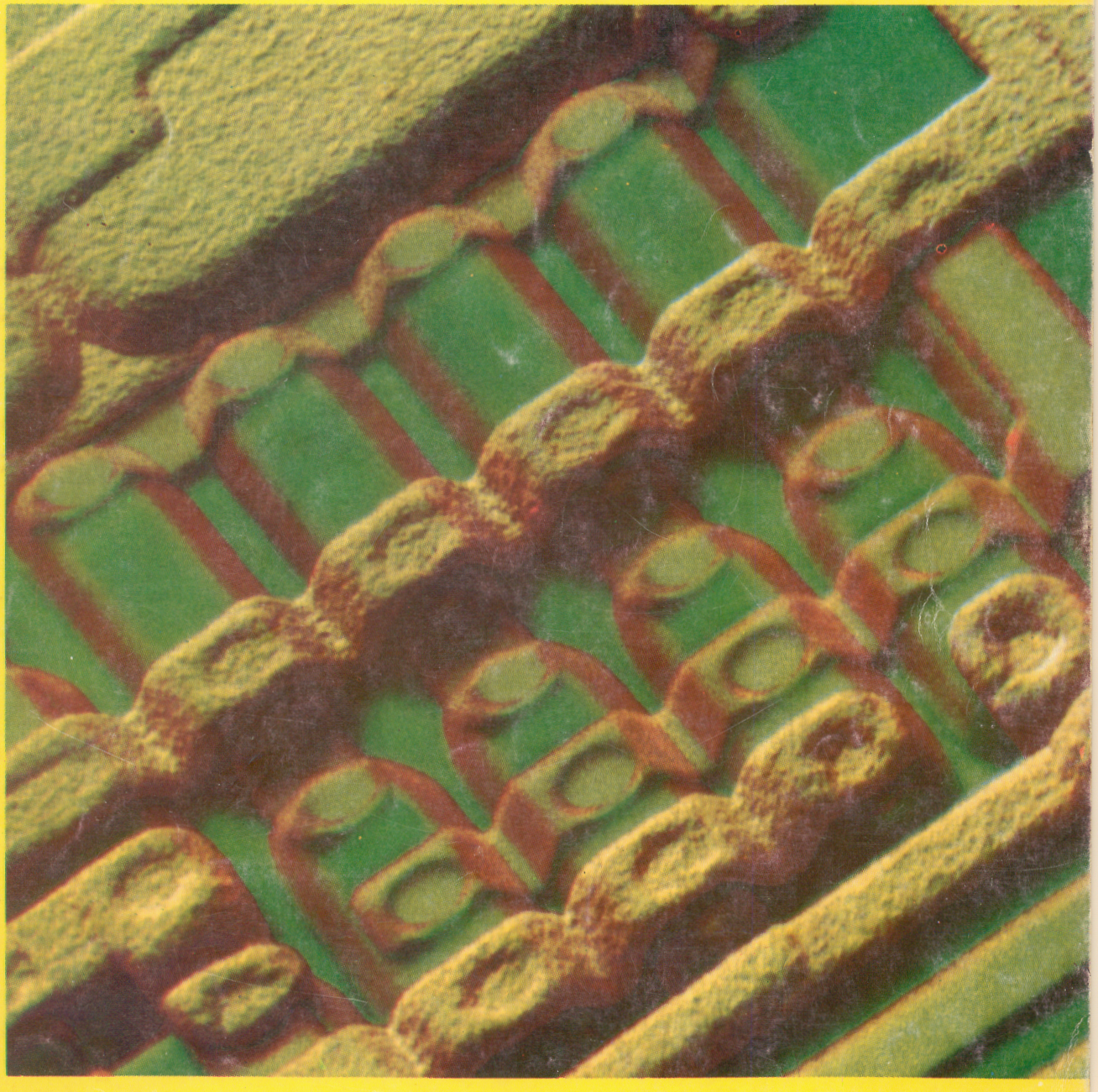
P. RENTON.

Fundamentals of SOLID STATE

An introduction to semiconductors and their applications

JAMIESON ROWE



AN 'ELECTRONICS AUSTRALIA' HANDBOOK

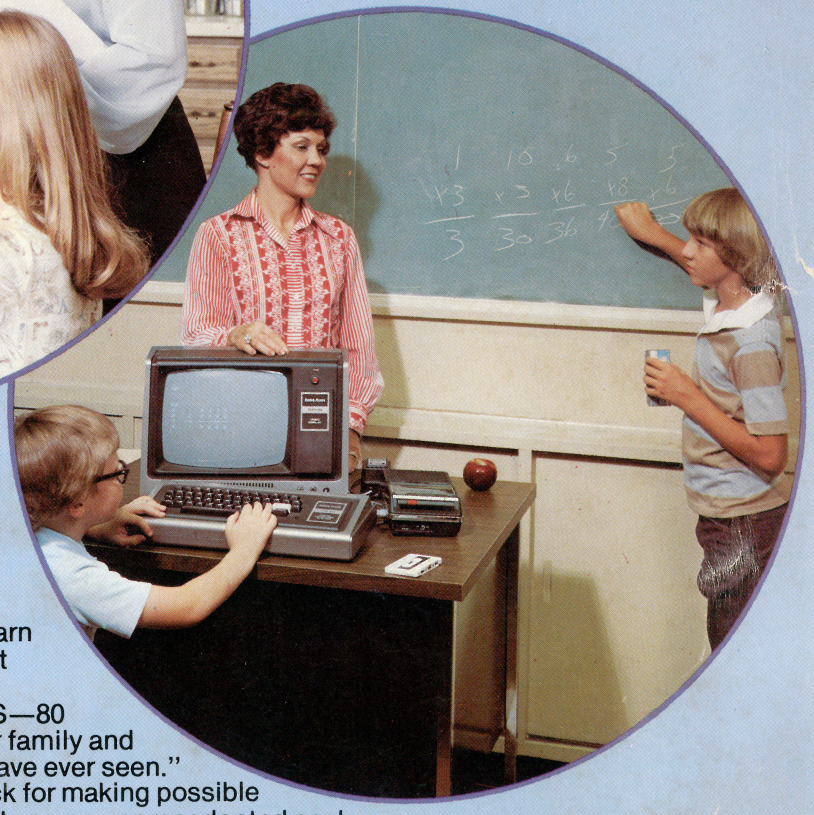
$3.50

The future is here!

Tandy/Radio Shack TRS-80 MICROCOMPUTER SYSTEM

Ideal for Home, Office and School

The world is full of great things: guitars, electronic kits, cameras, and — when you're old enough — cars. But never EVER until this year was there a microcomputer available for under $1000 that was fully wired, ready to use, and in stock for immediate delivery.
Level-I with 4K RAM $799.95



Here are just a few things it can do — with your help, of course. It can teach, remember, display, solve problems, play games like chess and vocabulary-building. It can "invent" new things to do when you learn how to program it. It doesn't need your TV set because it comes with its own 12" screen. A father wrote to tell us his investment in a TRS—80 "is one of the most significant in value to our family and to the future education of our child that we have ever seen." An educator wrote to thank Tandy/Radio Shack for making possible the tapping of human innovation and creativity on an unprecedented scale.

An electronic world based on computers IS ALREADY HERE. If your family doesn't START NOW to understand this world, the gap between understanding and ignorance will rapidly widen. Then the world will be too complex to cope with; the machines will run the people, instead of the people running the machine! Tandy — alone — has the TRS-80. The "tomorrow machine" that's a breakthrough in affordability, reliability, availability.

TANDY ELECTRONICS

Over 160 Stores and Dealers

INDEPENDENT TANDY ELECTRONICS DEALERS MAY NOT BE PARTICIPATING IN THIS AD OR HAVE EVERY ITEM ADVERTISED

Fundamentals of
SOLID STATE

JAMIESON ROWE

B.A. (Sydney), B.Sc. (Technology, N.S.W.),
M.I.R.E.E. (Aust.),
Editor, "Electronics Australia"

SECOND EDITION, COPYRIGHT 1979

Printed by Masterprint Pty. Ltd., of Dubbo, N.S.W., for the publisher, Sungravure Pty. Ltd., of 57/59 Regent Street, Sydney, N.S.W., Australia.

Preface

Much of the material in this book was first published in the monthly magazine "Electronics Australia", as a series of articles. In both the original articles and the present book I have attempted to provide a basic introduction to modern semiconductor devices and their operation. An introduction not just for the service technician and the electronics hobbyist, who perhaps may never wish to delve into the subject in greater depth, but also for the university or technical college student who may need a broad introduction to semiconductor concepts before plunging into the mathematics.

There are many other introductory books on this subject, but most fall into two broad groups. In one group are the very elementary books, which are very easy to read and understand but generally don't give you much more than a very superficial understanding. In the other group are books written for the college student, which tend to assume that the reader has a thorough grasp of solid state physics and advanced calculus.

In this case I have tried to steer a middle course. The book starts at a very basic level, and doesn't deal with the mathematics of solid state physics at all; yet at the same time it tries to present many of the concepts normally found only in the more advanced books. Concepts like the nature of a crystal lattice, energy bands, carrier diffusion and drift, and so on.

To a certain extent the inclusion of these concepts may tend to make the book less easy to read. However I believe this is justified by the richer and more satisfying insight they give into device operation. In any case I have tried to present these concepts in particular as clearly as possible, to minimise the additional effort required by the reader.

For this second edition, chapter 17 has been completely rewritten to bring the book up to date. The discussion of recent advances in fabrication technology and current development trends represents the situation in January, 1979, as accurately as I have been able to determine it from current overseas journals and a recent trip to California's "silicon valley", the world heart of semiconductor manufacture. The glossary has also been revised and updated, to make it of greater potential value.

Needless to say, no book of this type is the work of one person, and I should like to thank a number of people who played important parts in making the present book possible. Many thanks are due to Neville Williams, Editor-in-Chief of "Electronics Australia", Assistant Editor Phil Watson, and indeed the whole staff of the magazine, whose constructive criticism and friendly advice has surely helped to improve the quality of the text. And I would especially like to thank draftsman Bob Flynn, whose co-operation and involvement extended far beyond the preparation of diagrams.

— Jamieson Rowe
January 1979

Contents

Chapter 1

ATOMS AND ENERGY

Introduction — modern concept of the atom — electrons as both particles and waves — "allowed" orbits — electron energy and energy levels—energy level capacities—valence —excitation and energy "quanta"—radiant energy as both waves and particles.

The concept of an atom as a micro-miniature "solar system" is a familiar one to most people in electronics. According to this picture, atoms consist of a central and relatively heavy nucleus having a positive electrical charge, around which orbit smaller, lighter and negatively charged electrons whose number is such that in the "normal" state the atom carries zero nett charge. For each of the chemical elements, the atomic nucleus has particular values of mass and positive charge, and is accompanied in the "normal" state by the appropriate number of orbiting electrons.

Consistent with this picture is the idea that electrical conduction is a mechanism in which an applied electric field causes outer orbiting eletrons to be freed from their atoms, whereupon they can wander through the material to form the traditional "current flow." Conducting materials such as metals are thus understood as materials in which the outer electrons are "loosely bound" to the atomic nuclei, while insulating materials are in contrast those in which the electrons are "tightly" bound, and unable to wander.

For many years, this simple and quite easily grasped concept of atomic structure and its relationship to electrical conduction proved quite satisfactory for most purposes. It was generally adequate, for example, for an understanding of the operation of thermionic valves and the circuits in which they were used. However as science, and consequently technology, progressed it was found increasingly that the simple picture could not adequately explain many of the new discoveries. It became necessary, as with so many scientific concepts, to both revise and expand our conception of atomic structure, and with it our understanding of electrical conduction.

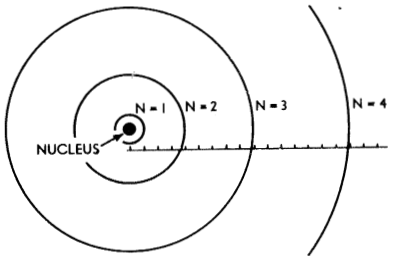
Hence it is that, in order to gain a clear knowledge not only of the mechanisms of electrical conduction as it is currently understood, but also and in particular of the operation of the many different semiconductor devices which are used in — and have virtually revolutionised — modern electronics, one must begin by becoming acquainted with the atom as it is now pictured.

Unfortunately, perhaps, a full understanding of modern atomic theory and the physics of electrical conduction requires a thorough grasp of the abstract and highly mathematical science of Quantum Mechanics; and this is beyond many professional engineers. However a full understanding in this sense is really only necessary for the scientist, research student and device development engineer. A somewhat more limited understanding at a basically "qualitative" level is usually found both adequate and satisfying for most other purposes, including that of preparation for further detailed study. It is at this level that the following treatment is pitched.

Perhaps the first thing to be noted about the modern view of the atom is that it is somewhat more "fuzzy" than before, and in consequence it tends to be less satisfying. Although disconcerting, this must unfortunately be accepted as a fact of life. The fact is that the apparent clarity of the simple "solar system" picture was an illusion, with no real justification on the basis of our actual knowledge.

We are unlikely to know for some considerable time, if indeed we will ever know, the "real" nature of electrons and other sub-atomic "particles," or of such fundamental things as mass, energy, time, electric and magnetic fields, and electrical charge.

The modern picture of the atom and its behaviour tries to take this lack of knowledge into account. In producing a theory which "works," in the sense that it can satisfactorily explain most of the little we do actually know, it aims at the same time at preventing us from kidding ourselves that we know more than this!



ALLOWED ELECTRON ORBITS
(N = QUANTUM NUMBER)

Figure 1.1

At this stage the reader may well be wondering if the modern picture of the atom bears any resemblance at all to the simpler one. The truth is that there is a resemblance, although only a general one.

In broad terms, the atom may still be pictured as consisting of a central positively charged nucleus, surrounded by a number of negatively charged electrons. As the nucleus plays no more than a nominal part in electrical behaviour, we need not concern ourselves here with its structure. Suffice to say that it is just as well that this is the case, because the closer physicists examine the nucleus, the more complex does it seem to become!

The electrons are still held to be the components of the atom which are responsible for its electrical and chemical behaviour. However, they can no longer be regarded simply as tiny physical particles orbiting around the nucleus, nor can the part which they play in electrical conduction be pictured as a straightforward one whereby an electric field "loosens" those in the outermost orbits and whisks them along to form a current flow. As with the nucleus, the closer the electron and its behaviour are examined the more complex—and in this case, the more elusive—does it become.

It has been found that, in some circumstances, the behaviour of electrons can indeed only be explained by visualising them as small particles. Yet, equally, there are other situations in which their behaviour can only be explained as consistent with that of small bursts of oscillations or "waves" of a type similar to, but different from, those responsible for sound, heat and light. In other words, an electron must now be regarded as a somewhat vague thing which sometimes behaves as a physical particle, and at other times alternatively behaves as a "packet" of some sort of waves.

As it happens, it is the wave aspect of their "personality" which seems to play the major part in determining the behaviour of electrons as they surround the nucleus of an atom. So that in place of the simple picture of a number of electron "planets" orbiting around the nucleus, we must now try to visualise a system of spherical and and elliptical "surfaces" at various distances from the nucleus, and each somewhat fuzzy and indistinct because of wavelike variations over the perimeter.

Whereas it would appear that, at the more familiar macroscopic level, planets may orbit around a sun at virtually any radius providing they have the appropriate orbital velocity, this does not occur in the microscopic level of the atom. Electrons are only able to "orbit" (the term is still used, for convenience) around the nucleus at **certain definite radii.** In terms of the wavelike aspect of the electron these radii can be interpreted broadly as those whose perimeter corresponds to an integral (or whole-number) multiple of a compatible electron "wavelength."

Although this concept may seem strange and rather hard to accept, the full reasoning behind it is quite abstract and involves mathematical "gymnastics" which we cannot deal

with here. However, for the present it may help to compare the situation with the more familiar one involving the production of standing waves in a stretched string: waves can only occur at frequencies at which the string length corresponds to a single wavelength, two wavelengths, three wavelengths, and so on.

The electrons of an individual atom, then, can only occupy orbits corresponding to certain "allowed" effective radii. This is illustrated by the diagram of figure 1.1. As may be seen, the various possible orbits are each assigned a so-called **quantum number,** commencing at "1" for the innermost. The effective radius of the orbits increases with the square of the quantum number, i.e., 1 unit, 4 units, 9 units, and so on.

Some idea of the size of the orbits may be gained from the fact that the innermost or N=1 orbit corresponds to an effective radius of approximately 5 $\times$ 10^{-11} metre, or about 50 millionmillionths of a metre.

Associated with each possible orbit is a corresponding **energy level;** i.e., an electron occupying a particular orbit will have a particular amount of energy. This will consist of both the kinetic or "motional" energy associated with its orbiting momentum, together with the potential or "latent" energy which it possesses by virtue of its position in the electric field surrounding the nucleus.

Because of the opposite charges of electrons and nuclei, an electron is attracted to the nucleus with a force which varies inversely with the square of its distance from the nucleus centre. In view of this, an electron at a particular point in the electric field surrounding the nucleus has a positive potential energy with respect to that nucleus, and at the same time a negative potential energy with respect to any point more distant from the latter.

If these polarities seem wrong, remember that positive potential energy corresponds to the ability to perform work or release "internal" energy, while negative potential energy implies a need for energy to be externally supplied.

From the point of view of the electron, therefore, the vicinity of the nucleus represents an area of lower or "more negative" potential energy than elsewhere. In fact the field around the nucleus forms what may be visualised as a potential energy "pocket" or well, with the nucleus at its centre and the "sides" sloping exponentially. Viewed in this light, a free electron wandering near the nucleus and attracted to it effectively "falls into the well." These ideas are illustrated in the diagrams of figure 1.2.

According to this view, an electron which is orbiting around the nucleus does so (rather than "fall") by virtue of its orbital momentum — in effect, it "rolls around" the walls of the potential energy well at a sufficient speed to prevent itself from falling. An orbiting electron thus possesses a positive kinetic energy, and because the required orbital momentum increases with decreasing effective orbit radius, the kinetic energy similarly increases. In fact it turns out that the positive kinetic energy follows the same exponential curve as that of the negative potential energy, but with opposite sign and with an amplitude half as great if spherical orbits are considered.

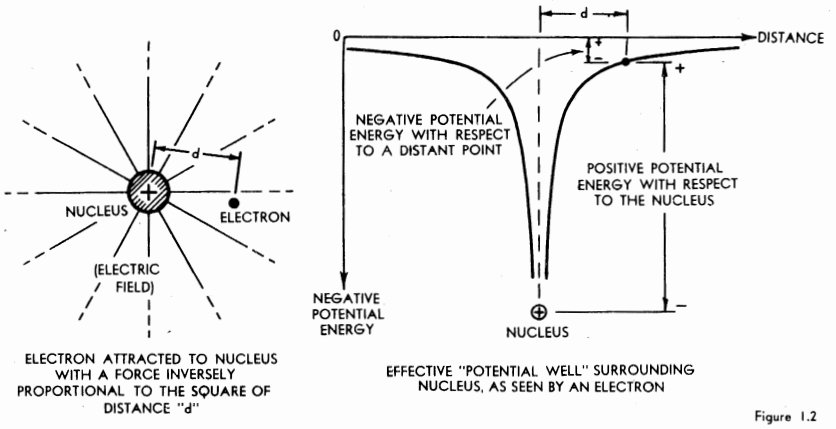


Figure 1.2

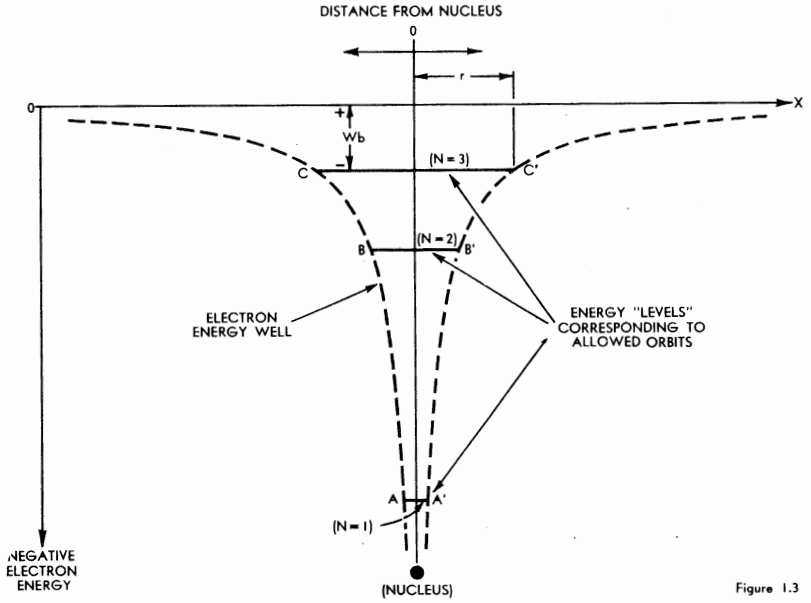


Figure 1.3

As the total energy of an orbiting electron consists of the algebraic sum of its potential energy (negative) and its kinetic energy (positive), and both these follow exponential laws with the former larger than the latter, the total energy will thus be negative and will also follow an exponential. In short, the vicinity of the nucleus represents for orbiting electrons a **total energy well,** similar to that for potential energy but "less deep." A two-dimensional representation of this well is shown by the dashed curved lines in figure 1.3.

An example may help in clearing up any possible confusion at this point. An electron in an orbit of effective radius "r" is seen to occupy an energy level represented in figure 1.3 by the line C-C', with a total negative energy of Wb. From the shape of the dashed outline of the energy well, it may be seen that the smaller the effective orbital radius, the greater the negative energy possessed by an electron in that orbit.

It should be fairly evident at this stage that removal of a particular orbiting electron from the influence of the nucleus (i.e., taking it to an effectively distant place) will involve doing positive work, to a degree which corresponds exactly to the negative energy level of the orbit concerned. Hence an electron occupying the energy level C-C' in figure 1.3, in order to be "freed" from the nucleus altogether, must acquire a positive energy equal and opposite in sign to Wb.

In short, the negative energy level of an orbit simply corresponds to the degree to which an electron in that orbit is "bound" to the nucleus — the orbital **binding energy.**

As we saw earlier, in an individual atom electrons can only occupy orbits having certain allowed effective radii. Hence, in terms of the energy level diagram of figure 1.3 an orbiting electron must occupy one of the discrete energy levels represented by such lines as A-A', B-B', C-C', and so on. Level A-A' might correspond to the N=1 orbit of figure 1.1, for example, and level B-B' to the N=2 orbit.

Although only three of the permitted energy levels are shown in figure 1.3, there is in fact a very large number, corresponding to allowed orbits with effective radii increasing rapidly with the squares of successive quantum numbers. Because of the exponential shape of the energy well around the nucleus the energy differences between successive orbits actually decreases, however, so that if further levels were shown in figure 1.3 they would be seen to form a series of horizontal lines with decreasing vertical spacing, above level C-C' and approaching the zero energy level represented by O-X.

One might perhaps imagine, from the foregoing, that in an individual atom of an element all of the electrons surrounding the nucleus would be found occupying the lowest (most negative) energy level, at least when the atom is in the **ground state** with no

additional energy or "excitation" received from external sources. However, this is not so.

In fact it is found that, in effect, each energy level has a definite electron "capacity"; only two electrons can occupy the N=1 level, only eight can occupy N=2 level and so on.

The maximum number of electrons which may occupy the first five allowed energy levels are 2, 8, 18, 32 and 50 respectively.

Although quantum mechanical theory provides an adequate explanation of the electron capacities of the various energy levels, the detailed arguments involved are beyond the scope of the present treatment. For the present it should be sufficient to note that in addition to their energy level, electrons in orbit have other important characteristics such as degree of orbit ellipticity, magnetic moment, and spin polarity. It is believed that only certain combinations of these characteristics are permitted at each energy level, and further that no two electrons at the same energy level can have the same combination. The latter "law" is held to apply to any unified system involving electrons, and is known as **Pauli's exclusion principle.**

In an individual atom in the ground state, then, the electrons occupy the lowest permitted energy levels to a degree determined by the various energy level capacities. For example in a boron atom, with five electrons, two occupy the N=1 level, which is thus "filled," while the remaining three occupy but only partly fill the N=2 level; the remaining levels are empty. Similarly the fourteen electrons of the silicon atom are disposed with two filling the N=1 level, eight filling the N=2 level, and the remaining four partly filling the N=3 level.

Table 1.1 gives the electron dispositions of the first 20 elements of the "periodic table," illustrating the way in which the various energy levels are progressively "filled."

It is those electrons in the outermost of the **occupied** energy levels of an atom which almost completely determine its external behaviour, both chemical and electrical. The electrons which may be present in any filled lower energy levels play little part in external behaviour, because they are relatively strongly bound to the nucleus. Accordingly the latter are usually called the "core" electrons, and can often be considered as "lumped together" with the nucleus, whereas the former are called the **valence electrons** (from the Latin "valere," meaning strength; an allusion to the part played in chemical bonding), and are almost always considered separately and in detail.

The energy level occupied by the valence electrons of an atom of a particular element is consequently known as the **valence level** for that element. Each of the allowed energy levels is the valence level for atoms of certain elements; for example the N=2 level is the valence level for atoms of elements such as boron, while the N=3 level is the valence level for elements such as silicon.

In the foregoing description of the structure of the atom as it is currently pictured, we have considered the atom in the so-called ground state. Actually, this state is a purely hypothetical one; it would only occur if an atom could be placed in a light-tight and radiation-proof container maintained at a temperature of absolute zero (-273°C). Let us therefore look briefly at the more usual situation, where an atom is at a somewhat more comfortable temperature and is accessible to light and possibly other forms of radiation.

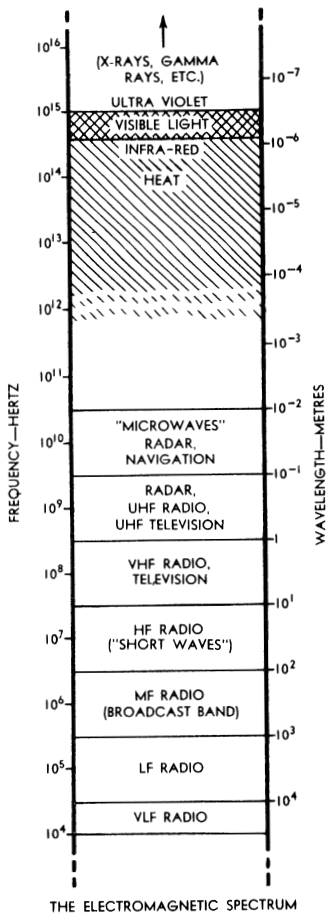


Figure 1.4

Most readers will probably be aware that light, heat and other forms of radiation such as X-rays are essentially energy, in the form of electromagnetic waves. As such, they are related to the familiar waves used for communication and for sound and television broadcasting. They differ from the latter almost solely in terms of frequency, or wavelength; in fact heat radiation corresponds virtually to "super-super-high frequency" radiation, or "ultra-ultra-short waves," while light and X-rays correspond to even higher frequencies and shorter waves again. These relationships are illustrated in figure 1.4, which shows the relevant portion of the electromagnetic spectrum.

In view of this, it should not be hard to understand that an atom which is in any practical situation involving light, heat and the other forms of radiation is virtually subjected to a constant bombardment of energy. And it should be no surprise that in such a situation the atom will tend to be found not in its ground state, but in one of many possible "excited" states which correspond to its having absorbed—at least temporarily — additional energy.

As one might perhaps guess, the mechanism by which an atom "absorbs" energy to become excited is a rather complex and obscure one; so too is the converse mechanism whereby the atom "ejects" energy to return to

TABLE 1.1

Element	Number of Electrons (Atomic Number)	Occupation of Orbits/Energy Levels					
		N=1	N=2	N=3	N=4	N=5	N=6
Hydrogen	1	1					
Helium	2	2					
Lithium	3	2	1				
Beryllium	4	2	2				
Boron	5	2	3				
Carbon	6	2	4				
Nitrogen	7	2	5				
Oxygen	8	2	6				
Fluorine	9	2	7				
Neon	10	2	8				
Sodium	11	2	8	1			
Magnesium	12	2	8	2			
Aluminium	13	2	8	3			
Silicon	14	2	8	4			
Phosphorus	15	2	8	5			
Sulphur	16	2	8	6			
Chlorine	17	2	8	7			
Argon	18	2	8	8			
Potassium	19	2	8	8	1		
Calcium	20	2	8	8	2		

either the ground state or a lower excited state. For a full explanation, as before, one must delve quite deeply into quantum mechanics. However, there is a basic and important principle involved, and one which we can consider here briefly.

Stated simply, the principle is as follows: The absorption of energy by an atom corresponds to the transfer of electrons to higher energy levels. Because there are only certain allowed energy levels in an atom, as we have seen, this means that energy can only be absorbed in "lumps" or **quanta** of definite sizes. The sizes of the quanta correspond to the energy differences between the various allowed levels.

Hence an atom can absorb an amount of energy corresponding to the transfer of an electron from the N=1 level to the N=3 level, for example, or to the transfer of perhaps three electrons at the N=2 level to the N=4 level. But, whatever the quantum of energy absorbed, it must correspond to the transfer of a whole number of electrons from one of the allowed energy levels to other, higher levels.

And the same principle holds for emission of energy, which as one would expect involves transfer of electrons from higher to lower allowed energy levels. An atom can only emit energy in quanta of fixed sizes, corresponding to the transfer of whole numbers of electrons from higher to lower allowed energy levels.

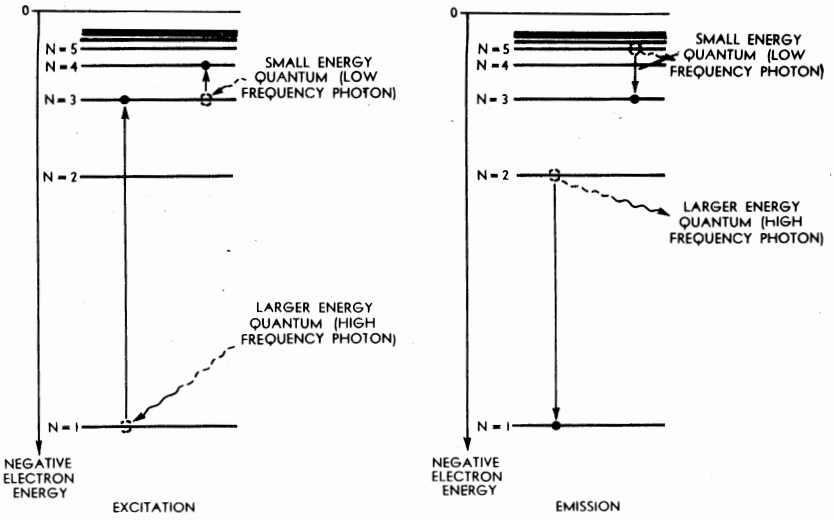
At this point the reader may well be asking how it is possible for atoms to be able to absorb and emit energy in discrete quanta, when the energy absorbed and emitted is in the form of supposedly continuous radiation such as light or heat.

The answer to this is that in fact electromagnetic energy, like the electron, behaves in many ways as if it too has a "split personality." In contrast with its continuous wavelike nature, it can equally readily behave as if it consists of small particles or quanta of energy. These particles have been named **photons.**

It happens that the amount of energy represented by a photon is independent of the intensity or "strength" of the radiation concerned; this only determines the number of photons present. **Rather, the energy of a photon is directly proportional to its frequency.** This is a very important relationship which was discovered by the physicist Max Planck in 1900 and developed by Albert Einstein in 1905.

According to this relationship, photons of "blue" visible light represent larger energy quanta than photons of lower frequency such as "red" light, and the latter in turn represent larger quanta than photons of heat radiation. Also, and very importantly for our present purposes, heat photons corresponding to higher **temperatures** represent larger energy quanta than those corresponding to lower temperatures. This arises because temperature is a direct function of frequency.

From the foregoing it may be seen that, because it is only capable of absorbing energy quanta of certain fixed sizes corresponding to electron transfers between allowed energy levels, an atom can effectively only be excited by radiation of particular frequencies (wavelengths). Each frequency will correspond to an electron transfer between a particular combination of



(HORIZONTAL AXES HAVE NO SIGNIFICANCE)

Figure 1.5

levels; hence a transfer from the N=1 level to the N=3 level might result from absorption of a photon of frequency f1, while a photon of another frequency f2 might produce a transfer of an electron from the N=3 to the N=4 level.

Similarly the ejection of energy by an atom dropping to the ground state or to a lower excitation state results in the emission of radiation only at particular frequencies. An electron transfer from the N=2 level to the N=1 level might result in the emission of a photon of frequency f3, for example, while a transfer from the N=5 to the N=3 level would result in the emission of a photon at a different frequency.

These concepts are illustrated in the diagrams of figure 1.5.

In practical situations atoms can thus be found tending to continuously absorb and emit radiation at a number of specific frequencies, each of which corresponds to one of the possible energy level transitions. It is this behaviour which accounts for the so-called "line spectra" obtained by analysis of the wavelengths of light and heat absorbed and emitted by atoms of the various elements under suitable conditions.

As one might expect, the number of specific photon frequencies involved in atom energy absorption and emission tends to be quite large, as there are many possible energy level transitions. This is particularly so with elements having many electrons surrounding the nucleus. However due to differences between levels concerning the allowed "secondary" electron characteristics of orbit ellipticity, magnetic moment and spin, some level transitions tend to be more prevalent than others, in a fashion which varies from element to element. As a result each element tends to have a characteristic pattern of "dominant" absorption and emission frequencies.

An atom in the excited state contains, as we have seen, electrons which are occupying higher energy levels than they would occupy in the ground state. It is interesting to consider whether we can make any inferences regarding which of the electrons will be more likely to be found at such higher levels.

As it happens, we can. Earlier, we saw that the energy differences between the allowed energy levels **decrease** with increasing orbit radius and quantum number. Hence somewhat greater energy would be required to transfer an electron from the innermost N=1 level to the next or N=2 level for example, than to transfer an electron from the N=3 level to the adjacent N=4 level. Thus even for transfer between adjacent levels, the electrons at the lower levels require larger energy quanta.

There is also the electron capacity of the various levels to be considered, i.e., Pauli's exclusion principle. As the capacity of the various levels does not alter with excitation, this means that a transfer of an electron to a particular energy level can only take place if there is a "vacancy" at that level. From this it can be seen that transfer of electrons from the higher levels is more likely to occur than from the lower levels, both because the lower levels are more likely to be "full" and also because the capacity of the levels increases with increasing energy.

We can say, then, that for a given degree of excitation, the "excited level" electrons will tend to be those which already occupy the higher levels in the ground state, rather than those from the lower levels. In particular, there will tend to be a high proportion of the electrons from the valence level of the atom concerned.

SUGGESTED FURTHER READING

BURFORD, W. B., and VERNER, H. G., **Semiconductor Junctions and Devices,** 1965. McGraw-Hill Book Company, New York.

MORANT, M. J., **Introduction to Semiconductor Devices,** 1964. George G. Harrap and Company, London.

SCROGGIE, M. G., **Fundamentals of Semiconductors,** 1960. Gernsback Library Inc., New York.

SHIVE, J. N., **Physics of Solid State Electronics,** 1966. Charles E. Merrill Books Inc., Columbus, Ohio.

SMITH, R. A., **Semiconductors,** 1959. Cambridge University Press.

Chapter 2

CRYSTALS AND CONDUCTION

Atoms in combination — energy interaction — crystalline solids and energy bands—the valence band—conductors and electrical conduction—insulators and semiconductors —the effect of excitation—electrons and holes—crystal conductivity and resistivity.

Having looked at the modern concept of atomic structure, and at what might be called the "internal" behaviour of individual atoms, let us now examine what happens when atoms link together to form molecules and "solid" matter. It should become apparent as we progress that knowledge of this "external" behaviour is essential for a clear understanding of electrical conduction.

We have seen that in an individual atom, the electrons surrounding the central nucleus can only occupy certain "allowed" orbits, each of which correspond to a particular value or level of total electron energy, and that in the unexcited or "ground" state the electrons of an atom are found occupying the orbits nearer the nucleus in numbers determined by the orbit capacities. We have also seen that in a practical situation involving light, heat and other forms of radiant energy, electrons are constantly transferred back and forth between allowed orbits as the atom absorbs and emits "lumps" or quanta of energy whose sizes correspond to the energy differences between the various levels.

Two individual and separate atoms of the same element will possess the same allowed orbit structure, or in other words the energy levels of their allowed orbits will be identical. Note that in saying this we make no reference to the electrons occupying the levels, but refer only to the allowed levels themselves. Hence it is not implied that at every instant of time each atom will have exactly the same excitation energy, with identical numbers of electrons at each level. In fact this would not be so even if their situations were equivalent, because the random nature of energy absorption and emission would produce instantaneous differences such that all we could say is that they had the same **average** excitation energy.

A most interesting thing happens if two such atoms are brought near to one another: the electric fields around the two nuclei interact in such a way that each of the allowed electron energy levels of both atoms progressively "splits" into a **pair** of levels (orbits), whose energy difference increases as the two atoms are brought closer together. At first sight, this may seem quite inexplicable; however a moment's thought should show that it is no more so than many other similar effects with which the reader is likely to be familiar.

Recall, for example, that when two resonant circuits tuned separately to the same frequency are coupled together, they interact such that in the coupled state neither is resonant at the original frequency, but both are effectively resonant at two new adjacent frequencies whose separation depends upon the degree of coupling. It is this effect which produces the familiar "double humping" associated with large coupling factors.

Another example occurs in the case of loudspeakers fitted into tuned enclosures. Here a loudspeaker cone system and an enclosure, having the same resonant frequency when separated, interact when together to produce the same sort of double resonance — which in this case is used to smooth the low-frequency response.

In fact, it is found that this sort of interaction effect is quite universal where oscillatory systems are concerned. Therefore it should not be surprising that it occurs between the allowed electron energy levels of "coupled" atoms, particularly as we have seen that each energy level corresponds to an orbit which represents a particular mode of "oscillation" associated with the wavelike aspect of electron behaviour.

As one might expect, it is the highest or least negative electron energy levels of two atoms which first split as they are brought nearer, because these correspond to the largest allowed orbits. For the same reason it will be the level pairs produced by these levels which will be found most widely separated for any given distance or spacing between the two atomic nuclei. This is illustrated in figure 2.1, which shows the splitting of the various energy levels as a function of the nucleus spacing.

Note particularly that this diagram applies equally to either atom, and that in the interests of clarity only the first four levels are shown. It may be seen that for large spacing, the levels are unaltered from their "individual atom" values, but as the spacing decreases they split progressively from the higher levels. At a spacing distance D1, for example, only the N=4 level has split, while at a smaller spacing D2 both the N=3 and N=2 levels have split also but by smaller amounts. At a still smaller spacing D3, the lower of the pair of levels corresponding to the N-4 level has moved below the higher of the N=3 pair. Such "overlapping" occurs more and more as the spacing is reduced.

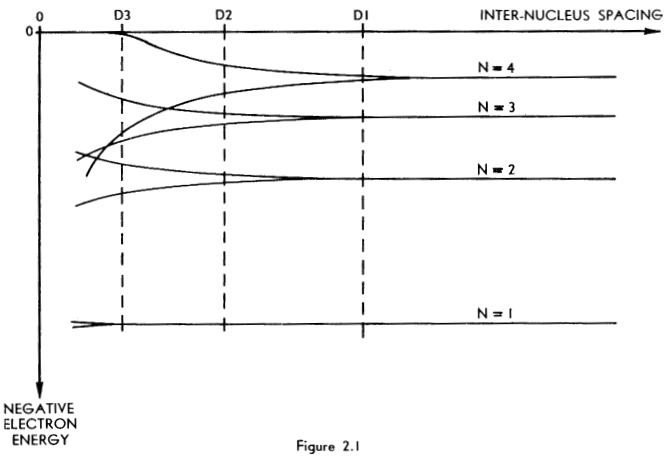


Figure 2.1

What does this mean? Simply that when two similar atoms are placed relatively near one another, their interaction effectively alters and increases the number of "allowed" orbits for the electrons surrounding each. Hence when the atoms whose behaviour is represented by figure 2.1 are spaced at a distance D2 apart, each has two new allowed orbits in place of each of the orbits corresponding to its previous N=4, N=3 and N=2 energy levels. As splitting occurs progressively from the highest levels down, this will also mean

that all of the higher levels not shown will also have split into two, so that each atom will have very many more allowed orbits than it would have had alone. (In fact the number of allowed orbits will have almost doubled, as in this example only the N=1 level has remained unaltered at a spacing of D2.)

It so happens that, in the same way that the energy levels of two atoms split into pairs when they are brought together, the energy levels of larger numbers of relatively close atoms are found to split into a corresponding number of new levels. With three atoms, the levels each tend to split into triplets; with four atoms, into quadruplets, and so on.

Accordingly, if we have a lump of an element comprising a large number "M" of atoms regularly spaced at a particular distance, certain of the "individual" energy levels will be found to have split into the same large number of M new energy levels, forming **bands.** The number of levels which will have split into such bands, and the energy width of the bands, will depend upon the atomic spacing, with the higher levels splitting before the lower and to a greater extent.

An example may help in picturing this situation. A cube of metal measuring one centimetre on each side typically consists of something like 10^{23} atoms — one-hundred-thousand-million-million-million. This means that in place of certain of the higher energy levels of an individual isolated atom of the metal concerned, each of the atoms of the metal cube will have bands each containing no less than 10^{23} extremely closely spaced individual levels. A cube one-hundred-thousand times smaller in volume will similarly have 10^{18} levels in each of the atomic bands—still a very large number!

In both cases the number of bands present, and their "width" in terms of energy levels, will depend as before only upon the inter-atomic spacing. In fact the number of bands and their width is exactly the same as the number of "paired" levels and the separation widths produced for the simple case of only two atoms, illustrated in figure 2.1. Hence, although the size of a lump of material determines the number of discrete levels making up each of the energy bands, it does not affect either the number or the width of the bands.

The type of atomic interaction which we have been considering occurs almost only in the "solid" state of matter, as opposed to the "liquid" and "gaseous" states, because it is only in the solid state that the spacing between atoms is sufficiently small and relatively fixed. And as one might expect, the solid materials whose behaviour most closely conforms to this picture are those in which the atoms are arranged in very regular 3-dimensional "lattice" patterns —the **crystalline solids.**

The electron energy relationships inside a typical crystal structure are illustrated in figure 2.2, which is a two-dimensional energy/distance representation of the same type as that for a single atom given previously in figure 1.3.

It may be noted that in this example the **lattice spacing** or distance between the atomic nuclei is such that the N=1 and N=2 energy levels have remained unaltered, while the N=3 and higher levels have split into the expected bands each comprising M closely spaced new levels. In fact overlapping of the N=5 and higher bands has produced virtually a single "higher band," extending right up to the zero energy level. Such overlapping tends to occur with the higher levels in crystalline solids, both because the splitting is greater for these levels, and also because as we have seen previously the energy differences between the original atomic orbit levels decrease with increasing distance from the nucleus.

In this example the N=3 band is shown as the **valence band,** which corresponds to the valence electron energy level of the individual atoms concerned. Although shown here as an isolated band, not overlapped by higher bands, the valence band is not necessarily so isolated, and is in fact overlapped in certain crystals.

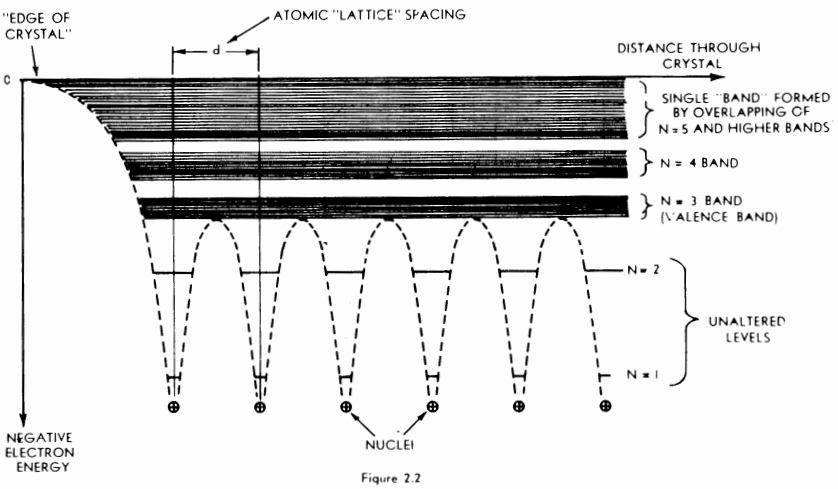


Figure 2.2

As a result of the interactions between the atoms of the crystal lattice, only the walls of the electron energy wells (dashed lines) surrounding the nuclei at the edge of the crystal rise fully to the zero energy level, as they do with an isolated atom. For the nuclei inside the crystal, the well "walls" fuse and cancel at a somewhat lower level, as shown. The level at which they fuse is in fact very close to the valence band, and this has considerable importance.

It may be noted that below the fusion level, the original electron energy levels are unaltered, and that they are shown as before separately for each nucleus. Conversely above the fusion level, all levels have become bands, and are shown extending continuously throughout the lattice. The significance of these distinctions is that electrons occupying energy levels below the fusion level are influenced almost solely by the individual atomic nuclei, whereas electrons occupying the energy bands above the fusion level are virtually uninfluenced by single individual nuclei, and are effectively "common property."

In other words, this means firstly that electrons having low or more negative energy can exist in the crystal lattice only in orbits closely surrounding the individual nuclei. These are the highly bound "core electrons," and they will be those occupying orbits corresponding to the unaltered energy levels represented in figure 2.2 by the N=1 and N=2 levels.

On the other hand, electrons having higher or less negative energy can occupy any of the many levels comprising the valence and higher bands, in which they are no longer the "property" of individual atoms but belong only to the crystal as a whole. Those whose energy places them within the valence band are thus "shared" equally by all the atoms of the crystal, and it is in fact these electrons which effectively bind the crystal together. Any electrons in the higher bands are even less restrained than these, having at the same time less negative potential energy and more kinetic energy, and these can accordingly move with increasing freedom anywhere inside the crystal.

It is those electrons in the "common property" valence and higher energy bands of a crystalline solid which are responsible for its electrical behaviour, and the part played in this regard by such electrons is very largely determined both by the relative positions of these bands, and by the distribution of electrons in them. Hence in order to gain an insight into electrical conduction in a crystal, we must look closely at both the bands themselves and the ways in which electrons can occupy them.

There are a number of different ways in which atoms can link or "bind" together to form crystal structures. Depending upon the type of atomic bond involved, and the size of the atoms, a particular crystal lattice will have a definite inter-atomic spacing, and thus an appropriate number of the atomic electron energy levels will be split into bands of appropriate width. The disposition of electrons in the allowed levels and bands will depend, as before, upon both their disposition in the ground state of an individual atom, and on the excitation energy of the atom concerned.

From this is may be appreciated that each crystal structure formed by the various elements will tend to have a different and unique overall energy pattern, with different energy band widths and spacing, and each different with respect to the number of electrons occupying the various levels and bands at a given temperature.

Despite this, it happens that most crystalline solids fall into only two

broad categories when one considers the electron energy situation associated with the valence and higher energy bands. One of these situations applies in the case of metal crystals which are excellent electrical conductors; the other applies in the case of crystalline solids which are basically either semiconductors or insulators.

The first type of situation is basically that in which the valence electrons of the various atoms of the crystal do not completely fill the valence band in the ground state, as illustrated in figure 2.3.

This situation can arise if the electrons of an individual atom of the element concerned do not fill the original valence level; it can equally be caused by a crystal lattice spacing which results in overlapping of the "true" valence band by a higher order band or bands, to produce a much wider effective valence band. For our purposes it does not matter which factor is responsible, the essential point being that the valence band is not completely filled.

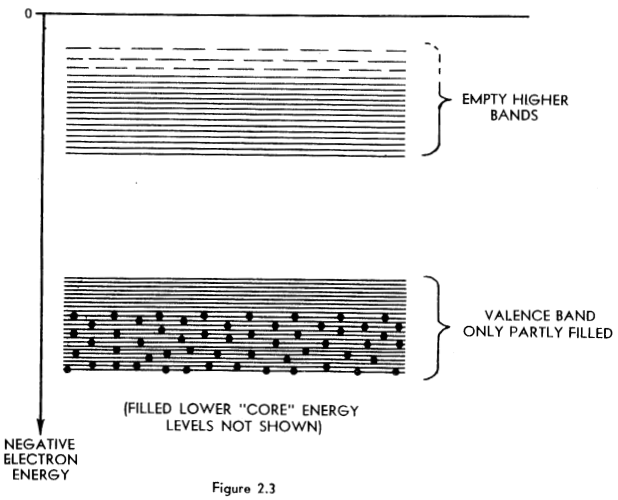


Figure 2.3

In order to understand how this situation allows the crystal concerned to act as a good electrical conductor, consider for a moment what happens when an external source of EMF is connected across the crystal. Due to the applied EMF, an electric field is set up through the crystal; as a result one end of the lattice has an effective potential energy with respect to the other, so that the various electron energy levels and bands no longer remain horizontal but have a slope which corresponds to the electric field gradient. This is illustrated in figure 2.4, which shows the same valence and higher energy bands which were shown in equilibrium in figure 2.3.

Electrons are always in motion, and those in the valence band of a crystal are continually "sharing themselves around" among all the atoms of the lattice. The effect of the applied electric field, as one might expect, is to produce a tendency for the electrons to be accelerated in the "downhill" direction of the field, and slowed down or decelerated in the "uphill" direction.

Now deceleration of electrons by the field is in fact difficult, because this implies loss in kinetic energy, and falling of the electrons concerned to lower levels; yet the lower levels are filled. However, the converse process of electron acceleration is quite easy, because this involves the transfer of electrons to higher energy levels, and such levels are in this case readily available in the form of the remaining empty upper levels of the partly filled valence band.

Acceleration of electrons thus occurs readily under the influence of the field, and there is the "nett flow of charge from one end of the crystal to the other" which we define as an electric current. In moving through the crystal the electrons exchange negative potential energy for kinetic energy, jumping from level to level and effectively moving along the crystal energy diagram along paths such as that shown in figure 2.4.

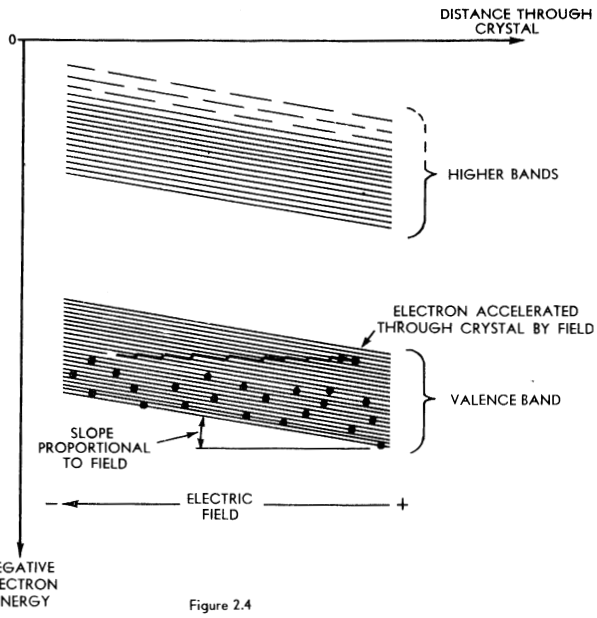


Figure 2.4

A solid material can thus be defined as an electrical **conductor** if its energy band situation in the vicinity of the valence band corresponds to that shown in figure 2.3. In other words, it is one in which the valence band is only partly filled with electrons. This is the situation which applies in the case of metallic conductors such as copper, gold, silver and aluminium.

The second type of energy band situation which can occur in the vicinity of the valence band of crystals in the ground state is that illustrated in figure 2.5. It may be seen that the only essential difference between this situation and that for a good conductor shown in figure 2.3 is that the valence band is here **completely filled.** The only energy levels of the crystal unoccupied in the ground state are thus those in the higher bands, separated from those of the valence band by a relatively wide "forbidden energy gap."

It may seem surprising, but a crystalline solid in which this energy band situation occurs is completely unable to conduct electricity when in the ground state. This is because a nett electron flow from one region of the crystal to another is impossible in any energy level which is completely filled with electrons, and in this case all the levels of the valence band are so filled.

The reason why a nett electron flow cannot occur in a completely filled energy level is that, for a nett flow to occur, there must be set up either an electron density or an electron velocity unbalance between one "end" of the level and the other. In a completely filled level a density unbalance is fairly obviously impossible; but so too is a velocity unbalance, because by definition all electrons in a given level have the same kinetic energy.

It may help in understanding this point if one imagines a filled level as something like a highway capable of carrying only a single lane of cars in each direction, and on which all the cars must travel at a fixed speed (corresponding to the particular energy level). If the highway is "filled" with both lanes carrying cars moving "bumper to bumper," there is no way in which more cars can travel in one direction than in the other; in other words, there can be no "nett car flow" in either direction.

The only ways in which a nett flow could occur would be either if the lanes of the highway were not filled, so that more cars could conceivably travel in one direction than in the other (a density unbalance), or if cars could travel at different speeds (a velocity unbalance), the latter implying the availability of additional "energy level" lanes.

From the foregoing it may be seen that if the valence band of a crystalline solid is completely filled, the crystal concerned will be an electrical **insulator.** ALL crystals whose energy band situation in the vicinity of the valence band corresponds to that shown in figure 2.5 in the ground state are thus strictly insulators in that (hypothetical) state.

Into this category fall both those materials normally known as "insulators" and those which have relatively recently become known as "semiconductors," as noted earlier. In fact, there is **no essential difference** between these

two groups of materials, only a difference in the degree to which their behaviour changes with excitation level. To clarify this point, let us now look at the effect of excitation on the basic situation shown in figure 2.5.

We have seen previously that an individual atom would only be in its ground state if it could be maintained at a temperature of absolute zero (−273 deg. C), shielded against all forms of radiant energy such as heat and light; whereas in actual fact, an atom in a practical environment is taking part in a continual process of energy absorption and emission, involving the transfer of electrons between its various allowed energy levels. As one might expect, the same argument applies to a crystal lattice composed of a large number of such atoms.

A crystalline solid in a practical environment involving heat, light and other radiant energy is therefore

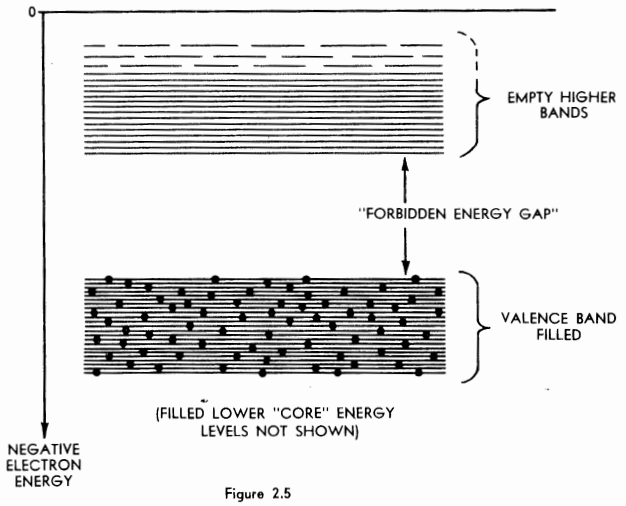


Figure 2.5

similarly involved in a continuous process of absorption and emission, with electrons now transferring both between levels within the crystalline energy bands, and also between the bands. The latter naturally involves absorption or emission of larger energy quanta than the former, as it involves transfer across the relatively large forbidden energy gaps between bands.

Under such conditions the "insulator" energy band situation shown in figure 2.5 will change. Absorption and emission of energy by the crystal lattice will reach a dynamic balance or equilibrium at an excitation level above the ground state, in which a small proportion of ever-changing electrons from the valence band have been transferred to higher energy bands. This is illustrated in figure 2.6.

The extent to which this will occur depends both upon the energy level of the environment in which the crystal finds itself, and also upon the width of the forbidden energy gap between the valence and next higher energy band. Naturally enough, the higher the temperature of the heat energy present in the crystal, the "bluer" the light incident on its surface, and so on, the greater will be the tendency of valence band electrons to acquire the energy necessary for them to be transferred to higher bands; but this granted, the proportion which do actually transfer will depend upon the energy width of the forbidden gap.

The width of the forbidden energy gap varies widely among the crystalline solids whose valence band situation is represented by figures 2.5 and 2.6. Accordingly, such materials also vary widely in the degree to which electrons are transferred from the valence to higher bands under the influence of excitation. And as we shall see shortly, this behaviour determines directly their electrical characteristics.

In a crystal of diamond, the binding between the constituent carbon atoms is such that the forbidden energy gap is very wide. It amounts to some 6 electron-volts (eV), where an electron-volt is a convenient unit of energy used in atomic physics and other fields; one electron volt is the potential energy acquired by an electron when it is moved through an electric field for a distance corresponding to an increase of one volt.

In comparison, the forbidden energy gap of pure germanium crystal is only 0.72eV, while that of pure silicon is only a little larger at 1.11eV. It may be seen from this that in such materials the proportion of electrons which have transferred from the valence band to higher bands will be much greater, at a given degree of excitation, than for a material such as diamond.

Because under "normal" practical conditions crystals of materials like germanium and silicon have significant numbers of electrons which have transferred from the valence band to higher bands, whereas crystals of materials like diamond have not, and because this results in significant differences in the normal-conditions electrical behaviour of the two groups of materials, it has become convenient to distinguish between them. Crystals of germanium and silicon are thus known as **semiconductors,** while those of materials such as diamond are known as **insulators.**

It should perhaps be stressed again that there is no distinct division between the two groups of materials; as we have seen, in the ground state both have the "insulator" energy band situation shown in figure 2.5, and both types of material behave as shown in figure 2.6 with excitation. The only difference is one of degree.

Hence by raising the temperature of an "insulator" crystal, for example, one could obtain a semiconductor, while conversely by cooling a "semiconductor" one produces an insulator.

From our earlier look at conduction in metallic crystals, the reader may by now have deduced that a semiconductor crystal in the excited state shown in figure 2.6 will become quite a good conductor, by virtue of the electrons which have transferred from the originally full valence band into the originally empty higher bands. And this is quite so, although it is only half the story.

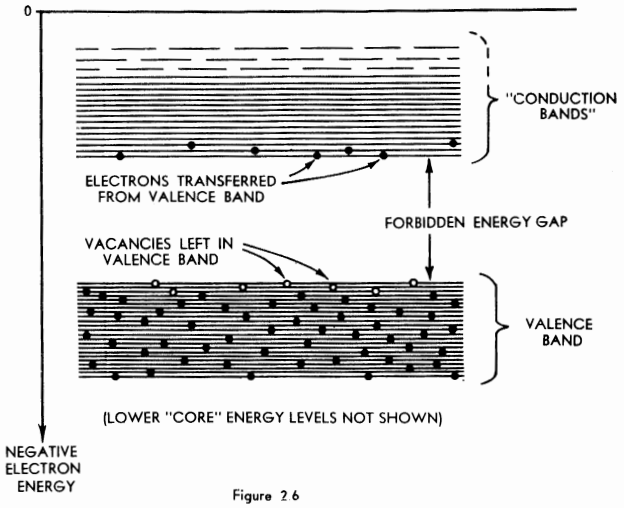


Figure 2.6

The electrons which have transferred into the higher bands, because these bands are largely empty, are certainly capable of forming a nett carrier flow through the crystal under the influence of an applied electric field. In fact because of this, the higher bands are usually called the **conduction bands,** as shown in figure 2.6. However, as it happens, the "vacancies" which are left by transferred electrons in the valence band are also able to contribute to conduction.

In order to understand this, consider that when an electron is transferred from the valence band to a conduction band, this is actually equivalent to the weakening of a valence electron bond between two adjacent nuclei of the crystal lattice. Instead of the usual two-electron "covalent" bond which each nucleus shares with each of its four adjacent nuclei, there is left in the place concerned a weakened bond having only a single electron. This is illustrated in the two-dimensional picture of figure 2.7, where the weakened bond is shown consisting of the single remaining electron together with a **hole** or vacancy in place of the missing electron.

Because of the missing valence electron, the electrical charge balance of the crystal lattice is upset in the vicinity of the weakened bond. The positive charges of the relatively fixed atomic nuclei are no longer exactly balanced by the negative charges of the surrounding electron population, so that a localised nett positive charge is produced.

In fact this positive charge is localised right in the "hole" originally occupied by the missing electron, and it has a value of charge equal and opposite to the negative charge of an electron. Neither of these facts are really surprising in view of the way in which the charge is produced.

The interesting thing is that such a hole is capable of moving through the crystal lattice, and as a moving positive charge it can thus effectively make a contribution to a current flow which is almost equal (but opposite) to that of an electron.

A hole tends to move through the crystal lattice because electrons in neighbouring valence bonds are attracted by its positive charge; when such an adjacent electron jumps across to "fill" the hole, it in turn leaves a hole in its own original bond to be filled by another electron, and so on. This "leapfrog" effect results in the effective movement of the hole through the lattice. Under the influence of an applied electric field, the hole movement will tend to take place in the direction opposite to that taken by a conduction band electron.

It may perhaps seem from this description that the concept of a hole is a redundant one, for the reason that "hole movement" in a particular direction through a crystal might seem to be "really nothing more" than a series of small jumps by electrons in the opposite direction; but this is not so. The fact is that the localised positive charge present in a crystal lattice at a weakened valence bond is no more and no less a reality than the "localised negative charge" which we are pleased to call an electron. It even has an effective mass, which is approximately equal to that of an electron.

To use an analogy, a hole in a crystal lattice valence band is rather like an air bubble in a test-tube almost filled with water. Both might be interpreted merely as "vacancies" whose effective movement takes place purely by means of movement in the opposite direction of something which superficially seems more "real" — like electrons, or water. Yet like the air bubble, a hole makes its existence apparent by means of its behaviour, in this case its electrical behaviour.

In a semiconductor crystal of the type whose valence band situation is shown in figures 2.5 and 2.6, then, for every electron which is transferred to the conduction bands and accordingly becomes available as a "negative current carrier," there is also produced a hole **which remains in the valence band** but is equally available as a "positive current carrier."

Because of this, it is usual to say that excitation of a semiconductor crystal lattice results in the production of **electron-hole carrier pairs.** Similarly the emission of energy by the lattice is visualised as a process whereby a wandering electron in the conduction band "accidentally" meets a hole wandering in the valence band, the two permanently cancelling or "annihilating" one another and producing a photon of appropriate energy. The latter process is usually termed **recombination.**

A pure or **intrinsic** semiconductor material such as we have been considering thus contains, in the excited state, equal numbers of conduction band electrons and valence band holes available for electrical conduction. However, the two types of carrier do not contribute to current flow in an exactly equal manner, because holes are in the valence band and cannot move through the material at the same rate as conduction band electrons. In effect, whereas the electrons in the conduction band can move speedily through the lattice without having to conform to any orbit requirements, the holes in the valence band must "weave" their way through the crystal binding orbit system, and therefore travel at a slower rate.

This means that while the numbers of free electrons and holes present in an excited intrinsic semiconductor at any one time are equal, any nett current flowing through the material is carried more by the faster-moving conduction band electrons moving from negative to positive than by the slower-moving holes moving from positive to negative.

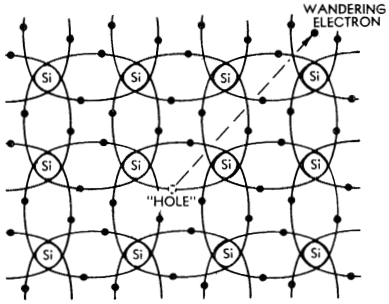


Figure 2.7

To use the analogy of a highway introduced earlier, but in a slightly different sense, the situation is now like a two-lane highway in which both lanes are packed with cars travelling in opposite directions, bumper to bumper but in this case at different speeds (corresponding to the two different energy bands). Although any given length of highway will contain equal numbers of cars in the two lanes, there will still be a greater car "flow" in the faster lane than in the slower lane.

Because the generation of electron-hole carrier pairs depends upon the excitation level of the crystal lattice, the number of such carriers available for conduction varies directly with the excitation level. Hence the **conductivity** of an intrinsic semiconductor crystal similarly varies directly with excitation. In the ground state, as we have seen, it will be zero; in more practical circumstances it will rise to a value which will depend directly upon both the temperature and the frequency/intensity characteristics of any light incident at its surface.

At this point it is perhaps worthwhile to pause briefly and note the contrast between the current picture of semiconductor - insulator conduction, which we have been examining, and earlier ones which held that these materials were merely those wherein the valence electrons were "harder for the electric field to pull free." It may be seen that the latter idea was quite wrong, because in fact such materials **cannot conduct at all** under the influence of an electric field alone; they become capable of conduction only when excited. Neither this fact nor the existence of holes as additional current carriers in these materials could be explained by the earlier theories.

TABLE 2.1

Material	Resistivity, Ohm-cM
Copper	1×10^{-6}
Bismuth	$1 \times 10^{-}$
Germanium	47
Silicon	200,000
Mica	1×10^{16}
Diamond*	3×10^{42}

*Theoretical resistivity. In fact unmeasurable.

In talking about the electrical behaviour of a semiconductor at a particular excitation level, reference is often made to the **resistivity**, which is simply the reciprocal of the conductivity. Resistivity is usually defined as the resistance in ohms between opposite faces of a cube of material measuring one centimetre on each side; this gives units of ohms/cM/square cM, or **ohm-cM**.

As the conductivity of an instrinsic semiconductor rises from zero with excitation, this means that the resistivity effectively falls from a value of infinity. Table 2.1 gives the approximate resistivity figures for pure silicon and germanium under "normal" conditions, and also gives the equivalent figures for typical metallic conductors and insulators.

The fact that the resistivity of intrinsic semiconductors falls sharply with excitation is exploited by using them in thermistors, or temperature-dependent resistors which have a negative coefficient. This is in fact the main use of intrinsic semiconductors as such, their resistivity being rather too high and too temperature-dependent for direct use in most other semiconductor devices.

SUGGESTED FURTHER READING

BURFORD, W. B., and VERNER, H. G., **Semiconductor Junctions and Devices,** 1965. McGraw-Hill Book Company, New York.

MORANT, M. J., **Introduction to Semiconductor Devices,** 1964. George G. Harrap and Company, London.

SCROGGIE, M. G., **Fundamentals of Semiconductors,** 1960. Gernsback Library Inc., New York.

SHIVE, J. N., **Physics of Solid State Electronics,** 1966. Charles E. Merrill Books Inc., Columbus, Ohio.

SMITH, R. A., **Semiconductors,** 1959. Cambridge University Press.

Chapter 3

THE EFFECTS OF IMPURITIES

Doping and impurity semiconductors—donor impurities and N-type impurity semiconductor—majority and minority carriers—doping concentration and its effects—acceptor impurities and P-type impurity semiconductor—resistivity and excitation—Fermi level and the Fermi-Dirac distribution—compensation.

As we have seen, electrical conduction cannot take place in intrinsic semiconductor materials such as pure silicon and germanium when they are in the ground state, because of the completely filled valence band. However excitation of the crystal lattice results in the production of electron-hole pairs, of which the electrons become available as negative current carriers moving in the conduction bands, and the holes become available as positive current carriers moving in the valence band.

Increasing the excitation of the lattice, by raising its temperature, for example, thus causes the conductivity of such materials to increase. Or looked at in another way, their resistivity falls. At room temperature their resistivity has typically fallen to a value which, while quite high compared with metallic conductors, is still low compared with an insulator such as diamond.

Actually semiconductors such as silicon and germanium only exhibit this so-called **intrinsic behaviour** when they are extremely pure—something like 99.9999999% pure, in fact, with any other elements present in the crystal lattice as "impurities" kept to less than one part in 10^9. Even microscopic amounts of certain impurities can radically alter their electrical behaviour, and in different ways.

From this may be judged the degree of precision which has been evolved by modern semiconductor technology, which is not only concerned initially with the production of extremely pure materials such as silicon and germanium, but also and consequently with the controlled alteration of their electrical behaviour to an accurate extent. The latter technique, which is known as **doping,** involves the addition of precise microscopic quantities of selected impurities. Typical concentrations range from a few parts in 10^9 to a few parts in 10^7.

As we shall see, the presence of impurities in a semiconductor results in the availability, under normal conditions, of many more current carriers than are available in an intrinsic semiconductor. As a result the resistivity of such an **impurity semiconductor** is typically considerably lower than that of an intrinsic semiconductor, while the influence of temperature and other forms of excitation is less pronounced —again under normal conditions. As figure 3.1 shows, the resistivity is still infinite for zero excitation (the ground state), and still drops proportional to excitation at very high levels; but at moderate excitation levels there is a "plateau" not present in the characteristic of an intrinsic semiconductor.

Although all impurities tend to alter the broad electrical behaviour of a semiconductor in this fashion, there are in fact two different and somewhat complementary mechanisms by which this can occur. Each mechanism is associated with a particular group of impurity elements, so that when used for doping the elements of the two groups produce two different "types" of impurity semiconductor material. The differences between these two types of impurity semiconductor are vital for the operation of virtually all semiconductor devices, so that we should now examine each in turn.

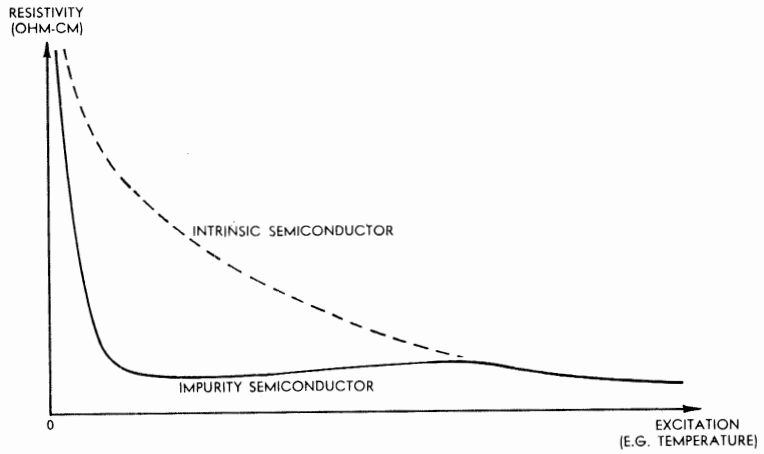


Figure 3.1

We have seen that the atoms of a silicon or germanium crystal lattice are bound together by the valence electrons, of which both silicon and germanium atoms have four. Each atom in the lattice is bound to its four neighbouring atoms by a so-called "covalent" bond, involving one valence electron of each atom in a common "shared pair" orbit. A simplified two dimensional representation of this was given previously in figure 2.7.

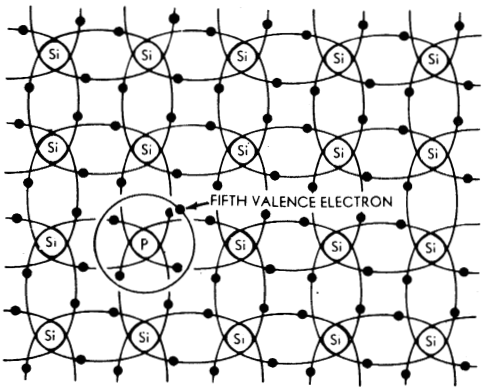
When atoms of elements such as phosphorus, arsenic, antimony or bismuth are present as impurities in such a crystal lattice, they are for the most part incorporated into the lattice structure in a simple "replacement" or substitutional manner. Four of their valence electrons are engaged in covalent bonds with the neighbouring "host" atoms, so that in this respect an impurity atom is quite equivalent to a host atom.

Of course an impurity atom cannot be fully equivalent to a host atom, because it will have both a different nucleus mass and positive charge, and a correspondingly different number of surrounding electrons. The latter is of particular importance because in the case of phosphorus, arsenic, antimony and bismuth there are in fact **five valence electrons,** one more than is present in silicon or germanium.

Because of this, when an atom of these elements is present as an impurity in a silicon or germanium crystal lattice there is one valence electron "left over" after the atom has engaged itself in covalent bonds with its neighbours. This is illustrated in figure 3.2, where the "left over" fifth electron is shown occupying an orbit surrounding its parent phosphorus nucleus in a silicon lattice.

Although the electron is shown in an orbit surrounding its parent impurity nucleus, it may be remembered that electrons at the valence and high energy levels in a pure crystalline solid tend to be the "common property" of all the nuclei in the lattice. Thus while the additional positive charge on an impurity nucleus does produce a small local "dip" in the electron energy pattern of the lattice, with a consequent tendency for the fifth valence electron to remain, this effect is in fact quite slight. Very little energy is required in order to free the electron, so that even when the lattice is only slightly excited such electrons are virtually all freely wandering around the crystal and available as negative current carriers.

Because of this effective "donation" of electrons as additional negative current carriers to the basic simiconductor lattice, impurity elements such



(NUCLEI AND CORE ELECTRONS NOT SHOWN)

Figure 3.2

as phosphorus, arsenic, antimony and bismuth are known as **donor impurities.** And because with such donor impurities present there is an excess of negative current carriers, in contrast with the equal numbers of positive and negative carriers present in an excited intrinsic semiconductor lattice, a crystal lattice which has been doped with a donor impurity element is termed an **N-type impurity semiconductor.**

The energy band diagram of such an N-type impurity semiconductor is shown in figure 3.3. It may be seen that in the ground state the fifth valence electrons of the donor impurity atoms occupy localised and relatively isolated segments of a single energy level, which is only slightly below the bottom of the lowest conduction band. The electrons occupy a single new level rather than a multi-level band because, being relatively isolated from one another, they are not subject to coupling interaction effects.

The small gap between this "donor level" and the bottom of the conduction band represents the small energy increment required to free the electrons from their ground-state orbits. It may be seen that only a slight excitation of the crystal lattice will cause most of the donor level electrons to be transferred to the conduction band levels, so that the resistivity of the material will fall rapidly with excitation to a value which is many times lower than an intrinsic semiconductor under normal conditions.

At this point the reader may perhaps be wondering whether the electrons which transfer from the donor level to the conduction band leave holes behind. The answer to this is no, because the donor level simply corresponds to the isolated "fifth valence electron" orbits shown in figure 3.2, rather than to a complete binding orbit system, and the concept of a hole has little if any meaning except with reference to a complete binding system. To extend an earlier analogy, an empty isolated orbit is somewhat like a test-tube emptied of water, in which an "air bubble" can scarcely have any meaningful existence.

When a fifth valence electron leaves its parent donor impurity atom to wander through the crystal lattice, then, it does not leave behind a hole. But this is not to say that the parent impurity atom then becomes indistinguishable from any of the host atoms; this can never occur, because it may be remembered that the nucleus and core electron system of the impurity atom will be always different from that of the neighbouring silicon or germanium atoms.

In fact, a donor impurity atom which has lost its fifth valence electron to the lattice will have a nett positive charge. This being the case it should strictly no longer be called an "atom," but given the name by convention applied to a charged particle — an **ion.** It will be a positive ion, naturally, and will be **fixed** rather than movable because of its covalent bonding with the neighbouring host semiconductor atoms.

Under moderately excited "normal" conditions, then, N-type impurity semiconductor material contains two types of localised electric charge whose presence can be attributed to the addition of impurity atoms to the crystal lattice. On the one hand are an appreciable number of electrons moving through the crystal with an energy level which places them in the conduction band, and which are therefore available as negative current carriers; while there are also an equal number of positively charged donor impurity ions which are fixed and therefore not themselves available as current carriers. We will see later on that while the fixed impurity ions cannot act as current carriers themselves, they can despite this play an important part in controlling the behaviour of the carriers.

Although the electrons "donated" by the donor impurity atoms are the main

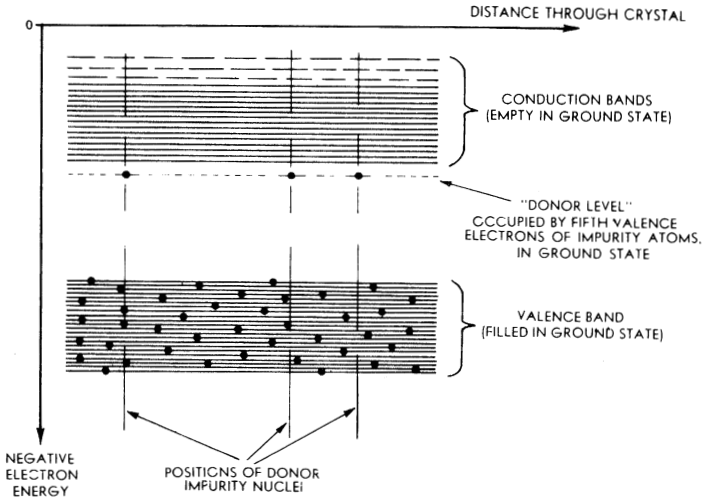


Figure 3.3

current carriers in N-type impurity semiconductor material, they are **not** the only available current carriers. The reason for this is that there will still be electron-hole carrier pairs produced by excitation of the lattice in the same fashion as in an intrinsic semiconductor.

As one might expect, a particular degree of excitation of an impurity crystal tends to produce as many electron-hole carrier pairs as in an intrinsic semiconductor crystal at the same degree of excitation. However in an impurity semiconductor the **effective** number of such carrier pairs present at any degree of excitation is considerably lower than in intrinsic material.

In the case of the N-type impurity semiconductor material which we have been considering, the reason for the reduction is that with a considerable number of donor-derived conduction electrons already wandering through the crystal lattice at the conduction band levels, there is an increased probability that wandering holes and electrons will meet to annihilate one another by recombination. Naturally such recombinations "remove" equal numbers of conduction-band electrons and valence-band holes from the crystal lattice, so that the numbers of both types of carrier effectively available in addition to the donor-derived conduction-band electrons will be somewhat smaller than the numbers of carrier pairs available in intrinsic material under the same conditions.

The total population of current carriers available in N-type impurity semiconductor material under normal conditions thus consists mainly of conduction-band electrons, with a small minority of valence band holes.

In this material electrons can thus be termed the **majority carriers,** and holes the **minority carriers.** Both these terms serve to emphasise the contrast with the equal-numbers-of-electrons-and-holes situation which applies with an intrinsic semiconductor.

As one might expect, increasing the number of donor-derived conduction band electrons in the material further reduces the effective additional 'proportions of "intrinsically produced" electron-hole pairs. And not surprisingly, the number of donor-derived electrons is in turn directly proportional to the number of donor impurity atoms originally added to the lattice — the **doping concentration.**

Hence we can say that in N-type impurity semiconductor material, the proportion of total available current carriers represented by the majority carriers — in this case electrons — is directly proportional to the doping concentration. Highly or heavily doped material can thus be considered to be "more N-type" than lightly doped material, because it will have a higher proportion of majority-carrier electrons and a lower proportion of minority-carrier holes.

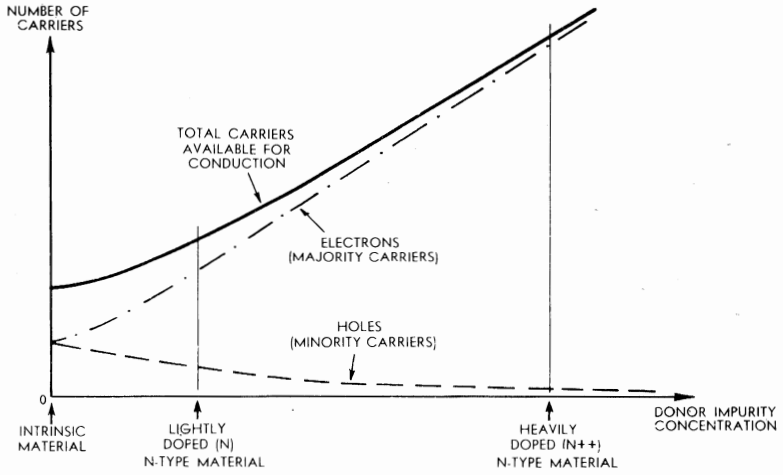


Figure 3.4

Figure 3.4 illustrates the foregoing by showing the effective numbers of electron and hole carriers which will normally be present in a crystal sample for various doping concentrations. It may be seen that for intrinsic material with zero donor impurity, there are present equal and modest numbers of electrons and holes — the "intrinsic" electron-hole pairs. With the progressive addition of donor impurity the number of electrons rises rapidly while the number of holes falls, so that while the total number of carriers available for conduction rises rapidly with donor impurity concentration, it progressively becomes composed more and more of electrons or majority carriers, and less and less of holes or minority carriers.

Having looked fairly closely at one of the two types of impurity semiconductor material, let us now examine the second type. We may well expect to find a similar but complementary set of situations involved, and this in fact turns out to be the case.

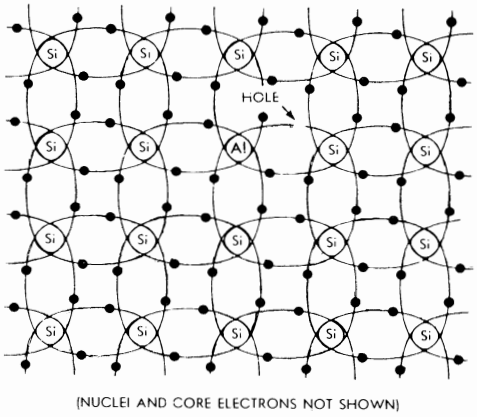
Those impurity elements which act as electron carrier donors to an intrinsic semiconductor crystal lattice, it may be remembered, are pentavalent. They have five valence electrons, in other words, one more than the four possessed by intrinsic semiconductors such as silicon and germanium. As one might perhaps expect, there also exists a second group of important impurity elements which are in contrast trivalent — possessing only three valence electrons, and hence in this case one less than silicon and germanium. Elements which fall into this group include boron, indium, aluminium and gallium.

When atoms of one of these elements are present as impurities in a semiconductor crystal, they are for the most part incorporated into the lattice in much the same "substitutional" manner that applies in the case of donor impurities. However, having only three valence electrons, they are able to enter into the required covalent bonds with only three of the neighbouring host atoms. With the remaining neighbour atom they can form only a weaker "non-contributory" bond involving a single electron.

As the illustration in figure 3.5 shows, the weakened fourth bond is of exactly the same type which we saw to be present in an intrinsic semiconductor lattice bond when an electron has been removed by excitation. In short, the impurity atom—in this case aluminium—effectively brings with it into the crystal lattice nothing other than a positively charged valence-band hole.

Although in the common-property valence electron binding system of the crystal lattice, the hole has a weak tendency to remain in the vicinity of its parent impurity nucleus. This is because of the lower positive charge or "relative negativity" of the impurity nucleus compared with the neighbouring host nuclei. However, as with the fifth valence electron of a donor impurity, the hole binding is very weak, and very little energy is required for the hole to effectively move away through the crystal in the manner which we previously examined.

This means that even for quite low levels of excitation, the holes introduced into the lattice by the impurity atoms will be found wandering through the crystal and available as positive current carriers. At the same time the impurity atoms themselves, having gained a valence electron, will have become fixed negatively charged ions.



(NUCLEI AND CORE ELECTRONS NOT SHOWN)

Figure 3.5

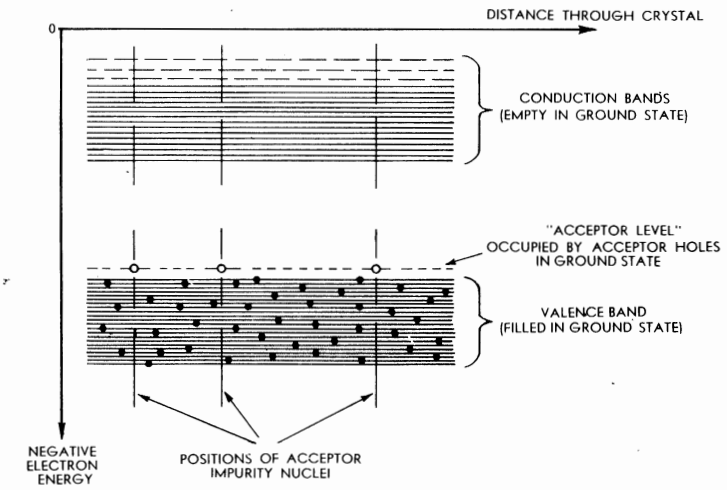


Figure 3.6

It may be seen that in contrast with the behaviour of donor impurities, the impurity atoms have in this case effectively "accepted" valence electrons from the crystal lattice. To distinguish this behaviour from that of donor impurities, elements such as boron, indium, aluminium and gallium are known as **acceptor impurity elements.** And be-

cause with an acceptor impurity present a semiconductor crystal has an excess of positive current carriers under normal conditions, compared with intrinsic material, a crystal which has been doped with an acceptor impurity is termed a **P-type impurity semiconductor.**

The energy band diagram of a P-type impurity semiconductor is shown in figure 3.6, and the reader may care to compare it with that for N-type material shown in figure 3.3. It may be seen that the holes which "belong" to the acceptor impurity atoms in the ground state again occupy localised and isolated segments of a single energy level, but that in this case the impurity level is slightly above the top of the valence band.

The small gap between the "acceptor level" and the top of the valence band represents the small energy increment required for electrons in the valence band to transfer into this level, "filling" a hole but leaving behind another in the valence band itself. Only slight excitation of the lattice is therefore required for most of the acceptor level holes to be filled, leaving many holes behind in the valence band to act as positive current carriers. The resistivity of P-type material thus falls rapidly with excitation in almost exactly the same fashion as with N-type material, and like the latter it has, under normal conditions, a resistivity many times lower than intrinsic semiconductor.

Just as the donated electrons are not the only carriers present in N-type impurity semiconductor, so the holes derived from acceptor atoms are similarly not the only carriers present in P-type material. As before there will be "intrinsic" electron-hole pairs produced by the normal excitation mechanism, although again the effective numbers of these carriers is lower than in intrinsic material.

The reason for the reduction is again carrier loss by recombination, due in this case to the relatively large number of holes moving through the crystal lattice at valence band level. As before this means that the numbers of both types of "intrinsic" carrier effectively fall with increasing doping concentration.

Accordingly the effects of doping concentration for a P-type impurity semiconductor may be represented as in figure 3.7, which is a similar diagram to that of figure 3.4 for an N-type semiconductor. Here the holes are the majority carriers and the electrons are the minority carriers, but otherwise the relationships are very similar. As before the total number of carriers available for conduction rises rapidly with impurity concentration, and progressively becomes composed more and more of the majority carriers and less and less of the minority carriers.

