to the other forms.

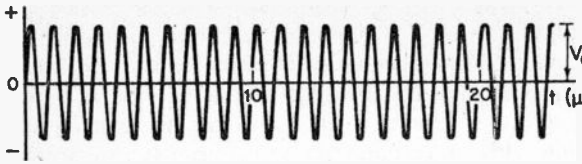
Modulation is the process by which information is impressed on a *carrier* wave, this being at a higher frequency than the information signal and used for transmission either by radio or cable. The peak value, frequency or phase of the carrier wave may be modified by the signal. The first is known as *amplitude modulation* and is the one we consider, frequency modulation (FM) and phase modulation are equally important

but are not necessary in a simple discussion of rectification.

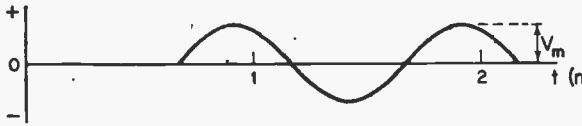
Fig. 3.4 shows at (i) an *unmodulated* carrier wave, to give it some practical significance, assume it to be at a frequency of 1 MHz, somewhere in the middle of the radio medium-wave band. It is a sine wave so that it carries no harmonics to interfere with other radio transmissions. It carries no information except whether it is being transmitted or not.

Now, suppose we wish to transmit a 1 kHz *tone* (e.g. as used for the Morse Code) as shown in (ii) [Fig. 3.4 demonstrates the principles only, to one cycle of tone there are 1000 cycles of carrier wave, this cannot possibly be shown on a small diagram, nor can sine waves be accurately drawn.] A frequency of 1 kHz cannot be transmitted by radio so it is impressed on the carrier wave in the modulator circuit at the transmitter. The amplitude of the carrier wave then varies according to the amplitude of the tone as shown in (iii). For a tone having a peak value V_m volts and a carrier of V_c volts, the variation of the carrier amplitude on modulation is between $(V_c + V_m)$ and $(V_c - V_m)$. In this form the wave is transmitted to the receivers.

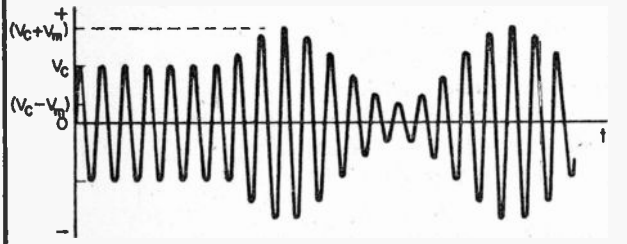
A radio receiver is capable of selecting (tuning) a particular carrier wave and amplityfing it to around one volt for application to a *diode detector*. This is usually a half-wave circuit [as in Fig. 3.1(i)] followed by a single capacitor and load resistance. The reactance of the capacitor is high at the modulating frequency but low to the carrier frequency thus bypassing the latter but developing the modulation voltage across it. The rectified wave is shown at (iv) and it is evident that half-cycles of the carrier frequency are present, as would be expected from Section 3.1.1.1, the filtering by the capacitor results in a waveform as shown in (v). This has the appearance of the original modulating frequency but not only does it fluctuate with traces of the carrier (exaggerated in the diagram) but also it is lifted above the X (time) axis. The fluctuations may be minimized by changing to a resistance–capacitance filter section as shown in Fig. 3.2(iii) if required, giving a complete demodulation circuit as in Fig.



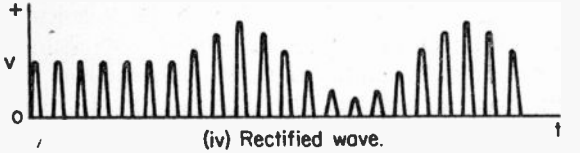
(i) Unmodulated carrier wave at 1MHz.



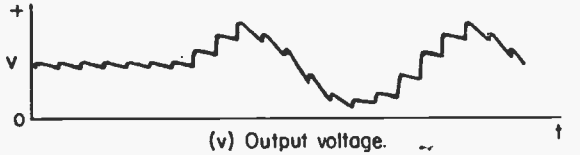
(ii) 1kHz modulating frequency.



(iii) Modulated carrier wave.



(iv) Rectified wave.



(v) Output voltage.

Fig. 3.4 Modulation and demodulation.

3.5. Electrolytic capacitors are not needed at radio frequencies because capacities are small.

Since the carrier wave has been rectified, there must be a d.c. component (in fact this is what *power* rectification is all about) and this accounts for the shift of axis. This component may be removed if required by a final series capacitor (C_2) which *blocks* it and hence passes on a replica of the original modulating frequency of Fig. 3.4(ii).

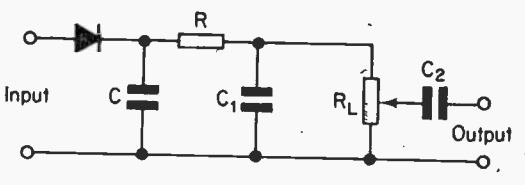


Fig. 3.5 Demodulation (detection) circuit.

At these frequencies C and C_1 may typically be of the order of 0.01–0.05 μF, R, 470 Ω and R_L 5,000 Ω. R_L may have a sliding connexion (*potentiometer*, for adjusting electrical potential) acting as a volume or gain control feeding into an audio stage which amplifies the 1 kHz tone for reproduction by phones or a loudspeaker.

3.2 AMPLIFIERS

Change in electronic technology continually gathers momentum so that whereas in the not too distant past amplifiers and the like were designed stage by stage to well known principles, as time progresses the design process changes more and more to acceptance of an integrated circuit *package* (discussed in Chapter 4) containing many diodes and transistors and capable of doing a complete job. Such packages are treated as black boxes with the electrical characteristics for the whole box

quoted by the manufacturer and the designer adds those components which at present cannot be part of such a package as, for example, large value capacitors. In view of this changing technique we will not follow transistor amplifier design in detail but instead concentrate on the basic concepts which will not change with time and which are of importance whether a single transistor amplifier stage or a complex integrated circuit is involved.

3.2.1 Expectations

Going back to Fig. 1.4(ii) showing the general symbol for an amplifier gives a first idea of what we expect from such a device, a small signal applied to the input resulting in a larger signal at the output. That the output magnitude should be greater characterizes the transistor as an *active component* (that is, one which gives rise to a power gain). Alternatively, resistors and other devices which contain no source of power are known as *passive components.*

To complete the design requirement the output signal should be *clean*, implying that the amplifier must be discouraged from adding ingredients of its own, generally known as *noise*. The maximum noise tolerable is seldom quoted as an absolute value, the amplifier is more likely to be assessed by its *noise figure* which involves a *signal-to-noise ratio*. This ratio is also a suitable way of stating the requirements of a system and varies widely, for example the human ear is remarkably tolerant to noise when it wants to be, it can withstand a ratio of less than 1 (noise louder than the signal) because of its capability of picking out the signal, yet for an enthusiast judging a hi-fi system a ratio of 100 will not be satisfactory. Electronic systems do not in general have the discriminating qualities of the human ear and are therefore frequently less able to accommodate a low signal/noise ratio.

Summing up our expectations of a perfect amplifier, we might therefore say that its output must resemble the input signal in all respects except for an increase in magnitude, and with nothing added.

The term "increase in magnitude" may need some further definition. When the output current, voltage or power is divided by the appropriate input value, a number results which is called the *gain* of the amplifier. An output power of 1 W resulting from an input power of 1 mW shows that the amplifier gain = 1 W/1 mW = 1000.

A useful and commonly used system for expressing gain or loss is by *decibel notation*, but people do slightly irregular things with decibels when quoting amplifier gains which tends to make this technique confusing and it is therefore best left for futu.e studies.

3.2.2 Limitations

Anybody who listens to a hi-fi system may perhaps feel that amplifiers have reached the stage of perfection but it must be remembered that much design effort has been expended in reducing the several limitations that are inherent in the ordinary transistor amplifier. Some of these are outlined below.

3.2.2.1 Distortion

This simply means that the output wave is not an exact replica of the input, that is, wave distortion is present. This can occur either through the addition or subtraction of harmonics. Fourier analysis has taught us two important features about treatment of waveforms:

(i) changing the shape of a sine wave adds harmonics, for example when the wave is completely squared, all odd harmonics to infinity are produced;
(ii) alternatively, subtracting harmonics from a rapidly rising or falling wave reduces its rate of rise or fall because a square-type wave is being changed towards sine shape.

Condition (i) therefore shows that for minimum harmonic generation, the output/input relationship of the system should be linear, that is, there should be no *non-linearity distortion*. This is demonstrated in Fig. 2.7(i) where a curve

relating an output to an input is non-linear and accordingly the input and output signals are of different shapes, the positive half-cycle of the output having increased more than the negative. Such waveform distortion results in a change of harmonic content which in a hi-fi system for example, might be audible through unwanted sounds. The ideal output/input characteristic is therefore straight.

Condition (ii) is demonstrated by a gain/frequency response curve for an amplifier as shown in Fig. 3.6. This is obtained by applying to the input terminals of the amplifying system sine wave signals of constant amplitude but with frequency varying over the band in question. The output is measured across a load resistance of the correct value for the particular amplifier. The load used is non-reactive and is substituted for the normal load to avoid secondary effects due to any variation with frequency which may be present in the latter. The gain is simply output voltage/input voltage and in this case we are not so much concerned with the actual gain as with how it varies with frequency. To add some reality the frequency range shown is that for an audio system and we are now using *logarithmic* graph paper. This has two advantages, it condenses the frequency scale and also adds a touch of realism because the human ear which this amplifier serves has logarithmic qualities in that each octave rise in the musical scale, although doubling in frequency each time, seems to result in the same change in pitch.

Fig. 3.6 shows that the gain falls off at both ends of the frequency range. At the higher frequencies the fall is caused by unwanted but inevitable shunt capacitances across the signal path. Fig. 2.10 is a reminder of the shunting effect of transistor capacitance, so causing a signal loss which increases with frequency. The fall in the characteristic at the lower frequencies will be understood when we study a practical amplifier circuit later. Evidently for any wave passing through the amplifier which contains harmonics, some of the higher ones will be amplified more or less than their fundamentals, so destroying the original harmonic/fundamental relationship. As an example a fundamental wave at 6 kHz

will find its 3rd harmonic amplified only 90 times whereas for itself the amplification is 100 times. The ideal characteristic for a 50 Hz to 20 kHz audio amplifier is therefore as shown dotted in the figure, with this, all frequencies in the design range are amplified equally.

The principles outlined apply to all amplifiers whatever their frequency ranges.

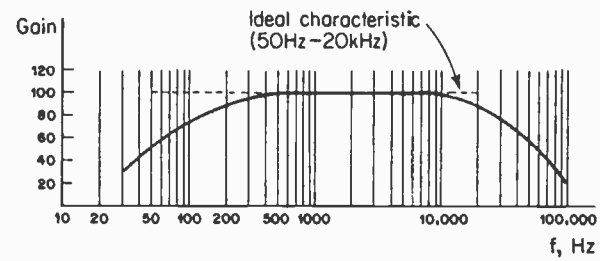


Fig. 3.6 Typical gain/frequency characteristic of audio amplifier.

3.2.2.2 Noise

In discussing rectification it was found that a d.c. supply obtained from the power mains may contain traces and/or harmonics of the power frequency and this was classed as *noise*. Such noise is detrimental to amplifier efficiency because noise in the power supply at the first stage of an amplifier is followed by the full gain of the amplifier [amplifiers consist of one or more *stages* of amplification in tandem, generally each stage consists of a single transistor]. Hence a tiny noise voltage results in an appreciable amplified one at the output. Battery driven equipment does not suffer from this.

There is another source of noise which, unlike the above cannot be eliminated and does in fact set one of the limits to amplification. Consider the end of a piece of wire as an example. There, as at all other parts of the wire, electrons

are in motion and at any instant the number at the surface will be different from the number there at any other instant. This can be likened to a pool, heavily stocked with tiny fish, they are darting about in all directions and the number at the surface is continually changing. With the electrons the total charge at the surface must therefore be varying in a random manner, resulting in tiny varying voltages at all frequencies. This is known as *white noise* – just as all colours add up to white light, so all frequencies add up to white noise. It may seem rather odd that a piece of wire connected to nothing has a voltage at its ends but the value can actually be calculated and will make itself heard in an audio system of sufficient gain as a hissing sound in the loudspeaker. The formula is

$$v_n = \sqrt{4kTR(f_2 - f_1)}$$

where v_n is the r.m.s. noise voltage, T is the temperature in °Kelvin, R is the conductor resistance, $(f_2 - f_1)$ is the frequency band in Hz and k is a constant of value 1.38×10^{-23} joule/°K and is known as *Boltzmann's constant* (after Ludwig Boltzmann, an Austrian physicist).

That the noise voltage increases with temperature is to be expected from the fact that electrons have greater energies at higher temperatures. The noise is generated equally at all frequencies therefore v_n depends on the width of the band considered.

In addition to the above *resistance noise* which is fundamental to all materials, other noise is generated by carrier movements in transistors, for example, holes travelling across the base in a pnp transistor, their movement is random so that the number collected varies from instant to instant, this variation appears as noise. Recombination also produces noise and the sum of all noise voltages will appear amplified at the system output. To what degree noise can be tolerated depends much on the type of system of which the amplifier forms part. The amplifier itself may be rated by its *noise figure* which is the ratio of the signal/noise power ratio at

the input to that at the output.

EXAMPLE:
What noise voltage is developed within the frequency range 10–20 MHz across a 10 kΩ resistor running at 180°C?

$$(f_2 - f_1) = (20 - 10)\text{MHz} = 10^7 \text{ Hz}$$

$$180°\text{C} = 273 + 180\,°\text{K} = 453°\text{K} \; (= \text{T})$$

$$R = 10^4\,\Omega$$

Then,

$$v_n = \sqrt{4 \times 1.38 \times 10^{-23} \times 453 \times 10^4 \times 10^7}$$

$$= \sqrt{4 \times 1.38 \times 453 \times 10^{-12}} = 10^{-6}\sqrt{4 \times 625}\ \text{V}$$

$$= \underline{50\,\mu\text{V}}.$$

3.2.3 h-Parameter analysis
There are several methods of representing transistors and their associated circuits as electrical networks for analysis. Two which have their own particular fields of usefulness are the T and hybrid-π Equivalent Networks, the latter being the more complicated but capable of representing the transistor more easily at high frequencies. However a third method which has the advantage of needing only a few straightforward measurements is in more general use and has found favour with many manufacturers, mainly because of the ease of measurement. It is known as the *h-parameter* method and because much data is published in this form it is the one we will briefly examine.

A "parameter" is a quantity which is constant for the particular device under the set of conditions being considered but which varies in other cases, so h-parameters are fixed for a certain type of transistor but are different for other types. h arises from the term *hybrid*, meaning in this instance a

"mixture", because the several parameters involved are in different units or are ratios.

The h-parameter technique is used here with transistors but because it works on the black box principle, it can be applied to any system which has two input and two output terminals. Consider Fig. 3.7 which shows a 4-terminal black box fed at the input by a voltage generator v_s with internal resistance R_s, and terminated by a load R_L at the output. Let us consider one h-parameter of which we already have some experience, the *input* resistance, h_i. Because we are finally interested in the treatment of a small signal applied to the input terminals, this is an a.c. value and strictly is an impedance rather than a resistance. The value can be obtained from the slope of the static curve which relates direct voltage with current at terminals 1 and 2.[A4] However, given the equipment, it is more usual to make a direct measurement of a.c. resistance at some chosen frequency, usually 1 kHz. Certain conditions of measurement are stipulated for each h-parameter, for example, the definition of h_i requires that the output terminals are short-circuited to a.c. Since a metallic short-circuit may interfere with collector d.c. supplies, using a large capacitor instead effectively short-circuits a.c. but not d.c.

Similarly the remaining three h-parameters are determined from the slopes of the appropriate static curves at the operat-

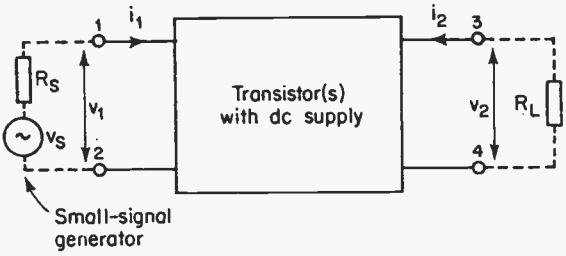


Fig. 3.7 h-parameters, general circuit.

ing points or alternatively obtained directly by a.c. measurements. The parameters, including h_i, and with reference to Fig. 3.7, remembering that v, i, etc. are small-signal a.c. quantities, are

h_i is the input resistance $= \frac{v_1}{i_1}$ (with output short-circuited)

h_f is the forward current transfer ratio (current gain) $= \frac{i_2}{i_1}$ (with output short-circuited)

h_o is the output conductance $= \frac{i_2}{v_2}$ (with input on open-circuit)

h_r is the reverse voltage transfer ratio $= \frac{v_1}{v_2}$ (with input on open-circuit).

These are the ones particularly suited to analysis of transistor circuits and a second subscript is added to each to specify the configuration, for example, h_{ob} = output conductance, common-base, h_{oe} is the same but for common-emitter.

h_f is not new to us since we have already considered current gain in the form of α_b and α_e. h_o is a conductance, that is the reciprocal of resistance, it is measured at the output terminals just as h_i is measured at the input terminals. h_r needs explanation. It was mentioned in Section 2.2.3 that, for example, with common-base, the collector voltage does have a small effect on the input characteristic. Although the reference was to the static characteristic, it must also apply to the a.c. condition, that is, v_2 at the output (Fig. 3.7) has a small effect on v_1 at the input. We may understand this better by the simplified idea that the transistor is basically a resistance, complicated though it may be, thus a voltage at the output end produces a current which flows back through the transistor and through the input resistance to create an internal *feedback* voltage there.

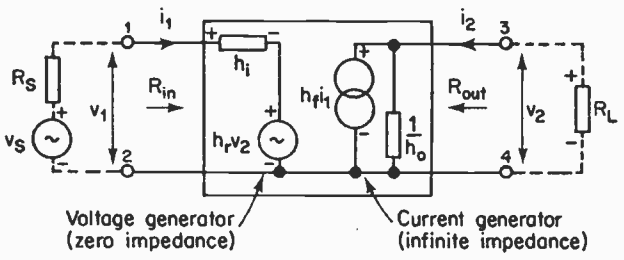


Fig. 3.8 h-parameters, equivalent circuit.

We next draw an *equivalent* circuit to represent what goes on electrically inside the box, in this case presumed to contain a transistor circuit. Fig. 3.8 shows this to include:

(i) a voltage generator of zero impedance. No matter what current this generator causes to flow, its internal voltage-drop is zero, hence the terminal voltage remains constant;
(ii) a current generator of infinite impedance. To provide the current a voltage source is assumed to be acting through the impedance. Other resistances or impedances connected in series with an infinite impedance have no effect, therefore the current remains constant.

These artifices may seem unreal but from the analysis point of view, they hold good. The transistor is now seen as two separate parts. The input at terminals 1 and 2 consists of a resistance h_i in series with a feedback voltage $h_r v_2$. The latter acts in series with v_1 which it modifies slightly according to the magnitude of v_2 from whence it arises. Transistor action is seen as a current generator feeding the output circuit, the current value being $h_f i_1$, that is, i_1 magnified by the current gain h_f. The full current from the generator does not reach the output terminals however because of the output resistance of the transistor itself. h_o is the output conductance, therefore $1/h_o$ is the output resistance, some of the current is by-

passed through this, showing that the practical current gain is not simply $h_f i_1$ but something less. We first determine the current gain as an example of the usefulness of the equivalent circuit approach.

Current gain (i.e. current in load/input current) is denoted by K with subscript i. (Subscripts v and p are used for voltage and power gains respectively.)

Current gain: Thus i_2 comprises the current from the generator plus the current in the output resistance (e.g. the collector-base junction in a common-emitter circuit), i.e.

$$i_2 = h_f i_1 + \frac{v_2}{1/h_o} = h_f i_1 + h_o v_2$$

Now if the instantaneous current i_2 flows as shown in Fig. 3.8, i.e. downwards through the generator, the latter and R_L have polarities as indicated. At the output terminals the polarities are in opposition and we therefore class the voltage across R_L as negative:

$$\therefore \quad v_2 = -i_2 R_L$$

$$\therefore \quad i_2 = h_f i_1 - h_o i_2 R_L$$

$$\therefore \quad i_2(1 + h_o R_L) = h_f i_1$$

$$\therefore \quad \text{Current gain } K_i = \frac{i_2}{i_1} = \frac{h_f}{1 + h_o R_L}$$

EXAMPLE:
The h-parameters of a transistor connected as an amplifier (in common-emitter) with a load resistance of 20 kΩ are, h_{fe} 33, h_{oe}, 25 μS. What is the current gain?

$$R_L = 20 \times 10^3 \Omega$$

Current gain:

$$K_i = \frac{h_{fe}}{1 + h_{oe}.R_L} = \frac{33}{1 + (25 \times 10^{-6} \times 20 \times 10^3)}$$

$$\doteq \frac{33}{1 + 0.5} = \underline{22}$$

Note that the input resistance (h_{ie}) and the reverse voltage transfer ratio (h_{re}) do not affect the current gain since although these factors may change the magnitude of the input current, this has no effect on its ratio with the output current.

Input resistance: Looking into the input terminals of Fig. 3.8, the net voltage is seen as the voltage drop across h_i, plus the feedback voltage $h_r v_2$, thus:

$$v_1 = h_i i_1 + h_r v_2$$

Now $v_2 = -i_2 R_L \qquad i_2 = K_i i_1$

$\therefore \quad v_2 = -i_1 K_i R_L$

$\therefore \quad v_1 = h_i i_1 + h_r(-i_1 K_i R_L) = i_1(h_i - h_r K_i R_L)$

$\therefore \quad R_{in} = \frac{v_1}{i_1} = h_i - h_r K_i R_L$

and substituting for K_i

$$R_{in} = h_i - \frac{h_f h_r R_L}{1 + h_o R_L}$$

thus the input resistance is h_i modified by a term which includes h_r. Also R_L is included showing that the load-impedance affects the input resistance.

Output resistance: i_1 flows through the resistance of the source, therefore $v_1 = -i_1 R_s$ (negative because the voltage developed across terminals 1 and 2 opposes that of the source generator) and since as we have seen above:

$$v_1 = h_i i_1 + h_r v_2$$

$$\therefore \quad -i_1 R_s = h_i i_1 + h_r v_2$$

$$\therefore \quad -i_1 (h_i + R_s) = h_r v_2$$

$$\therefore \quad i_1 = -\frac{h_r v_2}{h_i + R_s}$$

But $$i_2 = h_f i_1 + h_o v_2$$

(see under "Current gain")

$$\therefore \quad i_2 = -\frac{h_f h_r v_2}{h_i + R_s} + h_o v_2$$

$$= v_2 \left(h_o - \frac{h_f h_r}{h_i + R_s} \right)$$

$$\therefore \quad \frac{i_2}{v_2} = \left(\frac{h_o (h_i + R_s) - h_f h_r}{h_i + R_s} \right)$$

which gives the output conductance

or Output Resistance, $R_{out} = \dfrac{h_i + R_s}{h_o (h_i + R_s) - h_f h_r}$

thus the output resistance is modified by the source resistance.

Voltage gain: Defined as v_2/v_1, and as already found, $v_1/i_1 = R_{in}$,

$$\therefore \quad v_1 = i_1 R_{in} \qquad \text{Also, } v_2 = -i_2 R_L$$

$$\therefore \quad K_v = \frac{v_2}{v_1} = \frac{-i_2 R_L}{i_1 R_{in}} = \frac{-K_i R_L}{R_{in}}$$

$$\text{since} \quad \frac{i_2}{i_1} = K_i$$

or

$$K_v = -\frac{h_f R_L}{h_i(1 + h_o R_L) - h_r h_f R_L}$$

by substituting for K_i and R_{in}.

The following example reminds us that circuits are seldom wholly resistive and also shows how the overall or *external voltage gain* can be calculated. R_s is no longer a resistance but an impedance.

