DEPARTMENT OF WEAPONS TRAINING
LOWRY AIR FORCE BASE
COLORADO

T R A N S I S T O R S

IN-SERVICE TRAINING COURSE

INTRODUCTION TO SEMICONDUCTOR DEVICES

CHAPTER 1

22 NOVEMBER 1960

COPY ____ OF 100 COPIES

FOR INSTRUCTIONAL PURPOSES ONLY

CHAPTER 1

INTRODUCTION TO SEMICONDUCTOR DEVICES

Author: A/1C Robert J. Widlar

Abstract -- The idea of conduction in solids by both holes and electrons is introduced. The existence of two distinct types of current carriers is demonstrated using the Hall effect. Without explaining the origin of these current carriers, an elementary description of junction diodes and two-junction transistors is given. Finally, it is shown how a transistor can be used in an amplifier circuit.

INTRODUCTION

It is difficult to find a piece of modern electronic equipment that does not incorporate some kind of semiconductor device, such as; a selenium rectifier, a crystal diode, or a transistor. This is true of almost all commercial, industrial, and military equipment because these semiconductor devices are generally smaller, more efficient and in some cases, more reliable than components used in the past.

The existence of semiconductors has been known for many years; but, until recently, they were considered a useless transition between conductors and insulators. They first enjoyed widespread use in the early days of radio when it was discovered that the contact between a fine wire and some substances - such as galina, iron pyrite, and silicon carbide - exhibited rectifying properties. The reason for this effect was not known. Therefore, vacuum tubes pushed semiconductors into the background, because conduction in a vacuum was well understood; and an intelligent approach could be used in the design and perfection of tubes.

The influx of radar and other complex electronic systems during World War II spurred the development of semiconductor diodes; but, because of the pressing need, an experimental approach was used. This effort resulted in practical silicon and selenium rectifiers; however, it did not provide a major breakthrough, as the phenomenon of conduction in solids was still not completely understood. The major advance came in 1948 with the development of the transistor, a semiconductor amplifying device. It appeared that the transistor could perform the same function as a vacuum tube, while using one hundredth the power and occupying far less space. This added impetus to semiconductor research, and workable theories on the operation of semiconductors were evolved.

The knowledge gained from the research on transistors was applied to other devices, and soon silicon and germanium diodes were made with ratings exceeding those of the best copper oxide and selenium rectifiers. Also, the silicon solar cell was developed which could convert the energy of the sun directly into electricity with reasonable efficiency. Diode amplifiers for example, the varactor and the tunnel diode were made which could perform at frequencies above the range of transistors. These and many other devices benefited from semiconductor research.

It can be seen that semiconductors have assumed an important position in the electronics industry. ~~They are being used with increasing frequency because some semiconductor device might be able to replace a whole circuit of conventional~~ components. It is therefore necessary for anyone connected with electronics to become familiar with semiconductor theory just as he was required to learn the theory of vacuum tubes.

One of the most important fundamentals in the theory of semiconductors is the mechanism of conduction in solids. But in order to understand this phenomenon, it is also necessary to learn the structure of solids. The chemistry and physics of a few important semiconductor materials will be discussed in chapter two.

Furthermore, the characteristics of these substances important to the understanding of practical semiconductor devices will be brought forth. In later chapters, the operation of transistors and other PN junction devices will be given a detailed explanation. Emphasis will be placed on the junction diode and the two-junction transistor because these will illustrate the more significant results that can be realized with semiconductors. Still later, the techniques used in the manufacture of these devices will be introduced; and most of the processes currently used in the production of diodes and transistors will be briefly described. The last chapter will be devoted primarily to the circuit applications of transistors although other devices of current interest will be mentioned.

Before going into a more exact explanation of semiconductor phenomenon, a beief preview of conduction in solids will be given and then related to the operation of the junction diode and the transistor. The purpose of this is to give the reader an opportunity to get an overall view of the material to be covered. Exact explanations and detailed proofs will not be used here.

CONDUCTION IN SEMICONDUCTORS

Semiconductors are a class of materials having conductivities somewhere between those of metalic conductors and insulators. The conductivity of these materials is several orders of magnitude greater than insulators but still very much less than metalic conductors. A simplified picture of conduction in metals attributes conduction to the existence of free electrons which can be moved under the influence of an electric field. It then seems reasonable to assume that free electrons are also responsible for conduction in semiconductors. This is the case for some semiconductor materials, but it has been shown that conduction can also take place by what appears to be a positive electron, or hole. The properties of semiconductors of concern at this point are that the current carriers can be either holes or electrons and that the concentration and type of current carrier can be controlled during production of the material. This much will be assumed here, but to help substantiate these statements a demonstration of the existence of holes follows:

It was found that the current carriers in semiconductors traveled at greater velocities than those in metals. It was thought, then, that these current carriers could be deflected appreciably from their normal path within the semiconductor by the application of a magnetic field. This was indeed the case as is shown in Figure 1.1. A current was passed between two metalic contacts on a block of semiconductor material. Two other electrodes were placed at right angles to the current flow as can be seen from the figure. If the current carriers were electrons, the results shown in Figure 1.1a. could be expected; with no magnetic field applied, the number of electrons reaching the two electrodes, at right angles to the current flow would be equal, and there would be no potential difference between these electrodes. When a magnetic field is applied into the page as shown, the electrons would be deflected upward; and more electrons would reach the upper contact than would reach the lower. The upper contact would then become more negative than the lower, and a potential difference could be measured between them. This was found to be the case for some semiconductor materials. However, with other materials, the opposite effect was observed: When the magnetic field was applied in the same direction, relative to current flow, the upper contact became positive with respect to the lower. This could not happen with electron current carriers. It was then assumed

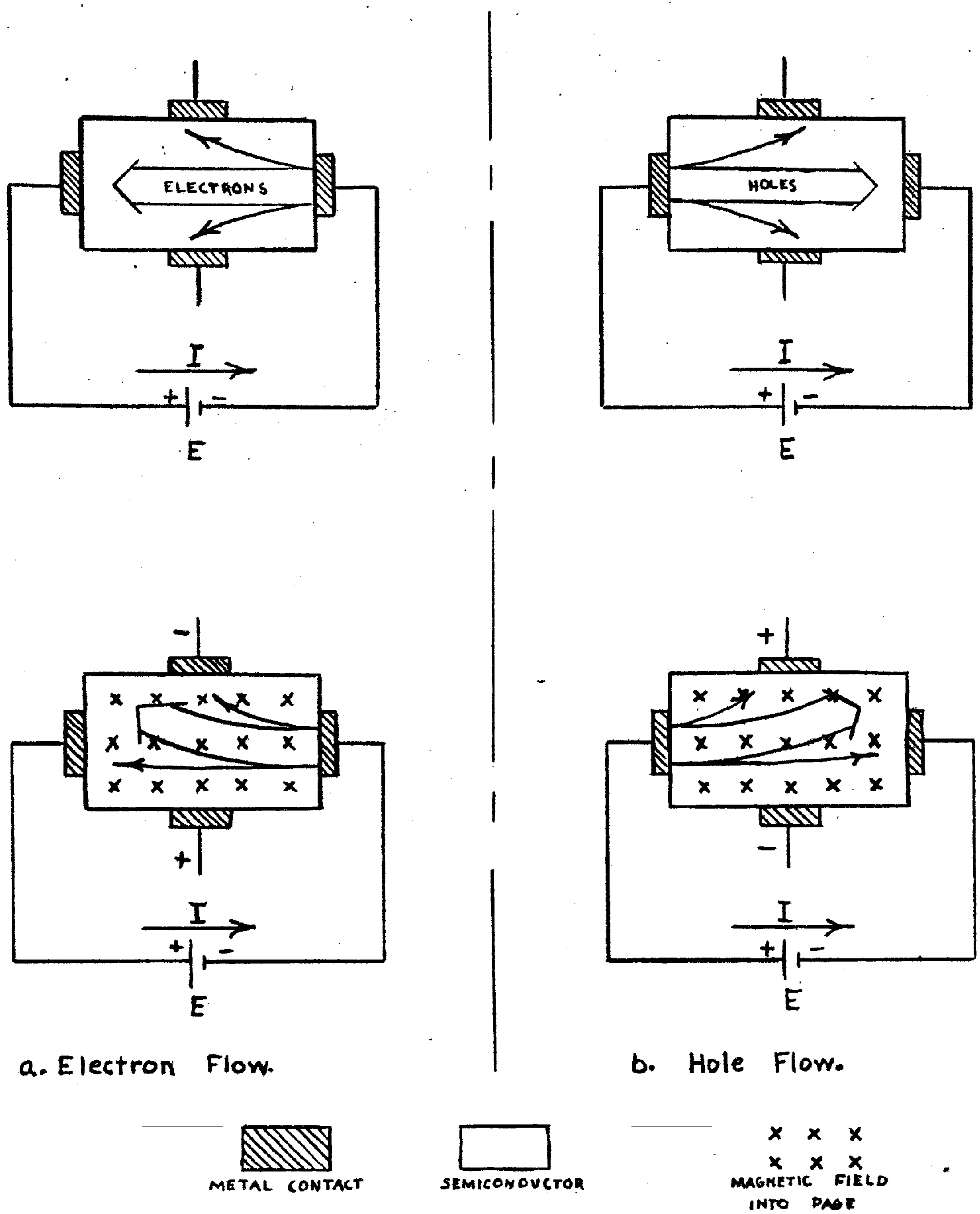


Figure 1.1. Using the Hall Effect to Demonstrate Existence of Holes.

that there were two distinct types of current carriers in semiconductors, electrons and holes. This would explain the effect observed in Figure 1.1b.

This is known as the Hall effect, and in practice it is used to measure steady magnetic fields for the detection of submarines; for tape recorder heads; and in circulators, gyrators, and isolators where the direction of current flow is altered by a magnetic field.

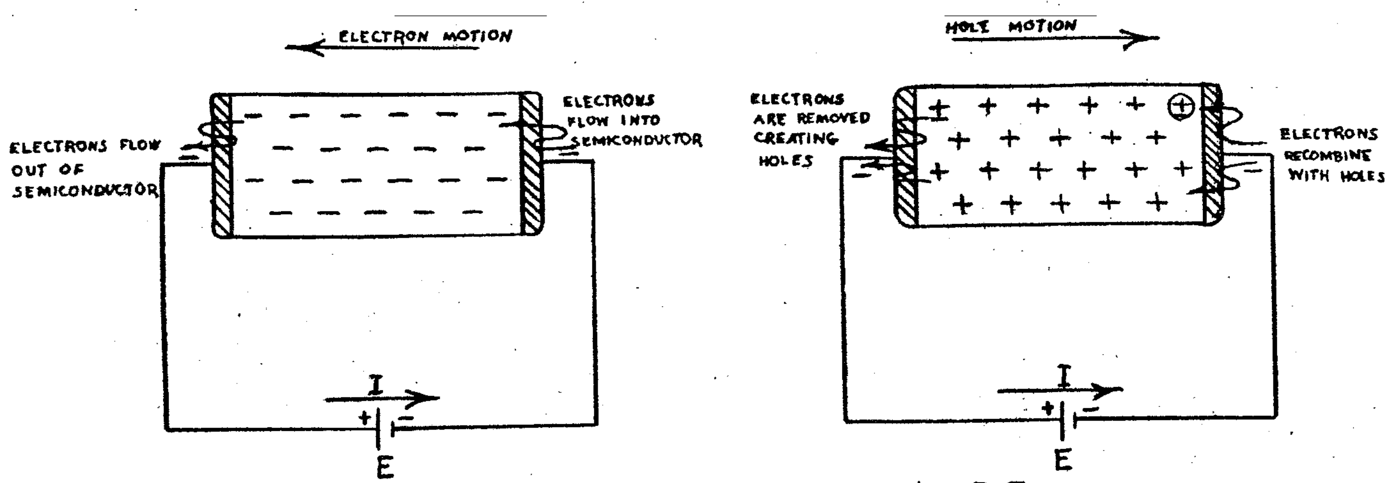
When the current carriers in a semiconductor are free electrons (Figure 1.2a) conduction is relatively easy to visualize. When a voltage is applied across the semiconductor, the electrons will move from negative to positive and into the external circuit. When an electron passes out of the semiconductor into the external circuit at the positive terminal, another electron immediately flows into the semiconductor at the negative terminal.

When holes carry the current in a semiconductor, there must be a transition at the circuit connections. This is because the current carriers in the external circuit are electrons while the current carriers in the semiconductor are holes. If an electric field is applied, the holes will flow from positive to negative until they reach the negative terminal as shown in Figure 1.2b. When a hole does reach the negative terminal, it captures a free electron from the external circuit and the hole disappears. At the same time the external circuit recovers an electron from the positive terminal, creating another hole. This new hole will travel to the negative terminal repeating the process. By this mechanism, current is carried by holes within the semiconductor, and by electrons in the metalic conductors of the external circuit.