It may be seen from the foregoing descriptions of the two types of impurity semiconductor that both types have available under normal conditions

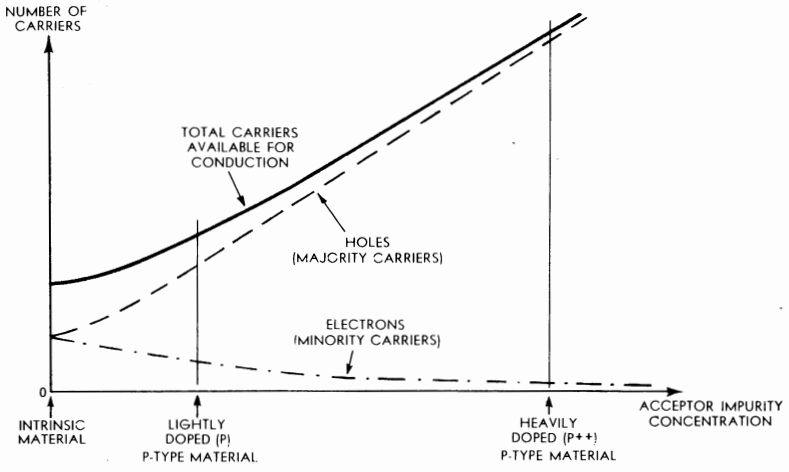


Figure 3.7

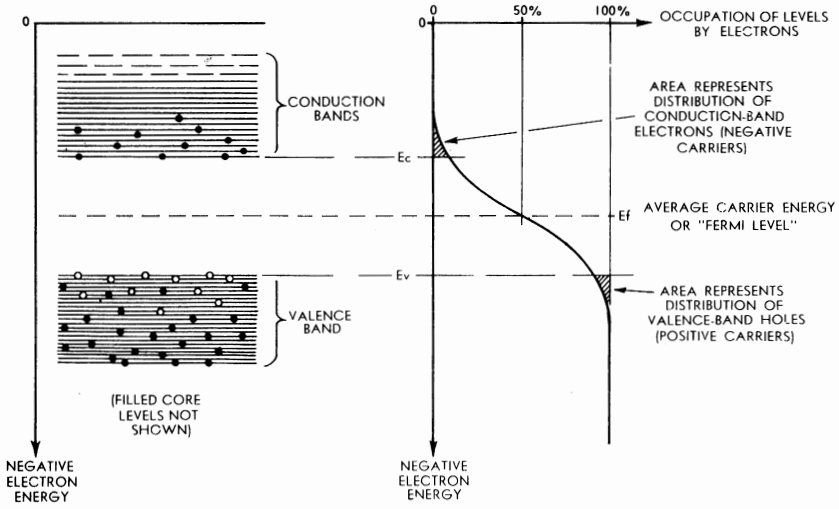


Figure 3.8

of excitation considerably greater numbers of current carriers than are available in intrinsic semiconductor material. The numbers are of different composition in each case, to be sure, but the total numbers are in both cases greater—by an amount proportional to the concentration of the appropriate doping impurity.

With applied excitation the resistivity of both types of impurity semiconductor thus tends to fall much more rapidly than with intrinsic material, and this explains the steeper initial slope of the solid curve given earlier in figure 3.1. However, increasing excitation rapidly results in the situation where virtually all the electron or hole carriers derived from the impurity are available for conduction; at this point the resistivity tends to flatten out.

Further increase in excitation tends to produce little if any reduction in resistivity, because the tendency for increased numbers of electron-hole pairs to be produced is largely balanced by a corresponding increase in recombination. In fact the resistivity of the material tends to increase slightly, because with increasing activity within the crystal lattice the motion of the carriers becomes impeded by an increasing number of "collisions." This **reduction in carrier mobility** explains the slight upward slope of the plateau in figure 3.1.

If the increase in excitation is continued still further, a point is eventually reached where the production of "intrinsic" electron-hole carrier pairs simply swamps the recombination mechanism. When this happens the majority-minority carrier situation gives way to the equal numbers situation, while resistivity again begins to fall. Thus in effect both N-type and P-type impurity semiconductor materials revert back to "intrinsic" semiconductor at very high excitation levels.

From the foregoing it may be seen that both the total number of carriers available in a semiconductor, and the proportions of negative and positive carriers making up that number are determined by three factors. These are the presence and concentration of any impurities present, the type of impurity and the degree of excitation.

It has been found of considerable value to describe this rather complex situation using two very useful concepts: that of an "average carrier energy level," and that of a statistical "spread" or distribution of the carriers above and below the average level. As with some of the concepts introduced earlier, a full understanding of these concepts requires considerable background in quantum mechanics and is thus beyond the present discussion. However, the basic ideas involved are not unduly difficult, and can help considerably in understanding practical semiconductor device operation.

As we have seen, conduction in semiconductor materials takes place by movement through the crystalline lat-

tice of two types of carrier—negative carriers which consists of electrons possessing an energy which places them in in the conduction band, and positive carriers which consists of hole possessing an energy which places them in the valence band. Because of this, the most useful measure of the excitation level of the material from an electrical viewpoint is one which takes both types of carrier into account, in terms of both numbers and energy distribution. We may thus talk meaningfully of an "average carrier energy level" of a semiconductor crystal, representing the average of the energy levels of all the carriers available in the crystal lattice.

In the case of an intrinsic semiconductor it may be recalled that for any degree of excitation the number of conduction band electrons and valence band holes are equal. Hence the average carrier energy level for such material will be exactly midway between the valence and conduction bands. This is illustrated in figure 3.8, where the average carrier energy level is given its more usual name of **Fermi level** (in honour of the physicist Enrico Fermi), and labelled Ef.

It has been found that the distribution of carriers in the various energy

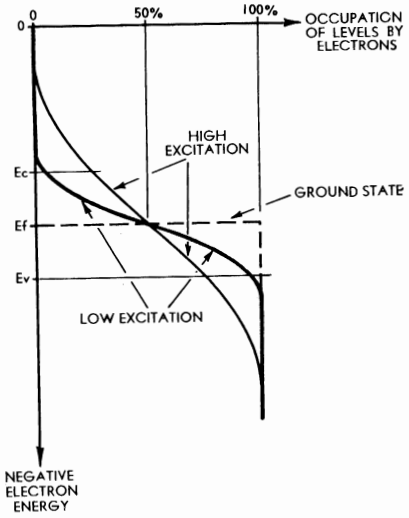


Figure 3.9

bands above and below the Fermi level can be described quite accurately by the type of curve shown. The shape of the curve corresponds to what mathematicians call the **Fermi-Dirac distribution.**

As may be seen from figure 3.8—which, it should be remembered, corresponds to an intrinsic semiconductor only — the curve represents a plot of the relative occupation by electrons of any allowed energy level, expressed as a fraction or percentage of the level capacity. Hence the curve has a value of 100 per cent for the lower filled levels, then slopes over to a value of 0 per cent for the uppermost empty levels.

Note that the continuous nature of the curve is not intended to imply that electrons are occupying levels other than the allowed levels of the various bands. Hence the portion of the curve between level Ec, marking the bottom of the conduction band, and Ev, marking the top of the valence band, is essentially a theoretical interpolation or "fill in." It is arranged so that the curve is symmetrical above and below the Fermi level Ef, with the intersection at Ef corresponding to the theoretical point of 50 per cent level occupation.

In figure 3.8 the small cross-hatched area above the level Ec represents the distribution of electrons in the conduction band — i.e., the number and distribution of negative carriers. Similarly the lower small cross-hatched area below level Ev represents the distribution of electron vacancies or holes in the valence band levels — i.e., the number and distribution of positive carriers. Note that the two areas are equal, and equal in shape.

The shape of the Fermi-Dirac curve changes to describe the way in which the number of carriers available in the material varies with excitation. Its shape as shown in figure 3.8 corresponds to a moderate degree of excitation, where the "tails" of the curve indicate a modest number of each type of carrier.

In figure 3.9 is shown the way in the shape of the curve varies

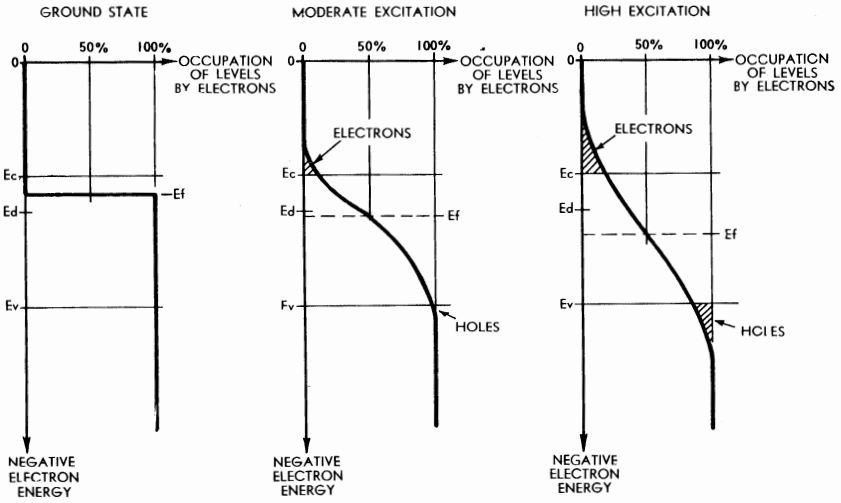


Figure 3.10

with excitation. For the ground state or zero-excitation case, it is not a curve at all, but a sudden "step;" as excitation increases the "corners" of the step round off, producing longer and longer "tails." It may be seen that this results in larger and larger areas above Ec and below Ev, corresponding to the increased numbers of carriers available with increasing excitation.

It should be noted that both figures 3.8 and 3.9 are drawn for intrinsic material, in which as we have seen the Fermi level is fixed and exactly midway between Ec and Ev. Naturally this same situation cannot be true with either of the two types of impurity semiconductor, because in these cases there are not only unequal numbers of negative and positive carriers, but the ratio between the two varies with excitation.

Actually it turns out that the Fermi level of each of the two types of impurity semiconductor is different, and also that it varies both with the type of impurity and the excitation. However although this is the case, the new and changing distributions of carriers are still described by the Fermi-Dirac distribution curve, providing its 50 per cent point is kept in alignment with the Fermi level.

Figure 3.10 shows the Fermi level positions and carrier distributions for N-type impurity semiconductor at three degrees of excitation. The energy band structure of the material is not shown, but as before Ec and Ev represent the bottom of the conduction band and the top of the valence band respectively. The donor level is represented by Ed.

As may be seen, in the ground state the Fermi-Dirac curve is again a step curve with the "step" at the Fermi level. But the latter is now at a somewhat higher level than in the case of intrinsic material. Its position will naturally vary with the donor impurity doping concentration, to take account of the changing carrier ratio illustrated in figure 3.4; thus the position between Ed and Ec shown in figure 3.10 will correspond to a quite heavily doped N-type material. With lower doping concentrations Ef will be lower down than this, although it will always be higher than the forbidden-gap-midpoint position — which as we have seen corresponds to intrinsic material.

With moderate excitation, illustrated in the centre diagram of figure 3.10, two things have happened. Probably the most obvious thing is that the carrier distribution curve has developed "tails," as before, and that because the Fermi level is higher than the forbidden gap midpoint, the curve tails indicate the expected majority/minority carrier unbalance. But the more subtle thing that has occurred is that the Fermi level Ef has started to fall, slightly but perceptibly, to correspond to the effect of "intrinsic" (balanced) carrier generation.

The third diagram of figure 3.10 shows what happens at a very high degree of excitation. The Fermi-Dirac curve has spread well out, as before, while at the same time the Fermi level itself has fallen almost to the forbidden gap midpoint. Hence while there are large numbers of carriers, it can be seen that they are now made up of

almost equal numbers of electrons and holes — showing that the material has almost completely reverted to an effective "intrinsic" semiconductor.

In figure 3.11 are shown equivalent diagrams for a P-type impurity semiconductor, and it may be seen that the situation is here very similar. The only difference is that the Fermi level in this case occupies in the ground state a position somewhat lower than the forbidden gap midpoint, and moves up with excitation. As before its ground-state position is determined by the doping concentration; the position shown between the acceptor level Ea and the top of the valence band Ev corresponds to a quite heavily doped P-type material.

another **when they are present in small quantities.** Due to the compensation effect, the effective type and impurity concentration of a practical semiconductor material is really the resultant or net effect of whatever types of impurity are present in the lattice.

Hence in practice an N-type impurity semiconductor is one in which a donor impurity is present in greater proportion than any other impurities, and a "heavily doped" N-type material is one in which this dominance is even greater. Similarly P-type material is material in which an acceptor impurity is dominant, again to a degree which determines the effective doping concentration.

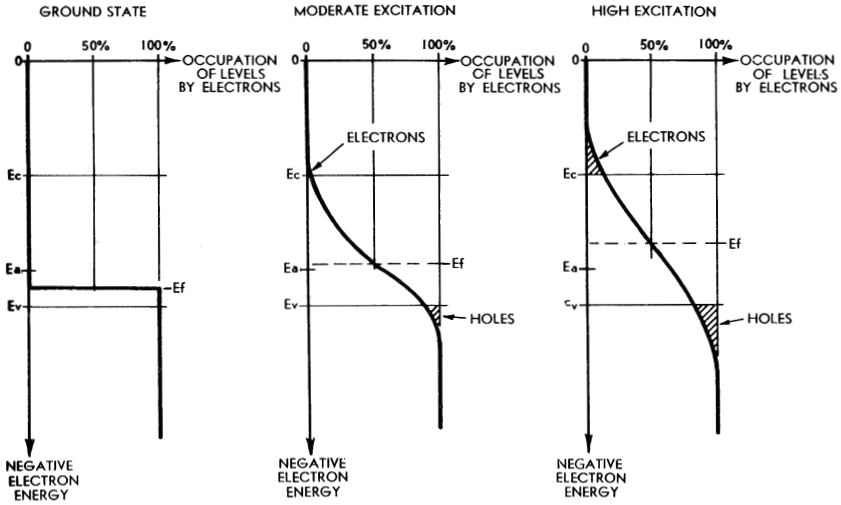


Figure 3.11

Thus far, in considering impurity semiconductor materials we have assumed that only one type of impurity is present. Although modern semiconductor technology can approximate this situation, this is all that can be done. In practice, a number of different impurity elements are almost always present, in electrically significant amounts. The reader may therefore well wonder what effect such "spurious" impurities have on the concepts which we have looked at in the foregoing.

The answer to this is that there occurs an effect called **compensation** whereby opposite types of impurity element tend to "cancel out" one

The same argument applies in the case of "intrinsic" semiconductor material. If a material has equal and minute amounts of opposite types of impurity, mutual compensation cancels out their effect so that in practice the behaviour of the material is indistinguishable from a perfect intrinsic semiconductor. The success of modern semiconductor technology in producing "pure" samples of intrinsic semiconductors such as silicon and germanium is therefore not due solely to reduction of impurity levels, but also to the development of ways of ensuring that the inevitable residuals of impurities compensate one another to a highly accurate degree.

SUGGESTED FURTHER READING

BURFORD, W. B., and VERNER, H. G., **Semiconductor Junctions and Devices,** 1965. McGraw-Hill Book Company, New York.

MORANT, M. J., **Introduction to Semiconductor Devices,** 1964. George G. Harrap and Company, London.

SCROGGIE, M. G., **Fundamentals of Semiconductors,** 1960. Gernsback Library, Inc., New York.

SHIVE, J. N., **Physics of Solid State Electronics,** 1966. Charles E. Merrill Books, Inc., Columbus, Ohio.

SMITH, R. A., **Semiconductors,** 1950. Cambridge University Press.

Chapter 4

THE P-N JUNCTION

Non-homogeneous semiconductors—carrier diffusion—"inbuilt" electric fields—drift currents—equilibrium and the Fermi level—the P-N junction—equilibrium, forward and reverse bias conditions—depletion layer width—junction "breakdown"—the semiconductor diode.

In our examination of semiconductor materials in the foregoing chapters, we have, for simplicity, looked only at the properties and behaviour of what might be called "homogeneous" samples—lumps of crystalline material in which the composition is uniform throughout. Thus we have considered, separately, uniform lumps of intrinsic semiconductor and of both N-type and P-type impurity semiconductor. In each case, by considering only a simple homogeneous sample of the material concerned, we have been able to isolate and examine its "basic" properties.

As the reader might suspect, however, such homogeneous samples of semiconductor materials are in fact mainly of academic interest. A large majority of practical solid-state devices depend for their operation upon further interesting properties and aspects of behaviour which arise in the more complex type of situation wherein the semiconductor crystal concerned is not homogeneous, but effectively composed of regions of different types of semiconductor material.

In order that the reader might gain a clear understanding of the operation of practical solid-state devices, it is therefore necessary that the basic concepts of semiconductor properties and behaviour developed earlier are expanded to cover the additional properties and behaviour of non-homogeneous samples.

With this aim in view, the present chapter will introduce and discuss some of the further basic concepts which apply to non-homogeneous semiconductor samples in general, and will then deal at some length with the extremely important "special case" of the P-N junction. Later chapters will show and explain how such P-N junctions, singly or in combination, and in one or another of a variety of physical forms and configurations, form the basis for almost every type of practical solid-state device.

To begin, then. Probably the most basic situation involving a non-homogeneous semiconductor sample, from the theoretical point of view, is a lump of impurity semiconductor crystal in which the impurity doping has not been made uniformly, but rather in a gradually increasing manner from one end of the specimen to the other. This situation is represented in simplified form in the upper diagram of figure 4.1, which shows a crystal of N-type material, whose donor impurity concentration has been arranged to increase from a low value at one end to a considerably higher value at the other.

We have seen in the preceding chapter that each donor impurity atom in a semiconductor crystal lattice effectively "splits," with excitation, into two parts—each of which plays a different part in determining the electrical behaviour of the crystal. The fifth or "excess" valence electron constitutes one part, leaving to wander through the lattice as a potential current carrier; the remainder of the donor atom is naturally fixed in the lattice, but having lost one of its original complement of electrons, it becomes a fixed positively charged ion.

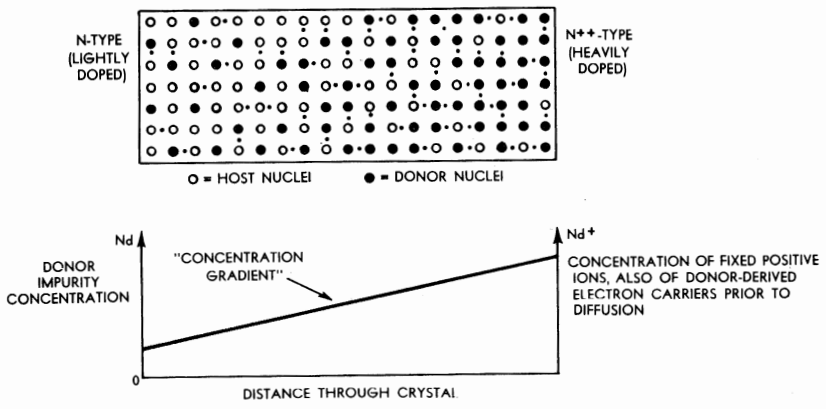


Figure 4.1

Just as the donor impurity concentration therefore quite naturally determines the number of donor-derived electron carriers and fixed positive ions in an excited **homogeneous** sample of N-type semiconductor, it similarly also tends to determine the number of these particles present at any point in an excited non-homogeneous sample. Thus, in a lightly doped region of such a sample, there will tend to be relatively few donor-derived electron carriers and fixed positive ions, while in a heavily doped region there will tend to be a considerably larger number of both.

Hence the donor impurity "concentration gradient" of the sample in figure 4.1 **tends** to result in identical gradients for both donor-derived electron carriers and fixed positive ions. This is shown in the lower diagram of the figure.

As a result of the impurity concentration gradient, one might therefore expect to find in such a sample, when it is excited, a gradual increase in the number of fixed positive ions from one end to the other, matched by an exactly equal gradual increase in the number of negatively charged donor-derived electron charges. The charges of the two types of particle would therefore cancel in every part of the crystal sample, and, as the only other effective charges present would be "intrinsic" electron-hole carrier pairs, the sample would be electrically neutral thoughout its length.

If the distribution of donor-derived electron carriers in such a sample was determined only by the impurity concentration, as it is for a uniformly excited homogeneous sample, this satisfying picture would indeed represent the situation. However, the impurity concentration is not the only factor which applies for non-homogeneous material, so that in actual fact the situation is a little more complex.

It may be remembered that in an excited semiconductor crystal lattice, electron and hole carriers do not remain fixed, but move around "at random" as a result of acquired excitation energy. In so doing, they act in a very similar fashion to gas molecules in a container at room temperature. And it happens that, just as this type of motion tends to result in the uniform

diffusion or "spreading out" of gas molecules throughout a container, a similar diffusion of both electron and hole carriers tends to occur in any excited semiconductor sample.

This **diffusion effect** occurs in all excited crystalline lattices, although in the case of a homogeneous semiconductor it cannot be detected if the material is uniformly excited. The reason for this is that in such a case the excitation itself produces both carriers and fixed ions which are uniformly distributed throughout the sample. The effect of diffusion can be made apparent in a homogeneous semiconductor only if the excitation is applied in a non-uniform manner.

For example: if one end of a bar of uniformly heavily doped P-type impurity semiconductor is heated, while the remainder of the bar is kept at a low temperature, it will be found that the heated end acquires a negative electric charge with respect to the rest of the bar. This occurs because, while the localised excitation at the heated end produces equal numbers of positive hole carriers and fixed negative acceptor impurity ions, the positive hole carriers tend to diffuse throughout the bar while the acceptor ions remain fixed at the heated end. The heated region thus acquires a net negative charge due to excess ions, while the remainder acquires a positive charge due to excess holes.

In an excited semiconductor lattice, then, both electron and hole carriers tend to diffuse themselves throughout a sample. Hence if, for one reason or another, a localised concentration of carriers tends to be produced in some part of a sample, there will accordingly be a tendency for such a concentration to diffuse away. This will occur irrespective of whether the localised carrier concentration is due to localised excitation, as in the case of our heated bar example, or due to a localised impurity concentration as in the non-homogeneous sample of figure 4.1, or due to any other possible cause.

Further, and most importantly, the tendency for a concentration of carriers to diffuse away and spread evenly through a sample is in itself quite independent of any electric field or fields which may be acting through the material, being dependent only upon the excitation level and the degree of carrier concentration. The presence of electric fields can only influence diffusion indirectly, by modifying energy levels in the material in a way which determines the energy necessary for carriers to participate in diffusion in any particular direction.

Because electron and hole carriers are electrically charged, their motion through the crystal lattice constitutes a current regardless of its cause. Hence the motion of carriers due to the diffusion effect may be quite accurately described as **diffusion currents.**

In a uniformly excited homogeneous semiconductor sample, there will fairly obviously be no net diffusion current as all carrier movements will on the average cancel. However, in the previous example of a P-type rod heated at one end there is, in contrast, a net diffusion current of holes from the heated end.

From the foregoing, it may be expected that in our graded-doped specimen of figure 4.1 any tendency for a concentration of electron carriers to be produced at the heavily doped end as a result of the larger numbers of donor impurities will be opposed by an electron diffusion current toward the lightly doped end. And this is, in fact, exactly what happens.

However, as in the case of the heated bar, the effect of the diffusion current is to upset the electrical neutrality of the specimen. In this case, the diffusion of electrons away from the heavily doped end leaves an excess of positively charged donor ions, while at the same time producing an excess of negatively charged electron carriers at the lightly doped end. The heavily doped end of the specimen thus becomes **positively** charged, while the lightly doped end becomes **negatively** charged. A potential difference is thus set up between the ends of the specimen and an electric field appears.

It should perhaps be noted that the potential difference set up in the specimen has exactly the **opposite** polarity of that which one might intuitively predict from the fact that the heavily doped end has been given a larger proportion of electron DONOR impurity. Surprisingly, perhaps, it is this end which acquires the positive charge!

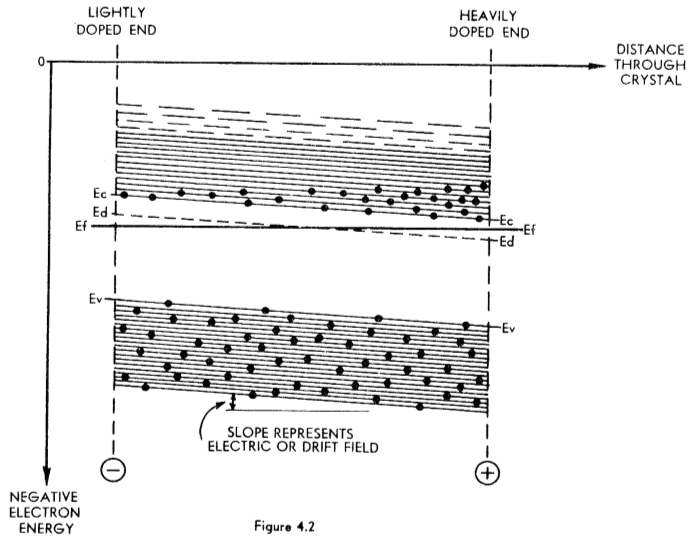


Figure 4.2

In an example of graded doping such as that of figure 4.1, therefore, the combined effect of the impurity concentration gradient and the carrier diffusion current is to set up in the material an "inbuilt" electric field, acting in the same direction as the concentration gradient.

We have seen in an earlier chapter that the effect of an electric field acting through a semiconductor lattice is to cause any available current carriers to be accelerated in the appropriate direction. Naturally, this will be the effect of the "inbuilt" field set up in our specimen.

Hence, there will be a tendency for the very electrons which diffused away from the heavily doped end of the material, setting up the electric field, to drift back again under its influence.

Note that the term "drift" has been used here to describe the effect of the electric field on the carriers, suggesting a relatively modest influence. This is quite intentional, because, in fact, although the acceleration produced by practical electric fields acting through semiconductor crystals at normal levels of excitation may be quite significant, it always remains quite small relative to the acceleration due to excitation energy. It is because of this that the motion of carriers through a crystal lattice due to an electric field is usually described as a **drift current.**

From the foregoing, it may be seen that when the specimen of figure 4.1 is excited, the electron carriers present in the material are subjected to two opposing tendencies. Firstly, there is the tendency to diffuse uniformly throughout the specimen, which in this case means to **diffuse** away from the heavily doped end. And, secondly, there is the opposing tendency to **drift** back in the direction of the heavily doped end as a result of the charge unbalance and electric field set up by the diffusion.

What does this mean? Simply that the specimen will reach an **equilibrium,** in which an electron diffusion current from the heavily doped end to the lightly doped end is balanced by an equal electron drift current in the opposite direction. And as part of this equilibrium there will be a potential difference between the ends of the material and, accordingly, an electric field through it.

In saying that the specimen reaches equilibrium, it is not implied that when this occurs all current in the specimen ceases. This cannot occur, for the simple reason that the very conditions which would result in cessation of the diffusion current are those which would result in maximum drift current, and vice-versa. For zero diffusion current the carrier concentration would have to be constant throughout, giving a maximum charge unbalance and hence maximum drift current; conversely, for zero drift current the carrier-fixed ion charges would have to be balanced throughout, giving a maximum carrier concentration gradient and therefore maximum diffusion current.

By its very nature, then, the equilibrium must be and is a dynamic one. Both the diffusion and drift currents continue to flow indefinitely in the specimen, although as their magnitudes are equal and their directions opposite, they have no measurable net resultant. Their continued presence in the specimen can only be inferred by the measurable potential difference set up between the ends of the specimen as part of the equilibrium process. The

magnitude of the potential difference will naturally depend upon the semiconductor involved and the doping gradient present; for P-type or N-type silicon it could amount to as much as 500 millivolts.

Perhaps it should be noted in passing that while the potential difference generated "inside" such a semiconductor specimen is measurable, it can only be measured using extremely sensitive equipment such as an electrometer. The reason for this is that the equilibrium mechanism involved cannot supply significant power to any "external" circuitry without itself being disturbed.

The example of figure 4.1 illustrates what has been found to be a most important general principle, one which applies to all cases involving non-homogeneous semiconductors. This is that wherever there exists a gradient of doping concentration, an inbuilt electric or "drift" field is set up along that gradient.

Further important aspects of the principle may be appreciated by referring to the energy band picture for such a non-homogeneous semiconductor. The relevant part of the energy band diagram for the graded-doped specimen of figure 4.1 is shown in figure 4.2, and it may be seen to reveal a number of interesting points.

Perhaps the most obvious point is that the energy bands are tilted, in exactly the same way which we saw in an earlier chapter to apply when an electric field is set up through a semiconductor specimen by the application of an external potential difference. And, as in such a case, the slope of the tilting is directly proportional to the intensity of the field and the effective potential difference between the ends of the specimen.

What may not be quite so obvious is that here the slope of the bands is precisely such that the average carrier energy level—the Fermi level—remains constant throughout the material, despite the large number of conduction band electrons at the heavily doped end. This may be seen from the fact that the line Ef, representing the Fermi level, has zero slope.

Although this may seem somewhat fortuitous, it is really nothing more than the natural outcome of the dynamic equilibrium which we have just seen to be set up in the material due to a balancing of the opposing effects of diffusion and drift. As we have noted, the equilibrium occurs when diffusion current of electron carriers in one direction is balanced by an exactly equal and opposite drift current in the other direction; this implies that there is then no **net** carrier flow in either direction, and consequently that the average carrier energy is constant throughout.

It is found that all non-homogeneous semiconductors, in equilibrium, conform to this pattern. In other words, the electric or drift fields which are set up inside such materials as a result of impurity concentration gradients are always such that the Fermi level—the average carrier energy level—remains constant throughout the material.

Looked at conversely, this fact provides a most important general principle, and one which we will find most useful in understanding the operation of the various solid-state devices which we will meet in later chapters. This is simply that, for all semiconductors—whether homogeneous or non-homogeneous—we can describe a specimen of material as being in electrical equilibrium if, and only if, the Fermi level is constant throughout the specimen. In actual fact this principle is quite fundamental and applies not just to semiconductors, but to all materials.

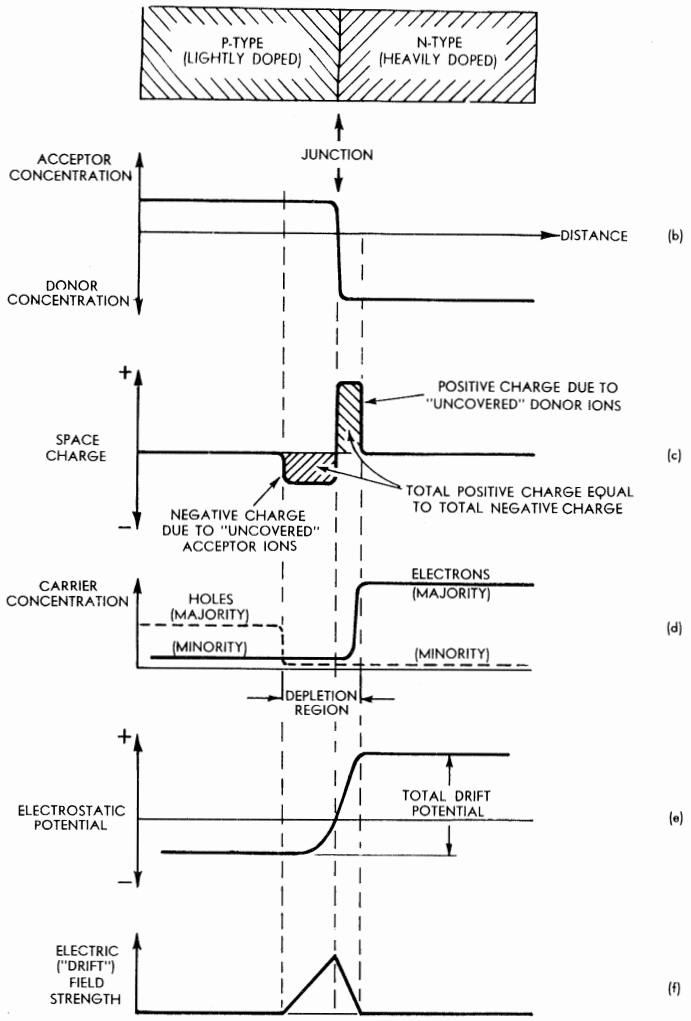


Figure 4.3

Before leaving this general discussion concerning non-homogeneous semiconductors, we should perhaps note that a very useful conclusion may be drawn regarding the intensity of the electric drift fields set up in such materials as a result of impurity concentration gradients. This is simply that, because a high concentration gradient will tend to produce a correspondingly high diffusion current, it will naturally also tend to result in the setting up of an appropriately strong internal drift field, in order to produce the high reverse drift current necessary for equilibrium.

In other words, the intensity of any electric fields set up in non-homogeneous semiconductors, in equilibrium, is directly proportional to the impurity concentration gradients with which they are associated. Thus high gradients, produced by relatively large changes in doping concentration over short distances through the material, set up quite high electric field intensities. Conversely low gradients, produced by either small changes in doping level, or changes spread over relatively long distances, or both, set up relatively low field intensities. We will find in later chapters that this fact has many implications for solid state device design and operation.

For the present, however, let us turn to consider what is probably the most important basic "special case" of a non-homogeneous semiconductor, knowledge of which is virtually essential for an understanding of the operation of almost any solid state device. This is the **P-N junction.**

In its most basic form a P-N junction, as the name suggests, is a place in an impurity semiconductor crystal at which there is a relatively abrupt transition between a uniform P-type region and a similarly uniform (but not necessarily equal in resistivity) N-type region. Such a situation is illustrated in figure 4.3(a), which shows a junction between a lightly doped P-type region and a relatively heavily doped N-type region.

There are quite a variety of methods by which this type of situation may be produced in a semiconductor crystal, and a number of the appropriate techniques will be described in a later chapter. However, for our present purposes the method used to produce such a junction is not important. The essential requirement is that we have a crystal specimen in which one region has been uniformly doped with an acceptor impurity to produce P-type material, while closely adjacent to this region is another which has been

uniformly doped with a donor impurity to produce N-type material.

Although both regions of the specimen of figure 4.3(a) are uniformly doped, they are of opposite "type," so that the specimen is therefore not homogeneous. This much the reader may have deduced already; however, a fact which may be less obvious is that the specimen also has a steep impurity concentration gradient, despite the uniform doping on either side of the junction.

The fact is that the concentration gradient occurs right at the junction itself, because here the impurity concentration changes rapidly and effectively "reverses polarity" over a very short distance. This is shown clearly by the impurity concentration curve of figure 4.3(b).

From the foregoing discussion of impurity concentration gradients and their effects, one might predict that the steep concentration gradient represented by a P-N junction would result in a high carrier diffusion current, and consequently an equally high reverse drift current and an associated high-intensity electric field. And this is, in fact, exactly what happens.

Because of the large number of conduction band electrons in the N-type material relative to the number of such carriers in the P-type material, there will tend to be a diffusion current of electrons across the junction in the N-P direction. Similarly, because of the greater number of valence band holes in the P-type material relative to the N-type material, there will tend to be a hole diffusion current across the junction in the P-N direction.

As before, the effect of these diffusion currents is to upset the electrical neutrality of the specimen. The electron diffusion current in the N-P direction leaves an excess of positively charged donor ions in the N-type material, while also tending to create an excess of conduction band electrons in the P-type material. Conversely, the hole diffusion current in the P-N direction leaves an excess of negatively charged acceptor ions in the P-type material, while also tending to create an excess of valence band holes in the N-type material.

The P-type material thus tends to gain an excess of both conduction band electrons and fixed acceptor ions, both of which are negatively charged, while at the same time the N-type material tends to gain an excess of both valence band holes and fixed donor ions —both of which are positively charged. A potential difference is thus set up between the two types of material, with the P-type material negatively charged with respect to the N-type, and hence an intense inbuilt electric "drift" field is set up across the junction.

From the fact that the only impurity concentration gradient present in such a semiconductor sample is confined to the narrow junction region itself, it might be expected that the drift field set up would similarly be confined to this region. However, this is not the case; in fact, the field "spreads" slightly to either side of the actual junction region, to an extent depending upon the doping concentration of the material concerned.

What happens is that, in diffusing across the junction, both holes and electrons effectively leave regions in which they are the majority carriers, to enter regions in which they are minority carriers. There is thus a very high incidence of carrier recombination on either side of the junction —so high, in fact, that few, if any, free carriers of either type are found near the junction on either side.

As a result of this effective depletion of carriers from the regions immediately adjacent to the junction, there are no electric charges available in these regions to compensate for the fixed charges of ionised impurity atoms. Hence a negative space charge is set up in the region on the P-type side, due to ionised acceptor atoms, while conversely a positive space charge is set up in the region on the N-type side due to ionised donor atoms. It is these space charges which are, in fact, responsible for the drift field set up across the junction.

The total charge unbalance produced by the two space charge regions is just sufficient to produce a drift field such that carrier drift back across the junction balances the diffusion currents. And because the space charge regions are effectively only the result of redistribution of charge within the semiconductor specimen, and not the result of a gain or loss of charge by the specimen as a whole, the net charges contained in the two regions must be equal and opposite.

Because of this, each region is found to extend into the material concerned to a distance just sufficient to "uncover" ionised impurity atoms equal to half the necessary total charge unbalance. If the P-type and N-type materials have equivalent doping concentrations, this will mean that the space charge regions will extend equally on either side of the junction, to a distance inversely proportional to the value of the doping concentration. A high doping level will thus result in narrow space charge regions, and a low doping level in relatively wider regions for the same degree of excitation.

If the doping concentrations of the P-type and N-type materials are dissimilar, as in the example of figure 4.3(a), the space charge regions will naturally extend by differing amounts. This is illustrated by the curve of figure 4.3(c), where the negative space charge region in the lightly doped P-type material is seen to have extended further than the positive region in the heavily doped N-type material, in order to "uncover" an equal number of ionised impurity atoms.

The curves of figure 4.3(d) show the carrier concentrations which correspond to this type of situation. It may be seen that the two space charge regions together constitute a region, extending from either side of the junction, which is nearly exhausted of carriers and thus virtually "intrinsic" semiconductor. From this it should not be surprising to learn that it is usual to call this region the **depletion layer.**

In figure 4.3(e) is shown the curve of electrostatic potential for the P-N junction of figure 4.3(a), illustrating that the potential difference which appears in the specimen as part of the equilibrium is set up entirely within the depletion layer region. In other words, under equilibrium conditions there is virtually no change in electrostatic potential throughout the remainder of the specimen. Hence, as shown in figure 4.3(f), the electric drift field is confined to the depletion layer region, and reaches its maximum intensity at the junction proper.

Further insight into the P-N junction in equilibrium may be provided by the energy level diagram of figure 4.4. Here the particular junction represented for the purpose of illustration again has an asymmetric doping concentration profile, but the ratio has been reversed from that of the specimen of figure 4.3. In other words the junction is here visualised as between heavily doped P-type material and lightly doped N-type material.

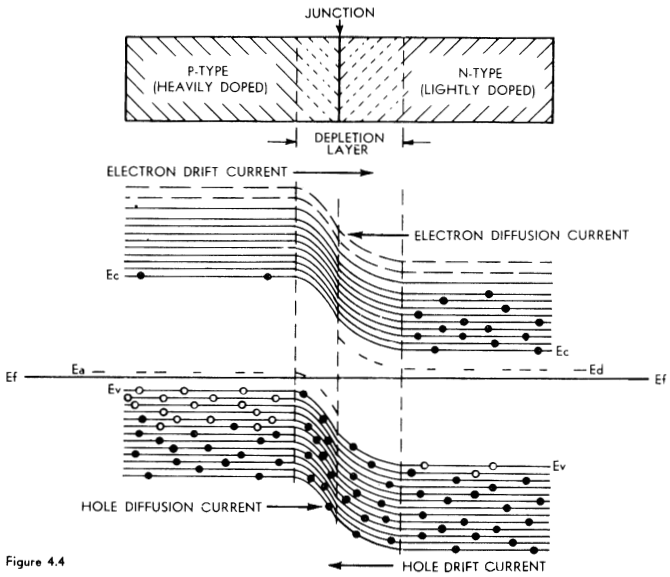


Figure 4.4

As may be seen, the equilibrium set up between diffusion and drift currents of both conduction band electrons and valence band holes has set up in the specimen the expected potential difference between the P-type and N-type materials, with a value just sufficient to make the Fermi level Ef constant throughout the specimen. The electric field associated with this potential difference is confined to the depletion layer region, as expected, this being shown by the fact that the energy levels slope appreciably only in this region.

At this stage it is hoped that the reader has gained a reasonably clear and satisfying picture of the P-N junction "in equilibrium" — which is, naturally enough, the situation which applies when such a semiconductor specimen is "left to itself" and not disturbed by the application of external electric fields.

Understandably this situation, while basic for an understanding of P-N junction operation, is of little direct interest where solid state device is concerned. Hence we should now turn to consider what happens when the junction is disturbed by external potential differences. However, before doing so it may be worthwhile to conclude the foregoing section with a brief summary which draws attention to the important points.

As we have seen, the steep doping concentration gradient present at a P-N junction results in carrier diffusion currents across the junction, with majority carriers from either side diffusing across to the other side and becoming minority carriers. A high incidence of carrier recombination thus tends to occur in the vicinity of the junction, which leaves a region of low overall carrier concentration and resultant "uncovered" impurity ions extending from the junction on either side. This region is the depletion layer, and corresponds to a layer of effectively "intrinsic" semiconductor material.

The "uncovered" impurity ions in the depletion layer result in a charge unbalance, and an electric "drift" field is set up across the junction. This results in drift currents of carriers across the junction in the reverse directions to the diffusion currents, and an equilibrium is set up when the two types of currents balance.

The higher the doping concentrations of the materials from which the junction is formed, the greater tends to be the concentration gradient at the junction, and the larger the diffusion currents. However the densities of impurity ions in the materials are directly proportional to the doping concentrations, with the result that the overall width of the depletion layer actually decreases with increasing doping concentration. Thus a junction between heavily doped materials tends to be relatively narrow, while a junction between lightly doped materials tends to be wide. The same factors result in unequal depletion layer widths on either side of a junction formed between materials of differing doping concentration, as we have seen.

It may be noted that the diffusion currents are effectively composed of majority carriers, because the carriers concerned are drawn from the majority carrier populations on each side of the junction. In contrast with this, the reverse drift currents are effectively composed of minority carriers, being drawn from the minority carrier populations of each material.

Let us now turn to consider what happens when a P-N junction is disturbed by the application of external potential differences. We shall find that its behaviour will depend quite markedly upon the polarity of the applied potential difference.

In figure 4.5 is shown the effect of connecting to a P-N junction specimen an external "bias" voltage, supplied by a battery whose positive pole is connected to the P-type semiconductor material, and whose negative pole is connected to the N-type material. This situation is normally called **forward bias.**

We have seen earlier that the effect of a potential difference applied to a semiconductor specimen is to set up an electric field along its length, and effectively raise the energy levels of the end of the specimen connected to the positive polarity relative to those of the end connected to the negative polarity. And this is what happens here, although the situation is complicated by the fact that the effective doping concentration — and hence the electrical resistivity — varies along the specimen.

Whether or not the P-type and N-type materials at either end of the specimen have differing resistivity will depend upon their doping concentrations, of course, and this will vary from specimen to specimen. However, regardless of the doping concentrations of these regions, the effective doping concentration of the depletion layer region is, as we have seen from the equilibrium case, very low. In effect, it behaves as "intrinsic" material, and has a relatively high resistivity.

Because of the high resistivity of the depletion layer region relative to the end regions, a major proportion of the applied potential difference is applied across the former. Hence the main effect of the applied forward bias is to tend to set up across the depletion layer a second electric field, in addition to the "inbuilt" field set up in equilibrium.

It may be seen that the polarity of this new field is opposite to that of the inbuilt field; that is, the two fields oppose. The effect of the forward bias is therefore to reduce the strength of the inbuilt field acting across the depletion layer, by partial cancellation. And, as shown by the electrostatic potential curve of figure 4.5(b), this has the result of effectively reducing the "potential barrier" opposing majority carrier diffusion across the junction. The majority carrier diffusion currents are therefore allowed to increase beyond their equilibrium values.

The minority carrier reverse drift currents in opposition to the diffusion currents cannot increase proportionally to maintain a balance, because they in contrast are almost completely limited by the numbers of minority carriers generated in the bulk of the material by the familiar "intrinsic" excitation mechanism. In short, the drift currents are "saturated," and virtually independent of the actual value of the potential difference across the depletion layer. (For this reason they are often called the **saturation currents** of the junction.)

When forward bias is applied to a P-N junction, then, the majority diffusion currents increase beyond their equilibrium values while their opposing minority drift currents remain substantially unaltered. A net current flow therefore takes place across the junction, with conduction band electrons moving from the N-type material

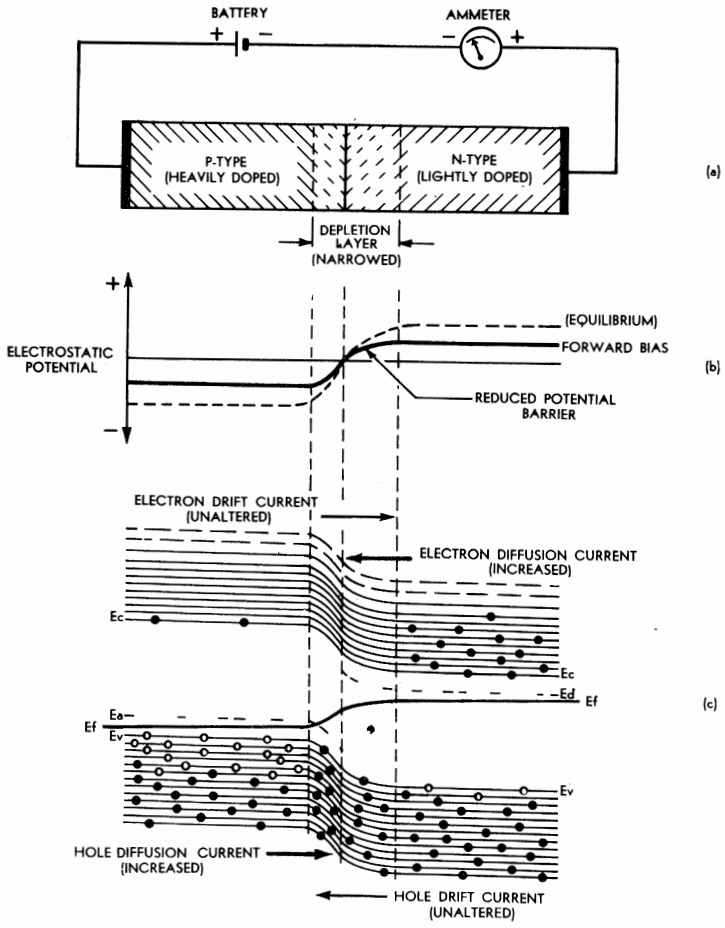


Figure 4.5

to the P-type, and valence band holes moving from the P-type material to the N-type. As the applied forward bias voltage is increased, the predominance of majority diffusion currents increases rapidly as the "inbuilt" potential barrier of the junction is progressively eliminated.

The current passed by a forward biased P-N junction thus increases quite rapidly with applied voltage, its resistivity falling rapidly to a very low value. This is illustrated by the right-hand half of the diagram shown in figure 4.7.

The energy level diagram for such a forward biased junction is shown in figure 4.5(c). Note that the Fermi level Ef is not constant throughout the material, an immediate sign that equilibrium conditions have been upset. The relatively steep slope of Ef in the depletion layer region indicates the degree to which the "inbuilt" field has been attenuated, while the corresponding slope in the energy bands themselves indicates the extent to which this field remains.

The composition of the current passing across a forward biased junction will naturally depend upon the impurity doping concentrations of the P-type and N-type materials. If the doping concentrations are equal, the current will be composed of equal numbers of conduction band electrons and valence band holes; however, if one material has a higher doping concentration than the other, the corresponding majority carriers will predominate.

Hence the forward biased junction current of a heavy-P/light-N junction such as that of figure 4.4 will consist mainly of holes, while that of a light-P/heavy-N junction such as that shown in figure 4.3 will consist mainly of electrons. But it should be remembered that conduction band electrons have a greater mobility than valence band holes, and this fact will also influence the exact ratio of currents flowing across the junction.

In should perhaps be noted, in connection with the foregoing discussion of the composition of forward biased junction current, that the composition of the junction current in no way determines the composition of the current current entering and leaving the semiconductor specimen from the external circuit. As we have seen, conduction in metallic conductors is effectively composed entirely of conduction band electrons; hence all current entering and leaving the P-N junction as a whole is of this form. What happens is that the "composition" of the current changes in the bulk of the material, due to the complementary mechanisms of "instrinsic" carrier generation and carrier recombination.

A further point to note regarding the forward biased P-N junction is that the width of the depletion layer region of a junction is narrower under forward bias conditions than for the equilibrium situation. This occurs because as we have seen the space charge of "uncovered" impurity ions in the depletion layer is intimately associated with the electric field and potential barrier. Hence when the latter are reduced in value, the space charge also reduces to correspond. The depletion layer thus contracts, leaving a smaller number of ions "uncovered."

If the external bias voltage connected to a P-N junction specimen is connected with the polarities reversed from the situation which we have just considered, its behaviour is somewhat different. This alternative arrangement is known as **reverse bias** and is illustrated in the diagrams of figure 4.6.

From figure 4.6(a) it may be seen that reverse bias involves the connection of the negative polarity of the external voltage to the P-type end of the specimen, and the positive polarity to the N-type end.

As before, a major proportion of such an applied potential difference is applied directly across the depletion layer region, because of its high resistivity, and a second electric field tends to be set up across the depletion layer in addition to the "inbuilt" field. But in contrast with the forward bias case, in which the two fields opposed, here the two fields are acting in the same direction. The field across the depletion layer is therefore increased in intensity from its equilibrium value, rather than decreased.

The effect of this increase in field strength is to effectively increase the height of the potential barrier which majority carriers must surmount in order to diffuse across the junction. This is illustrated in figure 4.6(b). As a result, few if any majority carriers of either type are able to cross the junction, and the diffusion currents fall considerably from their equilibrium values. Increasing the reverse bias voltage beyond about 0.5V effectively extinguishes the diffusion currents altogether.

As before, the minority carrier drift currents are virtually unaltered by the new situation, because they are "saturated" or limited by the numbers of minority carriers generated in the material by excitation. However, the magnitudes of the minority drift currents are actually very small—with silicon P-N junctions of modern manufacture, they together usually amount to but a small fraction of a microamp.

In the equilibrium condition, of course, the majority carrier diffusion currents are of equally small and opposite magnitude. However as we have seen, the diffusion currents fall away very rapidly with reverse biasing, virtually extinguishing for applied voltages greater than about 0.5V. For reverse bias voltages above this level the only currents drawn by a P-N junction are therefore the unopposed but very small minority carrier drift currents—the saturation currents.

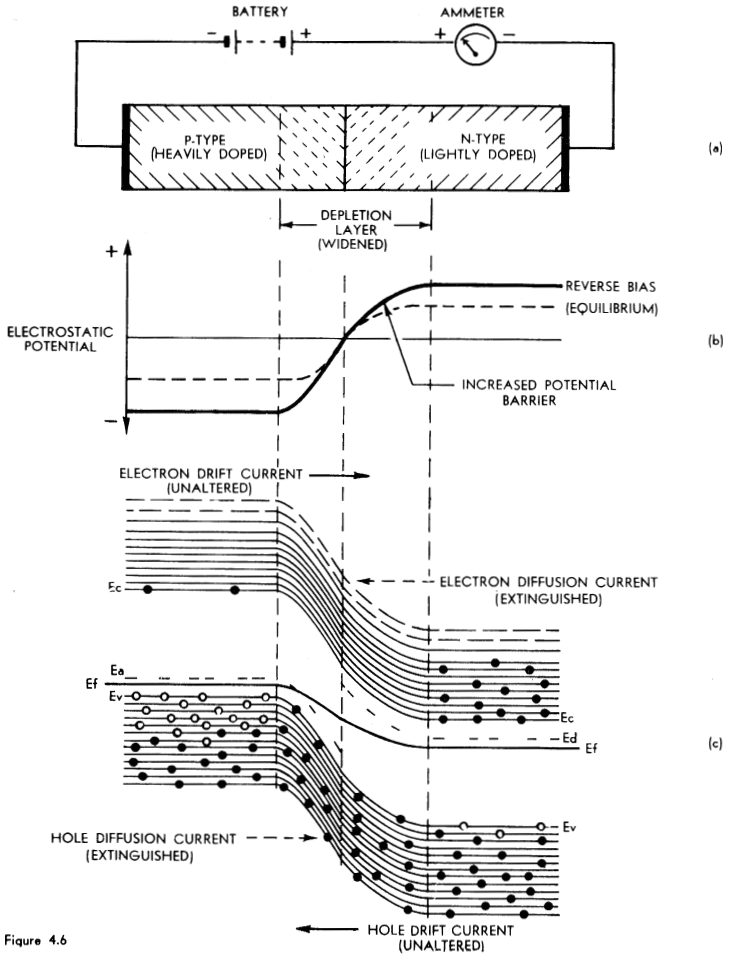


Figure 4.6

The current drawn by a reverse biased P-N junction thus tends to increase only very slightly with increasing voltage, rapidly reaching a constant and very low value corresponding to the sum of the saturation currents. This is illustrated by the left-hand portion of figure 4.7.

The energy level diagram for such a reverse biased junction is shown in figure 4.6(c). Again it may be seen that the Fermi level Ef is not constant throughout the material, indicating non-equilibrium conditions. The steep slope of Ef again occurs in the

depletion layer region, here signifying the extent to which the "inbuilt" field and the potential barrier of the junction have been increased. The full extent of the field present at the junction is indicated by the energy bands themselves.

In contrast with the situation under forward bias conditions, it may be noted that the depletion layer of a reverse biased junction is actually wider than for the equilibrium case. As before this occurs because of the intimate association between the space charge of "uncovered" impurity ions and the potential barrier. When the potential barrier is increased due to external reverse bias, the depletion layer therefore widens in order to "uncover" a correspondingly greater number of ions.

Because of this widening of the depletion layer the electric field intensity in the region does not increase as rapidly as it would if the layer width remained constant. However, it does steadily increase with increasing reverse bias, and inevitably a point is reached where one or another of a number of "breakdown" mechanisms occurs. When this occurs the effective resistivity of the junction again falls rapidly, and the current increases sharply from its basic "saturation" value.

The various mechanisms which may be involved when a reverse biased junction "breaks down" are each rather complex, and in fact not entirely understood; hence it will not be appropriate to examine them here in any detail. However, in broad terms the two main mechanisms involved are so-called field emission or Zener breakdown, and avalanche breakdown.

The **field emission or Zener breakdown** mechanism is usually that responsible for the breakdown of very heavily doped P-N junctions, which generally enter breakdown at reverse bias levels below about 6V. Due to the heavy doping concentrations in such junctions the depletion layer in very narrow, even under reverse bias conditions, and as a result of this the peak electric field intensity at the junction can be extremely high — in the order of 10^6 volts per cM, even at the low reverse voltages concerned.

When the electric field intensity reaches this order of magnitude, valence electrons may be effectively ripped from their orbit system, producing both a conduction band electron and a valence band hole. In short, the field itself produces electron-hole carrier pairs, and this explains the term "field emission." The carrier pairs thus produced in the depletion layer region are immediately swept away in opposite directions by the field, and as a result the junction current increases sharply from its saturation or "leakage" level.

Avalanche breakdown, the other main breakdown mechanism, is that usually responsible for breakdown in lightly doped junctions — generally those breaking down at reverse voltages above about 10V. As the name suggests, it is a mechanism whereby the minority carrier drift or saturation current itself effectively increases, due to an avalanching or "snowball" action.

In this type of breakdown the depletion layer is wide, both because of the light doping concentrations and as a result of the appreciable reverse bias voltage. Because of this the minority carriers drifting across the junction are ultimately able to develop sufficient momentum that, when each collides with a fixed atom, it is effectively able to ionise that atom by "knocking out" one or more new carrier pairs.

Such "ionisation by collision" involves a net gain in the number of carriers crossing the junction, because each carrier upon collision with a fixed atom can effectively produce two or more carriers. Hence as a result the junction current again rises sharply from its saturation value.

It should be noted that neither of the "breakdown" mechanisms just described involves **inherent** damage to the P-N junction: in themselves, they are merely mechanisms whereby the current drawn by the junction under reverse bias conditions increases markedly from its low saturation value when a particular voltage level is reached. It may be seen that they are thus rather different from the type of "breakdown" which occurs when excessive voltage is applied to dielectric materials such as paper or plastic.

Whether or not a junction sustains damage when it enters "breakdown" is primarily determined by the very same factor which determines whether or not it sustains damage in the forward bias mode: the power dissipation. If the power dissipated in the semiconductor material — most of which is dissipated in the depletion layer region, because of its greater voltage drop — is sufficient to cause overheating and disturbance to the crystal lattice structure, then damage generally results. But if this level is not reached, then the junction will sustain no damage. Some junctions in practical semiconductor devices are in fact designed to operate continuously in the "breakdown" condition, as we shall see in later chapters.

We have seen in the present chapter that the P-N junction behaves in rather different ways when external bias voltage is applied, depending upon the polarity of that applied bias. In one direction it tends to conduct readily, whereas in the other direction it tends to conduct only very slightly. No doubt the thoughtful reader will have already realised that this behaviour is virtually identical to that of the familiar thermionic diode valve, and will have noted the resemblance between the curve of figure 4.7 and the voltage-current characteristic of a diode valve.

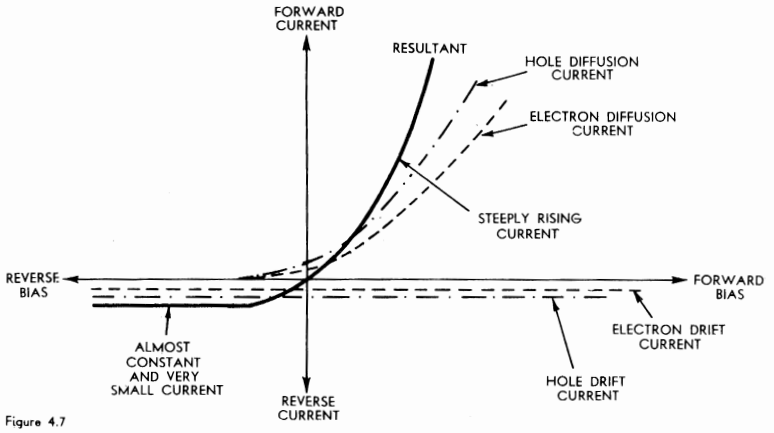


Figure 4.7

It should therefore come as no surprise to learn that the P-N junction is in fact the heart of the modern semiconductor or "crystal" diode, a device used in large numbers in almost every branch of modern electronics. At the same time, P-N junctions either singly or in combination also form the basis of almost every other modern semiconductor device, so that in the foregoing discussion of the P-N junction we have not only been describing the theory of crystal diode operation, but also laying the theoretical groundwork for many of the later chapters.

In the next chapter we take a look at the practical aspects of semiconductor diodes, examining both their various physical forms and their applications. However, before passing to this material the reader might perhaps be well advised to glance back over the material which has been presented in the present chapter, to ensure that he has fully grasped the important concepts involved.

SUGGESTED FURTHER READING

BURFORD, W. B., and VERNER, H. G., **Semiconductor Junctions and Devices,** 1965. McGraw-Hill Book Company, New York.

MORANT, M. J., **Introduction to Semiconductor Devices,** 1964. George G. Harrap and Company, London.

SHIVE, J. N., **Physics of Solid State Electronics,** 1966. Charles E. Merrill Books, Inc., Columbus, Ohio.

SMITH, R. A., **Semiconductors,** 1950. Cambridge University Press.

Chapter 5

THE JUNCTION DIODE

Diodes and semiconductor materials — reverse bias current — temperature effects — forward bias characteristics — high temperature operation — power rating — surge current rating — reverse breakdown — peak inverse voltage rating — switching speed — package capacitance — junction capacitance — charge storage — diode applications.

The basic P-N junction, whose behaviour was described in the preceding chapter, effectively forms the functional "heart" of almost every type of semiconductor diode. However, as the reader may already be aware, practical semiconductor diodes are encountered with widely differing electrical ratings. They are also found in circuits performing a variety of rather different tasks, and seen in an almost bewildering array of different physical forms.

In order to provide the reader with a satisfying explanation of these wide divergences between practical semiconductor diodes, it is necessary to expand the concepts of basic P-N junction operation already developed, and this will be attempted in the present chapter and in that which follows it. The present chapter will deal with what may be called "orthodox" diodes — that is, those devices which are designed to take advantage mainly of the unidirectional conduction properties of the P-N junction. Such diodes include those commonly encountered in circuits performing rectification, signal detection, mixing, switching, gating and clipping.

Chapter six will deal in turn with those diode devices which are designed to take advantage of aspects of P-N junction behaviour other than that of unidirectional conduction. Examples of this type of device are diodes used as voltage regulators and coupling elements, variable capacitors, oscillators and amplifiers, light detectors and energy converters.

Perhaps the first thing to be noted regarding practical semiconductor diodes is that, as one might perhaps expect, they are made from a number of different semiconductors. A very large majority of diodes in use at the present time are made from either germanium or silicon; the latter having been used to a lesser extent in the early days of semiconductor technology because of manufacturing difficulties, but now used very extensively and possibly to a greater extent than germanium. Other semiconductor materials which are becoming used for diodes include gallium arsenide, gallium phosphide and gallium antimonide.

The electrical behaviour and the ratings of a diode are both influenced significantly by the semiconductor material from which it is made. As we shall see, the semiconductor concerned plays a significant part, along with the doping level, in determining the voltage-current characteristics of a diode for both forward and reverse bias. It also determines the extent to which this behaviour varies with temperature, and the power which the device is capable of dissipating before this behaviour is permanently altered.

As we saw in chapter 2, all crystalline semiconductors are alike in the sense that, in the ground state, they behave as electrical insulators. The valence electron energy band is completely filled, while the empty conduction band is isolated from it by the "forbidden energy" gap. From an electrical viewpoint the essential differences between the various semiconductors arise mainly because this forbidden energy gap has a different width in each case.

Germanium, it may be remembered, has a forbidden energy gap width of 0.72eV (electron-volts), while silicon has a somewhat larger gap width of 1.11eV. The compound semiconductor gallium arsenide has a gap width which is even larger again at 1.39eV.

The width of the forbidden energy gap was shown earlier to control the conductivity of intrinsic semiconductor material, by determining the excitation energy required for electrons to be transferred to the conduction band. From this, and knowing that the generation of minority carriers in an impurity semiconductor material takes place by the same "intrinsic" mechanism, it should be fairly clear that the gap width also determines the number of minority carriers generated in an impurity semiconductor at any given excitation and doping level.

However, it is also true that the width of the energy gap controls, in a minor, but inverse manner, the relationship between minority carrier level and excitation level. Thus, although material with a wide energy gap tends initially to have a smaller number of minority carriers than material with a narrower gap, for the same excitation, its minority carrier population tends to multiply slightly more rapidly with increasing excitation.

Hence, while silicon impurity semiconductor material tends to have a considerably smaller minority carrier population than germanium material, at room temperatures, it also exhibits a slightly increased tendency for this population to grow as the temperature is increased. Despite this the minority carrier population of typical silicon

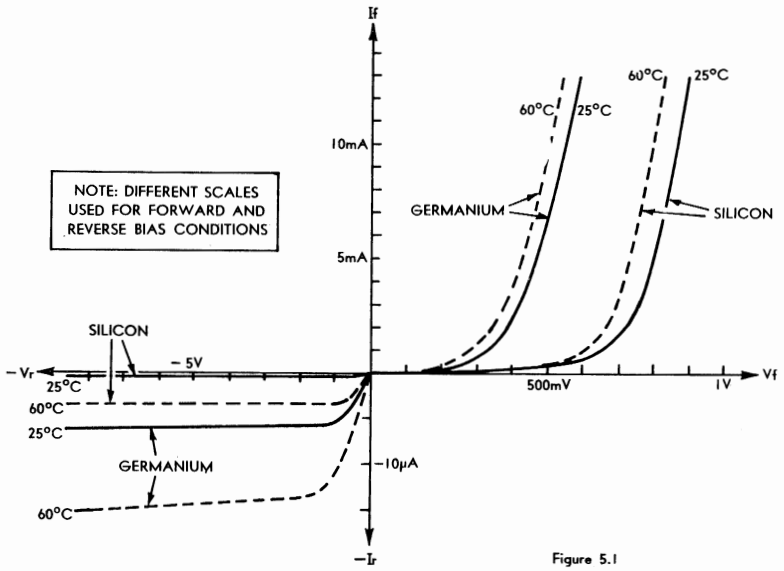


Figure 5.1

material does not even approach that of germanium until very high temperatures are reached, both because germanium has a larger initial population, and because this population itself increases significantly with temperature.

What effect do these differences have on the behaviour of practical P-N diodes? They have a significant effect upon the reverse-bias saturation currents, because it may be recalled that these currents are directly proportional to the minority carrier populations on either side of the junction.

In short, diodes made from a semiconductor material having a relatively wide forbidden energy gap, such as silicon or gallium arsenide, tend to have a very low reverse bias saturation current at normal temperatures. In comparison diodes made from a semiconductor material such as germanium, which has a relatively narrow energy gap, tend to have a somewhat larger saturation current at the same temperature. This despite the fact that in the former case the saturation current will tend to increase slightly more rapidly with temperature.

It is true that the total reverse bias current drawn by a practical semiconductor diode is not composed of the minority carrier saturation currents alone. It is very difficult, during the manufacture of practical diodes, to ensure that the surface of the semiconductor crystal element or "die" does not become contaminated in some way, and such contamination tends to result in additional reverse bias currents, which are commonly referred to as **leakage** currents.

Early in the history of semiconductor device development, these leakage currents were typically of the same order of magnitude as the saturation currents. However, in recent years, manufacturing techniques have been considerably improved, and leakage currents can typically be held to a very small fraction of the saturation currents. Hence, with modern semiconductor diodes and other devices, the reverse bias current drawn by an independent P-N junction is almost entirely composed of the minority carrier saturation currents.

In quantitative terms, the total reverse bias current of a typical modern silicon diode is of the order of a few hundred nA (nanoamps), at room temperature. Comparable germanium diodes typically have reverse bias currents of something like 100 times this figure, i.e., a few tens of uA (microamps). Because of the influence of excitation upon minority carrier generation these figures both increase as the temperature is raised, the silicon device current increasing slightly more rapidly.

Typically the reverse bias current of a germanium diode approximately doubles for every 8°C rise in temperature, while that of a silicon diode approximately doubles for every 5°C rise.

An illustration of the reverse-bias aspect of diode performance is provided by the left-hand portion of figure 5.1, which shows the reverse-bias currents of typical silicon and germanium diodes compared at room temperature (25°C) and at 60°C. It may be seen that at both temperatures the silicon diode has a considerably lower saturation current, even though the proportional increase may be larger over the temperature range concerned.

From the foregoing one might be tempted to infer that, because silicon diodes have lower reverse bias currents than germanium diodes under similar conditions, they would consequently be preferable for any application requiring a device whose performance should approach that of an "ideal" unidirectional conducting element. However, while this is true where reverse bias is concerned, unfortunately the reverse is the case under forward bias conditions. Here it is found that germanium diodes are somewhat closer to the ideal.

The reason for this is that, in addition to its influence upon minority carrier generation, and consequently upon saturation currents, the forbidden energy gap width of a semiconductor also plays an important part in determining the magnitude of the "inbuilt" drift field and potential barrier set up across a P-N junction in equilibrium. As a result the gap width also has a controlling influence upon the forward bias characteristic of such a junction, because it may be remembered that the forward bias current consists of excess majority carrier diffusion currents, which develop as the inbuilt potential barrier is surmounted.

For a semiconductor with a relatively wide forbidden energy gap, there will be a large energy difference between the Fermi levels of P-type and N-type material. Because of this, the potential barrier set up across a P-N junction made from the material will be relatively large under equilibrium conditions, compared with that across a junction made from a semiconductor having a relatively narrow energy gap. In turn this will mean that a relatively high external forward bias will be required before the internal barrier is surmounted.

Hence, because of the wider energy gap of silicon, a diode made from this material tends to require a higher applied forward bias than a comparable germanium diode for the same total forward conduction current. This is illustrated by the right-hand portion of figure 5.1, which shows the forward conduction characteristics of typical silicon and germanium diodes compared as before at 25°C and 60°C. It may be seen that the silicon diode is "harder to turn on" than the germanium device, and also that it has a higher voltage drop when in forward conduction.

It should be noted that both types of device "turn on" at a lower voltage, and have a lower conducting voltage drop, at the elevated temperature. The reason for this should become clear if it is recalled that the Fermi level of an impurity semiconductor moves toward the forbidden energy gap midpoint with increasing excitation, due to the increase in minority carriers. This means that the energy difference between the Fermi levels of the P-type and N-type materials becomes less as the temperature is raised, and accordingly the junction barrier potential also decreases. Forward conduction thus takes place at a lower applied voltage.

At this stage it should be fairly clear that when both forward and reverse characteristics are considered, neither silicon nor germanium diodes have a clear advantage. The silicon diode tends to have a somewhat lower reverse bias current, and therefore, more closely approximates the "ideal" diode in the reverse bias condition, but the germanium diode has a lower forward bias voltage requirement and thus represents the closer approximation to the ideal in the forward bias condition.

In terms of characteristics, then, the choice of the semiconductor material from which a diode is made depends largely upon the ultimate application and its requirements. If the application is one in which low reverse bias current is necessary or desirable, then a diode

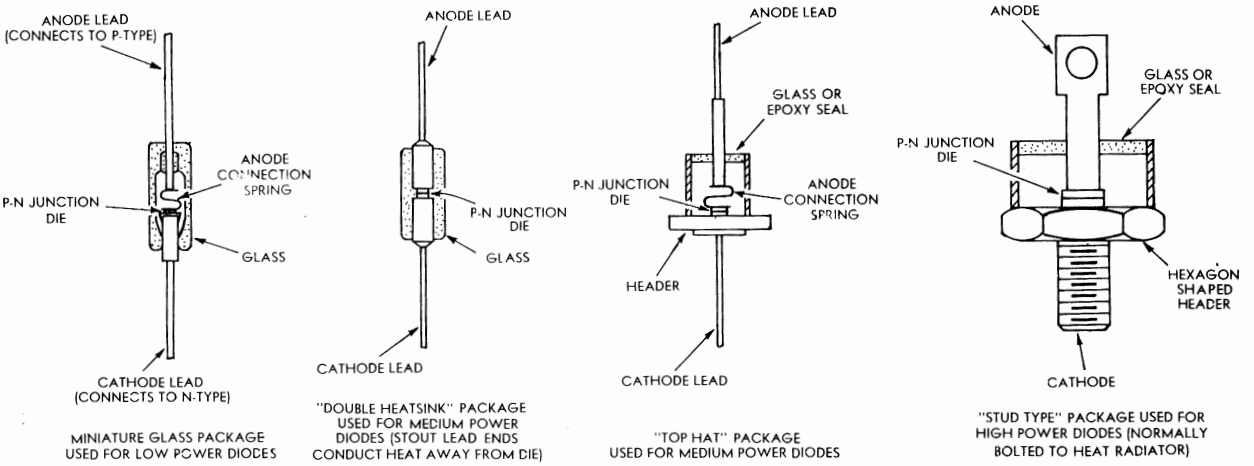


Figure 5.2

made from a wide energy-gap material such as silicon or gallium arsenide would be most appropriate.

Conversely if the prime requirement of the application concerned is for turn-on at a low voltage and minimum forward voltage drop in conduction, then the choice would favour a diode made from a narrow energy-gap semiconductor such as germanium. It is true that if either both forward and reverse bias behaviour were critical, or both were not unduly critical, the choice would be less straightforward. In such cases the decision might well be made on the basis of other factors, one of which would probably be operating temperature capability.

Generally a diode made from a semiconductor having a wide energy gap is more suitable for high temperature operation than a diode made from a semiconductor having a narrow energy gap. This is partly because of the somewhat lower reverse bias current at higher temperatures. However, a further reason is that the energy gap of a semiconductor plays a part in determining both the temperature at which the electrical structure of the device begins to alter permanently, due to thermal diffusion of the actual impurity atoms and ions, and also the crystal melting point. The wider the energy gap, the higher these temperatures tend to occur.

In practice the manufacturer of a semiconductor diode or other device usually rates his product in terms of the maximum allowable **junction temperature.** This is done in order to take into account the fact that both the ambient temperature and the power dissipated by the device contribute to its internal operating temperature.

Typically, germanium devices are given a maximum junction temperature rating of around 80-90°C, while silicon devices are usually given a somewhat higher rating of between 150-180°C. A silicon device would, therefore, be the logical choice for most applications involving high temperatures and/or very high power dissipation.

In order to allow the user to ensure that a device is operated within its maximum junction temperature rating at all ambient temperatures, the manufacturer must also normally provide information regarding the typical temperature rise of the device junction(s) with power dissipation. This informaton is usually given in terms of the **thermal resistance** of the device, expressed in units of (degrees C/watt dissipation).

Naturally the thermal resistance of a particular device depends upon both the size of the semiconductor crystal die itself, and the physical "package" in which it is mounted. Hence a device intended for very low power applications may have a very small die and be mounted in a small glass or plastic package having a fairly high thermal resistance, while a device for high power use will normally have a relatively large die and will be mounted in a large metal package of low thermal resistance.

In addition to thermal resistance, a crystal die and its package also possess thermal "capacitance" or inertia. Because of this, heating and cooling of the device involve definite thermal time-constants. Hence the heating of the die tends to be proportional not to the instantaneous power dissipation, but to the average dissipation taken over a short time interval — the interval length depending upon the crystal die itself, and on the package and its thermal time-constant.

As a result of this averaging effect, a diode is typically able to withstand short bursts or "surges" of power dissipation which may be considerably higher than its continuous or "steady-state" dissipation rating. This short-term capability is often expressed in terms of the forward conduction **surge current rating** of the device, which may be given a number of values for different time periods.

Depending upon the device itself and also upon the time period for which a surge rating is given, it may represent from about five times to more than 50 times the forward current corresponding to the continuous power rating of the device. The shorter the time involved, naturally enough, the higher tends to be the figure; however devices may be produced with the ability to withstand quite long surges of high amplitude, by appropriate thermal design.

Further discussion of thermal considerations will be given in a later chapter. However, from the foregoing it should be apparent that power dissipation requirements provide at least a partial explanation for the variety of packages in which semiconductor devices are found. Figures 5.2 and 5.3 show the basic construction of some of the diode packages in common use.

In general each of the various sizes and packages in which "orthodox" diodes of both the germanium and silicon type are made available are further subdivided into many individual device types differing from one another mainly in terms of two other important parameters. These are the reverse breakdown characteristic, and the switching speed, each of which will now be briefly discussed.

It may be remembered that if the reverse bias voltage applied to a P-N junction is increased, a point is eventually reached where the junction current rises rapidly from its low saturation value, and the junction is then said to have entered **"breakdown."** One of two main mechanisms is usually responsible for this behaviour, one being called field emission or Zener breakdown, and the other avalanche breakdown.

As was explained in the preceding chapter, the phenomenon of junction reverse breakdown does not involve inherent damage. However, it does constitute a potentially high-dissipation mode of operation, because under breakdown conditions a junction tends to maintain a relatively large voltage drop while at the same time being capable of heavy conduction.

It is also true that with practical P-N junctions, in diodes and other semiconductor devices, breakdown tends to occur unevenly and in a localised manner at some specific point on the crystal die. As a result, the increased current which flows is concentrated in a small area, and localised overheating and damage can occur with great rapidity at power levels considerably lower than the forward conduction continuous power rating of the device.

By exercising extreme control over

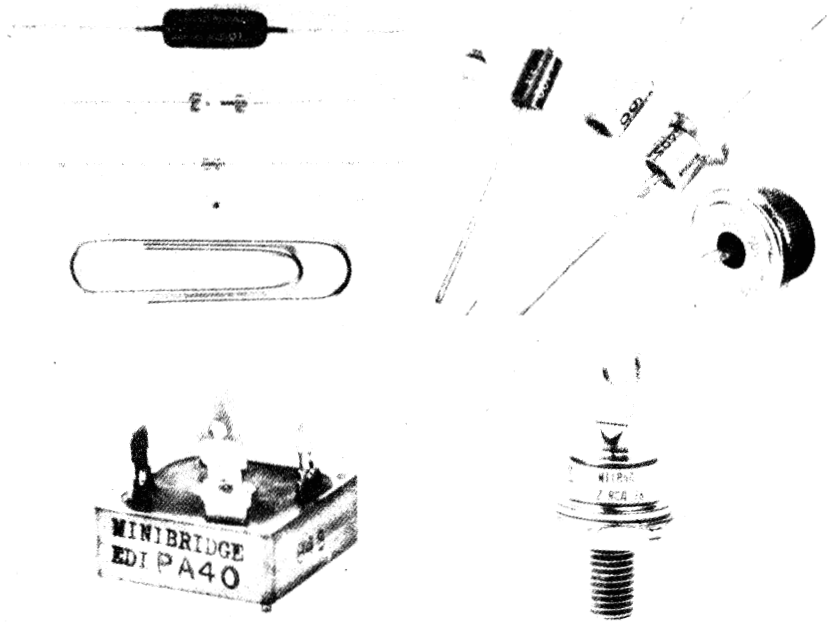


Figure 5.3. Typical semiconductor diodes. At upper left are various low-power or "signal" diodes, compared in size with a common paper clip. At upper right are four medium-power diodes as used in many receiver and amplifier power supplies, together with a power diode used in the rectifier within an automotive alternator. At lower left is an assembly containing four high-power silicon diodes, connected for bridge rectification. At lower right is a single stud-mounting high power silicon diode capable of handling an average current of 40 amps. All devices are shown approximately normal size.

cleanliness and such factors as doping uniformity during the various fabrication processes, device manufacturers have recently been able to effect a considerable reduction in this tendency for localised breakdown. However, the "transient protected" devices which have resulted from this effort are necessarily more costly than devices fabricated under less stringent conditions: and, of course, such devices still enter breakdown eventually, albeit in a uniform and evenly distributed manner.

Junction breakdown thus represents a condition which at the very least involves potential device damage. It should also be evident that quite apart from this, the rise in reverse current, which tends to occur at breakdown, represents in itself a significant departure from the ideal diode characteristic.

For a practical diode, therefore, the reverse breakdown characteristic is of considerable importance. It must be considered not only with relevance to the protection of the device itself, but also because of its possible consequences in the circuitry into which the device is connected.

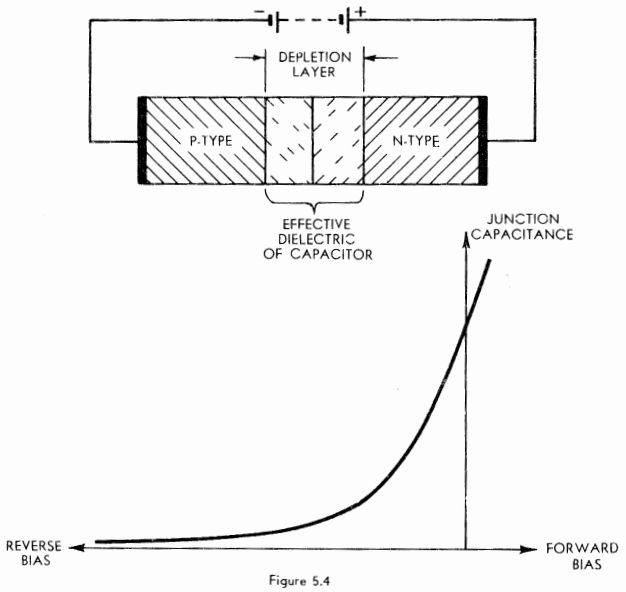


Figure 5.4

Usually the reverse breakdown characteristic of a semiconductor diode is specified in terms of a **peak inverse voltage or "PIV" rating,** which in effect represents a specific value of reverse bias voltage at or below which no device of the type concerned should enter breakdown. Some types of device may be given a number of different PIV ratings, to cover both steady-state and various reverse transient conditions. The "transient protected" diodes mentioned earlier are examples of devices normally given such multiple ratings.

Both silicon and germanium diodes may be manufactured to exhibit a wide range of breakdown voltages. However, devices required to have a very high breakdown voltage rating are usually made from silicon or some other semiconductor having a similarly wide energy gap. This is because the relatively high reverse saturation current of a narrow-gap semi-conductor such as germanium tends to make it very difficult to delay the onset of avalanche breakdown.

Germanium diodes are typically available with PIV ratings ranging from less than a volt to about 150V. Silicon diodes are available with PIV ratings ranging from about 3V to more than 1500V. Still higher PIV ratings can be produced by connecting a number of individual silicon dice in series; devices with PIV ratings in excess of 50KV have been produced using this technique.

As noted earlier, a further important general parameter of practical semiconductor diode behaviour is **switching speed.** This basically describes the ability, or otherwise, of a device to rapidly follow any changes in external circuit conditions. As diodes are often found in circuits involving rapid reversal of the bias voltages applied to the device, this parameter can be of considerable importance.

One of the main factors determining the switching speed of a diode is its **shunt capacitance,** which is simply the total effective capacitance present between the two device electrodes. Because it is effectively in parallel with the actual diode element, this capacitance can have a considerable influence upon the overall high-speed performance. For example, it tends to draw a current component which is purely proportional to the rate of change of applied voltage, regardless of polarity; behaviour which fairly obviously represents a significant departure from that of an ideal diode.

Naturally enough the diode package alone will contribute to the total shunt capacitance, as some finite package capacitance is unavoidable with practical devices. However, by careful design manufacturers have been able to produce packages with very low shunt capacitance, and these are normally employed for those devices intended for extremely high speed operation.

Quite apart from the package capacitance, however, an important component of the total shunt capacitance is provided by the inherent capacitance of the diode P-N junction itself. This capacitance is known as the "depletion layer capacitance," "barrier capacitance," "space charge capacitance," "junction capacitance," or "transition capacitance."

Although it may seem surprising at first that the P-N junction itself acts as a capacitor, the reason for this should become evident after a moment's consideration. Essentially, a capacitor consists of two conductors separated by a dielectric, and in the P-N junction we have, after all, two quite high conductivity semiconductor regions separated by a low conductivity depletion layer region. The latter is largely devoid of carriers, yet provided with the facility for charge storage in the form of ionised impurity atoms; small wonder, therefore, that it acts as a very effective dielectric.

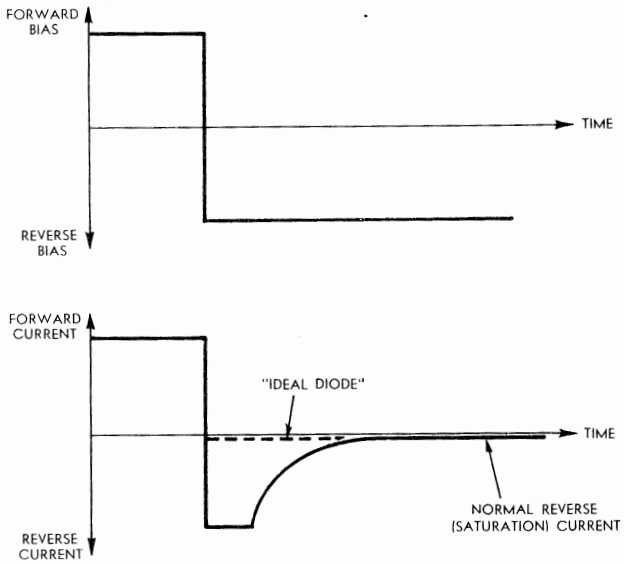


Figure 5.5

Of course the width of the depletion layer varies with applied voltage, as we have seen. Under equilibrium conditions, with zero applied bias, it has a width determined by the semiconductor concerned and by the doping levels. If reverse bias is applied, the depletion layer widens to uncover more impurity ions, and conversely if forward bias is applied it narrows to reduce the number of uncovered ions.

Because of this width variation, the junction capacitance is not static but also varies with applied voltage. This is illustrated in figure 5.4, where it may be seen that the junction capacitance of a typical diode varies inversely with reverse bias voltage, and directly with forward bias voltage.

The junction capacitance of a device may be minimised by using the smallest crystal die capable of handling the required power, and by using low doping levels to result in a relatively wide depletion layer. Naturally the latter technique involves a compromise, as low doping levels also increase the resistivity of the material and hence tend to increase the forward voltage drop and consequently lower efficiency.

As will be discussed in the next chapter, some semiconductor diodes are expressly designed to exhibit a very high junction capacitance. Such diodes are intended not for use as unidirectional

circuit elements, but rather as voltage-controlled variable capacitors.

Yet another important factor which influences the switching speed of a semiconductor diode is the phenomenon known as **charge storage** or **minority carrier storage.** This is particularly relevant where a diode is required to switch rapidly between the forward conducting or "on" state and the reverse-biased "off" state.

When a P-N junction is conducting due to forward bias, it may be remembered, excess majority carrier diffusion currents are flowing in both directions across the junction. At the same time the depletion layer has a width somewhat less than that for equilibrium conditions, and the potential barrier a somewhat lower value.

If the voltage applied to the device is changed, these conditions must also change to achieve a new dynamic balance. Thus if the forward bias is increased, additional carriers must be swept across the junction to set up higher diffusion current levels, while at the same time some of the previously ionised impurity atoms must be neutralised to reduce the depletion layer width and reduce the potential barrier.

Conversely, if the bias is reduced or reversed in polarity, the number of carriers crossing the junction must fall, while additional impurity atoms must be ionised to widen the depletion layer and increase the potential barrier.

In both cases, significant time must elapse before the new conditions stabilise. The depletion layer changes involve movement of carriers through a finite volume of material, and this necessarily takes time. Hence there is an inevitable delay involved before the new balance conditions are reached, and during the delay period the behaviour of the device may differ considerably from that of an ideal diode.

For example, figure 5.5 shows what tends to happen if the polarity of the applied voltage is suddenly switched from a forward bias value to a reverse bias value. Ideally when this occurs the diode current would drop immediately to its very low reverse saturation current value; however, it can be seen that what in fact happens is that the current swings rapidly to a high reverse value, and only subsequently falls back exponentially to its saturation value.

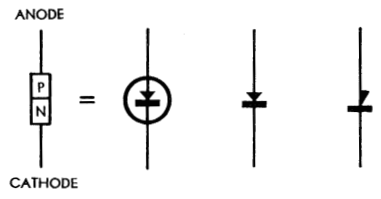
The reason for this is that at the instant of bias reversal, a considerable number of carriers of both types are stored or "trapped," in the depletion layer region and also in the adjacent P-type and N-type material as injected minority carriers. Before normal reverse-bias operation can be achieved, these carriers must all be removed, generally by being swept back across the junction in both directions. It is the removal of these stored carriers which results in the temporary high reverse current.