EXAMPLE:
What is the approximate output voltage of the amplifier shown in Fig. 3.9 when the input is 10 mV at 1592 Hz? The h-parameters are h_{ie} 1100, h_{fe} 90, h_{oe} 40 μS, h_{re} negligible.

Reactance of 10 μF capacitor in series with output terminal is equal to

$$\frac{1}{\omega C} = \frac{10^6}{2\pi \times 1592 \times 10} = 10\ \Omega$$

which is negligible compared with R_L.

$$\therefore \quad \text{Voltage gain } K_v = \frac{v_2}{v_1} = \frac{h_{fe} R_L}{h_{ie}(1 + h_{oe} R_L) - h_{re} h_{fe} R_L}$$

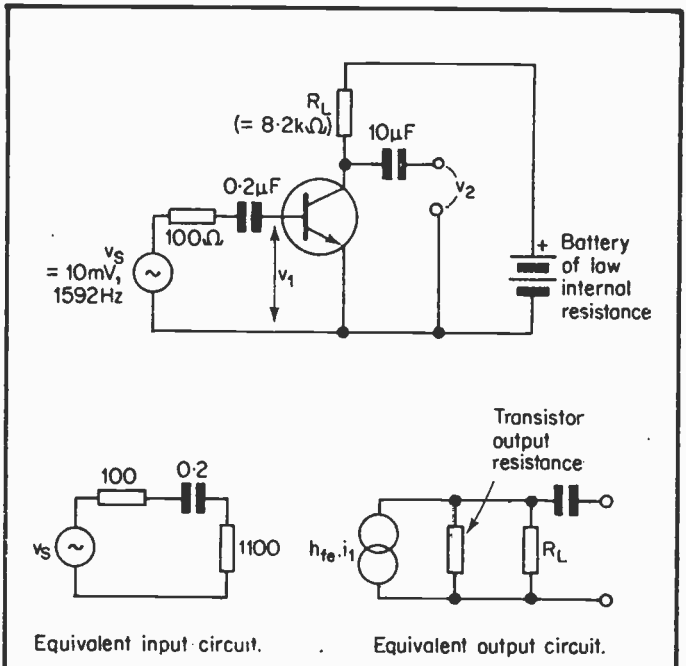


Fig. 3.9 Calculation of output voltage of transistor amplifier.

h_{re} is negligible therefore the second term in the denominator = 0.

$$\therefore \quad K_v = \frac{90 \times 8200}{1100(1 + 40 \times 10^{-6} \times 8200)} = \frac{90 \times 8200}{1100 \times 1.328}$$

$$= 505$$

Reactance of 0.2 μF capacitor is equal to

$$\frac{1}{\omega C} = \frac{10^6}{2\pi \times 1592 \times 0.2} = 500\ \Omega$$

$\therefore$ Total impedance of input circuit

$$|Z_{in}| = \sqrt{R^2 + X_C{}^2} = \sqrt{1200^2 + 500^2} = 1300\ \Omega$$

Then input current $= \dfrac{v_s}{|Z_{in}|}$

and

$$v_1 = \frac{v_s}{|Z_{in}|} \times 1100 = \frac{10 \times 10^{-3}}{1300} \times 1100 = \frac{11}{1300}\ V$$

$\therefore$ Output voltage is equal to

$$v_1 \times K_v = \frac{11}{1300} \times 505 = \underline{4.27\ V}$$

Thus we demonstrate the value of the system. If, as is often the case, a manufacturer quotes h-parameters for common-emitter only, the common-base and common-collector parameters may be derived from them by using standard formulae.

3.2.4 The two-stage amplifier

Section 2.2.1 describes and Fig. 2.6 illustrates the elements of single-stage amplifiers, operating correctly in principle but not fully practical because two separate batteries are employed whereas one is sufficient. But this now allows us to jump directly to a two-stage amplifier so that stage coupling arrangements can be included, that is, how the output of one stage is fed or *coupled* into the next. Common-emitter is one of the most widely used configurations and is therefore used as the example. Fig. 3.10 shows a typical practical two-stage transistor amplifier, suitable in this case as an *output* amplifier driving a small loudspeaker or headphones. Two npn transistors are shown and biassing arrangements (discussed next) have been chosen to illustrate the various methods. Calculation of the overall gain follows the principles outlined, remembering that the output resistance of one stage forms the source resistance of the next.

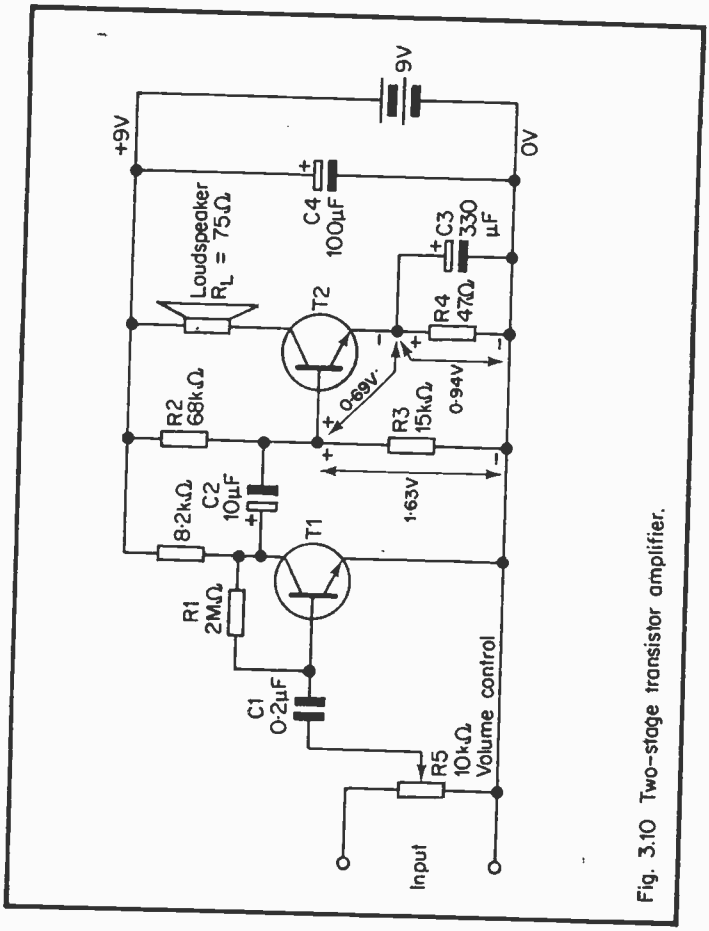


Fig. 3.10 Two-stage transistor amplifier.

3.2.4.1 D.C. bias

Reference to Fig. 2.7 reminds us that an optimum operating point on the input characteristic has to be chosen. It is produced in the practical circuit by a direct voltage applied to the base. Simply connecting a resistor from the base to the battery supply rail is not normally sufficient because although the correct bias may be obtained, the method does not guard against the change in transistor characteristics with tempera-

ture, nor against the fact that the parameters of transistors even of the same type, vary widely. It is important that the operating point should not shift with replacement of a transistor otherwise undesirable distortion may arise. Section 2.3 discusses this aspect with accent on the removal of the heat, circuit biassing arrangements are complementary to this by automatically guarding against a current rise as temperature increases. Because a current rise causes a further temperature rise it has a cumulative effect which may result in destruction of the transistor. In the circuit of Fig. 3.10 transistor power dissipations are such that heat sinks are unnecessary, nevertheless stabilization by d.c. biassing arrangements is.

For transistor T1, R1 the biassing resistor connected between collector and base is of such a value that the desired forward voltage is obtained on the base, i.e. $V_C - R1 \times I_B$. The value of R1 is considerably greater than the resistance of the base-emitter junction and R1 therefore mainly controls the base current. Should the temperature rise and I_C increase, the voltage-drop across the load resistance increases, leaving a lower voltage applied to R1 and therefore I_B falls, so reducing I_C, not to its original value, but very nearly.

The second transistor T2 is biassed by a different method, a resistor chain R2, R3 connected across the supply has resistances of such value that the base is held at the desired voltage. In addition an emitter-resistor R4 is used. Imagine R2 to be disconnected, then I_E flowing through R4 makes the emitter positive to the 0 V rail, hence via R3 the base is negative to the emitter. Connexion of R2 now holds a positive potential at the base in excess of that developed by R4 by the bias value required, thus, using values shown on Fig. 3.10 as an example, if I_E = 20 mA and R4 = 47 Ω, then emitter voltage relative to common is equal to

$$+\frac{20}{1000} \times 47 = +0.94 \text{ V}$$

and base voltage from potential divider R2, R3

$$= + \frac{15}{68 + 15} \times 9 = +1.63 \text{ V}$$

(neglecting the small base current flowing). These two potentials are shown in the figure from which, being in opposition, the net bias voltage is $+1.63 - 0.94 = +690$ mV. Note the relatively high values of R2 and R3 to ensure that their current drain on the power supply is not excessive.

If I_C and therefore I_E increase, V_E increases, thus reducing the net base voltage, reducing I_B and therefore I_C, so compensating for the original variation. These are d.c. conditions, we do not wish to neutralize changes in the a.c. signal, therefore the emitter resistor must be shunted by a *by-pass* capacitor of sufficiently low reactance (say about one-tenth of the resistance it shunts) at the lowest frequency, e.g. for C3 = 330 μF, X_C at 100 Hz, is equal to

$$\frac{1}{\omega C} = \frac{10^6}{2\pi \times 100 \times 330} = 4.8\ \Omega .$$

There are other bias and stabilization circuits, those shown are fairly common and serve to demonstrate the principles.

3.2.4.2 The dynamic load line

This is a most useful tool for choosing the value of the load, the optimum operating point and subsequently for examining the output waveform to ensure that the transistor does not drive into distortion. Stage T2 of Fig. 3.10 demonstrates the technique.

The load, which would include R4 (via C4) if it were not short-circuited to a.c. by C3, is simply R_L. For this discussion we assume the loudspeaker to be non-reactive, that is, equivalent to a 75 Ω resistor. A family of output curves for

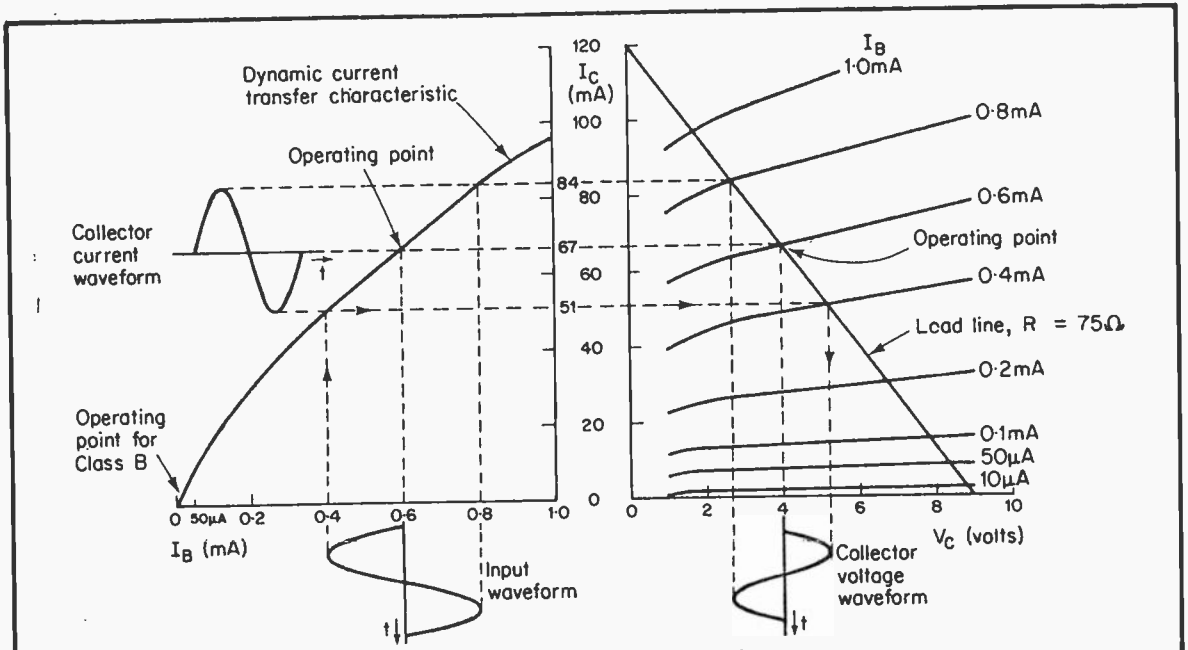


Fig. 3.11 Load line construction.

T2 might be as on the right-hand side of Fig. 3.11, linking I_C and V_C over a range of values of I_B. On these the load line for 75 Ω is drawn, i.e. between I_C = 9 V/75 Ω = 120 mA at V_C = 0 and at V_C = 9 V where I_C = 0. From the points of intersection of the load line with each characteristic, horizontal lines are projected to the left to define the *dynamic current transfer characteristic*, linking I_B and I_C for that particular load. For a different load the load line changes and a new transfer characteristic must be drawn. The bias operating point is now chosen with due regard to the magnitude of the input signal. We will assume a fairly small signal on the base of peak value 0.2 mA. A base bias value (operating point) of 0.6 mA appears suitable because it is about central in the straight part of the transfer characteristic. From the dotted lines linking the two characteristics this results in a steady I_C value of about 67 mA. Downwards from the load line to the V_C axis we see that V_C = 4 V. Thus we have determined the values of I_B, I_C and V_C for a load of 75 Ω when the signal current is zero. These are known as the *quiescent* (inactive) values.

Drawing the input waveform varying between I_B = 0.4 and 0.8 mA (on a vertical time axis) enables us to plot both the waveforms of collector current and collector voltage. Very little distortion is evident, but should the input signal be large so as to swing I_B over most of its working range then at the lower currents some distortion of the collector voltage waveform would occur and the operating point might need to be shifted. Note that when I_B increases, V_C falls and vice versa, this is equivalent to a phase-shift of 180°, a feature of the common-emitter circuit.

For an overall picture from base voltage rather than base current, the input characteristic connecting V_B with I_B can be added.

3.2.4.3 Interstage coupling

In the type of amplifier shown in Fig. 3.10 the a.c. signal developed across the load of T1 must be applied to the base of T2. A direct connexion, satisfactory from the signal

point of view, would however apply the collector d.c. potential of T1 directly to the base of T2, giving incorrect bias. The simplest answer to this is the series capacitor, in this use known as a *coupling capacitor*. It *blocks* the d.c. yet if of sufficiently low reactance will allow most of the a.c. voltage to be developed across the resistance of the following circuit in this case the base circuit of T2. Generally at the lowest frequency, the reactance of a coupling capacitor will be about one-tenth or less of the net resistance to which it is connected.

Taking Fig. 3.10 as an example, the net a.c. resistance of the input circuit of T2 as seen by C2 is the parallel combination of R3 with R2 (its top end is effectively connected to the 0 V rail by C4) and the input resistance R_{in} of T2. (R4 is in series with the latter but it is by-passed by C3.) Suppose R_{in} of T2 to be 1000 Ω, then if resistance seen by C2 = R,

$$\frac{1}{R} = \frac{1}{R2} + \frac{1}{R3} + \frac{1}{R_{in}} = \frac{1}{68{,}000} + \frac{1}{15{,}000} + \frac{1}{1000}$$

$$= 0.00108, \quad \therefore \; R = 925\ \Omega$$

For most of the signal voltage to be developed across R, X_C ought to be less than one-tenth, i.e. < 92.5 Ω. If the lowest amplifier frequency at which the gain should not fall appreciably from the maximum is, say, 200 Hz, since

$$X_C = \frac{1}{\omega C},$$

$$C = \frac{1}{\omega X_C} = \frac{10^6}{2\pi \times 200 \times 92.5} = 8.6\ \mu F,$$

– a 10 μF capacitor would therefore be adequate. At

frequencies below 200 Hz more of the signal voltage is developed across the increasing reactance of C2 leaving less across the input circuit of T2, hence the gain falls progressively as frequency falls. At frequencies above 200 Hz the opposite applies and the voltage loss due to C2 is less. Reference to Fig. 3.6 illustrates this for a complete amplifier.

For amplifiers working at higher frequencies the values of coupling capacitors are lower, for example, if the circuit of Fig. 3.10 were designed for 200 kHz upwards, C2 would be reduced to 0.01 μF.

3.2.5 Power amplifiers

So far discussion has centred on general amplifier features with some bias towards small signal amplifiers. With these, large voltage or current gains are possible, if insufficient gain is available from one stage then further stages may follow in tandem, the overall gain being the product of the individual stage gains. The final stage in an amplifier chain usually does some work as is required for example in moving the cone of a loudspeaker, operating a solenoid or even turning a pointer over a scale. Power is the rate of doing work, hence the term *power amplifier.* Fig. 3.10 illustrates this for T1 is a voltage amplifier, raising the input voltage to a level sufficient to drive a power amplifier T2 which, in this case, has a very moderate power output of the order of some 10 mW as can be estimated from the I_C and V_C waveforms of Fig. 3.11. T2 is also said to be a *Class-A* amplifier. This refers to the method of operation as set by the bias. In Class-A, output current is present under quiescent conditions and flows throughout the cycle of the input signal. In Fig. 3.11 Class-A conditions hold because I_C = 67 mA (quiescent) and varies between 51 and 84 mA when the signal is applied, thus current is always present. This is in contrast with *Class-B* operation where the output current under quiescent conditions is zero and current only flows during one half-cycle of the input signal, for example, by choosing a bias of zero or a few μA as marked on the transfer characteristic of Fig. 3.11. This may seem to be leading to disastrous distortion or even half-wave rectification but we shall see that special Class-B ampli-

fiers overcome this and moreover lead to greater efficiency than Class-A ones by drawing less current from the supply for the same a.c. power output.

One further method is used, *Class-C*, for this the steady bias is such that the quiescent value of I_C is also zero but current flows for less than one half-cycle. Class-C is used mainly in radio transmitters.

3.2.5.1 Push-pull amplifiers

Maximum power output for the circuit represented by Fig. 3.11 is obtained when the input waveform is of such magnitude that the transistor is driven over the whole of the usable part of the dynamic current transfer characteristic. The operating point lies midway between the values of I_B which result in zero collector current and the maximum allowable. From the figure it is clear that distortion in the output waveform occurs at low values of I_B and I_C because of the curvature of the transfer characteristic. This becomes worse when the input characteristic is also considered owing to its own non-linearity. To reduce distortion two transistors may be operated in a *push-pull* mode; the derivation of the term will become evident as we progress.

By moving directly to examination of the principles of push-pull amplifiers and of Class-B working at the same time, we shall in fact learn about the most commonly used system.

Consider the skeleton circuit of Fig. 3.12(i) which at this stage uses an input transformer with two secondary windings feeding the transistors T1 and T2 biassed (circuitry not shown) to a quiescent operating point as shown in Fig. 3.13. This is not quite as defined for Class-B because a small current flows in both transistors when no signal is applied but doing this avoids the most curved part of the characteristic where waveform distortion is greatest. Strictly the bias is known as Class-AB, that is, somewhere between the two.

The secondary connexions of the input transformer of Fig. 3.12(i) provide equal but 180° out-of-phase input signals

across base and emitter of both transistors, thus as the collector current of T1 moves in either direction, that of T2 moves in the opposite direction. Both currents flow through the load R_L and are therefore *subtractive*. Fig. 3.13 shows this in graphical form. The characteristics shown are I_C/V_B, we do not take the further step to V_C/V_B since this involves adding the output characteristics and load lines which are not essential to this particular discussion. The input wave results in collector current variations as shown. Subtraction of one wave from the other because they flow in opposition in R_L produces a current waveform in the latter as developed to the right in the figure (marked "current in load").

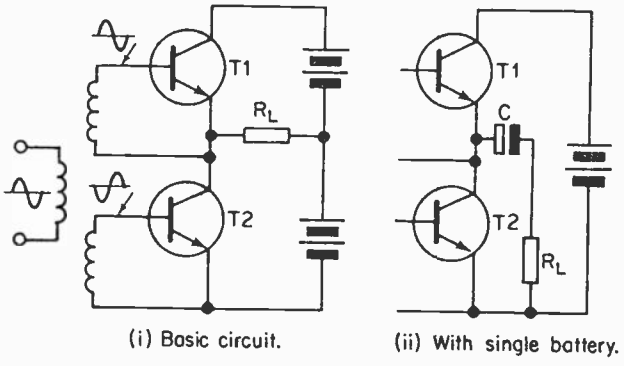


Fig. 3.12 Push-pull amplifiers.

Clearly improvements can be made to the circuit of Fig. 3.12(i) because not only are two batteries (or a centre-tapped one) undesirable but also transformers are components to be excluded from transistor circuits when this is feasible because of bulk and cost. For a single battery supply therefore, Fig. 3.12(i) gives way to (ii) in which the right-hand end of R_L is connected to the common rail which from an a.c. point of view is permissible since the impedance of the lower battery in (i) is very low. A capacitor C of low reactance compared with

R_L at the lowest frequency is then necessary to avoid shunting T2 from the d.c. point of view. Generally C will be of high value, for example, if R_L were an 8 Ω loudspeaker, C might be of the order of 1000 μF or more. The input transformer is eliminated by special *driver* circuits, discussed later.

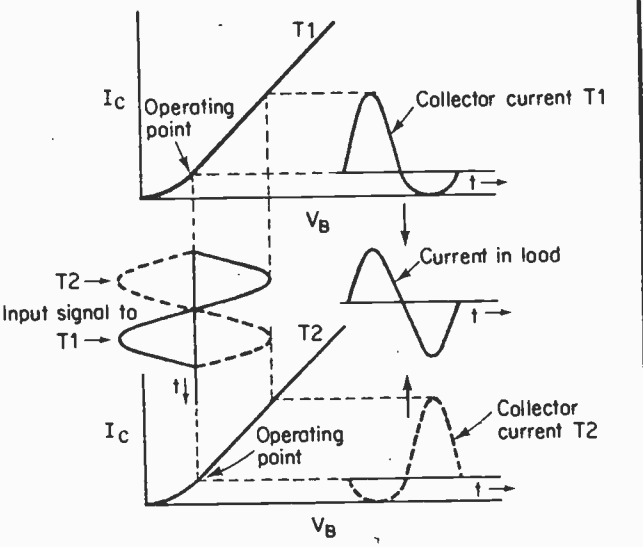
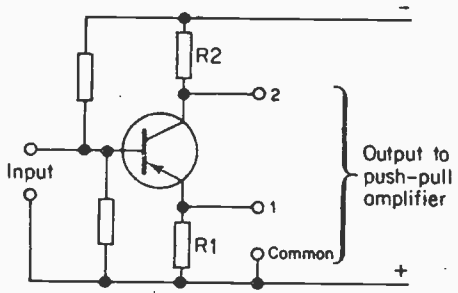


Fig. 3.13 Waveforms in a Class-B amplifier.

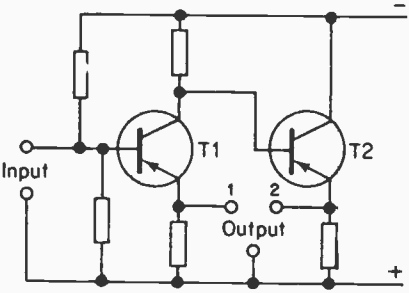
Some distortion of the two collector current waveforms is inevitable because the I_C/V_B characteristic may not be as linear as shown in Fig. 3.13. However there is a reduction of the resultant harmonics through cancellation in the load. Mathematically it can be shown that all even harmonics are completely cancelled provided that the two transistors have identical characteristics. For this reason matched pairs of transistors are available from manufacturers for push-pull working.

Apart from this, another very important feature of Class-B is its lower direct current drain, especially important when the power is supplied by a battery.

Driver circuits: This is the name given to a stage preceding a power amplifier, its output when the correct input signal is applied must be sufficiently large to drive the power amplifier fully. Examination of just one of the many types of driver circuits for push-pull working is sufficient to show the technique of *phase-splitting.* This produces two signals in



(i) Single stage.