Some helpful rules for predicting conduction in semiconductors will be stated here without proof. Whenever an electron is removed from a semiconductor under the influence of an applied electric field, it must soon be replaced by an electron from the potential source (external circuit) to keep the overall number of electrons constant. Similarly, whenever a hole is removed from a semiconductor by recombination with an external electron, the hole must be replaced by the removal of another electron to keep the overall number of holes constant. To appreciably change the number of holes or electrons would require high electric potentials which are never encountered in practice. Another detail concerns the production of holes by the removal of electrons: This can only occur to any appreciable extent at the terminals (contacts) of the semiconductor. Hole generation occurs at imperfections in the physical structure, and there are normally very few imperfections within the semiconductor. However, the surface conditions resulting from attaching a contact create many imperfections. Hole generation can therefore take place far more readily at the contacts than it can within the semiconductor.

JUNCTION DIODES

It can be seen then, that there are two distinct classes of semiconductors: N type in which the current carriers are free electrons, and P type in which the current carriers are equivalent to positive electrons (holes). This fact can be used to build many useful devices, one of which is the junction diode.



a. N Type. b. P Type.

Figure 1.2. Representations of P and N Type Semiconductors Indicating Mechanism of Conduction.

A junction diode is made by somehow joining together P and N type semiconductor materials. A diagram of a junction diode is shown in Figure 1.3. Contacts are made to both the P and N type materials so that the device can be connected to an external circuit. A device constructed in this manner will exhibit a low resistance to current flow in one direction and a high resistance in the other. A mechanical view of this action follows:

When P and N type semiconductors are joined, a junction is formed. The current carriers on one side of this junction are electrons, and those on the other are holes, as can be seen from Figure 1.3. When a voltage of the polarity shown in Figure 1.4 is applied across the diode, it will conduct. This is because the electrons and holes will be forced across the junction in opposite directions, establishing a current within the diode and also in the external circuit. More exactly, the applied voltage will move the electrons from negative to positive, across the junction, and into the P region. Once the electrons flow into the P region, two things can happen: The electrons can move through this region and into the external circuit via the positive diode contact; or they can recombine with the holes in the P region. If an electron does recombine with one of the holes, another electron will be drawn out of the P region, at the positive terminal, and flow into the external circuit; still another electron will flow into the diode at the negative terminal, to keep the overall number of holes and electrons constant in the P and N type regions. In either case, a current flow is established through the diode by the electrons.

Similarly, the holes will also contribute to current flow. Under the influence of the applied voltage, the holes will move across the junction into the N region. After crossing the junction, again two things can happen; the holes can move through the N region and recombine with electrons at the negative terminal of the diode, or they can recombine with electrons within the N region. In either case, when a hole recombines with an electron, another hole will be created at the positive terminal by the removal of an electron to the external circuit; and another electron will flow in at the negative terminal again to keep the overall number of holes and electrons constant. This action establishes a current through the diode by hole conduction.

It can be seen that a current will flow across the junction by both electron and hole conduction. Consequently, current will flow in the external circuit. The polarity of applied voltage that will cause this conduction is called forward or conducting bias.

If a potential of the polarity shown in Figure 1.5 is applied across the junction diode, no current will flow. The holes and electrons will be drawn away from the junction by the applied voltage, and no current carriers will flow across the junction. It can be seen from Figure 1.5 that the electrons will be influenced by the applied voltage to move away from the junction and toward the positive contact of the diode. However, a continuous current flow cannot be maintained by

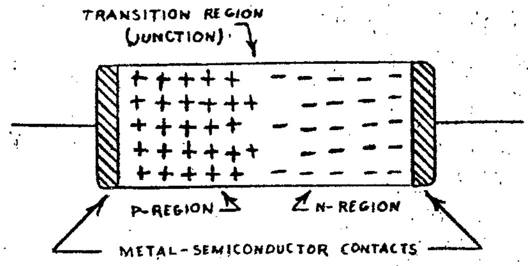


Figure 1.3. Diagram of PN Junction Diode.

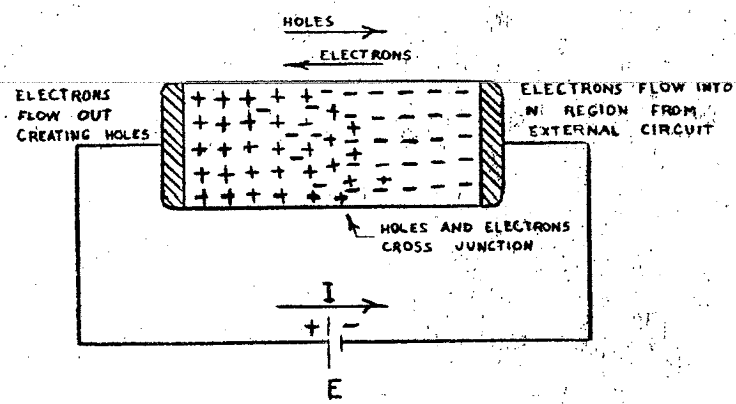


Figure 1.4. Action in a Junction Diode for Forward Bias.

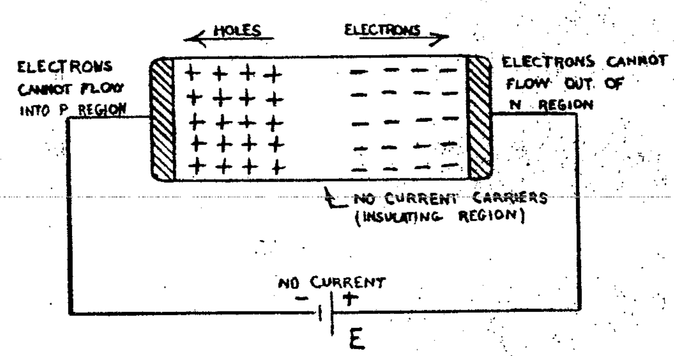


Figure 1.5. Action in a Junction Diode for Reverse Bias.

the electrons. If the electrons did flow out the positive terminal, they could not be replaced by electrons from the P region because there are no free electrons present.[1] Therefore, no appreciable amount of electrons can flow out the positive terminal without greatly reducing the number of electrons within the semiconductor.

A similar situation exists for the holes. The holes are moved away from the junction, and toward the negative terminal, under the influence of the applied potential. But again, the holes cannot recombine with electrons at the negative terminal to any appreciable extent. This would greatly reduce the number of holes because the holes cannot be replaced from the N region. Therefore, no current can be maintained through the diode by hole flow.

To summarize the description of a nonconducting P-N junction; When the voltage is first applied, a negligible current will flow. This makes the internal potential of the diode equal to the applied potential, by removing a minute quantity of holes and electrons. However, a continuous current cannot flow in the external circuit because there cannot be current flow within the diode: the holes and electrons cannot cross the junction. The holes and electrons will actually be pulled away from the junction forming an insulating region in which there are no current carriers. This condition is known as reverse bias.

The junction diode is used quite frequently in practice. It can replace vacuum tubes in a wide variety of circuits. This diode requires no filament power and physically, it is much smaller than equivalent vacuum diodes. The electrical characteristics of these diodes leaves little to be desired; the voltage drop across a typical junction diode is approximately 0.5 to 1.0 volt at rated current. This is much less than the 50 to 200 volt drop across high vacuum rectifiers and the 5 to 50 volt drop across gas diodes. This means that a small semiconductor diode functioning as a power rectifier can produce more D-C power at a higher efficiency (99 + percent) than can its larger vacuum counterparts. Semiconductor diodes are also used in low power and high frequency applications, and they are unchallenged in computer applications where thousands of diodes must be crowded into a small space.

Semiconductor diodes are made with a wide range of voltage and current ratings. Diodes are available with current capacities ranging from 10 ma to 500 amperes, and inverse voltage ratings of 10 to 1000 volts.

These diodes do have a small reverse current, but in most circuits it is of negligible magnitude; a fraction of a microampere for low current silicon diodes and a few milliamperes for high current germanium power rectifiers.

[1]In addition, the generation of appreciable numbers of holes and electrons cannot occur at the junction (or within the semiconductor) if it is properly made because there are few imperfections in the material.

TRANSISTORS

Although the development of the junction diode was an important contribution to the field of electronics, it represents only one component in a large family of solid-state devices. Another member of this family is the transistor, a semiconductor amplifying device. The operation of the transistor is closely related to the junction diode.

The first transistors made were point contact devices, but almost all of these are now obsolete. Hence, only junction transistors will be discussed here as they are by far the most popular type, both in theoretical considerations and in practical usage.

The physical construction of a PNP junction transistor is shown in Figure 1.6. A thin N type semiconductor wager is sandwiched between two larger P type slabs. Contacts are made to these three regions, and they form the basic elements of the transistor; emitter, base, and collector. The drawing is not made to scale; the base region is much thinner than is shown in the drawing. In a typical unit the cross section of the base region is 1/4 inch square while its thickness is only 0.001 inch. The emitter and collector are therefore separated by an extremely thin N type region. As will be seen, this is essential for efficient transistor operation.

As is indicated on the diagram, the current-carrier densities in the three regions are not alike. The emitter is made from a very high conductivity semiconductor. This means that there is a relatively large number of holes available for conduction. The base is a low conductivity semiconductor in that there are relatively few electrons available for conduction. The collector is a moderate conductivity P type material. The concentration of current-carriers in the collector is not too important for an elementary discussion. However, the high ratio of carriers between the emitter and base regions is essential as will be seen.

The collector-base and emitter-base junctions satisfy the requirements previously set down for a junction diode in that P and N type materials are joined together. The behavior of these junctions will indeed be similar to that of a diode, but they will be put to a somewhat different use.

In Figure 1.7 a voltage is applied to the transistor, negative on the collector and positive on the emitter. Under these conditions, the collector-base junction will behave like a reverse-biased diode: the holes and electrons will be pulled away from the collector junction, and no current will flow. This voltage is of the correct polarity to forward bias the emitter-base junction, but the emitter junction and collector junction are electrically in series. Hence, all the applied voltage appears across the collector junction and there is none left to forward bias the emitter junction. Under these circumstances, no current flows in the transistor; and it is said to be cut off.

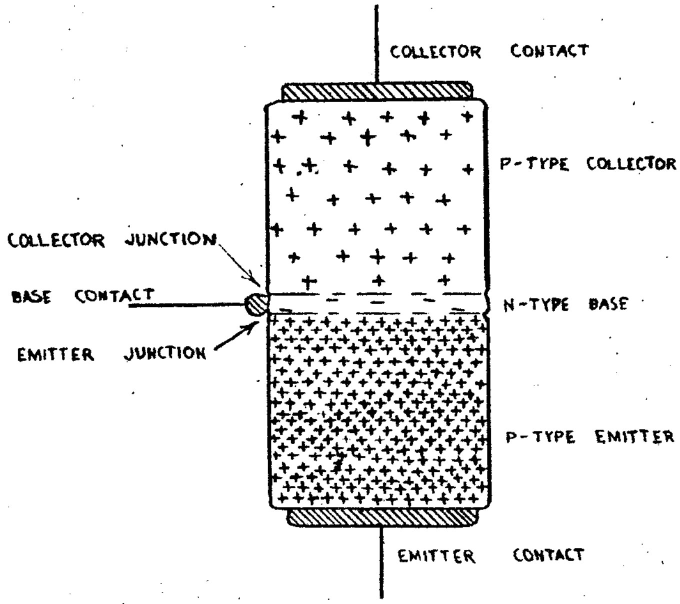


Figure 1.6. Diagram Showing Construction of PNP Transistor.

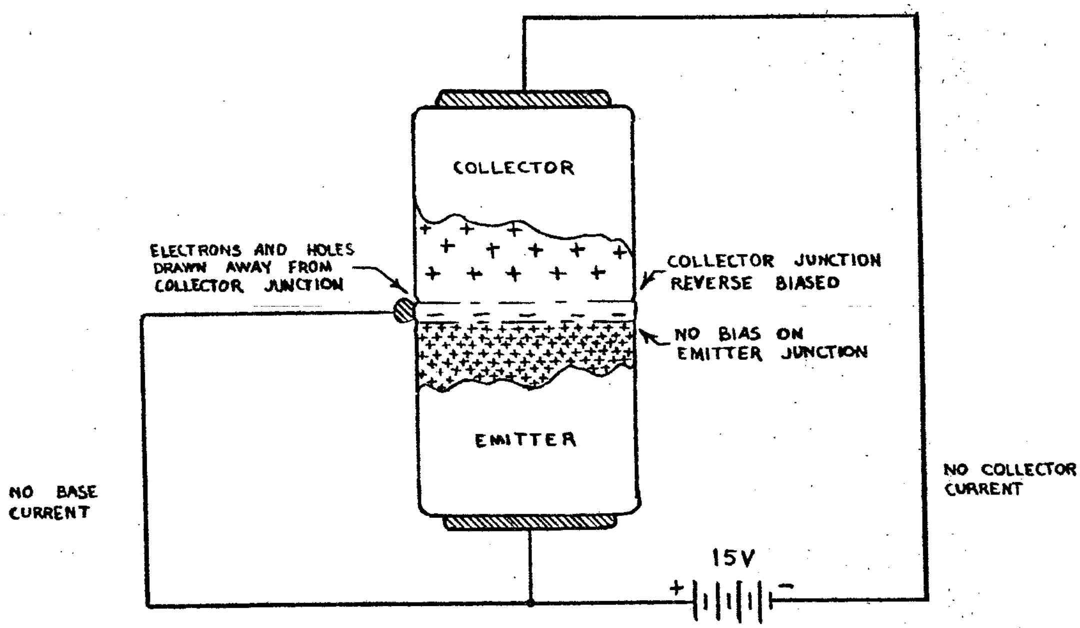


Figure 1.7. PNP Transistor in Nonconducting State.