The charge-storage mechanism can be controlled to a considerable extent by special techniques involving non-uniform doping and careful choice of impurities. The rather specialised diodes produced by such techniques include those called "step-recovery diodes," "snap-off diodes," "avalanche switching diodes" and "PIN diodes."

To conclude this discussion of "orthodox" semiconductor diodes, brief descriptions will be given of a small, but representative selection of the great many applications of these devices. Before the applications are discussed, however, brief mention will be made regarding diode symbols used in circuit diagrams, for the possible benefit of those readers as yet unfamiliar with the devices.

The circuit symbols most commonly used for semiconductor diodes are shown in figure 5.6, together with a simplified representation of the basic P-N junction shown for reference. Note that the symbols are all similar in that they use an arrow-head to represent the P-type material, and a bar or line to represent the N-type material. The arrow-head is actually intended to indicate the direction of forward or "easy" current flow according to the classical "positive charge" current convention.

For orthodox diodes the electrode connecting to the P-type material is normally labelled the "anode," as shown, while the N-type electrode is labelled the "cathode." However, these terms really depend upon the polarity of the applied voltage, and may be reversed in certain cases.



DIODE SYMBOLS

Figure 5.6

Probably the most familiar application of semiconductor diodes is in circuits used for the rectification of alternating current into unidirectional current. In fact they are particularly well suited for this task, because, despite the limitations discussed in this chapter, they still represent the closest available approximation to an ideal diode element.

There are numerous different rectifier circuit configurations, each of which has certain distinct advantages in specific situations. Four of the most common configurations are illustrated in figure 5.7.

The first of these is the "half-wave" rectifier, figure 5.7(a), which is the simplest circuit in common use. As the name suggests, this configuration employs a single diode element which is arranged to allow only alternate half-cycles of the AC input to reach the load circuit, while simply rejecting the half-cycles of opposite polarity.

The use of a reservoir capacitor "C" helps to smooth out the appreciable ripple which tends to be present in the output as a result of the "gaps" between the half-cycle pulses delivered by the diode. However, even with a relatively large reservoir capacitor the ripple tends to be high, and the output rather poorly regulated, as a result of the fact that the reservoir capacitor may be discharged continuously by the load, but can only be recharged by the diode on every alternate half-cycle. The half-wave rectifier circuit accordingly finds use mainly in very low current applications.

The limitations of the half-wave circuit are obviated in the "full-wave" circuit shown in figure 5.7(b). Here two diode elements are connected to a transformer effectively having two identical secondary windings connected in series. Each diode conducts only on alternate half-cycles, as before, but the two elements are arranged so that one conducts for the positive half-cycles and the other for the negative half-cycles, and both charge the reservoir capacitor in the same direction.

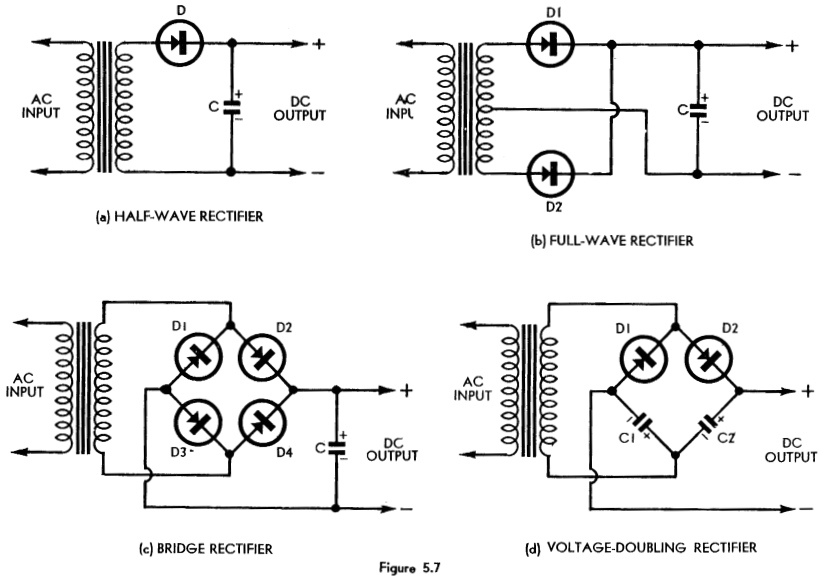


Figure 5.7

Because it effectively "uses" both the positive and negative half-cycles of the AC input, the full-wave rectifier tends to deliver less output ripple and possess better load regulation than the half-wave circuit. The ripple is also easier to filter out, having a fundamental frequency component of **twice** the AC supply frequency, whereas the ripple of the half-wave circuit has a fundamental component **equal** to the AC supply frequency.

The full-wave circuit is therefore better suited for high current applications; however it has the disadvantage that it normally requires a transformer having a double secondary winding. This requirement can be obviated by the use of the so-called "bridge" circuit, shown in figure 5.7(c).

Here a single transformer secondary winding is used, with two additional

diode elements used to effectively reverse both connections between the load and the AC supply on successive half-cycles. The circuit still performs full-wave rectification, and therefore tends to have low ripple and good load regulation. It differs from the "full-wave" configuration mainly in that transformer complexity has been reduced at the cost of two additional diodes.

The fourth configuration shown is the "full-wave voltage doubler" rectifier, figure 5.7(d), one of many configurations used to deliver an output voltage higher than the peak value of the input A.C. In this case the two diodes used are arranged to charge separate reservoir capacitors during their respective half-cycles, the capacitors being effectively connected in series with respect to the output and load circuit.

In a half-wave rectifier circuit, the diode used should normally have a current rating sufficient to allow it to carry the full value of the average load current. In contrast, the diodes used in full-wave, bridge or full-wave doubler circuits need only have a rating sufficient to allow them to carry approximately 50% of the average load current because in these circuits the average current is shared between elements. In other circuits a different sharing factor may apply, depending upon the number of diodes involved.

In each type of rectifier circuit the diodes used should also be capable of handling both the initial surge current which flows when power is applied with the reservoir capacitor(s) fully discharged, and also the repetitive current pulses involved because of the continuous discharge/periodic charge situation. The peak currents due to the latter effect tend to be higher with the half-wave circuit because of larger "gaps" between charging pulses.

The amplitude of switch-on current surges is limited by the effective impedance in series with the diodes, and typically this is mainly composed of the effective secondary impedance of of the transformer. If this impedance is too low, external low-value high wattage resistors may be added in series with each diode. Such resistors must also be used if devices are connected in parallel for increased current handling capacity, to ensure that they share the current properly.

The P.I.V. rating of the diodes used in rectifier circuits depends upon the configuration used. For the half-wave and full-wave circuits, for example, the diode P.I.V. rating should be greater than twice the peak no-load output voltage, whereas for the bridge and doubler circuits it need only be greater than the peak no-load output voltage itself.

Individual diodes may be connected in series to achieve a suitably high P.I.V. rating. However, unless "transient protected" devices are used, parallel R-C networks must be connected across each device to ensure that they share both repetitive and surge reverse voltages equally.

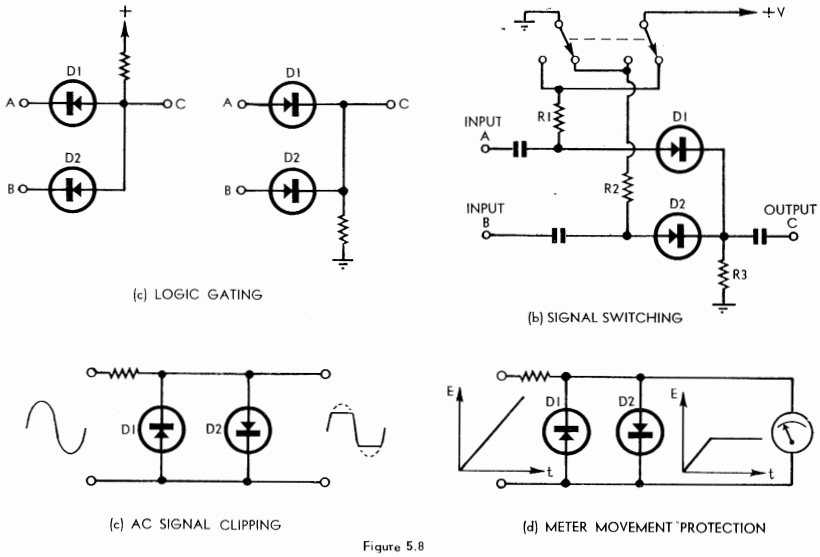


Figure 5.8

Many "rectifier" circuit configurations are in basic form suitable not only for power rectification, but also for detection—the process of extracting modulation information from a high frequency carrier signal. Hence signal detection circuits form another important application of semiconductor diodes, and account for many of the diodes found in radio and television receivers and test equipment.

A rapidly growing application for semiconductor diodes is in circuitry involved in logic gating and signal switching. Here the unidirectional properties of the device are used to effectively connect or disconnect circuit points in response to their relative voltage polarities.

Simple circuit configurations of this type are shown in figure 5.8(a) and (b). In the first circuit of (a), it may be seen that the diodes perform the "AND" operation, because point "C" will rise to a positive potential if, and only if, both points "A" **and** "B" are also raised to a positive potential.

Contrasted with this behaviour is that of the second circuit, which because of the changed diode connections performs the "OR" operation. In this case point "C" will go positive if either, or both, points "A" **or** "B" go positive.

The circuit in figure 5.8(b) illustrates the use of diodes for remote switching of AC signals. Here the circuitry is arranged so that when D1 is conducting and providing a signal path for input "A," diode D2 is reverse biased and held "off." Operating the switch reverses the situation, with diode D2 conducting and D1 held off.

Another important class of applications for semiconductor diodes includes circuits which take advantage of the fact that the forward characteristic of such devices is non-linear, representing a high initial resistance and subsequently a low resistance when the device reaches full conduction. Figure 5.8 also illustrates two of the many types of circuit which exploit this behaviour.

In the circuit of figure 5.8(c) it may be seen that two diodes are connected in inverse parallel across a source of sinewave signals, a resistor being used to limit diode current. During each half-cycle, one of the two diodes conducts; however, because of the forward bias characteristic, this conduction is effectively confined to that part of the half-cycle during which the signal exceeds the turn-on "knee." Hence the effect of the diodes is to effectively "clip" the signal to a known peak-to-peak amplitude.

The circuit of figure 5.8(d) shows how a similar diode configuration may be used to protect a delicate meter movement from damage due to overload. Here the non-linearity of the diodes effectively prevents the voltage applied to the movement from rising above the turn-on knee voltage, in either direction. Silicon diodes are normally used in this type of application, because their higher turn-on voltage and lower saturation current both ensure that normal meter operation is not disturbed.

SUGGESTED FURTHER READING

BRAZEE. J. G., **Semiconductor and Tube Electronics,** 1968, Holt, Rinehart and Winston, Inc., New York.

MORANT, M. J., **Introduction to Semiconductor Devices,** 1964. George G. Harrap and Company, London.

PHILLIPS, A. B., **Transistor Engineering,** 1962. McGraw-Hill Book Company, Inc., New York.

ROWE, J., **An Introduction to Digital Electronics**, 1967. Sungravure Pty. Ltd., Sydney.

——, "Transient Protected Rectifiers," in **Electronics Australia,** V.30, No. 10, January, 1969.

SMITH, R. A., **Semiconductors,** 1950. Cambridge University Press.

SURINA, T., and HERRICK, C., **Semiconductor Electronics,** 1964. Holt, Rinehart and Winston, Inc., New York.

Also "Solid-State Diodes," a special section in **Electronics World,** V.82, No. 1, July, 1969.

Chapter 6

SPECIALISED DIODES

Zener diodes — breakdown voltage — power dissipation — temperature coefficient — reference diodes — compensation — zener applications — varicaps — capacitance range — Q-factor — varicap applications — varactors — frequency multiplication — parametric amplification — tunnel diodes —back diodes —applications —photodiodes — light-emitting diodes — injection lasers.

Increasing the reverse bias voltage applied to a P-N junction diode eventually results in a phenomenon known as "breakdown," as we have seen in previous chapters. When breakdown occurs the normally very small and almost constant reverse bias current of the device suddenly and rapidly increases. It may be remembered that one of a number of mechanisms may be responsible for this rise in current, depending upon the doping levels and the construction of the device.

The mechanisms of breakdown do not involve inherent damage to the device, as we have noted. However, a diode which has entered this region of operation is capable of heavy conduction, while at the same time tending to maintain an appreciable voltage drop. The region therefore tends to be one of high power dissipation, and consequently of **potential** device damage.

In addition to the risk of device damage, there is the further consideration that in the breakdown region the behaviour of a device represents a significant departure from that of an "ideal" diode. It should therefore not be surprising that in a great many diode applications, considerable care is taken to ensure that device breakdown cannot occur.

Despite this there are certain applications in which diode breakdown is not avoided, but in fact intentionally planned. The reason for this is that, provided the device dissipation is kept below damage level, the voltage drop of a P-N junction in the breakdown region tends to be substantially constant, and independent of current level. A diode which is operating in the breakdown region may thus be used as a voltage regulating or limiting element, with applications similar to those of gas-discharge regulator tubes.

Although many "orthodox" semiconductor diodes may be used in this fashion, their usefulness as voltage regulators or limiters is generally rather limited. This is because with many devices there is a tendency, noted earlier, for breakdown to occur unevenly and in a localised manner at some specific point on the crystal die. Breakdown current thus tends to be concentrated in a small area, causing localised overheating and damage, even at relatively low power levels.

Some years ago, device manufacturers found it possible to obviate this problem by careful control of doping level, doping gradients and the cleanliness levels maintained during the various fabrication processes. This enabled them to produce devices designed specifically to be capable of continuous operation in the breakdown region. At first these devices were capable of only modest power dissipation but, in recent years, the techniques have been further developed and power capability has risen significantly. (The same development in techniques has resulted in the appearance of the "transient protected" rectifier diodes mentioned in the previous chapter.)

The names given to devices specifically intended for breakdown region operation are "breakdown" diodes, "regulator diodes," "reference" diodes, and "zener" diodes (often contracted to "zeners". The last of these terms should strictly only be applied to devices whose breakdown is due to the field-effect of Zener mechanism; however, it is widely used to describe all devices designed for breakdown operation.

Zener diodes are fabricated almost exclusively from silicon, because of the higher temperature/dissipation capability of this material compared with the other commonly used semiconductors. They are made in many of the physical packages used for "orthodox" diodes, including most of those shown in the previous chapter. The breakdown characteristic of a typical device is shown in figure 6.1.

By varying doping levels and gradients, device manufacturers are able to provide circuit designers with zener diodes having breakdown voltage figures ranging from about 3V to above 200V. For convenience, device types are usually given a **nominal breakdown voltage** designation according to the familiar logarithmic "preferred value" series used for resistors, capacitors and other components, and a similar tolerance system is used. Hence a particular device might have a rated breakdown voltage of (4.7V ± 5%).

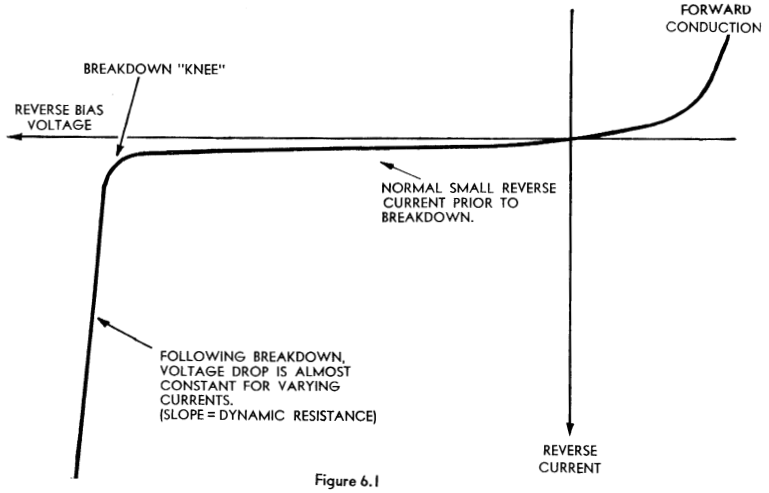


Figure 6.1

The nominal breakdown voltage of a zener diode is actually a somewhat arbitrary figure, because the voltage drop of a practical device in the reverse breakdown region is not entirely independent of current level. It also tends to be temperature dependent. With devices having a very low breakdown voltage there is also the problem that breakdown is not characterised by a sharp "knee" in the reverse bias behaviour, but by a rather gradual current increase.

Because of these factors, it is usual for the nominal voltage of a zener diode to be quoted for a particular current level, and for a specific am-

bient temperature. The behaviour of the device at other current levels and temperatures may then be described in terms of a current-voltage characteristic and/or a dynamic resistance figure, together with a temperature coefficient. The **dynamic resistance** of a device is the slope of the characteristic following breakdown, as indicated in figure 6.1; the temperature coefficient will be discussed shortly.

For most zener diode applications, a parameter of importance almost equal to that of nominal breakdown voltage is the device **power dissipation rating.** As with "orthodox" diodes, this rating determines the operating current levels at which the device may be operated for a given ambient temperature.

Currently available devices have continuous dissipation ratings ranging from a modest 200mW to more than 350 watts. High power devices have been developed with transient power dissipation ratings as high as 100KW for periods less than 100uS. The higher power devices often use a multiple-chip construction, with a number of crystal dice connected in parallel and/or series inside a common package.

Most manufacturers provide a number of ranges or "families" of zener diode devices, each range having a common package and an appropriate dissipation rating. Thus a manufacturer may provide a 400mW range, a 1W range, a 5W range, and so on, each range consisting of a series of device types covering the nominal breakdown voltage range.

It should be fairly clear that because of their differing breakdown voltages, the devices of a particular zener diode range having a common dissipation rating will have differing maximum current ratings. Thus in a 1W device range a nominal 10V device would have a maximum current rating of 100mA, while a 33V device would have a maximum current rating of only 30mA.

The **temperature coefficient** of zener diode breakdown voltage is quite often of importance, particularly in applications where a device is required to maintain a potential difference at substantially constant current over a wide temperature range. The name "reference diode" is sometimes reserved for devices intended specifically for this type of application.

It happens that the two main mechanisms responsible for P-N junction breakdown have **opposing** temperature coefficients. The field-effect or Zener mechanism responsible for low voltage breakdown (<6V) has a **negative** temperature coefficient: voltage falls with rising temperature. Conversely the avalanche mechanism responsible for high-voltage breakdown (>10V) has a **positive** temperature coefficient: voltage rises with rising temperature. In each case, a typical figure (absolute) is 5mV/°C.

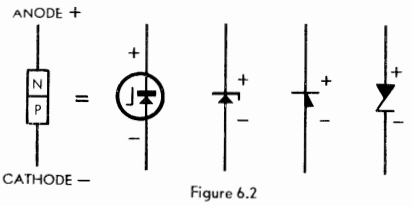


Figure 6.2

Because of the opposing temperature coefficients of the two mechanisms, cancellation tends to occur at the midpoint of the range of transition between the two which occurs at roughly 6V. Hence devices whose breakdown voltage is close to 6V tend to exhibit a very low temperature coefficient and are accordingly well suited for use as reference diodes.

Although many applications requiring a zener diode of low temperature coefficient can be arranged to employ a device with a breakdown voltage around 6V, this is not always the case. However, where the requirement is for a higher breakdown voltage, and this is a fairly common situation, there is fortunately another way of achieving high temperature stability.

It may be remembered that a forward biased P-N junction has a negative temperature coefficient, its forward voltage drop falling with temperature. Because of this, it is possible to effectively cancel the positive temperature coefficient of a zener diode of breakdown voltage higher than 6V by connecting in series with it one or more suitably designed or carefully chosen forward biased diodes. The combination will then exhibit a slightly higher effective breakdown voltage, but with a very low temperature coefficient.

Taking advantage of this idea, some device manufacturers have combined forward and reverse-biased dice inside standard packages to produce highly stable reference devices covering a wide range of nominal "breakdown" voltage. A typical device of this type has a nominal zener voltage of (11.7V ± 5%) at a specified current of 7.5mA, with a temperature coefficient of only 25uV/°C over the temperature range from −55°C to +150°C.

The same technique may be used with separately packaged devices, and circuit designers frequently combine zener and forward-connected diodes to obtain a low effective temperature coefficient at a certain voltage, using low-cost devices. Actually the technique is quite a flexible one because the forward-connected junctions used to compensate the zener need not be diode devices, but may well consist of any suitable junctions forming part of one of the more complex devices which we shall be meeting in later chapters.

The circuit symbols commonly used for zener diodes are shown in figure 6.2. It may be seen that in most cases the symbol attempts to indicate that the N-type side of the junction is connected to the positive supply polarity, in contrast with the connections for an "orthodox" diode. For a zener diode the N-type electrode is therefore the "anode," and the P-type electrode the "cathode."

Most applications for zener diodes are in power supply circuitry, where the devices are commonly used either as straightforward shunt regulators or as reference sources for feedback-type voltage or current regulating circuitry. Basic configurations for these applications are show in figure 6.3.

Figure 6.3 (a) shows a simple shunt regulator. Here the relatively constant voltage drop of the zener diode when operating in the breakdown region is used to provide a stabilised voltage source despite any variations in the unregulated input voltage and the load circuit current. In this type of circuit the resistor R is chosen so that the diode current has a value sufficient to permit the device to effectively "absorb" current changes due to loading or input variations, without exceeding the device dissipation ratings.

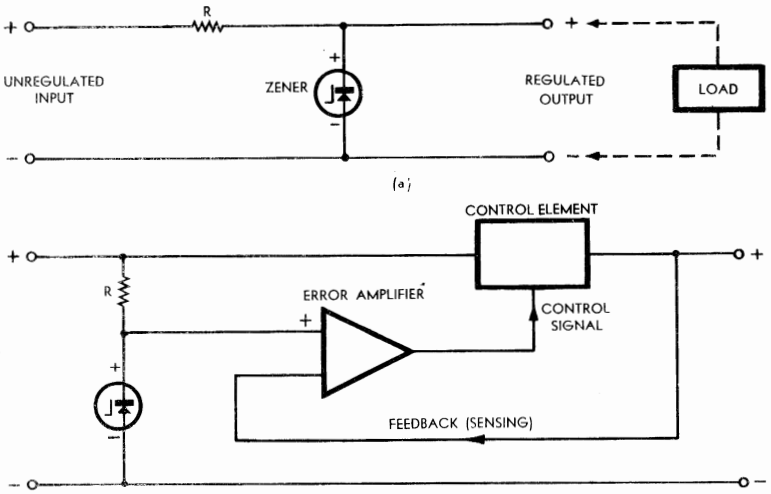


Figure 6.3

In figure 6.3 (b) is shown the somewhat more complicated scheme normally used where either very high accuracy regulation, or regulation at high power levels is required. Here the zener diode is used simply to provide a stable reference voltage source, against which the output quantity (in this case voltage) is compared. An error amplifier then provides a control signal proportional to any difference between the two, and this signal is used to correct the output signal by means of a power control element. As one might expect, a low temperature-coefficient "reference" device is often used in this type of circuit, to achieve the highest possible stability.

Zener diodes are particularly well suited for this type of application, possessing many advantages over the gaseous discharge regulator tubes formerly used. They are physically smaller and more rugged, and are available in a much wider range of nominal voltage and power dissipation ratings. Not only this but they have a lower dynamic resistance, giving

better regulation, and the further advantages that their characteristic lacks both the "ignition voltage" peak and the negative resistance segment which complicate the use of gas regulator tubes.

There are many uses for zener diodes other than as voltage regulators and reference sources. For example they are often used either singly or in combination for signal clipping and limiting, using configurations similar to that shown in figure 5.8 of the previous chapter. A single zener diode may be used for asymmetrical clipping, while two identical devices connected in inverse series may be used for symmetrical clipping.

Other applications include threshold circuits which change state when a voltage passes a critical level, circuits which effectively shift the zero reading of a meter movement to correspond to a finite applied voltage ("zero suppression"), and circuits in which

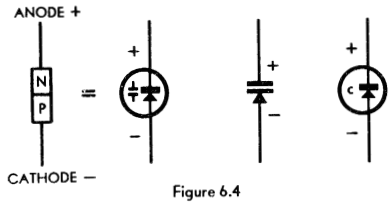


Figure 6.4

the device is used as a DC coupling element having substantially constant voltage drop.

Let us now turn from the zener diode to consider another important type of "special" semiconductor diode.

As we saw in the previous chapter, a P-N junction has inherent self-capacitance. The depletion layer which forms in the vicinity of the actual junction acts as a dielectric separating the remaining P-type and N-type regions, forming an "inbuilt" parallel-plate capacitor. The usual names given to this capacitance are "depletion layer capacitance" or "junction capacitance."

It may be remembered that the width of the depletion layer varies with applied bias voltage, so that the junction capacitance similarly varies. It has a high value at zero bias, when the depletion layer is relatively narrow, rising still further to an effective maximum at a value of forward bias just short of device "turn-on." Conversely as the depletion layer widens with increasing reverse bias, the junction capacitance falls and reaches an effective minimum at a point just short of reverse breakdown.

Provided that the voltage applied to a P-N junction is kept inside the range between forward conduction and reverse breakdown, this variation in junction capacitance in fact constitutes the main change in junction behaviour with applied voltage because, within this range, the net current drawn by the junction as a whole remains very small and almost constant. Broadly speaking, then, a P-N junction is potentially capable of acting as a voltage-controlled variable capacitor.

While most semiconductor diodes may be used in this fashion with some success, the usefulness of a typical "orthodox" device as a variable capacitor tends to be rather limited. Because the device has usually not been designed with this application in mind, the doping levels and gradients used do not generally result in smooth capacitance variation over a useful range. The crystal die structure and package construction also tend to introduce excessive series resistance and inductance, degrading performance at high frequencies. With germanium devices the saturation current also tends to be excessive.

Aware of the limitations of normal diodes for this type of application, and recognising the potential interest by circuit designers in devices which would lack the limitations, device manufacturers have in recent years developed diodes specifically designed to give optimum performance as voltage-controlled capacitors. These devices have become known as "varicaps," "varactors," or "variable capacitance diodes."

Commonly used varicap circuit symbols are shown in figure 6.4.

Probably the most important parameter of varicap diodes is the **useful capacitance range,** which is roughly the range available between the forward conduction and reverse breakdown points. In some devices the useful range may be less than this, because of non-linearity at low and forward bias voltages.

Depending upon the doping levels and doping gradients employed, the useful range of a varicap diode may cover a capacitance ratio of from 4:1 to about 35:1. A typical device has a range of from 260pF—10pF over the reverse bias range 0—10V.

A second important parameter of varicap diode performance is the device **"Q" factor or figure of merit,** which is a measure of the quality or "purity" of the capacitance provided by the diode. As one might expect, the Q factor of a device is inversely proportional to the losses, the main components of which are the effective series resistance and the saturation and leakage currents.

Most varicap devices are fabricated from silicon material, in order to achieve low saturation current levels. Careful control of cleanliness during manufacture is used to ensure that leakage currents are kept to a similarly low level. And by using appropriate doping gradients and construction techniques, the series resistance of the device chips and packages can also be kept to a very low level. As a result, typical modern varicap devices exhibit a Q factor of between 200 and 500 at medium and high frequencies.

At very high frequencies the Q factor of typical devices tends to fall, because series inductance contributed both by the crystal die and its package tends to reduce the effective device capacitance. To minimise this effect, varicaps intended for use at very high frequencies usually employ a very small crystal die mounted in a special low-inductance package.

As the width of the depletion layer of a reverse biased P-N junction is temperature dependent, the capacitance of a varicap is similarly dependent. In applications where this temperature dependence is a problem, zener or forward-biased diodes can be used to produce a compensating opposite temperature variation in the controlling voltage.

Varicap diodes have many applications, some of which are illustrated in basic form in figure 6.5.

Probably the most common use for the devices is to permit remote adjustment of the resonant frequency of tuned circuit, as illustrated by the circuit of figure 6.5(a). Here the varicap effectively forms a parallel tuned circuit with inductor L, the capacitor C being a high value bypass. The resistor R is used for RF isolation of the control voltage, which may be derived and varied at some distance from the tuned circuit using a simple battery and potentiometer combination. Varying the DC control voltage varies the capacitance of the device and hence alters the resonant frequency.

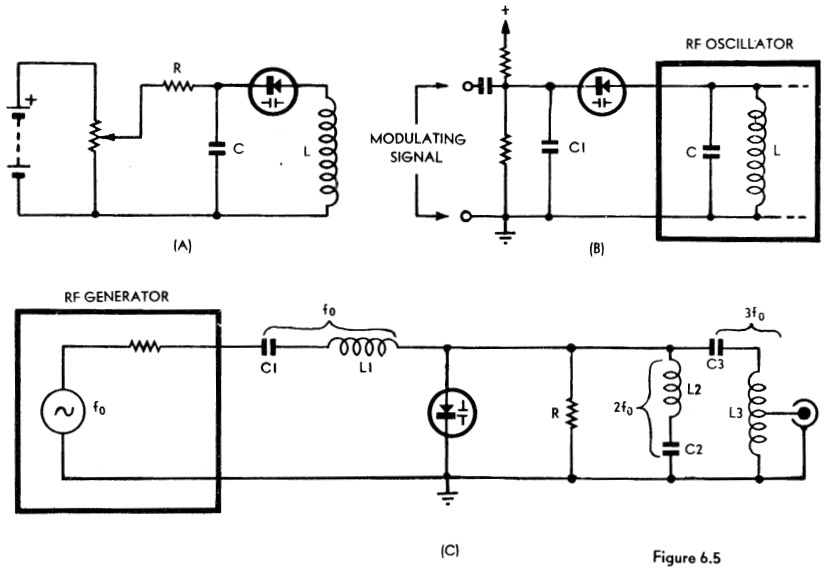


Figure 6.5

A similar configuration is often used for automatic frequency control (AFC) of oscillators. Here the varicap is usually connected not as the sole capacitance of the oscillator tuned circuit, but rather as a "trimming" element. The control voltage fed to the device is not derived from a manually adjustable source, but from a frequency comparator or other circuit used to monitor the oscillator frequency. The arrangement is such that if the oscillator frequency tends to drift away from its correct value, the control voltage fed to the varicap will

change by a suitable amount and in the appropriate direction to correct the tendency.

A related use for varicaps is in circuits designed for frequency modulation (FM) of RF oscillators. A basic circuit of this type is shown in figure 6.5(b). It may be seen that the device is here connected to the tuned circuit LC of the oscillator, with DC reverse bias applied via a resistive divider to bias it approximately at the midpoint of its capacitance range. The modulating signal is then superimposed on the DC bias, swinging the varicap capacitance above and below its quiescent value. Capacitor C1 acts as an RF bypass only, connecting the varicap effectively across the tuned circuit; its value is chosen to represent a negligibly high impedance at modulation frequencies.

By using a device having a carefully tailored voltage/capacitance law, and with careful circuit design, the swing in capacitance due to the modulation signal can be made to produce a linear swing in the resonant frequency of the oscillator tuned circuit. The oscillator output frequency is accordingly frequency modulated in a suitably faithful fashion.

This type of circuit has been used both as the heart of FM radio transmitters and also as the basis for linearly-swept RF signal generators used for tuning alignment of receivers filters and other RF equipment. In the latter case the modulating signal used to swing the oscillator frequency is generally a very low frequency sawtooth waveform, or alternatively a very low frequency sinewave, typically at just a few Hertz.

A class of varicap applications somewhat different from those illustrated in figure 6.5(a) and (b) are those in which the voltage/capacitance characteristic of the device is used, not to allow "external" variation of the capacitance present in a tuned circuit, but to allow the device to be used as a **non-linear reactance.** In this type of application the signals presented to the diode are deliberately made large enough to cause its capacitance to vary significantly during the signal cycle.

Devices used for this type of application are generally somewhat larger than those intended for the former class of application, and are expected to withstand somewhat higher voltage and current levels without damage. The term "varactor" is often used to distinguish them from the lower power devices. An example of a varactor device application is given in figure 6.5(c), which shows a passive frequency multiplication circuit. Such circuits are coming into common use at very-high and ultra-high frequencies, as they offer a simple, convenient and economical means of generating useful power levels at frequencies above those at which other devices operate at peak efficiency.

Basically this type of circuit relies upon the fact that the non-linear reactance of the varactor distorts the input signal, and thus generates strong harmonic components. A tuned circuit is then used to select the desired harmonic, which becomes the output signal. Because the harmonic generation is produced by a varying reactance, which is ideally a lossless circuit element, the conversion efficiency of such a multiplier tends to be quite high — in the order of 75 per cent, with modern devices.

Typical varactor diodes have a voltage/capacitance law which causes the even harmonics of the input signal to predominate. Thus the efficiency of varactors multipliers tends to be highest when they are used for frequency doubling, quadrupling, and so on. However, odd harmonic multiplication can be performed by using a circuit configuration which forces the diode to first generate a carefully chosen even harmonic near the desired odd multiple, and then act as a mixer to produce the desired output by heterodyning with the fundamental input.

The latter circuit is in fact illustrated by the circuit of figure 6.5(c), which shows a basic frequency tripler.

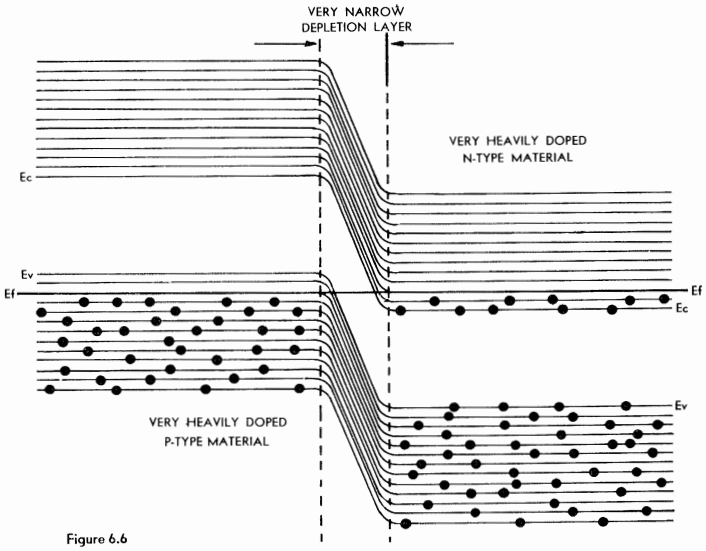


Figure 6.6

Here an input series tuned circuit L1-C1, tuned to the fundamental frequency Fo, is used to match the input RF generator to the very low impedance presented by the varactor. A second tuned circuit L2-C2 forms an "idler" circuit, being tuned to the second harmonic 2Fo and designed to ensure that a heavy current of this frequency flows through the diode in addition to the fundamental. The non-linearity of the device then causes the two to mix together, producing the third harmonic 3Fo, and this is selected by the third tuned circuit L3-C3 which transfers it as output signal into the load circuit. Resistor R is used as a DC return to permit the varactor to develop self-bias by conduction on signal peaks, in conjunction with C1, C2 and C3.

Using modern silicon varactor devices in this type of circuit, tripler efficiencies approaching 70 per cent can be achieved with careful design. A representative device is capable of delivering 27W into a well-matched load when driven by 40W input, and when tripling from 150MHz to 450MHz.

A further important use for varactor diodes is the parametric amplification of very weak RF signals, especially at ultra-high frequencies. Here, the non-linear reactance of the device is used to amplify the signals; while so doing, it contributes very little to the noise level because, as a reactance, it is ideally incapable of generating noise. In practice, some noise tends to be introduced by leakage currents and inevitable device and circuit resistances but, nevertheless, parametric amplifiers using varactors are capable of very low noise operation at extremely high frequencies.

The zener diode and the varicap-varactor diode are probably the most commonly encountered "special" semiconductor diodes. However these are by no means the only types which have been developed. The remainder of this chapter will accordingly be devoted to a brief look at some of the many other types of specialised diode device, and at their applications.

Tunnel diodes are diodes in which the semiconductor material forming the P-N junction is so heavily doped with impurities that the atoms of the impurity elements are sufficiently close together to be no longer isolated from one another. As a result the impurity-derived carriers no longer occupy in the ground state single donor and acceptor energy levels in the forbidden energy gap (figures 3.3, 3.6), but rather two multi-level bands which in fact extend to and blend with the valence and conduction bands of the host material. This may be seen in the diagram of figure 6.6.

Because of the blending of these "impurity bands" into the host material valence and conduction bands, there are in such material, in the ground state, effectively **filled** energy levels in the N - type material conduction band, and similarly there are effectively **empty** energy levels in the valence band of the P-type material. Because of this the Fermi level in the N-type material actually passes through the new widened conduction band, while that of the P-type material passes through the widened valence band. Consequently when a P-N junction is formed in such material it has the rather unique equilibrium energy diagram shown in figure 6.6.

There are two important things to note about this diagram. The first is that because of the heavy doping levels, the depletion layer is extremely narrow — typically .01uM or less. The second thing to note is that in both materials there are empty energy levels immediately above the highest filled energy levels. Because of this both materials are capable of exhibiting virtually **metallic** conduction, and hence have an extremely low resistivity.

The very narrow depletion layer and correspondingly abrupt potential barrier of this type of junction are of paramount importance, because it so happens that when two conducting regions are separated by an exceedingly narrow barrier, electrons are apparently able to transfer from one side to the other virtually instantaneously, and without having previously acquired the energy necessary to surmount the barrier in the usual way.

The mechanism responsible for this rather surprising behaviour is as yet imperfectly understood, although it can be accommodated by the rather abstract concepts of quantum mechanics. Because the effect is almost as if the electrons had "tunnelled through" the barrier, it has been given the name **electron tunnelling,** and hence the name "tunnel diode" used to describe a device which exploits the effect.

Because of the tunnelling effect, a device with the energy diagram of figure 6.6 conducts heavily if small voltages are applied to it, **in either direction.** If forward bias is applied, this results in the lifting of the occupied energy levels in the conduction band of the N-type material so that they become opposite the vacant energy levels in the valence band of the P-type material. A flood of electrons is thus able to tunnel through the depletion layer from N-type to P-type, and the current rises rapidly.

Conversely if reverse bias is applied, this effectively raises the occupied energy levels in the valence band of the P-type material so that in this case it is they which become opposite vacant energy levels in the conduction band of the N-type material. This again allows a flood of electrons to tunnel through the depletion layer, but in this case they flow from the P-type material to the N-type material. Once again the current rises very rapidly with applied voltage.

If the applied bias voltage is increased in the **forward** direction, the current passed by the device is found to reach a peak value and then decrease with increasing voltage—exhibiting a negative resistance characteristic. The reason for this is that as the occupied conduction band levels in the N-type material are effectively raised further, they are eventually raised beyond the level of the vacant levels in the P-type valence band and become opposite the forbidden gap in the material.

As this occurs, the current drawn by the device falls to a minimum. Then it eventually begins to rise again due to normal carrier diffusion from the occupied N-type conduction band to the vacant P-type conduction band. This becomes possible as the former band finally approaches the latter.

In the reverse bias direction, this type of action does not occur, as filled energy levels in the P-type material valence band are simply raised to become opposite a greater and greater number of empty levels in the N-type material conduction band. The current drawn by the device thus continues to rise steeply.

The net result of the foregoing is that a tunnel diode has the rather unique voltage-current characteristic shown in figure 6.7. In the reverse bias direction, it presents a very low and almost linear (or "Ohmic") resistance, while in the forward bias direction it first presents a very low resistance, then a negative resistance, and finally a roughly exponential characteristic similar to that of a conventional diode.

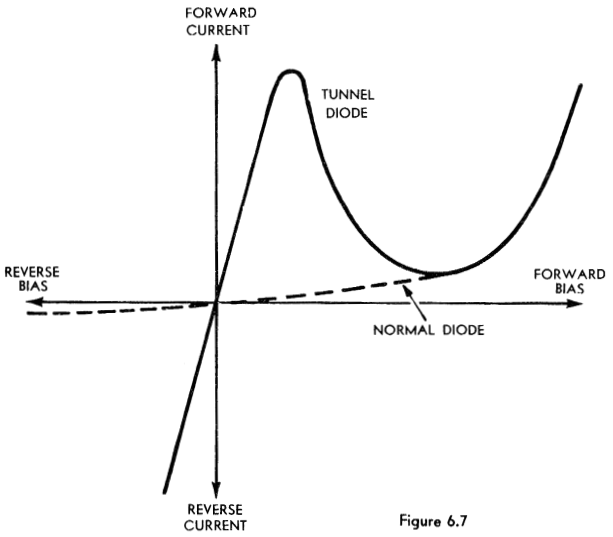


Figure 6.7

Note that in the foregoing description of tunnel diode operation we have spoken only of electron carriers. In fact, these are the only carriers involved in tunnel diode operation because, in the partially filled energy bands of such highly doped material, the concept of a hole has little meaning.

Many of the applications of tunnel diodes are designed to exploit the negative resistance behaviour which they exhibit between the "peak" and "valley" of the forward bias characteristic. By suitable biasing, and in appropriate circuitry, a tunnel diode can be arranged so that its negative resistance either amplifies small signals, or cancels the losses in a resonant circuit to produce continuous oscillation. Both these functions can be performed at extremely high frequencies.

Back diodes or "tunnel rectifiers" are closely related to tunnel diodes, differing only in that the doping levels and gradients employed are arranged to produce a negligible current peak in the forward characteristic. The device still retains the extremely high reverse-bias conductivity of the tunnel diode, however, so that for small signals it presents a very low resistance in the reverse direction and a relatively high resistance in the forward direction.

Because of this, the back diode actually provides a much closer approximation to an "ideal" diode than does any other device — for **small signal** excursions. The only catch is that it provides these desirable characteristics in "reverse" so that, for convenience, the the concepts of "forward bias" and "reverse bias" are applied in the opposite sense to normal: the N-type material becomes the "anode," and the P-type material the "cathode." The device characteristic then becomes that shown in figure 6.8, drawn for comparison on the same axes as the characteristic of an "orthodox" diode.

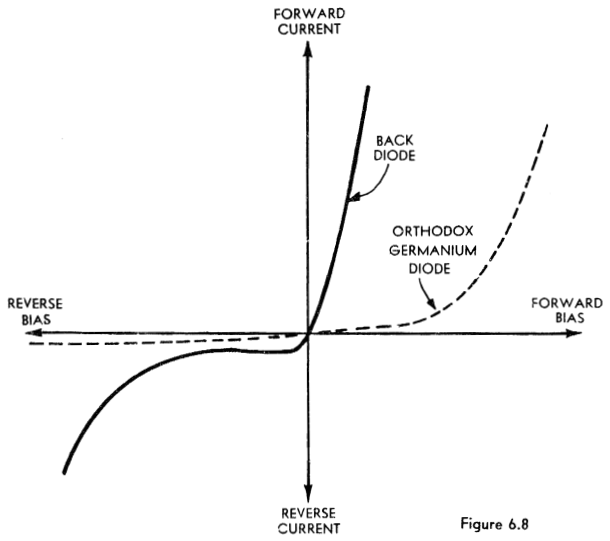


Figure 6.8

It may be seen that for small excursions either side of the equilibrium or zero bias condition, the back diode characteristic is somewhat closer to the ideal than the orthodox diode characteristic. For this reason back diodes find extensive use as rectifiers and detectors for very low amplitude signals, particularly at ultra-high frequencies.

The most commonly used circuit symbols for tunnel diodes and back diodes are shown in figure 6.9. Both types of device are normally represented by the same symbol, which may be reversed or otherwise modified in the case of the back diode.

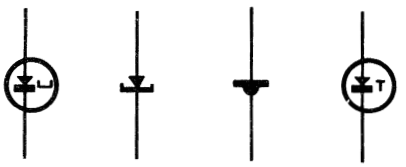


Figure 6.9

Junction photocells are P-N diodes constructed in such a fashion that light radiation may easily be allowed to

illuminate the depletion layer region. When this occurs electron-hole carrier pairs are created in the region by the incident light photons, and these light-produced carriers are swept in either direction respectively by the depletion layer field. The result is that the drift current of the junction exceeds the diffusion current, and equilibrium is disturbed.

One effect of this change is to cause a net EMF to appear across the terminals of the device, with the P-type material becoming positive because of surplus holes, and the N-type material becoming negative because of surplus electrons. A junction photocell may thus be used as a converter of light energy into electrical energy, and when used in this fashion it is usually called a **photovoltaic diode.**

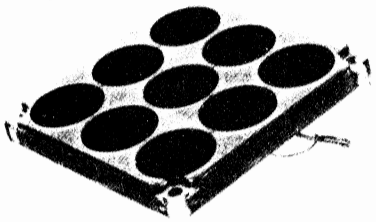
Arrays of large photovoltaic diodes are used to convert solar radiation energy into electrical energy to power electronic equipment. Such arrays are often called **solar cells.** Both photovoltaic diodes and solar cells are usually made from silicon material, as the wide forbidden energy gap of this material provides a higher output voltage than most other semiconductors.

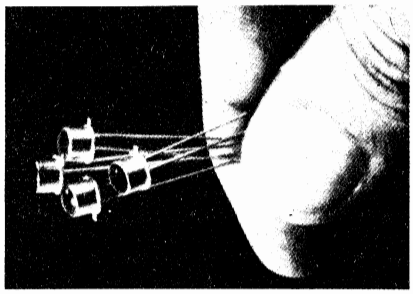
A second consequence of the change in equilibrium of a junction photocell, when it is illuminated, is that the conductivity of the device falls. Thus if such a diode is connected to a source of reverse bias, its reverse current will vary in direct proportion to the incident radiation. Diodes designed to be used in this way are usually called **photo-resistive diodes,** and find application in light-sensing situations such as punched-tape and punched-card scanning.

Light-emitting diodes or "LED's" may be regarded as devices which operate in opposite way to junction photocells. These devices are designed so that they can be operated at very high forward conduction current densities, in which condition large numbers of holes and electrons recombine in the depletion layer region to produce significant light radiation.

Light emitting diodes are used as sources in optical communications systems, as highly rugged and reliable "solid state lamps," and as the heart of compact and efficient numerical readout displays. Specially designed devices of this type may be operated at very high current densities, sufficient to create the conditions necessary for the production of coherent light output. Such devices are then called **junction or injection lasers.**

There are many other types of "special" semiconductor diode, including devices which employ a structure rather more complex than the simple P-N junction. Some of these are designed to act as very high-speed switches, or as variable resistors, or as specialised waveform shaping or frequency mixing elements. Yet another type is used as a magnetic field detector. Unfortunately space restrictions will not allow more than this brief acknowledgment of the existence of these devices here, and interested readers are referred to some of the references listed below.





At left is a small array of silicon solar cells, each measuring about one inch in diameter. Such arrays are used as energy sources for low-power electronic equipment, both on the earth and in space craft. At right are compact light-emitting diodes, for use as high reliability "solid state" lamps. (Courtesy N.E.T. Pty.Ltd., Hewlett-Packard Aust.)

SUGGESTED FURTHER READING

BURFORD, W. B., and VERNER, H. G., **Semiconductor Junctions and Devices,** 1965. McGraw-Hill Book Company, Inc., New York.

BRAZEE, J. G., **Semiconductor and Tube Electronics,** 1968. Holt, Rinehart and Winston, Inc., New York.

EVANS, J. P. (Ed.), **Voltage Regulator (Zener) Diodes,** 1966. Mullard Limited, London.

IVANOV, S. N., et al., **Physics of Microwave Semiconductor Diodes,** 1969. Iliffe Books Ltd., London.

MORANT, M. J., **Introduction to Semiconductor Devices,** 1964. George G. Harrap and Company, London.

ROWE, J., "Understanding Tunnel Diodes," in **Radio, Television and Hobbies,** V.22, No. 11, February, 1961.

————, "The Junction Laser," in **Electronics Australia,** V.28, No. 12, March, 1967.

SURINA, T., and HERRICK, C., **Semiconductor Electronics,** 1964. Holt, Rinehart and Winston, Inc., New York.

Also "Solid State Diodes," a special section in **Electronics World,** V.82. No. 1, July, 1969.

Chapter 7

THE UNIJUNCTION

The unijunction — basic construction — interbase current — intrinsic standoff ratio — the peak point — carrier injection — conductivity modulation — negative resistance behaviour—the valley point—static emitter characteristics — base current modulation — field effect — static interbase characteristics — temperature stabilisation — applications.

In the preceding chapters we have examined fairly carefully the operating principles and applications of the many varieties of P-N junction diode, which may be regarded for many purposes as the most basic type of semiconductor device in common use. Using the knowledge gained in these chapters as background, let us now turn our attention to a slightly more complex device: **the unijunction.**

The unijunction is quite a logical choice as the device type next examined after the basic P-N diode in a systematic treatment of semiconductor devices. It is probably the simplest of the three electrode devices and the device whose close relation to, and evolution from, the basic diode is most readily appreciated. Also an understanding of its operation involves important concepts, which are among those involved in understanding the more complex devices, so that a discussion of the device may provide a useful conceptual stepping-stone.

Although it is essentially a simple development from the basic P-N diode, the unijunction is capable of performing many other rather unique functions. It can form the basis of very simple relaxation oscillators, timers, threshold detectors, pulse generators and amplifiers, counters and information storage cells. Because of its flexibility it has found considerable use in electronic equipment of recent design, and particularly in pulse-handling and control equipment.

Other names for the unijunction are "unijunction transistor," "UJT" and "double-base diode." The latter name was that first given to the device when it was developed in 1953 at the Syracuse, New York laboratories of the General Electric Company.

Essentially a unijunction consists of a single P-N junction which differs from a normal semiconductor diode in that the material on one side of the junction is provided with not one, but two connection electrodes. This side of the junction is called the **base,** and its two electrodes are conventionally labelled the "base-1" (B1) and "base-2" (B2) electrodes. The material on the other side of the junction is called the **emitter,** and is provided with a single "emitter" (E) electrode.

At first sight it may seem rather surprising that a distinctly different and independently useful new semiconductor device may be developed from the basic P-N diode, not by radical re-arrangement of the junction, or by the addition of further junctions, but rather by the fairly straightforward addition of a second connection to one of its two semiconductor regions. Yet in basic terms this is really all that the unijunction involves. The fact is that

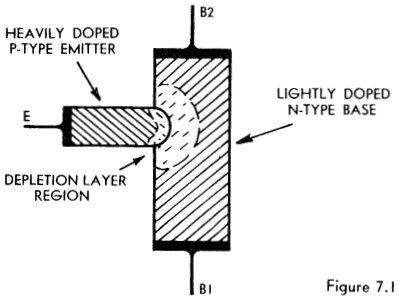


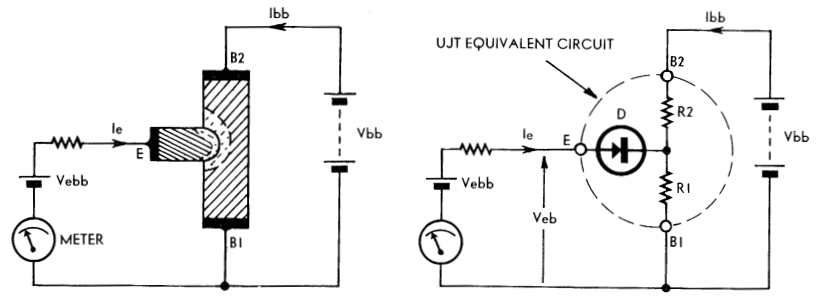
Figure 7.1

the second electrode attached to its base region allows the effective conduction characteristics of the device junction to be varied considerably from those of a normal diode, as this chapter seeks to demonstrate.

In theory, unijunctions may be made from both silicon and germanium, and with either the P-type emitter/N-type base configuration or its converse. The first devices to be produced were made from germanium, but the high minority carrier saturation currents of this material placed severe limits on device performance and stability. As a result, unijunctions are now made almost exclusively from silicon. Also, because devices with the N-type emitter/P-type base configuration present rather difficult manufacturing problems, the P-type emitter/N-type base version has become that most widely used.

The basic form taken by most unijunctions is shown in figure 7.1. The lightly doped N-type base material is usually in the form of a rectangular bar or cube, to which non-rectifying or "ohmic" connections are made at opposite ends to form the B1 and B2 electrodes. At a point between these two electrodes a junction is formed with the heavily doped P-type emitter material, with a third ohmic connection made to the remote end of this material for the E electrode. The junction is normally somewhat closer to the B2 electrode than to the B1 electrode.

Not surprisingly, under equilibrium conditions the junction of such a device behaves in exactly the same manner as that of a normal P-N diode which we examined in previous chapters. Majority carrier diffusion takes place over the junction, a drift field is set up, and a depletion layer appears in the material on either side of the junction proper. Naturally the depletion layer will extend further into the lightly doped base than into the heavily doped emitter, as suggested in the diagram.



EMITTER CONDUCTS WHEN $V_{eb} \geq V_d + \eta . V_{bb}$
WHERE $V_d \doteq 0.6V$ (NORMAL FORWARD VOLTAGE DROP OF P-N JUNCTION)
η = "INTRINSIC STANDOFF RATIO" = R1/(R1 + R2)
R1 + R2 = TOTAL INTERBASE RESISTANCE (Rbb)

Figure 7.2

If external bias is applied between the emitter and **either** of the two base electrodes, with the other base electrode left unconnected, the device will again behave exactly as a normal diode. Under forward bias (emitter positive) the device will conduct heavily as soon as the applied voltage is sufficient to produce significantly excess majority carrier diffusion currents —i.e., when the applied voltage exceeds about 0.6V, assuming silicon material.

Conversely under reverse bias (emitter negative) the device will draw only a small and almost constant current, composed primarily of the minority carrier saturation currents.

If one of the base electrodes is ignored, then, the unijunction behaves simply as a normal P-N diode. However by connecting both base electrodes into a circuit in a suitable manner this behaviour can be made to change markedly.

Typically, the circuit into which a unijunction is connected is arranged to apply a bias voltage between the two base electrodes, in addition to any bias which may be applied between emitter and base. The bias polarity is normally such that the B2 electrode is positive with respect to B1.

Impurity semiconductor material is capable of significant electrical conduction even at low excitation levels, it may be remembered, the resistivity being inversely proportional to the impurity doping level. Hence the base region of a unijunction, being composed of lightly doped and therefore fairly high resistivity N-type material, will possess a finite though moderately high resistance. This is normally termed the **interbase resistance,** symbolised Rbb. Typical values for Rbb range between 5K and 10K.

When bias voltage is applied to a unijunction between the B1 and B2 electrodes a small but significant **interbase current** thus flows, as a result of the finite interbase resistance.

Just as with any other resistor passing current, the base region of the device will have a distributed voltage drop. Any arbitrary point between the B1 and B2 electrodes will therefore possess a certain electrical potential with respect to each, that with respect to B1 being positive and that with respect to B2 negative. The magnitude of these potentials will depend upon the position of the chosen point along the electrical length of the base region.

The emitter junction, being placed at such a point on the base between the two end electrodes, will therefore possess such potentials. As typical devices have the junction closer to the B2 end of the base, this means that the positive potential of the junction with respect to the B1 end will be somewhat larger than the negative potential with respect to B2.

What does this imply? Simply that, if the B1 electrode is taken as reference, the current flowing through the base region between the two end electrodes effectively provides the emitter junction with an "internal" reverse bias. Even if the emitter electrode were shorted externally to B1, the junction would still have an applied (reverse) bias equal to the voltage drop in that section of the base between the junction and the B1 electrode.

Accordingly if an external forward bias is connected to the unijunction between emitter and B1, its magnitude must be increased to a level somewhat higher than for a normal diode junction before significant current flows. This, then, is the first important way in which the behaviour of a unijunction differs from that of a normal diode: the effective "turn-on" voltage of the emitter junction may be controlled by means of a voltage applied between the B2 and B1 electrodes.

Illustration of this behaviour is given in figure 7.2. In the left-hand diagram is shown a unijunction to which has been connected a bias voltage Vbb between the two base electrodes, resulting in an interbase current Ibb. If an adjustable source of emitter-B1 forward bias Vebb is connected in series with a suitable meter between the E and B1 electrodes, as shown, it will be found that the actual emitter-base volatge Veb must be increased to a value somewhat higher than the usual 0.6V or so, before significant emitter current Ie flows.

As the right-hand diagram of the figure shows, this behaviour of the unijunction allows us to draw a simple "equivalent circuit" for the device. The equivalent circuit consists of a diode D representing the emitter P-N junction itself, together with two resistors R1 and R2 representing the resistances of the base region between the junction and either end.

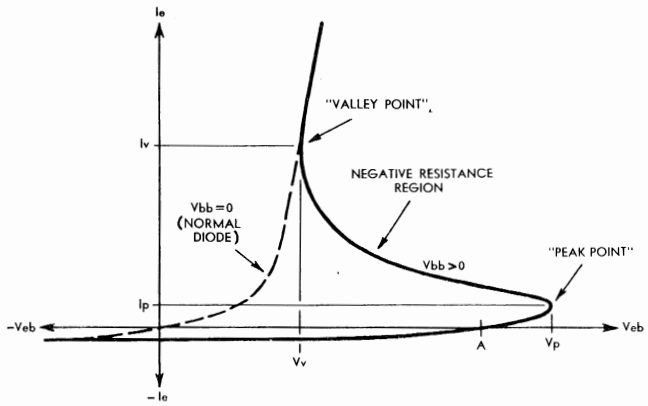


Figure 7.3

The reason for deriving the equivalent circuit for the device is that it enables us to formulate a simple expression for the emitter voltage required for conduction. It should be fairly clear from the right-hand diagram of figure 7.2 that conduction will occur only when Veb is increased to a level where it exceeds the sum of Vd, the "turn-on" voltage of the junction, together with the proportion R1/(R1 +R2) of Vbb.

As may be seen, the ratio R1/(R1 +R2) is known as the **intrinsic standoff ratio** of the unijunction, commonly represented by the Greek symbol Eta. As the intrinsic standoff ratio determines the proportion of the interbase bias Vbb which acts as "internal" reverse junction bias, and accordingly plays a major part in determining the conduction point of the emitter in a given circuit, it is an important unijunction parameter.

The inherent junction turn-on voltage Vd and the actual values of the interbase resistors R1 and R2 are all subject to variation between individual unijunction devices, being dependent upon doping levels and physical dimensions. However because such factors tend to influence both the R1 and R2 components of the interbase resistance equally, the intrinsic standoff ratio tends to be fairly constant for a given device type. Typical devices have an intrinsic standoff ratio of about 0.7, but special devices are made with values both higher and lower than this figure.

For a device with a certain intrinsic standoff ratio, it should be fairly apparent that the emitter-B1 voltage at which the emitter junction conducts—called the **peak point voltage** (Vp)—may be controlled by varying the interbase bias voltage Vbb. The higher Vbb, the higher the reverse bias effectively applied "internally" to the base side of the junction, and the higher Vp.

One important difference between the unijunction and a normal diode, then, is that its peak point voltage or emitter "turn-on" voltage may be electrically varied. However, this is not the only important difference between the two types of device, for other unique aspects of unijunction behaviour appear as soon as emitter current flows.

It may be remembered that the composition of the current passing across a forward biased junction depends upon the impurity doping concentrations of the P-type and N-type regions involved. If one of the regions has a higher doping concentration than the other, then quite naturally the junction current will consist mainly of the majority carriers appropriate to that material.

When emitter junction current flows in a unijunction it therefore consists mainly of valence band holes moving from the heavily doped emitter region to the lightly doped base region. Only a relatively small proportion of the total forward bias current consists of conduction band electrons moving in the reverse direction, because of the relatively low impurity doping concentration in the base material.

Often this situation is described by referring to the unijunction as a device wherein the doping levels are arranged to result in a high "emitter injection ratio." The latter term describes the proportion of total junction current formed by emitter-region majority carriers (holes) effectively **injected** as minority carriers into the base region.

Because of the high emitter injection ratio of the unijunction, then, the main result of the flow of emitter current is that a large number of holes are injected as minority carries into the base region from the emitter. The base therefore finds itself with an excess of holes in the vicinity of the emitter junction.

The holes ejected from the emitter leave that region with a nett negative charge. Accordingly, an appropriate number of electrons are repelled from the emitter via the E electrode, forming the emitter current Ie. Similarly, the excess holes injected into the base region give that region a nett positive charge, and this in turn causes an appropriate number of conduction band

electrons to be "sucked into" the device at the B1 electrode.

Because of the electric field present in the base region due to the inter-base bias Vbb, the holes injected into this region from the emitter drift toward the B1 end of the device. Similarly the electrons which enter the base at the B1 end to maintain neutrality drift in the opposite direction towards B2. The result is that the section of the base region between the junction and B1 finds itself with a high excess concentration of both minority carriers (holes) and majority carriers (electrons).

The presence of the excess carriers in this portion of the base region effectively lowers its resistivity, by providing a supply of current carriers additional to the relatively small number initially present in the lightly doped base material. In other words, the injected carriers cause the junction -B1 section of the base to behave temporarily as if it had been more heavily doped. This phenomenon is often referred to as **conductivity modulation.**

The result of the drop in base region resistivity is that there is actually a **decrease** in the reverse bias applied to the emitter junction "internally" via divider action from Vbb. In effect, resistor R1 in the unijunction equivalent circuit of figure 7.2 has been lowered in value, taking with it the proportion of Vbb appearing as reverse junction bias. This is despite the fact that additional current is flowing through R1, due to the emitter current Ie.

As the external emitter-B1 voltage drop of the device (Veb) is simply the sum of the voltage drops of the junction itself (Vd) and of the R1 portion of the base, this means that the decrease in the latter will cause Veb to similarly decrease. And if the emitter current Ie is allowed to increase from its initial value, the two voltages will decrease even further.

This is rather unusual behaviour, as the observant reader will no doubt have realised. Normally, when the current passing through a circuit element is increased, its voltage drop also increases; but here we have a situation where an increase in current results in a **decrease** in voltage drop. In short, we have an effective **negative resistance,** as we had with the tunnel diode.

Not only does the emitter-B1 circuit of a unijunction possess an adjustable "turn-on" point, then, but it also behaves as a negative resistance as soon as emitter current begins to flow. This behaviour is illustrated in the diagram of figure 7.3, which shows for comparison the effective emitter junction forward bias characteristics for both the Vbb=0 case, where the behaviour is virtually identical with a normal diode, and the case where Vbb has some definite value.

It may be seen that when there is an applied Vbb, the emitter current remains at a very low level for applied forward bias levels considerably higher than those necessary when Vbb=0. This is due to the "internal" reverse bias applied to the junction, as we have seen. The junction does not actually reach the equilibrium of "zero bias" condition until point "A" is reached, and accordingly until this point is approached it draws only the usual reverse bias current composed mainly of minority carrier saturation currents.

As the applied emitter voltage is increased to reach and exceed the level corresponding to point "A," majority carrier diffusion currents gradually appear and the junction current begins to rise. The junction then enters conduction, and the so-called **peak point** is reached.

The junction voltage drop at this point is called the "peak point voltage" (Vp), as we saw earlier, while the corresponding current is naturally called the "peak point current" (Ip).

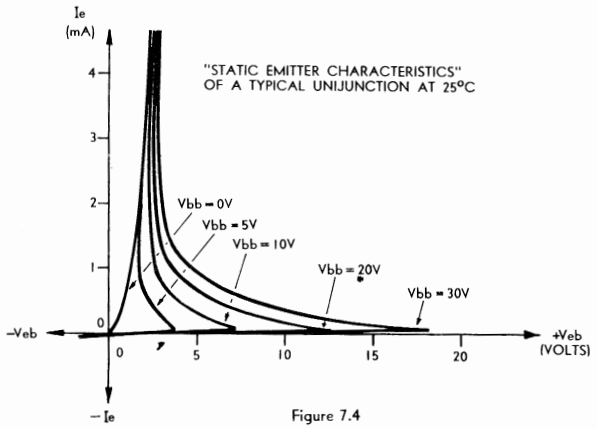


Figure 7.4

With typical devices Ip has a value in the order of 2uA.

If the junction current is allowed to increase beyond its value at the peak point, it may be seen that the effective junction voltage drops away; in other words, the device enters its negative resistance region. In this region the voltage continues to fall with rising current, as the resistivity of the emitter-B1 section of the base is falling at a faster rate than the increase in current.

Eventually, if the current continues to rise, the resistivity of the base region does not continue falling, but "flattens out" at a low **saturation level.** This occurs when the concentration of excess carriers in the base reaches such a level that further injected carriers merely result in increased carrier recombination, and do not effectively contribute to current conduction.

When this occurs the effective voltage drop of the emitter junction reaches a broad minimum, and then begins to rise again. The minimum is normally referred to as the **valley point,** as may be seen, and the corresponding voltage and current values as the "valley point voltage" and "valley point current" respectively. It may be seen that the emitter characteristic of the unijunction at current levels above the valley point is substantially the same as that of a normal diode, or that of the device itself for Vbb=0.

As we have seen, the peak point or effective emitter junction "turn-on" point is not fixed, but is controlled by the interbase bias Vbb. Hence the solid curve of figure 7.3 does not represent a single and fixed emitter characteristic, but in fact a whole "family" of characteristic curves, each corresponding to a different value of Vbb. The dashed Vbb=0 curve will represent the "limiting case" of the family.

Figure 7.4 shows such a family of **static emitter characteristic** curves, the values given being those for a typical general-purpose unijunction device.

It is mainly by virtue of the fact that the emitter junction of a unijunction is capable of behaving as a negative resistance over portion of its characteristic that the device is able to perform many of its unique circuit functions, as will be shown shortly. However before we progress to consider the applications of the device in practical circuitry, there are further aspects of its basic operation which should be briefly examined.

The reader may have noticed that in the foregoing discussion of unijunction conduction, reference was made only to the behaviour of the voltages and currents associated with the emitter. It may have been assumed from this that the interbase bias current Ibb of the device was unaffected by the mechanisms involved; however this is not the case.

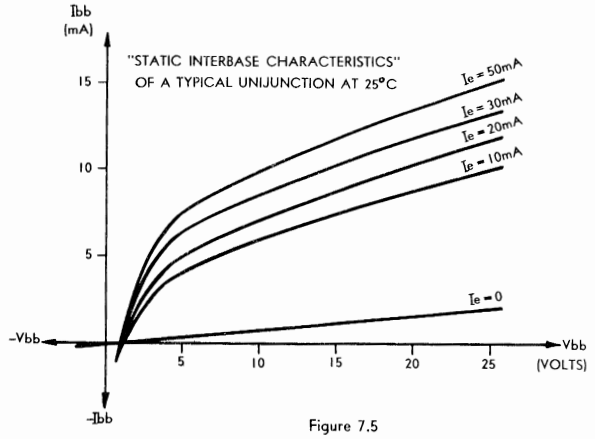


Figure 7.5

When emitter current begins to flow, the interbase current is found to increase to a small but significant extent. This is partly due, as one might expect, to the drop in resistivity of the lower portion of the base as a result of the injected minority carriers from the emitter. However, it is also partly a result of a separate conductivity modulation mechanism associated with the depletion layer surrounding the emitter junction.

When the emitter junction is reverse biased, i.e., when there is low external emitter voltage Veb relative to the

internal reverse bias, its depletion layer naturally extends to a significant extent into the material on either side. It tends to extend further into the base region, because of the lighter doping and higher resistivity of that material, and also in the base material itself it tends to extend further at the "top" or B2 side of the junction than at the "lower" or B1 side. This is because an electric field exists in the base due to Vbb, and the effective reverse bias is accordingly slightly greater at the B2 side of the junction than at the B1 side. (The shape of the depletion layer may be seen by reference back to figure 7.2)

A depletion layer, it may be remembered, is a region in a semiconductor which has been virtually stripped of available current carriers. As such, it is an effectively "intrinsic" region, capable of displaying only the rather poor conductivity of intrinsic semiconductor material. In short, it is a region effectively "converted" into very high resistivity material.

Prior to junction conduction in a unijunction, therefore, the base region of the device consists in part of effectively very high resistivity material — material considerably higher in resistivity than the remainder of the lightly doped base region. In effect, the actively conducting cross-section of the base material is virtually narrowed or "pinched" in the vicinity of the junction, as a result of the encroachment of the depletion layer.

When the emitter junction enters conduction, the depletion layer naturally contracts to correspond to the reduced potential barrier. The "pinching" of the base region is therefore reduced, and the actively conducting cross-section of the region widens. As a result of this widening the interbase resistance falls, and the current Ibb rises as a result.

It may be seen that this second mechanism responsible for the conductivity modulation of the unijunction base region by the emitter current is quite different from the minority carrier injection mechanism described earlier. It is in fact an example of a **field effect** mechanism, an important type of mechanism which is exploited in and forms the basis of a number of very useful semiconductor devices. The best-known example of these is the field-effect transistor, which the reader will meet in the next chapter.

Because the depletion layer of a unijunction emitter junction is basically associated with the potential barrier actually present across the junction, it is influenced both by the emitter voltage Ve and by the interbase bias Vbb — the latter not directly, but proportionally via the intrinsic standoff ratio. The interbase bias Vbb thus plays a part in determining the width of the depletion layer, and hence by means of the field effect mechanism it also influences the effective cross-section and conductivity of the base.

The interbase resistance of a unijunction is thus found to vary with the applied interbase bias voltage Vbb, an increase in Vbb causing a small but sometimes significant rise in interbase resistance from its initial value of Rbb. This effect is in itself quite distinct from those associated with emitter current flow, although when appreciable emitter current is flowing the narrowness of the depletion layer causes the effect to be somewhat reduced. Naturally the fact that the interbase resistance of a unijunction varies with Vbb tends to provide yet another source of variation in the interbase current Ibb.

From the foregoing it may be seen that the interbase current Ibb of a uninjunction is a rather complex function of both the emitter current Ie and the interbase bias voltage Vbb. It is usual to describe the relationship between Ibb, Vbb and Ie graphically, by means of the so-called **"static interbase characteristics."**

The static interbase characteristics of a typical general-purpose unijunction are shown in figure 7.5. As may be seen they consist, like the static emitter characteristics of figure 7.4, of a "family" of curves. In this case each curve shows the relationship between Ibb and Vbb for a specific value of emitter current Ie.

The lowest or Ie=0 curve shows the relationship between Ibb and Vbb when the unijunction is cut off — i.e., the initial slope of this curve represents the "nominal" interbase resistance Rbb. The remaining curves show how the interbase current increases moderately with increasing emitter current.

The circuit symbols usually employed for unijunctions are shown in figure 7.6. Note that the arrowhead on the emitter lead is used to symbolise the direction of forward emitter current flow according to the classical "positive charge" convention.

Being composed of impurity semiconductor material, the base region of a unijunction possesses a small but significant positive temperature coefficient of resistance at normal temperatures. It may be recalled from chapter 3 that this is due to the fact that once the impurity atoms are all ionised, further increase in excitation merely results in a reduction of carrier mobility, and accordingly a corresponding rise in resistivity.

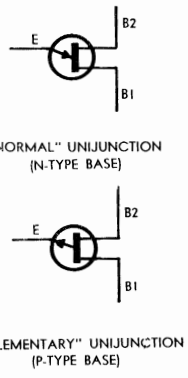


Figure 7.6

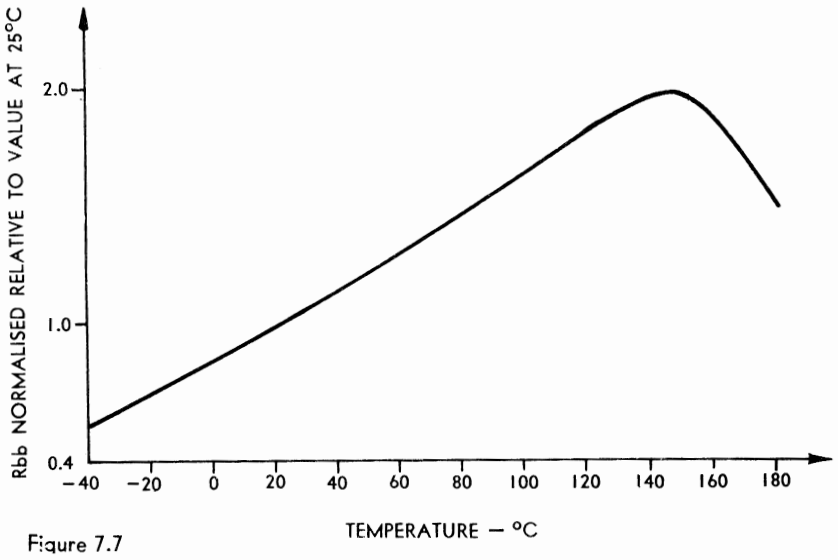


Figure 7.7

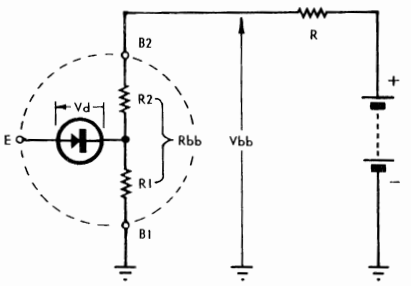


Figure 7.8

For a typical unijunction the interbase resistance Rbb increases linearly from about −40°C to about 150°C, with a temperature coefficient of about 0.8% per degree. This is illustrated in figure 7.7, where it may be seen that the value of Rbb at 150°C is approximately double its value at 25°C. Above 150°C the base resistivity begins to fall rapidly due to the increase in "intrinsic" carrier pairs.

As with a normal P-N junction, the inherent forward bias voltage drop (Vd) of the emitter junction of a unijunction decreases with temperature— i.e., it exhibits a negative temperature coefficient. Less forward bias is required to produce significant forward current at high temperatures than at low temperatures.

It may be seen from the foregoing that the two components of a unijunction most intimately responsible for determining the peak point voltage Vp, namely the junction itself and the base region, have temperature coefficients of **opposite** polarity. This is significant because it provides a means whereby

the peak point voltage may be simply and effectively stabilised over a wide range in temperature.

Figure 7.8 shows how simply peak point stabilisation may be achieved. The technique merely involves the addition of a suitably chosen resistor R in series with the connection between B2 and the interbase bias supply. The resistor and the device interbase resistance Rbb together then form a simple voltage divider.

Because of the positive temperature coefficient of the interbase resistance Rbb, the division ratio of this divider rises with temperature. Hence as the temperature rises the effective interbase bias Vbb rises also, and with it the proportion of Vbb presented to the base side of the junction by the "internal" (R1 + R2) divider.

If the value of resistor R is carefully chosen, the rise in voltage at the base side of the junction may be made almost exactly equal and opposite to the fall in junction voltage drop Vd. The emitter peak point voltage Vp will then remain substantially constant over a wide range in temperature. With typical devices this simple method may be used to stabilise Vp to within approximately .001% per degree up to about 100°C.

The astute reader may well have realised by this stage that the simple equivalent circuit given for the unijunction in figures 7.2 and 7.8 is valid only when the device is not conducting. In fact, the device is rather difficult to represent after conduction, and a complete equivalent circuit tends to be quite complicated.

To conclude this discussion of the unijunction let us now look briefly at some of the many applications of the device.

Probably the most common application of unijunctions is in simple **relaxation oscillators.** These may be used to generate sawtooth-wave and pulse signals over a considerable frequency range, and may also be synchronised to perform low-cost frequency division.

The basic circuit of a unijunction relaxation oscillator is shown in figure 7.9. It may be seen that the emitter electrode is connected to the junction of a capacitor C and a resistor R, which are connected in series across the supply Vbb. The base of the device is also connected across the supply, by means of resistors Ra and Rb. Resistor Rb is used primarily for temperature stabilisation of Vp, as explained earlier; the purpose of Ra should become clear in a moment.

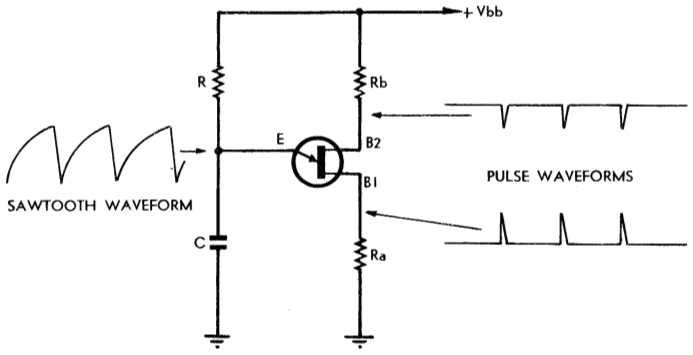


Figure 7.9

When voltage is first applied to such a circuit, the capacitor C is initially uncharged, and thus begins to charge from the supply via resistor R. The emitter voltage of the unijunction accordingly rises from zero in the familiar exponential fashion. Until the emitter voltage rises in this fashion to the device peak point voltage, the emitter itself draws negligible current, and does not significantly influence the charging operation.

As soon as the peak point voltage is reached, however, the emitter draws current, and its input resistance drops sharply through the negative region to the low resistance "saturation" region. This discharges the capacitor rapidly, feeding its stored energy as a pulse of current through resistor Ra.

Discharge current flows until the capacitor voltage drops below the value necessary to sustain the emitter current above the valley point value. The unijunction then turns off again, and the capacitor C begins to recharge via R. The cycle then repeats itself, and will, in fact, continue indefinitely as long as the supply is connected. The time taken for the capacitor voltage to reach the peak point voltage each time is determined both by the capacitor itself and the resistor R, so that the repetition frequency may be altered by varying the value of either of these components.

It may be seen that the circuit has the familiar "charge-discharge" action characteristic of relaxation oscillators. As such, it is very similar in operation to the familiar "gas tube" sawtooth wave generators using either neon lamps or gas-filled thyratron valves.

As shown in figure 7.9, a sawtooth output waveform is available at the emitter of the unijunction, while both positive and negative pulses are available at the B1 and B2 electrodes respectively due to the currents flowing during the discharge part of the cycle.

Naturally the sawtooth at the emitter, being part of an exponential charging waveform, will not be perfectly linear. However there are a number of ways in which the non-linearity may be corrected, many of which involve replacement of the resistor R with a circuit or device which supplies a controlled constant current.

For a particular capacitor value, the frequency range over which this type of oscillator may be varied by variation in the value of resistor R is quite wide, but limited in both directions. If the resistance is made too large, the slight leakage current drawn by the device emitter becomes significant compared with the charging current, and the capacitor will not charge up to the peak point voltage. On the other hand if the resistance is too low, the emitter current will not drop below the valley point current when the device conducts. In either case, oscillation ceases.

These restrictions are not severe, and with typical devices it is possible to achieve reliable operation over a resistance range (and a corresponding frequency range) of 1000:1. The upper limit of oscillation frequency for typical devices is approximately 150KHz.

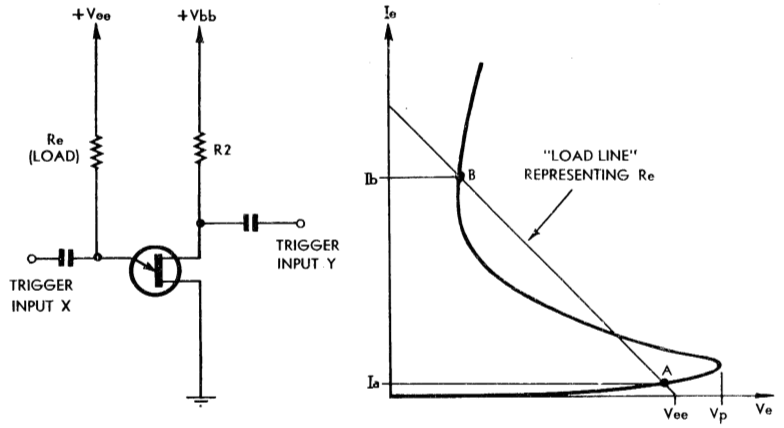


Figure 7 10

The basic unijunction relaxation oscillator of figure 7.9 may be synchronised to an external signal, providing its natural frequency is set to be slightly lower than that desired. Synchronisation is achieved by feeding a negative synchronising pulse to the B2 electrode of the device. The action of the pulse is to momentarily lower the effective interbase bias applied to the unijunction, so that if the capacitor is charged to a voltage even approaching the normal peak point voltage, it will conduct as a result of the temporary lowering of the peak point by the synchronising pulse.

This technique may be used to synchronise a unijunction oscillator at a submultiple of the synchronising frequency. An oscillator operated in this fashion may be used as a simple sweep generator for economy oscilloscopes or television receivers. A number of similar circuits may be cascaded to form a low-cost frequency divider system.

Actually a unijunction oscillator may be triggered into the conduction part of the cycle either by a negative pulse applied to B2, or by a positive pulse superimposed upon the capacitor voltage at the emitter. Either way, somewhat larger pulses than those necessary for triggering appear as output pulses at the B1 and B2 electrodes. Hence the circuit may be used with little modification as a **regenerative pulse amplifier.**

Because the basic unijunction relaxation oscillator may be arranged to oscillate at very low frequencies, it may be used as a **period timer.** Here the positive pulse output at the B1 electrode is normally used, being either amplified and arranged to drive a relay, or used directly to trigger in turn one of the more complex semiconductor switching devices to be described in a later chapter.

Typical unijunction timer circuits may be adjusted to any time period between a small fraction of a millisecond and a few minutes. More complex unijunction timers, still based on the simple circuit of figure 7.9, have been used to produce pulses spaced reliably at periods of up to one hour.

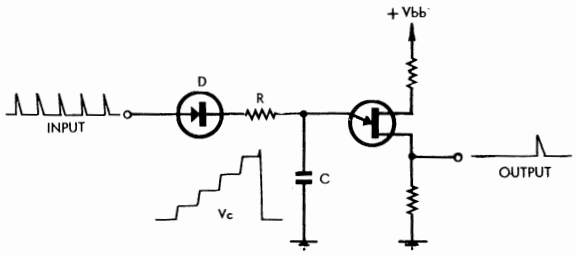


Figure 7.11

In a circuit not very different from that of the basic relaxation oscillator, a unijunction may be arranged to provide a simple **bistable storage element** which is capable of "remembering" the last of two types of switching pulses fed to it. A basic circuit for such a unijunction bistable element is shown in figure 7.10, together with a diagram which may be used to understand its operation.

The emitter of the device is here connected to a second fixed bias source Vee, via a load resistor Re which may in a practical case be a relay coil, or other device used to "read out" the state of the element. As before the B2 electrode is connected to an interbase bias source Vbb via a resistor R2, only in this case R2 is used not so much for temperature stabilisation but mainly as a decoupling resistor for triggering pulses applied to B2.

The emitter supply Vee is set at a value which is slightly lower than the peak point voltage Vp of the device, as determined by its intrinsic standoff ratio and the values of Vbb and R2. The value of Re is then selected such that **two** stable emitter operating points are possible — one on the "cutoff" portion of the unijunction emitter characteristic below the peak point, and the other on the "saturated" portion of the characteristic above the valley point.

These points are indicated in figure 7.10 as "A" and "B," respectively, the straight line joining the two being a "load line" representing the load resistor Re. It may be noted that both A and B correspond to stable operating points, as they are each situated on sections of the emitter characteristic having a "positive resistance" slope. The difference between the two points is that at A the emitter current and hence the load current are but a few microamps, whereas at B they may be in the order of tens of milliamps.

Which of the two operating points applies at any given time depends upon the last triggering pulse fed to the circuit via the triggering inputs "X" and "Y." If the last pulse to arrive was either a positive pulse fed to input X or a negative pulse fed to Y, then the operating point will be "B" as the unijunction will have been switched to the conducting state. Conversely if the last pulse to arrive was a negative pulse fed to input X, the operating point will be "A" because the unijunction will have turned off.

In each case the circuit will remain at point A or B until either the arrival of a pulse of the polarity necessary to switch it to the other operating point, or until power is removed. The circuit thus has the capability of being used for information storage.

A further adaptation of the basic unijunction oscillator is used for **pulse counting.** This is shown in figure 7.11; it may be seen that here the capacitor C is not charged up in a smooth fashion from the supply, but in a "staircase" fashion by individual input pulses applied via the diode D and resistor R.

By suitable choice of R and C, the capacitor voltage may be arranged to reach the unijunction peak point voltage only after the arrival of the last of a given number of input pulses— say five. The circuit will then deliver an output pulse for every five input pulses, and thus forms a simple pulse counter.

Yet another application for the unijunction is shown in figure 7.12. This is of interest because it does not take advantage of the switching or negative resistance aspects of unijunction behaviour, but rather of the fact that the interbase resistance Rbb varies with emitter current.

In effect, the unijunction is here used merely as a controlled-value resistor. Its interbase resistance is arranged to form an AC voltage divider with resistor R1, the divider controlling the proportion of an input signal applied to the input of the AC amplifier. The output of the amplifier is then rectified by diode D, which delivers only the positive half-cycles to capacitor C. The latter then discharges through the unijunction emitter circuit via resistor R2.

The idea is that when the output of the amplifier is low in amplitude, the voltage developed across capacitor C will be lower than the unijunction peak point voltage, and the device will be cut off. Its interbase resistance will accordingly be fairly high (around 8K), and most of the input signal will be fed to the amplifier. However, if the output voltage from the amplifier rises to the point where the capacitor voltage reaches the unijunction peak point, the latter will conduct and its interbase resistance will fall sharply. This will cause a smaller proportion of the input signal to be fed to the amplifier, and will tend to reduce the output.

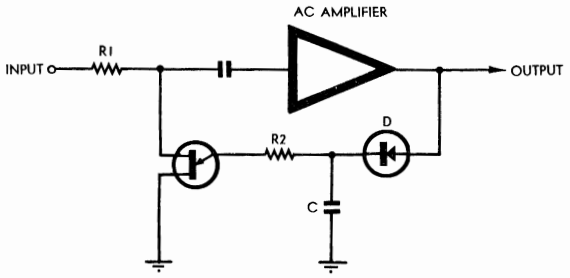


Figure 7.12

The system thus functions as an automatic output level control circuit, also called a **limiter.** As the interbase resistance of a typical unijunction falls to less than 100ohms at an emitter current of about 10mA, such a circuit can cope with a considerable range in input voltage, to maintain the amplifier output voltage substantially constant.

There are many unijunction applications additional to those briefly discussed in the foregoing, and some of these will be found in the literature listed below for suggested further reading. However, the few applications which have been given should help the reader to visualise the flexibility of the unijunction, and the way in which it lends itself to quite diverse applications.

SUGGESTED FURTHER READING

CLEARY, J. F., (Ed.) **General Electric Transistor Manual,** 7th Edition, 1964. General Electric Company, Syracuse, New York.