(ii) Double stage.

Fig. 3.14 Phase splitters.

antiphase from one input signal as does the transformer of Fig. 3.12(i).

Fig. 3.14(i) shows a phase-splitter using a single pnp transistor (an npn is equally suitable). A convenient way of looking at the process is by considering R1 and R2 to be equal (this is quite likely to be so in practice) and let the current through the chain be such that with a 9 V supply, 3 V is dropped across each. Then relative to the common rail, terminal 1 is at −3 V and terminal 2 at −6 V. Now suppose that the input signal swings negative so that the current through the chain (ignoring base current) rises to give a 3.5 V drop across both R1 and R2. Terminal 2 changes to −5.5 V, that is, it has moved 0.5 V positive (180° out-of-phase with the input signal) whereas terminal 1 becomes −3.5 V and so has moved 0.5 V negative (in phase). The opposite happens when the input signal goes positive. Terminals 1 and 2 can therefore drive a push-pull amplifier.

We ignored the base current however, not doing so would have shown a slightly greater voltage drop across R1 than R2, hence the phase-splitter is not perfectly balanced. This can be overcome by connecting the collector of T1 to a second transistor T2 as in Fig. 3.14(ii), the emitter voltage of T2 moves in phase, hence terminals 1 and 2 of this circuit are fed from similar impedances but still are 180° out-of-phase.

3.2.6 Direct-coupled amplifiers

It is evident from earlier comments on interstage coupling and by-pass capacitors that these have an undesirable effect at the lower frequencies of the design range of an amplifier. Capacitance values can be increased to accommodate lower and lower frequencies but when ultimately down to a few Hz or 0 Hz (direct current) the size of the capacitor is unpractical. *Direct coupling* obviates this and is feasible provided that the d.c. voltage on the collector of one stage can bias correctly the following stage. It is likely to be higher than is required but the excessive voltage may be cancelled by an opposing voltage developed across the emitter resistor as shown in Sect. 3.2.4.1. This is not a complete solution however for the problem of

drift remains. Drift refers to the slow change in characteristics with temperature and may produce a change in the output great enough to mask the change in the signal being amplified (we are now considering signals as slowly changing direct currents such as may be used in measurement systems). The general term "d.c. amplifier" is useful for the "d.c." reminds us of both direct coupling and direct current.

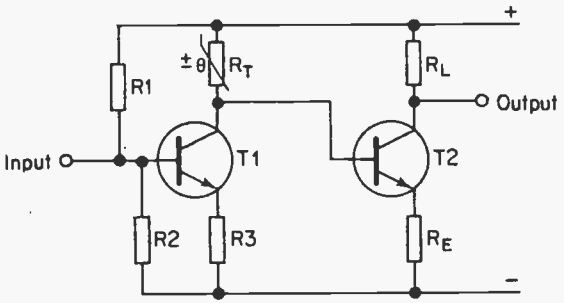
There are many methods of improving amplifier stability from the point of view of drift, some using special components but others employing ordinary transistor circuitry. one of each type is considered.

Fig. 3.15(i) shows two transistors directly coupled. T1 is biassed by R1, R2 and R3 while the emitter resistor R_E of T2 is of such value that the d.c. voltage from the collector of T1 is reduced to the operating value. The load of T1 is a *thermistor* (thermal resistor), R_T, a bead of material in which heat generated by a current through it causes its resistance to fall. A thermistor has a *negative temperature co-efficient* in contrast with most materials which increase their resistance with a rise in temperature.

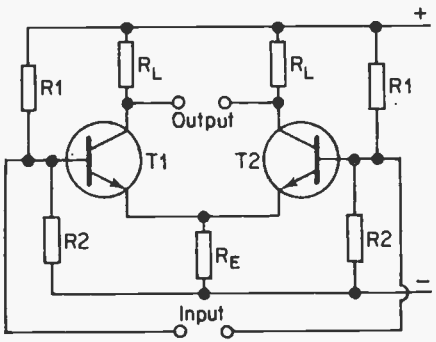
If temperature rises, the collector current of T1 rises, and the thermistor resistance falls. The increased voltage drop across R_T due to the increase in current is therefore compensated by the reduced voltage drop due to the fall in resistance, that is V_C for T1 remains almost constant. If temperature falls compensation is equally made. Because V_C of T1 remains constant, the bias applied to T1 is also held constant.

The *differential amplifier* or *emitter-coupled pair* has the two emitters directly connected with a common emitter resistor R_E as shown in the simplified circuit of Fig. 3.15(ii). Both transistors are forward-biassed by the same amount by resistor chains R1,R2 in conjunction with R_E causing equal collector currents to flow in the load resistors R_L and the combined emitter currents through R_E. No difference of potential therefore exists across the output terminals. If the temperature rises or falls R_E ensures that the bias and therefore

collector currents of both T1 and T2 change equally, thus collector potentials move in unison and again no p.d. exists across the output terminals. These are the static conditions, a potential or signal applied across the input terminals applies potentials, equal but opposite in phase to the two bases since the input resistance of both transistors is the same. The collector voltages swing in opposite directions, the output voltage being the difference between them, thus explaining the title of "differential" or "difference" amplifier. "Emitter-coupled pair" is also obvious.



(i) DC amplifier with thermistor compensation.



(ii) Differential amplifier.

Fig. 3.15 DC amplifiers.

3.2.7 Negative feedback

Feedback is a very important feature of electronics for although amplifiers can manage without it, they are much improved with it, while oscillators function because of it. Feedback occurs when a fraction of the output signal of a system is returned to the input just as an echo is a fraction of the voice returned to a talker. First consider an amplifier as a black box having a gain A as shown above the dotted line in Fig. 3.16. The amplifier output voltage is connected to a feedback network which inserts a fraction of it in series with the input circuit. We designate this fraction β (Greek, Beta) which might simply be derived from a potential-divider as shown separately.

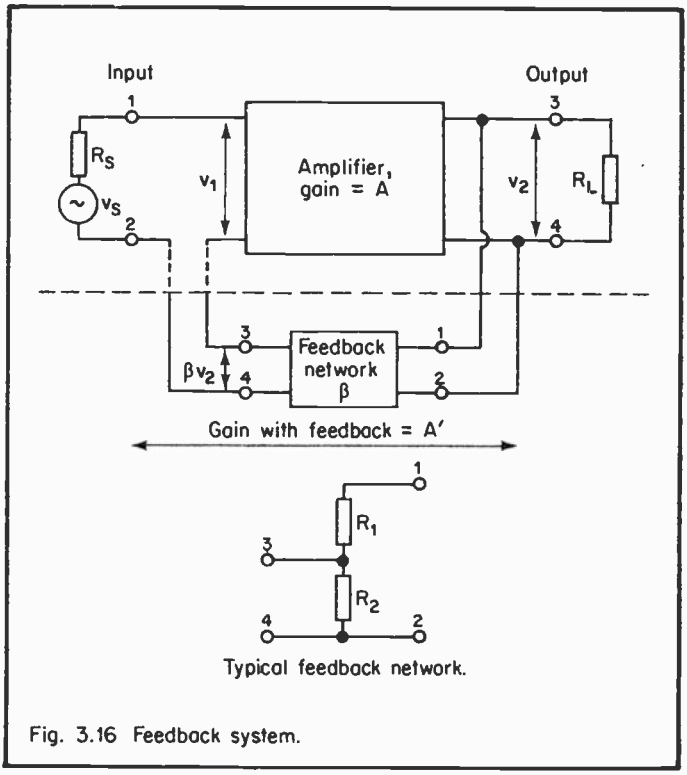


Fig. 3.16 Feedback system.

Now, v_2 is simply v_1 amplified A times, i.e. $v_2 = Av_1$ and v_2 at terminals 1 and 2 of the feedback network results in βv_2 at terminals 3 and 4. This is added in series with v_s, $\therefore v_1$ becomes $v_s + \beta v_2$ (ignoring the small effect of R_s) so that now feedback has been added, $v_2 = A(v_s + \beta v_2)$,

$\therefore v_2(1 - \beta A) = Av_s$ and the overall gain with feedback,

$$A' \doteq \frac{v_2}{v_s} = \frac{A}{1 - \beta A}$$

an important relationship applying to feedback systems generally.

Putting a few practical figures to this we might have an amplifier of gain (no feedback) 100, an input signal v_s = 10 mV, hence an output voltage of 10 mV x 100 = 1000 mV.

If R1 and R2 in the feedback network were preferred values of 56,000 Ω and 110 Ω respectively, β would be

$$\frac{R2}{R1 + R2} = \frac{110}{56{,}110} \approx 0.002$$

so feeding back 1000 x 0.002 = 2 mV.

Depending on which way round the network terminals 3 and 4 are connected in series with the input circuit, the feedback voltage of 2 mV will either add to v_s or subtract from it. If it adds (positive feedback), v_1 increases, v_2 increases, so does the feedback and provided that certain conditions are met, the whole system becomes electronically unstable and bursts into oscillation; discussion of this phenomenon is reserved for the next section (3.3). If the feedback voltage βv_2 subtracts from v_s, there is negative feedback and to maintain v_2 at 1000 mV and v_1 at 10 mV, v_s must be increased to 12 mV to satisfy the relationship $v_1 = v_s + \beta v_2$. For negative feedback β is negative, that is, in this example, −0.002.

With the negative feedback connected, the overall gain

$$A' \doteq \frac{A}{1-\beta A} = \frac{100}{1-(-0.002 \times 100)} = \frac{100}{1.2} = 83.3$$

which could also have been calculated from

$$\frac{v_2}{v_s} = \frac{1000}{12} = 83.3$$

The feedback circuit of Fig. 3.16 is known as *voltage output-series input*. Equally the feedback voltage may be applied across the input circuit (in shunt). Also the feedback signal may be made proportional to the output current and fed back in either of the two ways shown.

Negative feedback in amplifiers has several advantages, in spite of the fact that the overall gain falls.

3.2.7.1 Gain stability

If in the expression for gain with feedback, $A/(1-\beta A)$, βA is very much greater than 1, that is when A is high and β not too low, the denominator is approximately equal to $-\beta A$ and the expression to $A/(-\beta A) = -1/\beta$. The inference is that whatever happens to the transistor parameters (for example, by temperature changes) which would normally affect an amplifier gain, with feedback the gain does not change appreciably since it is always equal to $-1/\beta$. The latter is set by the designer. Evidently the greater the feedback, the better the amplifier stability.

EXAMPLE:

A multistage amplifier is designed for a gain of 20,000. What is the gain with feedback of $\beta = -0.001$? If a supply voltage variation reduces the gain by 25%, what is the percentage reduction in gain with feedback?

$$A = 20{,}000 \qquad \beta = -0.001$$

Then, gain with feedback,

$$A' \doteq \frac{A}{1-\beta A} \doteq \frac{20{,}000}{1+20} = \underline{952}$$

Gain with feedback with supply variation is equal to

$$\frac{20{,}000(1-0.25)}{1+0.001 \times 20{,}000(1-0.25)} = \frac{15{,}000}{16} = 937.5$$

∴ Percentage reduction in gain with feedback is equal to

$$\frac{952-937.5}{952} \times 100 = \underline{1.5}$$

Thus for a 25% fall in gain without feedback, there is only a 1.5% fall in gain when feedback is added.

3.2.7.2 Reduction of distortion

Distortion occurs within the amplifier and if a fraction of the harmonics present at the output is fed back into the input, then after being amplified the harmonics will appear at the output in antiphase to the original ones, thus tending to cancel. We can look at this a little more mathematically by considering the original harmonics present at the output as being due to an additional voltage generator v_h in series with the output so that v_2 which was originally equal to Av_s (Fig. 3.16) is now

$$v_2 = Av_s + v_h$$

In this case we consider adding feedback and raising the input level accordingly to see the change in harmonic voltage at the output, hence first determine the increased input signal required, v'_s.

Since gain with feedback

$$A' = \frac{A}{1 - \beta A},$$

then $$\frac{A'}{A} = \frac{1}{1 - \beta A} \quad \text{and} \quad \frac{A}{A'} = 1 - \beta A$$

and for equal output voltages, $A.v_s = A'.v'_s$

$$\therefore \quad v'_s = \frac{A}{A'} \cdot v_s = (1 - \beta A)v_s .$$

Then

$v_2 = A(v'_s + \beta v_2) + v_h$ and substituting for v'_s

$v_2 = Av_s(1 - \beta A) + \beta A v_2 + v_h$

$\therefore \quad v_2 - \beta A v_2 = Av_s(1 - \beta A) + v_h$

$\therefore \quad v_2(1 - \beta A) = Av_s(1 - \beta A) + v_h$

$$\therefore \quad v_2 = Av_s + \frac{v_h}{1 - \beta A}$$

showing that the distortion component in the output has been reduced from v_h to $v_h/(1 - \beta A)$, that is by $1/(1 - \beta A)$ and if βA is very much greater than 1, approximately by $-1/\beta$.

EXAMPLE:
An amplifier with a voltage gain of 400 and an output harmonic distortion content of 11% is to be provided with a negative feedback circuit to reduce the distortion to 1%. What feedback fraction (β) will be required and by how much must the preceding stage gain be raised?

Distortion with feedback is equal to

$$\text{Distortion without feedback} \times \frac{1}{1-\beta A}$$

$$\therefore \quad 0.01 = \frac{0.11}{1-(\beta \times 400)}$$

$$\therefore \quad \beta = \frac{-10}{400} = \underline{-0.025}$$

Gain of amplifier with feedback is equal to

$$\frac{A}{1-\beta A} = \frac{400}{1-(-0.025 \times 400)} = \frac{400}{11}$$

$\therefore$ Reduction in gain is equal to $400 \div 400/11 = 11$, meaning that the gain of the preceding stage must be raised by 11, or if this is not possible, then an additional stage having this gain is required. This is the price paid for reduction of the distortion.

3.2.7.3 Frequency response

The response of an amplifier over the frequency range for which it is designed (e.g. Fig. 3.6) can be improved (made flatter) by the addition of negative feedback. This can be proved mathematically but from what we have already learnt it is rather obvious because if β itself is made independent of frequency (e.g. by using a simple resistor divider as in Fig. 3.16) then for high values of βA, the amplifier gain, which approximates to $-1/\beta$ must also be relatively independent of frequency. Therefore the fall in gain usually experienced at the low and high ends of the frequency design range will be appreciably less.

3.2.7.4 Practical circuits

Three circuits, chosen to demonstrate many of the principles involved are given in Fig. 3.17. At (i) is an example of single-

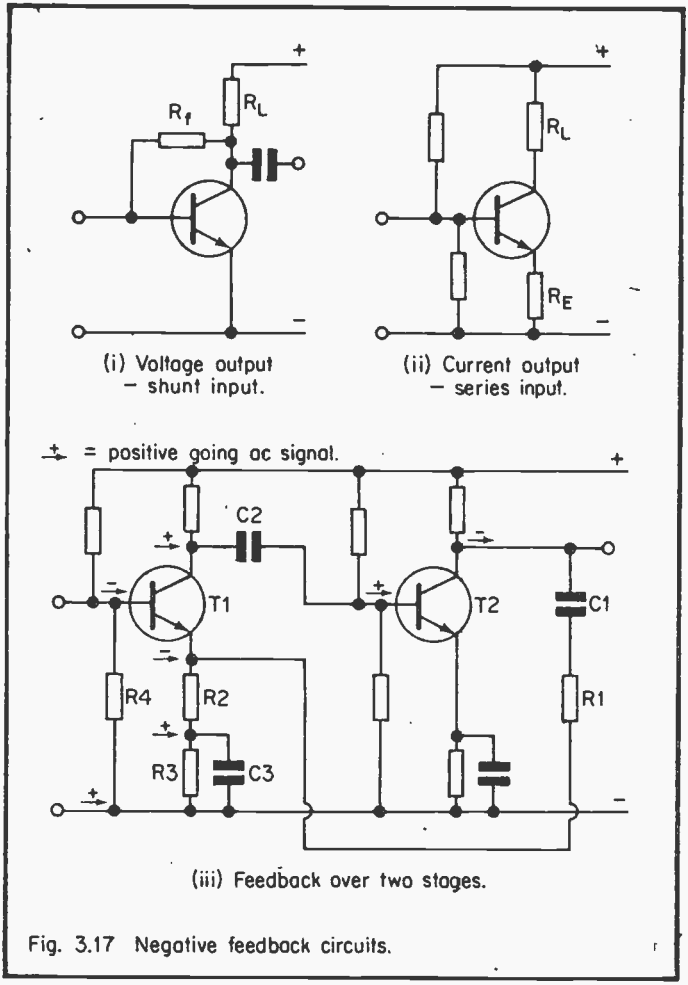


Fig. 3.17 Negative feedback circuits.

stage feedback, the circuit looking very similar to that for T1 of Fig. 3.10 because R_f not only provides feedback but also d.c. bias and stabilization. The feedback is negative because there is 180° phase difference between base and collector, for example, if the base swings positive, I_C increases, the voltage drop across R_L increases and hence V_C swings

negative – equally for a pnp transistor with directions changed accordingly. Analysis might proceed as follows :

$$i_C = h_{fe} . i_b$$

∴ Feedback voltage is equal to

$$i_C R_L = h_{fe} . i_b R_L$$

and feedback current is equal to

$$\frac{h_{fe} i_b R_L}{R_f}$$

(ignoring the small base-emitter resistance). This current is applied in parallel with i_b and opposes it, therefore the input current must have a total value of

$$i_b + \frac{h_{fe} i_b R_L}{R_f}$$

and the current gain with feedback is

$$\frac{h_{fe} i_b}{i_b\left(1 + \frac{h_{fe} R_L}{R_f}\right)} = \frac{h_{fe}}{1 + \frac{h_{fe} R_L}{R_f}} = \frac{h_{fe} R_f}{R_f + h_{fe} R_L}$$

Furthermore, if R_f is small compared with $h_{fe} R_L$ (i.e. the feedback is sufficiently large) this is approximately equal to R_f/R_L showing the gain to be relatively independent of transistor parameters.

Fig. 3.17(ii) shows an emitter resistor, one was used in Fig. 3.10 in the d.c. biassing system for T2 but by-passed by a capacitor so that the a.c. signal would not be affected. For negative feedback this capacitor is omitted and the a.c. signal current flowing through R_E develops a voltage across it which

is applied in opposition in the base-emitter circuit.

An analysis similar to that above shows that the gain may be made substantially independent of transistor parameters and approximately equal to R_L/R_E.

A complete 2-stage amplifier is shown in Fig. 3.17(iii) with feedback over the two stages. First we must check that the feedback is negative. Consider for example, a signal moving negative on the base of T1. Its collector goes more positive and so does the base of T2 (C2 has low reactance). Accordingly the collector of T2 goes negative. Ignoring the low reactance of C1 which is needed to block d.c. from T2 collector affecting T1 emitter, a fraction R2/(R1 + R2) of the output voltage of T2 is fed into the emitter circuit of T1 (C3 effectively puts the bottom end of R2 at common rail potential). Since the voltage across R2 is in such a direction as to oppose the original negative-going potential across R4, the feedback is negative. These polarity movements are indicated in the figure. The feedback angle will not be exactly 180° because of the capacitors in the chain. R3 in conjunction with R2 is part of the d.c. biassing arrangement for T1.

This is an example of voltage output – series input feedback.

3.3 OSCILLATORS

Paradoxically, in discussing the advantages of feedback in amplifiers in the previous section we discovered that positive feedback is detrimental and to be avoided at all costs. In contrast, in this section we examine the uses of positive feedback for actually creating instability in an amplifier so that in effect it changes into an *oscillator.* The basic principle follows from Fig. 3.16.

Let v_s result in a voltage v_1 at the amplifier input as shown, then $v_2 = Av_1$ and the voltage at terminals 3 and 4 of the feedback network = βAv_1.

We can visualize an oscillator as being an amplifier which supplies its own input, hence if v_1 can be obtained from the feedback network, v_s could be switched off, i.e. $\beta A v_1 = v_1$ which can only be true if $\beta A = 1$.

Due regard must be given to phase, so for positive feedback there must be a zero angle, hence the full condition for maintenance of oscillation is $\beta A = 1\angle 0^\circ$ (or 360° etc., mathematically expressed as $2n\pi$ where $n = 0, 1, 2, \ldots$).

However, oscillators do not need to be "started-up" by prior application of an external frequency, also there must be some control of frequency. Starting-up is automatic because any change (e.g. noise or a switching-on surge) which results in an output voltage, feeds back into the input and because in practice βA is made slightly greater than 1, results in an output voltage greater than the original one. The amplitude of v_2 therefore constantly increases until the amplifier becomes overloaded with a consequent fall in gain. The transit time around the amplifier plus feedback loop is so small that to an observer this would appear as an instantaneous process.

Control of frequency is usually by incorporation of a frequency-sensitive circuit (e.g. inductance and capacitance in a resonant circuit) or by a special arrangement of resistors and capacitors as shown later.

3.3.1 Resonant-circuit oscillators

Named after their originators, E. H. Colpitts and R. V. L. Hartley (both American transmission engineers), these oscillators maintain their frequencies by use of parallel, resonant (LC) circuits with a tapping connected to the emitter (common-emitter configuration) as shown in Fig. 3.18. The biassing arrangements are not included so that the drawing can show more clearly how zero phase-shift exists in the amplifier-plus-feedback loop. There is a 180° phase-shift from base to collector and again across the resonant circuit (shown in the figure for a negative-going signal at the base). It is only at resonance that the impedance of an LC

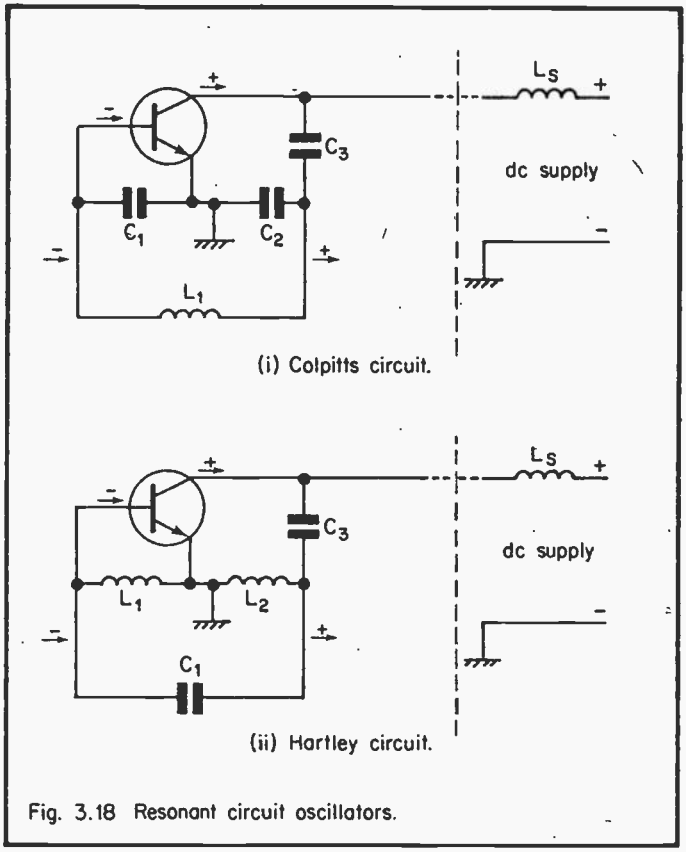


Fig. 3.18 Resonant circuit oscillators.

circuit is resistive with a phase difference across the whole circuit of 180°. At resonance therefore the signal presented at the input terminals of the amplifier has shifted through 180° + 180° = 360°, so is in-phase, giving positive feedback. Any reactance within the loop shifts the phase angle, hence the reactance of C3 which blocks collector d.c. potential from the base, must be negligible compared with the impedance of the resonant circuit at the working frequency.

L_s is an inductor in the d.c. supply of sufficiently high

reactance at the oscillation frequency to avoid the shunting effect of the supply on the oscillator itself.

With careful design so that βA is just above $1\angle 0°$ to maintain oscillation yet not excessively so for the transistor to be driven into distortion, these oscillators can produce reasonably accurate sine waves. Common-emitter configurations are shown in Fig. 3.18, the other two configurations are equally applicable. An advantage of the Hartley oscillator is that a single variable capacitor may be used for C_1 for variation of the oscillator frequency, for the Colpitts circuit, two variable capacitors are necessary.