If a forward bias is applied to the emitter junction, as in Figure 1.8, the amplifying action of a transistor becomes apparent. Feeding 0.2 volt at 1.0 ma into the base caused the collector current to increase from 0 to 50 ma. Furthermore, small changes in base current will result in correspondingly large changes in collector current. For example, if the base current is increased to 2.0 ma (approximately 0.4 volts base voltage) the collector current will increase to 100 ma. The base appears to exercise control over the collector current with a "current gain" of 50. The reasons for this must be investigated.

The amplifying action of a transistor takes place in a relatively small volume including the base and the two junctions. In Figure 1.9 this pertinent area has been redrawn in an expanded view to show this action more clearly.

When a forward bias is applied to the emitter junction, holes will flow across the junction from the emitter to the base; and electrons will flow from the base to emitter, as would be expected from the discussion of junction diodes. Here the similarity to diodes ends. After the holes flow across the emitter junction, there is a force acting on them that has not yet been brought forth. This force is the mutual repulsion that the holes have for each other because of their like charge.[1] When the holes are spilled across the emitter junction, they bunch up on the base side of the junction. Mutual repulsion will cause these holes to diffuse into the base. Because of the extreme thinness of the base, most of the holes will diffuse to the collector junction before they can recombine with electrons in the base or reach the base terminal. Once the holes reach the collector junction they are swept away into the collector by the negative collector voltage. When these holes flow into the collector, they will flow through the collector to the negative terminal. Thus, the forward bias on the emitter has resulted in a collector current.

It remains to be shown that the base current will be very much less than the collector current. Base current is caused by electrons flowing into the base from the external circuit for any of the following three reasons:

1. To replace electrons that have been forced across the emitter junction by the forward bias.

2. To replace electrons that have recombined with holes within the base region.

3. To recombine with holes reaching the base terminal.

This is where the ratio of conductivities between the emitter and base regions becomes important. It will be remembered that there is a far greater number of holes available for conduction in the emitter than there is electrons in the base. Therefore, when the emitter junction is forward biased, a greater

[1]This point will be discussed further in Chapter 2. The diffusion forces are of thermal rather than electrostatic origin.

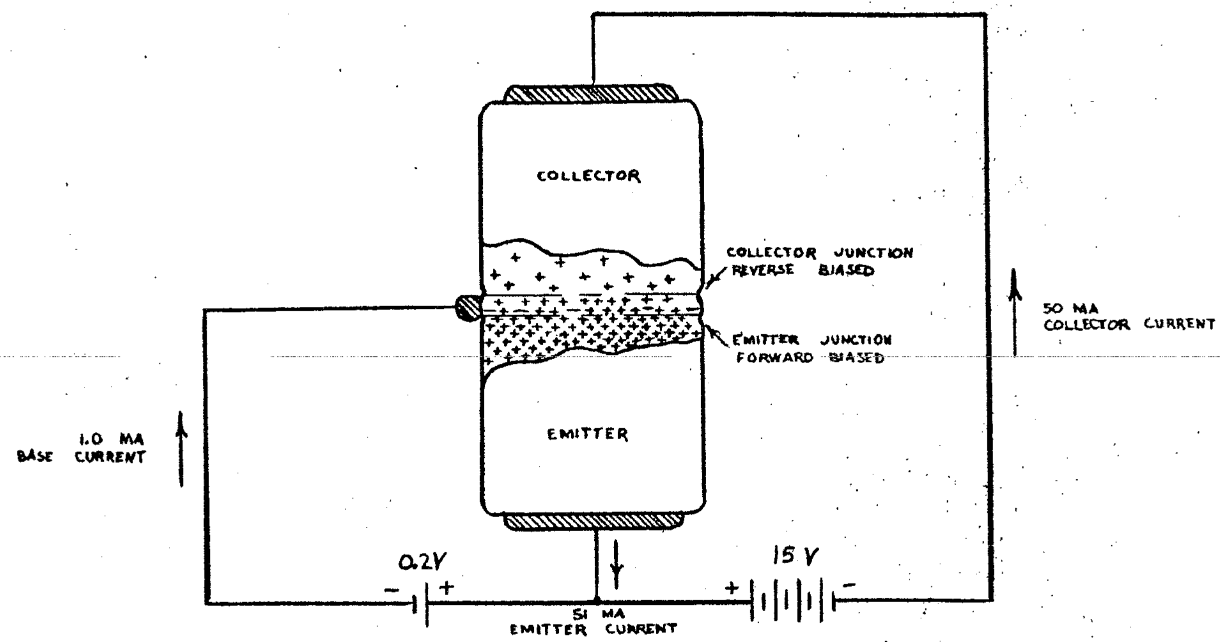
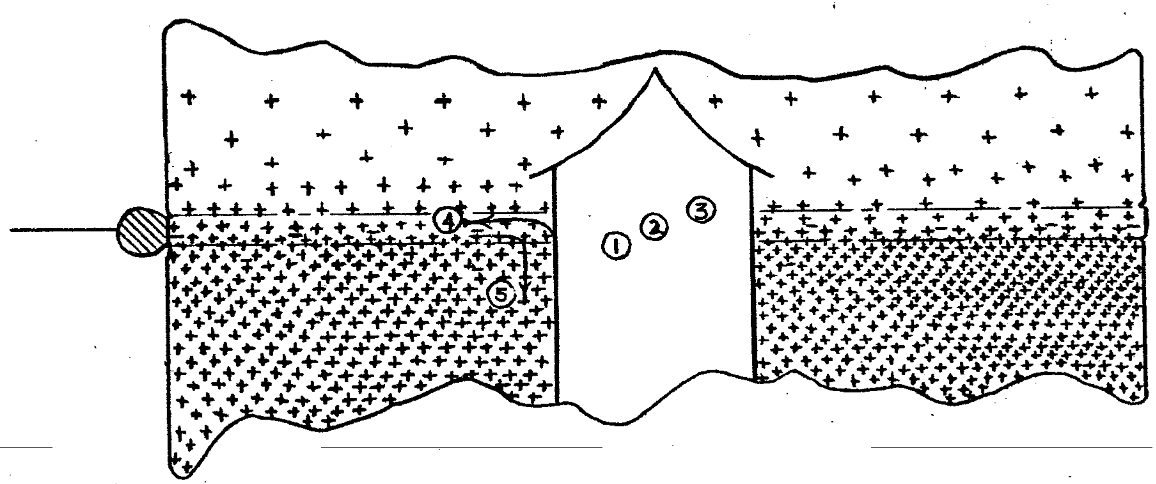


Figure 1.8. Conducting Transistor — Illustrating Control of Collector Current by Emitter-Base Junction.



1- Holes cross emitter junction because of foward bias.
2- Holes diffuse through base.
3- Holes are swept across collector junction by negative collector voltage.
4- Some holes do recombine with electrons or reach base terminal.
5- A far smaller number of electrons cross into emitter.

Figure 1.9. Expanded View of Base Region.

number of holes will flow across this junction into the base than will electrons into the emitter. Moreover, the holes in the base are not likely to recombine with electrons in appreciable numbers, both because of the extreme thinness of the base through which they must travel, and because of the small number of free electron carriers in the base. Finally, the number of holes, reaching the base and recombining with electrons, will be small because the holes will diffuse the short distance to the collector junction before traveling the comparatively long distance to the base contact. In this manner, small currents in the base can control large currents in the collector circuit. The ratio of the collector current to the emitter current is known as the beta (β) current gain of the transistor. In typical units it will range from 10 to 100.

The physics of the transistor will be set aside, for a moment, and its practical applications in electronic circuits will be observed. In particular, the use of the transistor as an amplifier will be discussed.

Figure 1.10 illustrates the PNP transistor connected as a medium-power amplifier. While a schematic representation of this transistor is used, the transistor elements and the directions of current flow are clearly labeled. The input and output waveforms are also shown.

As could be expected, -30 volts is applied on the collector to reverse bias the collector junction.[1] In addition, a small negative bias voltage is put on the base to forward bias the emitter junction. When the emitter junction is forward biased, a 50 ma collector current will pass through the 300 ohm load resistor dropping the collector voltage to -15 volts. The input signal is inserted in series with the bias source so that the forward bias can be increased and decreased alternately by the input signal. The changing forward bias will cause relatively large changes of collector current while there are only small changes in the base current. This varying collector current produces corresponding voltage drops across the series load resistor. The amplifying action of this circuit arises from the fact that small current and voltage changes in the base circuit can produce large variations in collector current. Further, because the collector is reverse biased, comparatively large voltages can be applied to the collector, so the current variations in the collector can cause large voltage changes. The tabulation of characteristics below give the type of performance that can be expected from a transistor amplifier. For this purpose the 300 ohm collector resistor is treated as the load:

1. Current gain --50.

2. Voltage gain --75.

3. Power input --0.025 mw.

4. Power output --94 mw.

5. Power gain --3760, or 36 db.

6. Maximum output --375 mw.

[1]Regardless of the configuration used, the collector junction of a transistor is always reverse biased.

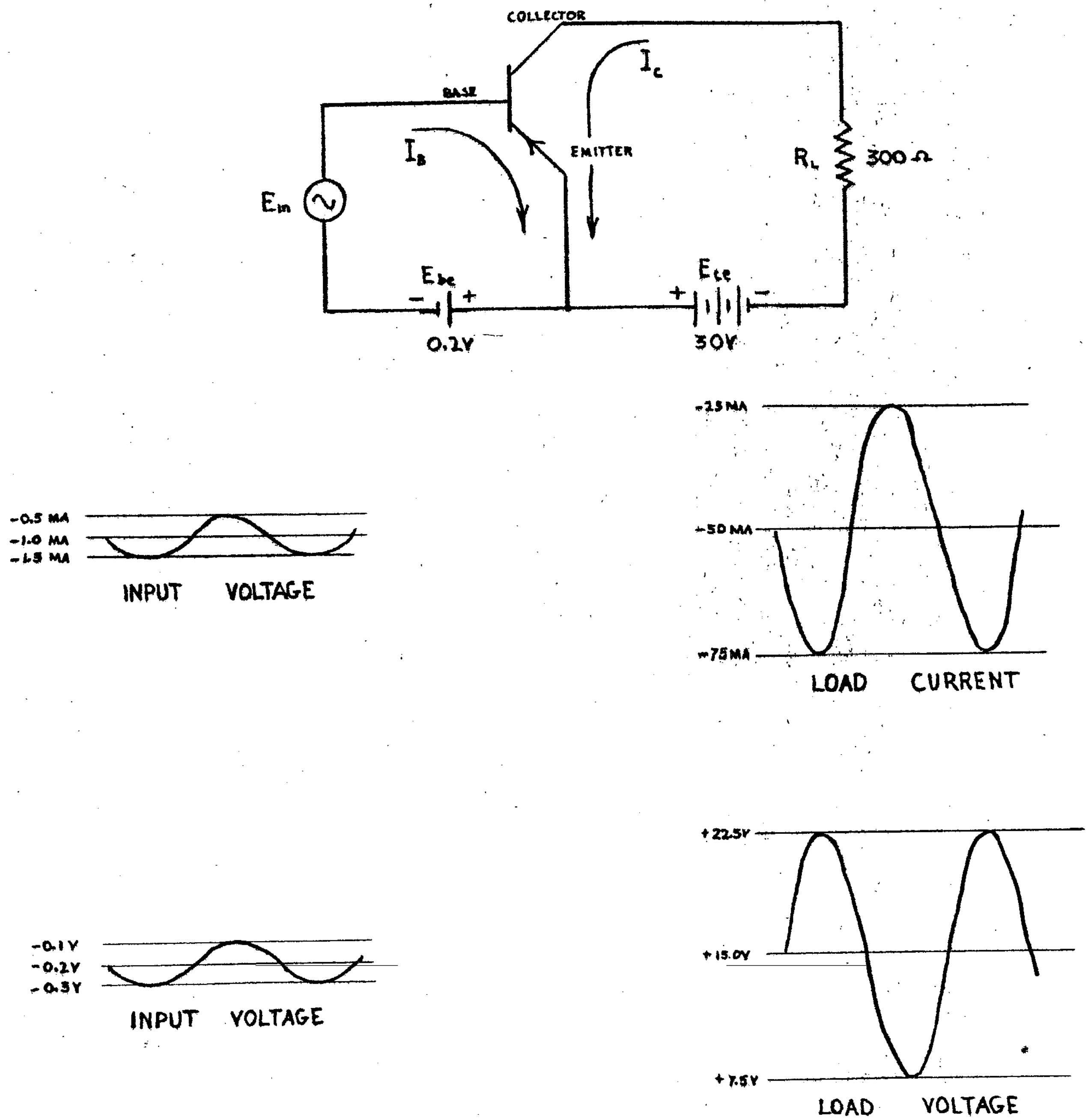


Figure 1.10. The Transistor as an Amplifier.