KYLE, J., "The Ubiquitous Unijunction," in **Electronics Australia,** V.29, No. 12. March 1968.

MILLMAN, J., and TAUB, H., **Pulse, Digital and Switching Waveforms,** 1965. McGraw-Hill Book Company, New York.

SPOFFORD, W. R. Jr., and STASIOR, R. A., "A Switch in Time," in **Electronics,** V.41, No. 4, February 19, 1968.

SURAN, J. J., "Double Base Expands Diode Applications," in **Electronics,** V.28, No. 3, March 1955.

Chapter 8

FIELD-EFFECT TRANSISTORS

Field-effect transistors — the junction FET — operation — cut-off and pinch-off — channel current "plateau" — static drain-source characteristics — triode and pentode operation, depletion and enhancement modes—the transfer characteristic — transconductance — other parameters and characteristics — constant current diodes — insulated gate FETs — the three basic types of MOSFET — the dual-gate MOSFET — insulation breakdown.

In the discussion of unijunction device operation given in the last chapter it was noted that one aspect of device behaviour involved a so-called "field effect" mechanism, in which the effective conductivity of one region of the device was modulated by the width of a depletion layer extending from an adjacent P-N junction. Mention was made of the fact that this type of mechanism is quite important, and that it actually forms the basis of a number of useful semiconductor devices. The best-known of these devices is the **field-effect transistor,** and it is appropriate that we now turn our attention to this device.

Like the unijunction, the field-effect transistor is a device whose complexity is only slightly greater than that of the basic semiconductor diode. However, even more so than in the case of the unijunction, the field-effect transistor is a device capable of performing many unique and highly useful functions. Because of this it has, in recent years, found use in many different applications, and it seems likely that it will be used to an even greater extent in the future.

In concept, the field-effect transistor was actually the first semiconductor amplifying device to be proposed. Far-sighted American engineer Julius E. Lilienfeld first proposed such a device as early as 1928, and patented the idea in 1930. Then in 1948 the pioneering semiconductor physicist William Shockley proposed a more practical form of the device—although his work at that time actually led to the development, with W. Brattain and J. Bardeen, of the bipolar transistor.

Despite the early theoretical predictions, it was not until 1958 that the first commercial field-effect transistor appeared. Called the "Tecnetron," it was developed by Polish scientist Stanislaus Teszner in the laboratories of the French firm, Companie Francais Thompson-Houston.

The Tecnetron was a germanium device and had rather limited performance; as a result, interest in field-effect devices did not really awaken until 1960, when the first commercial silicon device was produced by the American firm Crystalonics, Inc. Since then the devices have been developed to a stage where they are now highly competitive with the more established bipolar devices.

A number of different varieties of field-effect transistor have been developed, and although it is true that these all operated in a broadly similar fashion, the differences are significant enough to justify at least partially individual treatment. Accordingly, this chapter will adopt the procedure of dealing initially and primarily with the device which represents the most direct development from the basic semiconductor diode, namely **the junction field-effect transistor** or **"JFET."** It will use this device to develop most of the basic concepts and will then deal briefly with the other main types of device.

Other names which have been used for the JFET are "fieldistor" and "unipolar transistor," the latter term intended mainly to distinguish the device from the bipolar transistor.

In structure, the JFET is only slightly more complex than the unijunction, which we examined in the last chapter. It consists basically of a narrow strip or **channel** of lightly doped semiconductor material, whose effective conductivity is modulated by the width of the depletion layer or layers associated with one or more P-N junctions formed between the channel and adjacent heavily doped **gate** regions.

Like the unijunction, the JFET may in theory be made from either germanium or silicon; in practice, it is made almost exclusively from silicon because of the lower saturation currents and higher performance which this material offers. And as may be expected, it is possible to make "complementary" versions of the JFET—i.e., one can produce either a device having an N-type channel region and adjacent P-type gate regions, or alternatively a device with the opposite arrangement. Both types of JFET are in fact produced, and both are found in typical circuit applications.

Figure 8.1 shows the basic structure of a modern silicon JFET device of the "N-channel" variety. It may be seen that the lightly doped N-type channel of the device is roughly U-shaped, and that it is bounded on either side by heavily doped gate regions. The electrodes connecting to the gate regions are labelled here "gate 1" and "gate 2," but in most devices these connections are tied together internally and brought out as a single **gate** electrode.

The electrodes connecting to the ends of the channel region are conventionally known as the **source** and **drain** electrodes. However, in most JFETs the internal structure is symmetrical, so that these labels are actually interchangeable.

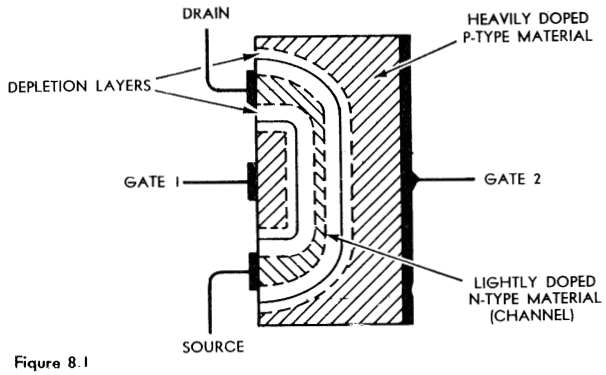


Figure 8.1

Naturally enough, even when such a device is in equilibrium with no external bias voltages applied to the electrodes, the familiar depletion layers will be set up in the vicinity of the P-N junctions along the sides of the channel. And because the channel material is intentionally doped rather lightly, compared with the gate regions, these depletion layers will extend further on the channel side of the junctions than on the gate side, as shown.

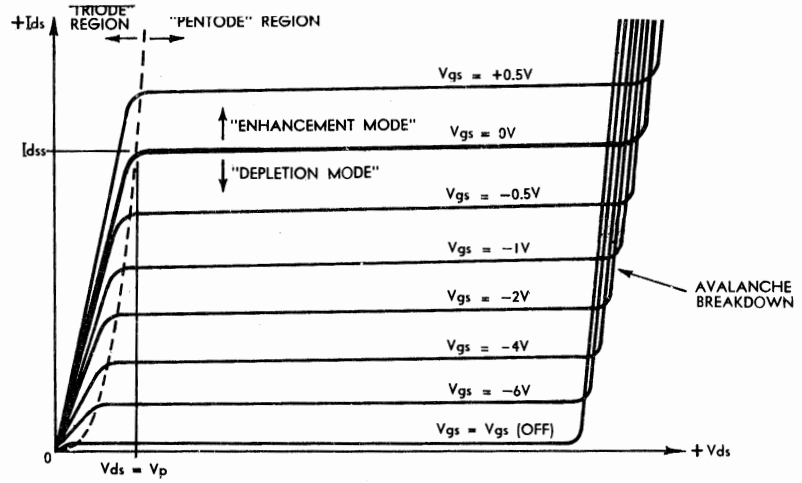
As we have seen in earlier chapters, a depletion layer is a region in a semiconductor which has been effectively "converted" into very high resistivity by the removal of all current carriers. Because of this very high resistivity, a depletion layer is actually closer to an insulator than to a conductor.

The depletion layers which extend into the channel region of a JFET thus represent areas in that region which are capable of only slight conduction relative to the remaining central strip. As a result the effective

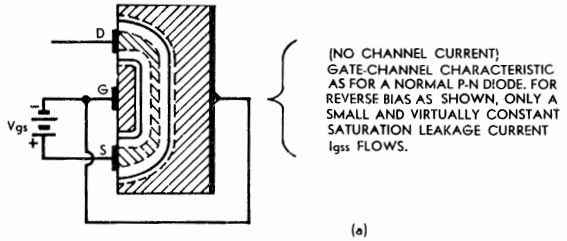
electrical width of the channel is somewhat less than its physical width, and its resistance is accordingly higher than would be the case if the depletion layers were not present.

From the discussions of P-N junction operation and depletion layer behaviour given in earlier chapters, it should be fairly easy for the reader to see that if an external bias voltage is applied to the JFET between the gate and channel regions, it will change the effective width of the channel region and hence change its resistance from the equilibrium value. An external voltage which reverses biases the gate-channel junctions will cause the depletion layers to widen, encroaching further into the channel to reduce its effective width still further and increase its resistance. Conversely, if the external voltage forward biases the junctions, the depletion layers will narrow, widening the effective width of the channel and lowering its resistance.

If another external voltage is applied to the device between the drain and source electrodes, the current drawn by the channel region will naturally depend upon both the applied drain-source voltage and upon the channel resistance. But the channel resistance is itself determined by the actual bias voltage present across the gate-channel junctions which will depend in turn upon both the external gate-channel bias and the drain-source voltage.



N-CHANNEL JFET "STATIC DRAIN-SOURCE CHARACTERISTIC"



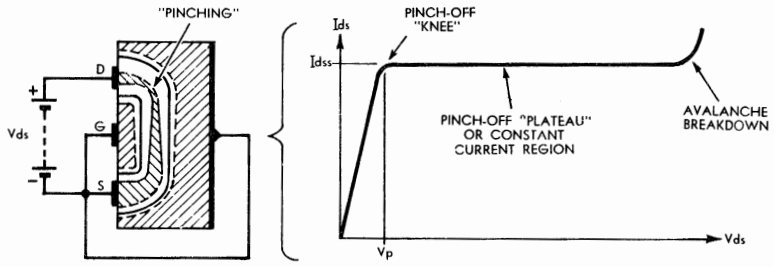


Figure 8.2

Hence the channel current which flows will be determined by both the gate-channel and drain-source voltages.

Although the relationship between channel current and the applied voltages may seem rather complex from the foregoing, it can be broken down into two quite simply understood mechanisms. One of these is associated with an "external" gate-channel junction bias component provided by the external gate bias voltage, while the other is associated with an "internal" bias component derived within the device from the applied drain-source voltage. The two mechanisms may be understood by reference to the diagrams of figure 8.2.

The diagram of figure 8.2 (a) shows the effect of an external gate-source bias Vgs applied to the JFET, with the drain electrode left unconnected. Here there is no longitudinal channel current, and hence no source of "internal" junction bias. The external bias simply causes the depletion layers of the junctions to adjust evenly to the altered conditions. In fact the gate-channel junctions of the device will behave in these circumstances exactly as a normal P-N diode.

If the polarity of Vgs corresponds to reverse bias of the junctions, as shown, the depletion layers will be found to extend considerably into the channel; at the same time only a small and almost constant gate-source saturation current **Igss** will flow. Conversely if Vgs were connected to bias the junctions in the forward direction, the depletion layers would be found to extend only a small distance into the channel. Of course if the forward bias were increased beyond the turn-on knee (0.6V in the case of a silicon device), significant gate-source current would flow; however, as will be seen later this significantly disturbs device operation, and is therefore not permitted to occur in the majority of JFET applications.

In the reverse-bias situation, there will naturally be a value of applied gate-source bias Vgs at which the depletion layers will have extended sufficiently to meet one another along virtually the **full length** of the narrow portion of the channel. For this and higher reverse-bias values of Vgs, therefore, the narrow portion of the channel will effectively consist of very high resistivity material— i.e., the channel will be effectively **cut off.**

The value of Vgs at which cutoff occurs is known as the **cutoff bias**, usually symbolised by **Vgs(off).** With typical JFETs it varies between about −1V and −10V, depending upon the doping levels and the device dimensions or "geometry."

When external gate-source bias alone is applied to the JFET, then, the depletion layers are uniform in width along the length of the channel, and the latter has a uniform width which is directly related to the gate-source bias. As soon as the applied voltage reaches the reverse-bias value Vgs (off) where the depletion layers meet, the channel is cut off. This illustrates the first of the two mechanisms responsible for JFET operation.

The second mechanism is that which is best seen when only drain-source bias is applied to the device, as illustrated in figure 8.2(b). Here the two gate regions are tied to the source electrode, so that in this case there can be no external component of gate-channel bias. However, because the drain-source bias voltage Vds is applied between the ends of the channel, there is a current and a voltage gradient in the latter, and this produces an **internal** gate-channel bias component.

Because of the voltage gradient in the channel, the gate-channel junctions will in fact be reverse-biased to an increasing extent along the channel length. The reverse bias will reach a maximum value at the drain end, where virtually the full value of the drain-source voltage Vds will be present as reverse bias.

As a result of the progressive increase in reverse bias, the junction depletion layers will increase progressively in width along the length of the channel as shown. At the source end they will have the modest width corresponding to equilibrium conditions, while at the drain end they will have widened to correspond to a reverse bias of Vds.

Because of this progressive widening of the depletion layers, a pronounced "pinching" occurs at the drain end of the narrow portion of the channel. Naturally the result of this pinching effect is that the effective channel resistance does not remain constant at its initially low value, but rises with increasing drain-source voltage. The

change in resistance is slow at first, but becomes more rapid as Vds rises.

If Vds is increased sufficiently, a point is eventually reached where the "pinching" of the channel at the drain end becomes virtually complete. The depletion layers effectively touch one another in the pinched region, converting this portion of the channel into high resistivity "intrinsic" material. Further increase in Vds then simply causes this "pinched off" portion of the channel to extend further down the channel towards source end.

Re-stating the situation, the result of this mechanism is that the drain-source current Ids drawn by the channel rises sharply with small values of Vds, then rises more slowly and finally flattens off as pinch-off is reached at the drain end of the channel. This is shown in the graph plotted on the right of figure 8.2(b), and it may be seen that the channel current has a distinct "knee" at the onset of pinch-off.

Not surprisingly, perhaps, the value of gate-channel reverse bias at the drain end of the channel which corresponds to the onset of pinch-off is known as the **pinch-off voltage,** symbolised **Vp.** Hence for the situation of figure 8.2(b) pinch-off occurs when Vds=Vp, because virtually the whole of Vds appears as reverse bias at the drain end of the channel.

With most devices the value of the pinch-off voltage Vp is almost exactly the same as that of the cutoff bias Vgs(off). A moment's thought should reveal why this is so: Vgs (off) effectively represents the junction bias necessary for the channel depletion layers to meet fully **throughout the length of the channel,** while Vp effectively represents the bias necessary at the drain end of the channel to cause the depletion layers to meet **in that region.** Providing the channel is reasonably uniform in width, therefore, one would expect the values of Vp and Vgs (off) to be identical.

Note, however, that this equivalence in value between Vp and Vgs (off) does not imply that the two have the same significance, or that "pinch-off" and "cut-off" are simply alternative names for the same situation. Vp and Vgs (off) merely have the same value because the two phenomena concerned each begin when the gate-channel depletion layers meet.

The important difference between pinch-off and cut-off is that in the cut-off situation the depletion layers have met throughout the length of the channel, converting the whole of the channel to high resistivity material, and preventing the flow of significant channel current even when drain-source voltage is applied; whereas in the pinch-off situation the meeting of the depletion layers involves only a relatively small portion of the total channel length, with the result that current flow is merely regulated.

The cutoff situation may actually be regarded as a special and "limit" case of pinch-off, as may become clear shortly. This is because the term "pinch-off" really applies to any situation in which the drain-source voltage Vds is equal to or greater than Vp.

It may be seen from figure 8.2(b) that for values of Vds above the pinch-off voltage Vp, the drain-source current Ids remains virtually constant, forming a "plateau" region. This is a result of the fact that drain-source voltages larger than Vp simply cause the pinched off portion of the channel to extend back toward the source end. The very high resistivity of the extending pinched off region thus effectively "absorbs" the additional voltage, maintaining the current constant at substantially its value at the pinch-off knee.

The drain-source current level corresponding to the constant - current "plateau" in the zero-external-gate-bias situation of figure 8.2(b) is known as the **zero-bias saturation current**, symbolised **Idss.** Like Vgs(off) and Vp, Idss is actually quite an important JFET behaviour paramater. It, too, varies with doping levels and device geometry, as one might expect, and with typical devices it ranges between about 1mA and 30mA.

Note that while the JFET pinch-off plateau current is termed a "saturation" current, it is a saturation current of a different type from that which flows through a reverse-biased P-N junction. As we saw in earlier chapters the latter type of current is "saturated" in the sense that it is limited by the number of available current carriers generated by the "intrinsic" mechanism; the JFET plateau current is limited, not by the number of carriers available, but by the pinching action of the channel depletion layers.

The channel current of the JFET remains substantially constant in the pinched-off region, then, over a wide range in drain - source voltage Vds. Significant increase in the channel current only occurs if Vds is increased to the point where the electric field strength in the depletion layers is sufficient to initiate avalanche breakdown. The current then rises sharply, as may be seen, and also the device dissipation.

As with the devices which were discussed in earlier chapters, the JFET can enter avalanche breakdown without necessarily sustaining damage. However, avalanche is a high dissipation region of operation, and like any other device a JFET has the usual continuous and short-term power dissipation ratings based on the allowable internal temperature rise. Accordingly, many low-power JFET devices may only be operated in the avalanche region, or even in some cases on the upper portion of the pinch-off plateau, for very short periods.

Although the two mechanisms involved in JFET operation have been treated separately in the foregoing discussion, and are shown separately in figure 8.2, they are generally both involved in device operation. Most JFETs are operated with both gate-source bias Vgs and drain-source bias Vds applied, so that the gate-channel junctions are presented with both "external" and "internal" bias components, and both mechanisms contribute to device operation.

The combined effect of the two mechanisms is basically a straightforward additive one. The external gate-source bias Vgs provides a fixed component of gate-channel bias, and hence contributes to widening (or narrowing) of the channel depletion layers in a uniform fashion, while the drain-source bias Vds provides a progressive internal reverse bias component, and hence a tapering or pinching contribution to the depletion layer width. The resultant width of the depletion layers is simply the sum of the two.

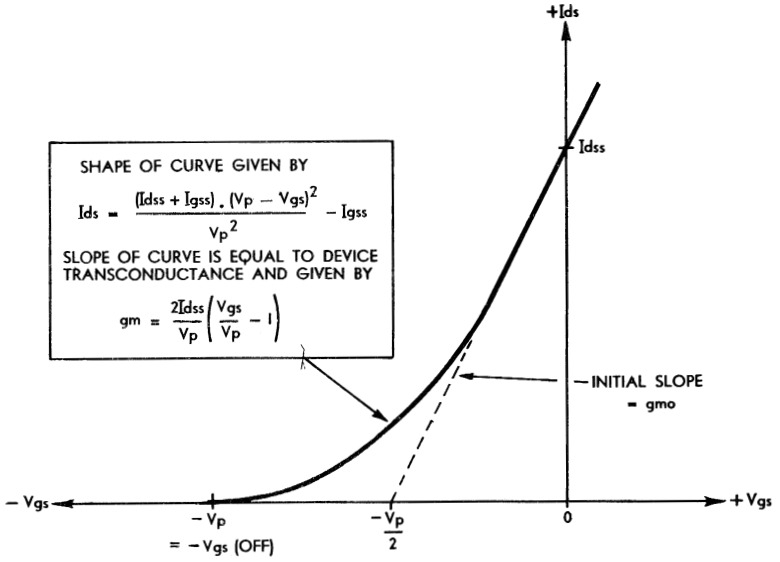


Figure 8.4 N-CHANNEL JFET "STATIC TRANSFER CHARACTERISTIC"

Pinch-off still occurs when the effective gate-channel reverse bias at the drain end of the channel is equal to Vp, the pinch-off voltage. However, this point will no longer in general correspond to the point where Vds=Vp, as in the zero-external-gate-bias case, but because Vgs also contributes to the depletion layer width it will now correspond to the situation

$$Vds - Vgs = Vp \qquad \ldots (8.1)$$

where the negative sign simply draws attention to the fact that the external gate bias is nominally of the opposite polarity to the drain bias.

In other words, the effect of a fixed negative bias component produced by Vgs is simply to lower the value of drain-source voltage Vds at which pinch-off is reached. The higher Vgs is made, the wider the uniform widening of the channel depletion layers and the lower the value of Vds at which the layers meet at the drain end.

Ultimately, of course, if Vgs is made equal to or greater than Vp, and hence

equal to or greater than Vgs(off), the device is in the pinch-off region of operation even when Vds = 0 —i.e., it is cut off. Hence the reason for regarding the "cutoff" condition as a special and limiting case of pinch-off.

Naturally the converse effect occurs if the applied gate-source bias is in the forward-bias direction. Here the effect will be to increase the value to which Vds may be raised before pinch-off is reached.

It should be noted in passing that in saying that the drain-source voltage Vds and the gate-source voltage Vgs both contribute to the width of the channel depletion layers, and hence to pinch-off, all we are really saying is that it is the **effective drain-gate voltage** present across the device which determines whether or not it has entered pinch-off.

In short, an alternative general requirement for pinch-off is that the drain-gate voltage Vdg must be equal to or greater than the pinch-off voltage Vp.

With either polarity of applied gate-source bias, the altered depletion layer situation also results in a value of pinch-off plateau current different from the value Idss corresponding to the zero-bias case. When Vgs is of the reverse-bias polarity the plateau current level is naturally lower than Idss, while with Vgs values of the forward-bias polarity (but below about 0.6V) the plateau current level exceeds Idss.

Because each value of Vgs thus results in both a unique value of Vds corresponding to the pinch-off knee, and also a unique value of pinch-off plateau current, it is convenient to represent JFET operation by a family of characteristic Vds/Ids curves of the type shown in figure 8.3. The polarities shown are for an N-channel device as shown in figure 8.1; for a P-channel device they would be reversed.

It may be seen that for each of the sample values of Vgs for which the curves are drawn, there is a different value of Vds appropriate to the pinch-off knee. In fact the knee points of the curves all lie on a parabolic locus (dashed curve), which is exactly what one would expect from the relationship given in expression (8.1). Similarly each curve has its current plateau at a different value of Ids.

The portions of the various curves to the **left** of the dashed knee-point locus may be seen to resemble fairly closely the familiar plate characteristics of a triode thermionic valve. For this reason this area of the JFET drain-source characteristics is often called the "triode region" of operation. Similarly because the remaining portions of the various curves resemble the plate characteristics of a pentode thermionic valve, this area of the characteristics is often called the "pentode region" of operation.

In most circuit applications JFETs are operated in the pentode region of operation—that is, at operating points to the **right** of the dashed curve in figure 8.3.

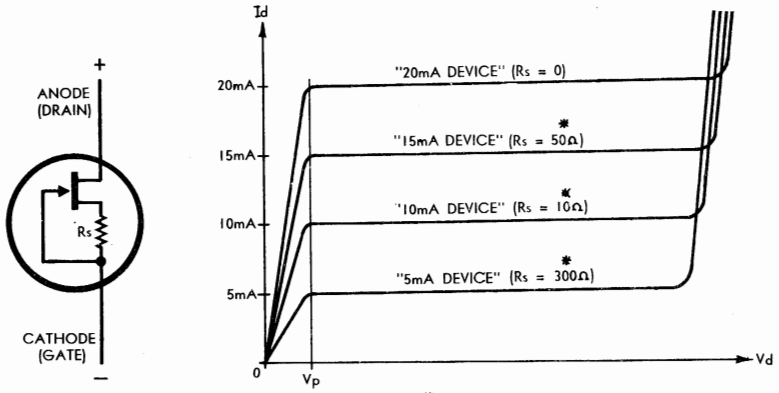
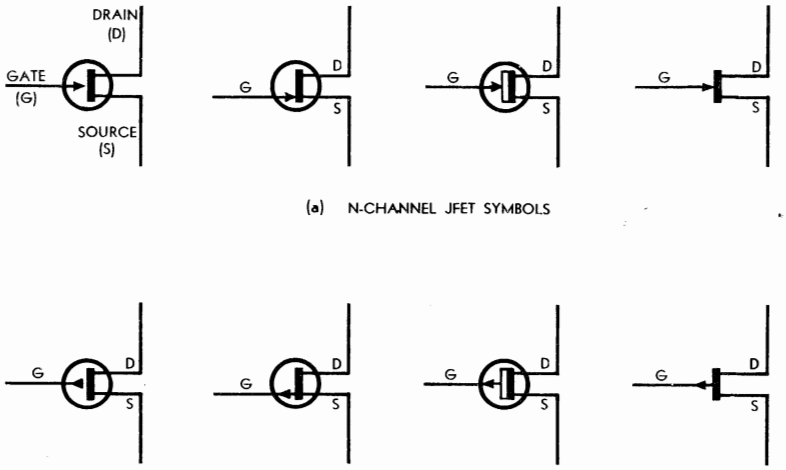


Figure 8.6



(a) N-CHANNEL JFET SYMBOLS

(b) P-CHANNEL JFET SYMBOLS

Figure 8.5

Because the narrower channel depletion layers produced by forward gate-source bias result in higher Ids (or "enhanced channel conduction") of a device in the pentode region of operation, relative to the zero-bias situation, this mode of operation is known as **enhancement mode.** This mode of JFET operation is represented in figure 8.3 by the curve marked "Vgs = +0.5V".

Fairly obviously the range of enhancement mode operation possible with JFETs is rather limited, because Vgs cannot be increased beyond the point where forward conduction current begins to flow through the gate-channel junctions. However, it will be shown later that other types of field-effect device are capable of more extended enhancement mode operation.

In contrast with forward gate-source bias, reverse bias produces wider channel depletion layers and results in lower Ids or "depleted channel conduction" in the pentode region of operation, again relative to the zero-bias situation. This mode of operation is accordingly known as the **depletion mode,** as shown.

JFET devices are almost always biased to a quiescent operating point in the depletion mode region, if only for the reason that this allows a device to be swung over a greater dynamic range before non-linearity occurs.

A further point which may be noted from figure 8.3 is that the drain-source voltage Vds at which a device enters avalanche breakdown reduces with increasing reverse gate-source bias Vgs. This is really only to be expected, because Vgs and Vds are additive in terms of the effective maximum reverse bias present across the gate-channel junctions at any time.

In effect, then, it is really the drain-gate voltage present across the device which determines whether or not it enters avalanche breakdown, just as this same voltage determines whether or not the device is operating in the pinch-off or pentode region. Hence a common way of rating a JFET in terms of its avalanche breakdown point is to quote its **drain-gate breakdown voltage,** usually symbolised **BVdgo.**

JFET "static drain-source characteristics" of the type illustrated in figure 8.3 show quite well the operation of the device, as may be seen. However, for design work they are often of less interest and lower utility than the so-called "static transfer characteristic," which is illustrated in figure 8.4. This curve shows the controlling action of gate-source bias Vgs upon the device drain-source current Ids, for the pentode region of device operation (only).

Note that whereas there is a whole family of curves comprising the static drain-source characteristics, the static transfer characteristic consists of but a single curve. This arises from the fact that the transfer characteristic by definition only applies to the pentode region of operation, where the constant-current nature of the drain-source characteristics makes the "transfer" or controlling effect of Vgs over Ids virtually independent of drain-source voltage Vds.

It may be seen that the transfer characteristic is a parabolic curve whose shape and slope are described by the expressions shown. The essential points to note are that the curve cuts the Ids axis at a value equal to Idss, the zero bias drain-source current, and that it becomes asymptotic to the Vgs axis at a value equal to both Vgs(off) and Vp.

The transfer characteristic describes the relationship between Ids and Vgs, so that its slope at any point represents the rate of change in Ids for a change in Vgs—i.e., the **transconductance** or "mutual conductance," usually symbolised **gm.**

Because of the parabolic shape of the curve, its maximum slope occurs in the region where it crosses the zero bias or Ids axis, at Idss. In other words, the transconductance of a JFET is greatest when the device is operating at zero or slight forward gate bias.

This being the case, device manufacturers usually specify the transconductance of a JFET for the zero-bias condition, where it is nominally at a maximum. This "maximum transconductance" is usually symbolised **gmo.** Because of the shape of the transfer curve gmo is closely approximated by the simple expression

$$gmo = \frac{-2.Idss}{Vp} \qquad \ldots (8.2)$$

which in graphical terms simply corresponds to the dashed line in figure 8.4 joining the Ids axis at Idss and the Vgs axis at $-Vp/2$.

An alternative to gm sometimes quoted on JFET data sheets is the **forward transadmittance,** symbolised **Yfs.** This is strictly a more general device parameter, including any susceptance (inverse reactance) components of the transfer behaviour in addition to conductance. However, in most cases it is specified at a low frequency (around 1KHz) where the zero-bias value of Yfs is generally almost identical with gmo.

For typical JFET devices in current production, gmo ranges from about 1000-8000 micromhos, or 1—8mA/V.

From figure 8.4 and from expression 8.2 it may be seen that the transconductance characteristics of a JFET are closely determined by the zero bias current Idss and the pinch-off voltage Vp. In fact, knowing these two parameters it is quite easy both to calculate gmo and to construct the transfer characteristic. This provides further evidence of the importance of the two parameters.

It may be worthwhile to summarise our present discussion of the JFET by drawing attention to those unique aspects of the device behaviour which are together responsible for its wide range of circuit applications, and which are accordingly of particular significance for circuit design.

Possibly the first thing which the reader may have realised from the foregoing description of JFET operation is that the device is one which, like the thermionic valve, is capable of **power amplification.** A small change in gate-source voltage Vgs is capable of producing a relatively large change in drain-source current Ids. Hence if a small AC signal is superimposed upon a suitable quiescent gate-source bias, an amplified AC signal can be obtained at the JFET drain electrode by placing a suitable load resistor in series with the Vds supply.

Because in normal operation its gate-source junctions are biased either only slightly in the forward direction, or more usually in the reverse direction, the JFET also has another important property in common with the thermionic valve: **high input resistance.** The only current which normally flows in the gate circuit is the junction saturation/leakage current Igss, mentioned earlier, which is typically in the order of but a few nanoamps. This gives typical devices an input resistance of around 1000 megohms.

As we observed from the static drain-source curves shown in figure 8.3, the Vds/Ids characteristics of the JFET in the pinch-off region are virtually "constant current" lines, having a very low current change/voltage change slope. In other words, then, the device resembles a pentode valve, possessing a **high output resistance.** Typical figures for JFET output resistance **rds** range from about 20K to 100K.

As with transconductance, some device manufacturers do not quote the output resistance rds on their JFET data sheets, but instead give values for **output admittance,** symbolised **Yos.** Usually this is quoted at a low frequency, say 1KHz, where its value is very close to the inverse of rds. Hence typical devices have Yos values in the range 10-50 micromhos.

Not surprisingly the depletion layers which separate the gate and channel regions of a JFET in normal operation behave as a dielectric, in this respect being no different from the depletion layer of a normal P-N junction diode. As a result there is a small but often significant capacitance between the gate and channel regions. The distribution of capacitance is naturally non-linear due to the tapering depletion layers, and also varies with the applied bias.

For convenience in circuit design the gate-channel capacitance is normally considered to consist of two main components: the **effective input capacitance** of the device as seen by the gate electrode, symbolised by **Cgs,** and the **reverse transfer or drain-gate "feedback" capacitance,** symbolised **Cdg.**

Typical modern JFET devices designed for low- and medium-frequency applications have Cgs figures ranging from 4—7pF, and Cdg figures ranging from 1—3pF. Devices intended for high frequency applications have figures somewhat lower than these.

The reverse transfer capacitance Cdg is often of particular significance for circuit design, because being coupled between the input and output of the device it can be effectively magnified in value by the familiar "Miller effect." Further discussion of this will be found in the next chapter.

The circuit symbols commonly used for JFETs of both configurations are shown in figure 8.5.

It is hoped that the foregoing discussion of the junction field-effect transistor has given the reader a basic understanding of the device and its operation. Let us now turn to consider briefly some of the other types of field-effect device in present use.

A device which is very closely related to the JFET is the so-called **constant current diode.** Although basically a very simple development from the JFET, this device is finding increasing use in many circuit applications in which current levels must be maintained despite voltage and impedance variations.

Basically the device consists of a JFET which is fitted with an "internal" self-bias resistor in series with the source, with the gate being tied to the remote end of the resistor. Figure 8.6 shows the basic arrangement, where it may be seen that only the drain and gate connections are brought out as device electrodes. These are labelled "anode" and "cathode" respectively.

As one might expect the operation of the device is again dependent upon the two basic JFET mechanisms discussed earlier. However, in this case the single bias voltage Vd applied to the device is connected directly between drain and gate, so that pinch-off simply corresponds to the situation where Vd=Vp. The pinch-off voltage is not dependent upon the value of Rs.

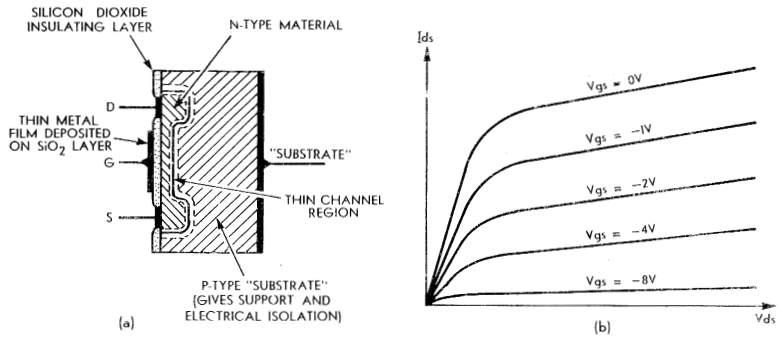


Figure 8.7 DEPLETION OR "TYPE A" MOSFET (N-CHANNEL)

The function of the resistor is to provide a "fixed" component of gate-source bias derived from the device channel current. This quite naturally has the effect of determining the value of device current at which the pinch-off plateau occurs. Thus a device fitted with no resistor might have a plateau current (in this case equal to Idss) of say 20mA, while a device fitted with a resistor of 100 ohms might have a plateau current of 10mA, as shown.

It should be fairly clear from this that the plateau current of such a device may be set to any desired value below the basic Idss for the internal structure, merely by fitting the appropriate value of resistor Rs. Hence it is possible to produce such devices with plateau currents covering quite a useful range, suitable for use in circuit applications as current regulating devices. In operation, the devices are merely arranged to operate on their pinch-off plateau, so that they tend to pass a substantially constant current despite variations in applied voltage.

No doubt the astute reader will have realised while reading the foregoing that virtually any normal JFET device could be used as a current regulating element, simply by connecting it into circuit with the source tied to the gate via a suitably chosen resistor. And in fact this forms the basis of many JFET circuit applications. However, semiconductor device manufacturers

have found it possible to provide a range of "custom-made" current regulating devices with specified current ratings, and accordingly circuit designers have been able to take advantage of the devices.

An important type of field-effect device which differs both in construction and in certain aspects of its operation from the JFET is the **insulated-gate field effect transistor,** or **IGFET.** Other general names for this type of device are MISFET, standing for "metal-insulator-semiconductor FET," and TFT, or "thin film transistor." The last of these names is usually reserved for devices which are in the form of elements within micro-circuits or "ICs."

In broad terms the operation of IGFET devices is very similar to that of the JFET device which we have already examined. As before, the effective conductivity of a semiconductor channel region is modulated by a control bias applied between the channel and an adjacent electrode termed the gate.

However, an important difference between the two types of device is that whereas in the JFET the gate electrode is isolated from the channel by a non-conducting P-N junction, in the IGFET this isolation is performed by a very thin layer of insulating material such as silicon oxide or silicon nitride. Also the gate electrode is a metallic film deposited on the surface of the insulating layer, rather than a semiconductor region.

Probably the most common type of IGFET device is the **MOSFET** or **metal-oxide-semiconductor** FET, in which as the name suggests the gate-channel insulation is performed by a thin layer of silicon dioxide. Other names for this device are "MOST," "MOS transistor" and "SCOUT" — the latter standing for "surface controlled oxide unipolar transistor."

Because the MOSFET relies upon an oxide layer for gate-channel isolation rather than the depletion layers associated with non-conducting P-N junctions, it is not inherently subject to the restriction on enhancement-mode operation which applies to the JFET. There are definite restrictions to the voltage which may be applied between gate and source, as will be explained shortly, and these restrictions are of paramount importance if a MOSFET is to be protected from damage; however in general they apply equally for both polarities of applied gate-channel voltage.

Taking advantage of this, device manufacturers have been able to provide three different types of MOSFET, each of which is designed to give optimum performance under different conditions. Thus there are (a) the depletion-mode or **normally on** MOSFET, designed to operate in a very similar fashion to the JFET; (b) the depletion/enhancement MOSFET, designed for operation at around zero bias, and capable of linear signal excursions into both the depletion and enhancement modes; and (c) the enhancement-mode or **normally off** MOSFET, designed for optimum operation in the "forward-biased" condition.

The three types of MOSFET are sometimes known respectively as type "A," type "B" and type "C" devices.

The basic construction of a depletion-mode or type "A" MOSFET is shown in figure 8.7(a). It may be seen that the device channel here consists of a very thin semiconductor layer linking the drain and source regions at the surface of a supporting or "substrate" region. The device shown is of the "N-channel" variety, with an N-type channel and a P-type substrate; however the complementary configuration is also made. Like JFETs, MOSFETs can be made in both "polarities," this applying to all three types of device.

As the channel and substrate regions of the device are of opposite type, the junction between the two is surrounded by the usual depletion layer even in equilibrium. However, in this case the depletion layer plays no part in the operation of the device, serving merely as an internal isolation medium for the channel. In typical circuit applications the substrate electrode of a JFET is simply tied to the source, to earth or to some other "cold" reference point.

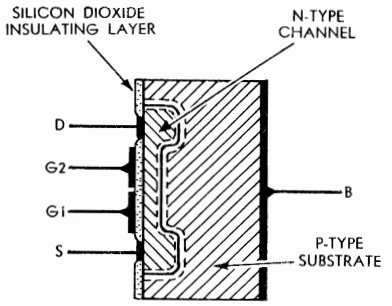
The surface of the MOSFET above the channel is covered, as may be seen, with a silicon dioxide insulating layer. The layer is very thin, typically in the order of 1,000 angstroms (.0001mM). Deposited in turn on the top of this layer, above the channel, is the gate electrode. This is simply a thin film of metal, usually aluminium.

Fairly obviously, because there is no "junction" as such between the gate and channel of the device, there can be no depletion layer at the top of the channel for zero bias, to correspond to the "equilibrium" depletion layers present in the JFET. However, the depletion mode MOSFET nevertheless operates in a very similar manner to that of the former device, as a result of the close electrical coupling between the gate and channel provided by the very thin oxide layer.

In fact although there is no depletion layer at the top of the channel for zero gate bias, such a layer begins to "grow" inwards from the top of the surface of the channel beneath the gate as soon as external reverse bias is applied to that electrode. The electric field between the gate and the semiconductor material causes carriers to be repelled from the surface, leaving a carrier-depleted region virtually identical to that associated with a P-N junction. (The repelled carriers are normally swept away by the longitudinal channel field, just as in the case of the JFET; they correspond to charging current of the gate-channel capacitance.)

As before, the encroaching depletion layer reduces the effective electrical thickness of the channel.

Not surprisingly, when gate-source bias Vgs and drain-source bias Vds are both applied, there is again a pinching action at the drain end of the channel, and channel current tends to reach a saturation or pinch-off level. Hence the depletion-type MOSFET has very similar Vds/Ids characteristics to those of a JFET, as may be seen from figure 8.7 (b). The "plateau" segments of the curves are not quite as horizontal as those of the JFET, because of the less intimate control exercised by the gate, but the behaviour of the device is very similar.



DUAL-GATE DEPLETION-TYPE MOSFET (N-CHANNEL)

Figure 8.9

In contrast with the depletion-mode MOSFET of figure 8.7 is the enhancement-mode type, whose basic construction and operation are shown in figure 8.8. The type shown is the "P-type channel" version, or more strictly the "**induced** P-type channel" configuration.

The construction of this type of device is similar to that of the depletion type, as many be seen, except that there is **no physical channel** between the two "islands" forming the drain and source regions. The substrate is continued right up to the oxide-covered surface between the two. Hence when no external gate bias is applied to the device, there can be no drain-source current except a small saturation/leakage current through the drain substrate and substrate-source junctions.

This explains why the enhancement-type MOSFET is often called a "normally off" device, in contrast with the "normally on" characteristics of the JFET and depletion-type MOSFET.

At this point the reader may well be wondering how the enhancement-type device can be persuaded to pass current. Actually the answer to this is fairly obvious — by the creation of an "effective channel" linking drain and source. And not unexpectedly, this effective channel is created at the surface of the substrate by the external bias applied to the gate electrode.

The idea is that "forward" bias applied to the gate produces an electric

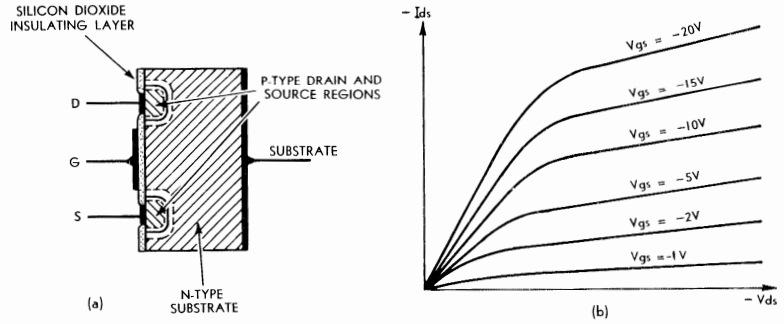


Figure 8.8 ENHANCEMENT OR "TYPE C" MOSFET (P-CHANNEL)

field at the surface of the substrate, and this in turn has two effects. One is that majority carriers in the substrate material are repelled away from the surface; the other effect is that minority carriers are attracted towards the surface. And the net result of both these effects is that the material at the surface of the substrate is effectively **inverted in type** to become what is termed an **induced channel** linking drain and source.

Hence the example shown, forward bias (gate negative) tends to repel electrons from the surface of the N-type substrate, and at the same time attract thermally generated holes. The surface is thus inverted in type to form an induced P-type channel linking the drain and source regions, and drain-source current is able to flow if a drain-source bias Vds is applied.

Naturally the greater the forward bias applied to the gate, the deeper the induced channel and the lower the drain-source resistance. However, as before the drain-source bias Vds tends to reverse-bias the drain end of the induced channel so that a phenomenon very similar to pinch-off occurs. Hence apart from the different gate bias

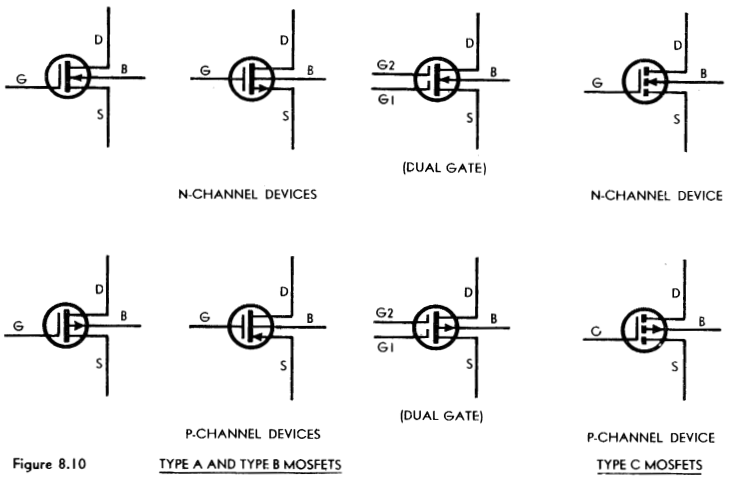


Figure 8.10

sense, the Vds/Ids curves of an enhancement-type MOSFET prove rather similar to those of a depletion - type device. This may be seen by comparing the typical curves given in figure 8.8(b) with those of figure 8.7(b).

Note that in the case of the depletion-type MOSFET the gate need not extend for the full length of the channel in order to achieve proper device operation, whereas with the enhancement-type device it is essential for the gate to extend the full distance between the drain and source in order to provide a link between the two. This tends to make the enhancement-mode device harder to fabricate, and also gives it a higher gate-channel capacitance.

The depletion-enhancement or "type B" MOSFET is very similar in construction to the depletion-type device shown in figure 8.7. The only difference is that the channel section is made particularly thin, allowing the gate bias to be used either to diminish its conductivity in the manner of a depletion-type device, or to enhance its conductivity in the manner of an enhancement-type device.

A further type of MOSFET device which should be briefly mentioned here is the **dual-gate MOSFET,** which as the name suggests is a device having not one, but two control gate electrodes. The two gates are arranged to act upon the channel conductivity in cascade, as may be seen from the diagram of figure 8.9. For practical reasons associated with both the fabrication and application of such devices they are normally made in either the type A (depletion) or type B (depletion/enhancement) variety—i.e., in "normally-on" form.

The two gate electrodes of this type of device make it very well suited for use as a controlled-gain amplifier, a "cascode" RF amplifier, and an RF mixer. Thus although the device is a relatively late development on the semiconductor device scene, it is already finding many applications.

The circuit symbols commonly used for the various types of MOSFET are shown in figure 8.10.

Because of the excellent insulating properties of the silicon dioxide layer insulating the gate of a MOSFET from its channel, the input resistance of these devices is typically some 1,000 to 10,000 times greater than that of a JFET — i.e., from 1 to 10 Teraohms (1 to 10 million Megohms). This is even higher than many thermionic valves, and is, in any case, independent of the polarity of the applied gate bias — in contrast with both the JFET and the thermionic valve. At the same time the gate-channel capacitance of the MOSFET is generally somewhat lower than for the JFET, due to the isolation associated with the oxide layer, and this gives lower values for both Cgs and Cdg.

Together with these advantages come problems, however. It proves difficult to reduce the thickness of the silicon dioxide layer separating the gate electrode from the channel sufficiently to achieve as high a transconductance with MOSFETs as can be achieved fairly easily with the JFET. Very thin oxide layers are not only difficult to achieve reliably during manufacture, but they also present stability problems; their insulation becomes more subject to imperfections due to trapped impurities, and a phenomenon known as "ion drift" can occur over a period of time due to migration of impurity ions from the oxide into the semiconductor channel.

Not only this, but the silicon dioxide layer of a MOSFET does not possess the same breakdown characteristic as that of the P-N junction insulating the gate of a JFET. Whereas the latter can enter avalanche breakdown without necessarily sustaining damage, the oxide layer of a MOSFET is only capable of the "punch-through" breakdown typical of dielectrics such as paper and plastic film. Hence if a critical field strength is exceeded the gate-channel insulation is punched through at a particular point, and the device may well be ruined.

Because of the very high resistance and low capacitance between the gate and channel, even slight "static electricity" charges reaching the gate of a MOSFET can produce permanent device damage in this fashion. Hence such devices are normally supplied by the manufacturer with all electrodes temporarily shorted together to preclude static charge effects, and the electrode shorting clips are normally left connected until the devices are wired into circuit ready for operation.

Recently MOSFET devices have been released featuring "internal" protection against gate insulation failure, by means of zener diode structures incorporated into the basic device. The diodes are arranged to enter non-destructive avalanche breakdown before the oxide punch-through voltage is reached. Naturally these devices provide a form of MOSFET which is somewhat more rugged electrically than the standard type; however because the protection diode P-N junctions are effectively in parallel with the gate-channel insulation, the input resistance of these devices is lowered to the level of approximately 1000M typical of JFET devices. Luckily this figure is still very high, and quite adequate for many applications.

SUGGESTED FURTHER READING

CHERRY, E. M., and HOOPER, D. E., **Amplifying Devices and Low-Pass Amplifier Design,** 1968. John Wiley and Sons, New York.

COBBOLD, R. S., **Theory and Applications of Field-Effect Transistors,** 1970. John Wiley and Sons, New York.

COHEN, J. M., "An Old-Timer Comes Of Age," in **Electronics,** V.41, No. 4, February, 19, 1968.

NOLL, E. M., **FET Principles, Experiments and Projects,** 1968. Howard W. Sams, Inc., Indianapolis.

SEVIN, L. J., **Field Effect Transistors,** 1965. McGraw-Hill Book Company, Inc., New York.

WALSTON, J. A., and MILLER, J. R. (Eds.) **Transistor Circuit Design,** 1963. McGraw-Hill Book Company, Inc., New York.

Chapter 9

FET APPLICATIONS

FET applications—parameter spread and its implications—design approaches — biasing — fixed and self bias — composite bias—audio amplifiers—configurations—DC amplifiers — RF amplifiers and oscillators — other applications.

We have seen in the last chapter that both the JFET and IGFET varieties of field-effect transistor effectively combine many of the worthwhile features of the familiar thermionic valve with the compactness, efficiency and reliability characteristic of semiconductor devices in general. It should therefore come as no surprise to find that both of these devices are of considerable value to the designer of electronic circuits and that, as a result, their applications are both numerous and rapidly growing.

It is proposed to discuss briefly in the present chapter some of the more common applications in which FET devices are found. However, before dealing with specific applications it will be worthwhile to examine some more general aspects of device usage which are associated with, and follow from, a seemingly unavoidable variation in parameter values from device to device.

It may be remembered that the behaviour of FET devices can be described in terms of three main parameters, which in the case of JFETs and depletion-type MOSFETs are the zero bias current Idss, the pinch-off voltage Vp, and the transconductance gm. Enhancement type MOSFETs may be described by equivalent parameters.

Basically these parameters are determined by such factors as the semiconductor impurity doping levels and the effective dimensions or "geometry" of the devise channel region. Because doping levels and device geometry are in practice variables, subject to inevitable variations during production, the parameters for a particular FET device type are subject to a corresponding variation about their nominal values. This type of variation in fact tends to occur in the parameters of all semiconductor devices, and is commonly known as **parameter spread.**

As a result of parameter spread, the voltage-current behaviour of the channel region of each device of a particular type of FET will tend to be different from that of every other device. Hence in practice such a device type cannot be represented by a single family of Vds/Ids curves as shown in figure 8.3 of last chapter. Rather, each individual device will tend to have its own family of curves, so that the device type as a whole could really only be represented by a complete "family of curve families."

Not only is such a "family of curve families" very difficult to represent graphically, but also it presents the information on both the nominal device parameter values and their spread ranges in a particularly unwieldy form. It is for this reason that the "static drain-source characteristic" illustrated in figure 8.3 is generally found to be of little use in practical FET circuit design.

Fortunately the required information can be both presented and represented in a form far more easily interpreted and used, by means of the "static transfer characteristic" which was shown previously in figure 8.4. Being a single curve for any particular device, this characteristic quite readily lends itself to representation of the limits of parameter spread applicable for a given device type.

The way in which the static transfer characteristic is adapted to represent the behaviour of a certain FET device type is illustrated in figure 9.1. Here the curve marked "nominal device" corresponds to an average or typical device having the nominal values of Idss, Vp and gm for the device type concerned, while the other two curves represent hypothetical "limit" devices representing the extremes of parameter spread permitted within that device type.

For any given FET device type, a set of curves similar to that of figure 9.1 can be drawn using information supplied by the manufacturer. Hence, by definition, any particular device of the type concerned should have a transfer curve lying somewhere between those for the upper and lower limit devices, with a majority of devices (ideally) falling close to the curve representing a nominal device.

The curves of figure 9.1 may seem to exaggerate the extent of spread variation in FET parameters, but this is not the case. Because of the narrow channel region involved in most FET devices, the parameters of these devices tend to be particularly sensitive to doping level and geometry variations. Hence the extent of spread variation in FET devices tends to be somewhat greater than for most other semiconductor devices.

As one might expect, this rather wide spread in FET parameters tends to complicate circuit design. The designer must generally arrange for his circuit to be particularly tolerant of device parameter variations, because each piece of equipment concerned will contain not simply a sample from a group of virtually identical devices of the required type, but rather a sample from a distribution of devices whose behaviour varies over a significant range.

Basically this forces the designer to think primarily in terms of the upper and lower limit devices, and adopt a "worst case" design procedure. Each important aspect of circuit performance is arranged to fall within acceptable limits for both the extremes of device variation, by this means ensuring that performance will also be acceptable for all devices in between the extremes.

Naturally those aspects of circuit be-

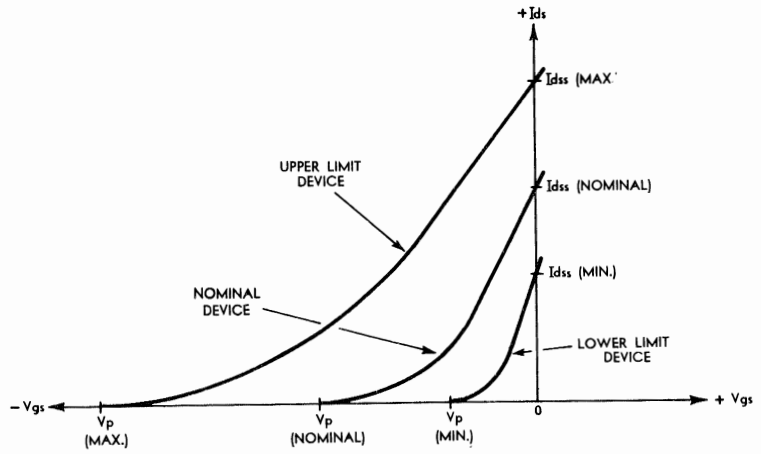
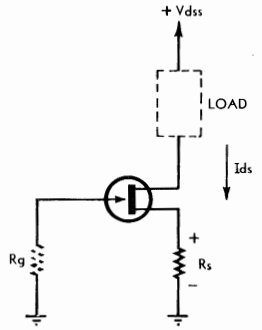


Figure 9.1 COMPOSITE TRANSFER CHARACTERISTIC (N-CHANNEL JFET)

haviour or performance which the designer may regard as important will depend upon the type of circuit involved. In the case of audio and RF amplifiers, it is often important to keep the transconductance of the FET devices within narrow limits at the quiescent operating point, in order to maintain the circuit gain; however, in RF mixers it may be more important to maintain the drain current within a narrow range, to ensure satisfactory signal-handling performance. In the case of DC amplifiers it may be important to maintain the drain or source voltage of a stage within certain limits, to ensure correct interstage coupling conditions, whereas in switching circuits the main concern may be to ensure that all devices switch reliably between cutoff and "full on."

Probably the most common method used to bias JFETs and type A and B MOSFETs to the appropriate DC quiescent operating point in linear circuits is the "self-bias" method, which is illustrated in figure 9.2. It may be seen that this method is very similar to the cathode-bias system often used for thermionic valves. The required gate-source bias voltage is generated as a voltage drop due to the flow of drain-source current Ids through a small resistor Rs connected in series with the source electrode.



EFFECTIVE GATE BIAS = −(Ids.Rs)
ASSUMING GATE LEAKAGE
CURRENT NEGLIGIBLE

Figure 9.2

With most modern JFET devices, and certainly with all normal MOSFETs, the gate leakage current Igss is sufficiently small to be negligible in all but the most critical circuitry. Hence the presence of any resistance Rg in series with the gate electrode does not provide any significant contribution to gate-source bias. As a result, the effective gate-source bias presented to a device in this type of circuit is simply given by −(Ids.Rs).

It may be seen that this bias method involves negative feedback, because the bias applied to the device depends upon the current drawn by the device itself. This explains the meaning of the term "self-bias."

The negative feedback involved in this method of biasing helps considerably to reduce the effect of device parameter spread upon the DC operating conditons. A device near the upper limit of the parameter spread range tends to provide itself with increased bias, while a device near the lower limit tends to provide itself with less bias—in both cases moving the operating conditions closer to those for a nominal device than would be the case if fixed bias were applied.

This "stabilising" action may be seen quite clearly if the behaviour of the self-bias system is compared with that of fixed bias using a composite transfer characteristic of the type introduced in figure 9.1. Such a characteristic has been drawn in figure 9.3 with bias lines for both fixed and self-bias.

With fixed gate-source bias, represented by the dashed line, it may be seen that the operating point of individual devices will vary widely. An upper limit device will operate at point A, drawing considerably greater drain-source current than a nominal device at point B. Conversely a lower limit device will operate at point C, with a considerably lower drain-source current. Quite apart from any embarrassment which this wide range in drain-source current may incur, there are often other problems.

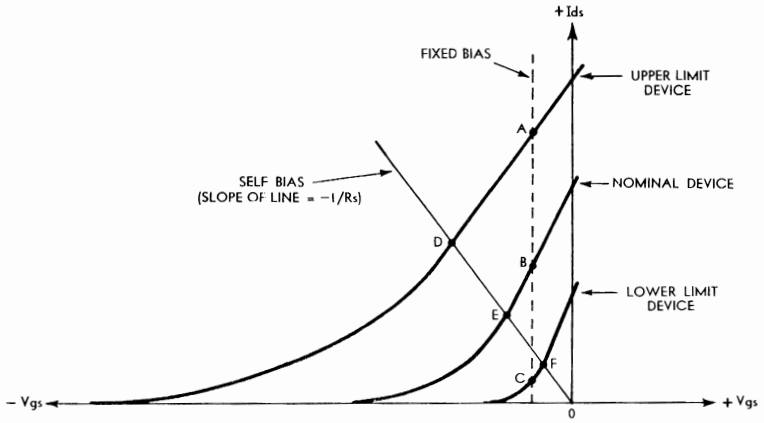


Figure 9.3

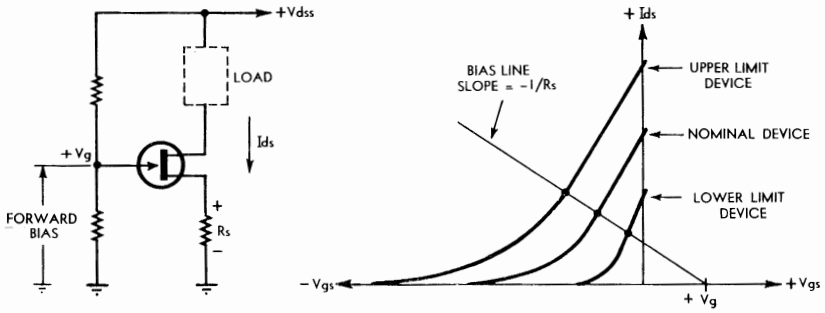


Figure 9.4 FIXED FORWARD BIAS TO PERMIT INCREASED SOURCE RESISTOR

In circuits wherein the drain is connected to the drain supply via a very low resistance, the high current drawn by an upper limit device under fixed bias conditions may be sufficient to cause excessive power dissipation and device damage. Alternatively, if the drain is connected to the supply via a significant resistance the voltage drop produced by the high drain current may cause the drain-gate voltage present at the device to fall below the pinch-off voltage Vp, causing such a device to operate outside the linear pinch-off or "pentode" region.

The low drain-source current drawn by a lower limit device admittedly does not produce dissipation or operating region problems. However, the working transconductance or gm of such a device will tend to be quite low, because of its proportionally greater bias. This may well reduce its signal performance to an unacceptable level.

The improvement in operating point stability gained by the use of self-bias is shown by the oblique solid line passing through the origin. The slope of this line is equal to (−1/Rs), representing the relationship between gate-source bias voltage and the drain-source current. The self-bias operating points for upper limit, nominal and lower limit devices are marked respectively as points D, E and F.

It may be seen that the range in Ids values between points D and F is significantly less than that between points A and C, showing that self-biasing helps considerably to reduce the effects of parameter spread on quiescent drain-source current. Also it may be noted that all devices are now biased to approximately corresponding points on their transfer characteristics, which generally reduces transconductance variations to a minimum and contributes to more uniform signal-handling performance.

As may become apparent later in this chapter from the various FET device applications described, the simple self-bias method illustrated in figures 9.2 and 9.3 is widely used for FET amplifiers, mixers, oscillators and other "linear" circuits. Assuming the use of modern FET devices having parameter spreads confined within reasonable limits, the designer can generally use this method to design his circuit for satisfactory operation with all devices of the type concerned.

There are, however, cases where the simple self-bias method does not provide the required degree of operating current stabilisation. In such cases it is often possible to achieve satisfactory results using an extension of the

method which is illustrated in figure 9.4.

Here a source resistor Rs is used as before, but in addition the gate of the device is provided with a fixed **forward** bias. This may be provided either by means of a resistive voltage divider from the Vdss supply, as shown, or alternatively by leaving the gate at ground or common potential and returning the source resistor to a suitable reverse-polarity supply line. In the case of an N-channel device as shown, the latter approach would involve the use of a second supply which provides a negative voltage with respect to ground.

As a result of the fixed forward bias, the source resistor Rs can be made larger in value than for simple self-bias. This increases the negative feedback action and hence gives a further improvement in operating current stabilisation. The effect may be seen from the graph in figure 9.4: the reduced slope of the bias line has further reduced the range in Ids values between the operating points of upper-limit and lower-limit devices.

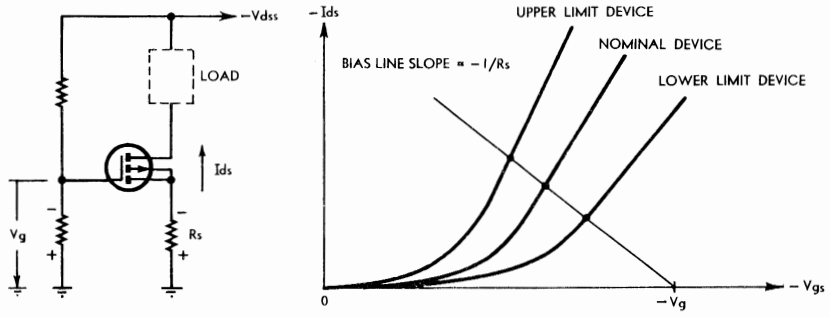


Figure 9.5 TYPE C OR ENHANCEMENT MOSFET BIASING (P-CHANNEL DEVICE SHOWN)

Note, however, that although an improvement is gained in terms of drain-source current stability, it is actually at the cost of transconductance stability. Because of the lower slope of the bias line, devices near the upper limit of spread range are biased at proportionally lower points than those near the lower limit.

Fairly obviously the bias requirements for minimum Ids variation are somewhat in conflict with those for minimum transconductance variation, so that the designer must generally select his bias conditions to favour whichever of the two is of greater importance in the application concerned. Where both are equally important it is usually necessary to arrange a "compromise" bias situation, and use selected FET devices whose parameters are confined to a sufficiently narrow spread range. Naturally this course is adopted only in critical applications, as selected devices are generally rather costly.

Because the application of forward bias to the gate of a FET tends to reduce the drain-gate bias, the designer using the bias method of figure 9.4 must be careful to ensure that all devices will operate in the pentode region with Vdg>Vp. This can sometimes pose problems, as the resistance of a load in the drain circuit tends to reduce the actual drain voltage, while the drain supply voltage Vdss is generally limited by device breakdown considerations.

The simple self-bias method of figures 9.2 and 9.3 cannot be used for enhancement or type C MOSFETs, because it may be remembered that these devices are normally "off." A fixed bias component is therefore always required for such devices when used in linear circuitry. However, this requirement nevertheless permits the use of the composite bias method of figure 9.4 and, in fact, this method is that generally used with type C MOSFETs because of the wide parameter spreads encountered.

Figure 9.5 shows this biasing method as used for type C MOSFETs, together with the corresponding composite transfer characteristic and bias line. The reader may care to compare these with figure 9.4.

Having considered briefly some of the general aspects of FET device usage, let us now turn to consider more closely some of the more common circuit applications of these devices.

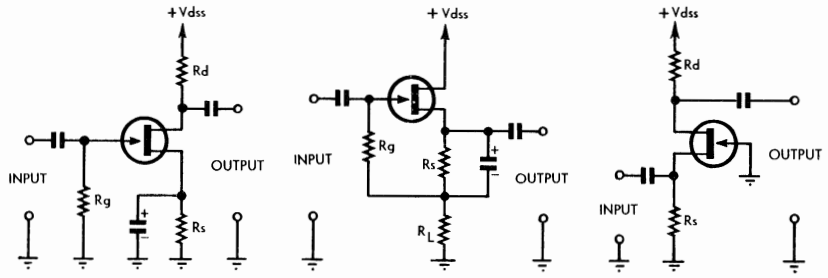


Figure 9.6

Possibly the most familiar type of circuitry in which FETs are found is that of **low frequency or "audio" amplifiers.** Both the JFET and the various types of MOSFET are eminently suited for use in this type of circuitry, their unique combination of high input impedance and significant power gain making them the solid-state equivalent of the familiar thermionic valve.

As with the thermionic valve, there are three general amplifier circuit configurations in which FET devices may be used. These are known respectively as the **common source, common drain** and **common gate** configurations, and are illustrated in basic N-channel JFET form in figure 9.6. It may be noted that the common drain configuration is also known as the "source follower" configuration.

The common source configuration may be seen to be the FET equivalent to the familiar common cathode thermionic valve stage. Here the input signal to the device is applied to the gate electrode, while a load resistor Rd in the drain lead develops the amplified output signal. The source resistor Rs is normally bypassed to ground as shown, to prevent signal negative feedback or "degeneration."

Typically the voltage gain which may be obtained from such a stage is rather modest compared with the corresponding valve configuration. This is both because current JFET devices have a somewhat lower transconductance, on the average, and because parameter spread and voltage breakdown limitations generally restrict the value of load resistor Rd to a relatively low value.

With JFET devices this type of stage also tends to have a rather restricted high-frequency response, due to effective magnification of the drain-gate capacitance Cdg by the Miller effect. For any signal at the gate of the device there is at the drain electrode an amplified replica of opposite phase, so that the total effective signal voltage across Cdg is (A + 1) times the signal amplitude at the gate. As a result the current drawn by Cdg is quite high, equivalent in fact to a capacitor of value (A + 1).Cdg connected from gate to ground.

Despite the limitations on gain and frequency response, the common source configuration is a very useful one, and is finding increasing use in modern audio equipment. It provides useful voltage gain while offering the high input impedance characteristic of thermionic valve stages, together with a very low noise level.

The common drain or source-follower configuration is very similar to the familiar cathode follower stage, and finds corresponding circuit applications. While providing slightly less than unity voltage gain it offers appreciable current gain, and is therefore well suited for impedance transformation. Typical JFET source follower stages may present input impedances of 30M or higher shunted by a few pF, and output impedances as low as 100 ohms. Higher input impedance values again may be obtained either by means of more elaborate source-follower circuits, or by using MOSFET devices.

The common-gate configuration is probably that least used for audio amplification, although it is quite frequently used in RF circuitry as may become evident shortly. It offers voltage gain but no current gain, and hence does not take advantage of the inherently high input resistance of the device itself. The main advantage offered is high frequency response, as the grounded gate acts as a shield

between input and output and prevents degenerative feedback.

It should be noted that although the basic circuits shown in figure 9.6 employ N-channel JFET devices, they are equally suitable for both P-channel JFETs and the various types of MOSFET. Also although the circuits are shown with the simple self-bias system of figure 9.2, which is often quite adequate, they may also be arranged with the composite biasing system of figures 9.4 and 9.5.

DC amplifiers form a second important application of FET devices. Such amplifiers are used in analog computers, control systems, electronic voltmeters, electrometers and other measuring instruments. FETs are quite well suited for this type of application by virtue of their high input impedance and the inherently good temperature stability of their parameters.

A common configuration used in DC amplifiers using FETs is the differential amplifier, also called the "long-tailed pair." This is a balanced circuit in which two devices are connected in basically common-source configuration, but sharing a single source impedance. The balanced nature of the circuit tends to cancel out any drift of the device parameters with aging and temperature variations, resulting in a high degree of stability. The shared source impedance may be either a resistor, for relatively non-critical applications, or alternatively another FET device connected as a constant-current load.

Another very important application of FET devices is in **RF circuitry,** where they are used as amplifiers, oscillators, mixers and frequency multipliers. JFET and particularly MOSFET devices are well suited for most RF applications by virtue of their high power gain, high input and output impedances (for low tuned circuit loading) and low internal feedback.

A further advantage of FET devices for RF amplifier and mixer applications is that they are generally capable of significantly improved cross-modulation performance compared with both thermionic valves and the more established bipolar transistors. This arises from the shape of the transfer characteristic, which for FET devices is significantly closer to the ideal "square law" relationship.

Examples of typical FET device RF applications are shown in figure 9.7. The circuits of (a), (b) and (c) illustrate RF amplifier configurations, shown here using various MOSFET devices, while the circuit (d) is of a simple self-excited L-C oscillator using a JFET device.

Of the RF amplifier configurations shown in figure 9.9 (a) and (b), which are those most commonly used for single-gate JFET and MOSFET devices, the common-source circuit generally makes best use of all the device parameters. However, with JFETs and to a lesser extent with MOSFETs, the feedback action of the internal drain-gate capacitance Cdg tends to make neutralisation necessary if full gain and adequate stability are to be obtained. This applies particularly at very high and ultra-high frequencies.

A common method of applying neutralisation in fixed tuned, narrow band amplifiers is by means of an inductor connected between the drain and gate electrodes of the device in series with a DC blocking capacitor, as shown. Basically the inductor produces parallel resonance with Cdg at the operating frequency, and thus cancels the feedback action.

The common-gate configuration of figure 9.7(b) needs no neutralisation, as the gate and substrate are both grounded and therefore act as a shield between input and output. However, this configuration has no current gain, and does not take advantage of the high input impedance of the device. It is used mainly at frequencies above the range in which stable operation may be obtained with the common emitter circuit.

Virtually all the advantages of the common-source and common-gate configurations are combined in the so-called "cascode" circuit, which is illustrated in figure 9.7(c). This is basically a combination of the two previous circuits, with the output of the common source stage untuned and connected directly to the input of the common-gate stage. It is a configuration for which the recently developed dual-gate MOSFET is particularly well suited, as may be seen, because this device is virtually two devices in one. However, the cascode circuit is also often used with single gate JFET and MOSFET device pairs.

The circuit of figure 9.7(d) illustrates a simple L-C oscillator using a JFET device. The configuration shown is that of the classical "Hartley" or tapped coil oscillator, and it may be seen that the JFET version of this oscillator is very similar to that using a thermionic triode. A "gate leak" R-C combination in the gate lead develops signal-derived bias as a result of gate current flow on signal peaks.

The MOSFET version of this and other oscillator circuits tends to be slightly more complex, as it may be remembered that MOSFETs cannot draw gate current without sustaining damage. It is generally necessary either to provide fixed bias, or to provide a diode detector circuit to generate signal-derived bias.

There are many interesting and important applications of FET devices in addition to those which have been mentioned in this chapter. These include ultra-long period timing circuits, sample and hold circuits, chopping and analog signal switching, and in circuits requiring voltage-controlled resistors and current regulating elements. Unfortunately space restrictions prevent discussion of these further applications here even in a brief and cursory manner; however, it is hoped that the foregoing will have given the reader at least a satisfying introduction to the many and varied applications of these devices.

Further information concerning both the applications discussed in the foregoing and those which it has not been possible to discuss will be found by interested readers in the references listed below.

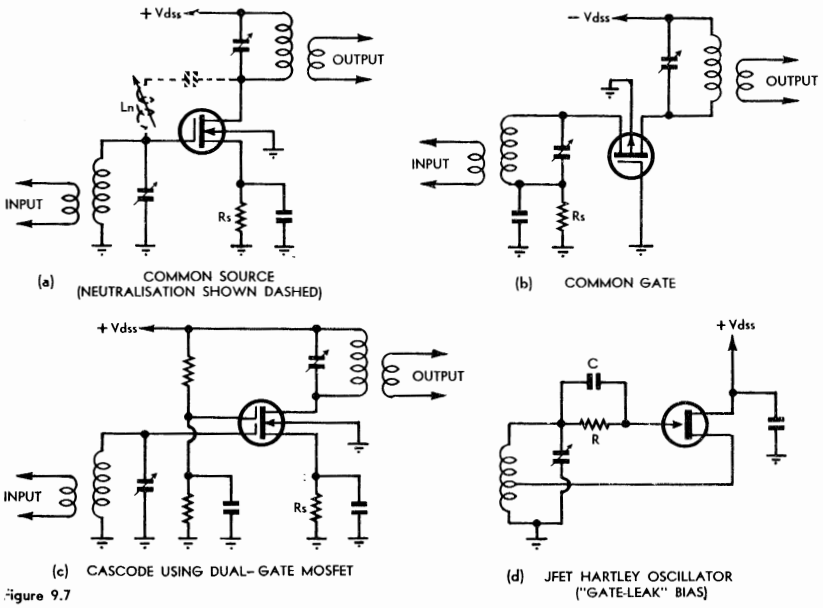


Figure 9.7

SUGGESTED FURTHER READING

BRAZEE, J. G., **Semiconductor and Tube Electronics,** 1968. Holt, Rinehart and Winston, Inc., New York.

EIMBINDER, J., **FET Applications Handbook,** 1967. Tab Books, Blue Ridge Summit, Pennsylvania.

GRISWOLD, D. M., "Understanding and Using the MOSFET," in **Electronics,** V.37, No. 31, December 14, 1964.

NOLL, E. M., **FET Principles, Experiments and Projects,** 1968. Howard W. Sams, Inc., Indianapolis.

SCHULTZ, J. J., "The Dual-Gate MOSFET," in **CQ,** V.24, No. 12, December 1968.

Chapter 10

THE BIPOLAR TRANSISTOR

The bipolar transistor — NPN and PNP forms — basic configuration — equilibrium conditions — collector-emitter bias alone — the effect of forward emitter-base bias — minority carrier injection — base diffusion, and collection — the device as a power amplifier — factors affecting the gain — the gain factors alpha and beta — characteristic curves — the common base characteristic — the common emitter characteristic.

Let us now turn our attention to a fourth basic semiconductor device, one whose development in fact marked one of the most important milestones in the history of the electronics industry: the **bipolar transistor.**

Although in earlier chapters we have examined other "transistor" devices such as the unijunction and the FET prior to the present introduction of the bipolar transistor, this order of presentation has been arranged by the author mainly in an effort to assist the reader by dealing with concepts and devices in a logical and progressive manner. In consequence, the presentation order is quite unrelated to the chronological order in which the devices made their appearance.

Actually the bipolar transistor was the first practical transistor device to be developed. The first crude working model was built in December, 1947, at the Bell Telephone Laboratories by physicists William Shockley, John Bardeen and Walter Brattain, and announced in the "New York Times" of July 1, 1948. This despite the theoretical proposition of the FET device some 20 years earlier, as noted previously in chapter 8.

In fact the name "transistor" was coined by Bell Labs expressly for the bipolar device. It was intended to describe the operation of the device, being a contraction of the words "Transfer" and "resistor." However, although the term "transistor" is still widely used to signify the bipolar device in particular, it has also become widely used as a generic name applied to any three-element semiconductor device capable of power amplification. The result has been ambiguity, and accordingly it has become common practice to use the term "bipolar transistor" to signify the specific device. For this reason the latter term will be that used in the present and subsequent chapters.

Basically, the bipolar transistor consists of a device having three functional semiconductor regions. The three regions are arranged such that there are two junctions, sharing one of the regions and approximately parallels a thin common region. This means that the unshared regions of the device must be of semiconductor **type,** because they will both be of opposite type to the shared common region.

From this definition and from the somewhat schematic representations shown in figure 10.1 it may be seen that, like the unijunction and the FET, bipolar transistors may be constructed in either of two complementary forms. One form is that having an "NPN" configuration, in which the shared common region is of P-type material while the outer regions are of N-type material. The other or "PNP" form has the converse arrangement.

As one might well expect from previous chapters, both varieties of the device operate in virtually identical fashion except that the polarities of operating voltages and currents are opposite, and the roles played by the various current carriers are reversed. In the case of the NPN variety, conduction band electrons play the major part in device operation, whereas with the PNP variety it is valence band holes which play the major role.

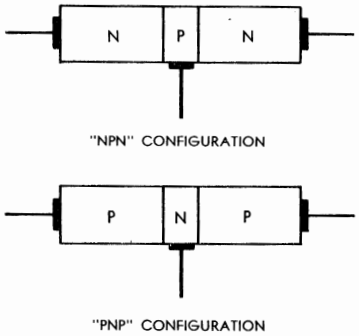


Figure 10.1

Because they operate in virtually identical fashion, the performance of NPN and PNP bipolar transistors of the same semiconductor material and geometry is almost identical. It is true, however, that there are subtle differences in performance due to such theoretical factors as the differing mobilities of electrons and holes. It is also true that there are practical manufacturing problems which tend to make it harder to achieve some aspects of device performance with one form compared with the other.

Despite these factors, at the present state of the semiconductor art there are many types of bipolar transistor which are available in NPN and PNP versions sufficiently comparable in performance to be virtually interchangeable in a large number of applications. The appropriate power supply polarities must be provided in each case, of course.

Although the foregoing might seem to suggest that one form of the device could perhaps be adopted for universal use, and the other virtually ignored, there are a number of reasons why this is not done. One is that there are many applications in which a particular combination of circuit configuration and power supply polarity strongly favours one form compared with the other. Broadly speaking, there are as many such applications favouring one form as there are favouring the other, so that in general it is both convenient and economical to have devices available in both forms.

A further reason is that in critical applications demanding high performance in terms of certain parameters, one device form can offer distinct advantages compared with the other as a result of the subtle differences between the two noted above. An example is in UHF power amplifiers, where NPN devices tend to be more attractive largely as a result of the higher mobility of electrons relative to that of holes. In other cases PNP devices may be more attractive.

There is also the important reason that with both varieties of the device available, it becomes possible to produce novel and highly efficient circuitry which employs both types of device and exploits their complementary behaviour. This is a very worthwhile and dramatic advantage which the bipolar transistor and other semiconductor devices, such as the unijunction and the FET, offer in comparison with the thermionic valve.

As with the other semiconductor devices which we have examined in previous chapters, the bipolar transistor may be fabricated from a variety of semiconductor materials; although to date only germanium and silicon have been used to any appreciable extent. Germanium was used for the first devices developed. and for those first marketed, and is still used to a small extent for very high current switching devices and very high frequency amplifiers. However, the great majority of bipolar transistors now manufactured and used are fabricated from silicon, largely because of the superior leakage behaviour and high temperature performance offered by this material.

Because of the similarity between the NPN and PNP forms of the bipolar

transistor, it is really only necessary to consider one form when examining the fundamentals of device operation. Once the basic concepts of device operation are grasped with respect to one form, the operation of the other form may be deduced simply by exchanging the roles of the current carriers, and reversing the polarities of all voltages and currents.

In the treatment which follows, the PNP form of the device is used as the basis for discussion, primarily to assist the reader in becoming more familiar with the behaviour of holes as current carriers. However, upon completion of the chapter the reader may care to deduce for himself the corresponding picture of NPN device operation, in order to test and reinforce his understanding.

An elementary PNP bipolar transistor is shown in figure 10.2 (a). It may be seen that the thin central N-type region shared by, and between the two P-N junctions, is called the **base** (B), while the two outer P-type regions are called the **emitter** (E) and **collector** C).

The same names are used for the corresponding regions of the NPN variety of the device. In both cases the term "base" refers to the thin central region, which is lightly doped, while the terms "emitter" and "collector" refer respectively to the heavily and lightly doped end regions. The significance of these terms should become apparent later in this chapter.

Superficially, as may be seen, the bipolar device consists basically of two P-N junctions arranged in a "back-to-back" or inverse series configuration, with the common connection brought out as the base electrode. From this the reader might be led to infer that its behaviour would be very similar to that of a pair of simple P-N diodes connected in a similar manner. However this is only true in a very limited sense indeed.

If one simply ignores either of the two device junctions and its associated P-type region, and proceeds to examine the behaviour of the remaining junction alone, then that behaviour will in general be identical with that of a normal P-N diode. The results of such a test will be the same for either junction, so that in this rather artificial and limited sense the device is in fact equivalent to a pair of simple diodes connected back-to-back. Yet almost the only practical significance of this fact is that it may be used as the basis of simple tests for device damage.

Of far greater significance is the fact that as soon as both junctions are permitted to play an effective part in determining device behaviour, its behaviour tends to depart quite markedly from that of a pair of simple diodes. And the reason for this is that the thin base region shared by both junctions causes a significant interaction between the two.

It is, in fact, this interaction between the behaviour of the two junctions which is responsible for virtually all of the highly useful properties of the bipolar transistor, and which therefore we must proceed to examine in some detail. But before we do so, it may be worthwhile for the reader to refresh his understanding of P-N junction behaviour by referring to the diagrams of figure 10.2 (b) and (c).

Figure 10.2 (b) shows the energy level diagram corresponding to the equilibrium condition of the elementary transistor shown in (a). As before, the equilibrium situation represented is that which applies with zero external bias applied to the device electrodes, and is thus the dynamic balance reached between the "internal" carrier movements due to the opposing mechanisms of diffusion and drift. The average carrier energy level or Fermi level (Ef) is constant throughout the material.

As part of the equilibrium, depletion layers are formed in the vicinity of each junction, extending into the material on either side to a degree inversely proportional to the doping levels. Drift fields are set up across the depletion layers, as before, with the P-type side of the depletion layers acquiring a negative charge, and the N-type side a positive charge. This produces a distribution of electrostatic potential throughout the device, of the shape shown in figure 10.2 (c).

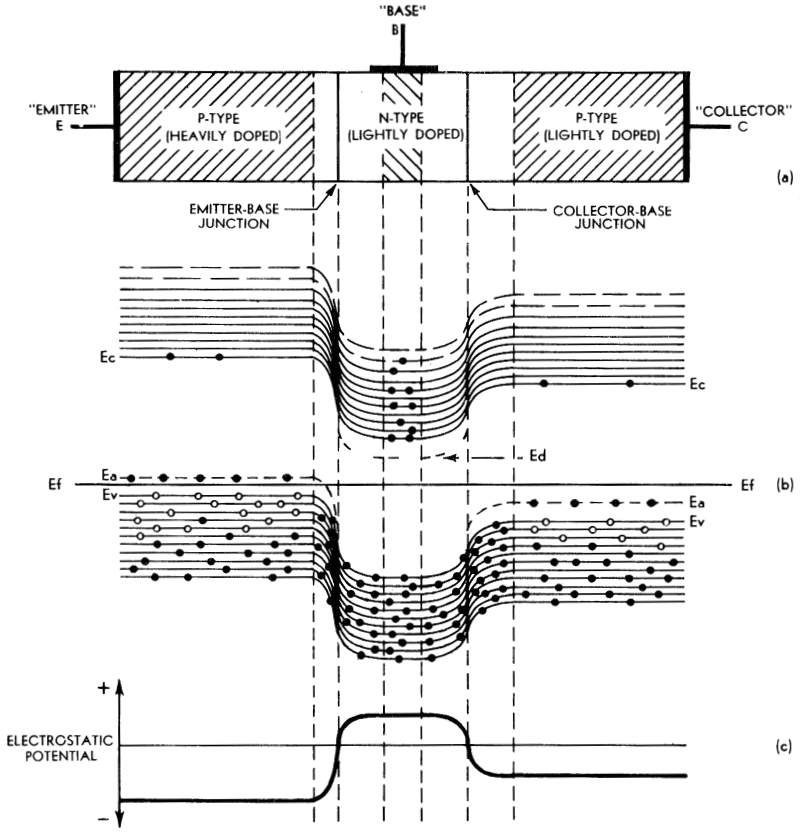


Figure 10.2

From previous chapters, the reader may be able to recognise that this potential distribution has two important implications regarding the movement of current carriers within the device.

One important implication is that, in the equilibrium situation, both P-N junctions of the device represent potential barriers to the MAJORITY CARRIERS in each of the three semiconductor regions.

Hence the majority carriers (holes) in both the emitter and collector regions of the PNP device shown effectively "see" the base region as a potential "mountain" separating these two regions, this "mountain" having the cross-section suggested by the curve of figure 10.2 (c) viewed normally. In order to pass from one P-type region to the other under equilibrium conditions, such carriers must first surmount the "up-hill" slope of the nearest junction.

Similarly the majority carriers (electrons) in the base region effectively "see" themselves at the bottom of a potential "valley" in the centre of the device, as may be seen if the diagram is viewed upside-down. In order to pass from the base to either of the two other regions, in the equilibrium situation, these carriers must surmount the "up-hill" potential slope of the junction concerned.

In short, then, any flow of majority current carriers across the device junctions constitutes a movement of these carriers in opposition to a depletion layer potential barrier. It may be recalled from chapter 4 that this type of majority carrier flow is normally known as a **diffusion current.**

The second important implication of the diagram of figure 10.2 (c) is really the converse of the first. In the same equilibrium situation, both P-N junctions of the device represent "downhill" potential slopes to the MINORITY CARRIERS in each of the three device regions.

Hence in the PNP device shown, minority carrier holes in the base region "see" themselves at the top of a relatively high potential "plateau," and can "roll down" a potential slope into the emitter or collector regions if they diffuse into the depletion layer region of the appropriate junction. Similarly, minority carrier electrons in the emitter and collector regions "see" the base region as a narrow "gorge" running between the two outer regions, and accordingly roll into the base to diffuse into the appropriate depletion layer.

Minority carrier flow across the junctions of the device may thus

scribed as **drift currents,** as they represent a carrier flow influenced by and in the direction of the "inbuilt" electric field across each junction depletion layer.

From previous chapters it may be recalled that the only minority carriers present in a doped semiconductor region under equilibrium conditions are those generated by the "intrinsic" excitation mechanism. Hence in the bipolar transistor for the equilibrium conditions, the minority carrier drift currents moving "down" each depletion layer potential slope will be at a very low level, and **saturated** in the sense that they are determined almost completely by the number of carriers produced by excitation.

Since by definition there can be no nett current flow in a device in the equilibrium condition, the foregoing implies that the majority carrier diffusion currents moving in opposition to the minority carrier currents, or "up" each depletion layer slope, must also be at a very low level.

Thus while there are currents flowing across both junctions of the bipolar transistor in the equilibrium situation, as an integral part of the equilibrium, these currents are quite small and consist in both cases of equal and opposite majority and minority carrier currents. In this respect the bipolar transistor does not differ from the other junctions semiconductor devices examined in earlier chapters.

It is hoped that the foregoing review of the equilibrium condition of the device has assisted the reader in refreshing his understanding of basic P-N junction behaviour. Now let us turn to consider what happens when the equilibrium of the bipolar transistor is disturbed by the application of external bias voltages.

If one connects a DC bias voltage between the **emitter** and **collector** electrodes of the device of figure 10.2, of either polarity, only a very small current is found to flow. This is perhaps not surprising, because in this case we are effectively connecting the external bias across both junctions in series, and for either polarity of the bias one of the junctions will be reverse biased.

The internal situation of the device will, in fact, be only slightly different from that for equilibrium. The depletion layer of the reverse biased junction will widen to correspond to the increased electric field across it. This in turn will cause a small nett current to flow, as there will be a reduction in the majority carrier diffusion currents while the minority carrier drift currents will remain fixed at their saturation values.

The small current which flows through the reverse-biased junction will cause a very small voltage drop to appear across the other junction, which will naturally be forward biased. Accordingly the depletion layer of this junction will narrow very slightly, and its majority carrier diffusion currents will increase to correspond to the small nett current flow.