The frequency of oscillation (f_0) of the Colpitts circuit is given by the formula for the resonant frequency of the $L_1 C_1 C_2$ circuit, hence if C represents the net capacitance of C_1 and C_2

$$f_0 = \frac{1}{2\pi\sqrt{L_1 C}}$$

and in terms of C_1 and C_2 this becomes

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{1}{L}\left(\frac{1}{C_1}+\frac{1}{C_2}\right)}$$

since

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2}$$

The Hartley circuit differs slightly because there may or may not be coupling between the two inductors. If none and if L represents the net inductance, i.e. $L = L_1 + L_2$

$$f_0 = \frac{1}{2\pi\sqrt{LC_1}} = \frac{1}{2\pi\sqrt{(L_1 + L_2)C_1}}$$

Coupling between the two inductors introduces M, the mutual

inductance. If two coils are connected in series and are mutually coupled, the combined inductance is given by $L_1 + L_2 \pm 2M$ depending on whether they are in series-aiding or opposition. The added quantity 2M arises from the fact that both coils are affected by flux-linkages with the other. It is unlikely that connexion would be made for series-opposition, therefore

$$f_0 = \frac{1}{2\pi\sqrt{(L_1 + L_2 + 2M)C_1}}$$

These formulae for f_0 contain some approximation because no account has been taken of the slight effect the transistor parameters have on the resonant circuits.

3.3.2 Resistance-capacitance oscillators

Since no inductance is involved, the principle of using a resonant circuit does not apply. Instead a resistance-capacitance network is used to provide the feedback with its correct phase-shift. Such oscillators are usually used at the lower frequencies. One common method of obtaining the required phase-shift which has the practical advantage that the frequency may be varied, is illustrated in diagrammatic form in Fig. 3.19. The amplifier has two transistor stages so that its overall phase-shift is $2 \times 180°$. The feedback network contains two equal resistors, R and two equal, simultaneously variable capacitors, C. These are arranged as shown in series and parallel combinations with impedances Z_s and Z_p, respectively and forming a potential divider across the amplifier output. The fraction of voltage fed back to the amplifier input is equal to

$$\frac{Z_p}{Z_p + Z_s} = \beta.$$

β can be analysed most easily using complex notation. Capacitive reactance is 90° out of phase with resistance and therefore is preceded by j, it is also given a negative sign. Each step is

shown as a reminder :

Impedance of parallel network:

$$\frac{1}{Z_p} = \frac{1}{R} - \frac{\omega C}{j} = \frac{j - \omega CR}{jR}$$

$$\therefore \quad Z_p = \frac{jR}{j - \omega CR}$$

Impedance of series network:

$$Z_s = R - j\omega C$$

Since

$$\beta = \frac{Z_p}{Z_p + Z_s},$$

$$\beta = \frac{\dfrac{jR}{j - \omega CR}}{\dfrac{jR}{j - \omega CR} + R - \dfrac{j}{\omega C}} = \frac{\dfrac{jR}{j - \omega CR}}{\dfrac{jR}{j - \omega CR} + \dfrac{\omega CR - j}{\omega C}}$$

$$= \frac{\dfrac{jR}{j - \omega CR}}{\dfrac{j\omega CR + (\omega CR - j)(j - \omega CR)}{\omega C(j - \omega CR)}}$$

$$= \frac{\dfrac{jR}{j - \omega CR}}{\dfrac{j\omega CR + j\omega CR - \omega^2 C^2 R^2 + 1 + j\omega CR}{\omega C(j - \omega CR)}}$$

$$= \frac{jR}{j - \omega CR} \times \frac{\omega C(j - \omega CR)}{3j\omega CR + 1 - \omega^2 C^2 R^2}$$

$$= \frac{j\omega CR}{3j\omega CR + 1 - \omega^2 C^2 R^2}$$

$$= \frac{1}{3 + \dfrac{1}{j\omega CR} - \dfrac{\omega^2 C^2 R^2}{j\omega CR}}$$

$$= \frac{1}{3 - \dfrac{j(1 - \omega^2 C^2 R^2)}{\omega CR}}$$

The impedance angle and therefore the phase-shift is zero when the imaginary (j) term is equal to 0, i.e.

$$\frac{1 - \omega^2 C^2 R^2}{\omega CR} = 0$$

$$\therefore \quad 1 - \omega^2 C^2 R^2 = 0$$

$$\therefore \quad \omega^2 C^2 R^2 = 1$$

and by taking square roots of both sides:

$$\omega CR = 1 \qquad \therefore \omega = \frac{1}{CR}$$

Hence

$$f_0 = \frac{\omega}{2\pi} = \frac{1}{2\pi CR}$$

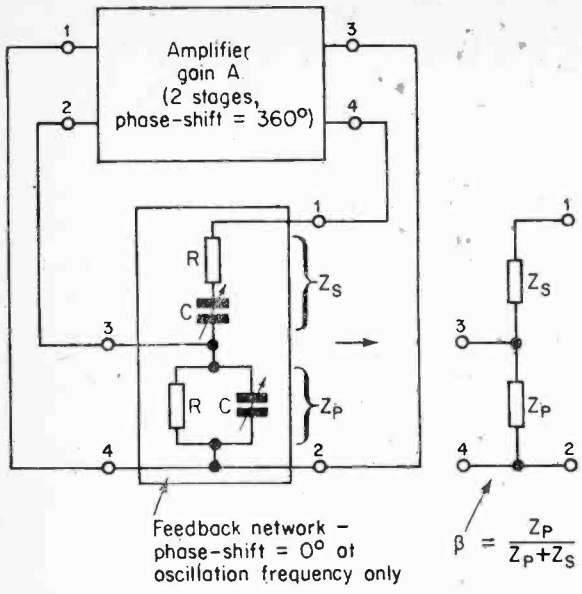


Fig. 3.19 Resistance – capacitance oscillator.

The gain A of the amplifier plus feedback network follows for at the frequency f_0, since the j terms are equated to zero, $\beta = \frac{1}{3}$ and therefore since $A\beta$ must equal 1, $A \times \frac{1}{3} = 1$,

$$\therefore \underline{A = 3}.$$

In this case two stages of the amplifier are necessary to provide a 360° phase-shift, they are not needed for high gain. This particular type of resistance-capacitance oscillator is also known as the *Wien Bridge* oscillator (after M. Wien, a German engineer).

3.3.3 Crystal-control

The above oscillator circuits have been selected to demonstrate

the main principles. It may at first be thought that *crystal-controlled* oscillators function on a different principle altogether, but this is not so. Several materials exist, the crystals of which exhibit the *piezo-electric* effect, particularly quartz and Rochelle salt. Quartz is the most commonly used and in fact has given rise to the general name *quartz crystal.* The mechanism is that if pressure is applied (hence, "piezo", from the Greek, to press) across two faces of a crystal, electric charges appear on the surface, conversely an electric field deforms the crystal. The crystal has a natural mechanical resonance frequency just as does a piano wire and an alternating electric field produced by a transistor on the crystal at this frequency produces the greatest effect and the varying electric charges may be fed back to the transistor under normal principles to maintain the oscillation. In fact, the crystal may be considered as an electronic resonance circuit having calculable resistance, inductance and capacitance. A quartz crystal can be cut to have mechanical features of very high stability therefore an oscillator employing one as its frequency control will suffer very little frequency deviation with time or temperature, this is evident from the ever increasing use of quartz crystal control of watches and clocks.

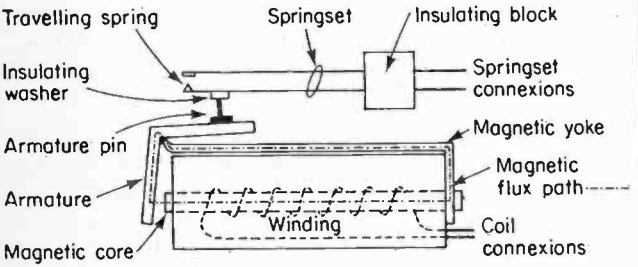
The oscillators described above are generally for sine waves, the generation of other waves such as square, pulse and saw-toothed which are mainly used in switching and control functions will be better appreciated from the next section in which transistors are less used in the *analogue* mode where the circuits are designed for following the magnitude of the input signal (e.g. the amplifiers of Section 3.2), but more as switches in a *digital* mode where the output usually has two states only, on or off, as we shall see in Section 3.4.

3.4 SWITCHING

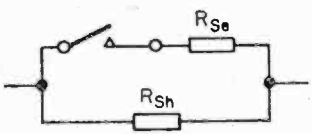
Section 1.4.4 introduced electromechanical switches, we might perhaps pause in our consideration of semiconductors to become a little more conversant with these because although semiconductors and this type are both labelled as

switches, there is considerable difference between them.

An electromechanical switch of the telephone switching type, more generally known as a *relay* is illustrated in Fig. 3.20(i). Since we have already studied the principles of electromagnetism it should be clear that when a current of sufficient magnitude flows in the winding, the magnetic flux set up attracts the armature, the armature pin rises and pushes the bottom spring contact upwards until the two contacts "make" and the switch is operated or "on". With a "break" springset the two springs are in contact until operation of the relay lifts the top spring and the circuit is "off". Various other springset arrangements are also used and a single relay can operate several springsets at once provided that the ampere-turns (NI) are sufficient to move the armature against the tension of the travelling springs.



(i) Relay.



(ii) Equivalent circuit of contacts.

Fig. 3.20 Electromechanical switch (relay).

If we examine the contacts electrically the equivalent circuit would appear as in Fig. 3.20(ii) where R_{sh} represents the shunt resistance across the contacts due to their being fixed in an insulating block and R_{se} represents the series resistance of the contacts themselves. For such a relay as shown the values approach the ideal, by use of high resistivity materials R_{sh} is usually tens of megohms upwards and with silver or platinum coated contacts R_{se} may be practically zero, certainly only a very small fraction of one ohm.

Diodes and transistors cannot match these features, however they have one great asset apart from smaller size, that of speed of operation. Compared with them relays are sluggish, not only does it take time to build up current in the coil to the operating value but further time elapses while the contacts are travelling between the rest and operated positions.

We can now move on to semiconductors generally and in a way assess their efficiency as switches by considering what sorts of values R_{se} and R_{sh} might have and the switch time of operation.

3.4.1 The diode as a switch

The diode operates as a *polarity-sensitive* switch, it is "on" with forward polarity, "off" with reverse. Unlike the relay where the operating and switched circuits may be completely separated, the diode is part of the circuit being switched.

3.4.1.1 Switching resistances

The circuit which is being switched must have some resistance, hence the load line comes to our aid as previously shown in Figs. 2.4 and 2.5, in fact the latter figure can be used as an example of determining approximately R_{se}. We first look at the switch in its closed position whereupon R_{sh} is almost inoperative. If in Fig. 2.5 with a voltage of 1.4 applied and $R_L = 3.5\ \Omega$, the circuit current is found to be 225 mA, then the total circuit resistance is equal to

$$\frac{1.4}{225} \times 1000 = 6.22\ \Omega,$$

therefore

$$R_{se} = 6.22 - 3.5 = 2.72\ \Omega$$

Again, when V = 1.2 V and $R_L = 2\ \Omega$, $R_{se} = 2.36\ \Omega$, a slightly lower resistance because of the greater forward current.

If in this second case the battery voltage is increased to, say, 1.6 V and a new load line drawn for 2 Ω, we should find the circuit current to be about 430 mA, giving $R_{se} = 1.7\ \Omega$, thus the higher the applied voltage the lower the value of R_{se}, a not unexpected result. The limit is at the maximum forward current as quoted by the manufacturer.

R_{sh} could similarly be obtained by the load line technique except for the fact that the scales may become unmanageable since the diode currents are usually of the order of μA or nA. Little is lost therefore by simply calculating the resistance of the diode at the applied voltage giving R_{sh} approximately but easily. Thus for a typical silicon planar epitaxial diode at a junction temperature of 25°C, the reverse current might be 2 nA at 1.5 V giving R_{sh} as 750 MΩ. At higher junction temperatures, R_{sh} may fall considerably for example, for the above diode at 150°C when V = 5 V, the current rises to 8 μA, giving $R_{sh} = 625$ KΩ. Diode resistance values vary over a wide range however.

3.4.1.2 Switching times

Time of switching, that is, for changeover from "on" to "off" or "off" to "on" is extremely short compared with that of a relay yet it cannot be neglected because computer and similar devices need to perform thousands of operations in a fraction of a second – obviously the relay cannot compete at all.

Delay occurs mainly on changeover from forward to reverse.

With a forward-biassed junction, holes and electrons diffuse to opposite sides as shown in Fig. 1.5(ii). This can be looked upon as a storage of charges and when the diode moves swiftly into reverse bias these must be removed to obtain the "empty" condition of Fig. 1.5(i). Thus for a short period of time, although the diode is reverse-biassed, current continues to flow as shown in Fig. 3.21. Because a charge is held, this implies capacitance, it is given the symbol C_f and is known as the *storage* or *diffusion* capacitance. This when discharging produces a current in the opposite direction, hence the swing in Fig. 3.21 of forward current from + to − at the reversal time t_r. The positive and negative currents are equal in magnitude as shown. It seems that we have now exceeded considerably the reverse saturation current, a condition only to be expected under avalanche conditions. However it is all part of the mechanism of the diode, is of extremely short duration and therefore has nothing to do with breakdown.

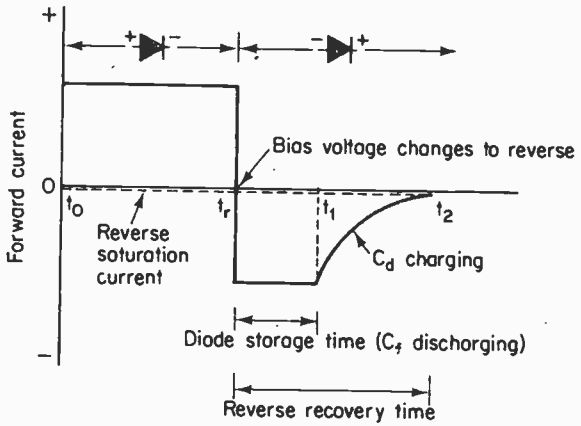


Fig. 3.21 Diode recovery and storage times.

The effect is rather complicated but we can look upon it in a simplified way by considering that from time t_r to t_1 the discharge maintains the current at a fairly constant value because it is controlled mainly by R_L in the external circuit. At t_1 therefore the diode forward capacitance is effectively discharged. Now because reverse bias causes an increase in the depletion layer (Fig. 1.5(i)) with consequent separation of charges, this is also equivalent to creation of capacitance (the reverse-bias capacitance, C_d) which will charge until the potential across it reaches the value of the applied reverse potential, hence the exponential curve from t_1 to t_2.

The total time $t_r - t_2$ is known as the *reverse recovery time*, designated by t_{rr} and is quoted by the manufacturer usually in ns together with the appropriate test conditions. t_{rr} varies from several hundred ns down to a few ns in specially constructed *high-speed* diodes.

Our explanation although somewhat simplified and incomplete shows how the time element in a diode, although of no consequence whatsoever in power mains rectification, could even preclude its use at high frequencies. At 1 MHz for example, the periodic time is 1 μs (= 1000 ns), thus a diode with a reverse recovery time of several hundred nanoseconds cannot switch correctly. Going one stage further, at 10 MHz such a diode would never reach reverse resistance because the applied waveform will have switched back to forward before the time t_2 is reached.

Thus, compared with a relay the diode as a switch is inferior in having greater "on" resistance, lower "off" resistance but it can switch very much faster, in a time of a few ns compared with the relay in ms. The diode is considerably more affected by temperature changes.

3.4.2 The transistor as a switch

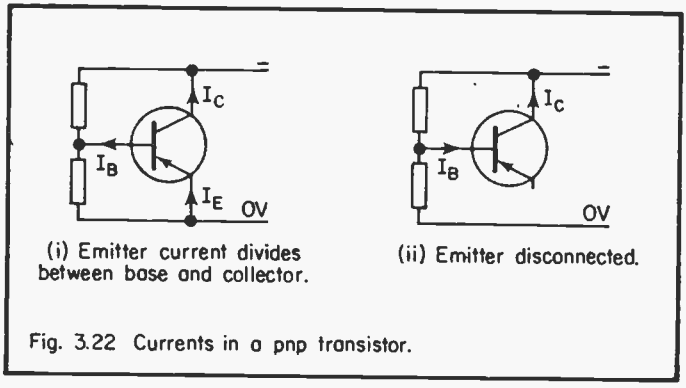
The limitation in the use of the diode as a switch mentioned in Section 3.4.1 in that it cannot switch a circuit separate from the one from which it derives its control applies much less for the transistor. For this device the controlling and switched

circuits are more but not completely separated, moreover the transistor is capable of switching a moderately large current in response to control from a very much smaller one. The relay on the other hand can also control a large current but needs considerably more power from the controlling circuit.

In considering the transistor as a switch we are no longer concerned with linearity to keep distortion at a minimum, in fact operation is entirely on the non-linear sections of the characteristics.

3.4.2.1 Switching resistances

It is first necessary to establish when a transistor is said to be "off", it would be convenient to say this happens when the collector current $I_C = 0$, but we have to contend with leakage currents. Fig. 3.22(i) shows the current directions in a pnp transistor and here we consider hole current, not electrons, so that we remain in conformity with the emitter arrow in the transistor symbol. I_B and I_C flow out of the transistor and are considered negative so that $I_E = -(I_B + I_C)$, the normal relationship (Section 2.2.2).



(i) Emitter current divides between base and collector.

(ii) Emitter disconnected.

Fig. 3.22 Currents in a pnp transistor.

But suppose the emitter is disconnected, hence $I_E = 0$ and $I_B = -I_C$. Fig. 3.22(ii) shows the new condition and we know that I_C does not fall to zero because of the collector-base leakage current I_{CBO}. For this small value of I_C to

continue it is obvious that current can only enter via the base in the form of a reversed base current. Therefore as I_E approaches 0, I_C falls to I_{CBO} and I_B changes from negative to positive, hence I_B must pass through 0. The graph of I_C plotted against I_B for this region is of the shape as shown typically in Fig. 3.23. When $I_B = 0$ the value of I_C must be the collector-emitter leakage current I_{CEO}. As I_B moves positive I_C falls to the lower value of the collector-base leakage current, I_{CBO}.

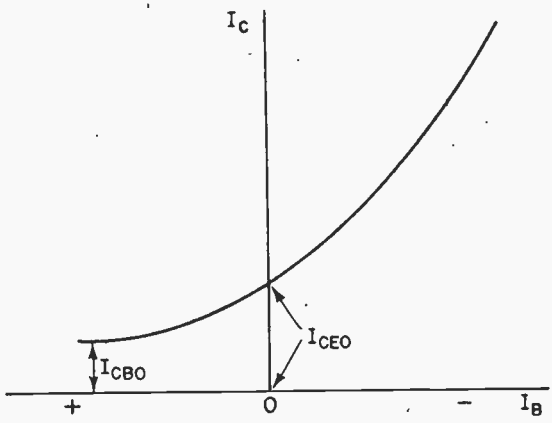


Fig. 3.23 Collector current as base current reverses (pnp).

What we have determined therefore is that for a pnp transistor, negative values of I_B give rise to the normal family of output characteristics, at $I_B = 0$ a small collector current still flows and at slightly positive values of I_B, I_C falls to I_{CBO} and cannot be further reduced. Similarly for an npn transistor with polarities changed accordingly.

From this we are now in a better position to understand the limiting conditions on the transistor as a switch. Consider the family of output characteristics of Fig. 3.24. The collector

currents for $I_B = 0$ and $I_B = I_{CBO}$ have been deliberately enlarged for explanation, in practice they are usually of the order of μA or less and therefore would not even show on a graph of these scales. As an example, a load-line for 100 Ω has been drawn for a supply of 20 V. As I_B is varied over its range from the small reverse value to 1.5 mA the collector current swings from some 10 mA to about 180 mA, that is the small base current change has switched the larger collector current from "off" to "on" (remember especially that the 10 mA figure is not realistic). The *cut-off* point shows that the "off" condition of the transistor does not have infinite resistance because some current flows. Modern switching transistors are, however, specially designed for low collector cut-off currents in the nanoampere range thus producing

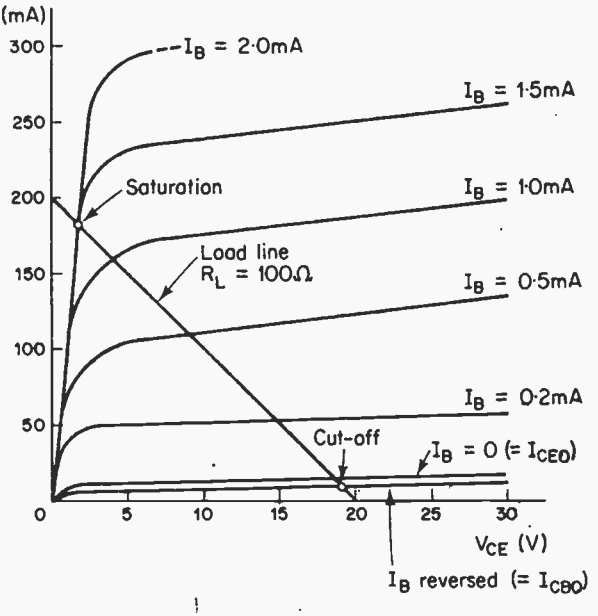


Fig. 3.24 Cut-off and saturation for switching transistor.

switch resistances (R_{sh}) of the order of tens of megohms upwards.

At *saturation* any increase in I_B above 1.5 mA has no effect on the saturation point, the transistor cannot operate further up the load line to reduce V_{CE} or increase I_C. The "on" resistance (R_{se}) at this point is of the order of 1.75 V/180 mA = 9.72 Ω, again a figure made artificially high by the need to draw curves for visual demonstration. In practice values lower than this are normally obtained. The *collector cut-off current*, I_{CBO} and the *collector-emitter saturation voltage*, $V_{CE(sat)}$ are usually quoted by the manufacturer.

Thus the transistor is usable as an efficient switch and with very small switching times as will be seen from the next section.

3.4.2.2 Switching times

As with the diode, switching delays occur in the transistor, naturally with two junctions compared with one for the diode the mechanism is more complicated. It will suffice therefore if, having examined conditions in the diode in some detail, we simply list those affecting the transistor with brief notes on their origins.

Fig. 3.25 shows an ideal voltage pulse applied to the base and below this is the graph of collector current (as would be shown on an oscilloscope).

The significance of the 10% and 90% collector current values is that these are the points at which the transistor may be considered to have switched "off" or "on" respectively. When the input voltage pulse v_b switches to "on" two delays affect the rise of i_c.

(i) the *delay time*, t_d. The two junction capacitances which make up the total reverse-bias capacitance C_d are being charged from below cut-off to the 10% value of i_c. The time t_d will be increased according to the drive beyond cut-off previously made to obtain the "off" condition.

(ii) the *rise time*, t_r. This is the time taken for i_c to rise from 10% to 90% of its final value. It is the time delay in charging the CR circuit comprising the collector-base capacitance and load resistance.

The total time $(t_d + t_r)$ is known as the *turn-on time*, t_{on}.

When the input voltage pulse switches back to "off" two further delays occur before the transistor is also considered to be "off":

(i) the *storage time*, t_s. The transistor has been at saturation point and a relatively large charge has accumulated in the base, this has to be removed before i_c can fall and is considered to have occurred at the 90% point.

(ii) the *fall time*, t_f. As with rise time, this delay is due to the discharge of the same CR circuit.