Although the listed power output of 375 mw may seem low, it must be remembered that this is about the same available from a large battery-powered portable radio.

This single circuit obviously does not begin to describe the possibilities of the transistor. Amplifiers and oscillators can be made that will operate with a d-c input power as low as a few microwatts or, using power transistors, with outputs in the order of several hundred watts. In addition, efficiencies approaching theoretical values can be attained with practical circuits in switching and audio applications. Furthermore, there are NPN transistors which operate in the same manner as PNP units except that the role of the holes and electrons are interchanged. The NPN transistors function with voltages of opposite polarity to the PNP types. This permits design of circuits that would not be possible with vacuum tubes, as this corresponds to having a tube that will operate with a negative plate voltage.

However, transistors do have many limitations that have not yet been mentioned, but these limitations can be overcome with proper circuit design. Therefore, any serious discussion of circuits using semiconductor diodes and amplifiers will be put off until a better understanding of the factors affecting semiconductor performance, mainly temperature and operating frequency, is gained.

DEPARTMENT OF WEAPONS TRAINING
LOWRY AIR FORCE BASE
COLORADO

T R A N S I S T O R S

IN SERVICE TRAINING COURSE

CHAPTER 2

CONDUCTION IN SOLIDS

8 DECEMBER 1960

FOR INSTRUCTIONAL PURPOSES ONLY

CHAPTER 2

Author: A1C Robert J. Widlar

CONDUCTION IN SOLIDS

__Abstract__ - The chemical behavior of the elements is related to the structure of the atom. Chemical reactions are discussed to introduce the covalent bond. The electronic structure of semiconductor crystals is analyzed to show the origin of both hole and electron current carriers. The temperature sensitivity of semiconductors is explained by considering thermal agitation. The mechanisms of conduction in semiconductors, both drift and diffusion, are given considerable attention. Finally, the important physical properties of many semiconductor materials are evaluated.

GENERAL

Now that the fundamental concepts of the diode and the transistor have been introduced, a more exact explanation of conduction in solids will be attempted. From this wil follow a discussion of semiconductor devices on a more exact basis.

Confusion often results when semiconductors are explained assuming the existence of holes and electrons then treating them as free particles unless certain restrictions are clearly understood. The operation of vacuum tubes can be adequately explained by considering the action of individual free particles, such as electrons and ions, because the separation of these charged particles is great enough that the interactions between them can be neglected. However, this is not the case when solid state devices are to be explained. The entire structure of the solid must be investigated as there are interactions between the mobile current carriers and the closely packed immobile atoms. For this reason, a certain amount of the physics and chemistry of solids must be discussed before an intelligent view of semiconductors can be had.

STRUCTURE OF THE ATOM

In the ensuing discussion it is assumed that the reader has a fundamental knowledge of atomic physics. The explanation of the atom will be conducted as a review. Only those points necessary for an understanding of solid state phenomenon will be covered in detail.

The atom can be thought of as a positively charged nucleus surrounded by a cloud of negatively charged electrons. The nucleus is composed of protons having a unit positive charge and neutrons having no charge. These two nuclear particles make up the major portion of the atomic mass. The electrons have a unit negative charge so an neutral atom has an equal number of protons and electrons. The electrons are extremely small particles and contribute little to the total atomic mass.

The negative electrons can be pictured as existing in orbits around the nucleus. These orbits are not random, but are arranged in some definite manner. Figure 2.1 illustrates this point with a diagrammatic representation of the first 20 elements on the periodic chart. Circular orbits are indicated at definite distances from the nucleus. This is not actually the case; the given orbits are meant to indicate that the electron possesses a certain energy (kinetic energy from its motion and potential energy from its proximity to the nucleus). The position of the electron is uncertain, but it can be thought to describe a three dimensional elliptical orbit (ellipsoidal shell). The energy is somehow related to the mean diameter of the orbit so, although it is not exactly correct, the orbit will be considered circular. This is done to simplify analysis, but it should be remembered that a large diameter orbit indicates a high energy electron, and only approximates the actual path of the particle.

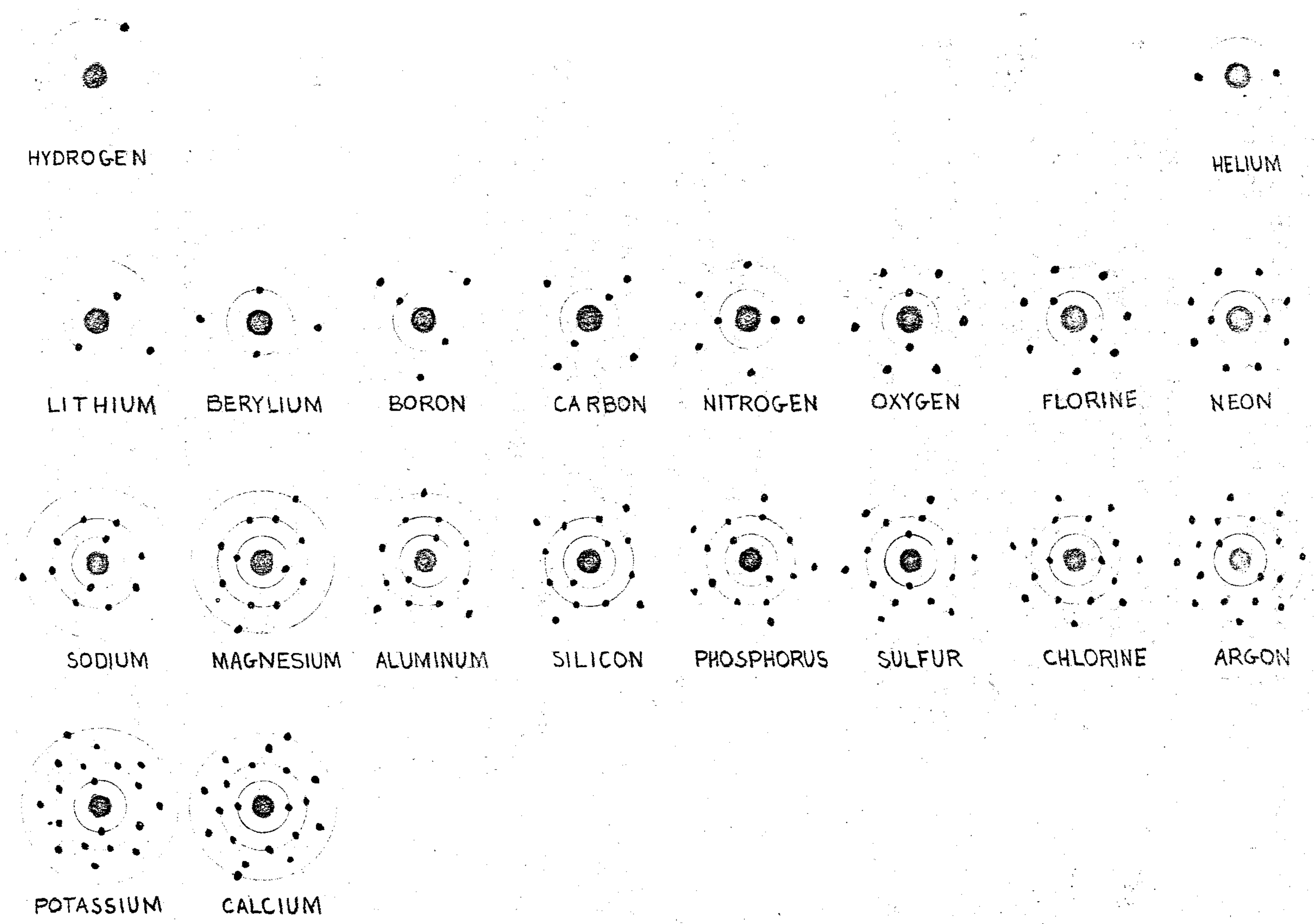


Figure 2.1. Electron Shell Configurations for the First Twenty Elements on the Periodic Chart.

The simplest element is hydrogen with a single electron in orbit about a proton nucleus. The next element, helium, has a nucleus, containing two protons and two neutrons, which shows a double positive charge. Two electrons orbit at equal distance from the nucleus.

The third element, lithium, has a nucleus with a charge of -3. There are three electrons circling the nucleus, but the third electron has assumed another larger diameter orbit. Continuing through the elements from lithium to neon, electrons are added in this second orbital path, or shell. However, with the sodium atom, a third shell is started.

CHEMICAL REACTIONS

It appears that these electron shells can only hold a certain number of electrons. Once these shells are filled, a new one is started. The filling of these electron shells and the behavior of partially filled shells is of particular interest because the difference in the chemical properties of different elements can be atributed, in part, to the arrangement of electrons in the outer shell.

Formation of Ions. Elemental lithium will react with florine gas to to form solid lithium floride. In this reaction, an electron is transfered from the outer shell of the lithium atom to the outer shell of the florine atom, as shown in figure 2.2a. The lithium atom is now lacking one electron so it will exhibit an overall charge of -1, while the florine atom will have a net charge of -1.

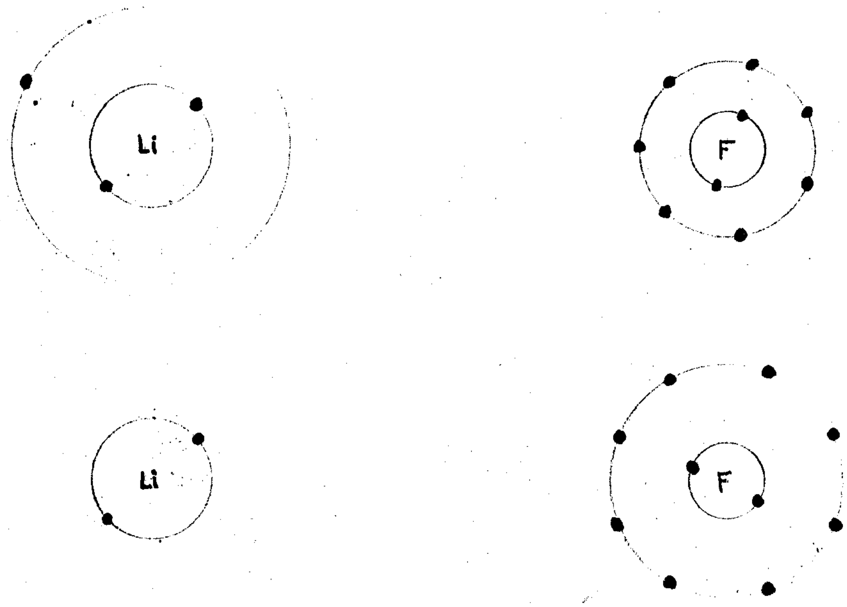
The fact that lithium will give up an electron to florine can be related to the configuration of its outer electron shell. In the lithium atom, the single electron in the second shell is shielded from the -3 charge of the nucleus by the two electrons in the first shell. Hence, this electron will be loosely bound to the nucleus and can easily be removed. However, in the florine atom, the two electrons in the first shell do not effectively shield the -9 charge of the nucleus so the 7 outer shell electrons are held by an appreciable portion of the nuclear charge. There is room for another electron in this second shell (to a maximum of 8) so, if an extra electron is available, it can go into the second shell where the electrons are far enough apart that the nuclear attraction for any electron in the shell is considerably greater than the mutual repulsion between them. This explanation accounts for the tendency of flourine to form a negative ion.

The reaction between berylium and oxygen, forming berylium oxide, supports this concept. Berylium will loose two electrons to oxygen as indicated in figure 2.2b. The reason that it looses these electrons is the same as that given for lithium, but this tendency is not so pronounced. The higher positive charge of the nucleus exerts a stronger influence on the second-shell electrons through the first shell. Berylium will generally loose both electrons because the attractive force on the second electron is not increased after the first is removed. The attraction that the nucleus exerts on an electron can be altered by the pressence of another electron only if this second electron is between it and the nucleus; not by an electron in the same shell. It follows that, if one electron could be removed, the second could be removed just as easily.