It may be seen that when a single external DC bias is connected between the collector and emitter electrodes of a bipolar transistor, the device behaviour for both applied polarities is basically very similar to that of a normal P-N junction diode when reverse biased.

However the small current which flows through the device in this situation is actually slightly greater than the reverse bias current of a normal P-N diode, or of either device junction alone. And although the difference in current levels is only slight, it is of great significance, for it provides the key to the really interesting and unique aspects of bipolar transistor operation.

The alert reader may have already noticed an interesting fact about the small current which we have seen to flow through the device when biased between emitter and collector. This is that the current which does flow effectively "changes its nature" during its passage through the device, even though its magnitude necessarily remains constant. When passing through the depletion layer of the reverse-biased junction, we saw it to consist of **drift currents** derived from the minority carrier populations of the two adjacent regions; yet when it passes through the depletion layer of the forward-biased junction it then consists of **diffusion currents** derived from the adjacent majority carrier populations.

At first sight this may seem not only confusing, but downright contradictory. Yet the explanation is really quite simple: when carriers move from one semiconductor region to another across a P-N junction, of necessity they undergo a reversal in terms of population statistics — simply because the majority/minority roles in each type of semiconductor material are the opposite of those in the other type.

Hence a carrier initially belonging to the majority population of one region will automatically become a minority carrier when it crosses a junction into a region of opposite type; and vice-versa.

Perhaps the most obvious implication of this, in the case of our foregoing biased transistor, is that the minority carrier currents flowing across the reverse-biased junction actually consist of the same types of carrier moving in the same directions as the majority carrier currents flowing across the forward-biased junction. Only the "labels" applied to the carriers change, in other words. This should clear up the apparent contradiction.

However, a second and more important implication is that those majority carriers (holes) which diffuse across the forward-biased junction into the base region will naturally then become minority carriers in that region, boosting the minority carrier population of the base above its normal level. And because the base region is relatively thin, many of these additional minority carrier holes will be able to diffuse across to the depletion layer of the re-

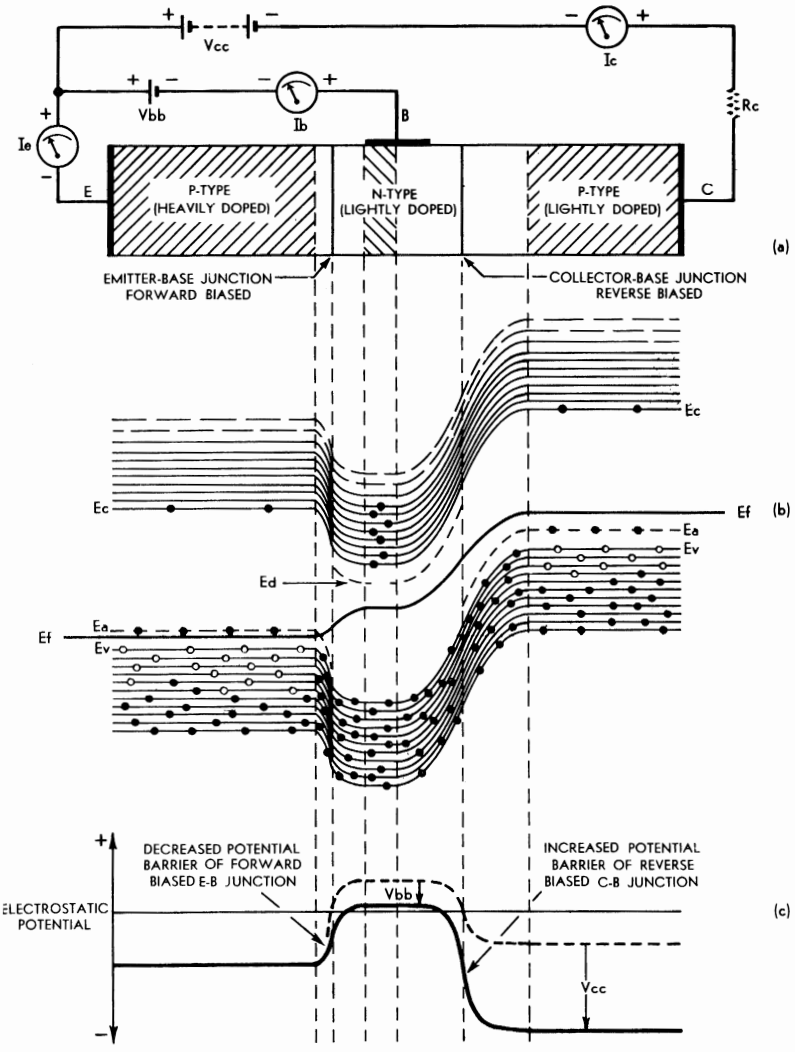


Figure 10.3

verse-biased junction before their passage can be interrupted by a recombination with a majority carrier electron.

Naturally those holes which do reach the reverse-biased junction depletion layer will be "grabbed" by the electric field of the latter and will accordingly drift into the far P-type region. The minority carrier hole current of the reverse-biased junction will thus be boosted to a value greater than its normal saturated value, and this explains why the net current passed by the device is higher than that of a normal reverse-biased junction.

The mechanism just described may be seen to involve a novel type of interaction between the two junctions of the device: the current passed by one junction in the reverse-biased state is increased significantly from its "normal" saturated value as a result of excess minority carriers injected into the narrow base region by the forward-bias of the adjacent junction.

This is, in fact, the precise "interaction" mechanism which is responsible for the really important behaviour of the bipolar transistor, and to which reference was made earlier. The ability of one forward biased junction to influence the conduction of the other reverse-biased junction gives the device the ability to perform power amplification, although of a slightly different type to that of the thermionic valve or the field effect transistor.

In figure 10.3(a) is shown an elementary PNP transistor with external bias voltages applied to it, in a fashion which should help in understanding how the device is capable of amplification.

One of the applied voltages, that labelled "Vcc," may be seen to correspond to the collector-emitter bias which we have just considered in the foregoing discussion. With this bias applied alone, as we have seen, the device will draw only a small current slightly larger than the current drawn by a reverse-biased P-N junction. We might thus expect the meter measuring collector current Ic to register only a current of this low order.

The other bias voltage applied to the device is labeled "Vbb," and may be seen to be applied between the base and emitter with a polarity which forward biases the base-emitter junction. Superficially one might therefore expect a considerable current flow to be registered by both the meter measuring base current Ib and that measuring emitter current Ie.

Perhaps surprisingly, this is not quite what is found. The meter in the emitter lead is certainly found to register an appreciable current, as expected. However, the meter in the base lead is found to register only a small fraction of the emitter current, while instead the meter in the collector lead is found to register virtually the same current as the emitter meter, despite the reverse bias on the collector-base junction!

The reason for this behaviour is simply that the majority carrier hole current passed by the forward biased base-emitter junction "injects" excess minority carrier holes into the base, as before. Because of the narrowness of the base region, a major proportion of these holes are able to reach the depletion layer of the reverse biased collector-base junction before being stopped by recombination with a majority carrier electron. Those which do reach the second junction will naturally "roll down" the potential gradient into the collector.

As the emitter region is relatively heavily doped whereas the base region is quite lightly doped, the majority carrier diffusion current passed by the emitter-base junction consists mainly of holes moving from emitter to base rather than electrons moving from base to emitter. Also the light doping in the base region results in a relatively low population of majority carrier electrons in the base, so that most of the minority carrier holes injected from the emitter escape recombination and do actually end up in the collector region. This explains why the emitter and collector currents are almost equal.

The base current is actually equal to the slight difference between the two, because in effect it consists only of the small number of electrons necessary to replace those majority carrier electrons in the base region which do actually meet and recombine with an injected hole. This accordingly explains why the base current is very small.

From the foregoing it may be seen that the operation of the bipolar transistor is basically a combination of three simple processes: injection, diffusion and collection. The forward-biased emitter-base junction **injects** minority carriers into the base, whereupon these carriers **diffuse** away through the base, because of the localised concentration. Those which manage to reach the reverse biased collector-base junction without recombining with a majority carrier are then **collected** by that junction.

The effect of forward bias Vbb applied across the base-emitter junction, then, is to cause a marked increase in the current flowing through the reverse biased collector-base junction. And as one might expect, the increase in collector current is highly dependent upon the magnitude of the applied base-emitter bias. In fact only a small bias Vbb is found sufficient to produce quite a large collector current Ic, and relatively small variations in Vbb tend to produce large variations in Ic.

As the collector-base junction is reverse-biased, the collector supply voltage Vcc may be considerably larger than Vbb. This allows a fairly large load resistor Rc to be connected in series with the collector, as shown. Changes in the low voltage/low current (and hence low power) conditions at the base of the device can thus be arranged to cause large changes in both the voltage and current levels at the collector. From this it may be seen that the bipolar transistor is capable of providing appreciable **power amplification.**

It should now be fairly clear why the names "emitter" and "collector" are used for the two end regions of the bipolar device. The name "emitter" is surely quite appropriate for the region concerned, which does effectively **emit** or inject carriers into the base region. Similarly the name "collector" is quite appropriate also, for the region concerned does effectively **collect** those of the injected carriers which diffuse across the base region.

The name "base" is applied to the central region of the device largely as a historical carry-over from the very first type of bipolar transistor, the long obsolete "point contact" variety. With this type of transistor the base region did in fact form the physical base of the functional device. However, with most modern forms of the bipolar transistor this is not the case, and the term "base" is merely a conventional label.

At this point the reader may find it worthwhile to examine the diagrams of figure 10.3 (b) and (c), which show respectively the energy band diagram and electrostatic potential distributions for the biased transistor of figure 10.3 (a). Comparison of these diagrams with the corresponding diagrams for the equilibrium situation, given in figure 10.2, may help in clarifying the foregoing discussion.

As we have seen, the base current of a biased bipolar transistor consists basically of the small number of electrons necessary to replace those majority carriers in the base region "absorbed" by recombination with injected minority carriers. As such, it is numerically equal to the difference between the emitter and collector currents. From this it follows that the larger can be made the fraction of emitter current reaching the collector, the smaller will be the base current and the higher the potential amplification of the device.

Naturally one factor which has an important effect on the proportion of emitter current reaching the collector is the actual composition of the emitter junction current. From the foregoing discussion, it should be fairly clear that it is only the component of emitter current which consists of carriers moving from emitter to base which plays a part

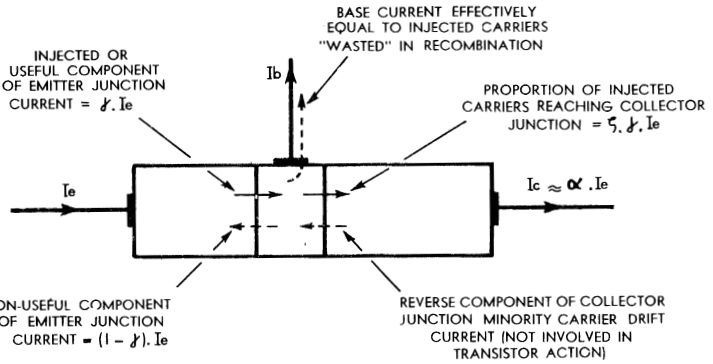


Figure 10.4

in the operation of the device; the component which consists of carriers moving in the opposite direction plays no useful part.

It is for this reason that the ratio between the emitter and base impurity doping levels is made very high, ensuring that the emitter junction current consists almost entirely of carriers moving from emitter to base. In the PNP transistor, as we have seen, these carriers are holes; conversely in the case of the NPN transistor they are electrons. The ratio of the emitter-to-base component of emitter current to the total emitter current is known as the **emitter injection efficiency or ratio,** which term the reader may recall was also used in chapter 7, in connection with the unijunction.

Apart from the emitter injection ratio, a second important factor governing device amplification is the number of minority carriers injected into the base which are able to diffuse through to the collector junction depletion layer without meeting with a majority carrier and "falling by the wayside" due to recombination. Expressed as a proportion of the total injected carriers, this is known as the **base transport efficiency.**

It is to ensure a high base transport efficiency that the base region is made very narrow; the shorter the distance which the minority carriers must travel through the base, the lower the likelihood of recombination. The base transport efficiency is also improved by making the impurity doping level of the base as low as possible, to ensure a relatively small population of base region majority carriers.

The schematic diagram of figure 10.4 may help the reader to visualise the significance of the emitter injection efficiency and base transport efficiency factors in terms of basic device operation.

It is possible to express the amplification action of the bipolar transistor in terms of two different gain factors, both of which are actually current increment ratios. One of these, and the earliest to be used, is called α (alpha), and is simply an expression of the ratio between **collector and emitter currents,** in terms of changes or small increments:

$$\alpha = \frac{dIc}{dIe} \ (dVcb=0) \quad . \ . \ . \ (10.1)$$

Here "dIe" is a small change in emitter current, "dIc" is the corresponding change in collector current, and the expression in brackets specifies that the collector-base voltage is defined as constant.

Fairly obviously alpha is always less than one, because the collector current is always less than the emitter current. In fact alpha is a fraction which approaches unity asymptotically as the amplification increases: a very low gain device may have an alpha of 0.91, while a very high gain device might have an alpha of 0.998.

The numerical values of alpha for practical devices tend to be restricted to the range between these two examples, and are progressively cramped for increasing orders of amplification. Because of this it is usually more convenient to use the second gain factor called alternatively β (beta) or hfe.

Beta is simply the device amplification expressed as the ratio between **collector and base currents,** again in terms of small increments:

$$\beta = \frac{dIc}{dIb} \ (dVce=0) \quad . \ . \ . \ (10.2)$$

Here "dIb" means a small change in base current, "dIc" again means the corresponding change in collector current, and the expression in brackets has a similar meaning to that in (10.1).

Naturally since Ib is simply the difference between Ie and Ic, alpha and beta have a fixed relationship to one another. A short calculation reveals that the relationship is as follows:

$$\beta = \frac{\alpha}{1 - \alpha} \quad . \ . \ . \ (10.3)$$

Beta normally has a value somewhat greater than unity, and does not become cramped at high orders of amplification. For example, devices having alpha values of 0.91 and 0.998 quoted above would have corresponding beta values of 10 and 500 respectively. The reader may care to work these out for himself using equation 10.3.

Just as there are two different gain factors which may be used to express the amplification action of the bipolar transistor, there are also two main ways of representing the behaviour of such a device in terms of graphical characteristics. These will now be briefly examined.

Figure 10.5 illustrates one of the two main types of bipolar transistor collector characteristic, which in this case displays the amplification behaviour of the device in terms of the relationship between collector current and voltage for various values of **emitter current.** As may be seen, this is known as the **common base characteristic.**

Perhaps the first thing which the reader may note from the diagram is that the major portion of each curve in the characteristic "family" is almost horizontal, at a collector current value which corresponds in each case to a fraction alpha of the emitter current value. The curves thus bear a strong resemblance to those of the FET and the thermionic pentode valve, and indicate that like these devices the bipolar transistor has a relatively high output resistance.

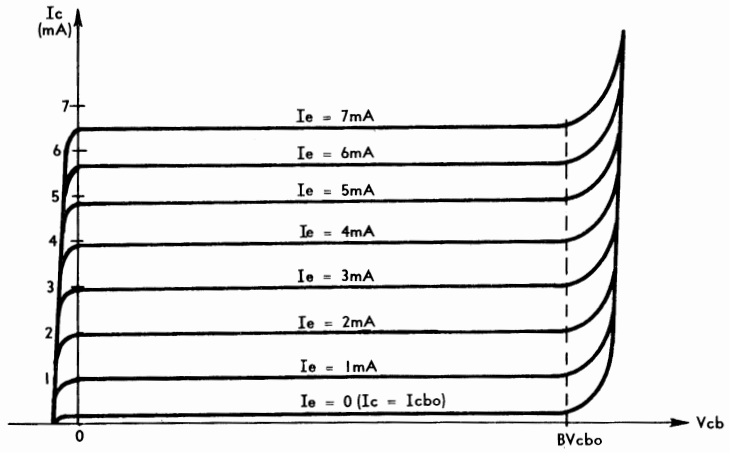


Figure 10.5 "COMMON BASE CHARACTERISTIC"

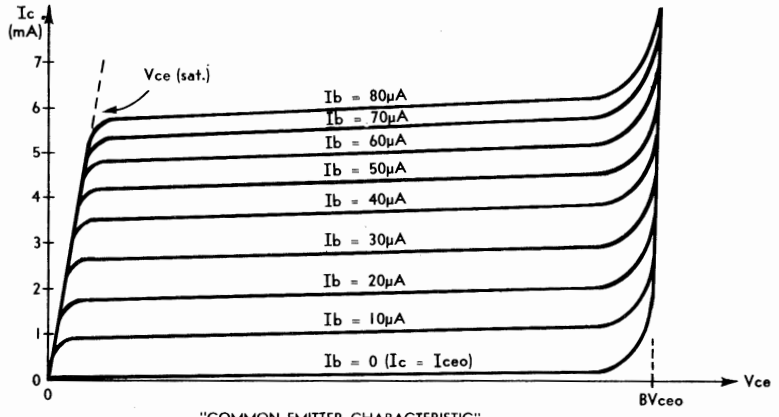


Figure 10.6 "COMMON EMITTER CHARACTERISTIC"

The curves have this basic shape because the injection-diffusion-collection mechanism which we have seen to be responsible for device operation is basically almost independent of the magnitude of collector voltage. As long as the collector-base junction is not forward biased, its depletion layer is capable of "collecting" virtually all of the injected carriers which diffuse across the base. The collector current which flows as a result of forward emitter-base bias is thus effectively "saturated."

If the collector-base junction is in fact forward biased, normal device operation does naturally cease. This explains why the curves of figure 10.5 drop down sharply on the left-hand side. All curves in the family fall to

zero current when the forward collector-base bias has been increased to approximately 0.3V for germanium, or 0.6V for silicon.

The other limit to the "pentode" region of device operation occurs when the reverse bias on the collector-emitter junction is increased to the point where this junction enters avalanche breakdown. As with any other P-N junction, the current then increases rapidly due to avalanche carrier collision. Normal device operation again ceases, and the device enters a high dissipation mode of operation. The collector-base voltage at which avalanche begins is called the **collector-base breakdown voltage,** symbolised BVcbo.

In the "common base" configuration in which the curves of figure 10.5 are measured, the collector-base junction is biased by a voltage applied directly between collector and base (Vcb). Because of this, the device does not amplify its own reverse bias current when the external base-emitter bias is zero (Ie = 0). The lowest of the curves therefore corresponds to the "normal" reverse bias current of the collector-base junction, known as the **collector-base saturation current (Icbo).**

It may be noted that the curves of figure 10.5 actually slope upward slightly in the direction of increasing collector voltage. The reason for this is that as the collector-base voltage is increased, the depletion layer of the collector-base junction naturally widens, and this reduces the effective thickness of the base. The amplification therefore increases slightly, due to lower carrier recombination and increased base transport efficiency.

A further point to note is that although the curves for low values of emitter current are spaced apart by almost exactly the same intervals of collector current, those for higher values become cramped together. In other words, the amplification action of the bipolar transistor droops at high values of emitter and collector current.

The primary reason for this is that at high current levels the emitter injection efficiency of the device tends to fall as a result of the large number of minority carriers present in the base region. In effect, the number of minority carriers becomes so great that the emitter-base diffusion current component tends to fall, while the base-emitter component rises to compensate.

The other main type of collector characteristic used for the bipolar transistor is that known as the **common emitter characteristic**, illustrated in figure 10.6. As may be seen, this displays the amplification action of the device in terms of the relationship between collector current and voltage for various **base current** levels.

The common emitter curves are rather similar to those for the common base configuration, as may be seen, the main difference being that the value of base current corresponding to each curve is very much smaller than the value of emitter current. In effect, the common-emitter curves display beta, whereas the common-base curves display alpha.

The curves may be seen to have a more pronounced slope than those of the common-base characteristic, and this is again due to the reduction of the effective base thickness as the collector-base junction depletion layer extends with increasing collector-base voltage.

An array of modern silicon transistors, reproduced approximately actual size. The largest device, when bolted to a suitable heatsink, is capable of dissipating up to 120 watts. The smallest shown is a subminiature type for VHF use, capable of dissipating only a few tens of milliwatts.

The effect is more dramatic in this case merely because beta is a more sensitive indicator of the amplification mechanism.

As before, the curves may be seen to cramp together at higher collector current levels, indicating the droop in amplification due to falling emitter injection efficiency. Again the effect appears more marked in this case, due to the greater sensitivity of beta as an indicator of the device amplification mechanism.

When the curves of figure 10.6 are measured, the collector bias voltage is applied between collector and emitter as shown in figure 10.3. As we saw earlier, this allows the device to amplify its own reverse bias current even when no external forward bias is applied to the base-emitter junction.

As a result the lowest (Ib=0) curve of the family in this case represents a collector current level somewhat greater than the normal reverse bias current of the collector-base junction. Known as the **collector-emitter saturation current** (Iceo), it is actually very close in value to Icbo multiplied by beta.

Whereas the curves of the common base characteristic maintained their value of collector current down to zero collector voltage Vcb, the curves of the common emitter characteristic may be seen to fall away at a low value of collector voltage, the value rising slightly with collector current. The reason for this is that because the collector bias voltage is applied between collector and emitter, it cannot be reduced below a certain value without effectively forward-biasing the collector-base junction. Naturally when the latter occurs collector action falls, so that collector current drops as soon as the effective collector junction voltage reverses.

The voltage level at which the collector current of the device begins to fall is known as the **collector saturation voltage,** symbolised Vce(sat). The slight rise in this voltage with collector current results from the reduction in effective collector junction voltage due to resistive voltage drop in the relatively high resistivity collector and base material.

As before the collector bias voltage of the device cannot be increased indefinitely, for at a certain voltage avalanche breakdown occurs and the device current rises sharply. The avalanche breakdown of a bipolar transistor in the common emitter configuration occurs at the **collector-emitter breakdown voltage,** symbolised BVceo.

Because the device amplifies its own leakage current in this configuration, the amplification and avalanche effects are cumulative and this generally causes BVceo to be somewhat lower than the collector-base breakdown voltage BVcbo.

At this stage of our discussion of the bipolar transistor it is hoped that the reader has gained a reasonably clear and satisfying concept of basic device operation. If this is not so, a glance back through the chapter may be advisable, as the concepts which have been presented are quite important and will be necessary for an adequate understanding of the following chapters.

SUGGESTED FURTHER READING

BURFORD, W. B., and VERNER, H. G., **Semiconductor Junctions and Devices,** 1965. McGraw-Hill Book Company, New York.

CLEARY, J. F. (Ed.) **General Electric Transistor Manual**, 7th Edition, 1964. General Electric Company, Syracuse, New York.

PHILLIPS, A. B., **Transistor Engineering,** 1962. McGraw-Hill Book Company, New York.

SHIVE, J. N., **Physics of Solid State Electronics,** 1966. Charles E. Merrill Books, Inc., Columbus, Ohio.

SURINA, T., and HERRICK, C., **Semiconductor Electronics,** 1964. Holt Rinehart and Winston, Inc., New York.

Also "The Transistor: Two Decades of Progress," a special review section in **Electronics,** V.41, No. 4, February 19, 1968.

Chapter 11

PRACTICAL BIPOLAR TRANSISTORS

Characteristics and ratings — collector-emitter breakdown voltage ratings — sustaining voltage ratings — punch-through — second breakdown —maximum collector junction temperature — thermal resistances and maximum power dissipation — packages and heat sinks — current ratings — emitter junction resistance and input resistance — current gain and current level — transconductance — frequency response — gain-bandwidth product.

Some of the behaviour characteristics and ratings of the basic bipolar transistor were introduced in the latter portion of the preceding chapter. The present chapter will build upon this material by examining further aspects of behaviour which relate to practical bipolar devices.

It may be recalled that avalanche breakdown was the reason given for the limitation of collector-emitter voltage applied to a bipolar transistor connected in the common emitter configuration. While in general and with modern devices this explanation is largely true, it is in fact a very simplified one which should now be expanded and qualified if the reader is to gain a satisfying insight into actual device behaviour.

As explained earlier, the collector-emitter breakdown voltage tends to somewhat lower than the collector-base breakdown voltage BVcbo, because in the common emitter configuration the device is capable of amplifying its own collector-base reverse bias current or "leakage" current. The amplification action provides more carriers to take part in avalanche multiplication, so that the two effects are cumulative. Yet the degree to which the device does in fact amplify its leakage current will naturally depend upon the effective bias conditions at the base-emitter junction, as the amplification action of the device involves both junctions.

The voltage at which collector-emitter avalanche breakdown occurs thus depends not only upon the internal geometry and doping levels of the device itself, but also upon the external bias and circuitry connected between emitter and base. In other words there is really no fixed and distinct "collector-emitter breakdown voltage" for a particular device, but rather a whole range of breakdown voltages corresponding to different emitter-base bias conditions.

Generally the lower limit of this range corresponds to the situation where the base of the device is effectively "floating," or open circuit. In this situation the device is able to amplify virtually all of its leakage current, as almost all carriers which reach the base region from the collector are effective in attracting carriers from the emitter. Avalanche breakdown thus occurs at a relatively low voltage.

It is actually this "base open" value of collector-emitter breakdown voltage which is given the symbol **BVceo** introduced in the last chapter. As the lowest value of collector-emitter breakdown voltage displayed by a bipolar device, BVceo is often of considerable importance for circuit design.

The upper limit of the breakdown voltage range generally corresponds to the situation where the base is reverse-biased with respect to the emitter. Here virtually no leakage current amplification can take place. The reverse bias on the base-emitter junction discourages carrier injection, while carriers reaching the base region from the collector are virtually "sucked" out of the base by the bias on the base electrode. Avalanche breakdown thus tends to occur at a relatively high voltage, approaching the collector-base breakdown voltage BVcbo.

The symbol usually employed to represent the "reverse biased" collector-emitter breakdown voltage is **BVcev.** In some cases manufacturers test devices for reversed-bias collector-emitter breakdown with a specified resistor in series with the reverse base bias, and in these cases the breakdown voltage may be symbolised **BVcex.**

Almost as high as the breakdown voltage for reverse base-emitter bias is that which corresponds to the situation where the base is effectively short-circuited to the emitter. Here the device is capable of only very slight amplification of its leakage current, as the base-emitter junction is virtually clamped in its equilibrium state.

A further symbol is used to represent this "base shorted" breakdown voltage, as one might expect. The symbol is **BVces.**

If external resistance is introduced between base and emitter, a higher proportion of carriers reaching the base region from the collector are able to attract carries fro mthe emitter, and the device begins to amplify its leakage current. The breakdown voltage thus falls from the "base shorted" value BVces, and as the circuit resistance is increased it falls ultimately to the value BVceo.

To provide the circuit designer with a measure of the rate at which the collector-emitter breakdown voltage of a device falls with increasing base circuit resistance, some device manufacturers quote a value of breakdown voltage which corresponds to a particular value of external base-emitter resistance. This is given yet another symbol: **BVcer.**

Depending upon base-emitter bias and circuit conditions, then, the collector-emitter breakdown voltage of a bipolar transistor can vary significantly over a range having a lower limit of BVceo and an upper limit of BVcev. This is illustrated in figure 11.1, where the breakdown characteristics of a typical modern silicon transistor are shown for each of the situations described in the foregoing. The value of collector-base voltage corresponding to BVcbo is

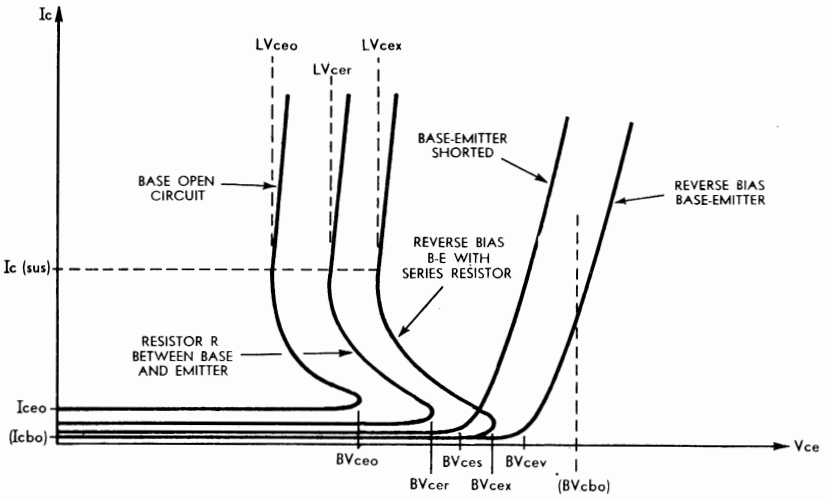


Figure 11.1 NOTE: LEAKAGE CURRENT LEVELS EXAGGERATED FOR CLARITY

shown as a dashed vertical line, for comparison.

From the shape of the BVceo, BVcer and BVcex curves, it may be seen that whenever the base circuit contains effective external resistance, the device breakdown characteristic enters a negative resistance region immediately following breakdown. Basically this occurs because although the device amplification action contributes to the onset of avalanche breakdown by means of the external base circuit resistance, it effectively ceases as soon as avalanching begins.

Because of this negative resistance behaviour, measurement of breakdown voltages BVceo, BVcer and BVcex can pose considerable problems. The negative resistance of the device tends to interact with device lead inductance and capacitances associated with the semiconductor chip and its package, generating oscillations which upset the measurement.

For this reason some device manufacturers tend to measure and quote not the actual breakdown voltages for these situations, but collector voltage values which correspond to the region where the breakdown characteristic has in each case passed through the negative resistance region and entered a second positive resistance region. These voltage values are known as **sustaining voltage ratings,** and as may be seen they are symbolised respectively as **LVceo, LVcer** and **LVcex.**

Note that when sustaining voltages are quoted for a device the corresponding collector current level must be specified. This is shown on figure 11.1 as Ic(sus). It may be seen that the sustaining voltage value in each type of situation is somewhat lower than the actual breakdown voltage, so that sustaining voltage ratings for a device may generally be regarded as quite conservative.

From the foregoing it may be appreciated that the collector-emitter avalanche breakdown behaviour of a bipolar transistor is somewhat more complex than that of the collector-base junction, its description involving the use of no less than eight different voltage measures of breakdown behaviour.

Unfortunately, perhaps, even this is not the full story, for in fact avalanche breakdown is only one of a number of mechanisms which can result in collector-emitter breakdown. Another mechanism is commonly called **punch-through** or "reach-through."

Punch-through occurs if the collector-emitter voltage applied to a device is increased to the point where the depletion layer of the reverse biased base-collector junction extends right through the narrow base region and reaches the emitter junction. Naturally when this occurs the current passed by the device rises rapidly, as the potential barrier of the emitter-base junction is destroyed, and the base effectively becomes nothing more than an accelerating field region linking the similar-material emitter and collector regions.

Like avalanche breakdown, punch-through is not inherently a destructive mechanism; it is merely a mechanism whereby the resistance of the device drops abruptly at a certain value of applied voltage. However, as with avalanche breakdown, it is a potentially high-dissipation mode of device operation, so that device damage can occur if the power dissipated by the device is not limited by the external circuit. The symbol usually employed to represent the punch-through voltage of a device is **Vpt.**

If avalanche breakdown occurs in a device at a voltage lower than that necessary for the collector junction depletion layer to extend fully through the base region, punch-through does not occur. The reason for this is that the collector junction depletion layer ceases extending when avalanche occurs. Hence in general terms a device breaks down due to either avalanche breakdown or punch-through, but not both.

Which of the two mechanisms is responsible for breakdown in any particular situation depends partly upon the internal geometry and doping levels of the device concerned, as these factors basically determine the voltage levels necessary to initiate each mechanism. For this reason some types of modern silicon device employing carefully controlled geometry and doping levels almost always enter avalance breakdown first, and punch-through is extremely rare.

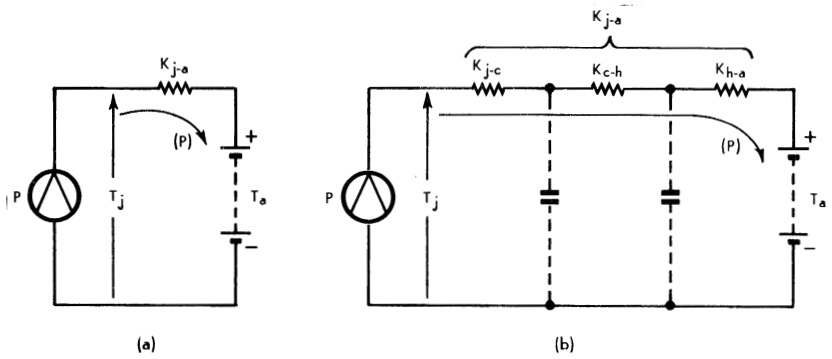
Naturally external circuit conditions can play a part in determining which of the two breakdown mechanisms occurs first, because as we have seen in the foregoing the avalanche voltage is quite dependent upon base-emitter bias. Hence with some dvices punch-through can occur if the base is reverse biased or effectively shorted to the emitter (BVcev $>$ Vpt, or BVces $>$ Vpt), but cannot occur if the base is effectively open circuited (Vpt $>$ BVceo).

There is a third type of bipolar transistor breakdown mechanism which is quite distinct from both the avalanche and punch-through mechanisms. This is the so-called **second breakdown** mechanism.

In contrast with the avalanche and punch-through mechanisms, which are basically voltage-dependent, the second breakdown mechanism is primarily a function of localised power dissipation and overheating in the collector-base junction depletion layer.

The cause of second breakdown lies partly in the fact that the lightly doped base region of all practical bipolar transistors possesses significant "transverse" resistance, or resistance in the effective cross-section of the base. This is true regardless of the particular doping levels and internal geometry employed. And one direct implication of the transverse base resistance is that any external bias applied to either device junction is never applied entirely uniformly; a potential gradient is always set up through the base region, causing the effective bias to be greater in some areas than in others.

Because of this effect, the current passing through practical bipolar devices is not distributed evenly throughout the cross-sections of the emitter and collector junctions, but tends to concentrate in a manner reflecting the non-uniform effective bias. Thus with most modern devices having an internal structure roughly circular in shape, current tends to concentrate around the periphery of the junctions under forward bias conditions, while conditions of reverse bias tend to cause a concentration at the centre of the junctions.



T_j = JUNCTION TEMPERATURE T_a = AMBIENT TEMPERATURE

$K_{j-a} \equiv \theta_{j-a}$ = THERMAL RESISTANCE BETWEEN JUNCTION AND AMBIENT

$K_{j-c} \equiv \theta_{j-c}$ = THERMAL RESISTANCE BETWEEN JUNCTION AND CASE

$K_{c-h} \equiv \theta_{c-h}$ = THERMAL RESISTANCE BETWEEN CASE AND HEATSINK

$K_{h-a} \equiv \theta_{h-a}$ = THERMAL RESISTANCE BETWEEN HEATSINK AND AMBIENT

P = POWER DISSIPATED AT JUNCTION

Figure 11.2

Being reverse biased in normal operation, the collector-base junction of a device accounts for most of the collector-emitter voltage drop. Hence it is the collector-base depletion layer which accounts for most of the power dissipated by the device, and in this region that most of the heat is generated. The non-uniform distribution of device current produced by transverse base resistance therefore results in uneven generation of heat in the depletion layer.

As well, minor doping variations tend to occur almost inevitably, and these tend to cause further localisation of current and power dissipation. The result is "hot spots," or small areas within the collector junction depletion layer which have significantly higher dissipation than the remaining areas of the layer.

It is these hot spots which are associated with the second breakdown mechanism. In effect, what happens in second breakdown is that the temperature at one or more of the hot spots reaches a level where melting of other permanent changes to the device structure can occur. Generally this results in a sharp rise in collector-emitter current, a fall in voltage drop and the ruin of the device.

As one might expect, the actual tem-

perature reached by the hot spots within a device depends not only upon the total power dissipation but also upon the doping non-uniformity, the transverse base resistance and the way this causes current concentration under various bias conditions. It also depends upon the effective duty cycle of the applied power, and the thermal behaviour of the device structure.

From this it may be seen that second breakdown is a rather complex mechanism, depending upon quite a number of factors. Some of these factors are under the control of the device manufacturer, and considerable research is being directed toward their more effective control. However other factors are determined by circuit conditions and biasing, and must be taken into consideration by the circuit designer.

Note in passing that in contrast with the avalanche and punch-through mechanisms, second breakdown is not merely a mode of potentially high power dissipation, but is rather a situation in which permanent device damage occurs. Because of this it is very difficult to test a device for second breakdown without ruining it in the process. "Second breakdown test sets" have been developed, but these are quite elaborate systems which are designed to detect slight changes in device behaviour occurring just before permanent damage ensues.

As with the other semiconductor devices which we have examined, bipolar transistors are rated by the manufacturer in terms of a maximum allowable internal operating temperature. Such a rating takes into account both the ambient temperature in which the device is situated, and the temperature rise within it due to power dissipation.

As the collector junction depletion layer generally accounts for a major proportion of the total device dissipation, bipolar devices are usually rated in terms of collector junction temperature, symbolised Tj(max). Typically Tj(max) for germanium devices lies in the range 80-90°C, and for silicon devices in the range 150-180°C.

Needless to say, the actual collector junction temperature of a device cannot easily be measured, as the junction itself lies buried within the device chip or die. However, the temperature may be deduced from a knowledge of the ambient conditions, the power being dissipated, and the thermal characteristics of the device and its immediate surroundings.

As we have seen in an earlier chapter, it is possible to describe the steady-state thermal behaviour of a semiconductor device and its package in terms of a **thermal resistance.** This is simply a measure of the temperature rise as a function of power dissipated, being expressed in units of (°C/watt). If transient thermal conditions are to be encountered it is necessary to supplement knowledge of the device thermal resistance with details of its **thermal capacitances,** either directly or in terms of the thermal time-constants.

Prior to the dissipation of power, the collector junction and all other parts of a device are normally at the so-called "ambient" temperature — i.e., the temperature of the surroundings, or more strictly that of those parts of the surroundings whose thermal capacity is so large that their temperature is for all practical purposes independent of any change in the thermal state of the device itself. When power is dissipated in the device, then, its internal temperature rises from this reference level rather than from absolute zero.

The extent to which the temperature rises above ambient is found simply by taking the product of the power being dissipated by the device and the **total effective thermal resistance** between the internal junction and the ambient surroundings. The latter parameter is often symbolised **Kj-a,** the letter "K" being a general symbol for thermal resistance (the Greek symbol **theta** is used alternatively, and perhaps more commonly; however, this symbol is not available for the present text).

Hence under steady-state conditions, the actual operating temperature of the collector junction of a bipolar transistor may be found by adding the temperature rise to the ambient temperature:

$$Tj = Ta + P.Kj\text{-}a \qquad \ldots (11.1)$$

Here Tj represents the junction temperature, Ta the ambient temperature, P the power dissipation, and Kj-a the total effective thermal resistance between the junction and the ambient surroundings. Note that Kj-a will include not only the thermal resistance of the device itself, but also that which is effectively present between the device package and ambient.

Often it is convenient to rearrange the expression of (11.1) into the following form, which permits easy calculation of the **maximum** power which a device may be allowed to dissipate for a given ambient temperature:

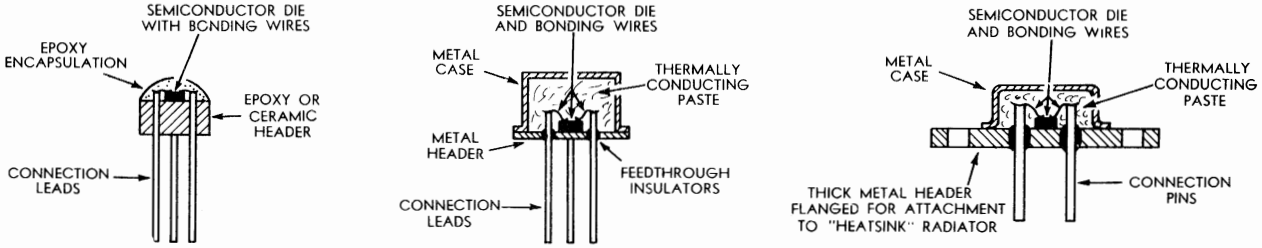
$$Pmax = \frac{Tj(max) - Ta}{Kj\text{-}a} \qquad \ldots (11.2)$$

Here Pmax is the required maximum dissipation figure, Tj(max) is the maximum junction temperature rating of the device, and Ta and Kj-a are the same as before.

The significance of expressions (11.1) and (11.2) may be seen quite clearly if the thermal situation involved is represented by a **thermal equivalent circuit.** This is a schematic diagram drawn using electrical symbols to represent thermal parameters, and based upon the fact that most thermal parameters behave in a very similar way to certain electrical parameters.

Thus the heat energy produced by power dissipation tends to "flow" through components in much the same way as an electrical current, interacting with the thermal resistances of the components to produce temperature drops in a very similar way to the voltage drops produced across electrical resistors.

The situation expressed in (11.1) may thus be represented by the simple thermal equivalent circuit of figure 11.2 (a). Here the constant current generator represents the constant power P dissipated in the device junction, while



MINIATURE EPOXY PACKAGE FOR LOW POWER DEVICES

SMALL METAL PACKAGE FOR MEDIUM POWER DEVICES

LARGE METAL PACKAGE WITH MOUNTING FLANGE FOR HIGH POWER DEVICES

Figure 11.3

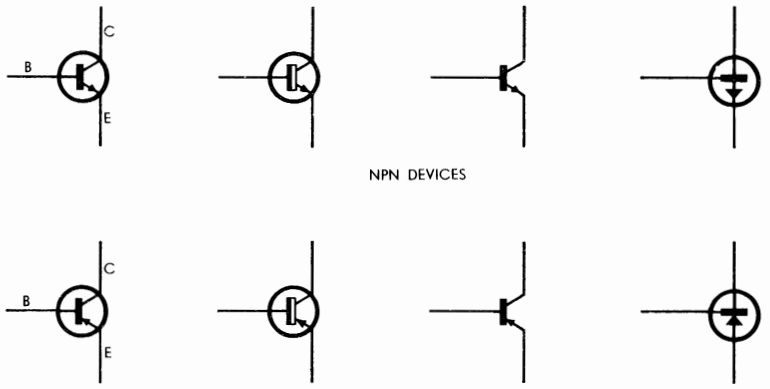


Figure 11.4

the battery represents the constant temperature Ta of the ambient surroundings. The resistor represents the thermal resistance Kj-a between junction and ambient. It may be seen that the junction temperature Tj, shown as equivalent to a voltage, will be equal to the sum of Ta and the temperature drop in the resistor, given by (P.Kj-a).

Typical packages used for bipolar transistors (and other devices) are shown in figure 11.3. As may be seen, small epoxy-resin packages are used for low power devices, while larger metal packages are used for higher power devices.

Low-power devices in small epoxy and metal cases are normally operated without any provision for heat removal additional to that provided by radiation and convection from the device itself, and because of this the thermal resistance figure generally specified by the manufacturer for these devices is, in fact, Kj-a, the complete "junction-to-ambient" thermal resistance. Naturally, this is an "average" figure representing a typical device in a typical thermal situation.

Because of the relatively low thermal coupling between a small package and the surroundings, the Kj-a for typical low-power devices tends to be rather high: in the range 250-600°C/watt. From expression (11.2) it may be appreciated that this tends to limit the power dissipation of even silicon devices to a few hundred milliwatts at normal ambient temperatures, and to proportionally lower power levels at elevated temperatures.

Higher power devices are not normally operated "free-standing," but rather with provision for additional heat removal via either a clip-on metal fin radiator, or a large "heatsink" radiator to which the device case is bolted. For these types of device the manufacturer therefore cannot in general predict the total effective junction-to-ambient thermal resistance Kj-a, because this will consist in part of the thermal resistance associated with the additional heat removal components.

This being the case it is usual for the manufacturer to specify for high power devices the **junction-to-case thermal resistance,** symbolised **Kj-c.** This parameter typically varies within the range 6—40°C/watt for medium power devices, and within the range 0.5 — 4°C/watt for high power devices.

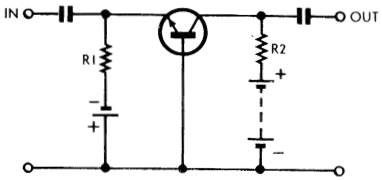
In order to calculate the operating junction temperature or the maximum power dissipation for a higher power device, using expressions (11.1) and (11.2), it is necessary to work out the total effective junction-to-ambient thermal resistance Kj-a. This is simply a matter of adding to the figure of Kj-c provided by the manufacturer the additional thermal resistances effectively present between the device case and ambient. Expressed symbolically:

$$Kj\text{-}a = Kj\text{-}c + Kc\text{-}h + Kh\text{-}a \quad \ldots (11.3)$$

where Kc-h represents any effective thermal resistance between the device case and the heatsink (mica insulating washers, etc.), and Kh-a represents the effective thermal resistance to ambient provided by the heatsink.

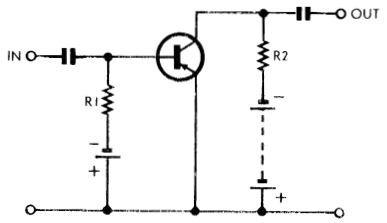
The significance of expression (11.3) may be seen in figure 11.2(b), where the simple thermal equivalent circuit of (a) is expanded to account for the distinct thermal resistances which together form Kj-a in the case of a high power device mounted on a heatsink. The individual thermal resistances are shown in series, to agree with the observed fact that their temperature drops are additive.

The approximate thermal resistance of different mounting configurations, insulating washers and heatsinks are given in many of the standard design manuals. This allows quite accurate predictions to be made of operating temperatures for higher power devices, and conversely it permits the designer to estimate quite accurately the type and size of heatsink required if a device is to be called upon to operate reliably at a given power dissipation.



COMMON BASE AMPLIFIER
(NPN DEVICE SHOWN)

Figure 11.5 (a)



COMMON EMITTER AMPLIFIER
(PNP DEVICE SHOWN)

(b)

It should perhaps be stressed that expressions (11.1) and (11.2) are only valid for steady-state conditions, as they take only thermal resistances into account. For accurate prediction of the temperature of a device under transient conditions, it is necessary to expand the foregoing discussion to take into account the effect of thermal capacitances.

The dashed capacitor symbols on figure 11.2 (b) indicate the basic effect of the thermal capacitances possessed by the device itself and the mounting arrangements. As may be seen, these introduce multiple thermal time-constants which will naturally tend to slow down any tendency for junction temperature Tj to change with changes in power dissipation P. Unfortunately a full discussion of the effects of thermal capacitance is beyond the scope of the present treatment, but this may give the reader some insight into the concepts involved.

Before leaving the topic of device temperature and power dissipation the reader may care to note that the foregoing discussion does not by any means apply solely to bipolar transistor devices. In fact it applies to virtually all electronic devices which dissipate power in operation, and hence to virtually all semiconductor devices. The concepts concerned have been developed in the present chapter simply because bipolar devices are those most often encountered at present in medium and high power applications.

From the preceding discussion of voltage breakdown, second breakdown, and temperature and power dissipation ratings for bipolar transisitors, the reader may perhaps have been led to infer that these devices might not be given specific ratings concerning current levels. On the surface this might seem a reasonable inference, based on the assumption that a device should not be damaged by any current level corresponding to operation within the second breakdown and Tj(max) ratings. However this is not the case.

As with most other semiconductor devices, bipolar transistors are in practice usually given both **average current ratings** and **surge current ratings.** In many cases, such ratings are given individually for each of the three device electrodes, to allow for situations in which the normal current relationships of the device are disturbed by transient conditions, breakdown or overdrive.

There is no single, specific breakdown mechanism associated with high device current levels as such. The current ratings which a manufacturer assigns to his devices are based upon consideration of one or more of a number of somewhat unrelated factors such as the fusing current of small internal bonding wires, and the fall-off in device amplification at high current levels due to dropping emitter injection efficiency and increased recombination in the base region.

Having looked in the foregoing at the main ratings which apply to bipolar transistors, let us now turn to consider further noteworthy aspects of normal device behaviour. To begin this section the reader may care to note the schematic symbols commonly used to represent bipolar transistors in circuit diagrams. These are shown in figure 11.4, where it may be seen that despite minor differences between symbols, the "arrowhead" on the emitter lead always points away from the rectangular bar base symbol for NPN devices, and toward it for PNP devices.

It may be recalled from the preceding chapter that the amplification action of the bipolar transistor essentially involves the modulation or control of collector junction current by the bias conditions at the base-emitter junction. Hence when considered as an amplifying device, it is the base and emitter electrodes which form the "input" terminals. As the base-emitter junction is normally forward biased, this means that the bipolar transistor is characterised by a relatively **low input impedance.**

The effective resistance of a forward biased P-N junction is a function of the current flowing, as the reader may care to determine by referring back to figure 5.1. The resistance is high at very low current levels, falling rapidly as the internal potential barrier is surmounted and the junction "turns on".

Surprisingly, perhaps, the actual resistance value of all forward biased P-N junctions as a function of absolute temperature and current flowing is remarkably consistent. It is virtually independent of doping levels, junction size and geometry. The theoretical reasons for this are a little beyond the scope of the present treatment; however the theory does predict that effective junction re-

sistance should be directly proportional to temperature yet inversely proportional to current, and this is in fact what is found.

The emitter junction of a bipolar transistor is no exception to this rule. Hence it is found that the effective resistance of the emitter junction of virtually any bipolar transistor at normal temperature (25°C) can be predicted quite closely by the simple expression:

$$Re = \frac{26}{Ie} \quad \ldots (11.4)$$

where Ie is the emitter current in milliamps.

When a bipolar transistor is used as an amplifier in the common-base configuration, as illustrated in figure 11.5(a), it is the value of junction resistance given by the foregoing expression which forms the effective input resistance of the device. This is because the current flowing in the input circuit is the full emitter current Ie. Hence in this configuration the device tends to have a very low input resistance; for example if the quiescent emitter current is a modest 1mA the input resistance will be only 26 ohms.

A somewhat higher, although still only moderate, input resistance is presented by the device when used in the common-emitter configuration of figure 11.5(b). Here there is an effective multiplication of the effective emitter junction resistance seen by the input circuit, because the current flowing in this circuit is the base current Ib, representing the relatively small current component Ie.(1 — α).

The effective input resistance of the junction itself in this configuration is thus equal to Re/(1 — α), and since alpha is very close to unity for most transistors, this is for practical purposes equal to (β.Re). For most practical devices one must add to this value the effective series resistance of the base region, so that the total input resistance of a bipolar transistor in the common-emitter configuration can usually be predicted quite closely by the expression:

$$Rbe = \beta.Re + Rbb \quad \ldots (11.5)$$

where Rbe is the common-emitter input resistance, Re is the junction resistance given by (11.4), and Rbb is the base region "spreading resistance."

From this it may be seen that the higher the gain of a device, the higher its input resistance in the common-emitter configuration. Also since Re is inversely proportional to emitter current Ie, according to expression (11.4), the input resistance tends to rise as Ie is reduced. However, the latter tendency is complicated by the fact that the amplification action of the bipolar transistor itself varies with emitter current.

As we saw in the preceding chapter, the amplification tends to fall at high current levels as a result of minority carrier concentration in the base region, and a consequent lowering in emitter injection efficiency. In fact, the amplification also tends to fall at very low current levels, particularly with silicon devices.

The full explanation of this is rather complex, and beyond the scope of the present treatment. However, in basic terms, what happens is that the small number of carriers injected into the base at very small emitter current levels cause a relatively weak concentration gradient, and thus tend to diffuse away through the base at quite a low velocity. This prolongs the time required to reach the collector junction depletion layer, and hence results in an increase in recombination with base region majority carriers. Hence the base transport efficiency is reduced, and with it the overall amplification.

This mechanism actually operates for both germanium and silicon devices. However, with silicon devices there is an additional mechanism which tends to reduce the gain at low current levels. The mechanism is associated with so-called "recombination centres" which tend to be present in the depletion layer region of the emitter junction, consisting of unwanted impurity atoms and various types of structural defect present in the crystal lattice.

The action of the recombination centres is to "grab" diffusion current carriers crossing the depletion layer from the emitter, and hold them captive so that they tend to be met by their opposite numbers travelling from the base. The net result is that the "useful" emitter-to-base injection component of emitter current is reduced, while the non-useful component in the opposite direction is increased; in other words, the emitter injection efficiency is lowered.

As the number of carriers involved in this mechanism is essentially fixed by the number of recombination centres present in a device, the effect upon emitter injection efficiency becomes significant only at low current levels where these carriers form an appreciable fraction of the total emitter current. At higher current levels the effect is swamped.

Because silicon devices offer many advantages in terms of low leakage currents and the ability to operate at elevated temperatures, device manufacturers have directed considerable effort toward reducing this effect. By stringent quality control of semiconductor materials and fabrication processes they have been able to reduce the number of recombination centres present in modern silicon devices to a very low level, resulting in β values as high as 300 at current levels as low as 10uA.

Naturally enough, devices capable of providing this order of current amplification at such low operating current levels are attractive from this viewpoint alone, as low operating currents generally mean higher efficiency and low circuit noise. However, reference to expressions (11.4) and (11.5) in the foregoing shows that such devices also offer the advantage of very high input resistance. At an emitter current level of 10uA, Re has a value of 2600 ohms, so that a device with a beta of 300 at this current level will display an input resistance of around 780K in the common-emitter configuration.

The variation of current gain β with emitter current level for a typical modern silicon bipolar transistor is illustrated in figure 11.6. It may be seen that the gain drops relatively slowly at low current levels, due to the influence of the recombination centres in the emitter junction depletion layer, and more rapidly at high current levels due to the effects of minority carrier concentration in the base.

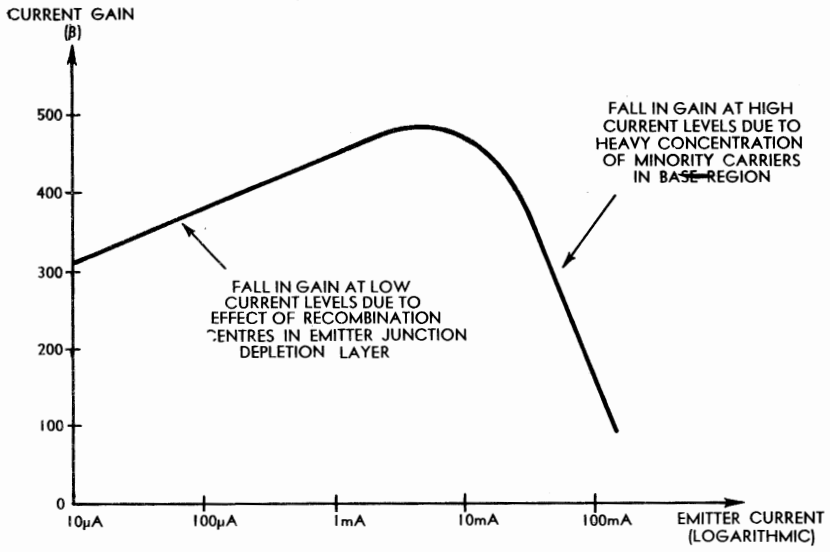


Figure 11.6

In passing, it may be worthwhile to point out that the **output resistance** of a bipolar transistor is basically the high resistance associated with the reverse biased collector-base junction. With modern low-leakage silicon devices this is typically around 1 megohm, whereas with the higher leakage germanium devices it is typically in the order of a few hundred kilohms.

Because of the relatively high input and output resistances displayed by modern silicon transistors, particularly in the common-emitter configuration, it is often convenient to visualise the amplification action of these devices not in terms of current gain, but rather in terms of an equivalent **transconductance** relating input base-emitter voltage with output collector current.

It is fairly easy to express the amplification of a bipolar device in terms of a transconductance, because the input voltage and current are related by the effective input resistance. In fact simple calculations based only on Ohm's law and expressions (11.4) and (11.5) show that for **both** the common-base and common-emitter configurations, the transconductance is almost exactly equal to the reciprocal of the emitter junction resistance Re.

The calculations for the common-

emitter configuration are as follows, shown for illustration:

$$Vin = Iin.Rin = \frac{Ib.Re}{(1 - \alpha)}$$

$$Iout = Ic = Ib.\beta = \frac{Ib.\alpha}{(1 - \alpha)}$$

$$\text{thus } gm = \frac{Iout}{Vin} = \frac{\alpha}{Re} \cong \frac{1}{Re} \quad \ldots (11.6)$$

The reader may care to verify that this result is also obtained for the common-base configuration.

What do these results actually mean? Simply that the transconductance of bipolar transistors, like the emitter junction resistance, is basically almost independent of device variations. The transconductance of virtually any bipolar transistor at normal temperature may thus be predicted simply by finding the reciprocal of Re, which from expression (11.4) is a simple function of the emitter current Ie. Hence at an emitter current of 1mA, the transconductance of any bipolar transistor is approximately 38.5mA/V, or in other words 38.5 milliMhos.

Actually, because the calculations leading to the expression of (11.6) are based on simplified theoretical assumptions, this expression tends to be over-optimistic in predicting gm. In practice it is found that the transconductance of most bipolar transistors is about 20 per cent lower than the predicted value, or equal to approximately (0.8/Re).

To conclude this discussion of the ratings and characteristics of practical bipolar transistors, let us now look briefly at the topic of **device frequency response.**

As with virtually all other "active" electronic devices, the behaviour of practical bipolar transistors is dependent upon frequency. In general, the performance of all devices tends to deteriorate as the operating frequency is raised. Various device types and individual devices differ only in terms of the rate of deterioration and the actual frequencies at which the performance is reduced to a nominal level.

The reasons for the fall-off in device performance at high frequencies are many. One important factor is that injected carriers take a finite time to diffuse across the base region — the so-called **base transit time.** At frequencies where this transit time becomes a significant proportion of the signal cycle, the carriers crossing the base region become "out of step" with the potential gradient across the region, resulting in a higher incidence of recombination. Base transport efficiency drops, and with it the device amplification.

The base transit time can naturally be lowered by reducing the physical thickness of the base region, and devices intended for use at very-high and ultra-high frequencies are generally provided with the thinnest base regions which can be reliably fabricated. It is also common to employ the NPN configuration for such devices, because the higher mobility of electrons results in a lower base transit time for a given base thickness than with the holes of a PNP device.

Other important factors influencing high-frequency performance are the space charge or **transition capacitances** associated with the depletion layers of the emitter and collector junctions. Together with the effective resistances of the junctions themselves, and also with the inevitable "bulk" or "spreading" resistances of the main emitter, base and collector regions, these depletion layer capacitances form R-C timeconstants which generally act as low-pass filters.

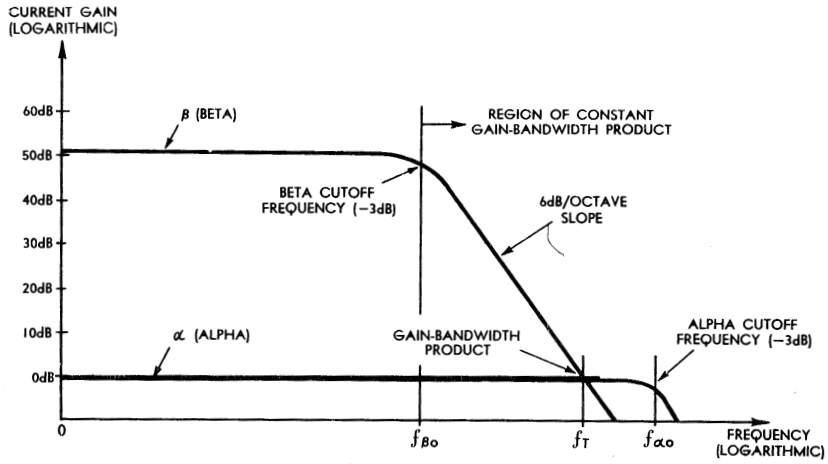
A further factor influencing bipolar device frequency response is the transit time taken by collected carriers to drift across the collector junction depletion layer. Although generally quite short compared with the base transit time, this further transit time can be significant with some very high frequency devices.

Possibly the perceptive reader will have realised from the foregoing that many of the factors which influence the frequency response of a bipolar device are themselves variables which depend upon the voltage and current levels at which the device is operated. Hence the frequency response of a device is not fixed, but is, in fact, dependent upon the operating conditions.

Thus, because base transit time depends upon the **effective** base region thickness, it is actually determined just as much by the width of the encroaching collector junction depletion layer, as by the physical base thickness. As a result, the base transit time of a device tends to increase significantly at low collector-emitter voltages, due to the relatively narrow depletion layer.

On the other hand, the widened collector junction depletion layer at high collector-emitter voltages itself tends to cause an increase in the depletion layer transit time. Base transit time and collector junction depletion layer transit time are thus complementary functions of collector-emitter voltage, each tending to cause a deterioration in frequency response at opposite voltage extremes.

Naturally, the collector-emitter voltage also determines the capacitance of the collector junction depletion layer, as this, too, depends upon the depletion layer width. Low values of collector voltage thus tend to reduce frequency response fairly rapidly because of the combined effects of increased base transit time, and increased collector junction capacitance. High values of collector voltage cause a somewhat less rapid deterioration due to rising collector junction transit time. The net effect is that the frequency response of a bipolar device tends to be highest at moderate collector voltage levels.

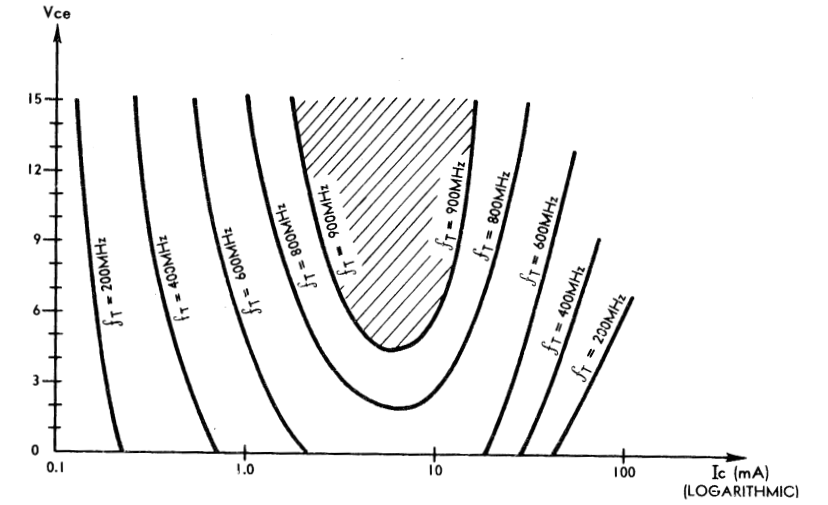


$f_{\alpha o} \equiv f_{\alpha co} \equiv f_{hfb}$ = ALPHA CUTOFF FREQUENCY

$f_{\beta o} \equiv f_{\beta co} \equiv f_{hfe}$ = BETA CUTOFF FREQUENCY

$f_T \equiv f_1$ = "GAIN-BANDWIDTH PRODUCT"

Figure 11.7



"CONTOURS OF CONSTANT GAIN-BANDWIDTH PRODUCT (f_T)"

Figure 11.8

The frequency response of a device tends to fall at low current levels, due to the rise in emitter junction resistance Re according to expression (11.4). This produces a long emitter junction time-constant, as the depletion layer capacitance of this junction is quite high under normal forward bias conditions.

There is also a slow drop in frequency response at high current levels, due primarily to the drop in effective collector junction bias caused by voltage drop in the semiconductor material of the collector region. The lower effective bias at the collector junction causes a contraction of the depletion layer as before, and a reduction in frequency response due to increased base transit time and collector junction capacitance.

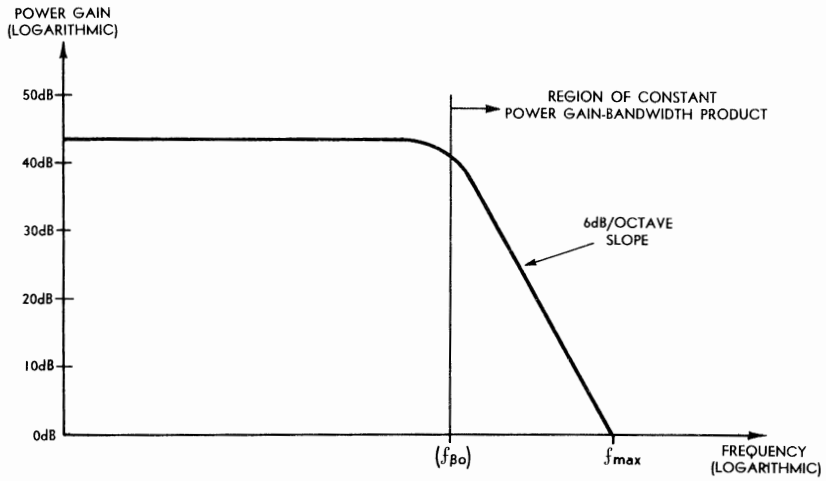


Figure 11.9

The frequency response of a bipolar transistor thus tends to be most favourable at operating points involving moderate voltage and current levels. At such operating points, the response tends to roll off smoothly in much the same manner as a simple R-C filter. The roll-off becomes apparent and/or significant in a number of ways, depending upon the circuit configuration in which the device is used, and the application.

In terms of the common-base amplification factor alpha, the performance of a device remains substantially constant up to a "corner" or "turnover" frequency, above which alpha falls logarithmically at the familiar 6dB/octave (20dB/decade) rate. This corner frequency, at which alpha has a value of 0.707 (−3dB) of its low-frequency value, is known as the **alpha cutoff frequency** of a device.

Like alpha, the common-emitter amplification factor beta also tends to remain constant up to a corner frequency, and then fall at 6dB/octave. However, because beta is a more sensitive indicator of device behaviour, and also because it is more sensitive to phase-shift effects, the **beta cutoff frequency** is generally very much lower than that for alpha. In fact, for typical devices it varies between $1/2\beta$ and $1/\beta$ of the alpha cutoff frequency. From this it may be seen that the higher the gain of a device, the lower tends to be its beta cutoff frequency as a fraction of the alpha cutoff frequency.

There are two broad implications from the foregoing. One is that because the beta cutoff frequency tends to be inversely proportional to beta itself, it is generally necessary to use low or medium-gain devices in a common-emitter amplifier stage in order to realise the maximum bandwidth. The other implication is that if the maximum bandwidth of a particular device is to be realised, it is generally necessary to use the common-base configuration in preference to common-emitter.

Because the beta of a device falls logarithmically above the beta cutoff frequency, the rate of gain fall-off in this region is such that the product of beta and frequency is always constant. Accordingly this region of device operation is often described as that wherein a device displays a constant "gain-bandwidth product."

The actual value of the gain-bandwidth product of a device in this region varies from device to device, and is in fact a very useful parameter of overall high frequency performance. At the same time it is conveniently measured because naturally enough its value is numerically equal to the frequency at which beta has fallen to unity. Because of this the latter frequency is often simply called the **gain-bandwidth product.**

The gain-bandwidth product of a device is generally in the same order as the alpha cutoff frequency, although usually below it. The relationship between the two parameters is not a simple one, however, and varies between devices and device types. The general relationships between alpha cutoff frequency, beta cutoff frequency and gain-bandwidth product are shown graphically in figure 11.7, together with the various symbols used for these parameters.

As noted earlier, the frequency response of a bipolar device depends not only upon the device itself, but upon its operating voltage and current levels. This dependence is conveniently expressed in terms of the gain-bandwidth product, as illustrated in figure 11.8.

Here are drawn the so-called **contours of constant gain-bandwidth product** for a typical modern silicon NPN transistor, expressing the way in which the gain-bandwidth product of the device varies with operating voltage and current. It may be seen that the maximum value of gain-bandwidth product for the device concerned is 900MHz, which may only be realised at operating points within the shaded region. Outside this region the gain-bandwidth product drops, as indicated by the frequency on the remaining contours.

Even at frequencies above the gain-bandwidth product and the alpha cutoff frequency, a bipolar transistor may be capable of useful power gain by virtue of the impedance step-up between input and output. In other words, a device can still have a useful power gain even at frequencies where its common-emitter current gain has dropped below unity.

In fact the power gain has a similar frequency characteristic to that for current gain, as may be seen from figure 11.9. Above the beta cutoff frequency, it falls logarithmically with frequency to give a constant power gain-bandwidth product. As before the power gain-bandwidth product is conveniently defined in terms of the frequency at which the power gain has fallen to unity (0dB), in this case known as the **power gain cutoff frequency.**

The power gain cutoff frequency is again a very useful parameter of high frequency performance, because it represents the highest frequency at which the device may be used to obtain power gain. It also represents the absolute maximum frequency at which the device concerned may be used in an oscillator, and for this reason it is alternatively known as the **maximum frequency of oscillation.**

There are other parameters used to indicate the high frequency performance of a bipolar transistor, including parameters which relate to the behaviour of the device as an "on-off" switch, as distinct from its use as a (nominally) linear amplifier. However, the parameters described in the foregoing are those most often encountered, and should give the reader at least a basic insight into device behaviour.

SUGGESTED FURTHER READING

AMOS, S. W., **Principles of Transistor Circuits**, 4th Edition, 1969. Iliffe Books Ltd., London.

CHERRY, E. M., and HOOPER, D. E., **Amplifying Devices and Low-Pass Amplifier Design,** 1968. John Wiley and Sons, New York.

CLEARY, J. F. (Ed.), **General Electric Transistor Manual**, 7th Edition, 1964. General Electric Company, Syracuse, New York.

GUNTHER, R. L., "Commonsense Transistor Parameters," in **Amateur Radio,** V.38, No. 1, January, 1970.

PHILLIPS, A. B., **Transistor Engineering,** 1962. Mc-Graw-Hill Book Company, New York.

STERN, L., **Fundamentals of Integrated Circuits,** 1968. Hayden Book Company, Inc., New York.

SURINA, T., and HERRICK, C., **Semiconductor Electronics,** 1964. Holt, Rinehart and Winston, Inc., New York.

Chapter 12

LINEAR BIPOLAR APPLICATIONS

Linear operation and the bipolar transistor — the load line and choice of quiescent operating point — biasing, parameter spread and temperature variations — conflicting bias supply requirements — the use of negative feedback — practical biasing circuits — bipolar amplifiers — the basic configurations — practical circuits — oscillators — other bipolar applications.

Having examined, in the two preceding chapters, both the basic theory of operation and the important practical characteristics and ratings of the bipolar transistor, the reader should now be in a position to consider the application of this device to typical circuitry. Accordingly, the present chapter and that which follows will discuss device applications. This chapter will examine the application of bipolar devices in so-called "linear" circuitry, while chapter 13 will deal with circuit applications in which they are used as switching elements.

Broadly speaking, "linear" circuits are circuits whose operation involves relatively smooth and continuous changes in voltage and current levels, and in which the active devices present are usually required to produce an "output" signal varying proportionally to the "input" signal over at least a significant part of the signal cycle. To satisfy this requirement it is generally necessary to arrange that the active devices are biased at a quiescent operating point which ensures that the device parameters remain as constant as possible over at least part of the range of circuit conditions involved.

Naturally enough, the exact position selected on the characteristic of each active device for the quiescent operating point will depend to a certain extent upon the requirements of each particular application. However in many cases the prime requirement is for the device parameters to remain constant for the largest possible output voltage and current signal excursions. This applies equally whether the active devices concerned are bipolar transistors, FETs or thermionic valves.

With bipolar devices this broad requirement is often satisfied by placing the quiescent operating point at a position similar to that marked "Q" in the diagram of figure 12.1 Here the family of curves shown are those of the common-emitter characteristics of a device, while the oblique line PQR is a **load line** representing the effect of the collector load resistance (or impedance) on the collector-emitter voltage.

The load line represents the locus of available operating points for the transistor, in terms of Vc and Ic. In operation, the device effectively slides up and down this line.

Point "P" represents a definite limit to linear operation in one direction along this line, representing the situation where the device current has dropped to almost zero and its applied voltage has effectively risen to the maximum available voltage. This situation is normally referred to as **cut-off.**

Similarly point "R" represents the limit of linear operation in the other direction along the load line, as this represents the situation where the device is **saturated** or "bottomed." It is drawing maximum current Ic(sat), while its applied voltage has fallen to the low saturation value.

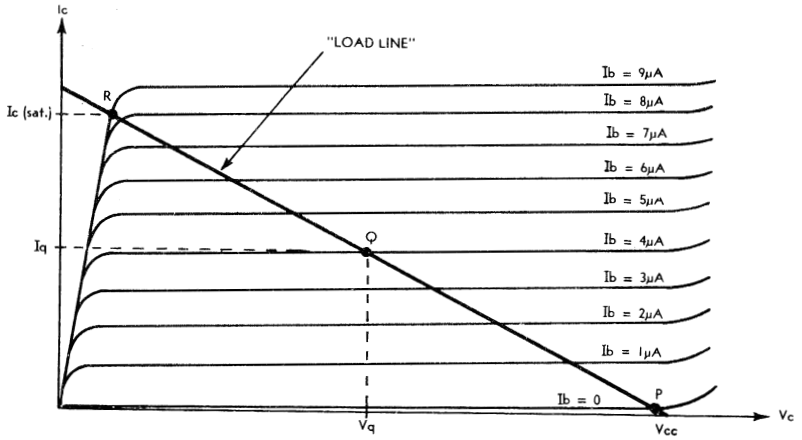


Figure 12.1

It is over the portion of the load line between cut-off and saturation that the parameters of most devices are relatively constant. Hence in a situation such as that in figure 12.1 it is the portion of the load line beween P and R which corresponds, at least nominally, to "linear" operation. With most devices the essential parameters vary only slightly over this portion of the load line, due to beta variation and other second-order effects.

Just where the quiescent operating point is placed on the linear portion of a load line depends upon the type of circuit involved, and upon the collector signal waveforms which must be handled. However, in a majority of amplifier and oscillator applications, the collector signal waveform involves fairly symmetrical excursions of both polarities about the operating point, and the usual design aim is to permit the device to operate linearly for the largest possible peak-to-peak collector voltage and current swings.

Placing the quiescent operating point Q at a position midway along the linear portion of the load line satisfies this requirement, as may be seen from figure 12.1. At this point the device is drawing a current Iq approximately equal to half Ic(sat), and its applied voltage Vq is approximately equal to half the available maximum voltage. With a resistive collector load the available maximum peak-to-peak voltage will approach the supply voltage Vcc, as shown, while with reactive load or a load reflected via a transformer it will be nearer twice this value.

Basically a bipolar transistor is placed at the desired quiescent operating point by the application of the appropriate forward bias to the base-emitter junction. Thus for the device whose characteristics are shown in figure 12.1, a bias which produced a base current Ib of approximately 4uA would be applied in order to set the operating point at Q.

While seemingly a simple matter, biasing of bipolar transistors is in practice complicated by a number of factors. One of these is that, like FETs, bipolar transistors are subject to appreciable parameter spreads. The common-emitter current gain beta is typically subject to a spread of about 3:1, for example, and this alone complicates biasing significantly.

As with FETs, the parameter spread causes each individual device of a particular transistor type to have its own unique family of Vc/Ic curves. Thus a device type cannot be represented simply by a single family of characteristic curves as shown in figure 12.1, but

could really only be represented by a whole "family of curve families."

Because of this, if one simply designs the biasing circuit of a transistor stage to supply the device with a fixed base current, the resulting operating point will depend very much on the gain of the particular device concerned. Only with a nominal or "bogie" device will it be near the optimum point, while with very high or very low gain devices it may be well away from this position.

Quite apart from parameter spread, there is a second major factor which complicates bipolar transistor bias design. This is that many of the key device parameters determining the operating point of a bipolar device are significantly **temperature dependent.**

Beta itself is temperature dependent to a moderate degree, usually tending to rise slowly with temperature. However, this is a second-order effect, and usually of far less practical significance

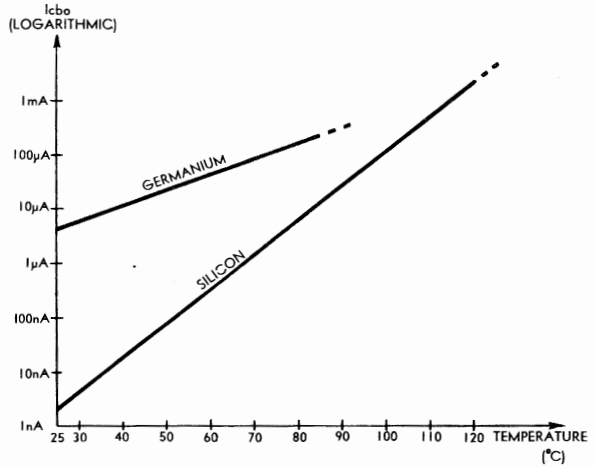


Figure 12.2

than the temperature dependence both of **Icbo,** the reverse bias saturation and leakage current of the collector-base junction, and of **Vbe,** the forward voltage drop of the base-emitter junction.

Being composed largely of minority carriers generated by the "intrinsic" mechanism, Icbo tends to rise rapidly and exponentially with temperature. For germanium transistors it approximately doubles in magnitude for every 8-10 deg.C rise, while for silicon devices it approximately doubles for every 5 deg.C rise.

As we have seen in an earlier chapter, Icbo tends to provide the base of the device with excess majority carriers, which in turn attract opposite polarity carriers from the emitter and so initiate device operation. In other words, Icbo tends to provide an effective "internal" forward bias component, acting additionally to any bias which may be applied to the device externally.

This means that because Icbo is strongly temperature dependent, there is a corresponding tendency for the effective bias on a bipolar transistor to rise with temperature, and the operating point to move accordingly. This is particularly true for germanium devices, where Icbo typically has a value at 25 deg.C of a few microamps. The effect is generally somewhat less evident with silicon devices, despite the higher temperature coefficient involved, because with these devices Icbo is typically some three orders of magnitude lower — only a few nanoamps at 25 deg.C. The relative magnitudes and temperature coefficients of Icbo for silicon and germanium devices are illustrated in the diagram of figure 12.2.

It is true that the extent to which Icbo does in fact supplement any external bias depends, as we have seen, upon the effective resistance connected externally between base and emitter. The lower this resistance, the greater the proportion of Icbo shunted around the base-emitter junction, and the smaller the influence of Icbo upon device operation. In order to reduce the effect of Icbo and its temparature dependence upon device biasing, one must therefore generally arrange the bias circuit connected between base and emitter to present the lowest practical source resistance.

The base-emitter forward voltage drop Vbe is also temperature dependent, being in this respect no different from any other forward biased P-N junction. However, in contrast with Icbo, the temperature coefficient is in this case negative, corresponding to the reduction in junction barrier potential as the Fermi levels in the P-type and N-type materials approach each other with increasing "intrinsic" carrier generation. With both germanium and silicon devices Vbe tends to decrease by approximately 2.5mV/deg.C, as shown by the typical curves of figure 12.3.

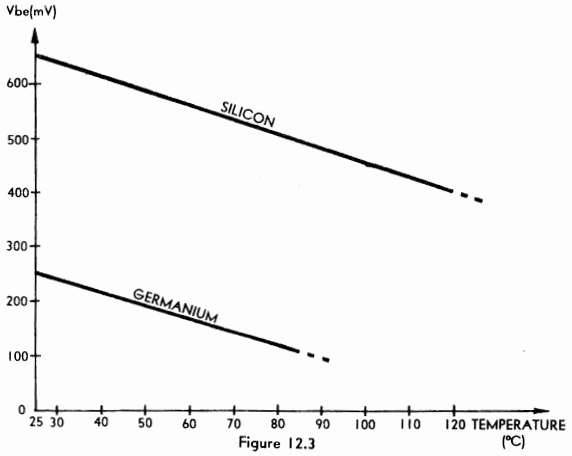


Figure 12.3

The negative temperature coefficient of Vbe tends to produce exactly the same type of change in operating conditions as the positive temperature coefficient of Icbo: a rise in base current Ib with temperature, and a corresponding change in quiescent current. And, unfortunately, the very same reduction in bias circuit source resistance which is desirable in order to reduce the effect of Icbo tends to accentuate the effect to Vbe. The lower the bias circuit source resistance, the closer the bias supply approaches the "constant voltage" situation, in which Vbe has maximum influence on Ib.

The effect of Vbe is actually inversely proportional to the bias circuit source resistance, and would fall to zero in the extreme case where the source resistance was increased to produce the "constant current" situation. Hence as far as bias circuit source resistance is concerned, there is a direct conflict between the requirements for reducing the effects of Icbo and Vbe.

Luckily, there are other means available for reduction of the effects of both Icbo and Vbe, so that this conflict does not lead to insoluble biasing problems. In general, practical biasing methods involve either supplementing the adjustment of the bias circuit resistance by the addition of negative feedback, or else adoption of the approach of deliberate temperature compensation, as will be shown in a moment.

In passing, it may be noted that because of the common tendency of Icbo and Vbe to cause device currents to increase with temperature, and the fact that there is a conflict between the biasing requirements for minimising the effect of these parameters, the bipolar transistor may be regarded as having an inherent tendency toward thermal instability or **thermal runaway.** The current tends to increase with temperature, and as an increase in current often tends to increase power dissipation and accordingly increase temperature, there is a definite positive feedback effect.

Unless the circuit is designed to stabilise device operation by reducing this positive feedback effect to a very low level, a bipolar transistor may either destroy itself, or at the very least cause its own operating point to slide up to the high-current "saturation" extreme of the load line.

The positive feedback and tendency toward thermal instability of the bipolar transistor are in sharp contrast with the behaviour of FET devices. Not only does the actual operation of the latter devices provide an inherent negative feedback mechanism which tends to stabilise the operating point, but also the temperature coefficients of the primary device parameters are such that they tend to cause FET devices to protect themselves by moving their operating point slowly towards cut-off as the temperature rises.

Some of the more commonly used bipolar transistor biasing circuits are shown in figure 12.4. These may be used to illustrate the basic concepts introduced in the foregoing.

The simplest method of bipolar tran-

sistor biasing is known as **current biasing** or "fixed biasing," and is shown in figure 12.4(a). As may be seen, it involves a single resistor Rb which is usually connected between the base electrode and the collector supply rail. The value of Rb is arranged to produce the required base current Ib, using Ohm's law: Ib = (Vcc − Vbe)/Rb.

If the supply voltage is greater than about 6V, the effect of Vbe in determining the bias current becomes insignificant, as Vbe is only about 0.65V for silicon transistors and about half this value for germanium devices. This is true in most applications, so that typically Ib is effectively determined only by Vcc and Rb, and is independent of the device itself; hence the description "fixed biasing."

The operating point stability provided by this type of biasing circuit is rather poor. The effects of Vbe and its negative temperature coefficient are reduced to a negligibly low level by the effectively fixed bias current Ib, to be sure, but on the other hand Icbo and its positive temperature coefficient generally assume maximum significance, due to the very high resistance of the bias source. The fixed bias current also tends to make the operating point significantly dependent upon beta, both in terms of spread variation and also in terms of temperature coefficient.

It is almost impossible to obtain adequate operating point stability using current bias with germanium transistors, due to the relatively high Icbo of these devices. Because of this, it is almost never used for such devices. The few exceptions are generally low power stages in very low cost equipment, intended for uncritical use within a restricted temperature range.

The very much lower Icbo levels of modern silicon transistors allow current biasing to be used to a somewhat greater extent, it is true, as with these devices the effect of Icbo is generally negligible at typical operating temperatures even with quite high bias circuit source resistance. However, the somewhat wide beta spread range of these devices still tends to restrict the use of current biasing to low cost applications, or to applications where either the bias resistors or the devices may be individually selected.

Some small improvement in operating point stability over that provided by current biasing may be obtained by feeding the base of the device from a resistive voltage divider, as illustrated in figure 12.4(b). Here the effective bias source resistance is equal to the parallel combination of Ra and Rb, and may thus be made very much lower than in the fixed bias case. The appropriate forward bias is applied to the device by manipulation of both the actual values of the resistors, and their ratio.

Because it provides a closer approach to a "constant voltage" bias source, **voltage divider biasing** generally allows the effects of Icbo to be made negligible. It also tends to stabilise the operating point against spread and temperature variations in beta.

It may be remembered from the preceding chapter (expression 11.5) that the input resistance of a device in the common-emitter configuration is directly proportional to beta. Because of this, changes in beta tend to cause a corresponding change in input resistance, which interacts with the essentially constant bias voltage provided by the bias divider to produce an opposite and compensating change in the input current Ib. Hence when beta is high, Ib tends to be low, and vice-versa.

Unfortunately while voltage-divider biasing does reduce the effects of Icbo and beta variation, it does not generally allow satisfactory stabilisation against Vbe variations. In fact, the lower is made the bias source resistance in order to stabilise against Icbo and beta variations, the more significant does Vbe become in comparison with the effective bias source voltage, and the greater the effect of Vbe variations. This illustrates the conflicting requirements for bias supply source resistance, noted earlier.

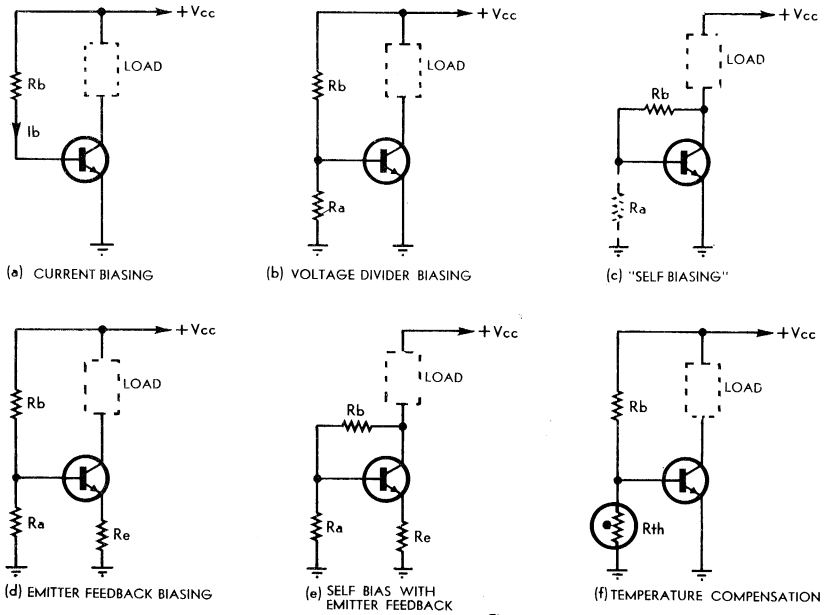


Figure 12.4

In some applications voltage divider biasing has the further disadvantage that, in order to achieve sufficiently low values of bias source resistance, the values of the divider resistors must be reduced to the point where the standing current drawn by the divider itself becomes comparable with, or can even exceed, the quiescent current of the transistor. In low-consumption battery equipment this can be very embarrassing where many stages are involved.