The total time $(t_s + t_f)$ is known as the *turn-off time.*

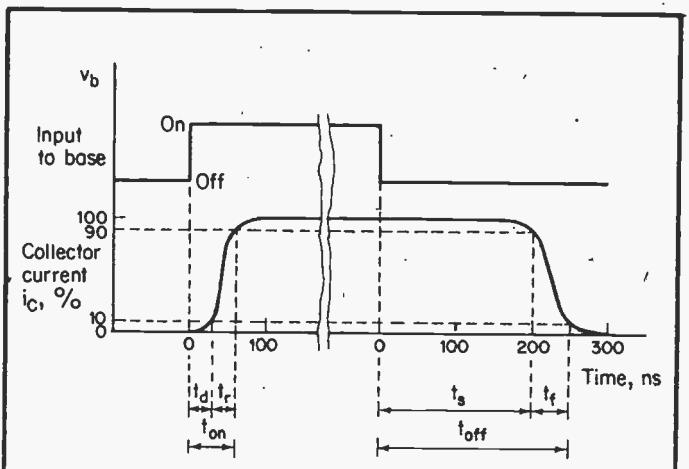


Fig. 3.25 Transistor delay times.

Fig. 3.25 shows typical practical delay times for a "general purpose" transistor, however in practice these vary over a wide range. Specially developed switching transistors may have turn-on and turn-off times as low as 10 ns (one hundred millionth of a second, too small for human comprehension, yet still undesirable in some applications).

3.4.2.3 The bistable multivibrator

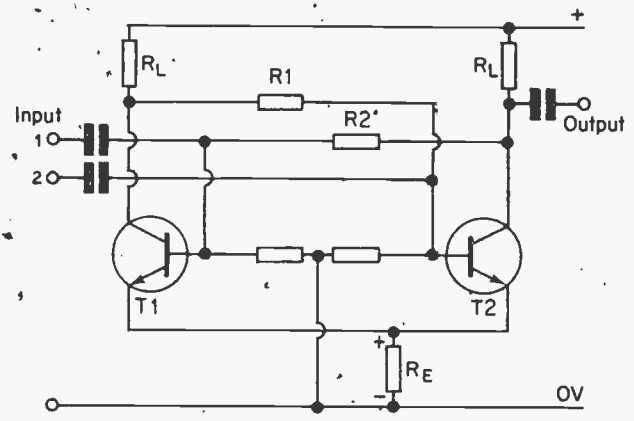
The switching sequence so far discussed can be likened to a bell-push where the bell rings while the button is pressed but stops as soon as the button is released. The bell-push is known as a *non-locking* switch. Alternatively an electric light switch is of the *locking* type, it stays where it was last put.

Additional circuitry can build this facility into transistor switching, a commonly used circuit being that of a *multivibrator*, this also gives us an opportunity to examine a practical circuit in which transistors are used as switches. As such they normally rest in their cut-off or saturated states (we will simply use the terms "on" and "off"), the load line between the two states being traversed within the very short times shown typically in Fig. 3.25.

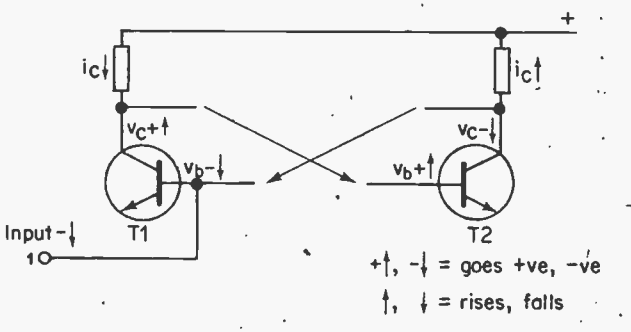
A basic multivibrator circuit is shown in Fig. 3.26(i). The circuit is classed as *bistable* because it is stable (stays put) in either of two states, T1 "on", T2 "off" or T1 "off", T2 "on". Let us start with the assumption that T1 is "on" and if necessary refer back to Fig. 3.24 to refresh our memories about "on" and "off" conditions.

Current flow through the common emitter resistor R_E is such that it provides a negative potential on T1 and T2 bases. For npn transistors this is reverse bias which would normally keep both transistors "off" (cut-off point on Fig. 3.24). However, because for example, T2 is "off", its collector is highly positive and this potential is transmitted to T1 base via R2, greatly exceeding the standing negative bias potential on T1 and biassing it "on".

At the same time the low potential on T1 collector fed via R1 to T2 base does not overcome the standing bias and T2 can remain "off". This is one of the stable states and the output terminal is at high potential.



(i) Basic circuit.



(ii) Change-over conditions.

Fig. 3.26 Bistable multivibrator.

Suppose next a short-duration negative pulse (generally known as a *trigger* pulse) is applied to input terminal 1 (that is, to the base of T1) and that the pulse is of sufficient amplitude to overcome the positive bias applied from T2. The collector current of T1 commences to fall, the bias to T2 rises positively and T2 collector current rises with a consequent fall in collector voltage, that is, it becomes less positive. Via the link R2 to T1 it therefore aids the fall of T1 collector current. The changing conditions are shown diagrammatically in Fig. 3.26(ii). The two transistors are therefore working together to bring T1 to "off" and T2 to "on" where they remain in the alternative stable state until an appropriate pulse arrives at an input terminal to swing them back. With T2 now "on", the output terminal is at low potential.

Thus the arrival of the pulse on terminal 1 has switched the output circuit from "on" to "off", it remaining thus even though the pulse has ceased. A negative pulse applied to terminal 2 will switch the output circuit to "on" Positive-going pulses are equally effective applied to terminal 1 to switch "on" and to terminal 2 for "off".

The multivibrator can also run freely, switching between the two states at a frequency determined by the circuit components. This is the *astable* circuit which is capable of generating a square waveform, that is, it is an oscillator. It is from this type that the name "multivibrator" was originally derived.

3.4.3 Switching logic

A brief excursion into the fundamental principles on which digital computers are based is also helpful in developing a wide-spread acquaintance with semiconductors for it is in the computer and automation fields that diodes and transistors are used in profusion. By having some understanding of *switching logic* (logic = science of reasoning) we shall begin to see how computers understand instructions, carry out their tasks and if so programmed, make their judgments. Unfortunately from the computer's point of view, it would have been better if man had chosen a numbering system based on

his two arms rather than his ten fingers. The *binary* system, based on 2 instead of 10 is described in Appendix 3. For those who have no acquaintance with the system it is suggested that this Appendix is read in its entirety next with reference made to it as necessary as this chapter proceeds.

3.4.3.1 Mathematical logic

Boolean Algebra (after George Boole, an English mathematician) is a system which combines logic with mathematics. It may be a somewhat confusing concept for the newcomer to conventional mathematics to grasp, so for that reason we refrain from delving too deeply and rather than getting involved too much with the algebra itself we will relate it to electrical switching circuits to understand the practical side as early as possible.

We use logic in our own thought processes and for many people the fascination of Sherlock Holmes and other great detectives is the logical reasoning through which they solved crimes. If two burglars A and B are both caught with stolen goods in their pockets we reason that both are guilty, for this we write in algebra A AND B (are guilty), shortened to A.B. This is a different algebra from the conventional sort so A.B does not mean A multiplied by B, it simply stands for A AND B. Now if there is doubt as to which of the two burglars blew the safe then it must be A OR B and this is written A+B, the plus sign nows meaning "OR". Of course each man denies his guilt, their separate statements being

NOT A, NOT B, written as $\overline{A}$, $\overline{B}$,

so A.B stands for A AND B
A+B stands for A OR B
$\overline{A}$ stands for NOT A
$\overline{B}$ stands for NOT B .

Throughout this discussion of burglary there have on each occasion been only two answers to the question, guilty or not guilty, a yes/no type of answer, or in electronics true or false,

designated 1 or 0. We now begin to see a link with the binary arithmetic of Appendix 3.

Let us return to the question as to who blew the safe. It is thought that one of the two burglars did the job but they could have had an accomplice. Assigning 0 for a false statement and 1 for a true statement, a *truth table* can be built up as follows:

A	B		
0	0	(both lying)	A.B = 0 (means that A AND B are lying) A+B = 0 (means that either A OR B is lying) $\overline{A}$ = 1 (means that A is NOT telling the truth) $\overline{B}$ = 1 (means that B is NOT telling the truth)
0	1	(A lying, B telling truth)	A.B = 0 (together the result is not the truth) A+B = 1 (one or the other is telling the truth) $\overline{A}$ = 1 (A is NOT telling the truth) $\overline{B}$ = 0 (B is NOT telling a lie)
1	0		as for A = 0, B = 1, except A and B reversed
1	1	(both telling truth)	A.B = 1 (together their statements are truthful) A+B = 1 (one or the other is telling the truth – in fact both are) $\overline{A}$ = 0 (A is NOT lying) $\overline{B}$ = 0 (B is NOT lying).

$\overline{A}$ and $\overline{B}$ are seen to be simply reversals of A and B respectively, i.e. if A = 0, $\overline{A}$ = 1, if A = 1, $\overline{A}$ = 0.

Sherlock Holmes would rightly have eschewed this truth table because, he would have said, it is all so obvious. But in complex electronic switching circuits truth tables have much to offer in design of switching systems for they ensure that all

unnecessary circuitry is avoided. An example follows later but now we move from simple logic expressed in a mathematical form to its application in switching.

3.4.3.2 Circuit logic

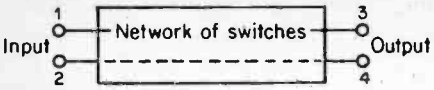
The above can without too much stretching of the imagination now be related to switches where 1 indicates a through circuit and 0 an open circuit. In the general diagrams we use, the switching is shown between terminals 1 and 3 only, as in Fig. 3.27(i), the remainder of the connexion between terminals 2 and 4 is omitted for convenience. Fig. 3.27(ii) shows two switches in a 0 condition, hence the circuit between terminals 1 and 3 is similarly 0. This agrees with the truth table that given $A = 0$ and $B = 0$, $A.B = 0$. The condition $A = 0$, $B = 1$ shows that switch B has operated but this still does not complete the circuit because it is broken at A, hence $A.B = 0$. The same holds for $A = 1$, $B = 0$. Finally $A = 1$, $B = 1$ means that both switches are operated (through), hence the circuit (terminals 1 to 3) is through.

Thus is demonstrated the AND condition. The OR arrangement in (iii) shows that if either A or B or both are operated, the circuit is through as evident from the truth table.

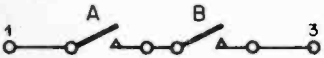
To obtain the reverse condition for NOT needs a contact which is normally made (through) but which disconnects the circuit when operated as in (iv). When $A = 0$ the circuit is through and when $A = 1$ the switch is operated and the circuit is open. A truth table for the two switches A and B in NOT functions will show that of the four different conditions possible, only $A = 0$, $B = 0$ gives a circuit of 1. In fact this produces the NOR function illustrated at (v).

The NOR has a truth table reversed compared with the OR ((iii) in the figure), for this reason it is classed as a negative OR or NOR. One more very likely to be met later is the negative AND or NAND. Its truth table is the reverse of that for AND.

Each of these functions is known in the practical circuit as a *gate*. Although the diagrams indicate that they comprise

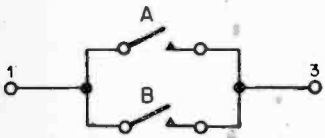


(i) General arrangement.



A	B	Circuit
0	0	0
0	1	0
1	0	0
1	1	1

(ii) Switches arranged A.B (AND).

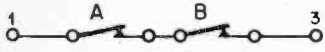


A	B	Circuit
0	0	0
0	1	1
1	0	1
1	1	1

(iii) Switches arranged A + B (OR).

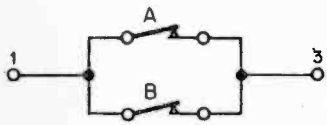
A	Circuit
0	1
1	0

(iv) NOT or inversion function.



A	B	Circuit
0	0	1
0	1	0
1	0	0
1	1	0

(v) NOR function.



A	B	Circuit
0	0	1
0	1	1
1	0	1
1	1	0

(vi) NAND function.

Fig. 3.27 Application of logic symbols to switches.

mechanical switch contacts, in electronic switching the function is provided in some way by a semiconductor circuit at much greater speed. This is the subject of the next section.

One simple example only is given to demonstrate the usefulness of switching logic to the circuit designer.

EXAMPLE:
The switching system in Fig. 3.28(i) has two relays, each with three separate contacts. Relay A has two make contacts and one break contact. Relay B has the same. Reduce the system to the simplest possible.

There are many theorems (algebraical rules) in Boolean Algebra which may be called upon for simplification of complex circuits, we need only 3 for this example:

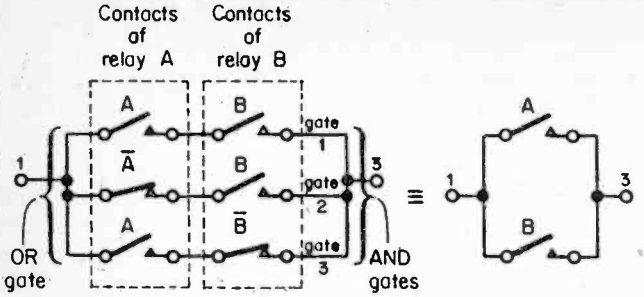
(i) $A.B + \bar{A}.B = B(A + \bar{A})$ [B AND (A OR $\bar{A}$)]
This appears to follow the normal (non-Boolean) algebraical rule, but we cannot take this for granted, so it should be proved. Set up the circuit as in Fig. 3.28(ii), that is, the two AND arrangements A.B and $\bar{A}.B$ connected in parallel to form an OR system. Applying 0 or 1 to A and also to B shows that for the circuit to be 1, whatever happens to A, B must be 1, therefore a single B contact is sufficient, giving $B(A + \bar{A})$.

(ii) $(A + \bar{A}) = 1$
This is shown in the figure at (iii) and it is evident that whether A is at 0 or 1 there is always a through path, i.e. the circuit is always 1.

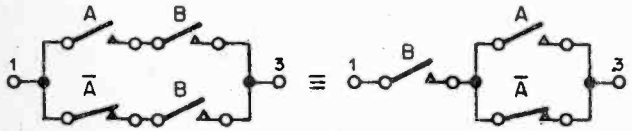
(iii) $A + \bar{A}.B = A + B$
The truth table in Fig. 3.28(iv) is seen to be identical with that for A + B (Fig. 3.27(iii)). Equally $A.\bar{B} + B = A + B$.

It is now possible to reduce the system which is of 3 sets of AND gates connected to an OR gate:

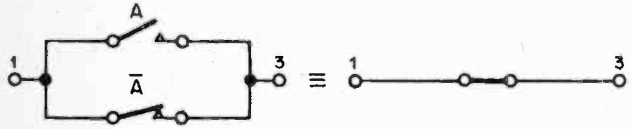
from (i): $A.B + \bar{A}.B + A.\bar{B} = B(A + \bar{A}) + A.\bar{B}$



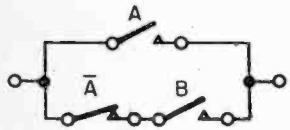
(i) System of switches and its equivalent.



(ii) Demonstration of $A.B + \bar{A}.B = B(A + \bar{A})$.



(iii) Demonstration of $A + \bar{A} = 1$.



A	B	Circuit
0	0	0
0	1	1
1	0	1
1	1	1

(iv) Truth table for $A + \bar{A}.B$.

Fig. 3.28 Reduction of o system of switches.

from (ii): $= B + A.\overline{B}$

from (iii): $= A + B$

i.e. only one contact per relay is needed as shown to the right in Fig. 3.28(i). This can be checked from the truth table:

A	B	Circuit	
0	0	0	
0	1	1	→ switched through by AND gate 2
1	0	1	→ switched through by AND gate 3
1	1	1	→ switched through by AND gate 1

Fig.3.28(i)

which is identical with that for the simple A+B gate of Fig. 3.27(iii). Thus no AND gates are necessary.

There is no point in our learning the three theorems used in this example, they have been stated and proved only so that we may gain confidence in the relevance of the technique.

3.4.4 Electronic gates

We have yet only discussed switches in terms of metal contacts whether mechanically or electromechanically operated. These spring contacts have two states only, a short-circuit or a disconnexion, theoretically 0 Ω and ∞ Ω. That these values are not absolutely achieved is of no consequence in most circuits. Semiconductor switches however are entwined with their bias supplies and switching conditions arise from changes in voltage (at say, the collector or emitter) relative to the common rail from practically zero (i.e. the same potential as that of the common rail) to a few volts relative to this, corresponding to the logic conditions of 0 and 1. Such a change in potential when the semiconductor switch operates is applied to the next stage in the chain. If 0 to 1 goes positive, it is called *positive logic* and vice versa. Each gate operates according to its type, AND, OR, NOT; these are the basic ones, NAND and NOR generally being derived from them. The elements of typical gate circuits follow, they are shown employing diodes and standard (pnp or npn) transistors only,

many specialized semiconductor components also exist from which gate circuits are designed. As already shown, operating times for semiconductor gates are usually less than a micro-second, extremely small times compared with those for relays.

The appropriate graphic symbols for the gates themselves are contained in Appendix 2.

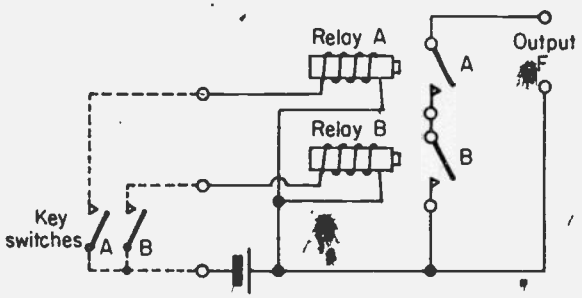
3.4.4.1 The AND gate

To appreciate a little more the practical side of gates, let us imagine that we have to design a circuit which ensures that a distant siren can only be sounded when two keyholders are present together, the keys being different – quite a practical situation. We ascribe to the two keyholders the letters A and B and it is clear that somewhere in the circuit an AND gate will be needed so that the siren is not sounded by A alone, nor by B alone, but when both insert their keys together.

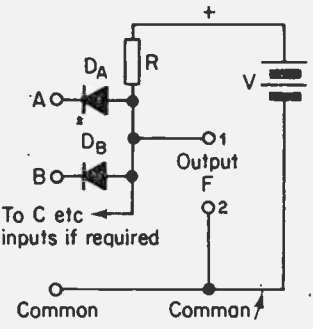
An electromechanical AND gate is shown in Fig. 3.29(i). Each key switches the battery via the line to the distant appropriate relay coil to operate it and the output circuit is arranged to switch a battery to an electromagnet coil (not shown) which operates the siren. Evidently for this to happen both relays A and B must be operated or in logic terms F = 1 when A and B are 1, i.e. F = A.B (F is the letter used to represent the output conditions).

A full electronic gate circuit for the same purpose is shown in Fig. 3.29(ii), "full" because the circuit includes the battery (or power supply) and common rail, these are not normally added to this type of diagram.

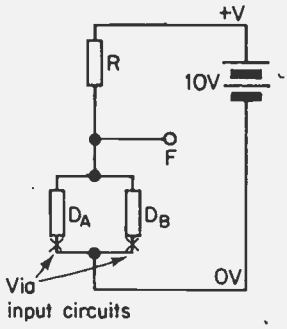
The diode circuit contains one diode per input, two only are needed in this example. V maintains a positive potential on diodes D_A and D_B. If no voltage relative to the common rail is applied to either A or B input terminals, then a current flows from the negative pole of the battery, through the input circuits (a modified form of those in (i)), diodes and resistor R to the positive pole, both diodes being forward-biassed.



(i) Electromechanical circuit.



(ii) Diode logic AND circuit.



(iii) Simplified equivalent circuit of (ii)

Fig. 3.29 The AND gate.

The total resistance in the circuit is mainly R as shown in the simplified equivalent circuit in (iii) where the resistances of D_A and D_B are low. Output terminal 1 is therefore held at almost the potential of the common rail, i.e. 0 V.

With a battery voltage of, say, 10, forward diode resistances of 4 Ω, R = 200 Ω and assuming that the resistances of the input circuits are low, the voltage drop across resistor R is

$$\frac{V.R}{R + R_D}$$

where R_D is the resistance of the two diodes in parallel (2 Ω)

$$= \frac{-10 \times 200}{202} \text{ V}$$

i.e. approximately 9.9 V, thus leaving point F at +0.1 V relative to the common rail which we class as effectively at 0 V. Thus A = 0, B = 0 gives F = 0.

Now a positive input at the A terminal, sufficient to reverse the bias, causes D_A to swing to high resistance. However D_B maintains a low resistance path in parallel so the voltage at F changes only to $+10 - (10 \times 200)/204 = +0.2$ V approx., a value still very near that of the common rail. Hence A = 1, B = 0, F = 0. Clearly for an input to B but not to A similar conditions apply and F remains at 0.

The last condition arises when the A and B inputs are both at voltages sufficiently positive that D_A and D_B are reverse-biassed and therefore of high resistance, say, 200 kΩ (the figures chosen are purely for convenience) then

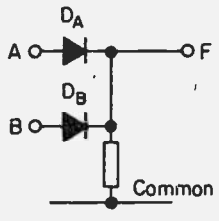
$$\text{Potential at point F} = +10 - \frac{10 \times 200}{100 \times 10^3} = +9.98 \text{ V}$$

i.e. it has risen from 0 V to very nearly +10 V. This is the condition which can now operate further switches to do the job required and A = 1, B = 1 gives F = 1. The truth table for an AND gate (Fig. 3.27(ii)) has been realized.

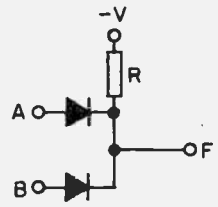
The gate is not limited to two inputs, it can have any number. For those inputs, F does not become 1 until all inputs are 1 together, hence the description sometimes used, *coincidence gate.*

3.4.4.2 The OR gate

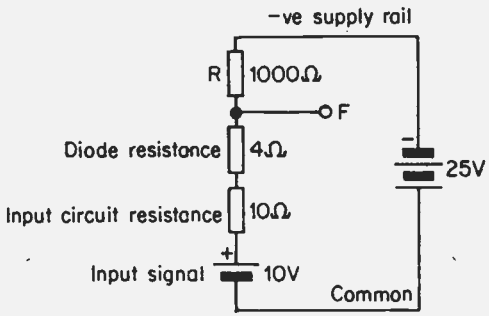
The OR gate can be quite uncomplicated, again using small diodes as switches. The diode-logic (DL) circuit in Fig. 3.30(i) shows the simplest form, just two diodes D_A and D_B feeding through to the output F (note that the drawings do not now contain those parts of the circuit which may be taken for granted). The diodes are connected for positive logic which means that a positive input of sufficient level, say 5–10 V, forward biasses the appropriate diode and brings F to this voltage level. The remaining diode becomes reverse-biassed and hence is ineffective, thus when A and



(i) Diode logic OR gate.



(ii) Gate with voltage supply.



(iii) Calculation of potential at F.

Fig. 3.30 The OR gate.

B = 0, F = 0 and when A or B or both are 1, F = 1, the truth table for the OR gate.

A faster operating circuit which minimizes delays introduced by circuit capacitances through its higher operating voltage, introduces a potential to the above circuit as shown in (ii). The voltage is appreciably greater than the logical 1 level. If the inputs are all at 0 V, D_A and D_B are forward biassed, therefore of low resistance and point F is connected to the input, so it is at 0 V. We could also look at this in the way shown for the AND gate, that is the relatively heavy current through R produces a voltage drop across it almost equal to V, again showing that F is almost at 0 V. If any input now rises to logical 1 (say +10 V), approximately that potential appears at F via the almost through path of its diode and this reverse-biasses the other or all other diodes.

The fact that the potential at F is not exactly that applied to an input terminal may bear a little closer examination, again using convenient values for ease of arithmetic. Suppose the resistance of the input circuit is 10 Ω, a logical 1 = +10 V and R = 1000 Ω. The equivalent circuit is then as in Fig. 3.30(iii). There are effectively two voltages in series-aiding, the 10 V input and the 25 V supply, together producing 35 V. Now, assuming that a forward-biassed diode has a resistance of, say, 4 Ω, Ohm's Law shows the circuit current to be

$$\frac{35}{1000 + 10 + 4} = 0.0345 \text{ A}$$

The voltage drop across R is therefore 34.5 V, positive at the point F relative to the negative supply rail or +34.5 − 25 = +9.5 V relative to the common, not quite the 10 V applied to the input terminal but very nearly so.