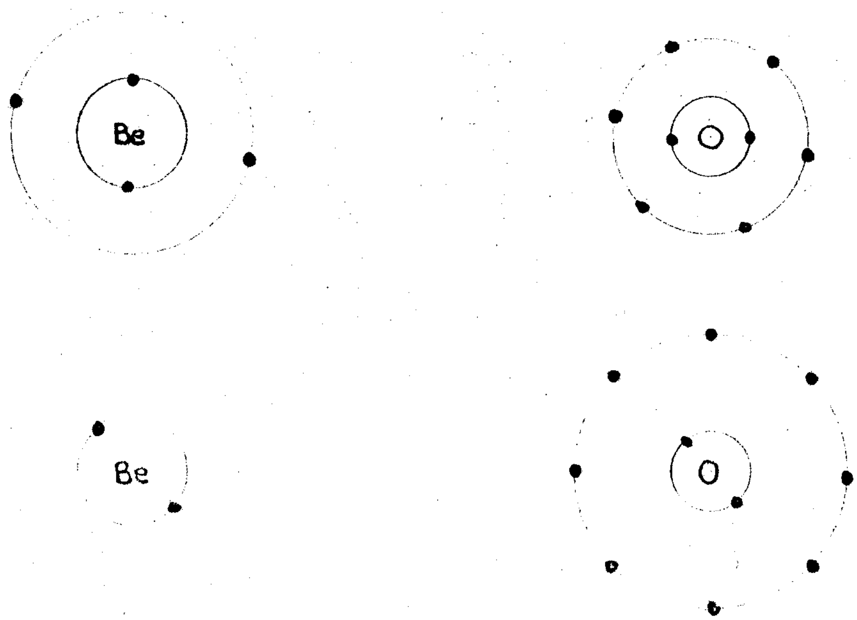
Oxygen can accommodate two more electrons in its outer shell. The addition of these two electrons will not increase the total number to the point where mutual repulsion between the electrons overcomes nuclear attraction. In the example being considered (figure 2.2b) oxygen captures these two electrons from berylium.

The tendency for oxygen to become a negative ion is not as strong as that for florine because the nuclear charge of oxygen is one unit less. The reason that it will take on two electrons is because it can accommodate them in the outer shell, not becuase its holding power is any greater than that of florine.

<u>Covalent Bonding</u>. It follows from the preceeding discussion that an exchange of electrons will take place between atoms with relatively few electrons in the outer shell (which form positive ions) and atoms with an almost full outer shell (which form negative ions). When the number of electrons in the outer shell of the reacting substances are nearly equal, the chemical reaction proceeds somewhat differently. For example, both florine and chlorine have a strong tendency to gain another electron because they both have seven in their outer shells. When the compound chlorine floride is formed, electrons are not completely transfered because neither atom has a strong enough affinity for an electron to remove it from the other atom. Or, conversely, the outer electrons are not so loosely held that they can be removed by either atom. What does happen is that the outer shell electrons are shared. The two atoms are joined together, and the outer shell electrons assume some complex orbit about both atoms instead of around the individual atoms. This gives rise to a force which binds the two together forming a molecule. This binding force is refered to as a <u>covalent bond</u>. The chlorine floride molecule is shown in figure 2.3a.



a. Lithium and Florine React to Form Lithium Floride.



b. Berylium and Oxygen React to Form Berylium Oxide.

Figure 2.2. Illustration of Chemical Reactions in which Complete Electron Transfer Takes Place.

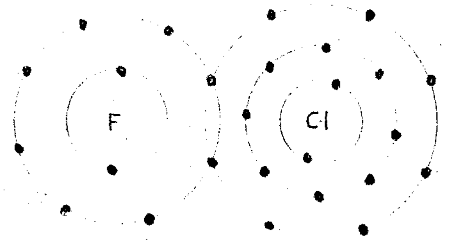
Other examples of covalent bonding are given in figures 2.3b and 2.3c. The first illustrates the result of a reaction between hydrogen and oxygen to form water. The second shows the product of a reaction between carbon and chlorine. In both cases, neither of the reacting substances will gain or loose electrons. The outer shell electrons are shared, not transfered. After the reaction, all the outer shell electrons apparently orbit around the entire molecule in a distorted path, rather than about the individual atoms.

Stable Electron Configurations. The inert gasses will serve as a foundation for the following discussion. It can be seen from figure 2.1 that the outer shell of the inert gasses listed (helium, neon, and argon) are apparently complete since the following element on the periodic chart starts a new shell. These gasses are chemically inactive in that they will not react with any known substance. Therefore, it appears that a completed electron structure is very stable.

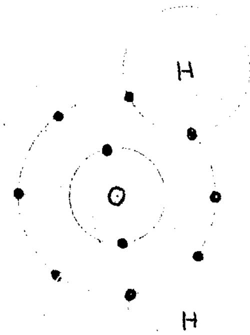
There is one thing in common in all the reactions mentioned thus far. The outer shell electrons were rearranged to produce a completed outer shell in all the reacting substances similar to that of the inert gasses. This is true whether there was a complete transfer of electrons to form ions or a sharing of electrons to form a molecule. For example, the lithium atom in figure 2.2a looses an electron thereby assuming the stable electron configuration of helium. The florine atom, which gained the electron, assumes the stable configuration of neon. The same is true for the berylium and oxygen atoms in figure 2.2b. Electron sharing produces similar results as can be seen from figure 2.3. In the first example (chlorine floride), sharing of the outer shell electrons produces a stable electron structure in both the chlorine and florine (that of argon and neon, respectively). The same is true in the other examples.

A chemical reaction will normally produce a stable electron structure in the reacting substances. If not, the reaction products are themselves unstable and will usually decompose into a compound which does have a stable structure.

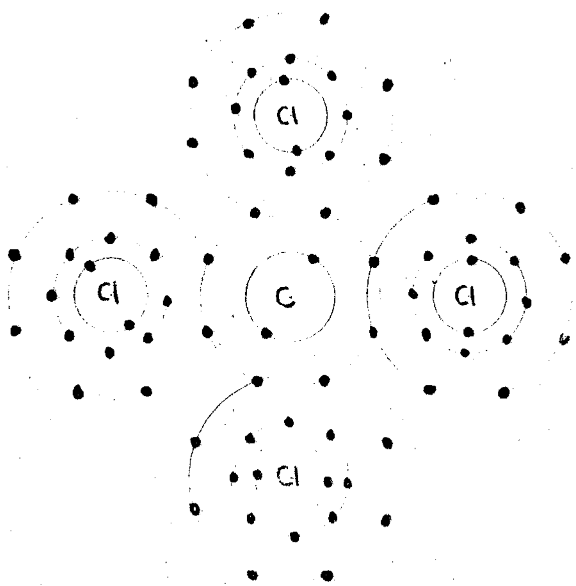
Other Chemical Properties. More information can be gathered from a further study of the atom. The chemistry of an element is not solely determined by the number of electrons in the outer shell, although this does appear to be the most important factor. The atomic number of the element is also important. For example, lithium, sodium, and potassium have one electron in their outer shell and all three tend to loose it readily. However, potassium looses it most easily while sodium is next and lithium last. This can be explained. The lone outer electron of potassium is better shielded by three completed electron shells than is that of either sodium or lithium, which have fewer completed shells. The same trend is evidenced with chlorine and florine. Both these atoms tend to gain an extra electron, but florine has the strongest tendency because



a. Chlorine Floride Molecule.



b. Water Molecule.



c. Carbon Tetrachloride Molecule.

Figure 2.3. Three Examples of Atoms Held Together by Covalent Bonding.

its nucleus is shielded by one less electron shell than is chlorine. Generally, in different elements with the same number of outer shell electrons, those with the higher atomic number will loose electrons more readily or gain them less easily. This fact becomes important later in explaining the difference between different semiconductor materials, namely germanium and silicon.

BINDING FORCES IN SOLIDS

When sodium reacts with chlorine, two ions will be formed. After the reaction has completed, these ions will join together to form solid sodium chloride. Their net positive and negative charges will set up an electrostatic force of attraction which holds the atoms together in the solid form. This binding force is known as an ionic bond.

Solids can be formed in several ways, ionic bonding being only one example. Water molecules (figure 2.3b) are held together in ice by forces of a different origin. Each molecule has a net charge of zero, but the molecule behaves like an electric dipole: Negative on one end and positive on the other. The outer electrons shared by both the hydrogen and oxygen are displaced somewhat toward the oxygen because of its greater affinity for electrons. This end of the molecule becomes negative with respect to the end containing the two hydrogen atoms. When water freezes to form ice, the molecules become lined up and are held in place by the polar forces.

When two neutral atoms are brought into close proximity, the electron cloud of each will cause a net repulsion. However, if they are brought even closer, the positive charges of the nuclei will act through the electron cloud and produce a net attraction: The nucleus of one atom will attract the electron cloud of the other. This is how some elemental solids are held together. For example, in one form of carbon, lampblack, the individual atoms are held together by this type of attractive force.

The kind of binding force that will be of primary interest from this point on is the covalent bond. Solids can be held together by one continuous covalent bond. This is the force that holds together practically all elemental metals and infermetalics. Carbon will serve to illustrate the characteristics of the covalent bond: As mentioned above, lampblack is held together by rather weak attractive forces. Another form of carbon, diamond, is held together by a continuous covalent bond. A third form, graphite, is held together by a combination of these forces. Graphite is made up of molecular layers of carbon atoms. In each layer the atoms are held together by covalent bonds, but between the layers the forces are the same as for lampblack. This accounts for the fact that graphite can be easily split along certain planes. An interesting point is that both lampblack and graphite are conductors while diamond is an excellent insulator.

CRYSTALS

The atoms of crystals are arranged in a definite, orderly, and continuous pattern. On the other hand the atoms of some solids known as glasses, are arranged in a haphazard fashion. The difference between these two substances is evident when they are melted and then cooled. A crystal will go from the liquid to the solid form at a definite temperature known as the freezing point, but a glass will become thicker and thicker before it finally hardens. This does not occur at any definite temperature. When a crystal freezes, it liberates a certain amount of heat (heat of fusion). If the crystal is then remelted, the heat of fusion must be supplied to the crystal before it will melt. This indicates that the binding forces in crystals are generally stronger than those of glasses since this energy in the form of heat must be supplied to break the bonds and melt the crystal.

Sodium chloride is an example of a crystaline solid. Each sodium ion is surrounded by six chloride ions, and each chloride ion is surrounded by six sodium ions. This pattern persists throughout the crystal, being broken only at the surface.

Water will form ice crystals as it freezes. The polar water molecules align themselves in an orderly pattern and produce a continuous crystal structure.

Most important in the study of semiconductors is the completely covalent crystal. In this case the entire crystal is held together by a single continuous covalent bond. Carbon, silicon, and germanium can form such crystals.

When carbon atoms join together to form a diamond crystal, the four outer shell (valence) electrons engage in a covalent bond. This is not unexpected considering that all these atoms have an equal tendency to gain or loose electrons. Figure 2.4 illustrates the bonding of a diamond crystal. Each carbon atom shares electrons with four of its neighbors in such a way that the electron structure of each atom becomes completed. A pictorial view of the space arrangement of the atoms in a carbon crystal is given in figures 2.5 and 2.6. Figure 2.5 shows the orientation of atoms about a central atom. This arrangement is repeated in the complete crystal as shown in figure 2.6. The covalent bonding and crystal structure of silicon and germanium is similar to this since they also have four electrons in their outer shell. The strength of the covalent bonds becomes progressively weaker going from carbon to silicon to germanium. This is probably due to the presence of an additional inner electron shell on each of these atoms.

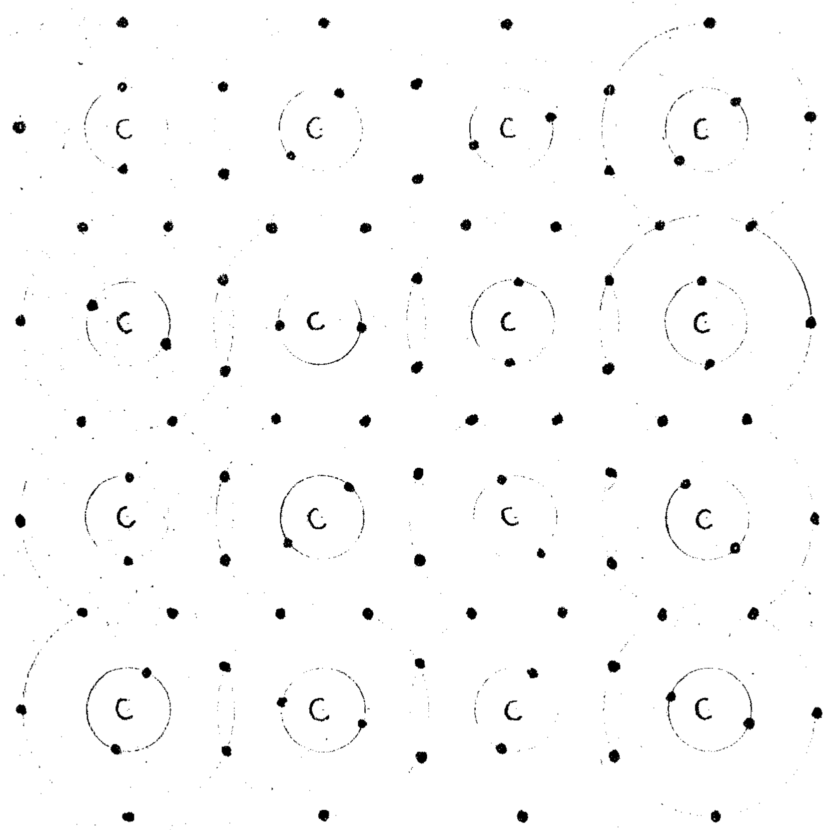


Figure 2.4. Schematic Representation of Covalent Bonding in a Diamond Crystal.