As mentioned earlier, negative feedback techniques may be used to overcome the conflict in bias source resistance requirements. One such method involves connection of the bias resistor Rb not to the collector supply rail Vcc, but direct to the collector of the device itself. This is illustrated in figure 12.4 (c), being known as **self biasing.**

Because of the finite resistance of the load in the collector circuit, the actual collector voltage of the device normally tends to vary inversely with collector current. By taking Rb, suitably modified in value, back to the collector, this voltage change can be used to automatically vary the bias in a direction which tends to counteract any change in collector current due to Icbo, Vbe or beta variations.

The basic self-biasing technique may be used with either single resistor current biasing or voltage divider biasing, as shown. The current biasing variant usually provides satisfactory stabilisation with silicon transistors, particularly in low power circuitry in which the collector load is a resistor. However the voltage divider variant is preferable, especially for germanium devices, because the lower bias source resistance tends to reduce the effects of Icbo and beta variations, leaving only Vbe variations to be compensated by the negative feedback.

It should be noted with regard to self-biasing that the resistor Rb connected between base and collector tends to produce negative feedback for "wanted" signal variations just as much as for unwanted changes in the quiescent operating point. As a result the effective gain of a device may be significantly lowered in cases where the input signal fed to the device comes from a relatively high impedance source. To prevent this effect, Rb is often split into two series components, and the junction of the two bypassed either to ground or to the emitter by means of a suitably high-value capacitor.

A second negative feedback biasing technique, quite distinct from self-biasing, involves an additional resistor Re connected in series with the emitter electrode. This is the **emitter feedback** technique, illustrated in figure 12.4(d).

Here the basic idea is that Re develops a voltage drop due to the emitter current Ie, and this voltage forms an effective component of base-emitter bias whose polarity is opposite to the forward bias applied to the base. The base voltage divider is arranged to provide a higher forward bias than in the case of 12.4(b), to compensate for this "bucking" component and produce the desired nominal emitter and collector currents. However in operation any tendency for Ie to change causes the voltage drop across Re to change accordingly, and this results in an automatic change in the effective base-emitter bias in the direction to counteract the tendency.

It may be seen that the negative feedback effect of Re is basically very similar to that associated with the "cathode bias" resistor of a thermionic valve, or the "self-bias" source resistor of JFETs and types A and B MOSFETs. However the action is not exactly the same, because whereas the thermionic valve and the field effect devices just mentioned are "normally on" devices, for which biasing may often consist solely of self-regulating negative feedback, the bipolar transistor is in contrast a "normally off" device. Like the type C JFET, it must therefore always be provided with some effective forward bias, to which may be added negative feedback components for the purpose of stabilisation.

As one might expect, the effectiveness of the negative feedback provided by the emitter resistor in stabilising the quiescent operating point is almost directly proportional to the ratio between the emitter resistor voltage drop and the resultant or effective base-emitter bias. If the feedback component is large compared with the resultant bias, the feedback will be very effective; but naturally if the feedback component is relatively small compared with the resultant bias, it will only be partially effective in counteracting current changes.

Generally the feedback component cannot itself be made very large, because the voltage drop across Re tends to reduce the available collector supply voltage and hence restrict the possible output voltage swing. To obtain effective feedback action, the forward bias applied to the base must therefore also be kept relatively low — or in other words, the base must be fed from a "voltage source" rather than a "current source." This implies either voltage divider biasing, as shown, or biasing from some other effective source of low voltage; current biasing cannot be used as this would tend to defeat the negative feedback action.

Note that the foregoing reasoning is actually identical with that given previously, in explaining why simple voltage divider biasing not only provides no control over Vbe variations, but in fact accentuates the effect of such variations. The only difference is that in the earlier case we were seeking to reduce the influence of Vbe, whereas in the present case we have been seeking to allow the negative feedback bias component to exercise the maximum stabilisation.

It is often found worthwhile to visualise the operation of emitter feedback biasing in terms of the effect of resistor Re upon the effective input resistance of the transistor as seen by the base bias source. Because of the amplification action of the device, Re will be seen by the base bias source as a resistor of value beta times its actual value, connected in series with the base-emitter junction. This very high effective resistance thus tends to produce pseudo-constant current biasing, by virtually "swamping" any tendency for Vbe to influence the base current Ib.

Because the action of the emitter feedback resistor may be visualised in this way it is often known as the "emitter swamping resistor."

Like bias resistor Rb in the self-biasing circuit, the emitter feedback resistor Re tends to introduce negative feedback for wanted signals just as much as for unwanted changes in the quiescent operating point. And as before, this can significantly lower the effective gain of the device. In this case the effect is not determined by the signal source impedance, however, but by the effective collector loading impedance.

The effective voltage gain in fact becomes stabilised by the negative feedback action, along with the quiescent operating point, becoming almost exactly equal to the ratio between the collector load and Re. Hence the larger Re is made relative to the load, the lower the effective voltage gain. In some applications this effect is deliberately used either to reduce the gain, or to stabilise the gain against parameter spread variations.

In other applications, of course, the gain reduction effect can be quite a nuisance, it being desirable to obtain full gain from the device. Happily this may be arranged simply by providing Re with effective signal bypassing, via a suitably high-value capacitor.

The emitter feedback biasing circuit shown in figure 12.4(d) is capable of providing very stable operation with both silicon and germanium transistors, as it may be designed to compensate almost completely for variations in all three parameters Icbo, Vbe and beta. For this reason it is the biasing circuit most commonly used for low and medium-power transistor circuitry.

There are cases, however, in which emitter feedback biasing alone cannot provide the desired order of operating point stability, due either to the need to stabilise over a very wide temperature range, or to the need to make compromises in setting the values of Re and/or the bias divider resistors. In such cases it is often found worthwhile to combine the self-bias and emitter feedback techniques, as shown in figure 12.4(e). By utilising two distinct sources of negative feedback, this combination circuit is generally capable of providing excellent stabilisation.

In many transistor circuits operating at high power levels the emitter feedback biasing method cannot be used, because an emitter resistor would reduce significantly the power fed to the load. This is often unfortunate, as such circuits usually operate at elevated temperatures where a high degree of stabilisation is desirable in order to guard against thermal runaway.

As self-biasing may not always be possible in such applications due to the type of load involved, while simple voltage divider biasing may not provide adequate stabilisation, some other means of maintaining the operating point must generally be found. Often this takes the form of a negative feedback system which monitors the temperature of the transistor, rather than its current.

Generally this approach involves the use of a device having a negative temperature coefficient, connected into the lower arm of the base bias divider, and placed in thermal contact with the transistor case. Thus, as the transistor temperature rises, the bias is automatically reduced. The temperature sensing element may be either a thermistor, as shown in figure 12.4(f), or a combination of one or more forward-biased P-N diodes. A thermistor is usually used with germanium devices, while diodes are usually used with silicon devices.

It may be noted that the biasing methods which have been discussed in the foregoing are all associated with a single transistor device, i.e., they are single-stage biasing circuits. As the reader might well have predicted, these are not the only possible biasing methods, for when devices are used in combination it becomes possible to arrange more complex biasing circuits involving direct coupling between a number of devices.

There are a great many variations possible with such multi-stage biasing methods, in some cases exploiting either the compensating temperature variations in complementary NPN and PNP devices, or the stabilisation action provided by negative feedback around many high gain devices connected in cascade. Unfortunately space limitations do not permit further discussion of such methods in the present treatment, and interested readers must be referred to the references given at the end of this chapter.

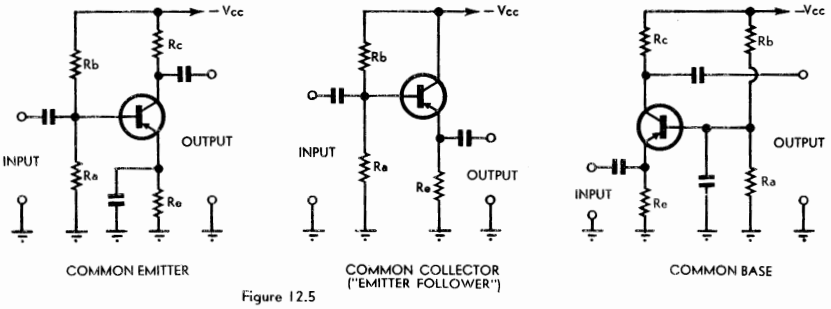


Figure 12.5

A final note which should perhaps be made before leaving the topic of bias stabilisation is that while the diagrams shown in figure 12.4 show NPN devices, this should by no means be taken to imply that any of the biasing methods described applies only to these devices. All methods apply equally to PNP devices, for which the supply polarity is simply reversed.

Having looked at the basic techniques used to bias bipolar transistors at a quiescent operating point appropriate for "linear" operation, let us now turn to examine briefly some of the very many applications of these devices in linear circuitry.

As with both field-effect devices and thermionic valves, probably the most common application of bipolar transistors is in amplifier circuits. The use of bipolar transistors in amplifier applications in fact far exceeds the use of FET devices at the time of writing, and has possibly now also exceeded that of

thermionic valves. This gives a good idea of the suitability of the bipolar transistor for many of these applications.

The variety of amplifier applications in which the devices are used is almost endless, including both small-signal and power amplifiers for audio, servo and other LF amplifiers, small-signal and power amplifiers for radio frequencies (RF), direct-current or DC amplifiers, operational amplifiers, wideband or "video" amplifiers, and instrumentation amplifiers. In almost every such application they may be used either alone or in conjunction with other devices such as FETs, and also either as a single type (NPN or PNP), or in mixed-type complementary circuitry.

Just as with FET devices and thermionic valves, bipolar amplifiers use only three basic device configurations. These are known respectively as the **common emitter, common collector and common base** configurations, and are illustrated in figure 12.5 as implemented for R-C coupled audio circuitry using PNP transistors.

The common emitter configuration may be seen to be the bipolar equivalent of the common cathode thermionic valve stage, and the FET common source configuration. The input signal is applied via a coupling capacitor to the base, while the output signal is taken via a similar coupling capacitor from the collector. Although the emitter feedback biasing method is shown, other methods may be used depending upon the specific application. Where an emitter resistor is used it is usually bypassed as shown, to prevent signal negative feedback.

This bipolar amplifier configuration provides a high order of voltage gain, useful power gain and a moderate input resistance. Typical stages may be arranged to give voltage gains in the order of 40-180 times, which compares very favourably with thermionic valve circuits. Current gain figures in the same order are also obtainable.

The input resistance of a common emitter amplifier stage consists of the input resistance of the device itself in parallel with the effective shunt resistance of the biasing network, as one might expect, and therefore tends to be somewhat lower than the input resistance of the device alone. The reader may recall from the preceding chapter that the input resistance of a bipolar transistor in the common emitter configuration depends upon its current gain and emitter current level, varying from a few ohms for a low gain power device operating at high current levels to many hundreds of kilohms for a high gain silicon device operating at very low current levels. Depending upon the device and the biasing circuit employed, therefore, a typical common emitter stage presents an input resistance of between a few ohms and a few hundred kilohms.

The output impedance of a common emitter stage is equal to the combination of the output resistance of the device itself in parallel with the collector load Rc. Generally the output resistance of the device is very much higher than Rc, however, so that in most cases the effective output impedance is almost exactly equal to Rc.

The common collector or "emitter follower" configuration is the bipolar equivalent of the cathode follower and source follower stages. Here the input signal is applied as before to the base by means of a suitable coupling capacitor, while the output signal is taken from the emitter. The collector is connected directly to the supply rail. The emitter resistor Re forms both the DC load resistor and the emitter feedback resistor, and as this dual function generally allows its value to be made somewhat higher than in the other configurations, the biasing stability of a common collector stage is usually excellent.

As with the corresponding thermionic valve and FET configurations, the common collector configuration provides no voltage gain but rather a slight voltage loss. However it provides a significant current gain, and also provides a very useful impedance transformation by virtue of a relatively high input resistance combined with a relatively low output impedance. Common collector stages are accordingly often used for isolation and impedance matching.

The input resistance of such a stage consists of the parallel combination of device and bias network resistances, as before, but in this case the input resistance of the device itself is much higher than in the common emitter configuration. It is in fact equal to beta times the sum of the base-emitter resistance of the device itself and the parallel combination of emitter resistor Re and the following AC load.

As a result of this increase in the effective input resistance of the device itself, the effective input resistance of a common collector stage is very often determined almost completely by the bias network. And because of the excellent thermal stabilisation provided by the large emitter resistor the bias network can often be arranged to present quite a high shunt resistance — as high as two or three megohms, if silicon transistors are used. Where germanium devices must be used or where even higher values of input resistance are required, it is possible to employ special techniques such as "bootstrapping" to produce effective multiplication of the bias network resistance, at signal frequencies.

The output impedance of a common collector stage is usually quite low, being equal to the output resistance of the device itself in parallel with the emitter resistor Re. The output resistance of the device is generally much lower than Re, being equal to the sum of the resistance of the base-emitter junction and a fraction $1/\beta$ of the effective resistance from base to ground provided by the bias network and signal source.

It may be seen from the foregoing that the input and output impedances of a common collector stage are not independent of each other, so that such a stage does in fact behave rather like an impedance "transformer." As such it provides less isolation between input and output circuits than either of the corresponding thermionic valve or FET configurations.

The common base configuration of a bipolar transistor corresponds broadly to the common gate FET stage, and to the "grounded grid" thermionic valve stage. It provides high voltage gain and a very slight current loss; however it also exhibits a very low input resistance, equal to the parallel combination of emitter resistor Re and the base-emitter junction resistance. These characteristics make the common base configuration of limited usefulness except at very high frequencies, where it becomes of interest because of the higher cut-off frequency associated with the common-base gain factor alpha.

Low level bipolar amplifier circuitry designed for audio and other LF applications generally uses emitter feedback

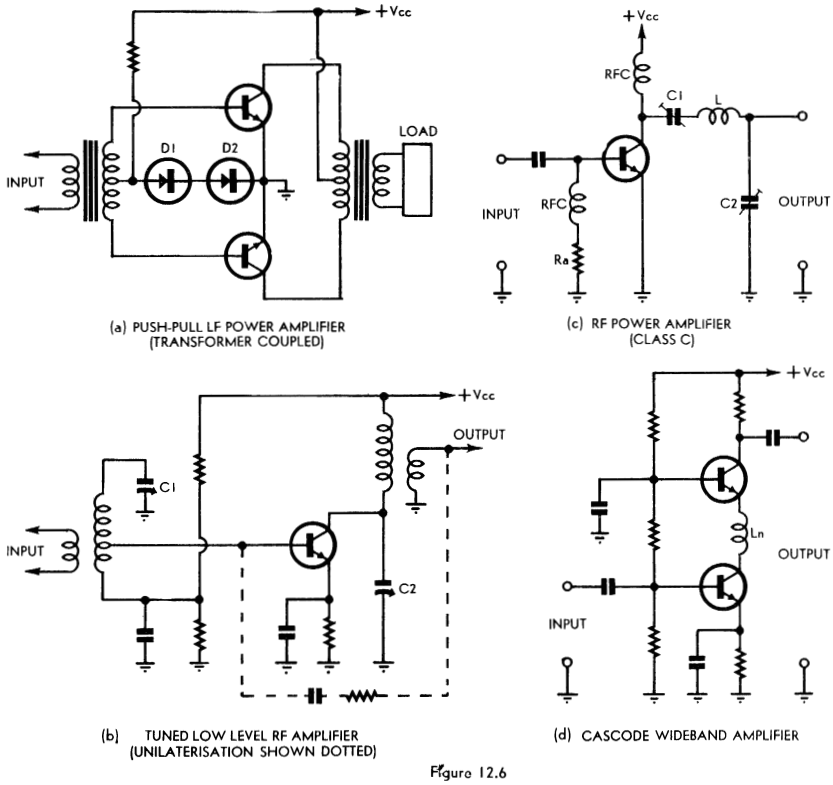


Figure 12.6

stabilised R-C coupled stages, of the type shown in figure 12.5. However amplifiers designed for different applications may use other types of coupling, and at times some of the other types of biasing circuit. Four representative examples of other types of amplifier stage are shown in figure 12.6, to briefly illustrate some of the many variations which may be encountered.

Figure 12.6(a) shows a transformer-coupled power amplifier stage of the type used in modest audio applications, and in high power servo amplifiers. As may be seen the stage is a push-pull type, in which two transistors are used in conjunction with centre-tapped windings on the input and output transformers.

For maximum efficiency such a stage is usually biased in either class B or class AB, the latter being used mainly in audio applications where it is desirable to reduce crossover distortion. For class B operation the devices are simply operated with zero base bias, the centre-tap of the input transformer secondary being taken directly to the grounded emitters. Being "normally off" devices the transistors then automatically operate only during alternate half-cycles.

For class AB operation a small forward bias is required, sufficient to allow each transistor to conduct for part of the other's primary half-cycles. While the resultant quiescent operating points of the devices are still quite near the cut-off end of the load line, however, it is usually very desirable to ensure that operation is well stabilised. This follows because such stages often involve considerable power dissipation and temperature rise.

An emitter feedback resistor generally cannot be used, both because of the drop in efficiency which this would introduce, and because it often proves extremely difficult to effectively bypass this resistor at the very low impedance levels involved. Hence the usual bias method chosen is that of temperature compensation using either a thermistor or diodes in the lower section of the bias divider. In the diagram diodes D1 and D2 perform this function, and would normally be arranged to be in thermal contact with the transistors.

Figure 12.6(b) shows a low level RF amplifier stage of the type found in many radio receivers, and in the early stages of transmitters. As may be seen it uses a single transistor connected in common emitter mode, with tuned transformer coupling at both input and output. Capacitor C1 tunes the secondary of the input transformer to the operating frequency, while C2 similarly tunes the output transformer primary.

Typically such a stage uses emitter feedback biasing, as shown, with the emitter resistor well by-passed at signal frequency, and the base bias divider connected to the by-passed "cold" end of the input transformer secondary. Note that whereas the high output resistance of the device allows the collector generally to be connected directly to the "hot" end of the output transformer secondary, the relatively modest input resistance necessitates the base being connected to a tap on the input transformer secondary, in order to preserve the input "Q." An alternative method is for the input transformer to have a tuned primary, with the base

connected to a low impedance secondary.

It may be recalled that the bipolar transistor possesses significant collector-base capacitance: the capacitance associated with the collector junction depletion layer. This provides a potential feedback path when the device is connected in common emitter mode, so that like the triode valve and the FET, it should ideally be neutralised.

In addition, the reverse-bias leakage and saturation current Icbo effectively constitutes a second "resistive" collector-base feedback component, so that for fully stable operation at high frequencies a bipolar transistor must strictly be **unilateralised.** This term means nothing more than the effective conversion of the transistor into an ideal "one-way" device, by neutralisation of both the capacitive and resistive feedback components.

Generally both neutralisation and the more complete unilateralisation can only be applied to fixed-frequency amplifier stages, as found in such applications as receiver IF stages and many low-power transmitter stages. Where the stage involved is tuned over a significant frequency range, it proves difficult to maintain constant compensation for the internal device feedback, and other techniques such as controlled mismatch must be used. One method of unilateralisation used for receiver IF stages consists of a series R-C combination, connected as shown in dashed form in the diagram.

An RF power amplifier stage of the type found in recent low power VHF transmitters is shown in basic form in figure 12.6(c). This type of stage generally employs special devices designed to provide useful power gain at many hundreds of Megahertz. The device usually operates in class C, conducting only on the tips of alternate half-cycles; the resulting collector current pulses are applied to a tuned circuit which then produces a smooth sinewave output by "flywheel" action.

In the type of stage illustrated the collector tuned circuit may not be immediately recognisable, consisting of inductor L and capacitors C1 and C2. It is basically a series resonant circuit, arranged in the form shown to act also as an impedance matching network and harmonic filter.

The reverse bias necessary to operate the device in class C may be applied either from a suitable bias supply, or by means of a "signal derived" bias system as shown. Here the conduction of the base-emitter junction of the device on signal peaks charges the input coupling capacitor, to an extent where the capacitor provides the required reverse bias. Resistor Ra is connected in series with the RFC base return to prevent the capacitor discharging significantly between charging peaks, ensuring effectively constant bias.

A further type of bipolar transistor amplifier application is shown in figure 12.6(d). This is a "cascode" amplifier stage, which like similar configurations of thermionic valves and FETs, often proves very useful in wideband amplifiers. The stage is effectively a combination of a common emitter stage and a common base stage, with the common emitter stage providing relatively high input resistance and significant current gain, while the common base stage provides high voltage gain with wide bandwidth.

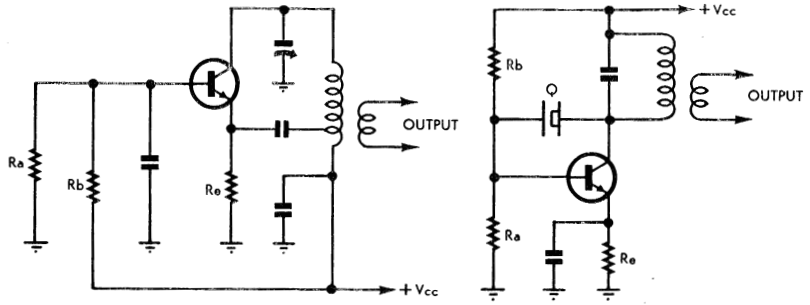


Figure 12.7

Although a cascode stage may consist of two devices with R-C coupling between, it is often possible to connect the two directly as shown and thus save components. The inductor Ln is a stabilising element often found necessary to prevent oscillation due to interaction between the collector and emitter impedances of the two devices at high frequencies.

Being capable of power amplification, bipolar transistors are naturally quite suitable for use in **oscillator circuits.** In fact they find use in oscillators generating signals spanning almost the full range electromagnetic frequency spectrum to which electronic circuitry is currently applied — from a small fraction of a Hertz to many Gigahertz (GHz).

At low and medium frequencies bipolar transistor oscillator circuits generally consist basically of R-C coupled amplifiers, either single stage or multi-stage, around which feedback loops are connected. The circuitry thus involves transistors connected in a manner usually very similar to that shown in the diagrams of figure 12.5. The feedback loops generally consist of R-C networks designed to provide a positive loop gain of unity at the desired operating frequency. In most cases additional circuitry is used to maintain a constant amplitude low-distortion sinewave output, by restricting the peak-to-peak oscillations to the linear portion of the transistor load lines.

High frequency oscillators normally employ either L-C tuned circuits, quartz crystals, tuned lines or similar resonant elements. Hence in this type of oscillator circuit, the transistor is basically used as a power amplifier which compensates for the resonant element losses. It is the resonant elements which oscillate, the transistor merely ensuring that the oscillations are maintained.

Two representative examples of high frequency bipolar transistor oscillators are shown in figure 12.7. In (a) is shown an L-C tuned or "self-excited" oscillator, in which the transistor operates as a common-base amplifier with feedback coupled to the emitter from a suitable tap on the tuned collector winding. Output is taken from the oscillator by means of a low impedance secondary winding. The biasing again employs the emitter feedback method.

Figure 12.7(b) shows an oscillator using a quartz crystal "Q" as the main frequency determining element. In this case there is also an L-C tuned circuit in the transistor collector circuit, for the circuit shown is an "overtone" type in which the crystal is forced to operate in a higher-order mode than the fundamental. The idea is that the collector tuned circuit is adjusted so that the transistor is only able to provide the loop gain necessary for maintaining oscillations at the desired crystal overtone. Hence it is at this overtone that oscillations occur, rather than at the fundamental or other overtone frequencies.

As may be seen the bias used is again of the emitter feedback type, while the output is again taken via a small winding coupled to the inductor of the collector tuned circuit.

There are many other applications of bipolar transistors in linear circuitry, in addition to amplifier and oscillator applications. Bipolar devices are used as detectors, mixers, harmonic generators and frequency multipliers, and also as controlled-value resistor elements in applications such as automatic gain control (AGC), modulators, series and shunt voltage regulators, and current regulators. Unfortunately space restrictions prevent more than a brief acknowledgment here of the existence of these applications, however, and interested readers must be referred to references such as those listed below.

SUGGESTED FURTHER READING

BRAZEE, J. G., **Semiconductor and Tube Electronics,** 1968. Holt, Rinehart and Winston, Inc., New York.

CHERRY, E.M., and HOOPER, D. E., **Amplifying Devices and Low-Pass Amplifier Design,** 1968. John Wiley and Sons, New York.

CLEARY, J. F., (Ed.) **General Electric Transistor Manual,** 7th Edition, 1964. General Electric Company, Syracuse, New York.

WALSTON, J. A.; and MILLER, J. R. (Eds.) **Transistor Circuit Design,** 1963. McGraw-Hill Book Company, Inc., New York.

WOLFENDALE, E., **Transistor Circuit Design and Analysis,** 1966. Heywood Books, London.

Chapter 13

THE BIPOLAR AS A SWITCH

Electronic switching, and the bipolar transistor — the OFF state, and the effect of transistor leakage — the ON state — saturated and unsaturated operation — device power dissipation — speed of response — delay, rise, storage and fall times — improving response speed — current mode switching — switching applications.

In addition to the multitude of linear circuit applications for which they prove suitable, bipolar transistors also have many applications in switching circuitry. In this chapter we will examine those aspects of device behaviour which are of basic importance in switching applications, and will then look briefly at some of the more commonly encountered applications of this type.

As the reader might well expect, it is normally desirable that any electronic device used to perform switching in a circuit should provide as close an approximation as possible to an "ideal" switching element. Hence in general such a device should exhibit as high a resistance as possible in its "switch open" or OFF state, and as low a resistance as possible in its alternative "switch closed" or ON state. Together with these basic requirements it should also possess the ability to be switched between these two states, in either direction, in as short a time as possible, as reliably as possible, and when so commanded by a control or "drive" signal for which the power requirements are relatively modest.

By suitable control of fabrication processes, the parameters of bipolar transistors can in general be arranged to meet these requirements rather well. When in the non-conducting or cutoff condition, a bipolar device typically exhibits a very high collector-emitter resistance, and thus provides a good approximation of an "open" switch. On the other hand, its resistance when in heavy conduction is usually quite low, giving an almost equally good approximation to a "closed" switch. And with a suitably designed device the transitions between these two states can be made reliably in a very short time, under the control of a relatively small input bias signal.

In basic terms, a bipolar transistor is used as a switch in exactly the same way as one uses a switch of the familiar mechanical variety: by simply connecting it across the source of supply, in series with the load whose current is to be switched on and off. In practice the load is connected in series with the collector, as shown in figure 13.1, with the transistor turned on and off by means of bias signals applied to the base via a series resistor Rb.

At this stage of our discussion of the bipolar transistor it should be almost unnecessary to point out that while an NPN transistor is shown in figure 13.1, the identical configuration is used with PNP devices. The only changes necessary if a PNP device is used are the usual reversal of supply and bias voltage polarities.

Essentially the operation of this basic circuit is quite straightforward. With zero bias or a reverse bias $-V_{bo}$ applied to the base via Rb, the device is cut off and draws negligible current; this is thus the OFF state of the circuit. Alternatively with a suitable forward bias Vbf applied to the base via Rb, the device conducts heavily and exhibits a low voltage drop; this is thus the ON state of the circuit.

As a bipolar transistor is a "normally off" device, it is at least nominally cut off with zero external bias applied to the emitter junction. However as we have seen in preceding chapters, a small collector-emitter current still flows when external forward bias is removed from the emitter junction, namely Iceo. This is an amplified version of the collector junction leakage current Icbo, and is accordingly dependent upon the semiconductor material involved, the temperature, the gain factor beta, and the resistance of the external circuit connected between base and emitter.

In theory, the mere existence of Iceo makes the bipolar transistor an imperfect switch, because it implies that the device never turns completely "off." However in practice the significance of Iceo depends very much upon its magnitude compared with the load current passing through the device in the ON state.

It is in cases where Iceo would be significant in comparison with the ON state current that it becomes particularly necessary to apply a reverse bias- Vbo to the base in the OFF state. The effect of the reverse bias is to prevent the device from amplifying the collector junction leakage current. Hence, when the reverse bias is used, the OFF state collector current passed by the device is not Iceo, but the considerably smaller Icbo.

Broadly speaking, from the point of view of minimum OFF state current, the very small value of Iceo passed by silicon devices makes it unnecessary to apply reverse bias when these devices are used, at least in applications which do not involve operation at high temperatures. In contrast it is usually necessary to apply reverse bias with germanium devices even at normal operating temperatures, because of the higher Iceo levels of these devices. However, with devices made from both silicon and germanium it is often desirable to apply reverse bias in the interests of operating speed, as will be explained later.

The OFF state of a bipolar transistor switch almost always corresponds to the situation where the device is at nominal cutoff. And as one might expect, the ON state always corresponds to a contrasting situation where the device is forward biased and conducting heavily. However, two different types of ON state operating point are possible: one is where the device has been driven completely into saturation, the other where the device is arranged to conduct heavily without quite entering the saturation region. Both of these types of operating point are used in practical switching circuits.

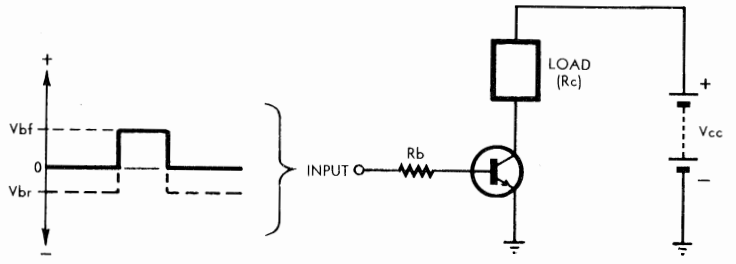


Figure 13.1

Circuits in which the devices are driven into saturation in the ON state are described as operating in the **saturated switching mode;** in contrast those which deliberately restrict the ON state operating point just short of saturation are described as operating in the **unsaturated switching mode.** Each of these modes of switching are illustrated graphically in figure 13.2.

In the saturated switching mode, as shown in (a), the forward bias Vbf applied to the base of the device via Rb is such that in the ON state the device operating point slides right up to

the intersection of the load line with the saturation locus of the device. The effective series resistance of the device thus falls to its minimum value, approximately equal to the "bulk" resistance of the collector-base and emitter regions.

When saturated, a bipolar transistor provides its closest approximation to a short circuit, and hence to an "ideal" switch in the closed position. It develops minimum voltage drop for the required load current, and hence wastes little power. Like cutoff, saturation is a low dissipation condition; in cutoff the device has relatively high voltage applied yet draws negligible current, whereas in saturation it passes considerable current yet develops negligible voltage drop.

Besides offering the advantage of low voltage drop and low power dissipation in the ON state, saturated mode switching also tends to be simpler and less costly than the alternative approach. Essentially only one additional component is required apart from the transistor itself and the load — the base resistor Rb.

The design of a saturated switch is generally quite straightforward: Rb and the forward bias Vbf are simply arranged to produce a base current Ib which **exceeds** that which would correspond to the required load current if the device were still in the active or "linear" region of operation. In other words, Ib is arranged to exceed the value of $Ic(sat)/\beta$, where Ic(sat) is the load current to be passed, and beta is the current gain of the device in the active region (strictly, the gain as measured just before saturation).

By Ohm's law Ic(sat) will be equal to (Vcc — Vce(sat))/Rc, where the term Vce(sat) is the saturation voltage drop of the transistor. Hence to ensure saturation, Ib must be arranged to satisfy the expression

$$Ib > \frac{Vcc - Vce(sat)}{\beta.Rc} \quad \ldots (13.1)$$

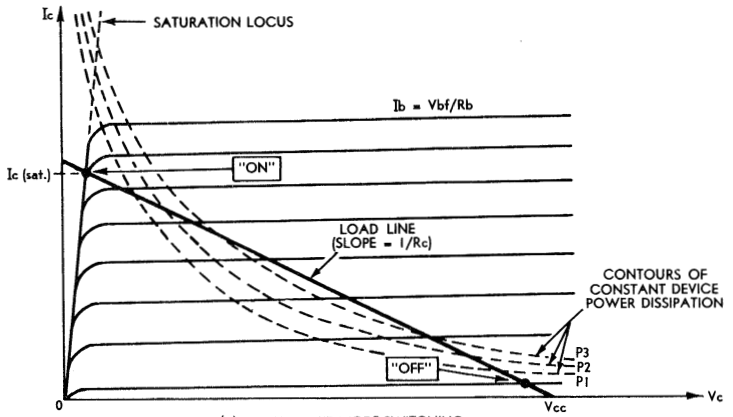
Usually Ib is arranged to be from 50% to 100% larger than the value of the right-hand side of this expression, when the latter is evaluated with beta equal to that of the lowest gain device likely to be used. This ensures that all devices should be reliably saturated.

Unfortunately, while saturated mode switching thus involves high static efficiency, low cost and relatively simple design, it also has the disadvantage of restricted operating speed. This is primarily due to the fact that, because of charge storage effects, a saturated bipolar transistor cannot cease conduction immediately upon removal of the base current drive. Further discussion of this phenomenon will follow shortly.

In contrast with saturated switching, unsaturated mode switching involves an ON state operating point which is near to, but not within, the saturation region. This is illustrated in figure 13.2 (b). While the collector current passed by the device is quite high and its voltage drop relatively low. operation is still in the "active" region where the device is capable of normal amplification action. Hence collector current Ic is still proportionally related to the base current Ib, according to the gain factor beta.

A device tends to dissipate higher power in the ON state of an unsaturated mode switching circuit than in a saturated mode circuit, because its voltage drop and effective resistance are both higher than if it were allowed to saturate. Hence in terms of static efficiency, an unsaturated mode switching circuit is less attractive than a saturated mode circuit.

The fact that the transistor is still "active" in the ON state of an unsaturated mode switch also produces an undesirable tendency for the load current to be dependent upon the gain of the device and the exact magnitude of its base drive signal, whereas ideally the load current should be determined solely by the load resistance and the supply voltage. Generally this means that unsaturated mode switching circuits cannot employ the simple circuit configuration of figure 13.1, but must employ more complex and more costly configurations which give adequate stabilisation against device and drive variations.



(a) SATURATED MODE SWITCHING

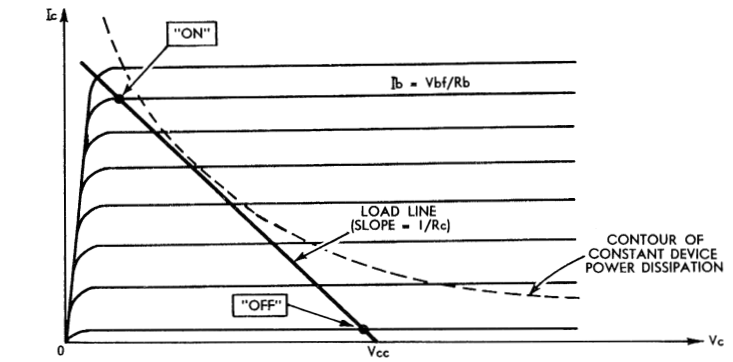


Figure 13.2 (b) UNSATURATED MODE SWITCHING

Despite these disadvantages, the unsaturated switching mode finds use because its offers the ability to operate at very high speeds. Because the transistor performing the switching is not driven into saturation, its operating speed for both turn-on and turn-off is basically only limited by the fundamental parameters which determine its "active" frequency response, and not by charge-storage effects. Unsaturated mode switching is thus used extensively in high-speed switching applications, particularly those where the ability to operate reliably at high speeds is very much less important than static efficiency or low cost.

In both the saturated switching and unsaturated switching modes, the power dissipation of the transistor tends to be relatively low in both the OFF and ON states. As a result it is generally true to say that switching circuitry tends to be relatively undemanding in terms of device dissipation rating. Quite small devices may be used even when appreciable power levels are involved in the load circuit.

While this is so, it is nevertheless true that in general the power dissipated by a bipolar transistor in a switching circuit tends to rise with the frequency with which switching operations are made. This follows because every time a device switches between the low-dissipation OFF and ON states, it necessarily spends a short but finite time in the intervening higher dissipation region.

It is in fact possible to draw contour lines on the collector characteristic of the device, representing constant device power dissipation, and examples of such contours are shown in figure 13.2. As may be seen the contours are of hyperbolic shape, corresponding to the fact that power dissipation is equal to the product of voltage and current. The distance between any contour and the Vc and Ic axes is directly proportional to the corresponding dissipation, so that in 13.2(a) contour P3 corresponds to a higher dissipation than P2, and P2 to a higher dissipation that P1.

The presence of the contours in figure 13.2(a) should allow the reader to verify the statement that a switching device necessarily passes through a region of relatively high dissipation in switching in either direction between the OFF and ON states. Note that whereas in both the OFF and ON states the device operating points are "below" the lowest dissipation contour P1, the load line crosses all three illustrated contours between these points, and for a significant part of its length is "above" the highest contour P3.

From this it may be appreciated that when a switching device is operated statically in either the OFF or ON

states, its average power dissipation remains quite low. However, the greater the frequency at which it is switched between these states, the greater the proportion of its total time is spent traversing the higher dissipation portion of the load line, and the higher its average dissipation tends to rise.

It is true that except in very high speed switching applications where a device may spend a relatively small proportion of its total time in the OFF and ON states, the average device power dissipation is usually somewhat less than the instantaneous dissipation at the centre of the switching load line. Because of this it is quite common for saturated switching circuits to be designed so that the load line actually crosses the contour corresponding to the maximum rated power dissipation for the device concerned. Hence in figure 13.2(a), contour P3 might in practice correspond to the Pc(max) rating for the transistor.

In unsaturated mode switching circuits this is generally not done, mainly because of the higher average dissipation produced by the non-saturated ON state operating point. In such circuits the contour of Pc(max) for the device might typically lie just above the central portion of the switching load line, as suggested by the dashed contour in figure 13.2(b).

Although device dissipation does tend to rise with switching frequency in both saturated mode and unsaturated mode switching circuitry, it is usually not the device dissipation rating which limits operating frequency. Rather, this is limited by the maximum speed at which the device can perform the required switching reliably in the circuit concerned: the **speed of response.**

Typically a bipolar transistor switching circuit responds to input drive changes in a manner illustrated in figure 13.3. Upon application of input drive, a short time elapses before the output current commences to rise. This is followed by a further period in which the output current rises to its full ON state value. Similarly, upon removal of the input drive a significant time elapses before the output current commences to fall, followed by a further period in which it falls to the OFF state value.

The short time required before the output current begins to rise after application of input drive is normally called the **delay time,** symbolised Td. For convenience of measurement this is defined as the time period between the application of drive and the point where the output current has risen to 10% of its ON state value.

The basic physical reason for the delay time is that before the device can commence conduction, charge must be supplied to the emitter junction depletion layer to reduce its width to that corresponding to the onset of "turn-on." In other words the initial flow of input drive current is effectively used to charge the emitter depletion capacitance, and does not result in any change in collector current.

The amount of charge required for this purpose depends upon both the area of the emitter junction of the device concerned, and also the conditions prevailing at the emitter junction in the OFF state. The larger the area of the emitter junction, the greater the charge required to alter the depletion layer width by a given amount, and the larger the delay time. Similarly if reverse bias $-V_{bo}$ is applied to the device in the OFF state, a greater change in depletion layer width is involved in preparing the junction for conduction than if zero bias is present in the OFF state, and the increased charge required accordingly tends to increase the delay time.

Hence from the point of view of minimising delay time, it is generally desirable to use a device with a small emitter junction area, and one which preferably does not require the application of reverse bias in the OFF state.

Following the delay time, the remainder of the turn-on time of the device consists of the time taken for the collector current to substantially complete its rise to the ON state value. This is the **rise time,** symbolised Tr. Conventionally the rise time is defined as the time taken for the collector current to increase from 10% to 90% of its ON state value.

The physical explanation for the rise time is that in order to increase the conduction of a bipolar transistor from the "just conducting" condition to that of full conduction, it is necessary to supply the device not only with the increased base drive current appropriate to the higher conduction state, but also with a further "lump sum" charge which is required to effect the appropriate change in internal conditions. Portion of the initial base current flow

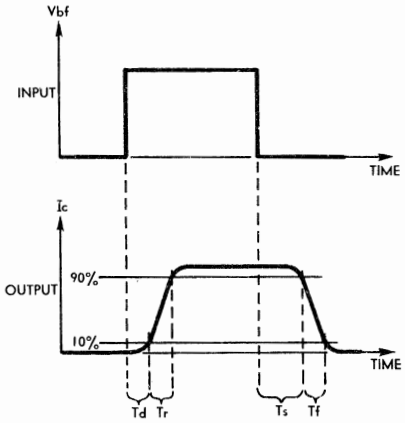


Figure 13.3

is used in supplying this "setting up" charge, so that until the device has obtained the charge and adapted to the new conditions, the full base current is not effective.

Basically there are three distinct components of charge which must be supplied to the device in this setting-up period. One component is the charge which must be supplied to the emitter junction depletion layer in order to narrow it to correspond to heavy conduction. In other words, the additional charge required by the emitter depletion capacitance.

A second component of charge is that required in order to set up the concentration gradient of injected carriers in the base region, necessary to produce a minority carrier base diffusion current equal to the full ON state collector current.

And the third component of charge is that which must be supplied to the collector junction depletion layer to reduce its width to correspond to the lower value of collector voltage present in the ON state. In other words, the charge required by the collector depletion capacitance.

All three components of the setting-up charge are determined partly by the internal geometry of the device, and partly by such circuit conditions as the supply voltage and the ON state collector current. Broadly speaking, the time required to supply each of the three components can be reduced for a given device by "overdriving," or considerably increasing the input drive current above that necessary to establish the ON state collector current. This may be done either on a "static" basis by increasing Vbf or reducing Rb, or on a "transient" basis, by arranging that an additional drive current is fed to the device only during turn-on.

The use of "static" overdriving naturally implies saturated mode switching, for it is only in this mode that overdriving does not essentially alter the ON state collector current. However, even in saturated switching circuits the use of static overdrive is not generally desirable, it causes a significant increase in the charge-storage effects to be discussed in a moment. Thus the most desirable way to reduce the rise time of a particular device is to use "transient" overdriving. One simple technique which achieves this end will be described briefly later in this chapter.

The significant time which elapses at turn-off before the collector current of the device commences to fall after the removal of input drive is called the **storage time,** symbolised Ts. This is defined by convention as the time period between the removal of drive and the point where the output current has fallen to 90% of its ON state value.

Basically, storage time is almost wholly associated with the previously mentioned charge storage effects produced when a bipolar device is driven into saturation. When a device is saturated, excess carriers are accumulated within the semiconductor lattice — carriers over and above those immediately involved in the conduction mechanisms. These excess carriers effectively constitute an inbuilt carrier "reserve" which allows the device to provide its own forward bias if external drive is removed. Hence upon removal of external forward bias the device continues conducting heavily until the stored carriers are exhausted.

With most bipolar devices, almost all of the excess carriers stored in saturation are located within the base region, being injected into this region from both the emitter and the collector. It should be fairly apparent that those injected from the emitter are basically excess carriers encouraged to take part in the normal emitter injection mechanism, as a result of the increased forward bias on the emitter junction. But the explanation for the additional injection of carriers from the collector may not be evident.

The clue to this behaviour is that, in the saturation situation, the collector-base junction of the device is effectively forward biased. In fact, as the reader may perhaps recall from chapter 10, the phenomenon of saturation occurs simply because the normal "collecting" action of the collector-base junction

breaks down if the collector voltage is allowed to fall to the point where this junction is no longer reverse biased.

In a situation such as that applying for our saturated switching transistor, where the device is passing a heavy current in saturation, carriers are obviously still crossing the collector junction in large numbers despite the breakdown in its minority carrier "collection" action. In fact the collector current in this type of situation consists of carriers moving across the junction in both directions as diffusion currents, encouraged by the forward bias conditions.

It is the component of saturated collector current comprising carriers moving from collector to base which provides the second source of carriers contributing to the excess accumulation in the base region. Hence it is basically these carriers, together with those excess carriers injected from the emitter, which provide the internal carrier "reserve" responsible for the continuation of device conduction during the storage time.

Like rise time, storage time is determined partly by the internal geometry of the device, and partly by the external circuit constants. An important factor within the device itself is the gain factor beta, to which storage time tends to be directly proportional. This follows because the higher the gain, the lower the effective base current required to sustain a given collector current, and hence the longer the period during which collector current can be maintained after removal of external bias by the accumulated carrier "reserve."

The external circuit factors influencing storage time are mainly the ON state forward bias current, which directly controls the amount of stored carrier charge accumulated within the device, and the OFF state bias circuit constants, which can assist turn-off by removal of stored carriers following the removal of forward bias.

For minimum storage time the ON state forward bias should be kept sufficiently low to ensure negligible accumulation of excess carriers within the device. In other words it should be prevented from saturating, as noted earlier. This explains the attraction of unsaturated mode operation at very high operating speeds. However, where saturated switching must be used, the base overdrive should fairly obviously be kept to the minimum level compatible with the requirements of expression (13.1), to prevent excessive storage time.

For a given device and ON state forward bias, the storage time is influenced by the effective constants of the bias circuit during the OFF state. A low impedance between base and emitter can reduce storage time, by providing a discharge path for the accumulated base charge. This effect is enhanced if a reverse bias —Vbo is used, as the accumulated carriers are then effectively "pulled" out of the base immediately forward bias is removed.

Following the storage time, the remainder of the time involved in transistor turn-off is that taken by the device in actually turning off. This is the **fall time,** symbolised Tf, and conf or the output current to fall from 90 % to 10 % of its ON state value.

The mechanisms responsible for the fall time are basically the converse of those responsible for the rise time. In this case, charge must be removed from the emitter and collector depletion layers, and also the minority carrier concentration in the base responsible for base diffusion must be dissipated.

Again, these mechanisms are influenced both by internal device geometry and by external circuit conditions. For minimum fall time the device used should possess small junctions having low values of depletion layer capacitance, and should ideally be forcibly turned off by means of a reverse bias —Vbo.

From the foregoing it may be seen that external circuit constants can play a significant part in determining a switching transistor's speed of response. Further illustration of this is provided by the diagrams of figure 13.4, which show some of the more common configurations used in practical switching circuits.

The circuit of 13.4(a) illustrates a technique often used to increase the operating speed of a simple saturated mode switch. The technique simply involves the connection of a capacitor Cb across the series base resistor Rb. The capacitor is often called a "charge-neutralising" or **commutating capacitor.**

The function of the capacitor is to lower the transient impedance of the bias source seen by the transistor. Thus at the onset of switch-on the capacitor effectively provides the device with a short pulse of overdrive which allows the emitter and collector junction depletion layer capacitance to charge up rapidly, and also allows the rapid setting-up of the minority carrier concentration gradient in the base. Hence both delay time and rise time tend to be significantly reduced.

Similarly at the onset of switch-off the charge acquired by the capacitor during the ON state tends to apply a transient reverse bias to the device, providing a means whereby the charge stored in the base region is rapidly drawn out. Thus storage time and fall time also tend to be improved, the former quite dramatically.

To achieve significantly higher switching speeds than those afforded by this technique, it is generally necessary to prevent the transistor from entering saturation. In other words, to adopt the unsaturated switching mode in preference to the saturated mode. Possibly the simplest way in which this may be achieved is illustrated in the circuit of figure 13.4 (b).

As may be seen the circuit consists basically of the elementary switch of figure 13.1, to which has been added two diodes. One diode is a silicon diode connected in series with the base electrode, while the other is a germanium type connected between the junction of Rb and the first diode, and the transistor collector. The configuration was first described by R. H. Baker in an MIT Lincoln Lab Report of 1956, and is often called the "Baker clamp."

The action of the diodes is to fix automatically the ON state operating point of the transistor just short of saturation. This action takes place as follows: As the transistor collector current rises, its collector voltage naturally falls due to the voltage drop across Rc. At the same time, the combined volt-

age drop of the silicon diode and the base-emitter junction of the transistor rises, as the base current increases.

The effect of these two voltage changes is to cause the germanium diode to become forward biased just before the collector voltage falls to the point corresponding to saturation. The diode thereupon conducts and effectively shunts all further increases in input current away from the base, and into the collector. This not only prevents the base current from reaching a value corresponding to saturation, but also provides additional current to the collector, to defer the onset of saturation.

By preventing the transistor from entering saturation, this circuit considerably improves the speed of response of the device itself. However, bias design tends to be somewhat more complex than with the simple saturated switch, as it is necessary to ensure that the diodes perform their function reliably for all possible parameter variations in both the transistor and the diodes. There is also the problem that the speed of response of the circuit now becomes highly dependent upon the response of the germanium diode, which must be a special high-speed type.

Probably the most satisfactory type of unsaturated mode switching circuit is the so-called **current mode configuration.** This is illustrated in basic form in figure 13.4(c).

It may be seen that the configuration differs from that of the simple switching circuit, in that the emitter of the transistor is now taken to a source of supply —Vee via a resistor Re. A diode D is also connected between emitter and ground.

The emitter voltage —Vee and resistor Re are deliberately chosen such that they provide an effectively constant source of current, whose magnitude is less than the value of emitter current corresponding to transistor saturation. This current flows into the transistor emitter when the base of the device is taken to a source of bias which is slightly positive with respect to earth.

In so flowing through the transistor, the current rigidly holds the ON state operating point of the device at a point outside the saturation region. Variations in the forward bias applied to the base, and in the beta of the device have virtually no effect on the operating point because of the controlling effect of the constant emitter current.

The purpose of the diode D is to act as an alternative path for the current from Vee—so that the transistor can be turned off! Changeover of the current from the transistor to the diode is simply arranged by taking the base of the device to a reverse bias sources which is slightly negative with respect to ground. This forces the transistor to attempt to reproduce a voltage at its emitter which is more negative than the potential at this point if the full current through Re were flowing through the diode; accordingly the transistor cuts off, and the current switches into the diode.

It may be appreciated from the foregoing brief explanation that current mode switching offers excellent DC operating point stability, is relatively easy to design, and is very insensitive to transistor parameter variations, while at the same time possessing the ability to operate at very high speeds which is characteristic of the unsaturated switching mode. For this reason current mode switching circuits are finding increasing use in high speed switching applications.

While the basic current mode configuration shown in figure 13.4(c) is quite practical, it is less commonly used in practice than the slightly modified configuration illustrated in figure 13.4(d). The operation of this circuit is virtually identical with that of the simpler circuit; the main difference is that the function of the diode D is now performed by a second transistor.

The base of the second transistor is supplied by a temperature-compensated

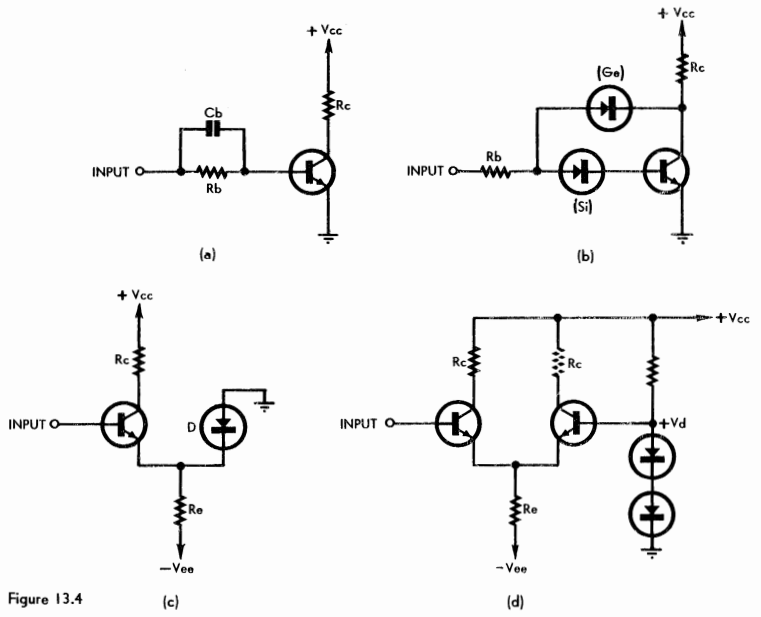


Figure 13.4

bias voltage Vd, developed across a series diode combination. This obviates the need to apply a reverse bias to the gate of the switching transistor in the OFF state. Simply taking the base of the first transistor to ground is now sufficient to cause the current from Re to switch to the second transistor, because of the latter's forward bias Vd.

To switch the first transistor to the ON state, it is again simply necessary to apply a small forward bias to the base. In this case the bias is only required to take the base slightly more positive than the voltage Vd present at the base of the second transistor. The current from Re then switches rapidly into the first transistor, again defining its operation very stably at a point outside saturation.

One of the advantages of this modified current mode configuration is that the second transistor may itself be used to perform switching, simply by inserting a second load Rc' in series with its collector as suggested by the dashed symbol in the diagram. Naturally enough, because this transistor is ON when the other is OFF, and vice-versa. it will act as a converse-acting switch. However this can be an advantage in logical switching circuits, where the logical converse or "complement" of a switching function is often required.

Having examined the basic aspects of bipolar transistor switching, let us now turn to look briefly at a small number of representative applications in which this mode of operation is involved.

One important application of this type is in power inverters and converters, used to produce high-voltage AC and DC respectively from low voltage DC sources. Here the bipolar transistor switches are used basically as automatic "choppers," to effectively convert the DC input into square-waveform AC capable of being fed to a step-up transformer. In this respect they perform a function very similar to the electromagnetic "vibrator" used previously in these applications.

The basic circuit for a typical DC-DC converter using two power NPN transistors is shown in figure 13.5 (a). It may be recognised as a push-pull blocking oscillator circuit in which the two transistors alternately drive each other into saturation and cutoff. Reliable starting is ensured by means of a small fixed bias applied to both bases via the divider formed by resistors Ra and Rb. The alternate switching between states is triggered by a breakdown in normal transformer action between the common collector and individual base windings due to the transformer core entering magnetic saturation, at the appropriate time after the previous switch-over.

The reversing magnetic flux in the transformer core produced by the transistors induces an appropriate AC voltage in the secondary winding, which in the case of an inverter feeds directly into the load. In a converter a rectifier and filter system are used to produce a high voltage DC output instead, as shown.

A bipolar transistor application not unrelated to the foregoing is that wherein the devices are used for modulation and demodulation in chopper amplifiers. As the reader may be aware, chopper amplifiers are basically AC coupled amplifiers which are fitted with an input modulator and output demodulator system which enables them to respond not only to the "AC" components of the input signal, but also to the "DC" component.

The basic configuration of a chopper amplifier using bipolar transistors for modulation and demodulation is shown in figure 13.5(b). Here both transistors are operated as saturated switches which are switched synchronously on

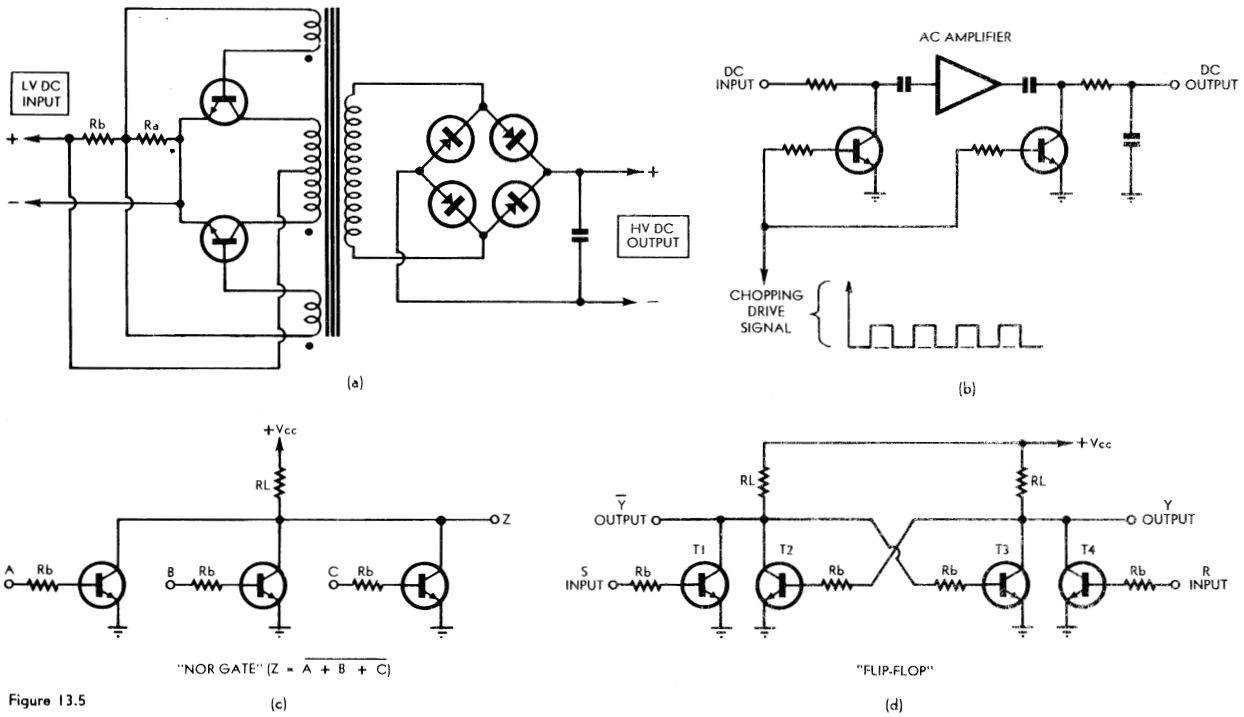


Figure 13.5

and off by a square-wave "chopping" drive signal. The first transistor effectively chops the input signal into an AC signal at the chopping frequency, of corresponding amplitude. This signal is then amplified by the amplifier in the normal way, so that a large AC square-wave signal whose amplitude is proportional to the original DC input signal appears at the amplifier output.

This signal is demodulated to produce a corresponding DC output signal, by the action of the second transistor switch. By effectively shorting the "load" end of the amplifier coupling capacitor to ground in synchronism with the action of the first transistor, this transistor forces the output coupling capacitor to acquire a charge which effectively restores the DC level of the square-wave signal. It is then only necessary to integrate the signal by means of a low-pass R-C filter, to remove the chopping frequency component and leave the original signal.

Although bipolar transistors find wide application in such switching applications as inverters, converters and chopper amplifiers, and also in pulse-width switching mode amplifiers and voltage regulators, perhaps the most rapidly growing of their switching applications is that of digital circuitry. Here transistor switches perform a wide variety of logical functions, ranging from simple logic gating to complex functions performed by elaborate configurations of transistor gates and transistor "flip-flop" storage elements.

Just two of the many transistor circuit configurations found frequently in digital applications are shown in the diagrams of figure 13.5(c) and (d).

In (c) is shown a simple logic gate consisting of three transistor switches sharing a common load resistor Rc. The idea is simply that the "output" voltage at point Z will only be at its "high" level if all three transistors are off — in other words, if none of the three input terminals A, B and C have forward bias applied. The application of forward bias to any one, two or all three of the inputs will cause the voltage at Z to fall to its "low" level, due to the conduction of one or more of the transistors.

This fixed relationship between the input and output conditions of the circuit allow it to be used to perform a variety of logical gating functions. For example if the inputs A, B and C are connected to three digital signal sources whose output is time-dependent, the appearance of a "high" output at point Z necessarily implies that at the instant concerned, none of the three sources are providing a "high" output. In this case the gate would be said to perform the logical "NOR" function.

The configuration shown in figure 13.5(d) is that of a simple "flip-flop" storage element. Here the idea is that because of the cross-coupling between the two central transistor switches T2 and T3, only one can be ON at the one time; the other must necessarily be OFF. Hence because the circuit is quite symmetrical, the circuit has two stable states — one with one transistor conducting, the other with the second transistor conducting.

The circuit may be forced to adopt either of these states at will simply by applying forward bias to either of the two additional transistors T1 and T4 connected in parallel with the cross-coupled pair. Hence forward bias applied to the "R" input causes the shunt transistor T4 on that side of the circuit to short the collector of its companion T3 to ground, removing the forward bias from the alternate device T2. This causes the latter device to cut off, so that forward bias is provided to maintain T3 in conduction after the removal of the external signal. The application of forward bias briefly to the "S" input produces the opposite effect, T2 being left in conduction with T3 cut off.

Because it possesses the ability to operate in one of two stable states, a flip-flop element such as that shown is eminently suitable to act as a storage medium for information in the form of binary numbers. Its state may be monitored at any time simply by examination of the voltage levels at the "output" terminals attached to the collectors on each side.

And with these brief comments we must bring the present chapter to a close. The survey of bipolar transistor switching operation and applications which has been given is necessarily very cursory and incomplete; to do full justice to this topic would require many weighty volumes. However, it is hoped that if nothing else the basic material presented will have given the reader an insight into the concepts involved, and may perhaps provide motivation for further reading in sources such as those listed below.

SUGGESTED FURTHER READING

MILLMAN, J., and TAUB, H., **Pulse Digital and Switching Waveforms,** 1965. McGraw-Hill Book Company, Inc., New York.

PHILLIPS, A. B., **Transistor Engineering,** 1962. McGraw-Hill Book Company, Inc., New York.

ROEHR, W. D. (Ed.), **Switching Transistor Handbook,** 4th Printing, 1967. Semiconductor Products Division, Motorola Inc., Phoenix, Arizona.

ROWE, J., **An Introduction to Digital Electronics,** 1967. Sungravure Pty. Ltd., Sydney.

WALSTON, J. A., and MILLER, J. R. (Ed.s), **Transistor Circuit Design,** 1963. McGraw-Hill Book Company, Inc., New York.

Chapter 14

THYRISTOR DEVICES

The PNPN thyristor structure — its behaviour — internal regeneration — current vs. gain, and the choice of silicon — methods of triggering — breakover, and the Shockley diode — gate triggering and the SCR — light triggering and the LASCR — related devices — bidirectional thyristors — device ratings — dv/dt and di/dt — applications.

The semiconductor devices which we have examined in the preceding chapters are all based on crystalline structures having either one, or at most two P-N junctions. We may now turn to consider a further important group of devices, based on a slightly more complex structure in which there are three main P-N junctions: the group of devices known as **thyristors.**

There are quite a large number of devices grouped under the designation "thyristors," and superficially some of these devices may seem very different. Despite this, virtually all thyristor devices are based upon a common fundamental three-junction structure, fabricated from silicon material. In its basic form, this structure has the PNPN configuration shown in figure 14.1(a).

Probably the most important characteristic of this structure is that it possesses the ability to operate in two stable conduction states. In one of these states, called the "off" or **blocking** state, it passes only saturation and leakage current, behaving in a very similar fashion to a reverse biased P-N junction. Conversely in the second state, called the "on" or **conducting** state, it is capable of passing very heavy current, its behaviour in this case being very similar to that of a forward-biased P-N junction.

Besides being able to operate in these two states, the PNPN structure is capable of switching extremely rapidly from the blocking state to the conducting state. This makes it very suitable for use as a power switching element, and also makes the structure a solid state equivalent of the older gas-filled thyratron switching tube. It was recognition of this equivalence which provided the rationale behind the term "thyristor."

As will be explained shortly, there are a variety of methods whereby the basic thyristor PNPN structure may be triggered into switching from the blocking to the conducting state. And although most thyristor devices are capable of being triggered by more than one of these possible methods, the majority of device types are designed to permit efficient and reliable triggering by one particular method. Hence it is broadly true that the wide variety of thyristor devices differ from one another mainly in terms of the provision made for triggering.

The basic operation of the PNPN thyristor structure may be understood by reference to figure 14.1. As shown in (a), the structure is normally connected so that the P-type end is connected to the positive polarity of the supply, becoming the device "anode," while the N-type end is connected to the negative supply polarity and becomes the "cathode." The load is connected in series with the device and the supply, usually in the anode lead.

It may be seen that this connection has the result that the outer P-N junctions, marked "1" and "3", are potentially forward-biased, but the centre junction "2" is reverse-biased. Hence because this reverse-biased central junction is in series with the other two, one would expect the device as a whole to behave in a very similar fashion to a reverse-biased diode. And this is precisely the way the structure does behave if the supply voltage Vf is slowly increased from zero to a moderate level. Only saturation and leakage currents flow, the magnitude of these being very small due to the silicon material involved. Fairly obviously, this corresponds to the "blocking" conduction state of the PNPN structure.

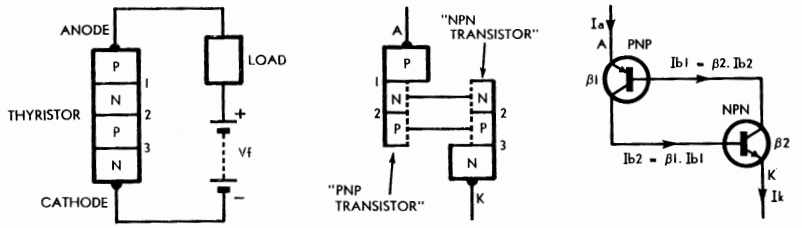


Figure 14.1 (a) (b) (c)

The conditions present within the structure in the alternative "on" state are perhaps less obvious. However, they may be visualised fairly readily by examining the mechanisms involved when the structure is triggered into switching from blocking into heavy conduction. Although a variety of possible methods exist whereby this switching may be triggered, as noted earlier, there are actually only two basic switching mechanisms involved.

One of these mechanisms involves an internal regeneration or positive feedback loop present in the PNPN thyristor structure.

Like the two junctions of a bipolar transistor, the three junctions of the PNPN thyristor structure are capable of interacting in such a way that the mechanisms of injection, diffusion and collection can produce current amplification. However, the presence of the additional P-N junction and the configuration of the resulting PNPN structure both have the additional effect that this amplification action is not only increased, but is also effectively formed into a continuous internal positive feedback loop.

This may be readily understood if the PNPN structure is visualised as effectively consisting of a PNP-NPN bipolar transistor combination, sharing a common collector-base junction such that the base region of each device is the collector region of the other. That this analysis is a valid one may be seen from figure 14.1(b), where the two "hidden transistors" within the PNPN structure have been separated.

As may be seen, the "PNP" transistor is effectively formed from the three upper regions of the PNPN structure, involving junctions 1 and 2, while the "NPN" transistor is effectively formed from the three lower regions and involves junctions 2 and 3. Junction 2 thus forms the collector-base junction of both devices.

A brief examination of the diagram should reveal that, because of the PNPN configuration, the "input" current of each of the two constituent transistors is formed by the "output" current of the other. Thus the collector current of the PNP device forms the base current of the NPN device, while the collector current of the latter in turn forms the base current of the former. This is demonstrated in the schematic diagram of figure 14.1(c).

From this it may be seen that the two transistors are effectively connected in a regenerative and potentially unstable feedback loop. Any current passed by one will tend to be amplified by the other, then passed back to the first to be amplified again, and so on, the device current tending to rise rapidly and without obvious limit.

One might thus expect that immediately following the application of supply voltage to the PNPN structure,

it would regeneratively amplify its own saturation and leakage currents in this fashion, and rapidly draw the maximum current possible from the supply.

To understand why such spontaneous amplification of saturation and leakage currents does not occur, it is necessary to consider the second basic mechanism involved in thyristor operation. This mechanism is associated not with the PNPN configuration of the device structure, but rather with the deliberate use of silicon as the semiconductor material rather than any other.

It may be recalled from chapter 11 that the current gain of a silicon bipolar transistor falls away at low current levels, primarily due to the effect of carrier recombination at so-called "recombination centres" in the emitter depletion layer. Thus like any other silicon bipolar transistors, the transistors constituting the PNPN thyristor structure tend to exhibit lower and lower amplification at reducing current levels.

As explained in chapter 11, the fall in current gain of normal silicon bipolar transistors at low current levels tends to be rather an embarrassment, as it limits the effective input resistance and gain of the device in typical amplifier applications. And, for this reason, silicon transistor manufacturers have directed considerable effort toward reducing the effect with these devices.

However with thyristor devices the effect is actually exploited, because it provides a means whereby the PNPN structure is able to remain stably in the low-current blocking state until intentionally triggered. By maintaining the gain of both the internal transistors of the PNPN structure below unity at the current level corresponding to the saturation and leakage currents, it thus prevents regeneration and current increase.

This should explain why thyristor devices are made almost exclusively from silicon semiconductor material. With other materials, such as germanium, not only is the fall-off in gain at low current levels somewhat less rapid than with silicon, but at the same time the saturation and leakage current levels tend to be somewhat higher at normal operating temperatures. Both these differences tend to make it very much harder to prevent a PNPN structure from spontaneously regenerating, so that thyristor devices made from these materials tend to be impractical.

It is the very low gain of the internal transistors at the low saturation and leakage current levels, then, which prevents the silicon PNPN structure of a thyristor device from regenerating, and allows it to remain stably in the blocking state. How then, the reader may well be asking, is the device triggered into regenerating and switching into its high conduction state?

This is achieved quite simply, by causing a brief intentional increase in the current passing through any one or more of the three device junctions. Provided that this increase is sufficient to raise the product of the current gains of the two internal transistors above unity, regeneration will then occur and the device will consequently drive itself rapidly into the heavy conduction state. Once this regeneration process begins, the initial cause of the triggering current increase may be removed without effect, because the regeneration process is self-maintaining once having been initiated.

In switching itself to the conduction state, the PNPN structure draws a rapidly increasing current, while at the same time its voltage drop falls sharply. In a typical switching circuit such as that of figure 14.1(a), this process ceases only when the current reaches a value where the two internal transistors of the thyristor enter saturation. When this occurs the regenerative action again ceases, because it may be remembered that saturation of a bipolar transistor involves a rapid drop in current gain.

Having entered the heavy conduction state, a thyristor thus remains stably in this state. Its voltage drop is basically that of the two internal transistors in saturation, being typically between 0.7V and 2.5V. The current level flowing through the device from anode to cathode (conventional current flow) is thus limited almost entirely by the supply voltage Vf and the load resistance of the load.

As the switching of a thyristor may be triggered by temporarily increasing the current through any one or more of the three device junctions, this makes it possible to trigger the device in a number of ways. As noted earlier, it is the consequent variety of possible triggering methods which has in fact resulted in the wide number of different thyristor devices in present use.

One possible way of triggering a device is simply to increase the effective anode-cathode voltage applied to the device, either steadily or with a short pulse superimposed upon the supply. By raising the anode-cathode voltage to the point where leakage current itself reaches the level required to raise the internal gain product above unity, regeneration is initiated as before.

Although this method of triggering may be used with almost all thyristor devices, it is virtually the only triggering method possible with one particular device. This is the **Shockley diode**, so named to commemorate its prediction from theory by physicist William Shockley. The first actual device was developed in mid-1956 by researchers Moll, Tannenbaum, Goldey and Holonyak of Bell Laboratories. Other names sometimes used for the Shockley diode are "PNPN diode," "four-layer diode," and "breakover diode."

As may be seen from figure 14.2 (a) where a simple diagram of a Shockley diode is shown together with its alternative schematic symbols, this device is basically identical with the elementary PNPN device shown in figure 14.1. It is thus the simplest of the thyristor device "family."

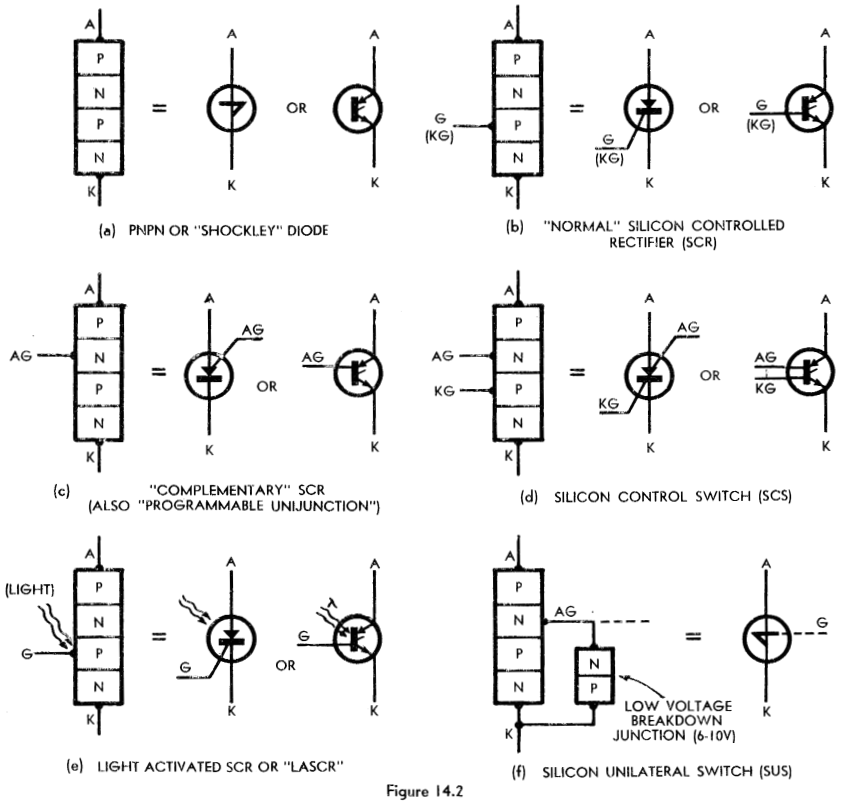


Figure 14.2

The characteristic of a typical Shockley diode is shown in figure 14.3 (a). It may be seen that upon application of forward voltage Vf the device remains initially in the low current blocking state. However if Vf is increased to the "breakover voltage" Vbo of the device, regeneration occurs and the device rapidly drops back through the unstable negative resistance switching region to reach the high current conduction (saturation) region.

The device will remain in the high current region unless, or until its current is forced by the external circuit conditions to drop below a certain "holding current," shown on the diagram as Ih. While in the high current region the device characteristic closely approximates that of a normal forward-biased P-N diode. When reverse-biased the device also behaves in a manner which closely approximates a P-N diode with reverse bias, the current remaining very low until one or both of the reverse-biased "outer" junctions enters avalanche breakdown.

A second possible way of triggering the basic PNPN structure of a thyristor is by injecting additional current carriers into either of the semiconductor

regions adjacent to the central P-N junction. This has the effect of supplying base current to one or other of the two internal transistors, resulting as before in the rise in device current levels necessary for the gain to rise and initiate regeneration.

Thyristor devices designed especially to be triggered in this way are provided with a third electrode connected to one or other of the two central semiconductor regions, to permit convenient injection of carriers. This electrode is generally referred to as a "gate," being alternatively designated a **cathode gate** when associated with the central P-type region, or an **anode gate** when associated with the central N-type region.

This type of device has become known as a **Silicon Controlled Rectifier,** or "SCR," although controlled rectification of AC forms only one of its many applications. The first SCR device was developed in 1957 by Gordon Hall, a semiconductor device engineer working at the General Electric rectifier plant in Clyde, N.Y.

Because it has been found easier to fabricate high power SCR devices in the configuration designed for triggering from a cathode gate electrode, as shown in figure 14.2 (b), this configuration has become known as the "normal" SCR configuration. Accordingly the alternative type of device having an anode gate has become known as a "complementary" SCR, as shown in 14.2 (c).

Low power devices having the same basic configuration as that of figure 14.2 (c) are also called **programmable unijunctions.** This term is used because they may be arranged quite easily, in a suitable circuit configuration, to perform the functions of an adjustable-parameter unijunction. Actually low power SCRs of both the "normal" and "complementary" configurations may be used in this fashion.

The characteristic of a typical SCR device is shown in figure 14.3(b). As may be seen, for the zero gate current case (Ig=0) it is basically identical with the characteristic of the Shockley diode shown in (a). However, in this case the switching or breakover voltage may be reduced from the value Vbo, by the injection of gate current. Increasing values of gate current Ig1, Ig2, Ig3 and Ig4 thus result in the reduction of breakover voltage to values Vb1, Vb2, Vb3 and Vb4 respectively.

In passing it should perhaps be noted that to trigger the PNPN structure by means of a cathode gate, a forward bias is applied between this electrode and the cathode, whereas triggering by means of an anode gate is achieved by application of forward bias between this electrode and the anode.

As one might perhaps expect, it is possible to construct a thyristor device having both an "anode gate" and a "cathode gate" — in other words, a device with gate electrodes connected to both the internal N-type and P-type regions of the PNPN structure. Such devices are made, being given the name **Silicon Controlled Switch** or "SCS."

Although generally only capable of operating at relatively modest power levels, SCS devices find many applications because of the flexibility offered by the two gate electrodes. The schematic symbols used for SCS devices are shown in figure 14.2(d), while the characteristic is very similar to that of the SCR shown in figure 14.3(b).

It may be noted that in the foregoing discussion of SCR and SCS devices, no mention has been made of any mechanism whereby the gate electrode(s) may be used to switch a device "off." The reason for this is that with most SCR and SCS devices the gate electrode(s) is functionally almost identical with the grid electrode of a gas-filled thyratron valve, being capable of initiating device turn-on, but incapable of producing turn-off once the device is conducting. Thus in normal use they are turned off by arranging for the anode-cathode voltage to drop below the value which produces the "holding current" Ih shown in figure 14.3(b).

By the adoption of special device geometries, by careful control of doping levels and by considerably reducing the current densities reached within the devices, manufacturers have in fact been able to produce thyristor devices capable of being turned off by a large reverse bias applied to a gate electrode. These have usually been called **Gate Turnoff Switches** (GTO) or **Gate Controlled Switches** (GCS). However, devices of this type have not become widely used, mainly because their function can generally be duplicated more economically using a silicon bipolar switching transistor.

A third available method of triggering the PNPN structure of a thyristor is to increase the excitation energy of the crystal lattice, by the application of additional light or heat. This has the effect of increasing the generation of "intrinsic" electron-hole carrier pairs, and thus results in an increase in the device saturation currents. Naturally this mechanism is again capable of initiating device turn-on, providing the current levels are increased to the level required for regeneration to take place.

Although relatively few applications would appear to exist for a thyristor device triggered by long-wavelength heat energy, a significant number of applications have been found for a device capable of being triggered by infra-red and visible radiation. Device manufacturers have accordingly been motivated to produce devices capable of being triggered by this type of radiation.

Generally such devices employ the basic PNPN thyristor configuration but with a modified, "flat" geometry designed to allow improved penetration of the semiconductor die by the triggering radiation. The case or package in which the device is encapsulated is provided with a "window" covered with mica, glass or a suitably transparent plastic material.

While it would be feasible to produce a diode device of this type, most light-triggered thyristors are in fact provided with at least one normal gate electrode. This is provided to allow electrical control of the radiation sensitivity of the device. Thus practical light-triggered thyristors are either of the **Light-Activated SCR** (LASCR) variety, having a single gate electrode as illustrated in figure 14.2(e), or of the **Light-Activated SCS** (LASCS) variety with two gate electrodes.

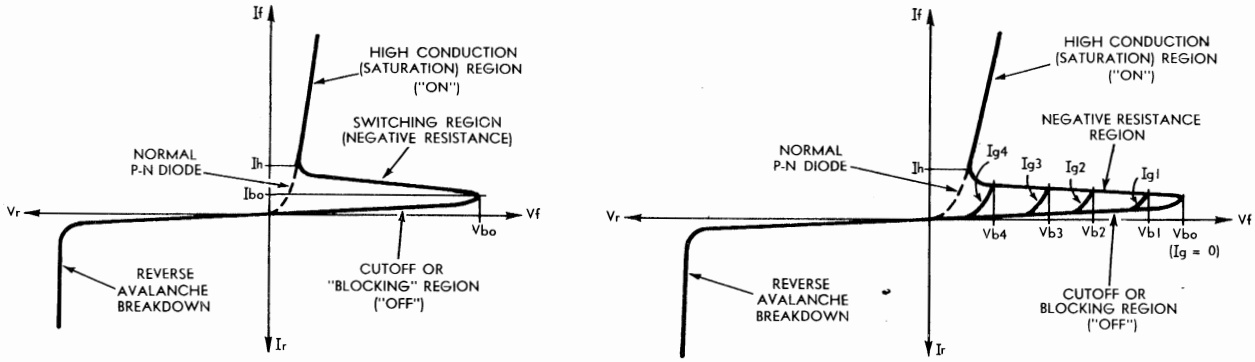


Figure 14.3 (a) PNPN DIODE CHARACTERISTIC (b) SCR CHARACTERISTIC

In addition to the thyristor devices which are designed to be triggered by one of the three basic methods just described, there have appeared a number of devices designed to be triggered in more complex ways. One such device is the **Silicon Unilateral Switch** or SUS, whose basic structure and schematic symbol are illustrated in figure 14.2(f).

As may be seen, this device is basically a complementary SCR with an inbuilt breakdown or "zener" diode junction connected between anode gate and cathode. The idea behind this is that the PNPN structure is triggered into conduction only when the voltage applied to the device exceeds that necessary to produce breakdown in the auxiliary junction. As the breakdown voltage of this junction can be quite accurately controlled, and made as low as 6-10V, the SUS can thus be used as a close-tolerance low voltage equivalent of the Shockley diode.

It may be noted that all of the thyristor devices described in the foregoing are **unidirectional** — i.e., their thyristor action applies for only one polarity of the applied anode-cathode voltage. This means that if such devices are to be used in applications where thyristor action is required for both supply polarities, as in AC circuits, it is generally necessary to use

either two devices in inverse parallel, or a single device in conjunction with some type of rectifier circuit.

Happily such circuit complication may be obviated in at least some AC applications, because there exists a further group of thyristor devices which are in fact capable of **bidirectional** operation. Three of these devices are in common use, one being a bidirectional diode device, another a bidirectional triode, and the third a symmetrical version of the SUS device. All three may be regarded as developments from the basic PNPN structure of figure 14.1.

The bidirectional diode thyristor consists of a modified PNPN structure which behaves as if it consisted of two Shockley diodes connected in inverse parallel. Thus for either polarity of applied voltage it behaves as a reverse-biased junction until its breakover voltage Vbo is reached, whereupon it regenerates and conducts as before.

The first device of this type was developed by the Hunt Electronics Corporation of Dallas, Texas, in the early 1960s. Currently a device of this type is marketed by the STC-ITT organisation under the name "Sidac." The basic structure of a typical device is shown in figure 14.4(a), together with the alternative schematic symbols, while the characteristic is represented by the heavy curve in 14.4(c).

The bidirectional triode thyristor or **Triac** device is similar to the diode device, but represents a further modification of the basic PNPN structure to allow triggering in both directions by means of a single gate electrode. Its behaviour is thus very similar to that of two SCR devices connected in inverse parallel.

The Triac was developed by General Electric in 1964. As may be seen from figure 14.4(b), its internal configuration is relatively complex. Because of this it tends to be rather difficult to produce.

The single gate electrode of the Triac controls its breakover for both polarities of the applied voltage. The control action is very similar to that of an SCR gate, with increasing gate current levels corresponding to reduced breakover voltages. This is illustrated by the dashed curve segments on the characteristic of figure 14.4(c).

The third type of bidirectional thyristor device in current use is the **Silicon Bilateral Switch** (SBS). This is again a General Electric development, being essentially an inverse parallel combination of two SUS devices of the type shown in figure 14.2(f). Hence by analogy with the relationship between the SUS and the Shockley diode, the SBS forms a close-tolerance low voltage equivalent of the bidirectional diode thyristor.

Although brief, the foregoing survey includes practically all of the thyristor devices in significant use at the time of writing this chapter. However, mention should perhaps be made in passing of a further device which — although not strictly a thyristor at all — is often included for convenience in the thyristor device "family."

This device is the **Diac,** which is a bidirectional breakover or trigger diode frequently used for triggering the Triac. Developed by General Electric, the Diac behaves in a rather similar fashion to the bidirectional diode thyristor; it switches into conduction when an applied voltage of either polarity exceeds about 30V. However the device is not a thyristor, being in reality a three-layer PNP structure rather like a symmetrical bipolar transistor without a base electrode. Its operation involves a relatively straightforward mechanism of amplified avalanche breakdown.

Like the basic performance parameters and ratings of the other semiconductor devices examined in previous chapters, those of thyristor devices are to a large extent controllable by manipulation of doping levels and device geometry. Thus it is possible to fabricate thyristors having breakover voltages falling over a very wide range, from as low as a few volts for some SUS and SBS devices to as high as 10,000V for specialised high-power SCR devices.

Current and power ratings are similarly controllable over a very wide range. Some very low power SCS and programmable unijunction devices are rated for operation at current levels in the order of a few tens of milliamps, while heavy-duty SCR devices intended for such applications as electric traction control circuitry may have current ratings as high as 1,600 amps.

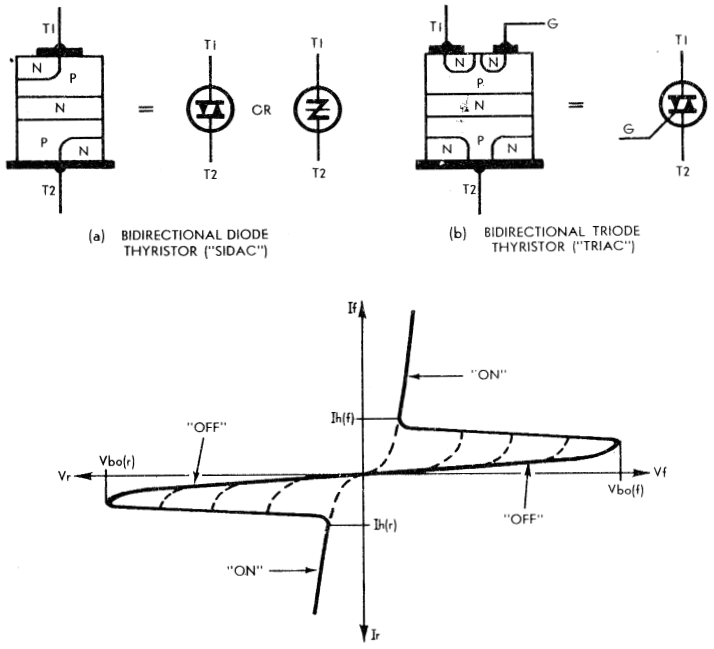


Figure 14.4

Apart from voltage and current ratings, however, there are two further ratings which play an important part in determining the suitability of a thyristor device for a given application. These ratings are rather unique to thyristor devices, both being concerned with the rates of change of voltage and current.

One of the ratings defines a maximum rate of change of the supply voltage applied to a thyristor device. This is known as the **dv/dt rating.**

The reason for the dv/dt rating is that any thyristor can be triggered into conduction from the forward blocking state, at a supply voltage far below its breakover voltage Vbo, if that supply voltage is applied sufficiently rapidly. This is the so-called **rate effect,** which is due to the fact that in the forward blocking state the reverse biased central junction of the PNPN structure has significant depletion capacitance.

In effect, this capacitance provides yet another mechanism whereby a thyristor may be triggered. Like any other capacitance, it tends to draw a reactive current proportional to the rate of change of applied voltage. Hence if the supply voltage is applied to the thyristor sufficiently rapidly, this reactive current will reach a value sufficient to initiate regeneration.