The appropriate circuit symbols are contained in Appendix 2.

3.4.4.3 The NOT gate

Unlike the gates which look at two or more inputs and give an output according to certain rules, this one merely inverts (changes over) the input condition. A transistor is suitable because it is an *inverting amplifier* when connected in common-emitter. Fig. 3.31 shows an npn transistor so connected. With 0 V input to the base, only a very small collector current flows, the voltage drop across R being correspondingly small, hence F is approximately at the supply line potential +V. Conversely for a 1 input to the base (say, +10 V) the collector current rises to saturation giving such a voltage drop across R that F is virtually at 0 V:

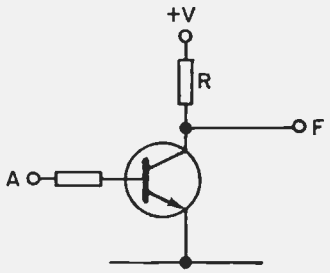


Fig. 3.31 npn transistor providing a NOT function.

Naturally there are many other semiconductor circuits capable of providing the above functions, those shown are by way of example only. Following an AND by a NOT circuit gives the NAND function, and similarly following an OR by a NOT provides a NOR. A NAND gate as used in an integrated circuit is considered in Section 4.3.2.1.

3.4.5 Computers

Intelligence is what sets people apart from animals and although the brain has many limitations, it is the capability of original thought which allows us to reign supreme. Now we have the computer to overcome some of the limitations

which include fatigue, slowness, unreliable memory and inability to handle large numbers, to name but a few. As yet the computer has ability but no wisdom of its own.

3.4.5.1 Computer organization

From the above it is not unreasonable to expect the computer to be organized in similar fashion to the brain. Looking at the latter first, only from the point of view of what it does, for we have yet to discover how it does it, somewhere inside there is a store of information or data, the *memory*. This is shown in Fig. 3.32(i). Feeding into the brain are several information-gathering devices, the eyes, ears and other sensory mechanisms. We remember some of the things we learn about each day but certainly not all and especially not the trivia for the memory has not the capacity for everything. The brain can therefore be imagined as having a built-in programme of instructions for dealing with the incoming information, a programme which differs from person to person.

Now information received can be (i) ignored and not even stored temporarily or (ii) stored in the memory or (iii) held and acted upon. We may perhaps see this sorting process as being carried out by a *control unit* presiding over all and backed up by its *stored programme* of instructions with two-way access to the memory or data store. Thus, information requiring no action other than storing in the memory needs only the functions shown in Fig. 3.32(i) which also caters for continuous "updating". Abundant and ever expanding as our memories seem to be, it is clear that information is removed regularly, often causing embarrassment to the "forgetful" owner. Some information is relegated to lower-level stores from which it may be recalled with varying degrees of success, the computer does not have this problem.

When action is required on information received, we can imagine the brain control unit, backed up both by its stored programme and the memory, signalling what is to be done to an *output unit* which controls arms, legs, voice etc. There may also be a need for calculations to be carried out and this is the function of a processor unit which must have its own

set of instructions (e.g. how to add, subtract, etc.) and also access to the memory, at least to recall multiplication tables. This is where the computer is really effective so we name this processor in computer terminology, the *arithmetic unit* as shown in Fig. 3.32(ii). The results are then fed to the output unit to distribute accordingly.

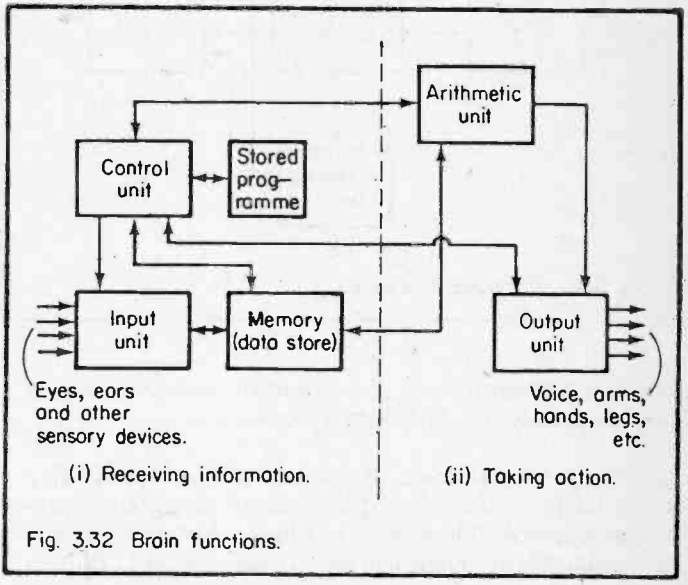


Fig. 3.32 Brain functions.

Slowly, through the ingenuity of designers the computer becomes more and more a copy of the human being for it can "see" light signals, "read" handwriting, "hear" audio signals and some even have arms (e.g. for handling radioactive materials) and legs (robots). Practically all computers "write" on screen or on paper.

This is not just a fanciful digression from the path of electronics for through this analysis of the functioning of the brain we shall remember more easily the organization of a computer, how it can be divided conveniently into units and the function of each. This is tidied up in Fig. 3.33; the

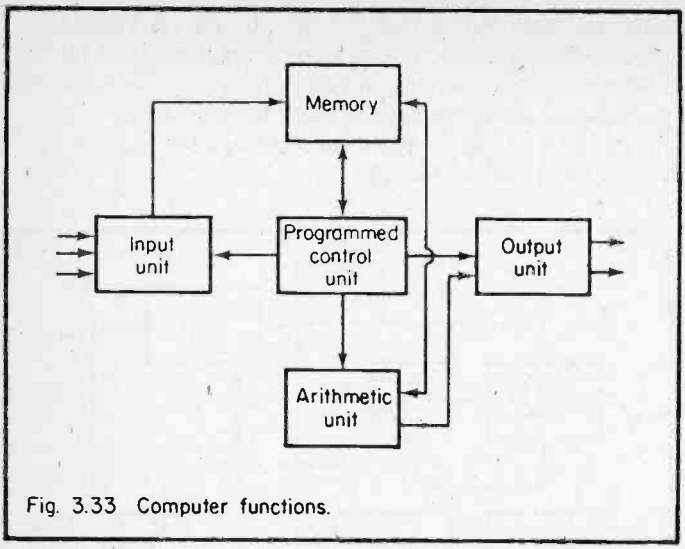


Fig. 3.33 Computer functions.

programmed control unit and arithmetic unit are generally combined under the title *Central Processor*.

We cannot hope to cover all these units even superficially, this is the province of the expert book on computers, nevertheless it is profitable to see how what we have learned about semiconductor switching and logic circuits comes to fruition in the arithmetic unit. Some ideas on memories also follow.

3.4.5.2 Computer arithmetic

There is no point in studying electronic gates if we do not appreciate their function in the complex networks of computers but we must be aware that design of computer circuits is highly specialized, needing much experience. Nevertheless, to build a little more on what we have done with binary arithmetic and gates so far it is worthwhile looking at the design of at least one type of circuit in the *arithmetic unit* of a computer (or calculator) so that at least we can feel that understanding of the basic elements of the computer is not unattainable.

From Appendix 3 the simple rules for addition of the binary digits in a column are developed as $0 + 0 = 0, 0 + 1$ or $1 + 0 = 1$, $1 + 1 = 0$ with a carry of 1 into the next higher column. The example given in the appendix is repeated here so that we may check as we go:

2^5	2^4	2^3	2^2	2^1	2^0	
1	0	0	1	0	1	(A digits)
	1	0	1	0	0	(B digits)
1	1	1	0	0	1	(S digits)

It is not only a question of adding 0's and 1's and sometimes carrying 1 into the next higher column, but considering any column on its own, a further question arises as to whether a 1 is to be absorbed from the next lower column. The total considerations in adding up any column are therefore the sum of the digits (S), the carry-in from the next lower column (C_i) and the carry-out to the next higher column (C_o). Thus in column 2^2 of the example, $S = 0, C_i = 0, C_o = 1$ and for column 2^3, $S = 1, C_i = 1, C_o = 0$. There is of course no C_i for column 2^0.

We can therefore set up a table showing the range of conditions which are possible for addition of any column (Table 3.1).

Now a 3-input AND gate gives an output of 1 when all its inputs together are 1 so by using AND gates to pinpoint those conditions where the sum of the digits is 1 ($S = 1$, conditions 2, 3, 5 and 8) and by changing any 0 in these to 1 by using a NOT or inversion gate, the AND gate operating conditions are determined as in the Table in Col. 6. Taking the first gate in this column as an example, $S = 1$ when $\overline{A}.B.\overline{C_i} = 1$ means that the AND gate gives an output of 1 when $\overline{A} = 1$ (i.e. $A = 0$) and $B = 1$ and $\overline{C_i} = 1$ (i.e. $C_i = 0$ – no digit is carried-in).

TABLE 3.1: POSSIBLE CONDITIONS IN BINARY ADDITION

	A (1)	B (2)	C_o (3)	C_i (4)	S (5)	S becomes 1 when (6)	C_o becomes 1 when (7)
(1)	0	0	0	0	0		
(2)	0	1	0	0	1	$\bar{A}.B.\bar{C}_i = 1$	
(3)	1	0	0	0	1	$A.\bar{B}.\bar{C}_i = 1$	
(4)	1	1	1	0	0		$A.B.\bar{C}_i = 1$
(5)	0	0	0	1	1	$\bar{A}.\bar{B}.C_i = 1$	
(6)	0	1	1	1	0		$\bar{A}.B.C_i = 1$
(7)	1	0	1	1	0		$A.\bar{B}.C_i = 1$
(8)	1	1	1	1	1	$A.B.C_i = 1$	$A.B.C_i = 1$

Of the four separate sets of conditions which cause S to be 1 and again the four which give $C_o = 1$, the output of the system for S will be 1 when any of the four equals 1, indicating that an OR gate is required, and similarly for C_o. Thus the Boolean expressions can be set up as:

$$S = \overline{A}.B.\overline{C_i} + A.\overline{B}.\overline{C_i} + \overline{A}.\overline{B}.C_i + A.B.C_i$$

(remembering that + in this algebra means OR), that is, there will be a 1 on the output S terminal of the system when any of the four AND gates has a 1 output.

Similarly:

$$C_o = A.B.\overline{C_i} + \overline{A}.B.C_i + A.\overline{B}.C_i + A.B.C_i$$

to determine whether a 1 will appear at the C_o output.

The following is a simple example, chosen to demonstrate the points made, the binary addition of 11 and 7.

2^4	2^3	2^2	2^1	2^0	
	1	0	1	1	(A)
		1	1	1	(B)
1	0	0	1	0	(S)

$C_o = 1$ $\quad$ $C_i = 1$

Consider the addition of the 2^2 column. There is a carry-in (C_i) of 1 from the 2^1 column and a carry-out to the 2^3 column (C_o) of 1. Thus A = 0, B = 1, C_i = 1, therefore S = 0 and C_o = 1 equivalent to line 6 in Table 3.

The Boolean expression above for C_o can be reduced by the apparently odd yet useful method of adding imaginary gates to the system, in this case $A.B.C_i + A.B.C_i$. This represents an OR gate comprising two similar AND gates which are also identical with one of the existing gates. If we draw them

out as switches as in Fig. 3.28 we will see that if the AND gate $A.B.C_i$ operates, the fact that there are two more similar gates in parallel has no effect. All these switch through together but the output of the OR system is still 1.

Thus

$$C_o = A.B.\overline{C}_i + \overline{A}.B.C_i + A.\overline{B}.C_i + A.B.C_i + A.B.C_i + A.B.C_i$$

and using theorem (i) of Section 3.4.3.2

$$C_o = B.C_i(A+\overline{A}) + A.C_i(B+\overline{B}) + A.B(C_i+\overline{C}_i)$$

and also using theorem (ii) of Section 3.4.3.2

$$C_o = A.B + A.C_i + B.C_i ,$$

a much reduced system.

So, what does all this look like in practice? Fig. 3.34 shows a schematic diagram of the gates required to satisfy the equations for S and C_o, the latter feeding out to the next column as C_i. A and B are the digits within the column being added with C_i injected from the next lower column. In the example $A = 0, B = 1, C_i = 1$, the appropriate potentials appear at the input terminals. The 0 on the A input passes through the inverter gate to appear on the output as $\overline{A}$. The $\overline{A}.B.C_i$ combination finds no AND gate which will operate to it hence there is no output from AND gates 1 to 4 into OR gate 1 and S remains at 0. The $B.C_i$ combination operates AND gate 7 and subsequently OR gate 2 and C_o changes to 1. Thus the binary adder has added 0 and 1 with the inclusion of a carry from the next lower column to give a sum of 0 with a carry of 1 to the next higher column.

Simplification of the Boolean algebra has enabled less complicated AND gates to be designed for C_o. From Col. 7 of Table 3.1 it is seen that only two inputs of the three avail-

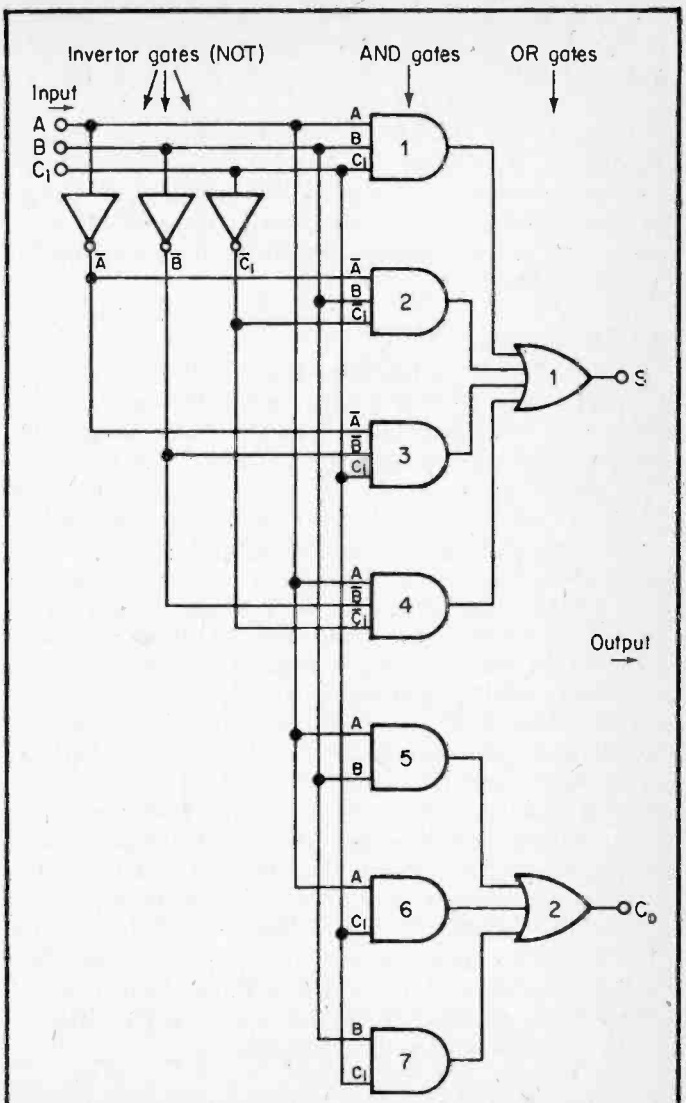


Fig. 3.34 Binary adder.

able are necessary, for example, in line 4, $C_o = 1$ when $A.B.\overline{C}_i = 1$, the $\overline{C}_i$ input not being used, in fact none of the inverted inputs is.

Considering that the circuit of Fig. 3.34 is needed for each column, we begin to see that a large number of diodes and transistors is required for a computer, hence the advantage of the integrated circuit which contains a multiplicity of semiconductor components, all correctly interconnected in a very tiny space.

3.4.5.3 Memories

Since in the binary system two states only exist, off or on, 0 or 1, etc., any device which can be switched electronically into one of two different conditions which can subsequently be recognized, can be used as a memory. The relay immediately appears as a likely candidate but not unexpectedly it fails badly on time, cost and bulk. However, we have already seen that the bistable multivibrator has two stable states and which one it is can be ascertained from the output terminal (Fig. 3.26(i)), hence replacing a relay by this device considerably reduces such limitations. We can get some idea of how such a memory would function by considering the problem of storing electronically a decimal number, say 12. This is binary form[A3] is 1100 showing that a 4-bit store is needed since 1100 consists of 4 binary digits (bits). Fig. 3.35 shows the basic arrangement. A negative *reset* pulse applied to the 4 input (1) terminals brings all output terminals to low voltage (just above 0 V). To set the memory, negative pulses are applied to those input (2) terminals where a 1 is to be stored, in this case in the 2^3 and 2^2 multivibrators. The output terminals of both of these then switch to high (almost to the positive supply rail potential, Fig. 3.26). The 4-bit memory now stores 1100 until cleared by a further reset pulse. The number held in the store can be read as often as is required from the 4 output terminals.

As mentioned in the previous section, the integrated circuit has so much to offer that it is taking over most complex computer circuitry and memories are no exception. The

multivibrator has been used as an example because we have already studied its operation, however special methods produce memory arrays capable of storing several thousands of bits in a package less than about 4 cm x 1.5 cm and research continually improves on this.

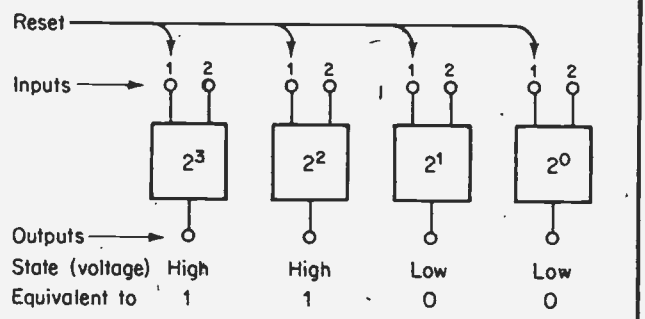


Fig. 3.35 4-bit memory using multivibrators.

Two other, perhaps obvious, types of memory are worthy of mention:

(i) magnetic tape: 1's and 0's are recorded along a flexible plastic tape coated with magnetic material by passing the tape over a *write* head which induces strong or zero magnetic spots. These can be read when the tape is passed over a *read* or *playback* head which develops a voltage when a strong magnetic spot passes. The principles are the same as those applying to home magnetic (cassette) recorders;

(ii) magnetic disc: as above but storage is on both faces of a rotating disc coated with magnetic material. Scanning of the disc by the writing and reading heads is similar to the tracking of the pick-up arm on a record-player.

4. MICROMINIATURE TECHNOLOGY

A short chapter but nevertheless an important one because it is through the enormous technological changes of micro-miniaturization that our whole way of life is being transformed.

As quantity production became more desirable the open style of connecting components together by (usually) insulated wires gave way to the technique of *printed wiring*. Here the process commences with an insulated board coated thinly with copper on one side, the copper being etched away where necessary so that what remains forms the circuit wiring. Similar boards can be produced in quantity and machine drilled so that component leads can be inserted through the holes and soldered to the copper. Although still using discrete (separate) components with many obvious manufacturing advantages, changing a component began to require special de-soldering techniques and the idea of throwing away a complete circuit instead of changing one or two components was born.

4.1 FILM TECHNIQUES

Then followed *thin* and *thick film* techniques with the term *microelectronics*. Both techniques use an insulating base, in the case of the thin film, insulating or conducting films are evaporated onto the base to produce wiring, resistors or "sandwiches" for capacitors. The thick film technique uses instead a metallized ink pattern formed onto the substrate by a silk-screen (as in printing) process. This film is some 0.025 mm thick, resistance material is applied where necessary and capacitors formed by deposition of one electrode, then the dielectric and finally the second electrode. A firing process fixes the film and deposited components and results in a highly stable completely wired circuit. Active components and those too large for application by these techniques are then wired in.

For the incredibly small device however, we next consider the *integrated circuit.*

4.2 INTEGRATED CIRCUIT TECHNIQUES

The terms "planar" and "diffusion" are explained in Section 1.6.1 and because the integrated circuit (IC) is formed entirely on a semiconducting wafer or *chip* using these processes, it remains (except perhaps at the surface) solidly uniform and hence is termed *monolithic* (Greek – as a stone). To get a moderately realistic idea of the actual size, Fig. 4.1 has been printed approximately full size and shows how a single slice of silicon some 75–80 mm in diameter is

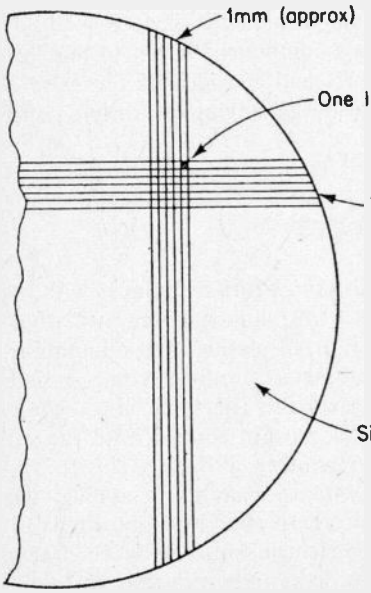


Fig. 4.1 Division of silicon slice into IC wafers (approx full size).

processed by being *diced* into as many as 4000 chips, each of the order of 1 mm square. On every one of these wafers will have been produced the particular integrated circuit, itself containing perhaps many thousand circuit elements depending on the technique employed. Study of the minute dot showing typically the size of an IC and realizing that wires are connected to it reminds us that the manufacturing process must involve scrupulous cleanliness, precision equipment for slicing (diamond saws used may have a cutting edge thickness of only 0.04 mm) and wire-bonding. Special techniques are also used for human operators to see what is going on. Clearly not all IC's pass the final electrical tests and some are therefore rejected. A complete IC package may contain several interconnected chips.

Integrated circuits are fabricated from both bipolar and MOS techniques. There are many variations in the processes used and these change with time as some find more favour than others, thus we confine ourselves to uncomplicated explanations to see how a seemingly impossible technique is carried out, with Fig. 4.1 always reminding us of the size of things. The depth of most layers is no more than a few microns (10^{-3} mm).

4.2.1 Transistors and diodes

If several elements are to be formed on a single substrate the problems of insulation and isolation must first be solved. Section 1.6.1 shows that insulation of layers is not difficult in view of the high resistivity and ease of production of silicon dioxide but in that section the requirement of isolation of elements from each other does not arise because each wafer contains one element (transistor or diode) only. To prevent the several elements on a single chip from being in contact with one another via the substrate, one commonly used method, known as *diode isolation* relies on the fact that a reverse-biassed diode has high resistance. This is evident in Fig. 4.2 which shows a diode in series with a transistor as a simple example of how elements are diffused in and interconnected.

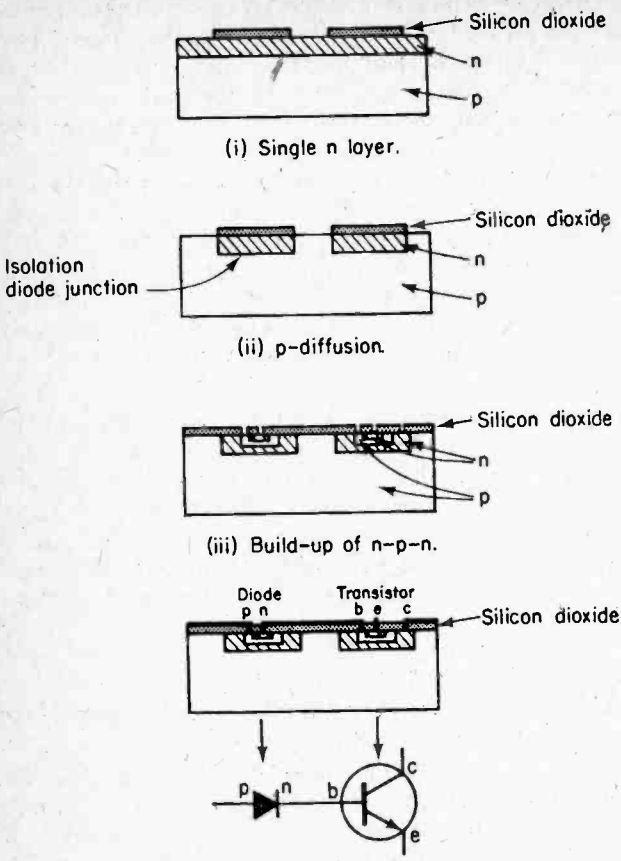


Fig. 4.2 Typical IC construction.