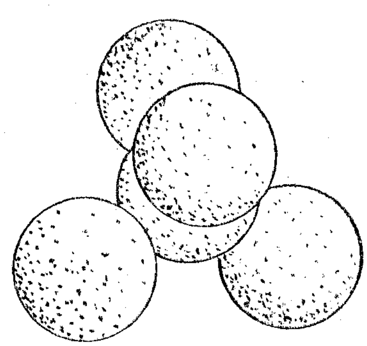


Figure 2.5. Illustration of Atomic Arrangement Around a Single Atom in a Carbon, Silicon, or Germanium Crystal.

When molten silicon or germanium (individual atoms) freeze to form a crystal, energy in the form of heat is released (heat of fusion). This energy must have gone into the formation of covalent bonds since the configuration of the outer electron shell is the only change brought about by this process. It follows that energy must be supplied to break the covalent bonds whether it be thermal energy in melting the crystal, mechanical energy in breaking the crystal, or electrical energy in ionizing electrons from the covalent bonds.

It appears that the electron structure of these crystals is quite stable. Electrons cannot be added to the completed valence structure, nor can they be removed without supplying energy to break the bonds.

Internal Electric Fields. The electric field close to a neutral atom is not zero. However, when many of these atoms are brought close together, the individual fields will neutralize, creating regions of negligible electric field. In the crystals being considered, the space between atoms is electrically neutral, as is the region occupied by the outer shell or valence electrons. This shows why the valence electrons are not attracted to any particular nucleus and are free to roam about the entire crystal; the only force holding them in the outer shell structure is the covalent bonds. If an electron is excited or broken away from the covalent bonds it will move about in the free space between atoms until it finds a vacancy in the covalent bonds. It will then give up the energy it gained in breaking loose and go back into the valence structure.

External Electric Fields. At low temperatures, the covalent bonds of carbon, silicon, and germanium crystals are complete. When an electric field is applied across these crystals, the valence electrons tend to move in the positive direction of the field. However, before the displacement becomes appreciable, they will be restrained by the covalent bonds. There is no continuous motion of charged particles and no current through the crystal.

Free Electrons. The covalent bonds are not so strong that they cannot be broken by the application of sufficiently high voltages. When very high voltages are applied across the crystal, the electric field tending to move the electrons becomes stronger than the restraining forces of the covalent bonds. When this happens, electrons are broken loose and introduced into the space between the atoms. Here, the only force acting on them is the externally applied electric field, so they move through the crystal under the influence of this field. This motion will not be in a straight line. The electrons will collide with the atoms in moving through the crystal. In these collisions, the electrons do not usually make physical contact with any part of the atoms; but they are deflected by electrostatic forces when they penetrate the outer electron shell. In the process, they give up some of their energy to the atom. This loss of energy accounts for the electrical resistance of the material. Normally, when the electrons collide with the atoms, they are not retained in the covalent bonds unless there happens to be a vacancy at that point in the crystal. Furthermore, the kinetic energy of the electron must be low enough so that it cannot break loose after it is captured. The motion of a free electron is illustrated in figure 2.6.

Holes. The movement of free electrons through the crystal accounts for only about half the total current observed. What was introduced as a hole in chapter 1 supplies the remainder.

When the covalent bonds are complete, the valence electrons are free to wander about the crystal: That is, they are not bound to a particular atom so any individual electron can drift from atom to atom as long as it is immediately replaced by another. When a moderate electric field is applied across the crystal, the valence electrons try to move under the influence of the field; but they are restrained by the covalent bonds. Any net motion would remove electrons from some atoms and give others too many, which will not happen because the outer shells of all the atoms are already completed.

If an electron is excited from the covalent bonds, this situation is altered. The removal of an electron from the bonds creates a vacancy, and the valence electrons acted upon in the proper direction by the external field will move in and fill this vacancy. Hence, the gap in the valence structure will move through the crystal in much the same way as the free electrons. It should be remembered, though, that this vacancy is not something tangible. It is just the condition of the covalent bonds at a particular point in the crystal. Conduction is actually caused by the net motion of bound electrons in the outer electron shells.

A good analogy of this action is the motion of a bubble in a glass of water. A bubble released at the bottom of the glass will appear to rise to the surface. However, an exact analysis will show that the water is actually moving downward under the influence of gravity. The bubble is nothing but a void that is filled by water in response to the force of gravity. Similarly, in the crystal, the vacancy seems to move through the continuous electron structure, but the valence electrons are actually moving under the influence of the applied field.

Since this vacancy unbalances the electronic charge distribution in the crystal, it is accompanied by a small positive charge. Therefore, considering only the electrostatic effects on the interior of the crystal, the hole appears to be a small positive charge. When an external field is applied, this charge seems to move toward the more negative portion of the field. Although it is an unbalanced condition that is being displaced, the net effect is the same as if a free positive charge were moving through the crystal.

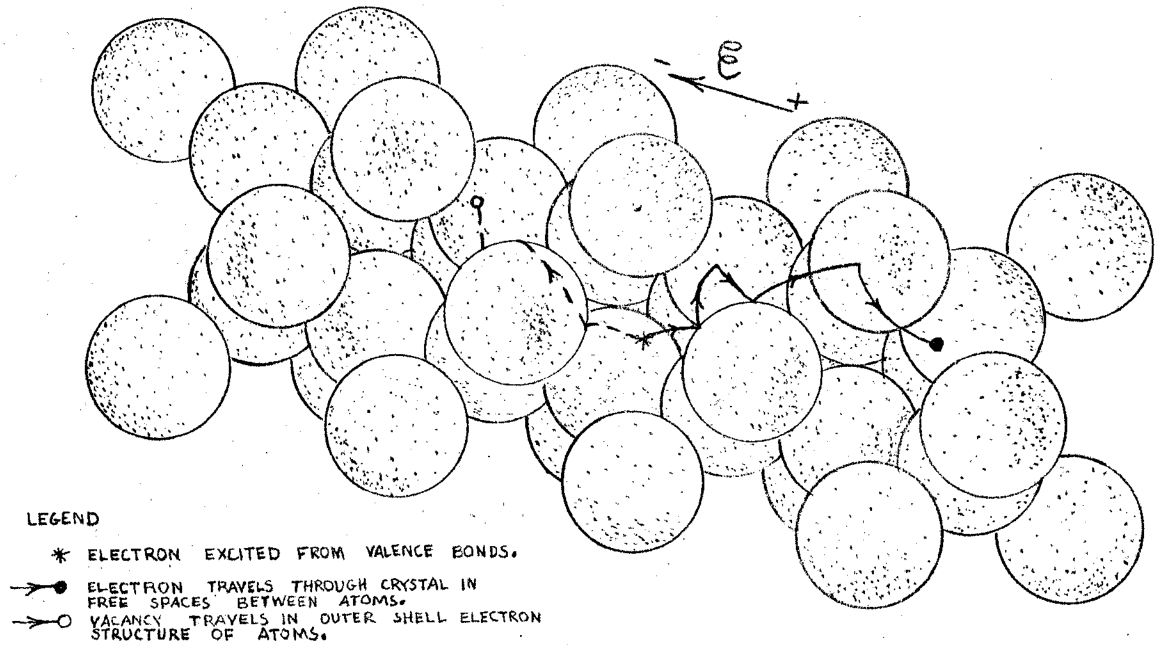


Figure 2.6. Paths of Free Electrons and Vacancies in the Valence Bonds Under the Influence of an Applied Electric Field.

Conduction by holes is illustrated in figure 2.6. The vacancy travels in the outer electron shells, but the electrical field from the net positive charge associated with this vacancy permeates the free space between the atoms. Any electric or magnetic fields that exist in the free space will react with this field so the net effect is again that of a free positive charge.

It can be seen that there are two distinct types of current carriers in the crystals discussed thus far: Free electrons and bound valence electrons. However, the valence electrons can only contribute to conduction if there is a vacancy in the covalent bonds. This condition was shown to be equivalent to a free positive charge within the crystal.

THERMAL AGITATION

The molecules of matter are in a continual state of motion. The magnitude of this motion is directly dependent on temperature. For example, if a gas is sealed inside a closed container and the temperature increased, the pressure on the walls of the container will increase. As the gas is heated, the thermal energy abosrbed will increase the kinetic energy of the gas molecules. These molecules will then strike the walls of the container with greater froce, producing the increased pressure.

Other physical evidence of thermal motion is not difficult to find. When a solid is heated above its melting point the individual molecules gain enough thermal energy to break loose from one another, and a liquid is formed. If the temperature is increased further, the molecules will gain enough energy to break away from the liquid surface and move out into the surrounding space to become a vapor or gas. If the temperature is increased still further, the molecules will shake themselves apart into individual atoms. Finally, at very high temperatures, the electrons of these atoms will absorb enough thermal energy to break away from their orbits, producing a gaseous mixture of ions.

From the above example, it appears that thermal motion manifests itself as an oscillation of the individual molecules against the forces binding them together. Furthermore, it appears that all the components of the solid take part in this oscillation: Not only do the molecules vibrate; but the atoms also vibrate, restrained by the forces binding them into molecules; and the electrons vibrate against the attraction of the nuclei.

It follows that in the crystals being considered (carbon, silicon, and germanium) the atoms are not in a fixed position at normal temperatures, but are oscillating about a mean position because of their thermal energy. Moreover, the valence electrons are oscillating against the restraining forces of the covalent bonds. In both these cases, the amplitude of oscillation is some function of temperature, increasing with increasing temperature.

Energy Distribution. When all the heat is removed from a substance, thermal motion ceases. The temperature corresponding to this condition is called absolute zero ($-273^{\circ}C$). At this temperature, the atoms of a semiconductor crystal will be fixed in a neat, orderly array as shown in figure 2.6. At any temperature above absolute zero, the atoms will oscillate about their original positions, the amplitude increasing with increasing temperature. The average thermal energy of all the atoms is well defined since only a certain amount of thermal energy was put into the crystal in raising its temperature. However, because it is possible for one atom to collide with another and give up part of its energy (whether the collision is physical or an electric field interaction), the energy of the individual atoms can vary. These collisions will be completely random so it is possible for some atoms to loose all their thermal energy and for others to gain energy from several collisions. The energies of the individual atoms can therefore range from zero to several times the average thermal energy.

Observation has shown that most crystals melt at a definite temperature which indicates that almost all the atoms gain enough thermal energy to break loose from the covalent bonds at approximately the same temperature. This means that the energy distribution about the average value will be relatively narrow. Very few atoms will have energies very much higher than the average value and very few will have energies much lower than average. The narrow distribution indicates that the frequency of collision between atoms is not too great.

The valence electrons will also have a certain amount of thermal energy in addition to their kinetic energy of motion in the outer electron shells. Again this energy is not the same for all electrons, but is distriubted about some average value as a result of interactions between electrons and between electrons and atoms. Hence, at any temperature above absolute zero, an individual electron may have any amount of thermal energy from zero to many times the average value.

Thermal Ionization. If an electron gains sufficient thermal energy, it is possible for it to break loose from the covalent bonds and be introduced into the free spaces between atoms. This process involving the thermal generation of hole-electron pairs is refered to as thermal ionization. It is similar, in effect, to the production of hole-electron pairs with strong electric fields, except that the ionization energy is supplied by the thermal motion of the electrons rather than an electric field.

At low temperatures there will be very few electrons with enough thermal energy to break the covalent bonds. As temperature increases, this number will increase quite rapidly. However, even at high temperatures, only a very small percentage of the outer shell electrons will become ionized.

Before going further, it would be enlightening to consider the magnitude of these thermal effects. For example, in a germanium crystal at -125°C, there are about 8 X 10^6 hole-electron pairs per cubic centimeter. This is quite a large number, but there is only about one thermal ionization for every 2 X 10^{14} germanium atoms. At this temperature, germanium is an excellent insulator. Near room temperature (30°C) there are about 2 X 10^{13} hole-electron pairs per cubic centimeter or one ionization in 5 X 10^{10} atoms. Even at the highest operating temperature of practical germanium devices (100°C), there are about 10^{15} ionizations per cubic centimeter or one in 10^9 atoms. At these temperatures, germanium is a poor isulator; but it is still far from being a conductor. In a conductor each atom donates several free electrons to the metalic crystal giving about 10^{28} free electrons per cubic centimeter.

After an electron is ionized from the covalent bonds, it does not remain stationary in the interior of the crystal; but it will wander through the crystal with a random, erratic motion, encountering frequent collisions with the atoms. This motion arises from the thermal energy of the electron. Similarly, the holes created by this process also move about the crystal with a radom motion because of the excess thermal energy of the valence electrons. However, in neither case does this motion constitute a current. On an average, there are as many current carriers moving in one direction as there are in another, so there is no net transfer of charge.

If an electric field is applied through the crystal, the random thermal motion of the holes and electrons will be modulated by a net motion in the direction of the field. The magnitude of the current for a given field strength (conductivity) will obviously depend on the number of carriers available and on their average forward velocity in the direction of the field. For a given field strength, the forward velocity of the current carriers will depend on the number of collisions suffered by the carriers in moving through the crystal. At low temperatures, the collisions will be relatively infrequent, being caused only by the forward motion of the carriers. At higher temperatures, the thermal motion of the atoms and electrons will increase the frequency of collision and, therefore, reduce the forward velocity.