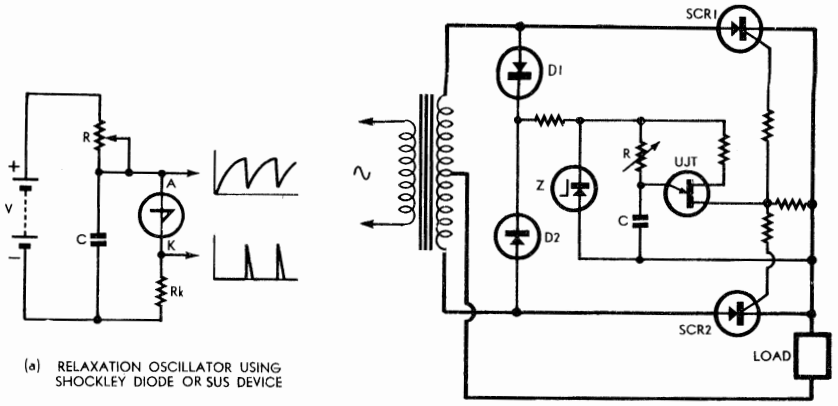
In a few thyristor devices, the rate effect is actually used as a means of triggering; an example is the "Sidac" bidirectional diode, which is usually triggered by a fast-risetime pulse superimposed on the AC supply. However, in most cases thyristors are intended to be triggered by one of the methods discussed earlier, and thus precautions must be taken to ensure that spurious additional triggering does not occur due to rate effect.

From this it may be evident that the dv/dt rating of a thyristor is equally important whether rate effect triggering is to be avoided, or to be exploited. If rate effect triggering must be avoided, then the dv/dt rating indicates the maximum allowable rate of change of applied voltage. Conversely if the device is to be triggered by this means, then the dv/dt rating represents the rate of change which must be adequately exceeded by the intended trigger pulse for reliable triggering.

The dv/dt rating of thyristor devices may be controlled by manipulation of the doping levels and geometry, and hence practical devices have dv/dt ratings which vary over a wide range to suit the intended applications.

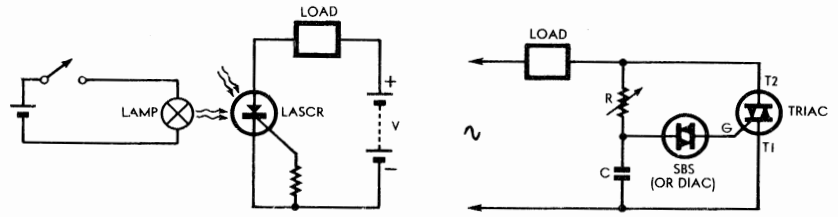
The second of the unique thyristor device ratings defines the maximum allowable rate at which the current drawn by the device may be permitted to increase when the device is triggered from the blocking state into conduction. This is known as the inrush current rate, or **di/dt rating.**

The reason for the di/dt rating is that no practical thyristor device is



(a) RELAXATION OSCILLATOR USING SHOCKLEY DIODE OR SUS DEVICE

(b) PHASE-CONTROLLED FULL WAVE RECTIFIER USING SCR DEVICES



(c) REMOTE SWITCHING RELAY (LATCHING) USING LASCR

(d) PHASE-CONTROLLED AC CIRCUIT FOR LAMP DIMMING, ETC.

Figure 14.5

capable of switching from the blocking to the fully conducting state instantaneously. A finite time is required for the new charge conditions appropriate to the fully conducting state to distribute over the device junctions.

When anode-cathode current initially begins to flow, it is localised in a relatively small area of the junctions. In the case of a device triggered by means of a cathode or anode gate, the current is initially localised in the area of the junctions adjacent to the gate contact, because of the bulk resistance of the various semiconductor regions. A short "spreading" time is required before the current distributes itself evenly over the full area of the device junctions.

Because of the initial current localisation, the maximum current which may be safely withstood by a thyristor device immediately after triggering tends to be only a fraction of its full rated current capacity. Only as the current distributes over the full area of the device junctions does the current level, corresponding to the threshold of overheating and damage, rise sufficiently to allow the device to accept its full rated current.

To prevent the device from being damaged, then, it is necessary to arrange that the circuitry associated with the thyristor limits the rate of increase or "inrush" of conduction current so that this does not exceed the rate at which the device junctions "turn on." And this is the significance of the di/dt rating specified by the thyristor device manufacturer.

Typical "standard" thyristor devices have di/dt ratings falling between about 30 and 200 amps/microsecond. However, high power SCR devices with di/dt ratings as high as 600 amps/microsecond have recently been developed by National Electronics Inc., of Illinois. These devices employ a special "regenerative gate" triggering mechanism, whereby the initial localisation of current in the device is itself arranged to promote current distribution and rapid turn-on.

In the remaining short space available in this discussion of thyristor devices, a brief survey will be given of some of the more common applications of the devices.

Because of its characteristic, the Shockley diode makes an almost ideal voltage-sensitive switching element. So too does the SUS device, which provides essentially the same characteristics at somewhat lower voltage levels. Both devices thus find use in many types of switching and pulse circuitry.

A common application is in simple R-C relaxation oscillator circuits, used for sawtooth wave and pulse generation. A simple circuit of this type is shown for illustration in figure 14.5(a), where it may be seen that the thyristor element performs a function identical with that of the unijunction of figure 7.9, or the familiar neon lamp.

Probably one of the most common applications of SCR devices is in controlled rectifier circuits, for which their gate-triggered facility makes them very well suited. In this respect the SCR forms a worthy successor to earlier discharge devices such as the hydrogen thyratron and the ignitron.

The diagram of figure 14.5 (b) illustrates a full-wave controlled rectifier circuit using two SCR devices (SCR1, SCR2). The conduction of the SCRs is controlled in this type of circuit by adjustment of the phase of the triggering pulses fed to the device gates. Hence by retarding the triggering pulses to a point relatively late in each half-cycle, the SCRs are arranged to conduct for only a small portion of the full half-cycle, and the DC load current is relatively small. Conversely, by advancing the triggering pulses to a point early in each half-cycle, the SCRs are allowed to conduct for a greater proportion of the time, and accordingly the DC load current is increased.

In the circuit shown the phase-control is achieved by deriving the SCR triggering pulses from a relaxation oscillator employing a unijunction transistor (UJT). The supply for the oscillator is derived from the AC supply across the transformer secondary, being full-wave rectified by diodes D1 and D2, and clipped to a suitable level by zener diode Z.

Because there is no filtering in the oscillator supply, its operation is synchronised with the AC supply. Hence at the beginning of each supply half-cycle, capacitor C begins to charge up to the firing point of the unijunction. By varying resistor R, the time taken to reach the unijunction firing point may be adjusted between a point very early in the half-cycle and a point very late. Hence R becomes the control which determines SCR triggering phase and average DC load current.

A simple but very useful application of light-triggered devices such as the LASCR is in remote switching relay applications, as illustrated in figure 14.5(c). Here the combination of a lamp and the LASCR essentially behaves in the same manner as a conventional electro-magnetic relay, offering complete isolation between control and load circuits. In addition the combination offers considerably improved reliability, increased operating speed and freedom from contact bounce.

If a DC supply is used in the load circuit, as shown, the relay is self-latching because the LASCR remains in the conduction state even if the lamp is subsequently extinguished after being lit. However if a non-latching relay is required, this can be achieved simply by employing an AC or unfiltered rectifier supply in place of the DC load supply.

Bidirectional devices such as the Triac, Sidac and SBS are very attractive for AC power control applications, their characteristics allowing considerable circuit simplification compared with other devices. This is well illustrated by the circuit of figure 14.5(d).

As may be seen, the use of a Triac device together with an SBS or Diac for triggering allows the circuit to be reduced to a bare assembly of four components. Together with the two semiconductors there is only the charging capacitor C and the variable resistor R used to vary the triggering phase. This provides a complete low-cost lamp dimming circuit which, in domestic applications, may be fitted if necessary into the wall cavity formerly occupied by the conventional flush switch.

SUGGESTED FURTHER READING

CLEARY, J. F., (Ed.) **General Electric Transistor Manual,** 7th Edition, 1964. General Electric Company, Syracuse, New York.

GUTZWILLER, F. W., (Ed.) **SCR Manual,** 4th Edition, 1967. Semiconductor Products Department, General Electric Company, Syracuse, New York.

HEY, J. C., "The Widening World of the SCR," in **Electronics,** V.37, No. 25, September 21, 1964.

ROWE, J., "The Regenerative Gate SCR," in **Electronics Australia,** V.30, No. 11, February 1969.

Chapter 15

DEVICE FABRICATION

Devices and their fabrication — refining the raw semiconductor materials — zone refining and float zone melting — growth of monocrystals — sawing into wafers — passivation — epitaxial deposition — selective diffusion — photolithography and oxide masking — multiple diffusions — contact metallisation — probe testing, scribing, cleaving, die and wire bonding — encapsulation and classification.

From the discussion of each of the various types of semiconductor device treated in the foregoing chapters, it may be apparent that the characteristic behaviour of each device type is very much a function of its particular configuration of semiconductor regions and junctions. Hence it is generally true that every device of a given type has the same basic configuration, this configuration in each case being very similar if not identical with that which we have used to explain basic device operation.

In dealing with each type of device, we have until now simply assumed the existence of its particular configuration of regions and junctions. This has been a justifiable assumption, because a discussion of actual device fabrication is not only unnecessary, but also largely irrelevant in a treatment of basic device operation and applications. It is a fact also that most of the techniques and processes involved in device fabrication are common to all modern semiconductor devices, so that rather than treat these in a piecemeal and distributed manner, it is really more appropriate to accord them a separate and unified discussion.

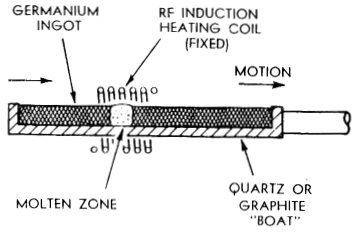
A very suitable place for such a discussion of device fabrication is provided in this treatment by the present chapter. At this point the discussion can at the same time both round out the treatment of discrete devices given in the preceding chapters, and also provide most of the foundation concepts necessary for an easy transition to the discussion of microcircuits to be given in subsequent chapters. Accordingly, the chapter will be directed towards providing the reader with a brief but useful introduction to fabrication techniques and processes.

As one might perhaps expect from earlier chapters, the first step in semiconductor device fabrication involves the preparation of extremely pure semiconductor material.

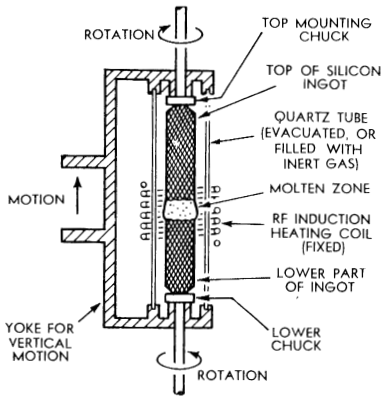
Because the performance of most semiconductor devices is highly dependent upon the actual impurity doping levels and gradients which are ultimately present in the various device regions, these doping levels and gradients must of necessity be tightly controlled and reliably maintained. To allow this to be achieved, it is generally essential that the semiconductor material used for device manufacture is initially purified such that it becomes virtually "intrinsic" material.

It may be recalled from chapter 3 that this involves the reduction of total impurity concentration in the material to less than one part in 10^9 — or in more familiar terminology, refinement to a level where the material is 99.9999999% pure.

This degree of refinement is considerably beyond that which could be achieved using the traditional physical and chemical methods available when the "semiconductor revolution" began in 1948 with the discovery of the bipolar transistor. Accordingly, the developing semiconductor industry has been forced to develop its own specialised refinement techniques, which are used to perform further extensive purification after the "raw" semiconductor materials have been refined to the limit of traditional techniques.



(a) ZONE REFINING



(b) FLOAT ZONE REFINING

Figure 15.1

One of the earliest of these special refinement techniques to appear was the **zone refining** technique developed in 1954 by W. G. Pfann of the Bell Laboratories. This technique is still used for the refining of germanium material, and is illustrated in basic form in the diagram of figure 15.1 (a).

As may be seen the process involves the placing of an ingot of chemically pre-refined germanium in a long crucible or "boat" of either graphite or quartz, which is then moved slowly and repeatedly in a horizontal direction through the load coil of an RF induction heater. The rate of motion is adjusted so that the portion of the ingot within the RF heating coil is maintained in the molten state, in effect creating within the ingot a zone of molten metal which is caused to sweep repeatedly in one direction along the entire length.

How does this operation reduce the impurity content of the ingot? The answer to this lies in an interesting and most fortuitous effect known as **segregation,** wherein virtually all of the impurities which contaminate the commonly used semiconductor materials prove to be less soluble in the solid phase of the host semiconductor material than in the liquid phase. The semiconductor material which is re-crystallising from the liquid phase thus tends to contain a lower impurity concentration than any remaining liquid, because the differential solubility results in a tendency for the impurity atoms to remain in the liquid.

Because of the segregation effect, the "sweeping molten zone" tends to accumulate the impurities from the ingot. It thus behaves rather like a small magnet drawn through a mixture of non-magnetic powder and iron dust. If the operation is stopped after a number of passes through the RF heating coil, with the molten zone at one extreme of the ingot, it is found that almost all of the impurities in the cooled ingot are concentrated at the end which finally solidified. This portion may then be cut off, leaving the remainder of the ingot in a highly refined state.

It is possible to repeat the zone refining process almost indefinitely, each time producing material having a lower impurity concentration. However, in practice economic considerations dictate that the process is continued only until the impurity concentration is re-

duced to a level sufficient to allow adequate control over ultimate device performance. As chemical tests are not capable of showing when the required degree of refinement has been reached, the indicator used to determine this is the rising electrical resistivity.

For a variety of reasons, the zone refining process in the form illustrated in figure 15.1(a) is not suitable for the refinement of silicon material. Silicon has a higher melting point than germanium (1420°C vs. 960°C), and tends to react strongly with both the atmosphere and a graphite crucible at temperatures near the melting point. On the other hand it is also incompatible with a quartz crucible, because "wetting" causes adhesion between the two, and this causes a problem due to the difference in temperature coefficients of expansion.

Because of these problems, silicon refining is performed using a modified zone refining technique developed by H. C. Theuerer of Bell Laboratories. This is the ingenious float zone technique, which is illustrated in figure 15.1 (b).

As may be seen, the float zone technique obviates the need for a crucible, by supporting the silicon ingot vertically between two rigidly separated chucks. The chucks are located at the ends of a large quartz tube which forms a protecting chamber around the ingot, this chamber being either evacuated or filled with an inert gas. The entire assembly of chucks, ingot and protecting chamber is then moved slowly up and down through the RF heating coil, to produce the same "sweeping molten zone" effect as before.

In this arrangement the molten zone is supported solely by its own surface tension, the length of the zone being adjusted carefully to ensure that it does not collapse. Although the ingot is rotated during the refining process by means of the chucks, to ensure even heating and thorough segregation, this is done very slowly to prevent disturbance of the molten zone due to centrifugal effects. As a further precaution the ends of the ingot are rotated in opposite directions.

Although it is initially necessary to refine the semiconductor materials for device fabrication using the foregoing techniques, the highly refined material obtained is not normally used in this state for device manufacture. Rather, it proves convenient to dope the material following refinement with carefully controlled quantities of a donor or acceptor impurity, to obtain uniformly doped N-type or P-type material having a known resistivity.

This preliminary doping process is generally performed during the next main fabrication step, which is the operation of **crystal pulling.** Here the semiconductor material is converted to a large single crystal or "monocrystal," having a consistent lattice structure throughout.

Figure 15.2: A germanium monocrystal boule being "pulled" from the melt, by the Czochralski technique, at left, while at right is a silicon boule grown in a similar fashion. (Courtesy Delco Radio, Fairchild Australia.)

It is essential that semiconductor material used for device fabrication be in the monocrystalline form, because in a multi-grain crystal structure the lattice discontinuities formed by the crystal grain boundaries produce spurious effects which completely swamp out the mechanisms responsible for normal device operation.

Germanium ingots produced by the zone refining process are generally in a polycrystalline form, and it is therefore essential that the further process of crystal pulling be used to convert the material into a monocrystal. In contrast, the silicon material produced by the float zone process is already in monocrystalline form. However, this material is usually also subjected to the crystal pulling process, if only to achieve the required pre-doping.

The crystal pulling process was developed by J. C. Czochralski. It involves the melting of the refined semiconductor material in a quartz crucible by an RF induction heater, which then maintains the melt at a temperature slightly above the melting point. A small single crystal of solid material is then introduced into the top of the melt, in a suitable crystalline orientation, and then slowly withdrawn. The crystal acts as a recrystallisation centre or "seed," and progressively grows into a large monocrystal. This may be seen in the photographs of figure 15.2.

The pre-doping of the material is achieved by adding a "pill" of pure dopant material to the melt prior to the crystal pulling operation. To ensure uniformity of doping in the final monocrystal, the melt is gently stirred just below the crystallising level, by slow rotation of the seed crystal during the pulling operation. The doping concentration in the liquid melt is deliberately made higher than that required in the final monocrystal, to allow for the segregation effect.

In general two levels of doping are used in the pre-doping process, each producing material intended for the fabrication of devices with particular characteristics. A relatively heavy doping level is used to produce low resistivity material (in the order of .005 ohm-cM), which is used in the fabrication of the so-called "epitaxial" devices to be described shortly. Conversely, a relatively light doping level is used to produce the high resistivity material (from 0.5 to 50 ohm-cM) used in the fabrication of non-epitaxial devices.

The monocrystal "boules" of doped semiconductor material produced by the crystal pulling process form the basic material from which most semiconductor devices are made. Typically a boule measures from 1¼in to 2¼in in diameter, and from 6 to 9 inches long; from it may be made many hundreds of thousands of individual devices.

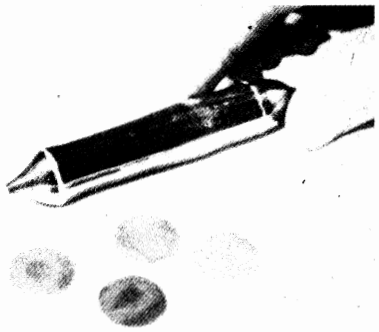


Figure 15.3: A silicon boule compared with the wafers cut from a similar crystal. (Courtesy Mullard-Australia.)

In the relatively few years since semiconductor devices were first produced on a commercial basis, many different alternative techniques have been used to fabricate semiconductor devices from the prepared boules of doped semiconductor material. These techniques have tended to produce a wide variety of different "versions" of most of the semiconductor devices which we have examined in earlier chapters.

Unfortunately no attempt can be made in this chapter to even briefly describe many of these techniques and device variations. To do so would involve considerable space which could scarcely be justified, because many of the techniques and devices are now regarded as obsolete and of purely historical interest.

The fact is that in recent years, one particular group of techniques has emerged as that most capable of achieving low cost, high-yield fabrication of reliable, high performance devices. This group of techniques has virtually eclipsed all others, and is now used almost universally for the fabrication of low and medium power discrete de-

vices of each of the types described in earlier chapters. Not only this, but the same techniques are used in the manufacture of most integrated microcircuits, being in fact the very techniques which made possible the development of these devices.

Accordingly it is these techniques, and the versions of each of the various types of device produced by them, which will be described in the remainder of this chapter. Interested readers will find descriptions of many of the older and now little-used techniques in the established texts, together with descriptions of the corresponding versions of the various types of device.

Throughout the description, silicon material and devices will be assumed, as this material is currently used for the vast majority of devices in commercial production. In many cases the techniques used for germanium and other materials are similar to those described, although some techniques used with silicon are not directly applicable to other materials.

Currently the first step in preparing a monocrystal boule for device fabrication is to slice it transversely into dozens of thin wafers, using a diamond saw. The thickness of the wafers is typically about 15 mils (.015in), or 375uM (microns); one mil being equal to 25uM. A typical boule and some wafers are shown in figure 15.3. Each

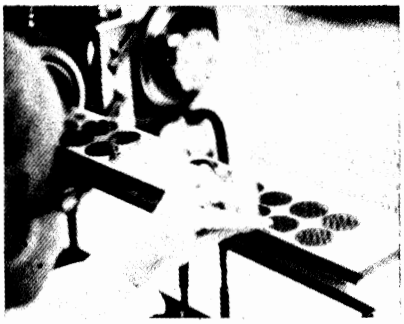


Figure 15.4: Silicon wafers being removed from an epitaxy reactor after deposition and passivation. (Courtesy Fairchild Australia.)

wafer ultimately becomes a complete two-dimensional array of individual devices, which are subsequently separated and individually packaged.

After slicing, one surface of each wafer is lapped and polished to a mirror finish. It is this surface of the wafer which is treated to produce the "active" regions of the device array, and the mirror finish is necessary to ensure precision during the various processes. After the polishing process the wafers are subjected to a chemical etch which removes all traces of sawing and polishing lubricants, and leaves the wafer in an extremely clean condition. Its thickness is now typically between 5 and 10 mils.

At this stage of the fabrication process, the lightly doped high resistivity wafers intended for non-epitaxial devices are simply subjected to the process of **passivation.** This involves the growth of a protective coating of inert material over the surfaces of the wafer, both to protect it from contamination during handling, and to prepare it for subsequent processing.

Typically the passivation coating is composed of silicon dioxide (quartz), or silicon nitride. The former is generally grown on the wafer by heating it to a temperature of 250°C or higher in an atmosphere containing either saturated water vapour, hydrogen peroxide vapour or pure oxygen. An alternative procedure involves heating the wafer to a temperature of between 900 and 1350°C, in an atmosphere containing hydrogen and carbon dioxide.

The heavily doped low resistivity wafers intended for epitaxial device fabrication are not passivated at this stage, but are instead subjected to the process of **epitaxial deposition.** In this process, a thin layer (typically from 5 to 20uM) of lightly doped silicon is grown on the polished surface of each wafer, in such a way that the crystal structure of the layer is aligned with, and virtually an extension of, that of the wafer itself. Each wafer is thereby provided with a high resistivity region surmounting the original low resistivity substrate, without disturbance to its monocrystalline structure.

The process of deposition is performed in an "epitaxy reactor," so named because in order to ensure reliable epitaxial growth of the new material on the wafers, it is necessary to arrange for the silicon to be formed directly at the wafer surface by a thermally triggered reaction between vapours.

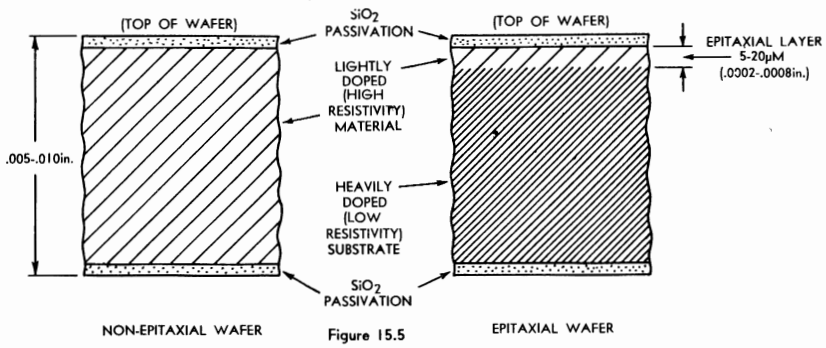


Figure 15.5

In the reactor, RF induction heating is used to heat the wafers themselves to around 1200°C, while the remainder of the reactor is maintained at a relatively low temperature. Dry hydrogen gas is then passed through the reactor for a short period, to chemically reduce any silicon dioxide which may be present on the wafer surfaces. Following this a carefully adjusted mixture of vapours is passed through, whereupon silicon material of the desired doping concentration is formed directly upon the hot wafer surfaces. This process is continued until the epitaxial layer grows to the desired depth.

Typically the principal vapour constituents used for epitaxial deposition are hydrogen and either silicon tetrachloride or silane (silicon tetrahydride), which react together at the hot wafer surface to produce the silicon itself. Doping of the epitaxial layer is performed by adding minute quantities of such dopant vapours as phosphine (phosphorous trihydride), diborane (boron hydride), or arsine (arsenic trihydride).

Immediately following the epitaxial deposition process the wafers are passivated, again both to protect them from contamination during handling, and to prepare them for subsequent processing. Only after the passivation is complete are the wafers removed from the epitaxy reactor, as illustrated in figure 15.4.

At this stage of the fabrication process the wafers for epitaxial and non-epitaxial devices have cross-sections as shown in figure 15.5. In both cases it is at the top of the wafer, and within the lightly doped high resistivity material, that the various functional areas of the devices formed from the wafer are produced in the subsequent processes.

The presence of the heavily doped low resistivity "substrate" region in the epitaxial wafer provides the means whereby a very low resistance connection may be provided to the lowest functional region of each device fabicated from this type of wafer. Hence this type of wafer is used in preference to the non-epitaxial type wherever such a very low resistance connection is required.

An example is bipolar transistors designed for switching applications, where the use of the epitaxial structure gives the devices a very low series collector resistance, and a correspondingly low saturation voltage, Vce(sat).

The various functional regions of the devices to be fabricated from the silicon wafers are currently produced by means of a series of **selective diffusion** processes. These processes involve the diffusion of dopant atoms into the crystal lattice from a concentrated vapour surrounding the wafers, in selected patterns controlled by "windows" formed in the silicon dioxide passivation layer.

Selective diffusion is made possible by the two mechanisms of dopant diffusion and oxide masking.

Dopant diffusion is a mechanism wherein the atoms of an impurity material, like the carriers in an excited semiconductor lattice, tend to diffuse themselves evenly throughout a medium, moving away from regions of high concentration and towards regions of low concentration. Hence if a high concentration of dopant atoms is created at the surface of a heated semiconductor crystal, for example by passing concentrated dopant vapour over the crystal, the dopant atoms will be found to diffuse into the surface of the crystal.

Not surprisingly, the rate at which the dopant atoms diffuse into such a crystal depends upon the dopant concentration produced at the surface, relative to the doping concentration already present in the crystal. In other words, the diffusion rate is proportional to the concentration gradient. It is also proportional to the temperature of the system, proceding more rapidly as the temperature is raised. The dopant distribution produced by diffusion is exponential in shape, decaying from the

surface at a rate proportional to the temperature and duration of the diffusion process.

The diffusion mechanism is potentially a very useful one, providing as it does a means whereby dopants may be introduced into a semiconductor crystal to form regions of any desired type adjacent to the crystal surface. This should become apparent shortly.

The second important mechanism which makes possible the process of selective diffusion is **oxide masking.** This is based on the happy fact that silicon dioxide, even in the form of a thin layer, is virtually "opaque" to almost all of the impurities normally used as silicon dopants.

areas corresponding to the desired diffusion "windows," the plate having been prepared by a high-reduction photographic step-and-repeat process such that it carries an array of many thousands of tiny identical images of a master pattern.

Following this exposure, which is shown in diagram (b), the photoresist is developed and the unexposed photoresist etched away. This leaves the desired "window" patterns as exposed areas of the silicon dioxide passivation, as in (c). The wafer is then immersed in a silicon dioxide etchant, such as hydrofluoric acid, which etches away the exposed passivation, leaving the wafer as shown in diagram (d). Removal of the exposed photoresist material then leaves the wafer with the silicon dioxide layer completely formed into the precision mask required for selective diffusion.

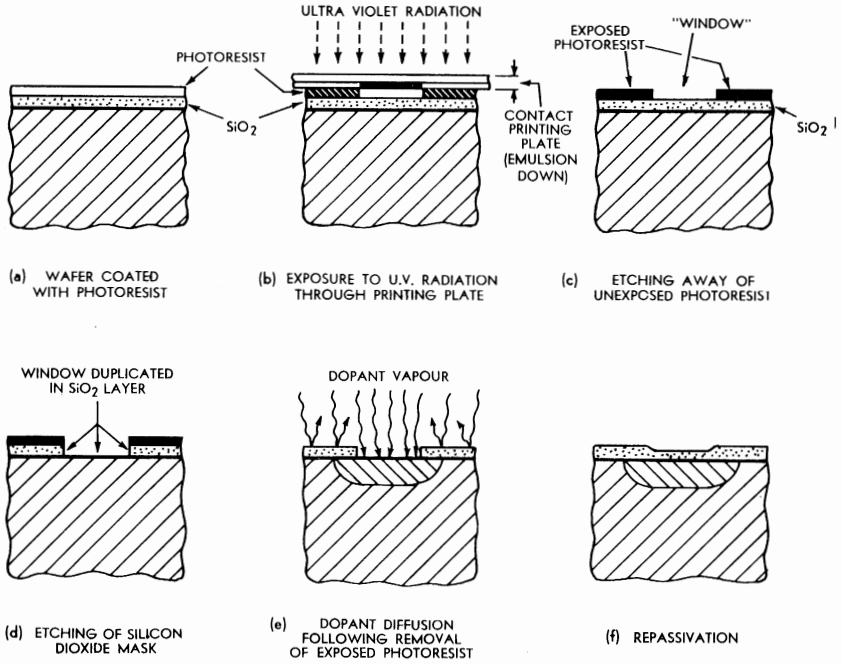


Figure 15.6

The fact that silicon dioxide is "opaque" to the dopants means that the silicon dioxide layer grown on the surface of the silicon wafers for passivation purposes can also be made to serve as a mask to control the dopant diffusion process. Hence the diffusion may be restricted to those areas on the wafer intended to become the active region of the individual devices, simply by etching away corresponding areas of the passivation layer using a **photolithographic process.**

The way in which the techniques of photolithography and selective diffusion are combined to convert the silicon wafers into arrays of completed devices will now be briefly described, with reference to the diagrams of figure 15.6.

The wafer is first given a thin coating of photoresist, as shown in diagram (a). The photoresist is a photosensitive material which, when exposed to ultraviolet light, becomes capable of resisting the etchant used for dissolving the "windows" in the silicon dioxide layer. The photoresist is applied as a drop of liquid to the wafer, which is then rotated rapidly in the horizontal plane to ensure even coating.

After drying, a contact printing plate is rigidly clamped to the sensitised surface of the wafer, and the assembly exposed to ultraviolet light. The emulsion of the printing plate has opaque

The diffusion process itself, illustrated in (e), is performed in a tubular electric furnace, at a temperature between 900 and 1300°C. The wafers are introduced into the furnace in a quartz "boat" crucible, and, after the temperature has stabilised, a carefully controlled mixture of dopant and inert "dilutant" vapours is passed through for a predetermined period.

Typical active vapours used for donor diffusion are phosphorous pentoxide and ammonium phosphate, while acceptor diffusion is usually performed using boron trichloride vapour. The inert dilutant vapour is either nitrogen or helium. Figure 15.7 shows silicon wafers being loaded into a diffusion furnace.

Often the diffusion process consists of two distinct phases. In the first and shorter phase, known as **pre-deposition,** dopant material is deposited through the silicon dioxide "windows" on to the surface of the wafer, as a thin solid film. This typically takes about 30 minutes. Then in the second phase, known as **baking,** the wafers are maintained at a constant high temperature while the dopant atoms diffuse into the silicon. This may take from 3 to 20 hours.

The depth and concentration of the diffusion is readily controllable by manipulation of conditions during the two phases. Thus a shallow but highly doped diffusion region is produced in the wafer if a high dopant vapour concentration is used in the first phase, to produce a relatively thick predeposition film, and then baking for relatively short time. Conversely a deep but lightly doped region may be produced by using a relatively dilute vapour concentration in the predeposition phase, and baking the wafers for a relatively long period.

As may be seen in figure 15.6(e), the semiconductor region formed beneath each "window" in the diffusion mask actually extends beyond the edges of the "window" itself. This occurs because the concentration gradient "seen" by dopant atoms upon entering the crystal extends both laterally and vertically, and thus causes diffusion to occur in both directions. To allow for this effect it is necessary to arrange that the master pattern and the opaque areas in the contact printing plate are somewhat smaller than the final area required for the diffused regions.

The final step in the selective diffusion process is repassivation, shown in figure 15.6(f). Here a new layer of silicon dioxide is grown on the wafers, both to cover and protect the wafer surface areas exposed during the diffusion, and to prepare the wafers for any subsequent diffusions. The repassivation is often performed in the diffusion furnace, during the last part of the diffusion baking phase.

The sequence of operations just described and illustrated in figure 15.6 may be performed a number of times during the fabrication of a semiconductor device, depending upon the configuration of functional regions required for each individual device. Thus it is common to speak of devices as having a "single diffused" structure, a "double diffused" structure, a "triple diffused" structure, and so on.

A specific device example may help the reader to visualise how a number of diffusions may be used to fabricate any desired device configuration. The diagrams of figure 15.8 show the relevant stages in the fabrication of a double-diffused NPN bipolar transistor.

In (a) is shown a small cross-section of the initial state at the top of the wafer used to fabricate such a device, together with a graph plotting dopant concentration against distance from the surface. As may be seen the material is lightly doped homogeneous N-type material, having a donor concentration which remains at a constant low value. This corresponds to either the bulk material of a pre-doped non-epitaxial wafer, or the doped epitaxial layer of an epitaxial wafer.

The corresponding situations following the first or "base" diffusion are shown in 15.8 (b). Here a relatively light but prolonged diffusion of acceptor dopant has been made, with a profile represented by the curve drawn in short dashes.

The resultant effective doping profile is represented by the heavy curve. It may be seen that the phenomenon of compensation has caused the region near the surface to be converted into P-type material, and a P-N junction to

be created at a distance D1 from the surface. This junction is that which ultimately becomes the collector-base junction of the completed device.

In the fabrication of this particular type of device the second or "emitter" diffusion operation is a short but relatively heavy one, in which donor dopant is diffused into relatively small areas within each of the "base" areas formed by the first diffusion. It is performed using the same procedures as the first diffusion, and produces the situation shown in figure 15.8 (c).

It may be seen that the diffusion of donor dopant has caused the region nearest the surface to be converted back to heavily doped N-type material, and a second P-N junction has been created at depth D2 — the emitter-base junction. At the same time, the high temperatures present during the second diffusion operation have caused the acceptor dopant from the first diffusion to move slightly further into the material, so that the first junction has moved to depth D1'. The area of the first diffusion has also increased slightly, for the same reason.

The combined effect of the two successive diffusion operations thus produces the NPN configuration required for the devices concerned, with a heavily doped N-type emitter region, a relatively short and lightly doped base region, and a lightly doped collector region separated from the base by a relatively large-area junction capable of appreciable power dissipation. Other types of device are fabricated in a similar fashion.

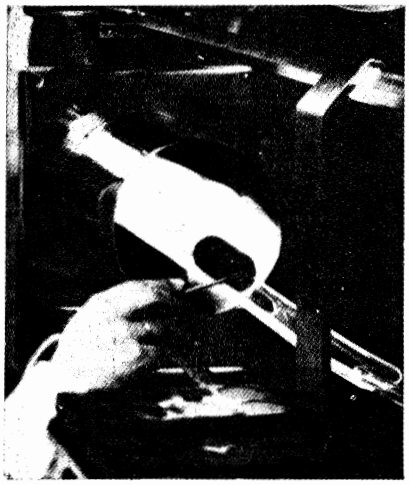


Figure 15.7: Silicon wafers being loaded into the quartz crucible of a diffusion furnace. (Courtesy Philips Industries Ltd.)

The last stage in the on-wafer phase of modern semiconductor device fabrication is **contact metallisation.** In this process each of the individual devices which have been formed on the wafer is provided with a set of ohmic contacts to those of its functional regions accessible from the top.

The sequence of operations involved in contact metallisation is as follows: Windows are first etched in the silicon dioxide passivation in the positions at which contacts are required, using the same photolithographic process used previously. Then a thin film of aluminium is deposited over the entire top surface of the wafer. This is achieved by placing the wafers in a vacuum vessel in which aluminium pellets are vaporised. Finally the excess aluminium is photo-etched away to leave the desired contact pads.

At the same time that the metallisation windows are etched in the passivation layer on the top of the wafer, the complete passivation layer on the lower surface is also etched away. This is done both to facilitate the next process of on-wafer testing, and also to prepare the devices for bonding and encapsulation.

It may be noted that for virtually the whole of the device fabrication sequence described, the devices on the silicon wafer are protected from contamination by the silicon dioxide passivation layer. The only areas not continuously protected in this way are those at which windows are etched for selective diffusion, and these areas are easily protected against contamination by impurities other than the desired dopants. Hence the devices fabricated using the foregoing procedures tend to exhibit very stable and consistent performance, and to be particularly reliable.

The suitability of a silicon dioxide layer as both a passivation layer and as a mask for selective diffusion was discovered in 1960 by Jean Hoerni, then chief physicist at Fairchild Semiconductor. As a result the use of a silicon dioxide layer for these purposes has been patented by Fairchild, and is called by them the Planar process. Devices which are fabricated using the foregoing techniques are thus often called "planar devices."

In passing it may perhaps be noted that the Planar process as described is not suitable for fabrication of germanium devices, for the reason that although it is relatively easy to grow an oxide layer on germanium, such a layer proves to be virtually "transparent" to impurities. It is thus incapable of performing the functions of passivation and diffusion masking.

Techniques have been developed in recent years to produce planar-type germanium devices, but these tend to be somewhat more complex than those required for silicon devices. This provides a partial explanation for the current popularity of silicon devices.

At the stage in the silicon device fabrication sequence just described, the individual devices formed on the original silicon wafer are still attached physically to one another. Typically, as many as 12,000 discrete devices may thus make up the "array" into which the wafer has been effectively converted.

The remainder of the fabrication sequence involves on-wafer testing of the devices, scribing and separation of the devices into individual chips or "dice," bonding of the dice to the package headers, bonding of connecting wires to the contact pads, and final encapsulation. These processes will now be briefly described, with reference to the photographs of figure 15.9.

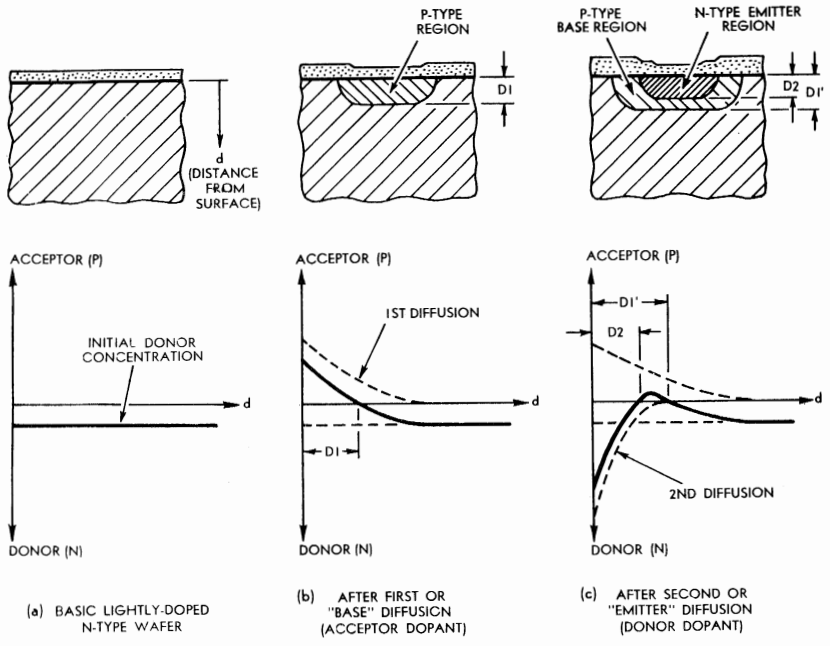


Figure 15.8

A small portion of a completed device array is shown in (a), prior to the commencement of further operations. The devices shown here are high frequency NPN bipolar transistors, each measuring approximately 25 mils square.

The first operation performed on the array is the testing of the devices, illustrated in (b). Here the wafer is mounted on a conducting table, which forms a master contact to the collector regions, and two micro-probe electrodes are applied to the contact pads of each device to check its operation. In modern manufacturing facilities this probe testing operation is done entirely automatically after initial set-up, under computer control.

During the testing, a drop of marking ink is used to identify any devices which prove to be unsatisfactory at this stage. This is shown in (c). The wafer is then precision scribed between the devices, as in (d), and broken up into individual dice as shown in (e). This operation is rather similar to that used in glass cutting. After division the marked reject dice are discarded.

The remaining dice are picked up individually by a vacuum chuck, as in (f), and bonded to the base or "header"

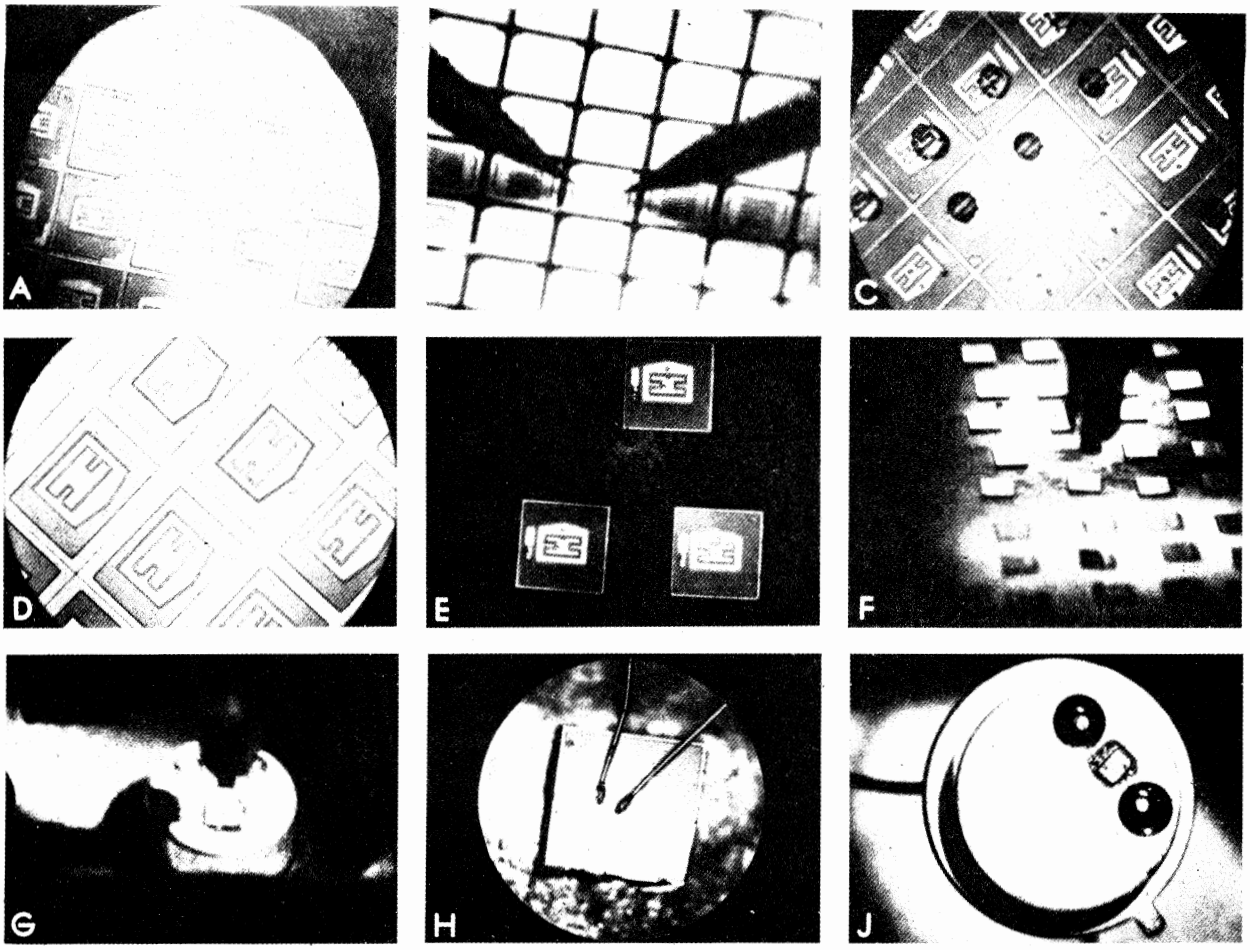


Figure 15.9: From wafer to assembled device. A—completed devices on wafer; B—probe testing; C—wafer showing inked rejects; D—wafer after scribing; E—separated dice; F—picking up die for bonding; G—bonding of die to header; H—connecting wires (.001in) bonded to metallised pads on die; J—completed TO-5 metal can device prior to cap sealing. (Courtesy STC Components Division.)

of the package in which they are to be encapsulated. With epoxy resin encapsulation the die is cemented to the header using an epoxy adhesive, whereas with metal encapsulation the die is generally bonded to the header by a gold soldering process, at about 400°C. The latter is illustrated in (g).

Connections between the contact pads on the die and the insulated terminal posts on the header are then made, this operation being known as wire bonding. The connections are performed using very fine wire, typically between 1 mil and 5 mils diameter. A variety of bonding methods have been used, but that currently favoured is ultrasonic welding using aluminium wire. Bonds of this type are illustrated in figure 15.9(h), and a completed header and die assembly is shown in (j).

Finally the completed device is sealed in its package, which protects it from both physical damage and the ingress of moisture. With devices in epoxy resin encapsulation, the final sealing is performed by covering the top of the header with a blob of epoxy. In metal package devices, a can or "cap" is welded to the header, the operation being performed in an inert atmosphere of ultra dry nitrogen.

The fabrication process is now finished, in that the devices are completed and encapsulated. However, before being marketed they are generally subjected to a series of quality control and reliability tests, to ensure that they meet published figures for both electrical and mechanical performance. The tests applied include impact resistance and hermeticity tests, electrical aging, and tests of such parameters as saturation voltage, breakdown voltages and current gain.

The fact that the fabrication of semiconductor devices involves a large number of extremely demanding, ultra-precise processes means that a large number of variables influence the behaviour of the final devices. Hence it is understandably difficult to fabricate devices having an accurately predictable and tightly controlled performance, although progress is continuously being made in this direction.

At present, however, most manufacturers operate using a system of post-fabrication classification. No attempt is made to fabricate a particular device type, but rather a group or "family" of related devices based on the same die size and configuration. Electrical testing after fabrication is then used to sort individual devices into the various device types of the "family."

In most modern production facilities this classification is performed by automatic equipment, under computer control. In addition to device classification, the same equipment is used to compile information on production yields, parameter distributions and fault analysis.

Because of space limitations the description of device fabrication given in this chapter has been rather brief. However, it is hoped that the material presented has given the reader a reasonably satisfying insight into the techniques capable of producing the devices which have been described in the preceding chapters.

SUGGESTED FURTHER READING

MYLES, D. D., "Silicon Planar Transistors," in **Electronics Australia,** V.29, No. 4, July 1967.

PHILLIPS, A. B., **Transistor Engineering,** 1962. McGraw-Hill Book Company, New York.

SITTIG, M., **Doping and Semiconductor Junction Formation,** 1970. Noyes Data Corporation, Park Ridge, N.J.

SITTIG, M., **Producing Films of Electronic Materials,** 1970. Noyes Data Corporation, Park Ridge, N.J.

STERN, L., **Fundamentals of Integrated Circuits,** 1968. Hayden Book Company, Inc., New York.

Chapter 16

MICROCIRCUITS or "IC's"

Microcircuits and their development — monolithic devices — general construction — transistor elements — diodes — resistors and capacitors — representative devices — complex devices — advantages and disadvantages of the monolithic device — thin film devices and their fabrication — active and passive thin-film elements — hybrid devices.

As a result of the very rapid progress made in semiconductor device fabrication, following development of the techniques of gaseous diffusion and epitaxial deposition in the mid-1950s, it very soon became possible to fabricate single or "discrete" functional device chips whose physical size was much smaller than the packages in which they were encapsulated. This despite the fact that device packages had already been reduced to a size approaching the limit for convenient handling, a size significantly smaller than existing thermionic valves.

The drive toward miniaturisation of electronic circuitry had already begun when this stage was reached, and perceptive device and circuit designers were quick to realise that semiconductor devices and fabrication techniques were going to contribute far more toward miniaturisation than had at first been realised. If single transistor and other semiconductor elements could be made much smaller than the smallest convenient packages, then presumably it was going to be possible to fit a number of devices—even perhaps a complex device array—into a single package, together with most of their interconnections.

Thus was born the concept of the microcircuit, or miniature integrated circuit (IC), consisting of a complete functional circuit assembly of active and passive elements, encapsulated together with their interconnections in a single package.

So great was the pressure for electronics miniaturisation, particularly from the avionics industry and the military in the U.S.A., that as soon as the possibility of semiconductor microcircuits was appreciated, work toward its practical realisation was given high priority at many device manufacturing laboratories and research institutions. By late 1958, prototype microcircuits had been developed by both Texas Instruments Inc., in Dallas, and Westinghouse Electric, in Youngwood, Pennsylvania.

The prototype microcircuit devices were rather crude, consisting basically of a number of separate semiconductor dice or "chips" carrying transistors, diodes and resistors, all mounted on a common header. Interconnections were made using fine wires bonded to the chip connection pads. The devices were time-consuming to assemble, and thus relatively costly. However, they demonstrated that microcircuits had become a practical reality.

The next main development in the microcircuit saga came in 1960, with the advent of the planar process and its technique of utilising the silicon dioxide passivation layer for selective diffusion masking. Using this technique, described in the preceding chapter, it almost immediately became possible to fabricate all of the components of a microcircuit assembly on a single chip, with interconnections formed by the final metallisation pattern.

The planar process thus provided a means whereby microcircuits could be fabricated as complete devices, using the same on-wafer "mass production" techniques used for discrete devices. And, being fabricated in this way, the devices had the same basic reliability as single discrete planar devices. Hence these **monolithic microcircuits** (from the Greek words "monos," meaning single, and "lithos," meaning stone) had two marked advantages: greatly improved reliability over equivalent discrete circuitry, and the ability to be produced at a cost only slightly above that of a single discrete semiconductor device.

The twin advantages of monolithic devices brought about an expansion of the whole concept of microcircuits and their development. No longer were these devices simply a part of the somewhat esoteric drive toward miniaturisation initiated by the military and the aero-space industry; rather they were seen to represent an important technological advance, whereby electronic circuits of almost every type could be given improved reliability, yet produced at significantly lower cost.

Accordingly, in the ten years since the first monolithic microcircuits were produced, tremendous effort has been directed towards microcircuit device development. This development has taken place not only in the field of monolithic devices, but also in connection with other related types of microcircuit device which have since been produced. The result is that electronics has undergone, and is still undergoing, a "microcircuit revolution" whose effects may prove further reaching than any previous developments in the history of the art.

Microcircuits of a variety of types have been developed to perform either switching (digital) or linear (analog) functions, and sometimes both together. As techniques have improved, increasingly complex devices have been made, capable of performing either single complex functions or multiple simple functions. Along with these developments there has occurred both a widening in microcircuit applications and a marked reduction in device cost.

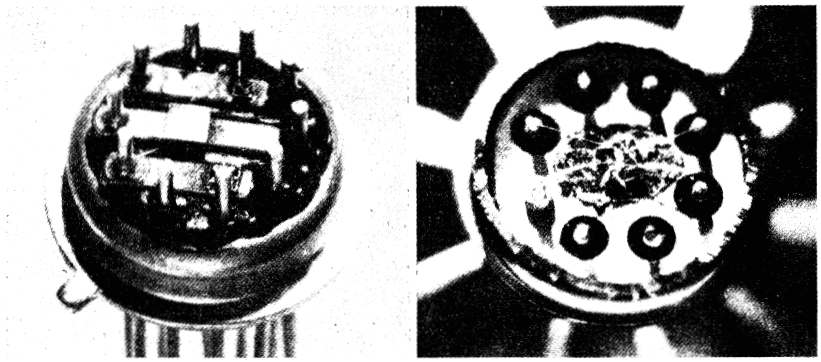


Figure 16.1: Representative microcircuits, with cans removed. At left is a relatively early multi-chip type, while at right is a more recent monolithic device. (Courtesy Mullard-Australia, Fairchild Australia.)

It is true that in terms of function, most microcircuits involve little that is new. Most devices consist essentially of a group of active circuit elements, similar to those which we have discussed in earlier chapters, interconnected in a fairly conventional manner with passive elements which contribute resistance, capacitance and inductance. Figure 16.1 shows two representative microcircuit devices, one an early multiple-chip type and the other a more recent monolithic type.

However, although they are functionally very similar to discrete circuits, microcircuits do differ from these more familiar circuits in one quite obvious respect: they are markedly smaller.

And this physical scaling-down involves significant differences both in terms of construction and in the corresponding fabrication techniques. It is therefore the construction and fabrication of microcircuits which will be discussed primarily in this chapter.

It is perhaps fitting that we should look first at monolithic devices, as it was this type of device which virtually triggered off the microcircuit "revolution." Also it is this variety of microcircuit which has been fabricated in the largest numbers to date, and which by virtue of its continuously improving performance/cost ratio has made the greatest impact, particularly in consumer product and industrial applications.

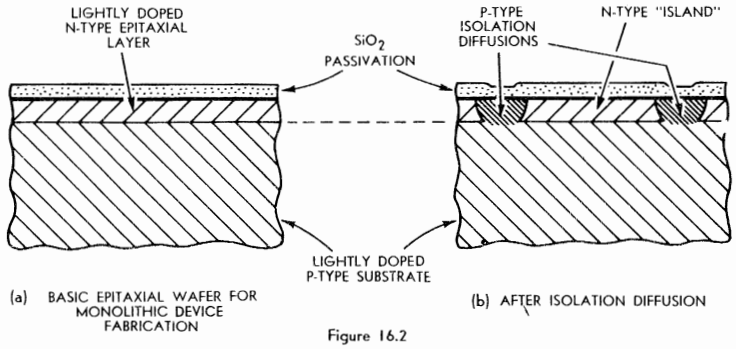


Figure 16.2

As already noted, monolithic or "single chip" devices consist essentially of complete circuit modules fabricated on or within single semiconductor chips. In the most basic type of monolithic device all of the circuit elements, both active and passive, are fabricated within the chip, with interconnections performed by a suitable pattern of conductors formed in the surface metallisation.

Because the desired interconnections between the various circuit elements formed within the monolithic chip are provided by the surface metallisation pattern, each element must be effectively isolated within the chip itself. Hence in general none of the elements may use the bulk or "substrate" of the chip as an effective electrode to one of its functional regions. This is in contrast with the situation which applies with discrete semiconductor devices, where, as we have seen, the substrate usually forms the means of connection to the lowest or "innermost" functional region.

The usual method of achieving element isolation is to fabricate monolithic devices from epitaxial wafers in which the lightly doped epitaxial layer is of **opposite type** to the pre-doped wafer material. An initial diffusion step is then used to effectively form "islands" in the epitaxial layer, and each of the individual circuit elements is subsequently fabricated within a separate island and isolated from the substrate by a P-N junction. By application of a suitable bias to the substrate of the completed device, all of the island-substrate P-N junctions are reverse biased to achieve element isolation.

The method is illustrated in the diagrams of figure 16.2. In (a) is shown the basic epitaxial wafer, with a lightly doped N-type layer deposited on a lightly doped P-type substrate. It may be noted that the wafer differs from those used to fabricate "epitaxial" discrete devices in two respects: the deposited layer is here of opposite type to the pre-doped substrate, rather than of the same type, and also the substrate is here only lightly doped rather than heavily doped. The reason for the light doping is to ensure that the P-N junctions which ultimately isolate the epitaxial islands from the substrate possess wide, low capacitance depletion layers.

The effect of the "isolation diffusion" is shown in figure 16.2 (b). Here it may be seen that a diffusion of acceptor dopant has formed narrow but heavily doped regions of P-type material which completely penetrate the N-type epitaxial layer, forming the latter into "islands" completely separated by P-type material connected to the substrate. It is within these islands that the various elements of the monolithic circuit are formed by subsequent selective diffusion steps.

Broadly speaking, all functional regions of the various circuit elements of a monolithic device, both active and passive, are fabricated simultaneously, using the same selective diffusion steps. The number of fabrication steps involved in producing a monolithic device is thereby kept to a minimum, which is in practice equal to or only slightly greater than the number of steps required to produce a discrete planar device. This explains why monolithic microcircuits are economically attractive.

Of the active devices which may be used in monolithic circuits, **bipolar transistors** are those most commonly used at present, although JFETs and MOSFETs have also been incorporated in recent devices. The construction of typical bipolar elements for monolithic circuits is shown in figure 16.3, where it may be seen that they differ slightly from the discrete devices described in the preceding chapter.

The basic structure of a monolithic circuit bipolar transistor is shown in (a). It may be seen that the main difference between the structure and an equivalent discrete device is that here all three functional regions are brought out to top surface contacts. Whereas in a discrete device the connection to the collector region is made via the substrate, in this case the isolation between the collector "island" and the substrate necessitates a third top contact.

While it satisfies the basic requirements for a monolithic bipolar element, however, the structure of 16.3 (a) has a potential weakness which arises directly from the substitution of a top collector contact for the underside contact of a discrete device. Carriers crossing the base-collector junction must move transversely through the collector region to reach the top contact, and since the collector region is only lightly doped in order to produce a relatively wide, high breakdown voltage collector depletion layer, it has appreciable bulk resistance. Hence the structure tends to possess a high collector series resistance, and accordingly a high saturation voltage Vce(sat).

It is possible to obtain a moderately low collector resistance within the basic structure by using a relatively large and distributed collector contact, and this is in fact done in relatively undemanding monolithic devices. However, where very low collector series resistance is essential, the structure is modified by addition of a highly doped low resistivity "buried layer" region at the lowest extremity of the collector region. The buried layer acts as an equipotential plane, providing a low re-

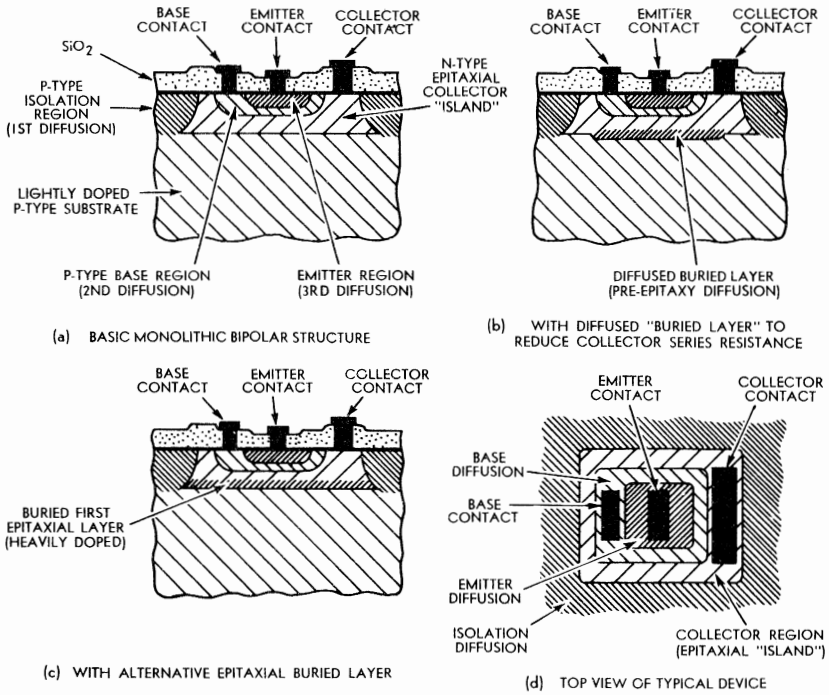


Figure 16.3

sistance path to the periphery of the collector island.

Two alternative methods are used to produce the buried layer, as illustrated in figure 16.3(b) and (c). One method is to selectively diffuse suitable areas on the semiconductor wafers prior to epitaxial deposition of the collector layer, producing the structure shown in (b). While in some ways this procedure is most satisfactory, it introduces a further source of registration errors, and also means that wafers become rigidly identified with a specific microcircuit even before epitaxy.

Because of this the alternative technique is often used, in which the wafers are prepared by a double epitaxy process wherein they are provided first with a very thin heavily doped layer, and then with the usual lightly doped "collector" layer. Both layers are of the same type, and opposite to that of the wafer itself. This type of water produces the bipolar structure shown in figure 16.3(c), the required buried layer being formed from the thin initial epitaxial layer.

From the surface all three versions of the monolithic bipolar structure look rather similar, appearing as shown in figure 16.3(d). Note that although the electrode contacts are shown in the diagram as simple rectangles, these normally form part of the metallisation interconnection pattern of the device.

Next to bipolar transistors, probably the element most often used in monolithic circuits is the **P-N junction diode.** In general, one of the two functional regions of this type of element is fabricated during the bipolar transistor "base" diffusion, while the second region is formed by either the "collector" epitaxial island provided for the diode element, or by a further region fabricated during the "emitter" diffusion. Thus diode elements are basically equivalent to either the collector-base or base-emitter portion of a bipolar transistor element, as illustrated in the diagrams of figure 16.4.

Because the depth and doping levels of the various regions available for fabrication of each element of a monolithic circuit are common to all elements of the device, and may not be varied individually, the selection between the alternative forms available for any particular monolithic diode element plays an important part in determining the characteristics of the element concerned. Hence a diode formed using the "collector" island and the "base" diffusion tends to have a relatively high reverse breakdown voltage, but also a relatively high charge storage and a significant stray capacitance to the substrate. Conversely, a diode formed using the "base" and "emitter" diffusions tends to have low charge storage and low stray capacitance to the substrate, but also a relatively low reverse breakdown voltage.

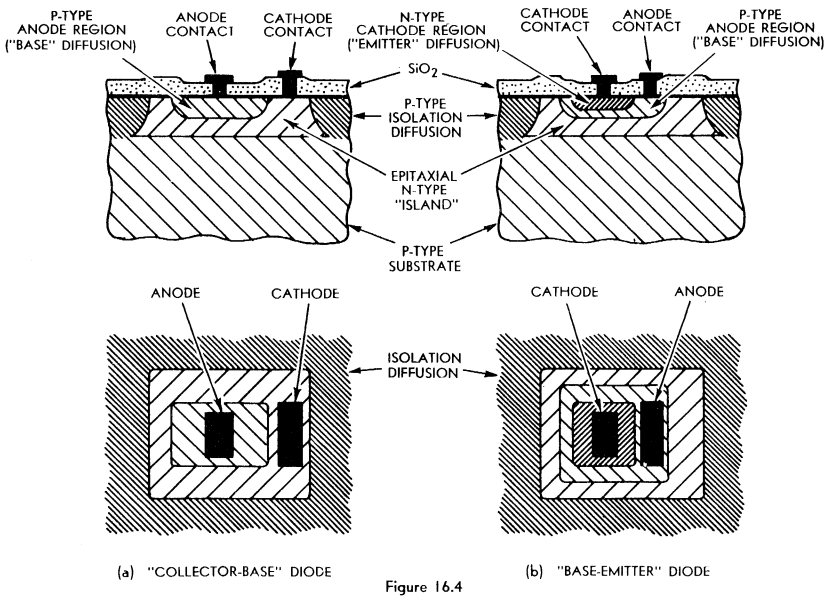


Figure 16.4

Of the passive circuit elements used in monolithic devices, **resistors** are those used in the largest numbers. Generally, monolithic resistor elements consist of a single rectangular diffused region formed within a suitable "collector" island, as shown in figure 16.5 (a). Normally the resistor region itself is formed during the "base" diffusion, as show, but if very low value resistors are required, the more heavily doped "emitter" diffusion is used.

Whichever diffusion is used, the depth and doping level of the element is naturally fixed, so that only the length and width may be varied in order to achieve the required resistance value. Even these parameters may only be varied over moderate limits, due to process limitations and the need to prevent the device chip from becoming excessively large. Where high value resistor elements are required, the element is often folded back beside itself one or more times, to produce a more compact format.

Capacitor elements are found in monolithic circuits, although comparatively rarely. The reason for this is that relatively large chip areas are required to fabricate capacitors of even modest value. Because of this designers of monolithic circuits seek to either obviate the need for most capacitors in their circuit, or arrange for the capacitors to be connected externally.

For those capacitor elements which cannot be avoided two different constructions may be used. The first of these is the "diffused capacitor," which is basically nothing more than a large-

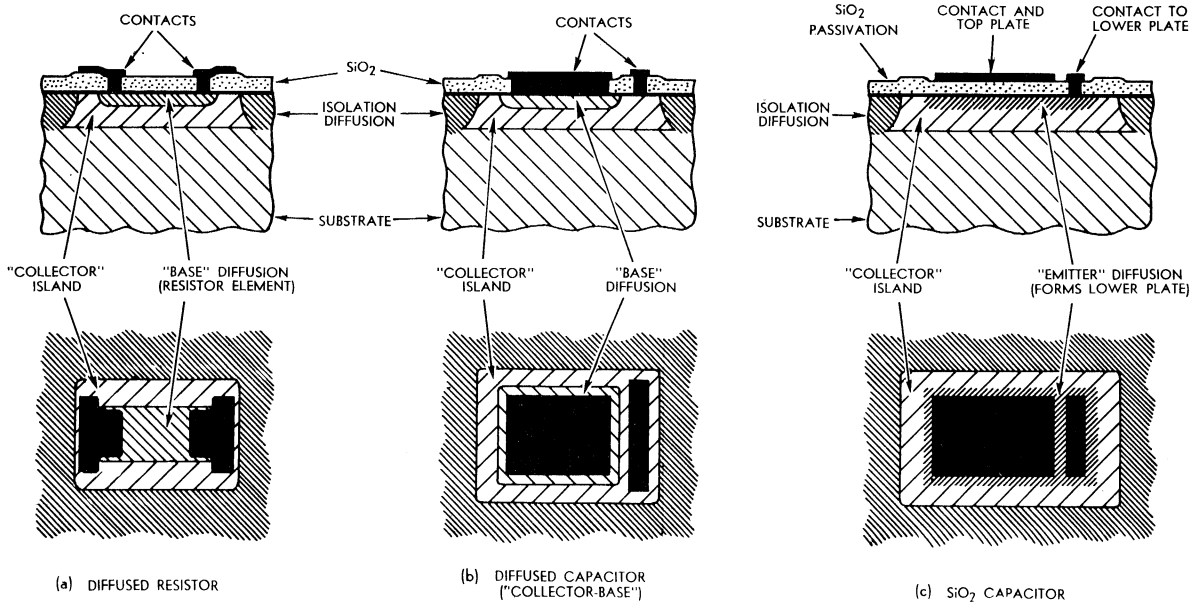


Figure 16.5

area P-N diode which is operated under reverse bias conditions. This type of element may be formed using either the "collector" island and a "base" diffusion, as shown in figure 16.5(b), or the "emitter" and "base" diffusions. Because of the higher doping of the "emitter" diffusion, the latter form possesses a narrower depletion layer and hence a larger capacitance per unit area. However, it also has a lower breakdown voltage, so that the choice between the two varieties of diffused capacitor depends upon the applied voltage as well as capacitance/space considerations.

While useful for non-critical applications such as bypassing, this type of capacitor element has two rather marked shortcomings, which both arise because the element is based on the depletion layer capacitance of a P-N junction. The first of these is that because the junction cannot be forward biased, the capacitor is subject to the same fixed-polarity restrictions which limit the flexibility of a conventional "electrolytic" capacitor. The second shortcoming is that as the width of the junction depletion layer varies with applied voltage, so too does the capacitance, making the capacitor voltage-variable and of non-constant value.

Luckily there is an alternative type of monolithic capacitor element available, which possesses neither of these disadvantages. This is the silicon dioxide capacitor, which uses the silicon dioxide passivation layer as dielectric instead of a junction depletion layer. The usual construction of this type of capacitor is shown in figure 16.5(c), where it may be seen that one plate is formed by a simple rectangle of metallisation on the top of the dielectric, while the other is formed by a low resistivity "emitter" diffusion beneath it.

Because of the silicon dioxide dielectric, this type of monolithic capacitor is non-polarised, and because the capacitance is virtually independent of applied voltage, it is also constant in value. The dielectric strength of the silicon dioxide passivation layer also gives the capacitor a relatively high breakdown voltage, typically in the order of 50V. Furthermore, because the "plates" of the capacitor are composed of metal film and highly doped low resistivity semiconductor respectively, the element possesses a lower effective series resistance and hence a higher "Q" than is possessed by typical diffused capacitor elements.

Small **inductor elements** may be provided on monolithic circuits, if required, being formed not in the semiconductor chip itself but purely as a spiral in the metallisation pattern. However, like capacitor elements, they tend to require relatively large chip areas and are hence avoided wherever possible.

As mentioned previously, the various circuit elements which form a monolithic microcircuit are fabricated simultaneously, using a common sequence of selective diffusion processes. These processes are basically the same as those for discrete planar transistors described in chapter 15, and as with the discrete devices the monolithic devices are fabricated in complete arrays on the semiconductor wafers. The only difference is that each "device" in the wafer array is a group or complex of elements, rather than a single element, and the final metallisation process is used to provide element interconnections in addition to the contacts for external device connections.

The contact printing plates used in the various photolithography steps are produced, as before, by high-ratio photographic reduction from precision master templates. Figure 16.6 shows the preparation of such master templates, and the precision step-and-repeat photographic reduction process used to produce the multiple-image contact printing plates.

After the various diffusion steps and the final metallisation, the completed wafers are subjected to a probe test to determine if any of the devices on the wafer are faulty. This probe test is similar to that performed on discrete device wafers, but because the microcircuits are somewhat more complex than discrete devices, each must be subjected to a larger number of tests to check for correct operation. A modern computer controlled monolithic probe test station is illustrated in figure 16.7. Following the probe test the devices are then scribed, separated and packaged in a very similar fashion to that already described for discrete devices.

Monolithic microcircuit devices range in complexity from simple digital logic gates, through more elaborate linear operational amplifier ("op amp") devices, to very complex "MSI" (medium-scale integration) and "LSI" (large-scale integration) devices involving many thousands of component elements. The largest LSI devices are virtually complete functional systems or sub-systems, containing all of the functional circuitry required for an equipment module such as a data processor or memory store.

Some idea of the range in monolithic device complexity may be conveyed by the photomicrographs of device chips shown in figure 16.8. The relatively simple device chip shown in (a) is accompanied in (b) by the equivalent circuit diagram, which should enable the reader to identify the various circuit elements using the preceding diagrams for guidance.

The simple device concerned is a dual two-input NOR logic gate, using resistor-transistor logic (RTL) circuitry. As may be seen it involves only four bipolar transistor elements and six diffused resistor elements, which are arranged on a chip measuring approximately 25 mils square.

It may be seen from the photomicrograph that the two transistor elements of each gate are fabricated within a single collector island, which is in each case rectangular and surrounded by diffusion isolation. The six diffused resistors of the circuit are all fabricated from "base" diffusion regions, and all share a common "collector" island used for isolation. This island may be seen to connect to the "+" supply metallisation near the contact pad in the top left-hand corner, this being done to en-

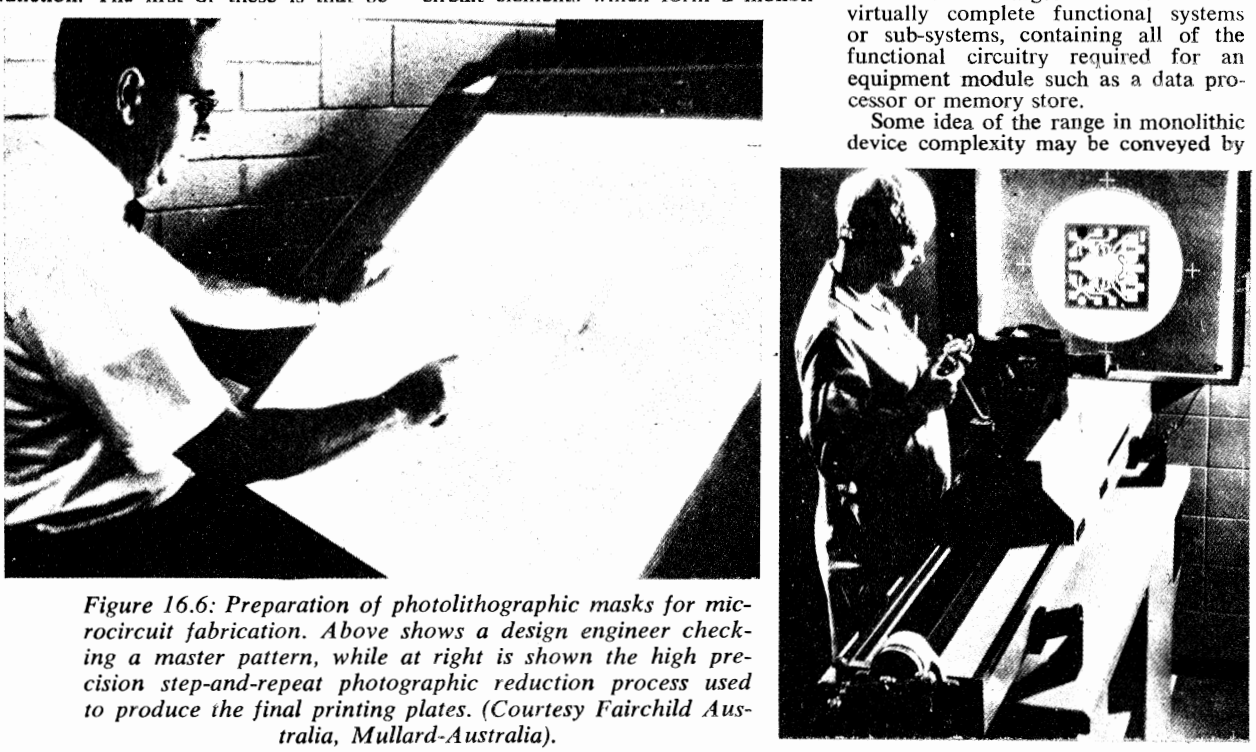


Figure 16.6: Preparation of photolithographic masks for microcircuit fabrication. Above shows a design engineer checking a master pattern, while at right is shown the high precision step-and-repeat photographic reduction process used to produce the final printing plates. (Courtesy Fairchild Australia, Mullard-Australia).

sure that the junction between each P-type resistor element and the N-type island is always reverse biased.

Similarly, to ensure that the junctions between each N-type "collector" island and the P-type isolation diffusion and substrate regions are also always reverse biased, the latter regions are connected to the "earth" supply metallisation. This connection is visible adjacent to the contact pad in the lower right-hand corner.

It may be noted that the conductors of the metallisation pattern pass over many of the circuit elements, this being possible because of the interposed silicon dioxide passivation. The reason why the outline of the elements appears to be visible through the metallisation is that the outlines are not really those of the elements themselves, but in fact correspond to the edges of the shallow "troughs" left in the silicon dioxide passivation layer after the various diffusion processes. The continuity of the outlines through the metallisation is due to the fact that the thin metallisation layer follows the surface contours of the passivation layer.

Figure 16.7: A modern computer-controlled microcircuit probe test station, capable of applying a large number of tests in rapid succession to each on-wafer device. (Courtesy Philips Industries Ltd.)

In rather dramatic contrast with the simple device of figure 16.8(a) and (b) is the device shown in the photomicrograph of (c). This is a complex MSI device, a dual 4-bit digital comparator based on diode-transistor logic (DTL) circuitry, and comprising many hundreds of circuit elements on a chip measuring only 80 x 110 mils.

The device is manufactured by Fairchild Semiconductor, and is representative of their "Micromatrix" range of MSI devices. The devices in this range are fabricated from a relatively small number of different "stock" chip designs, each of which contains a large array of component elements. A variety of different complex-function MSI devices are fabricated from each type of "stock" chip, merely by using different metallisation patterns to interconnect the array elements. In many of the devices, including that pictured, the metallisation is in two layers, a second passivation layer and metallisation pattern being deposited on the wafer after the normal metallisation, to permit a more complex interconnection system.

Techniques similar to the "Micromatrix" approach have been adopted by many manufacturers for MSI and LSI device fabrication, in an effort to maintain some flexibility in the fabrication of these devices.

There is a natural tendency for microcircuits to become increasingly specialised in application as they are made more complex, simply as a consequence of the increased internal circuitry. Because of this tendency, the potential applications of most devices become narrower as the devices themselves become more complex. Yet, at the same time, the development cost of devices tends to rise with complexity, making it necessary to manufacture increasingly larger numbers of devices if the individual device cost is to be maintained at an attractive level.

By using a technique such as "Micromatrix," the device manufacturer is able to produce a number of different devices from a single type of complex chip, thereby increasing the potential chip applications and the manufacturing volume. At the same time the effective development costs for each device type are reduced, as each chip is "shared" by a number of devices.

A further advantage of this type of approach is that it becomes possible to produce "custom" complex devices at short notice and at a relatively low cost, even for small quantities. Providing the desired "custom" devices can be fabricated using "stock" array chips, they may be produced at a cost little more than that associated with design and deposition of the required metallisation.

Recently this approach has been carried one step further, with the use of computer-controlled drafting techniques to produce automatically the precision custom metallisation mask contact printing plates, direct from the customer's device specifications. The computer system is also fed information from the probe tests of each wafer of "stock" arrays, and uses this information to produce individually tailored "custom" metallisation masks for each wafer. This **discretionary wiring** feature lowers costs significantly, by ensuring that each wafer produces the maximum yield of good devices.

The monolithic type of microcircuit, which we have been discussing in the foregoing, has two marked advantages, both of which arise from the fact that it is fabricated using virtually the same processes used for discrete planar transistors: it is **relatively inexpensive,** and it is **highly reliable.** These advantages have been responsible for the wide acceptance of this type of microcircuit in such fields as digital computing and consumer appliances.

It is true, however, that monolithic devices possess a number of disadvantages which weigh heavily against their use in certain other applications.

Not the least of these disadvantages is that the dependence upon a single set of diffusions to simultaneously produce the regions of the various diffused circuit elements places rather severe design constraints on these elements. Thus, in some cases, optimum component value or performance is not readily available for certain elements, simply because the diffusions which these elements must share with the other circuit elements do not provide suitable region depths or doping levels.

A disadvantage which is often more embarrassing than this, however, is that virtually all elements of a monolithic circuit which incorporate "within chip" diffusion or epitaxial regions are inevitably accompanied by superficially hidden **parasitic elements,** which are formed by interaction between the regions of the elements themselves, the surrounding isolation "islands," and the base substrate.

As noted earlier, each element of a monolithic circuit possesses significant parasitic or "stray" capacitance to the substrate, by virtue of the depletion capacitance of the reverse biased P-N junctions used for isolation. However, in addition to this stray capacitance, each element structure tends to be coupled effectively to the substrate via a parasitic bipolar transistor structure.

Thus in the case of the basic monolithic NPN transistor shown in figure 16.3(a), the intended NPN transistor element is accompanied by a parasitic PNP transistor, formed by the "base" diffusion, the "collector" epitaxial island, and the P-type substrate. Similar parasitic transistor structures are present for the more elaborate structures of figure 16.3(b) and (c), for the diode structures of figures 16.4, and for the diffused resistor and capacitor structures of figure 16.5(a) and (b).

Although the doping levels and geometry of the regions forming these parasitic transistor structures are such that the parasitic elements have very low gain, the effect in each case is to increase the leakage current to the substrate, and hence reduce the efficiency of the intended monolithic circuit component. Even with extremely careful design it is not possible to completely obviate this effect, which can seriously limit the performance of a monolithic device, particularly at high voltage and low current levels, and in low noise applications.

A further disadvantage of monolithic circuits at present is that fabrication process limitations results in fairly wide **spread variation** in active device parameters and passive component values. Thus the gain of bipolar transistor elements tends to vary over a range of 5:1 or greater, while it is difficult to produce diffused resistors having a consistent absolute value tolerance of better than plus/minus 20%; although the process variations tend to cause resistors on the same chip to vary together, giving a relative value tolerance approaching plus/minus 5%.

Because of these rather pronounced parameter and component value spreads, monolithic devices are basically somewhat better suited for switching and "digital" applications than for analog or "linear" applications. It is true that many quite satisfactory linear devices have been produced, and improved devices are being continuously developed; however it is also true that these tend to be somewhat harder to design than digital devices, and generally subject to lower fabrication yields. In many cases linear devices are provided with active or passive circuit elements having special structures designed so that parameter and value changes tend to be mutually compensating.

Yet another disadvantage of monolithic circuits is that most of the circuit elements, being formed largely from semiconductor material, exhibit a significant **temperature coefficient.** This is most pronounced when the element is formed primarily from lightly doped material, such as a "collector-base" diode or diffused capacitor, or a diffused resistor formed during the "base" diffusion, because of the increased significance of "intrinsic" carrier generation in lightly doped material. However the temperature coefficient of elements formed largely from heavily doped material is still significant, and in many linear device applications can prove quite embarrassing.

In most commercial and industrial applications, as noted earlier, the cost and reliability advantages of monolithic circuits far outweigh their disadvantages. However in critical and demanding applications the reverse tends to be the case, and it is because of this that techniques have been developed to produce other types of microcircuit device.

Broadly speaking, most of the other types of microcircuit which have been produced rely upon the techniques of **thin film technology,** which techniques permit the controlled and selective deposition on substrate materials of thin films of metals, insulators and semiconductor materials. In this context, the term "thin film" is used to denote films of thickness less than 1uM (micron), to distinguish this type of film from those of greater thickness which are also used in electronic device fabrication. The latter type of film are sometimes referred to as "thick films."

Thin film techniques are used either in place of the techniques used for monolithic device fabrication, or, alternatively, in conjunction with these techniques. Hence they are used to produce either complete thin-film circuits containing both passive and active components, or only partial circuits containing most of the passive components and interconnections, to which are added semiconductor chips containing the active components and remaining passive components.

Naturally enough, microcircuits which are entirely fabricated from thin films are termed **thin-film devices,** while those in which thin-film techniques are only used to fabricate passive components and interconnections to be used with semiconductor chips are known as **hybrid devices.**

In general, the thin metallic films used in fabricating thin-film components and wiring are deposited using techniques similar to the vacuum deposition process used for contact metallisation of planar discrete and monolithic devices. The insulating and semiconductor films are deposited using techniques such as reacting vapour reduction, as used for the deposition of epitaxial silicon layers on planar wafers.

The main types of component element used in thin-film devices are illustrated in basic form in figure 16.9. In (a) is shown a resistor element, in (b) a capacitor element, and in (c) a MOSFET element, this being to date the only type of active thin-film element to have been used in production devices.

As may be seen the **thin-film resistor** element consists basically of a thin stripe of resistive material film deposited between two spaced metallic film electrodes, the whole being supported on an insulating substrate of ceramic, glass or sapphire. The resistive material used for the element itself is typically either tantalum nitride or nichrome (nickel-chromium alloy), although other materials such as chromium, tantalum oxide, titanium, tin oxide, cermet and carbon resin inks have also been used. The thickness of the resistance element film is typically about 0.12uM, or 1,200 Angstroms.

Incidentally, in dealing with thin films of materials it is convenient to specify their electrical behaviour not in terms of resistivity, but in terms of the so-called **sheet resistance.** This term is used to denote the resistance between opposite edges of a square of material in film or sheet form, of specified thickness. Being defined in terms of a square of material, the sheet resistance is independent of the size of the square involved. It is normally measured in units of ohms/square.

The sheet resistance of the films used for thin-film resistive elements is typically about 60 ohms/square, which with suitable element geometry variations allows the production of resistors with values between a few ohms and about 100K. Values outside this range may be obtained using films of higher or lower sheet resistance.

Using a construction similar to that shown, in order to minimise the effect of registration errors, it is possible to produce thin-film resistors with absolute values falling within a tolerance range considerably narrower than with diffused monolithic resistors. The relative value tolerance tends to be considerably narrower also, both improvements being due to the fact that resistor value is here determined almost solely by the composition and thickness of the element deposition film. It is also possible to produce precision thin-film resistors, by "trimming" completed elements to exact value using electrolytic etching or laser-beam vaporisation techniques.

Quite apart from the closer tolerances possible with thin-film resistors, these components possess further advantages compared with their diffused

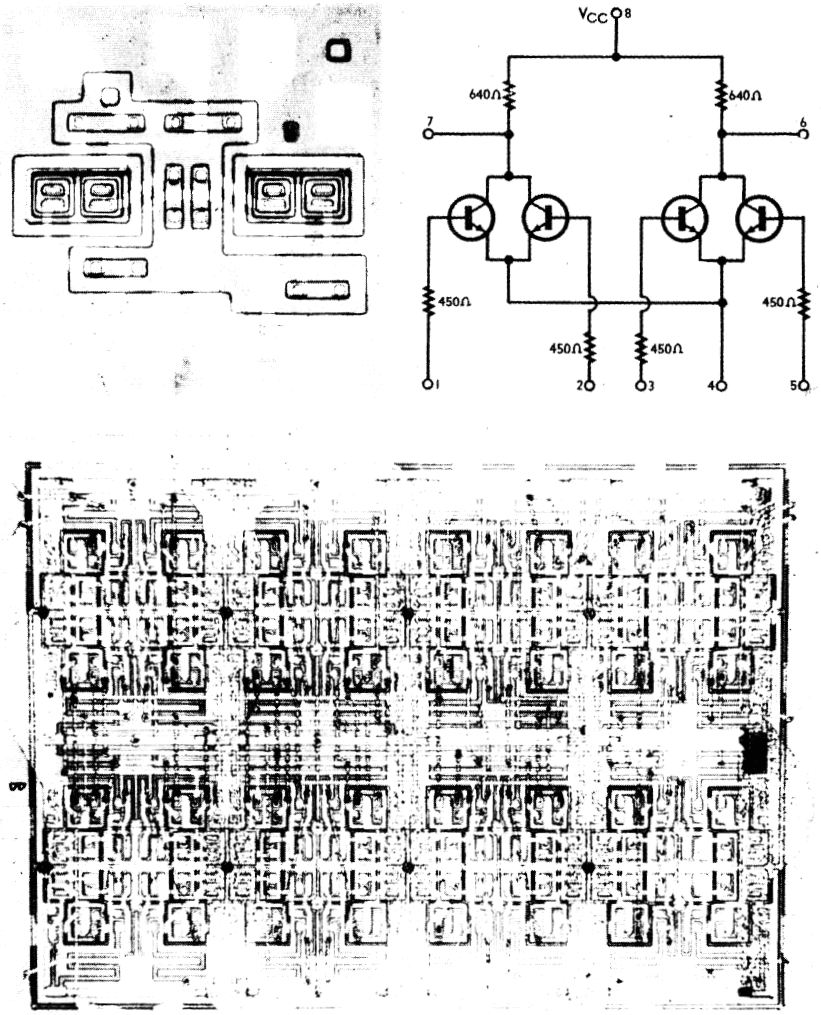


Figure 16.8: Simple and complex monolithic device chips. In (a) at top left is a simple RTL logic gate, reproduced by permission of Motorola Semiconductors, Phoenix, Arizona, and accompanied by its equivalent circuit alongside in (b), top right. Compare this device with the complex MSI array in (c), immediately above, reproduced by courtesy of Fairchild Australia.

monolithic counterparts. They tend to be virtually free from parasitics, being deposited on an insulating substrate rather than formed within a semiconductor crystal. And, provided a suitable material is chosen for the resistive element film, they can be fabricated with an extremely low temperature coefficient.