(i) at an early stage a p-type substrate has an n-type layer diffused on it. A layer of silicon dioxide has windows made in it by the photoresist process.

(ii) p-type diffusion through the windows sufficient to

reach the p-type substrate leaves the n-regions surrounded by p-type material, hence pn junctions are formed. If when in operation a small reverse bias is applied, the junctions become of high resistance thus effectively isolating the n-regions from the substrate.

(iii) by continuation of processing within the n-regions, further p diffusions and then n diffusions are applied. A silicon dioxide layer is added with holes etched for connexions to the elements.

(iv) shows the connexions added with the circuit diagram underneath.

To these interconnexions fine wires are bonded where external connexions are required and connected to tags solidly formed into the outer case or to an adjacent chip as appropriate.

4.2.2 Passive components

Mention has already been made of the formation of capacitors and resistors within monolithic IC's, nothing has been said about inductors or transformers. However, it is obvious from the nature of these components that IC technology cannot embrace them, hence circuit design aims at eliminating inductance if possible (e.g. using an RC instead of a resonant-circuit for an oscillator – Section 3.3), if not, then the item must be added externally. The same considerations apply to the higher values of capacitance and resistance, for example, none of the capacitors of Fig. 3.10 could be embodied in an IC. However the designer has recourse to direct-coupling and frequently is able to employ active elements instead with the result that the circuit diagram appears unconventional and difficult to understand without circuit notes.

Resistors may be formed from silicon or an alloy deposited in a narrow strip to a depth according to the value required. Equally for the p-type substrate of Fig. 4.2 a p-type strip may be diffused onto the first n-layer and connexions made to its two ends.

Capacitors up to some 100 pF may be formed by a metal film process using silicon dioxide as the dielectric, the area of plates

being adjusted to obtain the value of capacitance required. Frequently a more attractive method of reverse-biassing a diode is used (Section 3.4.1.2) and some control over the capacitance value is given by the fact that the capacitance falls as reverse-bias is increased.

4.3 INTEGRATED CIRCUITS

The term "circuit" being often used in a rather loose sense may be a little misleading because, frequently the integrated circuit is not complete, all possible is put into the package but a few components may have to be connected outside. What must be remembered is that the IC is indivisible, we cannot get inside to change or repair it. This would seem to indicate a multitude of different circuits to suit all purposes but this militates against the fact that it is only by the ultimate need of large quantities that the design and "tooling-up" become economic. Thus many IC's are available as "general-purpose" devices which can be adjusted to suit different requirements by the addition of other external components.

Very broadly, IC's may be divided into two types, *analogue*, where the signals within the circuit have similarity with the input signal (linear) and *digital*, where practically all work is on the 0 and 1 principle. Each can be subdivided as shown below.

4.3.1 Analogue circuits

There are two main classes, the first being a general one of amplifiers adjustable for many purposes, the second a range of slightly more specialized circuits, although there is no precise dividing line between the two classes.

4.3.1.1 Operational amplifiers

The *operational amplifier* figures prominently in the first class. Originally such amplifiers were used for mathematical "operations", but the name has now come into general use for amplifiers which can be adjusted by external circuitry to do a

whole range of jobs. The IC may contain some 10–30 transistors with a gain up to and even exceeding 200,000. The input impedance is high, typically 2 MΩ or more, it causes almost no *damping* therefore on an external input circuit.

Thus we choose a *buffer amplifier*, the main feature of which is to present a high impedance to an input circuit, as an example of the use of an operational amplifier. Any one of many other types of amplifier might equally be used as an example but this particular one retraces a little more our earlier thoughts on negative feedback.

We know from earlier studies that a voltmeter, for example, should not draw current from the circuit being measured otherwise it changes the circuit conditions and an inaccurate reading results. The circuit which finally moves the needle across the scale may have sufficient power output yet not have the required high input impedance, hence the use of a buffer amplifier which can provide this facility, its gain being unimportant.

A typical operational amplifier might have characteristics as follows:

Gain without feedback	200,000	all contained within a can some 10 mm diameter x 5 mm deep
Frequency range (at unity gain)	0–10MHz	
Input impedance	20 MΩ	
Supply voltage range	6–12 V	

The figure for the gain looks frightening when we may only need a gain of 1 or slightly more. Actually such gains as 200,000 are unusable for there is always the danger of positive feedback causing instability (Section 3.3). Two adjacent wires or one running parallel to the case or 0 V rail for even a short distance can easily have a capacitance of a picofarad or so and such a capacitance somewhere between output and input (X_C of 1 pF at 10 MHz is about 16 kΩ) will at some frequency produce $\beta A = 1\angle 0°$

and hence instability. Thus the operational amplifier is deliberately designed to have very high gain which in use will be reduced to the value required by negative feedback, as a bonus giving the negative feedback benefits outlined in Section 3.2.7.

Operational amplifiers have two separate input connexions, marked + and –. These signs indicate the relativities of the input with the output, + indicating that the output is in phase with the input, – that the output is 180° out-of-phase. The signs must not be confused with d.c. supply polarities.

The essential external circuitry required is shown in Fig. 4.3. R1 and R2 provide d.c. bias to the Input + terminal as required by the manufacturer and C1 isolates the input from this bias, this capacitor has high insulation resistance (an electrolytic is not suitable) so that the series combination C1,R2 does not affect the high input impedance. C2 and C3 block d.c. in the output circuit of the IC from the external circuit.

Feedback is applied from output to input, it must be 180°

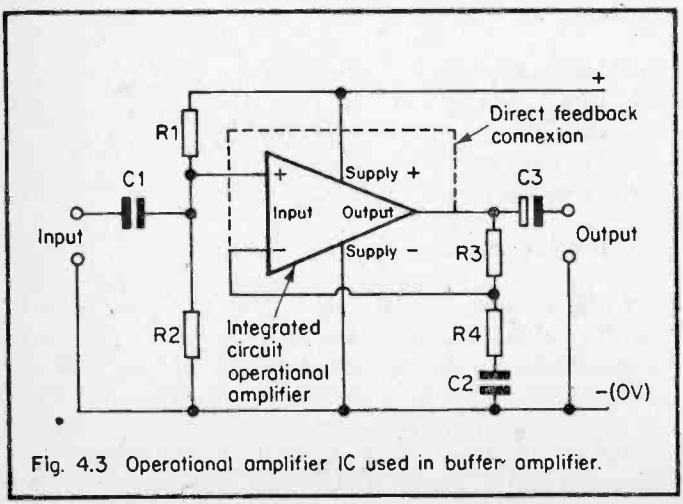


Fig. 4.3 Operational amplifier IC used in buffer amplifier.

out-of-phase, and this is provided for by the Input – terminal. The simple resistor chain R3,R4 decides the feedback fraction β as R4/(R3 + R4). Then the gain with feedback (Section 3.2.7.1) is equal to $A/(1 - \beta A)$ and since $\beta A >> 1$, this is equal approximately to $-1/\beta$, i.e. $-(R3 + R4)/R4$, so the overall gain can be set at any value we wish below, in this case, 200,000; the more feedback applied, the greater the feedback benefits. At the limit, where in the case of a buffer amplifier the interest lies in the high input impedance not the gain, R3,R4,C2 might be omitted altogether with a direct connexion from output to input (shown dotted). Then $\beta = 1$ and the overall gain is unity.

4.3.1.2 Special purpose circuits

Although standard operational amplifiers may be employed instead, special *microphone amplifiers* with fixed gain (or perhaps two settings) are available as IC's with gains around 200. *Pre-amplifiers*, that is, voltage amplifiers used for increasing the level of a tiny signal, for example, from certain types of microphone are also used. These raise microphone output levels from around 0.1 mV to a value suitable for driving a power amplifier and have gains of around 10,000. Power amplifiers themselves are also available although there are limitations on output power. Clearly large output powers, for example, 20 watts or more for a home audio system would involve excessive junction temperature rises.

For the manufacturers of television, radio and audio systems, many specialized IC's are produced in ever expanding variety and quantity.

4.3.2 Digital circuits

Very broadly these can be divided into three main classes, Logic, Memories and Microprocessors. We can appreciate the need of all these from our studies in Section 3.4 if we look upon the term *microprocessor* as being roughly allied to what goes on inside a computer. However memories and microprocessors are beyond our scope here, even the basic elements warrant separate volumes.

4.3.2.1 Logic gates

The various gates discussed in Section 3.4.4 and shown simplified in Figs. 3.29–3.31 with an example of their use in larger numbers in Fig. 3.34 indicate how appropriate these are to integrated circuit technology. The main circuit components used in a gate form the label which indicates the gate type, for example, RTL (resistor–transistor logic), DTL (diode–transistor logic) and TTL (transistor–transistor logic). These have evolved in the order given with the last tending to supersede its predecessors. The TTL system may use a *multiple-emitter transistor*, a technique combining several transistors into one which enables many more gates to be fabricated on one chip. As an example, Fig. 4.4 shows a simplified single TTL 3-input NAND gate.

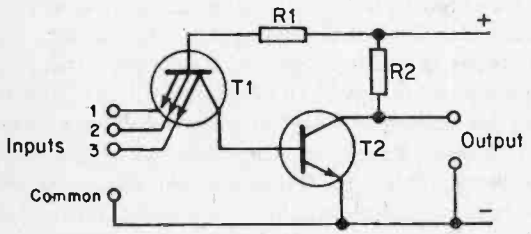


Fig. 4.4 Simplified TTL NAND gate.

In manufacture it is possible to diffuse several separate emitters into a single base region and in Fig. 4.4 T1 is shown having three. The positive bias applied to the base is such that if all inputs and therefore emitters are at "high" positive potential, the base-emitter junctions are not forward biassed, hence the collector is nearly at the supply voltage, so biassing T2 "on" and the voltage drop across R2 causes the output to be at "low". Connexion of any of the emitters to the 0 V rail forward-biasses the base-emitter junction, the collector voltage falls and T2 cuts off, that is, the output terminal switches to "high". It can be seen that this process agrees with the truth

table for the NAND function in Fig. 3.27 (input C not shown).

This is just a single example, all logic gates are obtainable in integrated circuit form using TTL.

5. A PAUSE FOR BREATH

So we reach a turning point in our studies of the *elements* or *first principles* of electronics. We cannot say we have *completed* our elementary studies for much has had to be omitted, three small books cannot possibly cover all the elements of such a vast subject. Nevertheless, having come so far, hopefully with success, the reader should have formed solid foundations on which to build and much of the frightening terminology yet to be met will be less so because in fact it is nearly all built on these same foundations.

The author sincerely hopes that readers have enjoyed the book or books and what is more, gained knowledge and confidence from them. The world so badly needs electronic engineers that if, through the medium of these books, no more than a handful of young readers make this their career, then the work will have been justified.

APPENDIX 1

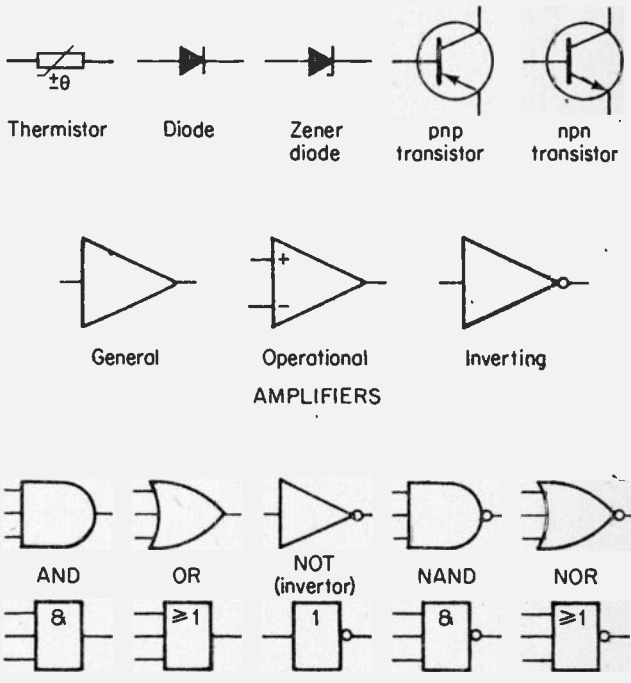
ABBREVIATIONS

Fig.	Figure
i.e.	Latin, id est – that is to say
e.g.	Latin, exempli gratia – for, by the way of, example
vice versa	Latin – by transposing the main items in the statement just made
a.c.	alternating current
approx.	approximately
cct.	circuit
d.c.	direct current
IC	integrated circuit
col.	column
eqn.	equation
e.m.f.	electromotive force
l.h.	left-hand
r.h.	right-hand
hi-fi	high fidelity
max.	maximum
min.	minimum

p.d.	potential difference
r.m.s.	root mean square
$\equiv$	is the same as
$\approx$	approximately equal to
$> \gg$	greater than, very much greater than
$< \ll$	less than, very much less than
+ve	positive
$\xrightarrow{+}$	positive-going
−ve	negative
$\xrightarrow{-}$	negative-going

APPENDIX 2

CIRCUIT SYMBOLS



FET and MOST –

some of the symbols used are shown in figs. 2.11 and 2.13.

Fig. A2.1 Circuit symbols.

APPENDIX 3

BINARY ARITHMETIC AND ITS APPLICATION TO COMPUTERS

This appendix gives only the briefest resumé of binary arithmetic with indications of how computer and calculator processors set about the problems. It is intended as an introduction to the subject and to the elementary circuits developed in the main text.

A3.1 NUMBER SYSTEMS

We are all conversant with the denary (of ten, decimal) system of numbers for it is the one we use every day. Many also recognize Roman numerals although not all understand how a number is derived, mainly because the symbols have not *positional* value as in the decimal system. In this by "positional" is meant that the place of any figure in the number indicates the power of 10 by which it is multiplied to bring out its full value, for example, the number 1,000 which can be expressed as 10^3 is derived by working to the left from the decimal point (in this case not shown because we know where it is), giving 0×10^0, 0×10^1, 0×10^2 and 1×10^3.

Again, 866,258.33 which has figures also to the right of the decimal point is derived from

$$\begin{array}{cccccccc} 8 & 6 & 6 & 2 & 5 & 8 & 3 & 3 \\ \times & \times & \times & \times & \times & \times & \times & \times \\ 10^5 & 10^4 & 10^3 & 10^2 & 10^1 & 10^0 & 10^{-1} & 10^{-2}, \end{array}$$

all added together. 10 is called the *radix* (root or base) of the decimal system and there must in this case be 10 different symbols, i.e. 1, 2, 3, . . . , 8, 9, 0.

Digital transmission systems and computers almost invariably

use a *binary* system with a radix of 2, thus requiring only two different symbols, for convenience we use a pair from our everyday system, 0 and 1. Like human beings, electronic circuits can more easily recognize on or off than any state between the two. As an example, consider the petrol tank of a motor car. The motorist knows only too well when the tank is empty for the engine will not run and again when the tank is full for at the petrol or gasolene pump, either an automatic device cuts the flow or the precious liquid overflows.

The two states empty and full (0 and 1) cannot be misinterpreted. But suppose the fuel gauge shows half-full, is it really so? The car might be on a slope or the gauge inaccurate, there is an element of doubt. Similarly with electronic circuits, there are many distortions a signal may suffer when transmitted over a line or through other circuits, so there may be doubt as to its exact intended value. Therefore instead of trying to work with 10 different states or levels, there is much less likelihood of error if only the two which are recognized with the greatest certainty are used. The additional cost in complexity of manipulation (binary numbers nearly always have more digits than denary) is outweighed by the greater reliability.

We must be able to put decimal numbers into a computer and also have decimal numbers displayed or printed at the output so conversion from decimal to binary and vice versa is necessary. Just as we analysed the decimal number above, so the same can be done with a binary number. As an example, the number 6258 in binary form is 1100001110010, a number filling us with misgivings at its sheer size, using 13 digits to describe only 4 in the decimal system. But it must be appreciated that electronic circuits can handle these in nanoseconds, which is certainly more than we ourselves can do with decimals! To show the similarity between the denary and binary systems:

$$6258 \text{ in radix } 10 = (6 \times 10^3) + (2 \times 10^2) + (5 \times 10^1) + (8 \times 10^0)$$

$$
\begin{aligned}
6258 \text{ in radix } 2 = {} & (1 \times 2^{12}) + (1 \times 2^{11}) + (0 \times 2^{10}) \\
& + (0 \times 2^{9}) + (0 \times 2^{8}) + (0 \times 2^{7}) \\
& + (1 \times 2^{6}) + (1 \times 2^{5}) + (1 \times 2^{4}) \\
& + (0 \times 2^{3}) + (0 \times 2^{2}) + (1 \times 2^{1}) \\
& + (0 \times 2^{0})
\end{aligned}
$$

and obviously any other radix could be used, for example, there is an octal system with radix 8.

To convert decimal to binary it is useful to have a table showing powers of 2 as in Table A3.1, and 6258 might be converted by the following procedure:

(i) the largest power of 2 in 6258 is $2^{12} = 4096$
$\therefore 6258 = (1 \times 2^{12}) + (6258 - 4096) = (1 \times 2^{12}) + 2162$

(ii) the largest power of 2 in 2162 is $2^{11} = 2048$
$\therefore 6258 = (1 \times 2^{12}) + (1 \times 2^{11}) + (2162 - 2048) =$
$= (1 \times 2^{12}) + (1 \times 2^{11}) + 114$

(iii) the largest power of 2 in 114 is $2^{6} = 64$ (note that we have skipped 2^{10}, 2^{9}, 2^{8} and 2^{7})
$\therefore 6258 = (1 \times 2^{12}) + (1 \times 2^{11}) + (0 \times 2^{10}) + (0 \times 2^{9})$
$+ (0 \times 2^{8}) + (0 \times 2^{7}) + (1 \times 2^{6}) + 50$
and the 50 is broken down similarly.

TABLE A3.1 POWERS OF 2

$2^{0} = 1$	$2^{7} = 128$	$2^{14} = 16{,}384$
$2^{1} = 2$	$2^{8} = 256$	$2^{15} = 32{,}768$
$2^{2} = 4$	$2^{9} = 512$	$2^{16} = 65{,}536$
$2^{3} = 8$	$2^{10} = 1{,}024$	$2^{17} = 131{,}072$
$2^{4} = 16$	$2^{11} = 2{,}048$	$2^{18} = 262{,}144$
$2^{5} = 32$	$2^{12} = 4{,}096$	$2^{19} = 524{,}288$
$2^{6} = 64$	$2^{13} = 8{,}192$	$2^{20} = 1{,}048{,}576$

The above is set out in detail for explanation, the conversion might simply be carried out as follows:

		Binary number (read downwards)
6258		
4096	$= (1 \times 2^{12})$	1 (actually followed by 12 0's but we need not write these)
2162		
2048	$= (1 \times 2^{11})$	1
114		
	(0×2^{10})	0
	(0×2^{9})	0
	(0×2^{8})	0
	(0×2^{7})	0
64	(1×2^{6})	1
50		
32	(1×2^{5})	1
18		
16	(1×2^{4})	1
2		
	(0×2^{3})	0
	(0×2^{2})	0
2	(1×2^{1})	1
0	(0×2^{0})	0

It is useful to remember that adding 1 to the largest power index of 2 in the full number gives the number of digits in the binary number.

Conversion from binary to decimal is now perhaps obvious:

1	1	0	0	0	0	1	1	1	0	0	1	0
2^{12}	2^{11}	2^{10}	2^{9}	2^{8}	2^{7}	2^{6}	2^{5}	2^{4}	2^{3}	2^{2}	2^{1}	2^{0}

From Table A3.1:

4096+2048+ 64 + 32 + 16 + 2 = 6258

Each of the 13 digits in the binary number above is known as a *bit* (from *b*inary dig*it*).

There are other techniques of conversion which may be slightly quicker but the above method, by reminding us of the similarity with decimals in use of the basic positional system, is less likely to be forgotten.

A3.2 BINARY ARITHMETIC

The arithmetic of the binary system is much less difficult than might be expected. It is most easily explained if we keep in mind the decimal equivalents as we proceed for this gives continuous confidence that the technique is sound.

A3.2.1 Addition

Consider the addition of binary numbers 100101 (37) and 10100 (20). So that we keep clearly in our minds what each binary digit represents in decimal, set out as follows:

2^5	2^4	2^3	2^2	2^1	2^0	
1	0	0	1	0	1	= 37
	1	0	1	0	0	= 20
Addition						
1	1	1	0	0	1	

Starting with the 2^0 column, add 1×2^0 to 0×2^0 ($= 1 \times 2^0$) giving a digit of 1. In the 2^1 column, 0 is added to 0, giving a digit of 0. In the 2^2 column 1×2^2 is added to 1×2^2 giving 2×2^2, but we cannot have 2 as the addition because the system only allows 0 and 1. But 2×2^2 is equal to 1×2^3, hence we write 0 and carry 1 into the 2^3 column giving in this column $0 + 0 + 1$ ($\times 2^3$). The 2^4 and 2^5 columns are straightforward. Simple rules follow:

Adding 0 and 0 gives 0
0 and 1 gives 1
1 and 0 gives 1
1 and 1 gives 0 with a carry of 1 into the next higher column.

A computer contains *binary adding* circuits, one per digit, and it adds the separate columns of digits (the 2^0, 2^1, 2^2 digits etc.) with the carry function when required. The outputs of the adding circuits give the answer in binary form.

A3.2.2 Multiplication

The other main computer component for arithmetical operations is known as a *shift register*, why the term "shift" is used will be clearer when we have considered the computer method of multiplication. Consider for example 23 x 10 or in binary 10111 x 1010.

The *multiplier* is in fact $(1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (0 \times 2^0)$ so by multiplying the *multiplicand* (the number to be multiplied) by each of these bracketed terms and adding the results together, the answer is obtained. Each shift of the multiplicand one place to the left in the following table multiplies it by 2 (i.e. two shifts multiplies by 4 etc.), therefore the following rule applies. Shift the multiplicand one place to the left for each 0 in the multiplier and add nothing. Shift one place to the left for each 1 in the multiplier and add the multiplicand.

This may seem complicated but the following sum should help.

Multiply 23 x 10 in binary:

10111	(23 – multiplicand)
1010	(10 – multiplier)
00000	– first move: the 2^0 value in the mulitplier is 0, therefore add nothing;

10111	– second move: the 2^1 value in the multiplier is 1, therefore add multiplicand shifted one place to left (equivalent to multiplying by 2);
00000	– third move: the 2^2 value in the multiplier is 0, therefore shift to left and add nothing,
10111	– fourth move: the 2^3 value in the multiplier is 1, therefore shift to left and add multiplicand (total shifts = 3, i.e. equivalent to multiplying by 8).
11100110	– fifth move: add. Result = 230.

Looking back on what we have done, the first and third moves add 0 to the result but create the necessary shift (unnecessary on first move). The second and fourth moves are effective in the multiplication process and are equivalent to multiplying by 2 and 8 respectively, the results added together effectively multiply by 10.

We have used the term "shift" so many times that it can be appreciated why the term "shift register" (register = device in which entries are made) is used and why such equipment is necessary.

Addition and multiplication are the main processes carried out in the arithmetic unit of a computer or processor. Suitable basic circuits are developed in the main text. Although there is no need to go into detail we cannot stop here without realizing that subtraction and division can be accomplished with little additional equipment.