This affords another distinction between pure semiconductors and metalic conductors: In a conductor, there are generally several outer shell electrons per atom that are not required to complete the covalent bonds. These electrons are free to move through the crystal under the influence of an externally applied electric field and can therefore support a current. The number of current carriers will remain essentially constant with temperature since there is already a huge number available and any small addition from thermal ionizations would be insignificant. Hence, as temperature is increased, the only effect on the conducting properties of the crystal is the restricted motion of the free electrons arising from the increased thermal motion. It follows that the conductivity will decrease with increasing temperature. For most metals, the conductivity is found to decrease at the rate of about 0.3 per cent per degree increase in temperature.

In a semiconductor, the number of current carriers does not remain constant, but increases enormously with temperature. As temperature is increased, the increased concentration of carriers affects the conductivity much more than does the restricted motion. Hence, a semiconductor has a negative temperature coeficient: That is, the conductivity increases with increasing temperature. Furthermore, the magnitude of this change is considerably greater than the change in metals. For germanium near room temperature, there is about a 5 per cent increase in conductivity per degree increase in temperature.

Thermistors. In diodes and transistors, the temperature sensitivity of semiconductor materials is moderated by the addition of impurities as will be shown in the next section. There are, however, a class of semiconductor devices known as thermistors (negative temperature coeficient resistors) that make use of this large change in conductivity with temperature. Thermistors have found application as temperature sensors and as compensating elements in transistor circuits.

EFFECTS OF IMPURITIES ON ELECTRICAL CONDUCTIVITY

It was shown in chapter 1 that diodes and transistors were constructed of P-type materials, having holes for current carriers, and N-type materials, having electrons for current carriers. In the pure semiconductor crystals considered thus far the number of holes and electrons were equal; however, these pure crystals can be given P or N-type characteristics by the addition of suitable impurities.

N-Type Crystals. If a small amount of phosphorous, arsenic, or antimony is added to a germanium crystal during the growth process, its conductivity at room temperature will be much greater than that of a pure crystal. Furthermore, tests carried out on this crystal will show that there are far more free electron current carriers than there are holes. This can be explained by considering the effect of these impurities on the electronic structure of the crystal.

The outer electron shells of phosphorous, arsenic, and antimony atoms contain five electrons. If these atoms are introduced into the crystal in small quantities, they will replace germanium atoms in the crystal structure; and their outer shell electrons will take part in the covalent bonding in much the same way as the germanium atoms, except that there will be one electron left over which is not required to complete the bonds. This is illustrated in figure 2.7.

At very low temperatures this fifth electron is held to the impurity atom by the extra positive charge of its nucleus. However, the electrostatic forces holding this electron are much weaker than those of the covalent bonds. At relatively low temperatures, the thermal energy of this electron will overcome the electrostatic forces binding it to its parent atom so it will break loose and wander through the crystal with a random thermal motion. Thus, free electrons are introduced into the crystal without creating holes; and a N-type crystal is produced.

P-Type Crystals. When impurities such as aluminum, gallium, or indium are added to a germanium crystal, the conductivity will also increase greatly. However, the current carriers will be holes; the number of free electrons will be negligible. This behavior is not difficult to reconcile. These impurities (aluminum, gallium, and indium) have three outer shell electrons. When they are added to a germanium crystal in small quantities, they will take the place of a germanium atom in the crystal structure; but, in the formamation of the covalent bonds, there will be a gap left near the impurity atom because of the absence of a forth outer shell electron (figure 2.8). At very low temperatures, this vacancy in the covalent bonds will be fixed to the impurity atom because of its lesser net charge. However, even at low temperatures, the thermal energy of the valence electrons will become great enough to overcome this differential in net atomic charge. Other valence electrons will then be able to move in to fill this vacancy, and it will range through the crystal with a random thermal motion. Hence, free holes are produced in the outer shell electron structure, giving a higher concentration of hole carriers than electron carriers.

It can be seen that the addition of suitable impurities to a semiconductor crystal will produce either P or N type material. The process of adding controled amounts of impurities is refered to as doping. The group V impurities (phosphorous, arsenic, and antimony) are called donars because they donate a free electron to the electronic structure of the crystal. The majority carriers in this N-type crystal will be electrons. The group III impurities (boron, aluminum, gallium, and indium) are called acceptors because they can accept an electron to complete their covalent bonds. These impurities will, therefore, produce vacancies in the outer shell electron structure, or hole current carriers. In such a crystal (P-type) the holes are called majority carriers.

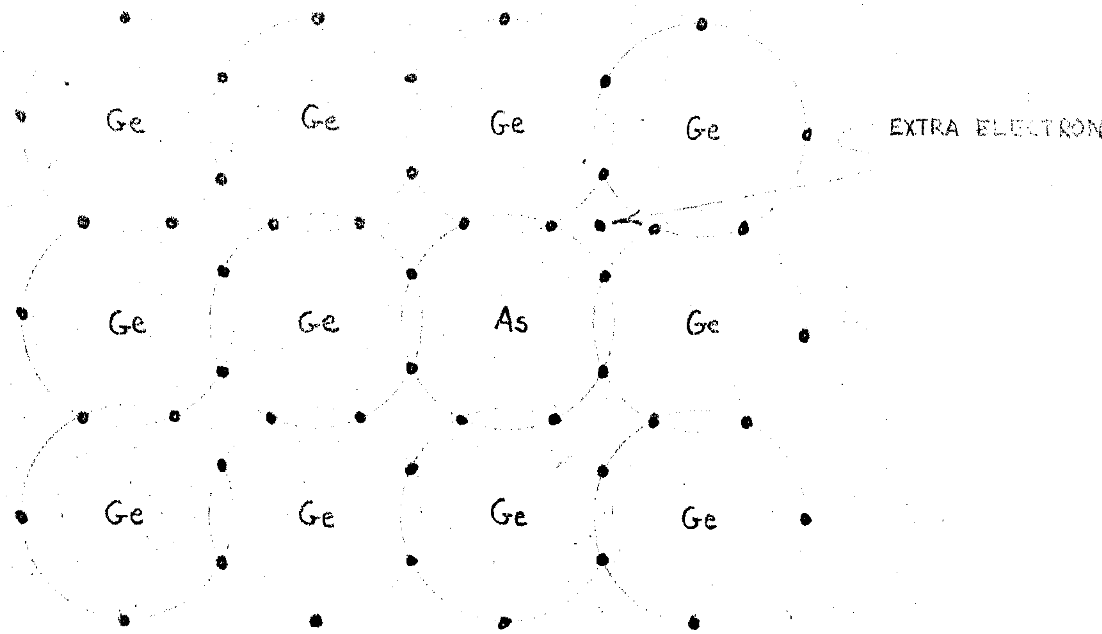


Figure 2.7. Covalent Bonding of an Arsenic Impurity in a Germanium Crystal.

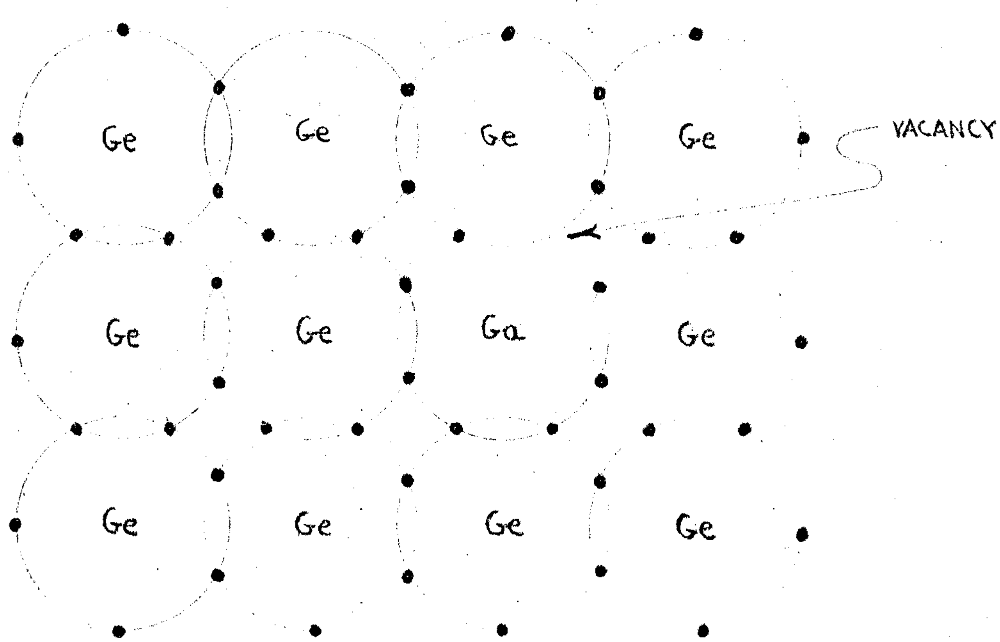


Figure 2.8. Covalent Bonding of a Gallium Impurity in a Germanium Crystal.

Only a small amount of impurities can be added if a continuous crystal structure is to be maintained. The usual concentration is around 10^{16} impurity atoms per cubic centimeter, or about one in 10^8 germanium atoms. If the impurity concentration is made much greater than one part per million, the desirable properties of the crystal will be destroyed by the local electric fields associated with the impurity atoms.

Effect of Temperature on Doped Crystals. It has already been stated that the conductivity of a pure semiconductor crystal increases rapidly with temperature. This characteristic is modified by the presence of impurities as is shown in figure 2.9 where the variation of conductivity with temperature is plotted for both doped and undoped crystals.

At temperatures near absolute zero very few of the current carriers provided by the impurity atoms will be available for conduction. The thermal energy of these carriers is not sufficient to overcome the unbalanced electric field near the impurity atoms (the net charge of a donor impurity is one unit greater than the atoms of the host crystal and that of an acceptor is one unit less). As temperature is increased, the carriers will gain enough thermal energy to break away from the impurity atoms. This process is called impurity ionization and is responsible for the sharp increase in conductivity shown in figure 2.9. A temperature will soon be reached (-225°C to -175°C) where practically all the impurities become ionized; further increases in temperature will not significantly increase the number of current carriers. In this region (extrinsic region) the conductivity will drop somewhat because increased thermal motion improves the probability of collision between the current carriers and the immobile atoms, thus reducing their average velocity for a particular applied field.

At high temperatures (about 100°C for germanium, or 200°C for silicon), the number of thermal ionizations from the covalent bonds becomes appreciable. This will increase the number of current carriers so the conductivity will again increase as shown in figure 2.9. In this region, thermally generated hole-electron pairs, as well as the majority carriers provided by the impurities, will contribute to any current through the crystal. Since the operation of most semiconductor devices depends on the existence of a majority carrier, the semiconductor is rendered useless in this region. This phenomenon is responsible for the sometimes severe temperature limitations imposed on diodes and transistors.

Figure 2.9 shows that the extrinsic region can be extended over a wider temperature range by increasing the impurity concentration. This improvement is greater than might be expected: At a given temperature the minority carrier concentration is lower in a heavily doped crystal. The presence of a larger number of majority carriers increases the probability of recombination between majority and minority carriers (holes and electrons or vice versa) and therefore suppresses the minority carrier concentration. This fact is utilized to improve high temperature performance of semiconductor devices, but it cannot be considered a cure-all since an increased impurity concentration may deteriorate other desirable characteristics.