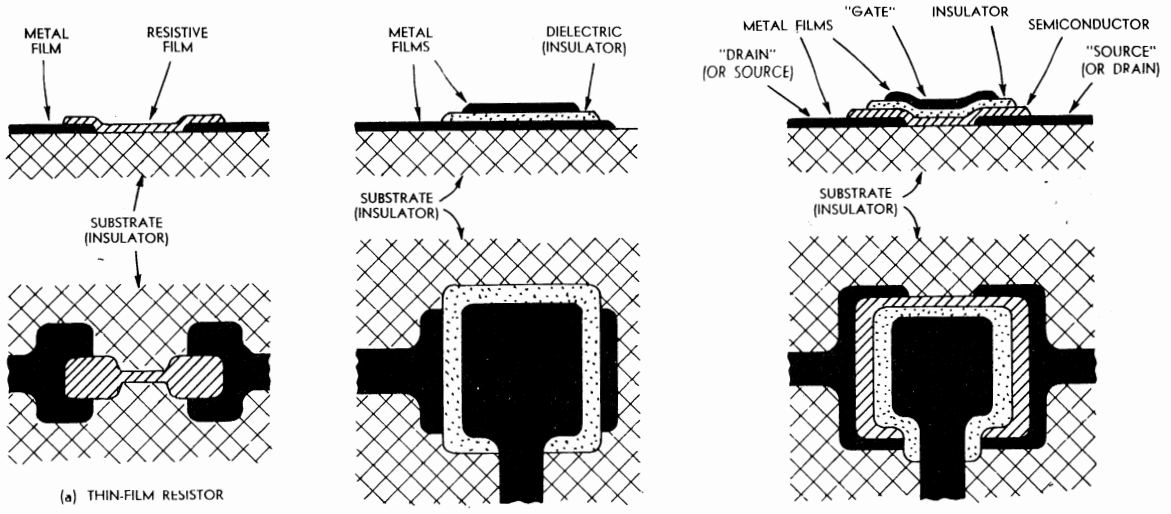
Thin-film capacitor elements are typically constructed as shown in figure 16.9(b), consisting basically of two metallic film plates separated by a thin insulating film as dielectric. The dielectric material is typically either tantalum oxide, titanium oxide, or a silicon oxide, in a film approximately 0.2uM thick. This provides approximately 1000pF of capacitance per square millimetre, with a breakdown voltage in excess of 50V, and thus allows quite useful capacitor elements to be formed in a very small area.

Again, the use of an element construction similar to that shown can be used to minimise the effects of registration errors on capacitor value. Thin film capacitors may thus be fabricated to within quite close tolerances, the element value being determined primarily by the thickness of the dielectric insulating film. In this and in most other respects their characteristics are rather similar to the monolithic silicon dioxide capacitor of figure 16.5(c), yet with the further advantage that like the thin-film resistor, they are free from parasitics.

As mentioned previously, the only type of active thin-film element to have been used in production devices to date is the **thin-film MOSFET transistor** or "TFT," which is usually fabricated as shown in figure 16.9(c). The first TFT devices were developed in 1961 by P. K. Weimer and C. Feldman.

It may be seen that this element is rather like a discrete MOSFET device, except that the semiconductor channel portion of the element is now simply a thin film of material deposited between two metallic electrodes. A further film of insulating material is deposited on the semiconductor film, and upon this again a final metallic film which forms the gate electrode.

To date most practical TFT elements have employed either cadmium sulphide, cadmium selenide or cadmium telluride as the channel semiconductor material, and have achieved transconductance figures ranging between about



(a) THIN-FILM RESISTOR (b) THIN-FILM CAPACITOR (c) THIN-FILM MOSFET

Figure 16.9

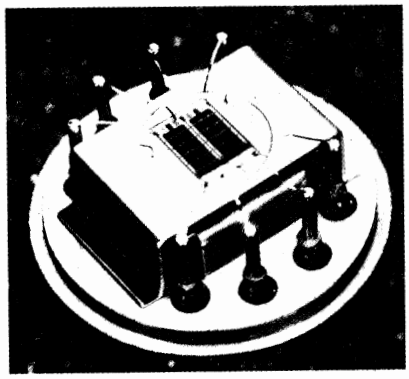


Figure 16.10: A thin-film thermopile device, used to perform "true RMS" AC-DC conversion for a digital voltmeter. (Courtesy Hewlett-Packard Australia.)

4 and 25mA/volt. However, the power dissipation capability has been rather low, typically about 25 milliwatts.

Other types of thin-film active element have been investigated, and some progress has in fact been made both with metal-semiconductor or "hot carrier" triode structures, and with tunnel triode structures. However, only limited success has been achieved to date with these alternative elements.

A typical thin-film microcircuit device is shown in figure 16.10. The device is a thin-film thermopile array, developed by Hewlett-Packard to provide "true RMS" measurements for their model 3480A digital voltmeter.

Because of the difficulty encountered in producing suitable thin-film active components for many applications, fully thin-film microcircuits are at present comparatively rare. Much more common are the hybrid devices, in which thin-film passive components and wiring are combined with semiconductor chip active devices. This type of device is rapidly becoming the preferred microcircuit form for high performance specialised devices.

Three different types of hybrid microcircuit have evolved to date. The first of these to appear was the **discrete hybrid device,** in which the active elements consist of fairly conventional discrete planar diode and transistor chips, bonded to suitable electrode pads provided on the thin-film passive structure.

To date most of the chips used in discrete hybrid devices have been standard discrete planar chips, bonded to the thin-film structure in a manner very similar to that in a discrete device, with thin wires used to connect to the top electrode metallisation. However, recent devices have used modified chips designed for inverted or "face down" mounting direct to the thin-film structure, without additional wiring. This type of chip is fitted with integral top contacts/mounting feet which are formed in a variety of ways, and described variously as "solder bumps," "solder balls" and "beam leads." The chips themselves are usually termed either "flip-chips" or "LIDs," the latter term being an acronym for "leadless inverted device."

A second type of hybrid microcircuit is the so-called **monobrid device,** which is similar to the discrete hybrid device except that here the semiconductor device chips incorporated with the thin-film structure are complete monolithic circuits rather than single discrete devices. Hence the monobrid type of device is a most flexible type of device, combining both monolithic and thin-film techniques.

Increasing use is being made of the monobrid format to produce small quantities of both extremely critical high performance linear devices and highly specialised complex LSI devices. For the latter purpose the monobrid approach is virtually ideal, permitting large numbers of pre-tested "stock" monolithic subassemblies to be assembled rapidly on a custom-prepared thin-film structure containing both wiring and any necessary further components, such as precision resistor networks.

The third type of hybrid microcircuit in current use is the **compatible hybrid device,** which, like the monobrid device, combines both monolithic and thin-film techniques. However, in this case the two are combined directly, as the device consists of a monolithic semiconductor chip containing the active components, to which thin-film wiring and passive components are added by deposition upon the surface passivation.

Probably the main advantage of compatible hybrid devices is that, like monolithic devices, they may be fabricated entirely as on-wafer arrays, involving no special operations to mate the thin-film and monolithic circuit elements. This gives them an edge in terms of reliability, and also seems likely to make them the most attractive future hybrid format from an economic viewpoint.

And with those brief comments on hybrid microcircuit devices, this chapter must unfortunately be drawn to a close. There are many significant aspects of microcircuits which have not been discussed, due to inevitable space limitations. Among these aspects are the important matters of device packaging and interconnection systems, testing procedures, and the many developments which have recently been made in fabrication techniques. However, it is hoped that the material which has been presented has given the reader at least a familiarity with the basic concepts of microcircuit technology, sufficient to provide a groundwork for further reading.

SUGGESTED FURTHER READING

CARROLL, J. M. (Ed.), **Microelectronic Circuits and Applications,** 1965. McGraw-Hill Book Company, New York.

CURRAN, L., "In Search of a Lasting Bond," in **Electronics,** V.41, No. 24, November 25, 1968.

GORE, W. (Ed.), **Microcircuits and Their Applications,** 1969. Iliffe Books Ltd., London.

RIGBY, G. A., "Establishing an Australian IC Facility," in **Radiotronics,** V.34, No. 4, December, 1969.

STERN, L., **Fundamentals of Integrated Circuits,** 1968. Hayden Book Company, New York.

WARNER, R. M., and FORDEMWALT, J. N., **Integrated Circuits — Design Principles and Fabrication,** 1965. McGraw-Hill Book Company, New York.

Chapter 17

PRESENT AND FUTURE

In this final chapter the author gives a survey of the current "state of the art" in solid state technology. Described are recent developments in fabrication technology, achievements and trends in the various device areas, and new types of device currently emerging.

Developments are now taking place so rapidly in the field of solid state technology that any survey which attempts to capture the current "state of the art" is bound to date very rapidly. Even in the brief time involved between writing and publication it is possible that events might easily relegate many items in such a survey to the limbo of recent history. Yet despite this, it would surely be unsatisfactory to end any basic introduction to the subject without at least a brief discussion of recent developments and trends.

While the survey which follows has been made as up-to-date as possible, it should therefore be regarded primarily as a conceptual bridge, written to assist the reader in understanding and evaluating both current and future developments in the light of the basic concepts presented in earlier chapters.

In order to make maximum use of the available space and also to increase the potential value of the chapter for reference purposes, the material is sectionalised under a number of topic headings. The initial section deals with fabrication technology, while following sections deal with the various types of established devices — both discrete and IC. The final section looks at currently emerging devices and technologies, together with likely future trends.

To begin, then.

FABRICATION TECHNOLOGY: One of the most significant developments in the fabrication field has been the rapid growth of the ion implantation impurity technique, which has supplanted the diffusion technique in many cases.

In ion implantation, impurity dopant atoms are injected into the semiconductor crystal wafers by direct bombardment in a vacuum. The atoms are given an electric charge, forming ions which are accelerated towards the wafers by means of an intense electric field.

Ion implantation is not new, having been first envisaged by William Shockley in the early 1950s. However, only recently have the techniques been refined and developed to the stage where they provide a fully practical commercial alternative to diffusion.

Ion implantation offers three main advantages. Possibly the most important of these is that it is considerably faster than diffusion; typically the ion bombardment process itself takes only 4 or 5 minutes, although a further 10 minutes or so is required for a follow-up thermal annealing process. The latter is necessary to allow the crystal lattice of the wafers to "recover" from structural damage incurred during the bombardment, and to ensure that the dopant atoms are fully incorporated into the lattice system.

The total ion implantation process thus takes only about 15 minutes compared with roughly as many hours for the diffusion process. This makes the process very attractive from a commercial viewpoint.

A second advantage of ion implantation is that the semiconductor wafers need not be heated to temperatures anywhere near as high as those required for diffusion. The ion bombardment process itself generally takes place at room temperature, while the follow-up thermal annealing operation typically involves a temperature between 500 and 650°C, which is considerably lower than the 900-1300°C range required for diffusion. The lower temperature processing generally eases contamination problems and wafer damage due to thermal cycling, and manufacturing yields tend to be higher as a result.

The third advantage is that ion implantation is proving to be more capable of precise control than diffusion. Doping penetration tends to be more uniform, and selective doping is subject to neither the lateral growth which occurs with diffusion, nor the continued penetration which occurs with successive diffusions. It has thus proved possible using ion implantation to produce circuit elements which are significantly smaller than may be produced by diffusion, yet with more tightly controlled parameters and higher performance as well.

This advantage is of great importance in view of the continuing emphasis on producing IC devices with more and more circuit elements on the chip. Ion implantation is one of the developments which has made possible the "LSI" revolution — the production of large-scale integrated circuits with complete functional systems on a single chip. It is also playing an important role in the move toward the next step — "VSLI", or very-large-scale integration.

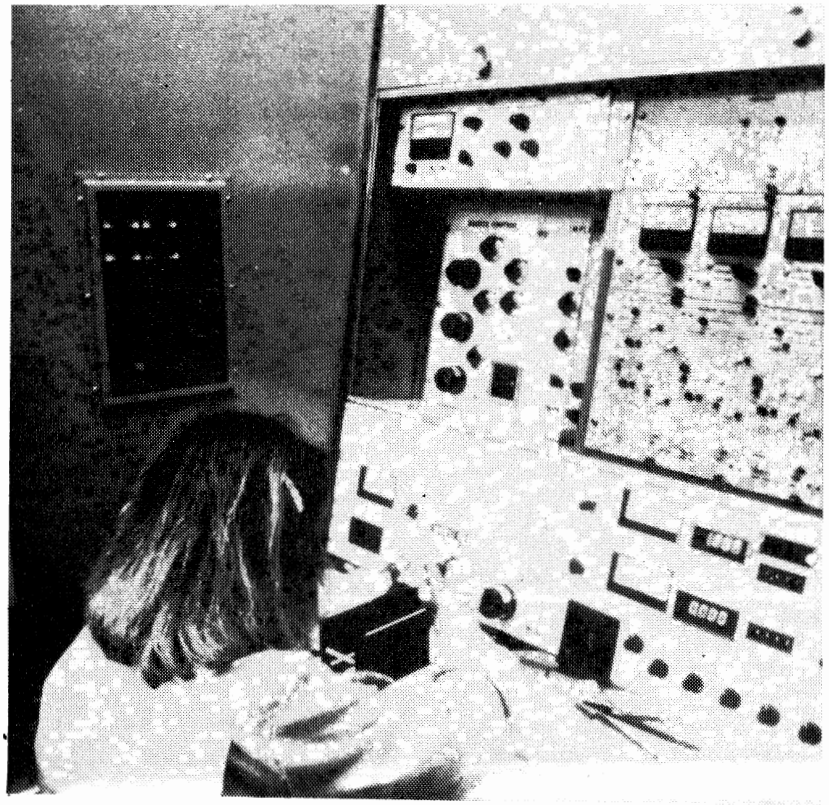


Figure 17.1: The control console of a modern ion implantation machine. (Courtesy Signetics Corporation.)

Ion implantation equipment is considerably more expensive than diffusion furnaces, but the increased throughput tends to counteract this disadvantage in actual production. And with implantation capable of greater resolution, the comparison is becoming of decreasing relevance.

Larger and larger silicon wafers are being used for device fabrication, to achieve greater output and reduce the labour content per individual chip. Currently wafers of 75mm diameter are still being used in some plants, although most have either changed or are in the process of changing over to use 100mm wafers. These offer an increase in active area of almost double the 75mm size, with a corresponding increase in manufacturing efficiency.

It is likely that wafers of 150mm diameter will be in use by 1980.

Another area in fabrication technology where important developments are taking place is photolithography.

Although contact printing is still being used to expose the photo-resist for wafer etching, diffusion/implantation and metallisation, it is rapidly being superseded by projection printing techniques. These have two main advantages: (a) the "mask" plates and the wafers being exposed do not touch, so both are less likely to be damaged; and (b) the ease of accurate alignment and mask registration is somewhat greater, making it possible to produce devices with smaller elements — and therefore greater packing densities — while still maintaining throughput.

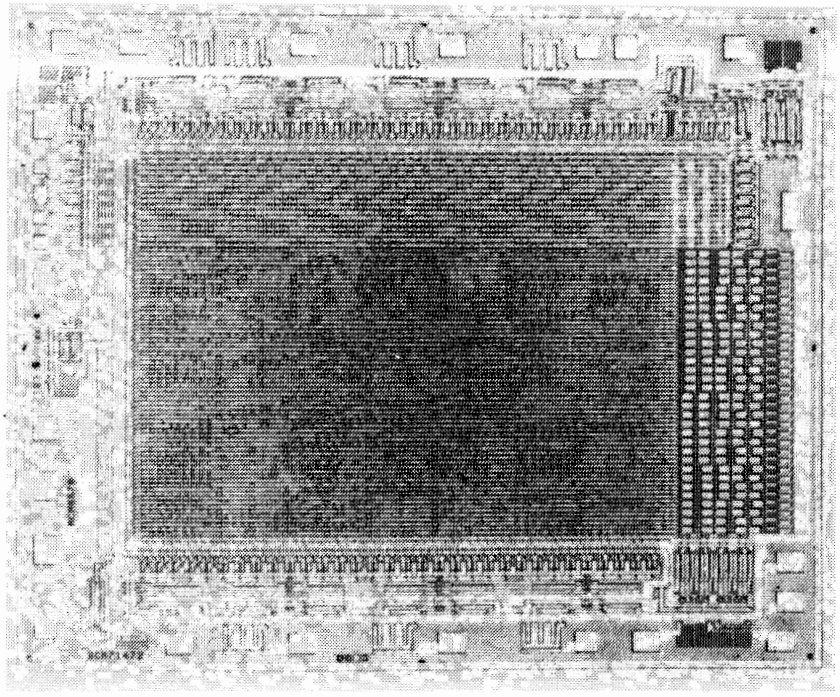


Figure 17.2: Ion implantation has made possible LSI devices like this 8192-bit ROM.

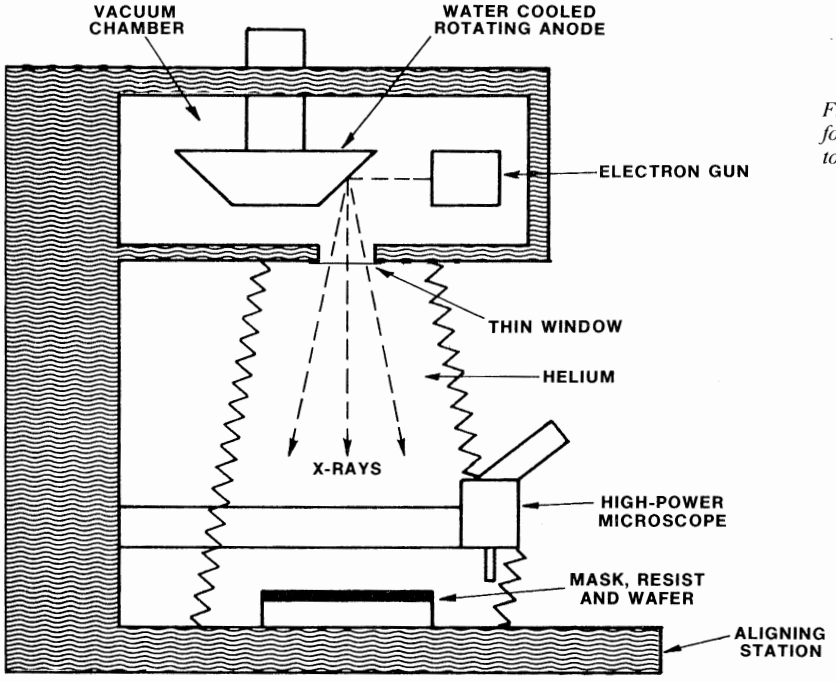


Figure 17.3: The basic system used for X-ray lithography of semiconductor wafers.

Projection printing is now being used increasingly in most fabrication plants, with ultra-violet light still being used for the actual exposure. The exposure masks themselves are still, in some cases, made by conventional micro-photographic means, but the movement towards narrower line widths is forcing the adoption of techniques capable of greater resolution.

As yet the main alternative to microphotography is electron beam or "E-beam" lithography, where a high energy electron beam is used to expose the masks. This produces masks capable of being used to create wafer line widths down to approximately 2 micrometres, which is somewhat better than can be done with conventional microphotography. Although E-beam equipment is very expensive — around $1.5 million per machine — it is therefore being used increasingly for making the masks for devices like LSI microcircuits.

By 1981 it is expected that wafer line widths of less than 2 micrometres will be required. One possible way of achieving this is to use the E-beam technique for direct writing on the semiconductor wafers themselves, cutting out the intermediate mask. This seems likely to achieve line widths of 0.5 micrometre or less.

An alternative approach is X-ray lithography. This is currently undergoing development, and seems likely to become a powerful fabrication tool in the mid-1980s. Already, experimental X-ray lithography systems have produced wafer line widths as narrow as 0.16 micrometre, suggesting that the technique has considerable potential.

As yet projection printing is not feasible with X-rays, as no focussing method has been found. The technique currently uses a slightly modified form of contact printing, with the exposure mask held above the wafer by a distance just sufficient to prevent mutual damage — around 25 micrometres.

The exposure masks themselves tend to use a thin layer of gold or chromium as the X-ray absorber, etched away in the required pattern. The mask substrate is typically a thin plastic material such as Mylar.

DISCRETE DEVICES (1) DIODES: A variety of semiconductor P-N diode known as the IMPATT diode has become increasingly used as a compact, low cost and highly reliable source of microwave energy. This type of device was first proposed by W. T. Read of Bell Telephone Labs in 1958, and is actually a refinement of the conventional avalanche breakdown or "zener" diode. The name IMPATT is an acronym for "impact avalanche and transit time".

The operation of the device depends upon two characteristics of the reverse biased P-N junction avalanche breakdown mechanism. One of these is that a finite time is required before avalanche current flows, after the application of breakdown inducing bias; the other is that avalanche current carriers take a finite transit time to cross the depletion layer. The total delay time produced by these two effects is sufficient to correspond to a phase shift of greater than 90 degrees at microwave frequencies, and hence such a junction tends to possess a negative component of AC resistance at such frequencies.

When the IMPATT diode is inserted into

a suitable microwave resonant circuit and biased at the threshold of avalanche breakdown, the negative resistance component of its AC junction resistance effectively cancels the resonant circuit losses, and hence the circuit produces continuous oscillations.

Because they are operated in avalanche breakdown mode, IMPATT diodes dissipate considerable heat energy when in operation. This poses problems for the device chip designer and for the package designer, because the device must be kept very small to be compatible with microwave circuitry. Accordingly recent devices have all been fabricated from silicon, and have used a special "upside down" chip format, with the junction itself at the very bottom to bring it into close proximity to the solid copper header of a miniature "pill" package.

This has produced IMPATT devices capable of providing up to 3.5 watts of continuous or "CW" power at 6GHz, with 10 per cent efficiency and conduction cooling at room temperatures. Pulse-optimised devices can provide 14W pulses at 10GHz, with 25 per cent duty cycle.

IMPATT diodes are not restricted to the low microwave region, either; in fact researchers at Bell Labs are confident of achieving operation at sub-millimetre frequencies (above 300GHz) in the very near future. Already devices have been produced capable of generating a power of greater than 30mW at 150GHz, with an efficiency approaching 3 per cent.

Closely related to the IMPATT diode is a high-efficiency microwave avalanche diode first discovered at RCA Laboratories in 1967. Known as an "anomalous mode" avalanche diode because it operates in a manner which is as yet only partially understood, this device is capable of generating microwave energy at frequencies considerably lower than would be explained by the normal avalanche delay and transit times. Experimental devices have been used to produce pulse power levels of greater than 1kW at 1GHz, with an efficiency greater than 25 per cent.

Another type of solid state diode now in fairly wide use is the so-called HOT CARRIER or "SCHOTTKY BARRIER" diode. This type of diode is not based on a semiconductor P-N junction at all, but rather on a junction between a semiconductor material and a metal. It relies upon the fact that there tends to be set up at such a semiconductor-metal junction a potential barrier very similar to that set up in a semiconductor P-N junction, giving the junction rather similar properties. The potential barrier is in this case known as a **Schottky barrier,** in honour of the physicist who first postulated its existence.

Although a metal-semiconductor junction may be formed using either P-type or N-type semiconductor material, N-type is normally used because this causes device operation to be mainly due to highly mobile conduction band electrons. If P-type material were used, operation would be due to the lower mobility valence band holes, and device operating speed would be lowered.

If external bias is applied to an N-type semiconductor-metal junction with the metal made negative with respect to the semiconductor, the Schottky potential barrier between the two is merely increased, and virtually no current flows. This bias condition thus corresponds to the "reverse bias" condition of a semiconductor P-N junction.

However, if the external bias is applied such that the metal is made positive with respect to the semiconductor, the barrier potential is reduced and electrons are injected into the metal from the semiconductor, causing a substantial current flow. This situation thus corresponds to the "forward bias" condition of a semiconductor junction.

In the forward bias condition, electrons are injected into the metal with an initial energy level substantially above that of the metal's own "free" electron population. These carriers are thus effectively "hotter" than the other electrons in the metal, and it is this fact which causes metal-semiconductor diodes to be called "hot carrier" diodes. Once in the metal the hot electrons give up their excess energy in an exceedingly short time — about one tenth of a picosecond — and immediately become indistinguishable from the other electrons.

Because of the characteristics of metal-semiconductor junction conduction, hot carrier diodes are capable of switching at extremely high speeds. Recent devices developed for UHF mixing and detection applications have achieved total switching times less than 50ps (1ps = 10^{-12}sec). They also tend to exhibit very low forward voltage drop, and this has resulted in their use for high-current low voltage power rectification.

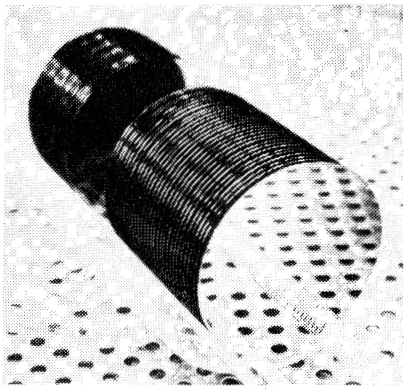


Figure 17.5: Most manufacturers now use 150mm diameter wafers.

A typical modern hot-carrier power diode handles 75A with a forward voltage drop of only 0.7V.

A further type of discrete diode device coming into increasing use is the "transient protected" or controlled avalanche rectifier diode, of which mention was made in chapter 5. This type of device is designed to enter avalanche breakdown in a distributed and controlled fashion, and is thus capable of sustaining short reverse transients of quite high amplitude without damage. Because of the controlled breakdown characteristic, such diodes may be connected in series for high-voltage applications without the need for additional components to ensure voltage sharing between devices. Recent devices have been connected in series stacks to produce assemblies capable of rectifying up to 100kV at current levels up to 10A.

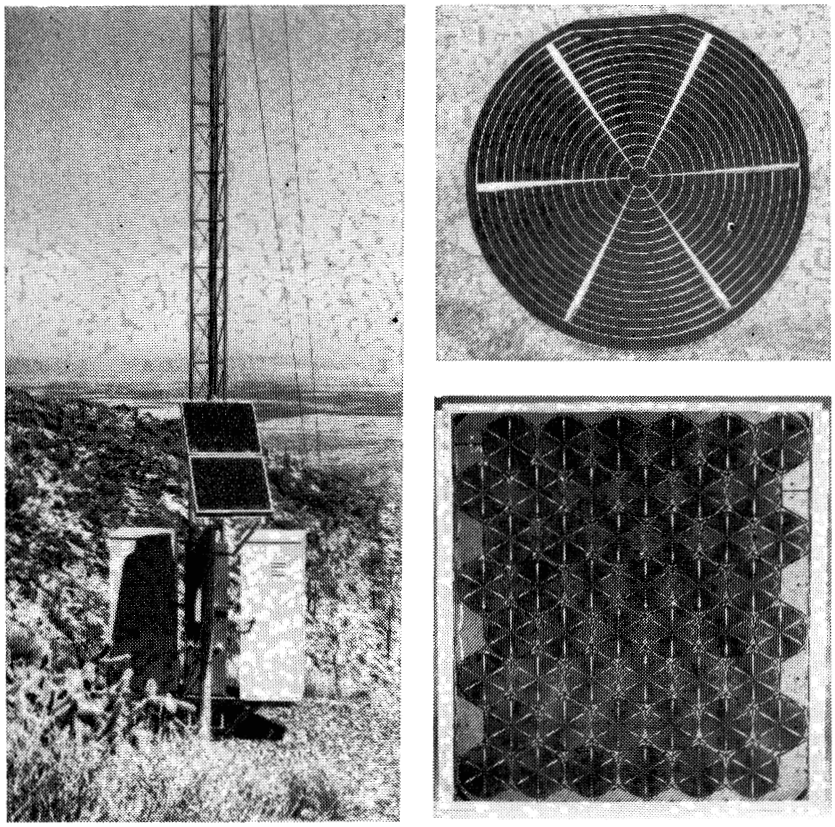
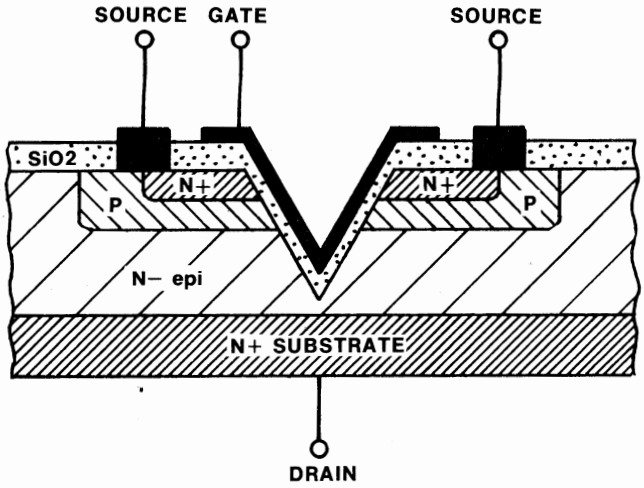


Figure 17.4: Silicon solar cells are growing in importance. A typical cell is shown at upper right; below it is a full array, and at left a typical application. (Motorola)

DISCRETE DEVICES (2)

TRANSISTORS: Much recent development work in discrete bipolar transistors has been directed towards improved high power devices. Many improved power devices have been produced for UHF and microwave applications; typical devices are now capable of delivering more than 50W at 175MHz, 20W at 470 MHz, and 10W at 2GHz. Other recent power devices have been designed for very high voltage or high current switching



CROSS SECTION OF A VMOS TRANSISTOR

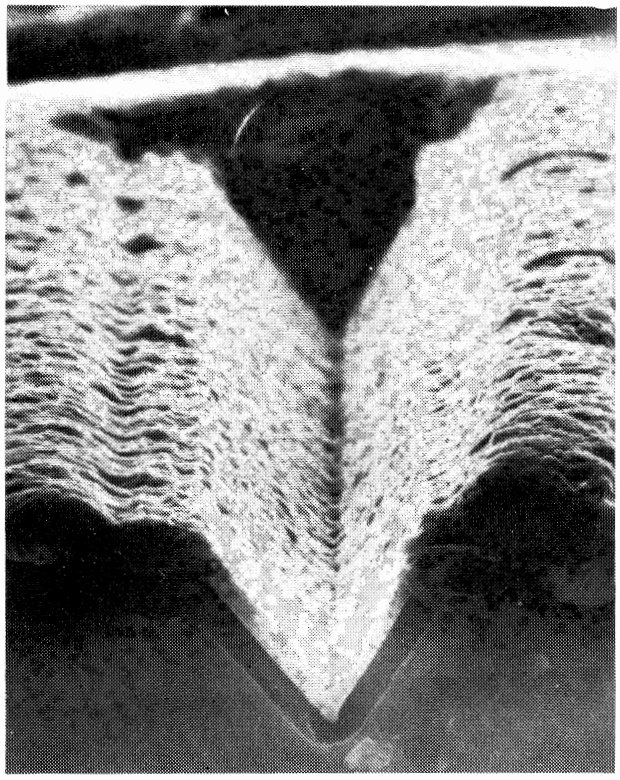


Figure 17.6: The construction of a VMOS power transistor is shown by the digram above and the scanning electron microscope view at right.

at low frequencies; one recently announced device offers a BVceo of 1000V, with a power rating of 100W, and a continuous collector current rating of 5A.

"Power Darlington" devices have been developed as one attractive answer to the problem of achieving high current gain from a bipolar device at high current levels. These devices are actually two cascaded bipolar elements in a single package, connected in the "Darlington" configuration — both collectors are connected together, with the emitter of the "input" device directly connected to the base of the "output" device so that the overall current gain becomes equal to the product of the two betas.

Recent devices of this type have employed IC techniques to fabricate both elements on a single monolithic silicon chip. This has both lowered device costs and improved performance: typical devices offer a minimum beta of 1000 at 3A collector current, with Icbo as low as 200uA at 60V for a device having an LVceo rating of 80V.

The most significant recent developments in discrete FET devices have been in the area of power handling capability.

In the early 1970s, pioneering work on power FET devices was done at the Japan Semiconductor Research Institute, and experimental devices were produced which were capable of controlling up to 40 watts of power. The devices used a lattice-pattern P-type gate, and an almost-intrinsic material N-type channel array between heavily doped N-type source and drain regions. It had an relatively high transconductance: 100mA/V.

Although the chips of the devices were only some 2.5mm square, samples were used to control currents of 200mA at up to 200V.

A different type of power FET was later developed at RCA Laboratories, in this case a microwave power amplifier device fabricated from gallium arsenide material. The device used a metal gate and a metal-semiconductor Schottky barrier for gate isolation. Experimental samples produced a power output of 5.6 watts at 2.2GHz, with 6.5dB power gain.

More recently a new type of power FET device has emerged — a power MOS transistor employing a vertical structure. The US company Siliconix Incorporated has done major work in this area, and calls the devices "VMOS power MOSFETs".

The VMOS transistors are fabricated from N-type epitaxial wafers, with the epitaxial layer much more lightly doped than the bulk of the wafer. Into the epitaxial layer are first diffused lightly doped P-type islands which form the ultimate channel regions. Then smaller but heavily-doped N-type islands are diffused into the centres of the former to form the source regions. (The heavily-doped N-type substrate and the epitaxial layer become the drain regions.)

A V-shaped groove is then etched out of the centre of each concentric island, down through both the source and channel regions and into the epitaxial layer. Oxide passivation is then grown both in the grooves and on the wafer surface, and aluminium metallisation deposited both on the surface to form the source connections, and in the grooves to form the gate electrodes and connections. Finally the wafers are given a further passivation to keep contamination from penetrating the gate oxide.

The finished VMOS transistors are enhancement-mode devices. With the metal gate electrode held at the same potential as the source, there is no conduction path between the N-type source region and the N-type epitaxial and substrate drain regions. However if the gate electrode is made positive with respect to the source, this induces N-type channels in the P-type region on either side of the groove, and allows current to flow.

The VMOS structure has a number of advantages. The length of the channels is determined by diffusion depths, rather than by masking as in a conventional MOS transistor; this gives a better width to length ratio, and allows higher current densities to be obtained. In any case the V-groove

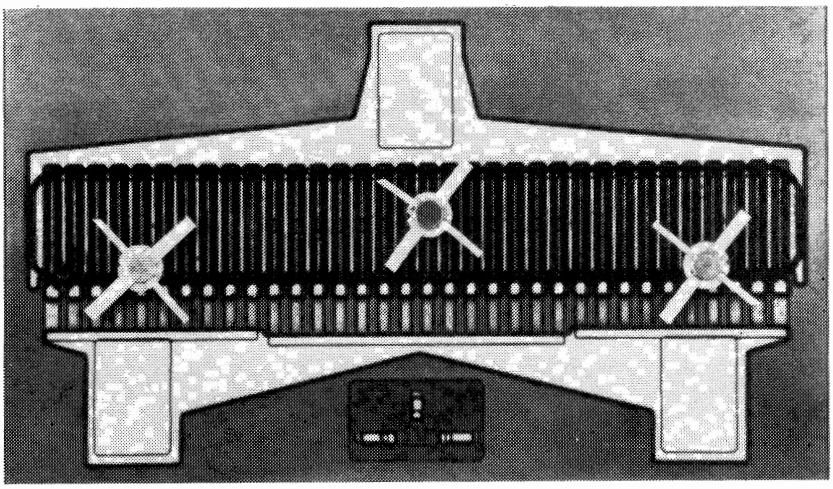


Figure 17.7: Three modern bipolar microwave transistors on a microphotograph of their chip pattern. Each device consists of many small transistors in parallel, with emitter resistors to ensure current sharing. (Hewlett-Packard)

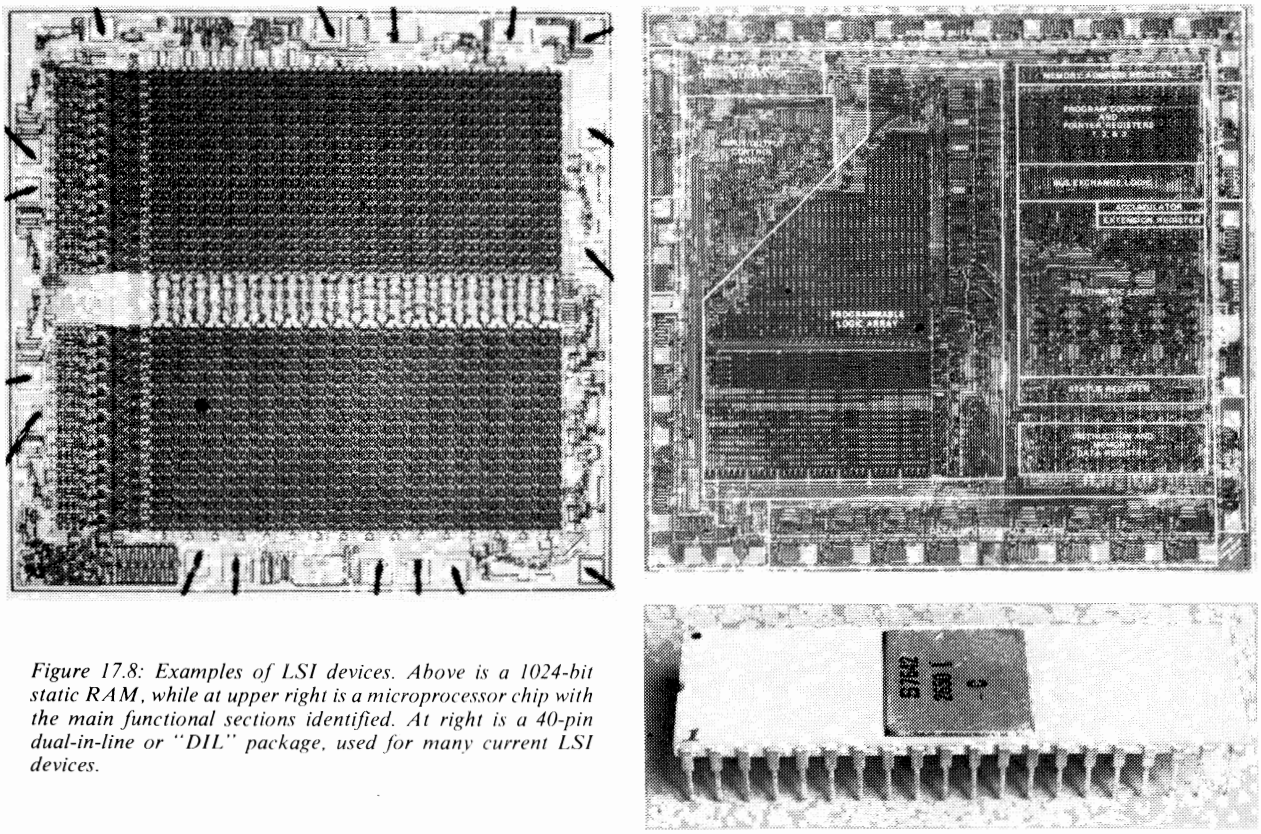


Figure 17.8: Examples of LSI devices. Above is a 1024-bit static RAM, while at upper right is a microprocessor chip with the main functional sections identified. At right is a 40-pin dual-in-line or "DIL" package, used for many current LSI devices.

produces two parallel channels, so that the current density is inherently doubled.

As the substrate forms the drain contact, no drain connection is needed at the top of the chip. This further reduces the chip area needed, and keeps saturation resistance low.

Since the gate only overlaps the drain at the bottom of the V-groove, gate-drain feedback capacitance is relatively low compared with a normal MOSFET.

The epitaxial layer of the VMOS device absorbs the depletion layer of the reverse-biased channel region-drain region junction, and thus gives the device a relatively high breakdown voltage — typically as high as 90V.

The VMOS device is capable of switching at high speeds and amplifying at high frequencies, both because it is a majority-carrier device and because it uses electrons rather than holes as the carriers. A typical VMOS switching device is capable of switching 1A on or off in about 4 nanoseconds — about 10 to 200 times faster than a bipolar power transistor. Similarly a typical VMOS device intended for RF amplifier applications is capable of delivering 20 watts at 150MHz with an input power of only 1 watt. Typical transconductance figures are from 100 to 250mA/V — very high indeed.

Other advantages of VMOS devices are very high input impedance, inherent thermal stability and the ability to share current evenly both within a single device and between devices in parallel. Like all FETs and unlike bipolar transistors which have an inherent tendency towards thermal runaway, VMOS devices tend to draw less current as temperature rises. This makes the device not only stable, but free from the current crowding and localised "hot spots" which cause secondary breakdown in bipolar transistors.

Actually VMOS devices seem to offer so many advantages over bipolar power transistors that they and similar FET devices seem destined to become the power semiconductor devices of the future, eventually supplanting bipolar devices in this area.

DISCRETE DEVICES (3) THYRISTORS: Planar epitaxial and diffused PNPN thyristor structures have been developed, and are becoming increasingly used both for low-current discrete devices, and for thyristor elements incorporated into monolithic microcircuits.

One interesting variation on the planar thyristor device theme is a reverse-blocking triode device (SCR) in which triggering is performed by means of the field effect. The P-type anode and cathode gate regions of the device are separated by a lightly doped N-type anode gate region into which a conducting channel is induced by a metal gate electrode deposited above it on the silicon dioxide passivation. The device thus combines a normal PNPN thyristor configuration with that of an enhancement or "type C" MOSFET, and offers very high power gain.

Steady but continuing progress is being made in the development of very high power thyristor devices. In Japan, Hitachi Ltd makes SCR devices capable of handling 400A at 10kV and 1600A at 2.5kV respectively, intended for such applications as electric traction. Triac devices of up to 200A capacity at 1kV have been produced by International Rectifier of California, for use in heavy duty AC static switching.

DIGITAL ICs: Undoubtedly the most spectacular developments in modern semiconductor technology have taken place in the area of digital ICs, particularly in LSI or "large-scale integration" devices such as large RAM and ROM memory arrays, microprocessors and microcomputers, and dedicated controller devices. Over the last few years there has been an almost constant stream of new devices, each representing a significant improvement over earlier devices in some respect or other.

Progress has been particularly dramatic in the area of random-access read-write memory devices or RAMs. Around 1970 the largest available RAM device was capable of storing 1024 bits of information on a single chip — the so-called "1k" device. This has now been increased to 8192 bits in the case of static RAM devices, where the information is stored in flip-flops, and 16,384 bits in the case of dynamic RAMs which store the information in tiny capacitors. By 1980 both these figures are likely to be increased by a factor of four, with a further four times increase almost certain before 1985.

Mask-programmed read-only or ROM devices have followed a similar course, if not perhaps quite as spectacular. Around 1970 these were available with up to 4096 bits of storage, while they are now available with capacitites up to 65,536 bits. By 1985 these devices will probably be available with capacities of around 524,000 bits, if not more.

Field programmable ROMs or "PROMs" have made considerable progress also. In 1970 the only type available was the fusible-link variety, programmed by burning away tiny metallisation links using short current pulses; these were available with capacities of up to 512 bits. Now this type of device is available with capacities up to 16,384 bits.

In 1971 an alternative type of PROM device, the ultra-violet erasible PROM or "EPROM" was developed by Dov Frohman-Bentchkowsky at Intel Corporation. This uses storage cells based on a MOS transistor with a floating gate electrode, onto which charge is stored by inducing an avalanche breakdown between drain and source, the so-called "FAMOS" or floating-gate avalanche-mode transistor.

Since its initial development the EPROM

has grown in capacity from around 2048 bits to 16,384 bits, with devices providing up to 65,536 bits likely to be available by 1980.

Around 1970, a new type of LSI digital IC emerged: the microprocessor. This offers virtually all of the computing or processing part of a digital computer, compressed into a single LSI chip.

The first microprocessors were relatively slow devices, which operated on 4-bit binary numbers and had a relatively small repertoire of instructions. Since then device capabilities have been increased in most respects — speed, word length and instruction power. Current devices are available which handle 16-bit numbers at speeds approaching that of minicomputers, and with an extensive repertoire of powerful instructions.

Also emerging are devices which combine microprocessors with RAM and ROM memory capacity, to provide a complete digital computer on a chip: the microcomputer. At present these are relatively modest devices, mainly suitable for low-level dedicated "controller" jobs, but by 1985 it is predicted that we will have computer chips offering the equivalent of one of today's minicomputers — on a single IC chip!

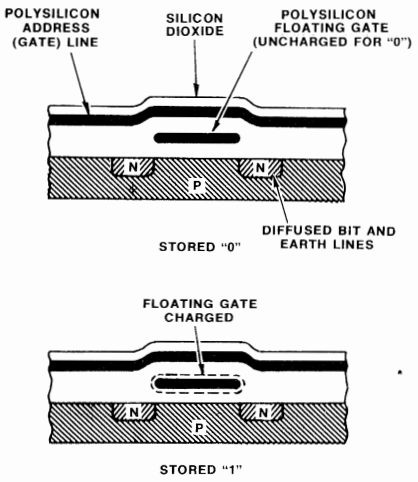


Figure 17.9: The UV-erasible PROM or "EPROM". Above is the basic structure of the FAMOS transistor storage cell, while a right is a typical current device with the chip visible under the quartz window.

LINEAR ICs: Developments in the linear IC area are generally not as spectacular as in the digital area. This may well be due to the fact that linear devices are generally harder to design and to fabricate than switching-type digital devices.

In the general purpose linear DC amplifier or "op amp" area, the early 1970s saw the development of higher performance devices with input stages using so-called "super-beta" transistors — bipolar transistors with common-emitter current gains approaching 10,000. The transistors were basically standard diffused planar devices with extremely thin base regions.

As well as giving the transistors high current gain, the thin base gave the transistors extended frequency response, an improved noise figure and low leakage. It also allowed the high current gain to be maintained down to very low emitter current levels: as low as 10nA.

The combination of very high current gain at low emitter current and low leakage allowed the super-beta transistors to provide the op-amps with high input impedance — up to about 2000 megohms. However with such a thin base region the devices had a very low breakdown voltage, which tended to limit the ability of the devices to handle large input signals.

More recently a new breed of high performance op-amp devices has emerged, based on the combination of bipolar and FET technologies. This is the "Bi-FET" op-amp, which uses FET input transistors and bipolar output circuitry on a common monolithic chip.

National Semiconductor has produced devices of this type using JFET input elements, while RCA has produced devices with MOSFET input elements. In both cases the amplifiers offer very high input impedance, around 1 Teraohm (1,000,000 megohms), coupled with high gain, wide frequency response, the ability to cope with large input signals, and the ability to swing the output voltage very rapidly (high slew rate).

In the realm of higher power devices, the Japanese Sony Corporation has produced a monolithic audio amplifier device capable of delivering an effective power of 18 watts into 8 ohms. The device operates from a 40V rail, and produces less than 10 per cent total harmonic distortion at full output; efficiency is 67 per cent, and the total size of the device chip only 59 x 69 mils. Other manufacturers such as Plessey and General Electric have produced monolithic devices capable of between 3 and 10 watts output.

Higher audio power output has been achieved using hybrid devices. The Japanese Sanken Company has produced a hybrid device capable of delivering 50W effective; while TRW Semiconductors of California have produced a hybrid switching mode amplifier, for servo applications, which delivers 200 watts output at 90 per cent efficiency and 0.1 per cent linearity.

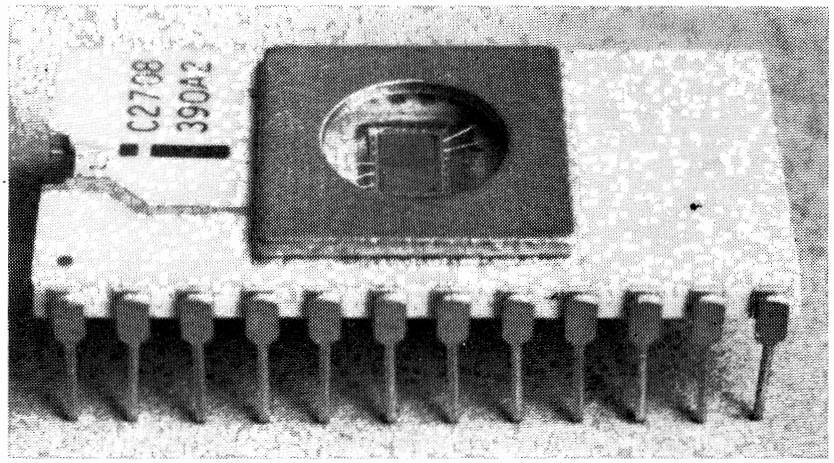
An important area of linear ICs is three-terminal voltage regulator devices, providing high-performance regulation circuitry on a single IC chip. Devices of this type, in a power-transistor style package, are now widely used for voltage regulation tasks in both digital and analog equipment. Typical modern devices are capable of regulating voltages up to about 30V, at currents of up to 5 amps. The reliability and performance offered are generally far better than can be obtained from discrete regulator circuitry of comparable cost, making these devices the logical choice in most applications.

NEW TECHNOLOGY: A number of new technologies are currently in the process of becoming established. Some are closely related to the solid state devices discussed earlier in this book, while others are more distantly connected if at all. Some appear to have a brighter future than others, although only time will perhaps tell which ones will be retained and which ones will fade into oblivion.

A new technology not very distant from the MOS devices discussed in chapter 8 is charge-coupled device or "CCD" technology, developed in 1971 by Drs Willard S. Boyle and George E. Smith at Bell Laboratories in New Jersey, USA.

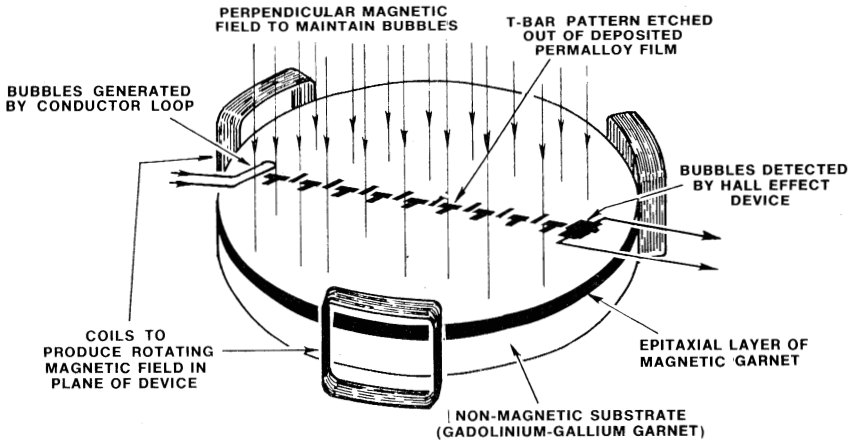
Although they use a metal-oxide-semiconductor structure like that of conventional MOSFET transistors, CCDs are considerably simpler than these devices because they have virtually no semiconductor junctions. The semiconductor material is essentially homogeneous. The devices operate by using bias voltages applied to a pattern of metallisation electrodes to manipulate small "packets" of charge carriers near the surface of the semiconductor.

The idea is that a CCD consists basically of an array of MOS capacitors, with the semiconductor crystal chip forming a common lower plate, and a series of metallisation electrodes the individual upper plates. Groups of carriers are held in the semiconductor beneath certain upper plates, because of "potential wells" formed due to bias voltages applied to these plates. The carrier groups may then be moved through the material as desired, by manipulating the bias voltages on the plates so that the potential wells transfer from one region to another.



CCD devices are essentially very simple in terms of construction, giving them the potential advantages of low cost and high packing density. And in some areas this potential has been realised already. CCD memory devices have been made with a capacity of 65,536 bits, on a single chip. Similarly CCD optical imaging arrays have been produced with as many as 185,440 visual detector elements in a 488-by-380 array — sufficient to give commercial-quality TV resolution, again on a single chip.

Unfortunately, in the memory device area it seems unlikely that CCDs will become really competitive with other technologies. Because a CCD memory device is basically a



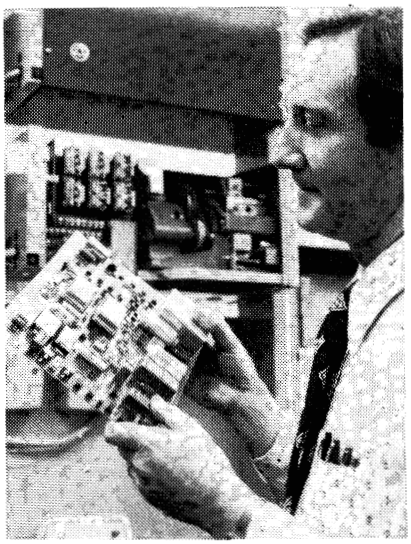


Figure 17.10: Magnetic bubble technology. Above shows the basic operation of a bubble device, with the pattern simplified for clarity. A wafer of devices is shown below, with a telephone message storage board using bubble devices at right. (Courtesy Bell Laboratories)

series of long shift registers, its operating speed and access time tend to be significantly slower than conventional RAM devices. At the same time the potential information packing density does not appear to be greater than conventional MOS or bipolar RAM devices, which are achieving higher and higher densities all the time.

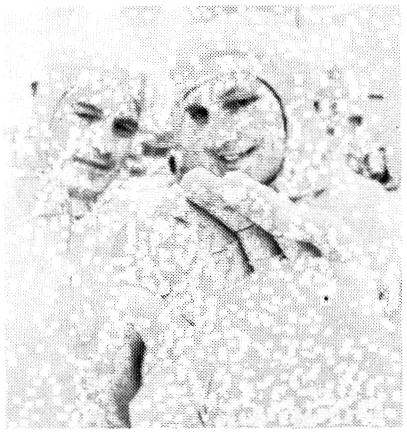
This being the case it seems likely that the main future of CCD technology will be in the area of optical imaging, sophisticated filters, and so on.

A new technology with perhaps more potential in the high-capacity memory area is magnetic bubble technology, pioneered by Bell Laboratories and the Autonetics Corporation of California. As the name suggests, this technology is based on tiny magnetic field "bubbles" or domains, which are manipulated around the surface of thin layers of magnetic materials such as the orthoferrites or "YIG" — yttrium indium garnet. The bubbles are created, manipulated, allowed to interact and also destroyed when required, by means of the interaction between a rotating magnetic "environment" field and suitably shaped microscopic thin-film patterns of permalloy.

Using this technology, Bell Labs, Autonetics and IBM have been able to produce shift registers, logic element arrays and memory arrays with information packing densities approaching 1.6 million bits per square cm. This is something like 10 times greater than the density possible with established techniques such as MOS memory arrays and magnetic disc recording.

At the same time, the energy required to manipulate the magnetic bubbles is about 100 times less than that required to manipulate information in an MOS memory array, so that magnetic bubble technology seems to hold considerable promise as a means of providing extremely compact, very low power consumption data processing equipment.

Bubble memory devices with a capacity of 250,000 bits on a single "chip" have been produced already. By 1980 it is predicted that the actual bubble sizes will have been reduced to below 1 micrometre in diameter, making it possible to produce memory devices with around 10 million bits capacity — at a cost which should make them very attractive as replacements for floppy discs and similar bulk storage devices. They may even be used to replace conventional audio and



video tape recorders by about 1985.

A more specialised area of new technology is that associated with so-called "bulk effect" devices, used to generate RF energy in the microwave spectrum. These devices owe their origin to a discovery in 1963 by J. B. Gunn of IBM, who found that the application of DC voltage across a simple homogeneous chip of gallium aresenide caused the chip to emit microwave energy.

Since Gunn's discovery, other researchers have found that oscillations can be produced by homogenous or "bulk" semiconductor material in a number of ways, the Gunn effect being in fact only one possibility. A mechanism known as the limited space-charge accumulation or "LSA" mode of bulk material oscillation was discovered in late 1966.

In basic terms, it would appear that bulk effect devices depend for their operation upon a tendency for electric field "bunching" to occur in certain semiconductor materials, when the applied voltage is raised above a critical level. The electric field "bunches" or domains move through the material at a fixed speed, so that the current passed by the device contains corresponding fluctuations. The fluctuations form the microwave AC energy which the device produces from the DC input.

To date the principal use of bulk effect devices has been as microwave energy sources. However, continuing research is being carried out into the mechanisms involved, and experimental results suggest that bulk devices may eventually be capable of ultra-high-speed waveform synthesis, logic and other operations.

A technology which received considerable acclaim when it was first publicly announced in November, 1968, is that based on amorphous glass semiconductors, also called "Ovonic" devices in honour of their acknowledged discoverer Stanford R. Ovshinsky.

Structurally the devices are very simple, in most cases consisting of nothing more than a blob of non-crystalline or "amorphous" glassy material between two metallic electrodes. The glass material has a carefully controlled composition, however, typical devices using a glass containing tellurium, germanium, and arsenic. Depending upon the exact ratio of these ingredients, the devices can be made to act as AC switching elements, threshold triggers, or bistable memory cells.

The exact mechanism responsible for amorphous glass device operation is not yet understood, and until this is known the future of this type of device must remain cloudy. However, some practical applications have already been found; the US Army is reported to have discovered that an experimental instrumentation amplifier constructed using amorphous glass devices had a greater resistance to high-intensity neutron radiation than any other equipment tested, the testing supply and monitor cables having disintegrated before the devices were affected.

Finally, a new technology which should be mentioned here is that associated with surface-acoustic wave or "SAW" devices. These devices depend for their operation on high frequency mechanical vibrations, which are generated and manipulated on the surface of piezo-electric materials by means of deposited metal electrodes. At present SAW devices are used for relatively specialised filtering applications, for which they seem particularly suited. However they may well find use in a wider range of applications in the future.

And with SAW devices this necessarily rather brief survey of modern solid state technology must end. Hopefully it has given you at least a broad idea of the many areas in which the various types of solid state devices are developing. For further information, and for news of future developments, I must refer you to the well-known technical magazines — including, of course, "Electronics Australia".

Appendix

A GLOSSARY OF TERMS

It is hoped that the following glossary of terms may assist the reader in understanding both the content of the foregoing chapters, and the concepts which may be encountered in further reading. However, it is by no means a complete inventory of the multitude of terms used in modern solid state technology.

ACCEPTOR IMPURITY: An element or compound whose atoms or molecules have fewer valency electrons than those of the intrinsic semiconductor material into which they are introduced in small quantities as an impurity or dopant. Because the acceptor impurity possesses fewer valency electrons its inclusion in the crystal lattice creates valency electron deficiencies (holes), so that material doped with an acceptor impurity is a P-type semiconductor.

ALPHA: One of the two main parameters used to express the current gain of a bipolar transistor. Alpha is usually defined as the ratio of a small change in collector current to the corresponding change in emitter current, when the collector-base voltage is maintained constant.

AVALANCHE: One of the mechanisms responsible for voltage "breakdown" of semiconductor junctions and devices. When avalanche occurs, carriers moving through the crystal lattice have achieved sufficient kinetic energy to knock further carriers from the lattice, producing a "snowball" increase in current level. Providing the current increase is limited externally, avalanche breakdown causes no permanent damage to the device.

BETA: The second of the two main parameters used to express the current gain of a bipolar transistor. There are many "versions" of beta, but all versions relate a change in collector current to the corresponding change in base current, with the collector-emitter voltage maintained constant.

BIPOLAR TRANSISTOR: A three-terminal active semiconductor device having two adjacent P-N junctions, and arranged so that there is a common P-type or N-type region shared by both junctions. In operation, one junction is forward biased and used to inject carriers into the other, which is reverse biased. This causes the device to provide a power gain.

CARRIERS: Entities which carry an electrical charge and are also able to move relatively freely through a crystal lattice. The two most commonly encountered carriers are conduction band electrons, which are negatively charged, and valency band holes, which are positively charged.

CHARGE-COUPLED DEVICES (CCDs): Semiconductor devices, usually integrated circuits, whose operation depends upon the manipulation of "packets" of electrical charge near the surface of the semiconductor region. The charge packets are manipulated by varying control voltages on a pattern of metal electrodes deposited on the oxide surface passivation.

CMOS INTEGRATED CIRCUIT: An integrated circuit which uses both types of MOS transistor — N-channel and P-channel — and takes advantage of their complementary electrical properties. CMOS devices generally offer markedly lower current consumption.

COLLECTION: The mechanism whereby the high potential gradient and intense electric (drift) field present within the depletion layer of a reverse biased P-N junction can cause the depletion layer to "collect" any carriers of appropriate type which happen to diffuse into it from the adjacent semiconductor regions.

COMPENSATION: The phenomenon whereby extremely small quantities of donor and acceptor impurities present in a semiconductor crystal tend to "cancel out" each other, so that the material tends to behave according to the dominant impurity only. If both types of impurity are present to an equal extent, the material tends to behave as pure "intrinsic" material.

CONDUCTOR: Any material whose valency energy band is only partially filled with electrons, so that empty levels are immediately available for a nett electron movement. Such materials conduct electricity readily, even at extremely low temperatures.

CONDUCTIVITY: The parameter of a material which indicates the extent to which it permits the flow of a net electrical current, and normally measured in terms of the conductance in reciprocal ohms (Siemens or Mhos) between opposite faces of a cube of the material, say one measuring one centimetre on each side. The conductivity of a material is the reciprocal of its resistivity.

CONDUCTIVITY MODULATION: The process whereby the effective electrical conductivity of a semiconductor region is modified by the injection of excess carriers. Thus excess majority carriers injected into a lightly doped region can cause the effective conductivity to be increased, simply by providing further carriers for current conduction. Conversely excess minority carriers injected into a heavily doped region can cause the effective conductivity to be reduced, by increasing the incidence of recombination and hence reducing the number of carriers available for conduction.

CRYSTAL: Solid material in which the atoms or molecules are arranged in regular three-dimensional "lattice" arrays.

CRYSTAL PULLING: A technique first developed by J. C. Czochralski, in which a monocrystalline "seed" is introduced into the top of a body of molten material, and then withdrawn slowly to grow or "pull" a large single monocrystal. This technique is used in semiconductor manufacture to produce the uniformly doped monocrystal boules from which most devices are fabricated.

CUTOFF, device: That condition of an electronic device in which its conduction is either zero or relatively insignificant. With semiconductor devices such as FETs, bipolar transistors and thyristors, cutoff is normally that condition in which the device passes only saturation and leakage currents.

DARLINGTON TRANSISTOR: A combination of two or more bipolar transistors, connected together in cascade to form what is effectively a single, high gain transistor. The transistors forming the Darlington may be either of the same type or complementary types, in which latter case the term "complementary Darlington" is often used.

DEPLETION LAYER: That region in the immediate vicinity of a semiconductor P-N junction which becomes exhaused or "depleted" of current carriers, in order to set up the internal potential barrier involved in either the balance between diffusion and drift currents present in the equilibrium case, or the imbalance between these currents present in a non-equilibrium situation. Being depleted of carriers, the depletion layer region is virtually composed of "intrinsic" material, irrespective of the doping levels of the P-type and N-type materials from which it is formed.

DICE (singular DIE): Also called "chips," or "pellets." The tiny slivers of processed semiconductor material which constitute the functional heart of each semiconductor

device, whether discrete or an IC.

DIFFUSION, of carriers: The tendency of entities such as current carriers to "diffuse" themselves, or move in directions which increase the uniformity with which their number occupy the available space. Hence carrier diffusion is a mechanism whereby carriers tend to move "downhill" along concentration gradients, away from regions of high concentration and toward regions of low concentration.

DIFFUSION, of dopant atoms: One method of modifying the impurity doping of a semiconductor crystal, which makes use of the fact that excited dopant atoms, like carriers, have a tendency to diffuse away from regions of high concentration and toward regions of low concentration. The technique involves prolonged exposure of the semiconductor crystal wafer to a concentrated vapour of the dopant at elevated temperatures, whereupon dopant atoms diffuse into the crystal structure. The resulting doping gradient is roughly exponential, with highest density at the surface.

DIP: A dual-in-line package. One of the most common types of package used for integrated circuits. The package itself is rectangular, with connection pins emerging from the two longer sides in parallel rows. Made with from 4 to 64 connection pins, in many sizes.

DISCRETE DEVICE: An electronic circuit element or component which is individually packaged or encapsulated, in contrast with "integrated" devices in which a number of elements and their interconnections are housed in a single common package.

DONOR IMPURITY: An element or compound whose atoms or molecules have more valency electrons than those of the intrinsic semiconductor material into which they are introduced in small quantities as an impurity or dopant. Because the donor impurity possesses more valency electrons its inclusion in the crystal lattice creates an excess of valency electrons, so that material doped with a donor impurity is an N-type semiconductor.

DOPING: The process whereby the electrical characteristics of an intrinsic semiconductor material are altered by the addition of precisely controlled but relatively small amounts of selected impurity elements or compounds called dopants. The resultant impurity semiconductor tends to have a considerably higher conductivity than intrinsic material, to a degree depending upon the doping level, or amount of dopant added.

DRIFT CURRENT: The relatively small directional bias which becomes superimposed upon the random motion of carriers in an excited crystal lattice under the influence of an applied electric field ("drift field").

DYNAMIC RAM: A semiconductor read-write memory device whose storage cells require frequent periodic "refreshing" in order to retain the stored information. Generally the cells consist of single capacitors, in which the information is stored as a packet of electrical charge.

E-BEAM LITHOGRAPHY: The use of high energy electron beams in the fabrication of semiconductor devices. At present this generally means the use of electron beams in making the photolithographic masks used for wafer etching, diffusion, etc, but in the future electron beams will very likely be used directly for wafer modification.

EPITAXIAL DEPOSITION (EPITAXY): The technique of growing a semiconductor layer upon an existing crystal by depositing it directly from reacting vapours, so that the structure of the new layer is isomorphic with, or simply an extension of, that of the original crystal. The deposited or grown layer may be of either intrinsic or impurity semiconductor, and if the latter it tends to have a relatively constant doping density throughout its thickness. Probably the most common use of epitaxy is to produce the so-called "epitaxial" wafers for certain silicon bipolar transistors and integrated circuits, consisting of a relatively thick heavily doped substrate and a thin lightly doped epitaxial layer which ultimately forms the collector regions of the completed devices. The word "epitaxy" is apparently derived from the Greek words "epi" (upon) and "teinen" (arranged).

EPROM: An erasable, programmable read-only memory device. A read-only memory (ROM) whose stored information may be "erased" when no longer required, usually by irradiating the device chip with intense ultra-violet light. New information may then be stored or "burnt in", by manipulation of supply voltages.

EQUILIBRIUM: In a semiconductor context, equilibrium is that state of a semiconductor crystal which obtains when there is no nett current flow through the crystal. A crystal is normally in this state when no external voltages or current are impressed upon it.

EXCESS CARRIERS: Any carriers present in a semiconductor material or region, in addition to those present in equilibrium.

EXCITATION: That energy which is present in a crystalline material as a result of its dynamic interaction with the external environment. This includes the energy acquired by the material in the form of sound, heat, light and other forms of radiation.

FAMOS TRANSISTOR: A floating-gate avalanche-mode MOS transistor. Developed for use as a storage cell in EPROM memories, the FAMOS transistor has an unconnected gate. Charge is stored on the gate by causing a controlled avalanche breakdown to take place in the channel. Once charged, the gate remains charged until "erased" by irradiation with ultra-violet light.

FERMI-DIRAC DISTRIBUTION: A mathematical description of the way in which the current carriers present in a crystalline material have energies distributed above and below the Fermi level, this distribution being a function of the excitation of the material.

FERMI LEVEL: May be broadly defined as the average carrier energy level of a semiconductor region. Hence by definition a semiconductor crystal in equilibrium has a constant Fermi level throughout.

FIELD EFFECT: A term used to describe the way in which the effective dimensions of an impurity semiconductor region are dependent upon the width of the depletion layers of adjacent P-N junctions. Hence if the depletion layers are widened due to increased reverse bias on the junctions, the growth of the "intrinsic" depletion layers inevitably reduces the effective dimensions of the impurity region. Conversely if the depletion layers are narrowed, the effective dimensions of the impurity region tend to increase.

FIELD-EFFECT TRANSISTOR (FET): A three-terminal active device whose operation relies upon the field effect. Generally has a semiconductor region called a "channel", whose conductance is effectively controlled by external manipulation of one or more depletion layers, using a "gate" electrode.

FLOAT ZONE REFINING: A modification of the zone refining process, in which the ingot of material being refined is supported vertically between two chucks. The zone of molten material is supported by its own surface tension, preventing contamination due to reaction with a crucible.

FORWARD BIAS: That polarity of external voltage applied to a semiconductor P-N junction which tends to counteract the internal potential barrier set up in equilibrium, resulting in a marked increase in the diffusion currents crossing the junction.

HEADER: That part of a semiconductor device package to which the actual chip or die is mounted. May consist of metal, ceramic or one of a number of plastics such as epoxy resin.

HOMOGENEOUS CRYSTAL: Crystalline material having a uniform composition. In the context of impurity semiconductor materials, a homogeneous crystal is one having a uniform doping concentration.

HOLE: A defect in the valency electron system of a semiconductor crystal lattice, equivalent to the absence of a single valence electron. Like a conduction electron, a hole is capable of moving through the crystal, and thus forms an effective current carrier having a positive charge. However, unlike a conduction electron the hole must remain in the valency bonding system of the crystal, and thus it has a lower mobility.

HYBRID CIRCUITS: Are strictly circuits, or microcircuits, which are fabricated using a mixture of techniques. More typically, a hybrid circuit is one consisting of a number of monolithic chips, or dice, mounted on a common header and with connections made using either fine gold wires or metallic film conductors.

IMPURITY: A "foreign" material present in a semiconductor material, usually in small quantities. Some impurities are unwanted,

and great pains are taken to extract them from the material. Others are intentionally added in small quantities to semiconductor material as dopants, in order to modify its electrical behaviour.

INJECTION OF CARRIERS: The introduction of excess carriers into a semiconductor region. This is often performed by means of a forward biased P-N junction.

INSULATOR: Any material whose valency energy band is completely filled with electrons, so that no empty levels are immediately available to facilitate a net electron movement. Such materials conduct electricity only when excited sufficiently to raise electrons into the higher conduction bands. Intrinsic semiconductors are strictly insulators, differing from "true" insulators only in that they possess a somewhat smaller "forbidden" energy gap separating the valency and conduction bands.

INTEGRATED CIRCUIT (IC): Strictly, this term simply refers to any circuit in which the component elements and wiring are grouped together within a common protective container or encapsulation. However, the term has become established as a synonym for "microcircuit," so that in practice it invariably refers to miniature integrated circuit devices.

INTRINSIC SEMICONDUCTOR: An element or compound which has the same electron energy band configuration as an insulator, but has a "forbidden energy gap" which is sufficiently narrow to permit transfer of electrons from the valency band to the conduction bands at normal temperatures. Conduction in an intrinsic semiconductor takes place via equal number of conduction band electrons and valence band holes.

INTRINSIC CARRIER GENERATION: The mechanism whereby excitation energy absorbed by a semiconductor crystal lattice causes a valency band electron to be raised into a conduction band, reacting a valency band hole carrier in addition to a conduction band electron carrier.

ION: An atom or molecule which is electrically charged, having lost or gained an electron. An atom which has gained additional electrons is thus a negative ion, while an atom which has lost an electron or electrons is a positive ion.

ION IMPLANTATION: A semiconductor fabrication technique in which a high-energy ion beam is used to implant impurity atoms in the semiconductor wafers. The technique may eventually supersede diffusion, being capable of greater resolution.

JUNCTION FET: A field-effect transistor in which reverse-biased P-N junctions are used to isolate the gate electrode from the channel, and also to provide the controlling depletion layers.

JUNCTION, P-N: A relatively abrupt transition between P-type and N-type semiconductor regions, within a crystal lattice. Such a junction possesses unique electrical properties, including the ability to conduct substantially in only one direction. Single and multiple P-N junctions form the basis for many semiconductor devices.

LEAKAGE CURRENTS: Those currents passed by a semiconductor device whose origin lies in spurious contamination of the crystal die, usually at its surface. Most modern semiconductor devices exhibit very low leakage current levels, but only because of extremely rigorous controls maintained during their fabrication.

MAGNETIC BUBBLE MEMORY: A solid state memory device in which information is stored in tiny magnetic domains or "bubbles", which are manipulated in a thin substrate of magnetic garnet. Basically a serial memory, with the potential for storing very large amounts of information.

MAJORITY CARRIERS. Those carriers in an impurity semiconductor material which are at least nominally in the majority. Hence in N-type material conduction band electrons are the majority carriers, whereas in P-type material the majority carriers are valency band holes.

MASK: This term is used to describe two different, but related things which are both involved in semiconductor device fabrication: (a) the master image plates used to expose the photoresist on the wafers, during photolithography; (b) the etched silicon dioxide layer, after lithography, which is then used to control impurity doping during diffusion or ion implantation.

MICROCIRCUIT: A complex semiconductor device consisting of a miniature assembly of component elements and their interconnections. This term is a general one and includes devices of the monolithic, thin-film and hybrid variety.

MINORITY CARRIERS: Those carriers in an impurity semiconductor material which are at least nominally in the minority. In N-type material valency band holes are the minority carriers, whereas the minority carriers in P-type material are conduction band electrons.

MNOS TRANSISTOR: Metal-nitride-oxide-semiconductor transistor, used for information storage in non-volatile static RAM memory devices.

MOBILITY: The facility with which a current carrier can move within a medium such as a semiconductor crystal lattice, under the influence of an electric field. Normally expressed in terms of the average drift velocity attained by the type of carrier concerned, per unit electric field intensity.

MONOCRYSTAL: A crystal of material which has a continuous lattice structure and orientation throughout its volume, in contrast with the multiple-grain structure of a polycrystal. Almost all semiconductor devices are fabricated from monocrystalline material.

MONOLITHIC CIRCUITS: Are circuits, or more usually microcircuits, in which all component elements and their interconnections are fabricated as patterns of P-type, N-type and instrinsic regions within a single chip of semiconductor crystal. The term "monolithic" is derived from the Greek words "mono" (single) and "lithos" (stone), and thus has the literal meaning "single stone."

MOS TRANSISTOR: A variety of field effect transistor in which the gate is a metallic electrode, isolated from the channel by the silicon dioxide surface passivation.

N-TYPE SEMICONDUCTOR: Impurity semiconductor material containing a predominance of donor dopants, and in which conduction band electrons normally form the principal current carriers.

OHMIC CONTACT: An electrical connection which passes current linearly in both directions. In the context of semiconductor device design, an ohmic contact to a semiconductor crystal is one expressly designed so that it does not possess any of the unilateral properties of a normal metal-semiconductor or P-type/N-type semiconductor junction. Usually all exterior electrode connections to a device chip are made by means of ohmic contacts.

PARAMETERS: Those indicators of device performance which relate one aspect of its behaviour with another. Hence input resistance is a parameter relating input voltage with input current, and current gain a parameter relating output current with input current.

PARAMETER SPREAD: The inevitable variation in value of the parameters of a given device type, due to manufacturing tolerances. Often expressed in terms of the various statistical distributions.

PASSIVATION: The technique of providing a semiconductor device chip with an isolating layer or "skin" which protects it from contamination by unwanted impurity atoms or molecules. With silicon devices, the isolating layer is usually composed of silicon dioxide (quartz) or silicon nitride, grown on the chip at a high temperature.

PHOTOLITHOGRAPHY: The process whereby patterns are etched into an oxide or similarly passive layer coating a semiconductor crystal wafer, using a photoresist process followed by an etchant. The remaining oxide material thus forms a precisely located miniature mask, used to control impurity doping or contact metallisation.

PLANAR: A semiconductor fabrication technique developed by J. Hoerni, of Fairchild Semiconductor, in 1960, and in which the semiconductor device chips are protected by an oxide passivation layer throughout the various stages of fabrication. The Planar process thus represents a synthesis of the separate oxide layer functions involved in the photolithographic etching of diffusion masking, and in chip passivation.

POPULATION INVERSION: In the context of semiconductors, this term describes any situation in which the normal majority/minority carrier ratio of an impurity semiconductor region is disturbed, to a degree such that the nominal "minority" carriers are actually present in larger numbers than the nominal "majority" carriers for that material.

P-TYPE SEMICONDUCTOR: Impurity semiconductor material containing a predominance of acceptor dopants, and in which valency band holes normally form the principal current carriers.

RAM: A random-access memory device. Usually, a random-access memory device of the read-write type, in contrast with a read-only memory (ROM).

RECOMBINATION: A "collision," within a semiconductor crystal lattice, between a conduction band electron and a valency band hole. The ability of each to function as a current carrier is lost, due to mutual cancellation, so that a recombination effectively "destroys" the hole-electron carrier pair.

RESISTIVITY: That parameter of a material which indicates the extent to which it resists the flow of a net electrical current, and hence the inverse of its conductivity. Resistivity is normally defined in terms of the resistance in ohms between opposite faces of a cube of the material.

REVERSE BIAS: That polarity of external voltage applied to a semiconductor PN junction which tends to reinforce the internal potential barrier set up in equilibrium, resulting in either a marked reduction or complete extinction of the diffusion currents.

SATURATION CURRENTS: Those currents passed by a reverse biased semiconductor P-N junction which are composed of minority carriers drifting across the potential barrier of the depletion layer. The term "saturation" is used because in material which is even moderately doped the number of minority carriers present in the material is almost solely determined by the lattice excitation, so that once all the minority carriers available at a given excitation level are involved in the reverse conduction, further increases in reverse bias produce virtually no increase in current.

SATURATION, of a device: Is generally that state in which the device is conducting most heavily for a given applied voltage. In many devices it is also a state in which the normal amplification mechanisms have become "swamped," and inoperative.

SAW DEVICE: A device whose operation relies upon the generation, manipulation and detection of surface acoustic waves.

SCHOTTKY DIODE: A semiconductor diode utilising the properties of a metal-semiconductor junction. The forward conduction involves mainly conduction band electrons from the metal, leading to the alternative name "hot carrier diode".

SEGREGATION: The phenomenon whereby a solute material such as an impurity in a semiconductor, exhibits a greater solubility in the solvent material when the latter is in the liquid form, than when it is in the solid form. Hence a crystal grown from a liquid solution containing a certain impurity concentration tends to have a lower impurity concentration, due to the differential solubility. The ratio between the concentrations in solid and liquid phases is known as the segregation or distribution coefficient. Many techniques of semiconductor purification rely heavily upon the segregation effect.

STATIC RAM: A random-access read-write semiconductor memory in which the storage cells are capable of retaining their stored information while ever the supply to the device is maintained. In other words, they do not need "refreshing" like those of a dynamic RAM device. In most static RAM devices the memory cells are flipflop latches using either MOS or bipolar transistors.

SUBSTRATE: The base or support layer of a transistor or monolithic microcircuit chip, which usually constitutes a major proportion of the total volume. When composed of ceramic, glass or sapphire, the substrate functions mainly as a support during the operations of fabrication and encapsulation. However, when composed of heavily doped semiconductor material it normally performs the additional furnction of a distributed low resistance connection to the physically lowest region of the device.

THIN-FILM CIRCUITS: Are circuits, usually microcircuits, in which the component element and interconnections are fabricated from thin deposited films of metal, semiconductor and dielectric materials, generally upon an insulating substrate such as ceramic or sapphire. The term "thin" is usually taken to imply films having a thickness in the order of 1 micron (micrometre).

TRANSCONDUCTANCE: That parameter of an active device which relates a change in output current to the change in input voltage producing it. Normally expressed in milliamps per volt (mA/V).

WAFER: The thin slice of semiconductor crystal, usually some 7.0 or more square inches in area, from which many hundreds of single monolithic device chips are ultimately obtained. Normally all techniques such as epitaxy, photolithography, diffusion and passivation are carried out on the wafer, before it is scribed and broken into individual dice.

X-RAY LITHOGRAPHY: The use of X-rays rather than light to expose the photoresist used in the preparation of etching and diffusion masks for semiconductor device wafers, during fabrication. X-ray lithography is likely to be increasingly used in the future, having greater resolution than optical and electron beam techniques.

ZENER BREAKDOWN: One of the mechanisms responsible for voltage "breakdown" of semiconductor junctions and devices. When breakdown occurs, the electric field intensity in the material has become so great that electrons are effectively "ripped" from the valency bonding system. Another name for this mechanism is "field emission." Providing the current increase which tends to occur is limited externally, zener breakdown causes no permanent damage.

ZENER DIODE: A general term used to describe any semiconductor diode intended to be operated in the reverse biased breakdown condition. Low voltage devices of this type do in fact exploit the zener breakdown mechanism, but with most devices having a breakdown voltage above about 6V, breakdown is in fact due to the avalanche mechanism.

ZONE REFINING: A technique used to reduce the impurity content of raw semiconductor materials to an extremely low level, relying upon the phenomenon of segregation. A zone of molten material is swept repeatedly through the ingot in the same direction, "collecting" the impurities.

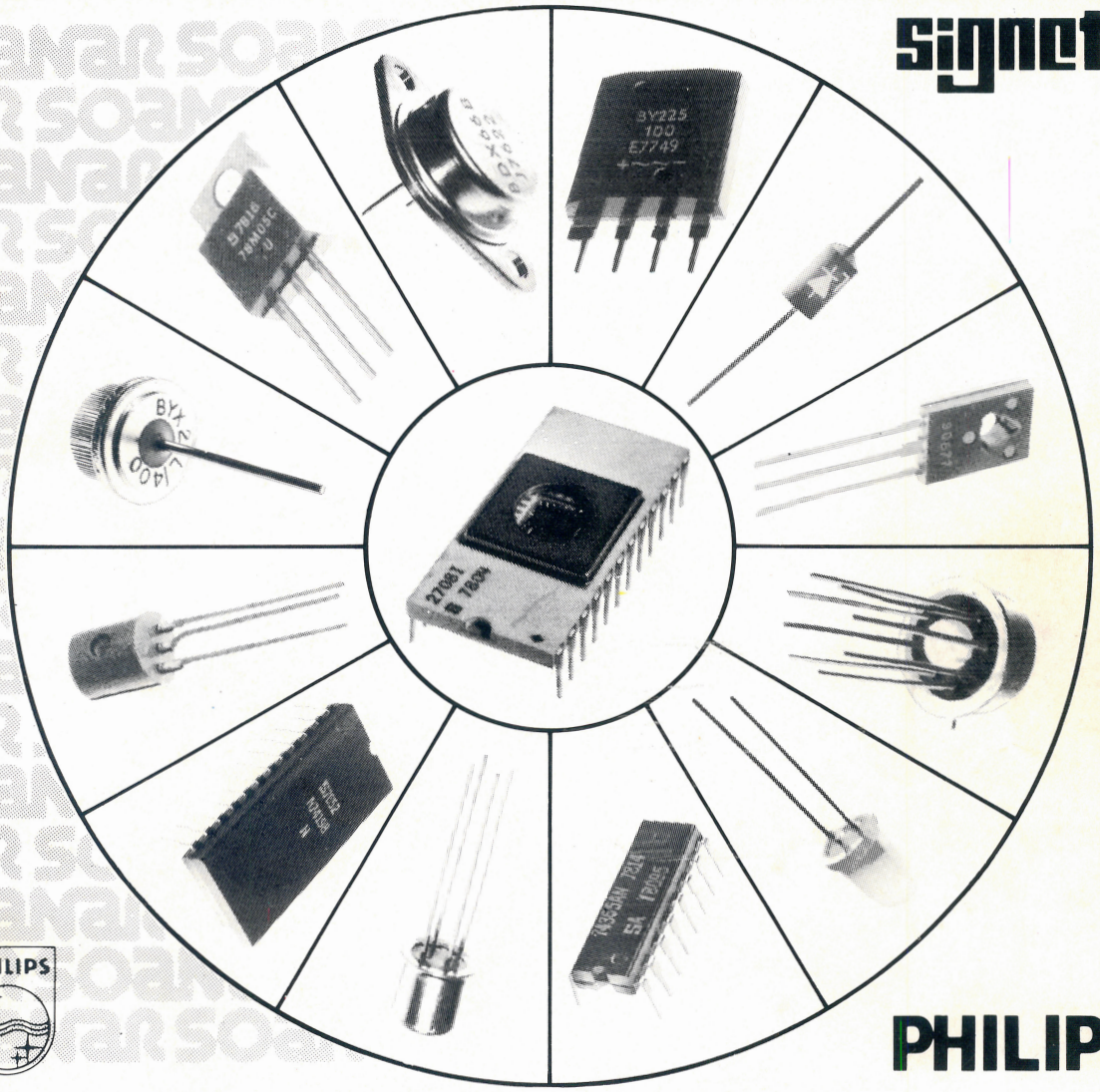
Index

Index — continued

NOW!

READILY AVAILABLE IN ALL STATES

signetics



PHILIPS

Transistors – Zeners – Rectifiers – LED's – Heatsinks
Solar Panels – Locmos – I.C's – Memories – Interface.

Microprocessors
Linear
Logic
Consumer

Write or phone for full technical data.

SOANAR GROUP

SOANAR

Soanar Electronics Pty Ltd
A member of the A & R Soanar Electronics Group
30 Lexton Road, Box Hill, Vic., 3128. Australia

SALES OFFICES	S. AUST.: 51 6981
VICTORIA: 89 0661	QUEENSLAND: 52 5421
N.S.W.: 789 6733	W. AUST.: 381 5500

LET'S FACE IT...

You only need to make one stop for all your electronic and hobby needs . . .

DICK SMITH ELECTRONICS

SYDNEY 125 York Street, SYDNEY. Ph. 290-3377
147 Hume Hwy, CHULLORA. Ph. 642-8922
162 Pacific Hwy, GORE HILL. Ph. 439-5311
30 Grose Street, PARRAMATTA. Ph. 683-1133

MELBOURNE 399 Lonsdale Street, MELBOURNE. Ph. 67-9834
656 Bridge Road, RICHMOND. **Ph. 428 1614**

BRISBANE 166 Logan Road, BURANDA. Ph. 391-6233

ADELAIDE 203 Wright Street, ADELAIDE. Ph. 212-1962

MAIL ORDERS P.O. Box 747, Crows Nest, N.S.W. 2065. Post and packing extra.

Dealers across Australia.

SHOPS OPEN 9AM to 5.30PM
(Saturday 9AM – 12 noon)
BRISBANE 1/2 Hour earlier

• ANY TERMS OFFERED ARE TO APPROVED APPLICANTS ONLY

Many lines available from the Dick Smith Electronics Centres at:

bankcard welcome here

DSE474