A3.2.3 Subtraction

Peculiarly enough subtraction can be changed mainly into a form of addition. The technique used is to first find the *complement* of the number being subtracted. The complement is obtained by changing 1's for 0's and 0's for 1's, adding 1 to the final result. It will then be found that when the number is added to its complement the result is 1 followed only by 0's. Then, if instead of subtracting the

number, its complement is added the correct result is given, for example:

Subtract 147 from 173 in binary:

We first find the complement of 147,

147 in binary is	10010011
Its complement is	01101100 + 1 = 1101101

Adding:

173	10101101
Complement 147	1101101
26	(1)00011010

The 1 preceding the 0's on the left is discarded, it arises through using a complement, giving the correct answer, 26, an ingenious system. Electronic circuits easily identify and change 1's for 0's and vice versa, after which the st[illegible]l binary adder takes over.

A3.2.4 Division

Division in binary follows the technique of multiplication but in reverse. The *divisor* (number by which another is divided) is repeatedly subtracted from the *dividend* (number which is divided by the divisor) with shifts to the right as appropriate. Subtraction as we have seen can be accomplished mainly by an addition process, hence the main components for division are yet again the binary adder and shift register.

APPENDIX 4

MATHEMATICS

We have at this point a fair amount of experience in the use of graphs as a pictorial representation of changing electrical conditions, frequently on a basis of time. Experienced though we may be with sine waves and their equations, the aim of this appendix is to take us a little further by looking at certain types of transistor and diode *smooth* curves, that is, those for which a mathematical equation can be derived to link the two variables so that given one the other can be calculated, in a way dispensing with the graph itself. But our intention is not to abandon the graph for it is always more informative than a string of symbols, but to find out just what other uses the equation may have. The first smooth curve we examine is, peculiarly enough, the straight line.

A4.1 THE STRAIGHT-LINE GRAPH

Consider a pair of axes 0X and 0Y as shown in Fig. A4.1 and

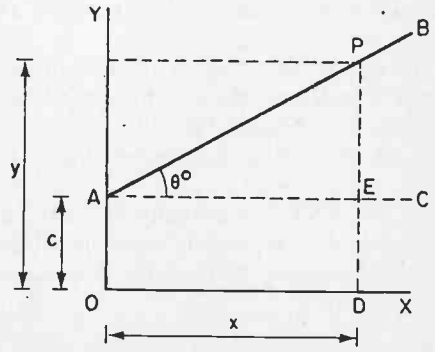


Fig. A4.1 The straight-line graph.

a point A on the Y-axis a distance c from O. Let a straight line AB be drawn at an angle θ° to the line AC parallel to OX.

Take any point P on the line AB and draw a line from P perpendicular to OX meeting OX at D, a distance x from O. Let PD meet AC at E and let PD = y. The *co-ordinates* of P are therefore x and y.

The equation to the line AB is the one which expresses the relation between the co-ordinates of any point on the line, that is, the relationship between x and y which defines the position of P irrespective of where it is along the line.

Then $$y = PD = PE + ED = PE + c$$

and since $$\tan\theta = \frac{PE}{AE}, \quad PE = AE \tan\theta = x \tan\theta$$

and if we call $\tan\theta$, m, then

$$y = mx + c$$

where m is the tangent of the angle the line makes with the X-axis and c is known as the *intercept* on the Y-axis.

Note that for any straight line on a graph, m is a fixed quantity, it determines the slope of the line while c shows its position relative to the X-axis. Both m and c may be positive or negative. When c = 0 the line passes through O.

The equation y = mx + c is naturally the least complicated of all curve equations, as soon as exponents, logarithms or trigonometrical functions affect x, the graph is no longer straight.

EXAMPLE:
Plot the graph of $y = x/\sqrt{3} + 4$ between x = 0 and x = 10.

This is immediately recognizable as the equation to a straight line where $m = 1/\sqrt{3}$ and $c = 4$, therefore only two points are needed:

at $x = 0$, $y = 4$

at $x = 10$, $y = \frac{10}{\sqrt{3}} + 4 = \frac{10}{1.732} + 4 = 9.77$

The graph is drawn in Fig. A4.2. It could be drawn between the two calculated points or at $\tan^{-1} 1/\sqrt{3} = 30°$ from the point $x = 0, y = 4$. Note that this angle only applies when the scales of x and y are the same.

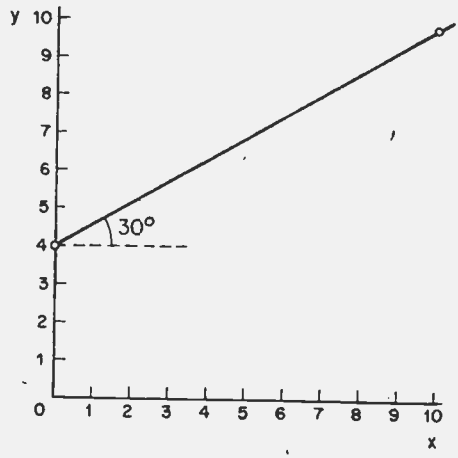


Fig. A4.2 Graph of $y = \frac{x}{\sqrt{3}} + 4$.

EXAMPLE:
Measurements made of the current in a system over a range of voltages are as follows:

V	0	1	2	3	volts
I	0.8	0.95	1.10	1.25	amps

Find the law connecting V and I.

If the graph is a straight line of the form $I = mV + c$ then:

$$1.25 = 3m + c$$
$$0.8 = 0m + c \quad \text{(subtract to eliminate c)}$$
$$0.45 = 3m$$

$$\therefore \quad m = 0.15$$

and $0.8 = 0m + c \quad \therefore c = 0.8$,

i.e. the first and last sets of figures suggest $I = 0.15\,V + 0.8$ amps.

Check at $V = 1$ volt, $I = 0.15 + 0.8 = 0.95$ A

Check at $V = 2$ volts, $I = (0.15 \times 2) + 0.8 = 1.1$ A

Both agree with the given figures, hence the equation satisfies all points.

A4.2 CALCULATION OF A.C. RESISTANCE

This section underlines the point made in Section 2.2.3 of the main text that the "a.c. resistance" of a device for which the current/voltage characteristic is available can be obtained graphically or by calculation. The mathematical considerations are not restricted to the type of example given but apply generally, stated briefly, they are that, given the equation to a curve, the gradient (amount of slope) at any point can be calculated. The gradient in this particular instance gives directly the "a.c. conductance", the reciprocal of which is the a.c. resistance. Gradient of a curve needs some comment first.

A4.2.1 Gradient of a curve

In Fig. A4.3 A and C are two points on the graph at $v_1 i_1$ and $v_2 i_2$ respectively. $(i_2 - i_1)$ is represented by BC and is the net

change in current for a change in voltage ($v_2 - v_1$), represented by AB. Hence the current change over this range of voltage might appear to be at a constant rate BC/AB. It is not constant however because the characteristic is curved, a constant rate would imply a straight line between A and C, such a straight line joining two points on an arc is known as a *chord*. BC/AB represents the gradient of the chord AC, that is, the amount of rise over a certain horiziontal distance.

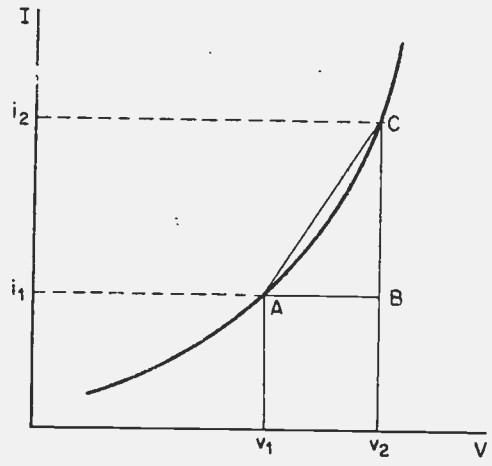


Fig. A4.3 Gradient of a curve.

If we now reduce the length of the chord AC by taking v_1 and v_2 closer together, the curve and the chord between these two points become less dissimilar until eventually when v_1 and v_2 coincide, the chord becomes a tangent which being at a point only, accurately represents the gradient of the curve at that point.

A4.2.2 Graphical method

Consider the transistor input characteristic in Fig. A4.4. It is

an exponential curve, the equation of which can be expressed by

$$I = I_s(e^{kV} - 1)$$

where I is the current at voltage V and I_s is the reverse saturation current.

Taking the point P as an example, the graphical method of measuring the gradient of the curve is to draw the tangent carefully with a ruler, complete a triangle such as ABC and calculate BC/AB. The triangle should be as large as possible for accuracy of measurement. Its size does not affect the

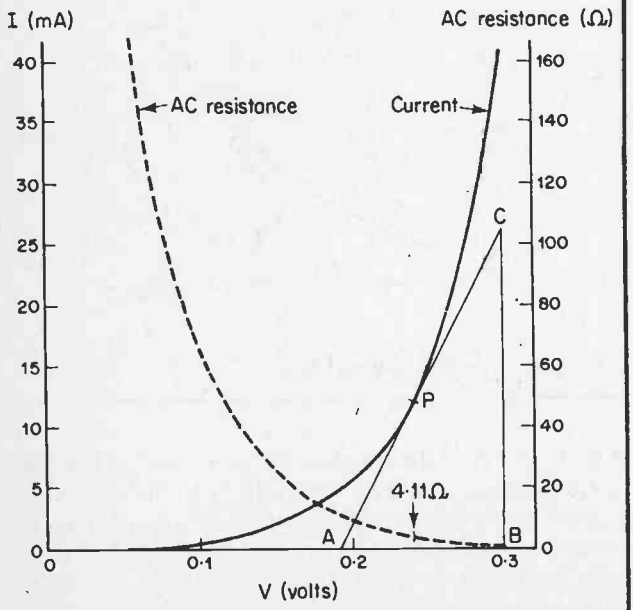


Fig. A4.4 Transistor input and calculated AC resistance characteristics.

result because we are only interested in the ratio of the two sides and the ratio does not change. In this case BC measures 26.5 mA, AB 0.108 V giving an a.c. conductance of

$$\frac{26.5 \times 10^{-3}}{0.108} = 0.245 \text{ Siemens,}$$

the reciprocal giving the a.c. resistance as

$$\frac{1}{0.245} = \underline{4.08\ \Omega}.$$

Note that the steeper the curve the lower the a.c. resistance, easily remembered because a small change in voltage is effective in creating a large change in current.

A4.2.3 By calculation

For the equation to the curve to be of any use we must first determine the values of k and I_s. This is done by taking any two points and solving the two equations produced:

At 0.2 V, $I_1 = 5.4$ mA $\quad \therefore 5.4 \times 10^{-3} = I_s(e^{0.2k} - 1)$ (1)

At 0.3 V, $I_2 = 40.2$ mA $\quad \therefore 40.2 \times 10^{-3} = I_s(e^{0.3k} - 1)$ (2)

These equations as they stand are not easy to solve, but it is fortunate that we can simplify them because over the part of the curve in which we are interested, e^{kV} is much greater than 1, hence we can neglect the -1 with very little loss of accuracy giving

$$\frac{5.4 \times 10^{-3}}{I_s} = e^{0.2k} \quad \ldots\ldots\ldots\ldots (1)$$

$$\frac{40.2 \times 10^{-3}}{I_s} = e^{0.3k} \quad \ldots\ldots\ldots\ldots (2)$$

Getting rid of the 1 was an artful move, another is to take natural logarithms of both sides of the two equations:

$$\log_e 5.4 \times 10^{-3} - \log_e I_s = 0.2\,k \quad \ldots\ldots\ldots\ldots (1)$$

$$\log_e 40.2 \times 10^{-3} - \log_e I_s = 0.3\,k \quad \ldots\ldots\ldots\ldots (2)$$

(we recall that $\log_e x/y = \log_e x - \log_e y$ and $\log_e e^x = x$).

Then

$$-3.214 - \log_e I_s = 0.3\,k \quad \ldots\ldots\ldots\ldots (2)$$
$$-5.221 - \log_e I_s = 0.2\,k \quad \ldots\ldots\ldots\ldots (1)$$

$$2.007 = 0.1\,k \qquad \therefore k = 20.07.$$

(subtract eqn. (1) from eqn. (2) to remove the terms containing I_s).

Next substitute for k in either eqn. (1) or (2) to obtain I_s:

$$-5.221 - \log_e I_s = 0.2\,k$$

$$\therefore \quad -5.221 - \log_e I_s = 4.014$$

$$\therefore \quad -\log_e I_s = 9.235$$

$$\therefore \quad \log_e I_s = -9.235$$

$$\therefore \quad I_s = \text{antilog}_e -9.235 = 0.0000976\ \text{A} = 0.098\ \text{mA}.$$

[We may need a reminder here on using Napierian logarithms since $\text{antilog}_e -9.235$ is not given directly by tables. We can work either in characteristic plus mantissa form or in full negative numbers, the latter may lessen the possibility of getting mixed up with a process which may not be all that easy to follow without practice:

Tables give antilog$_e$ $\overline{1}0.7897$ (= −9.2103) as 10^{-4}

and $\overline{1}2.4871$ (= −11.5129) as 10^{-5}

−9.235 lies between the two bracketed numbers, hence the answer lies between 10^{-4} and 10^{-5}.

Now −9.235 is positive to −11.5129 by 2.2779, the latter value is convertible directly from the tables, i.e. antilog$_e$ 2.2779 = 9.756 which when multiplied by 10^{-5} becomes 0.00009756, the answer required. Check: $\log_e 9.756 + \log_e 10^{-5} = 2.2779 + \overline{1}2.4871 = \overline{1}0.7650 = -9.235$.]

Hence the mathematical equation to the curve of Fig. A4.3 is

$$I = 0.098 \times 10^{-3} (e^{20.07V} - 1)$$

(the inaccuracy from neglecting the 1 from the original formula can be shown to be less than 2%).

One of the early surprises on first meeting the method of calculation known as *calculus* is that determination of the gradient of a curve at any given point is easy. We cannot develop calculus to this point in a short appendix so we accept that if the equation to the curve is

$$I = I_s(e^{kV} - 1)$$

then from calculus, the rate of change of I relative to V at any point (the gradient of the curve – expressed as dI/dV) is

$$kI_s e^{kV}$$

where V is the voltage at that point. The −1 disappears since it has no effect on the slope, it only shifts the curve.

In this case, gradient of curve is equal to

$$= 20.07 \times 0.098 \times 10^{-3} . e^{20.07\,V}$$

$= 0.001967\, e^{20.07 \times 0.24}$ (since for point P, V = 0.24 V)

$= 0.243$

$\therefore$ a.c. resistance $= 1/0.243 = \underline{4.11\ \Omega}$.

One might be forgiven for wondering whether the result is worth all the complicated mathematical juggling but we now have a single formula which can be used for any value of V (it is not even necessary to calculate I) and in fact a complete resistance/voltage (or current) curve can easily be calculated. With a scientific calculator the whole process takes no more than a few minutes. The curve is shown dotted on Fig. A4.4 with the a.c. resistance scale on the right.

Many curves of the transistor input and diode type conform reasonably well to the above equation or equally to another exponential type, $I = kV^n$. The latter is easier to handle and should be tried first to see if a good fit can be obtained. To do this calculate the constants k and n at the two ends of the curve and use the results to calculate a point in the middle to assess how well the equation fits. As an example, the equation to the transistor input characteristic of Fig. 2.8(i) in the main text is obtained as follows:

Read off points near the ends of the curve, e.g.

at $V_{BE} = 0.5$ V, $I_B = 0.1$ mA

at $V_{BE} = 1.2$ V, $I_B = 2.0$ mA

We need not bring I_B to amperes provided that it is not forgotten later on:

$\therefore \quad 0.1 = k \times 0.5^n$ ..... (1)

$2.0 = k \times 1.2^n$ ..... (2)

Taking common logarithms of both sides of each equation (no need for natural logarithms as e is not involved):

$$\log 0.1 = \log k + n \log 0.5 \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\log 2.0 = \log k + n \log 1.2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$\therefore \quad 0.3010 = \log k + 0.0792\, n \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$
$$-1.0 = \log k - 0.3010\, n \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$1.3010 = 0.3802\, n$$

[subtract eqn. (1) from eqn. (2)]

$$\therefore \quad n = \frac{1.3010}{0.3802} = 3.4219\ (\approx 3.42)$$

Also $$0.3010 = \log k + 0.0792 \times 3.4219$$

$$\therefore \quad \log k = 0.3010 - 0.2710 = 0.03$$

$$\therefore \quad k = \text{antilog } 0.03 = 1.07$$

A possible equation to the curve is therefore

$$I_B = 1.07\, V_{BE}^{3.42}\ \text{mA}\,.$$

Check at $V_{BE} = 1.0$ V:

$$I_B = 1.07 \times 1.0^{3.42} = 1.07\ \text{mA}$$

which agrees very well with the curve (Fig. 2.8(i)).

Now elementary calculus tells us that for a curve of equation $I = kV^n$, the slope (dI/dV) is given by nkV^{n-1}, thus at point P in Fig. 2.8(i) at 1.0 V:

$$\frac{dI}{dV} = 3.42 \times 1.07 \times 1.0^{2.42}$$

but this is in terms of mA/V, for resistance calculations we

must work in A/V otherwise Ohm's Law does not apply, therefore

$$\frac{dI}{dV} = 3.42 \times \frac{1.07}{1000} \times 1.0^{2.42}$$

$$= 0.003659$$

and the a.c. resistance is derived from the reciprocal of the slope, i.e.

$$\frac{1}{0.003659} = 273\ \Omega$$

We can summarize the two likely equations and their slopes in general form as:

Equation	Slope (dI/dV)
$I = A(e^{kV} - 1)$	kAe^{kV}
$I = kV^n$	nkV^{n-1}

(dI/dV) is known in calculus as the *differential coefficient of I with respect to V*. It is actually the ratio between extremely small differences in both I and V at the same point.

Please note overleaf is a list of other titles that are available in our range of Radio and Electronics Books.

These should be available from most good Booksellers, Radio Component Dealers and Mail Order Companies.

However, should you experience difficulty in obtaining any title in your area, then please write directly to the publishers enclosing payment to cover the cost of the book plus adequate postage.

If you would like a copy of our latest catalogue of Radio and Electronics Books then please send a Stamped Addressed Envelope to:–

BERNARD BABANI (publishing) LTD
The Grampians
Shepherds Bush Road
London W6 7NF
England

BP1	First Book of Transistor Equivalents and Substitutes	60p
BP2	Handbook of Radio, TV and Ind. & Transmitting Tube & Valve Equiv.	60p
BP6	Engineers and Machinists Reference Tables	50p
BP7	Radio and Electronic Colour Codes and Data Chart	25p
BP11	Practical Transistor Novelty Circuits	40p
BP14	Second Book of Transistor Equivalents	1.10p
BP22	79 Electronic Novelty Circuits	1.00p
BP23	First Book of Practical Electronic Projects	75p
BP24	52 Projects using IC741	95p
BP25	How to Build Your Own Electronic and Quartz Controlled Watches & Clocks	85p
BP26	Radio Antenna Handbook for Long Distance Reception & Transmission	85p
BP27	Giant Chart of Radio Electronic Semiconductor & Logic Symbols	60p
BP28	Resistor Selection Handbook (International Edition)	60p
BP29	Major Solid State Audio Hi-Fi Construction Projects	85p
BP30	Two Transistor Electronic Projects	85p
BP31	Practical Electical Re-wiring & Repairs	85p
BP32	How to Build Your Own Metal and Treasure Locators	1.00p
BP33	Electronic Calculator Users Handbook	95p
BP34	Practical Repair & Renovation of Colour TV's,	1.25p
BP35	Handbook of IC Audio Preamplifier & Power Amplifier Construction	1.25p
BP36	50 Circuits Using Germanium, Silicon and Zener Diodes	75p
BP37	50 Projects Using Relays, SCR's and TRIAC's	1.25p
BP38	Fun & Games with your Electronic Calculator	75p
BP39	50 (FET) Field Effect Transistor Projects	1.25p
BP40	Digital IC Equivalents and Pin Connections	2.50p
BP41	Linear IC Equivalents and Pin Connections	2.75p
BP42	50 Simple L.E.D. Circuits	75p
BP43	How to make Walkie-Talkies	1.25p
BP44	IC 555 Projects	1.75p
BP45	Projects on Opto-Electronics	1.25p
BP46	Radio Circuits using IC's	1.35p
BP47	Mobile Discotheque Handbook	1.35p
BP48	Electronic Projects for Beginners	1.35p
BP49	Popular Electronic Projects	1.45p
BP50	IC LM3900 Projects	1.35p
BP51	Electronic Music and Creative Tape Recording	1.25p
BP52	Long Distance Television Reception (TV–DX) for the Enthusiast	1.45p
BP53	Practical Electronic Calculations and Formulae	2.25p
BP54	Your Electronic Calculator and Your Money	1.35p
BP55	Radio Stations Guide	1.45p
BP56	Electronic Security Devices	1.45p
BP57	How to Build your own Solid State Oscilloscope	1.50p
BP58	50 Circuits using 7400 Series IC's	1.35p
BP59	Second Book of CMOS IC Projects	1.50p
BP60	Practical Construction of Pre-Amps, Tone Controls, Filters and Attenuators	1.45p
BP61	Beginners Guide to Digital Techniques	95p
BP62	Elements of Electronics – Book 1	2.25p
BP63	Elements of Electronics – Book 2	2.25p
BP64	Elements of Electronics – Book 3	2.25p
BP65	Single IC Projects	1.50p
BP66	Beginners Guide to Microprocessors and Computing	N.Y.A.
BP67	Counter Driver and Numeral Display Projects	N.Y.A.
BP68	Choosing and Using Your HI-Fi	N.Y.A.
BP69	Electronic Games	N.Y.A.
126	Boys Book of Crystal Sets	25p
160	Coil Design and Construction Manual	75p
196	AF–RF Reactance – Frequency Chart for Constructors	15p
200	Handbook of Practical Electronic Musical Novelties	50p
201	Practical Transistorised Novelties for Hi-Fi Enthusiasts	35p
202	Handbook of Integrated Circuits (IC's) Equivalents and Substitutes	1.00p
203	IC's and Transistor Gadgets Construction Handbook	60p
205	First Book of Hi-Fi Loudspeaker Enclosures	75p
207	Practical Electronic Science Projects	75p
208	Practical Stereo and Quadrophony Handbook	75p
210	The Complete Car Radio Manual	1.00p
211	First Book of Diode Characteristics Equivalents and Substitutes	1.25p
213	Electronic Circuits for Model Railways	1.00p
214	Audio Enthusiasts Handbook	85p
215	Shortwave Circuits and Gear for Experimenters and Radio Hams	85p
217	Solid State Power Supply Handbook	85p
218	Build Your Own Electronic Experimenters Laboratory	85p
219	Solid State Novelty Projects	85p
220	Build Your Own Solid State Hi-Fi and Audio Accessories	85p
221	28 Tested Transistor Projects	95p
222	Solid State Short Wave Receivers for Beginners	95p
223	50 Projects using IC CA3130	95p
224	50 CMOS IC Projects	95p
225	A Practical Introduction to Digital IC's	95p
226	How to Build Advanced Short Wave Receivers	1.20p
227	Beginners Guide to Building Electronic Projects	1.25p
228	Essential Theory for the Electronics Hobbyist	1.25p

BERNARD BABANI BP64

Elements of Electronics

Book 3

Semiconductor Technology

■ The aim of this book and others in the series is to provide an inexpensive but comprehensive introduction to modern electronics. The scope of the series is not restricted by being written to a specific college or course syllabus but concentrates on enabling the reader to gain a thorough understanding of the fundamental principles involved.

■ Although written especially for readers with no more than ordinary arithmetical skills, the use of mathematics is not avoided as so often happens. All mathematics required is taught as the reader progresses and to cater for different levels of knowledge, all study material ancillary to electronics is included in appendices to which the main text refers. This enables readers who are already conversant with the principles involved not to have to digress from the main subject matter.

■ This series is an essential addition to the bookshelves of all those involved, in anyway, with electronics, be it as a career, as a hobby or as a field of study. Much is offered to those who need revision or to be brought up to date with fundamentals, those with grown-up children asking awkward questions or preparing for entry to a college or even those who may find a different approach refreshing.

Cover designed by the Nicholls Design Unit

ISBN 0 900162 84 8

BERNARD BABANI (Publishing) LTD
The Grampians
Shepherds Bush Road
London W6 7NF
England

£2.25