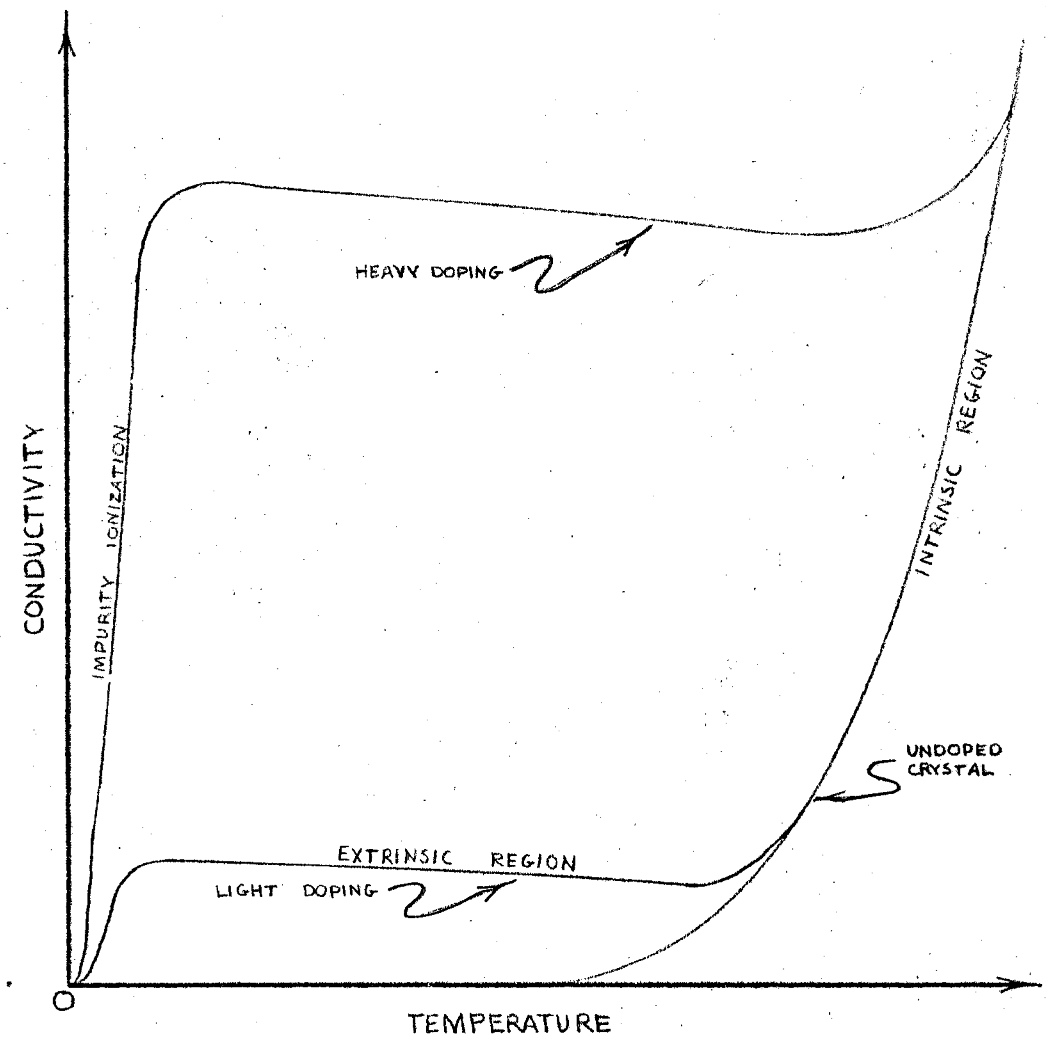


Figure 2.9. Variation of Conductivity with Temperature for Semiconductor Crystals Having Different Impurity Concentrations.

MECHANISMS OF CONDUCTION IN SEMICONDUCTORS

If an electric field is applied across a semiconductor crystal containing mobile current carriers, a net motion in the direction of the field will be superimposed on the random thermal motion of the carriers. This is illustrated in figure 2.10. The average velocity in the direction of the field will depend on the electric field intensity and the ease with which the carriers can move through the material. This latter quantity is defined as the mobility of the current carriers. Mobility will depend on the number of collisions suffered by the carriers in moving through the crystal and will, therefore, be a function of temperature. The average forward velocity is called the drift velocity, and the process of conduction by mobile carriers in an electric field is refered to as drift in a potential gradient where the potential gradient is the change in electric potential with distance through the crystal, or the electric field intensity. Drift is only one possible mechanism of conduction; it is also possible to establish a current in the absence of an electric field.

Thermal Diffusion. If a volume contains a large number of particles moving with a random thermal motion, there will be a force present to equalize the distribution of particles throughout the volume. Considering only a small element of the total volume, the number of particles moving out of the element will be porportional to the number of particles contained by the element since the motion is entirely random. Similarly, the number of particles moving into the element will be porportional to the particle density in the surrounding volume. It follows that if the density of one of these regions differs from the density of the other, there will be a net motion of particles from the region of high density to that of lower density. This process will continue until the densities become equalized; then there will be as many particles moving in one direction as there is in the opposite direction and no net motion.

The process involving the net transfer of particles from regions of high density to regions of low density, driven by thermal forces, is called diffusion. The rate of diffusion will be porportional to the density gradient which is the variation of density with distance through the volume. The diffusion rate is also dependent on the magnitude of the thermal motion and is, therefore, dependent on temperature.

The diffusion process is illustrated in figure 2.11. Specific examples are not difficult to find: If a gas is released in an evacuated container, it will diffuse through the container until a uniform distribution is reached. It should be evident that this is a thermal diffusion process since the gas will condense into a liquid if the temperature is lowered enough. The laws of diffusion also hold if there is a nonuniform distribution of one substance within another.

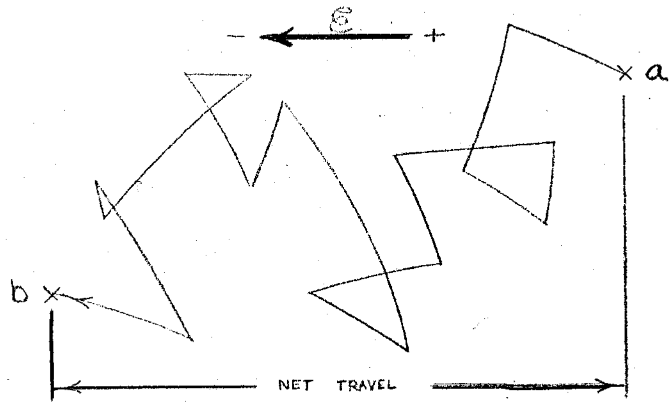
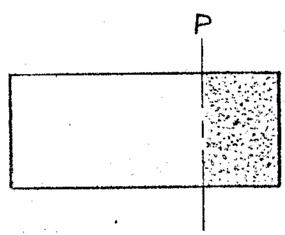
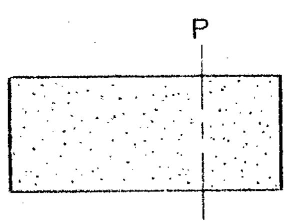


Figure 2.10. Modulation of the Random Thermal Motion of a Hole by an Electric Field.



a. Before Diffusion.



b. After Diffusion.

Figure 2.11. Thermal Diffusion of Uncharged Particles to Equalize Distribution.

For example, if ammonia gas is released into the atmosphere it will diffuse to equalize its distribution. The same is true for a drop of ink in a glass of water, or even for two solids in contact. However, in these examples, the diffusion rate will not only depend on the density gradient and the temperature, but it will also depend on the ease with which one substance can move through the other.

Diffusion of Charged Particles. The discussion thus far refered to the thermal diffusion of uncharged particles. Diffusion forces will also act on charged particles. However, the electric fields produced by the displacement of the particles must be considered. If a doped semiconductor crystal is brought into contact with an undoped crystal, current carriers will diffuse from the doped to the undoped crystal because of the density variation; but the current carriers moving into the undoped crystal will set up an electric field which will oppose further diffusion. This is only one example of a combination of thermal and electrical forces acting on charged particles; others will become apparent in later chapters.

When holes are injected into a N-type crystal one end of which is connected to a sink (ground) as shown in figure 2.12a, an electric field will be set up by the excess positive charge. (Note: How this injection takes place is of little importance here, but an example has already been given in chapter 1 where holes were injected into the N-type base of a transistor by forward biasing the emitter base junction.) Both the injected holes and the free electrons will be acted upon by this field and will move through the crystal. The end result is illustrated in figure 2.11c. Before the holes can move very far into the crystal, the free electrons will move in and neutralize the electric field; the extra electrons needed to accomplish this being drawn in from the sink. This phenomenon is known as space charge neutralization since it neutralizes the excess charge produced by carrier injection (analagous to the space charge around the cathode of a vacuum tube). The suppression of this charge permits the injection of large currents at relatively low voltages and is partly responsible for the high efficiency of semiconductor devices.

Diffusion Current. After space charge neutralization takes place, there will be no net electrostatic force acting on the current carriers. Within the crystal, the holes are neutralized locally by the presence of excess electrons; however, these electrons will not recombine immediately with the holes (fall into the valence vacancies) since their energy is too high. In the absence of an electric field, the holes will be acted upon solely by thermal diffusion forces and will diffuse through the crystal toward the sink, since the hole density is lower in this region. If the hole injection is continued, a current will be established through the crystal without the aid of an electric field; this is called a diffusion current.

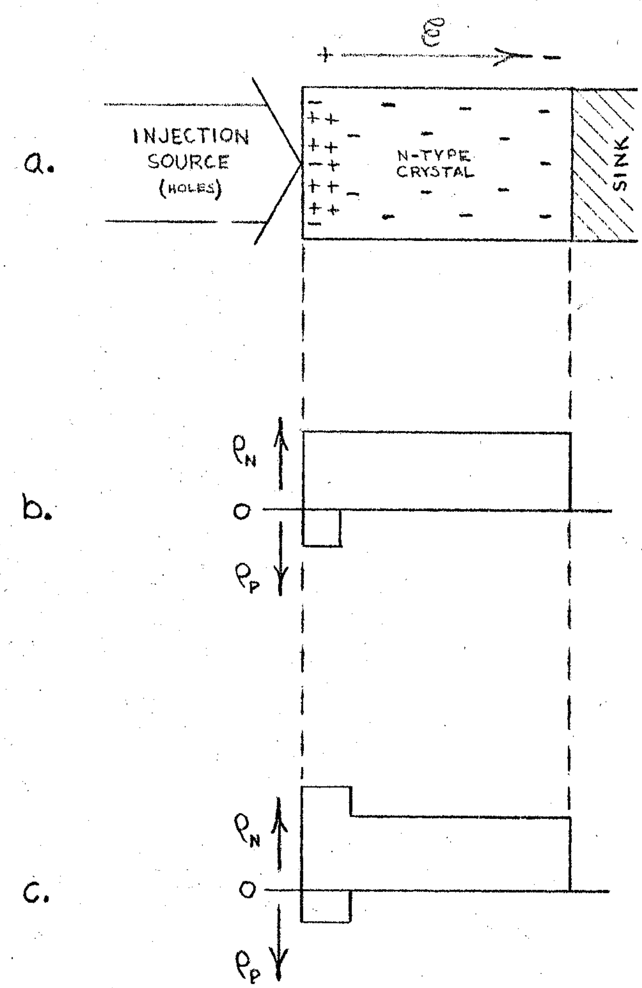


Figure 2.12. Electric Field Produced by Injection of Minority Carriers (a). Density of Holes (ρ_P) and Electrons (ρ_N) with Distance through Crystal before Space Charge Neutralization (b). Neutralization of Electric Field by Displacement of Majority Carriers (c).

If holes are injected at a constant rate, they must also diffuse through the crystal at a constant rate to prevent an accumulation of charge. When the number of recombinations is negligible, this requires linear variation of hole density with distance through the crystal so that the diffusion rate will be constant. This condition is shown in figure 2.13a. The entire current is carried by the diffusion of holes: The electrons remain fixed to neutralize the excess charge.

Recombination. If the diffusion path is made sufficiently long, the number of recombinations will become significant, particularly if the crystal contains many defects at which the free electrons can loose their excess energy and fall into the valence vacancies. The average time that a minority carrier can exist in a crystal containing a majority carrier is defined as the minority carrier lifetime and is a measure of crystaline perfection. For crystals of good quality, the minority carrier lifetime ranges from 10 to 1000 ys.

Under conditions of appreciable recombination, the density distribution shown in figure 2.15b can be expected. Here, current is carried both by holes diffusing into the crystal from the injection source and by electrons moving in and recombining with the holes. The sum of the hole and electron currents must be equal to the total current at any distance down the crystal: Current is carried almost entirely by holes near the injection source and almost entirely by electrons near the sink. The average distance that a minority carrier can travel into the crystal before recombination occurs is defined as the diffusion length. This quantity is of particular importance in a transistor; the base width must be much less than the diffusion length to minimize the number of recombinations in the base region.

Conclusions. It has been shown that there are two mechanisms of conduction in a semiconductor: Drift in a potential gradient and diffusion in a density gradient. It is possible for both of these mechanisms to act simultaneously, even with one opposing the other. In any event, the drift current will be directly porportional to the potential gradient, the number of current carriers involved, and the carrier mobility; and the diffusion current will be a direct function of the density gradient, the number of carriers setting up this gradient, the absolute temperature, and the carrier mobility. These mechanisms are of equal importance. The operation of semiconductor devices cannot be adequately explained unless both are given due consideration.

SEMICONDUCTOR MATERIALS

Thus far, the discussion of semiconductor materials has been confined to those elements having four valence electrons which form a completed electron structure as shown in figure 2.4. These are the group IV elements listed in table 2.1. Nonetheless, combinations of elements will sometimes exhibit semiconductor properties; and the semiconductor properties can usually be predicted by analyzing the electron structures of the combining atoms. For example, if a crystal is composed of an equal number of group III and group V elements, it can exhibit semiconductor properties. This happens because the atoms with five valence electrons will donate an electron to the atoms with three valence electrons in the formation of covalent bonds. The electronic structure will then be similar to that of a group IV crystal. Semiconductor properties also arise from combinations of group II and group VI elements. In this case, the atoms with six valence electrons donate two to the atoms with only two valence electrons, again forming the completed electron configuration of group IV crystals. Compounds of group I and group VII will not show semiconductor properties since a complete transfer of electrons takes place forming ionic rather than covalent bonds. This will sometimes happen in the group II-VI and group III-V compounds.

In group IV crystals, it is necessary to add impurities to produce majority carriers; however, in semiconductor compounds, this can be accomplished by increasing the concentration of one element with respect to the other. To illustrate, if there is an excess of the group III element in a group III-V compound, there will be a deficiency of electrons in the covalent bonds so the crystal will be N-type. If there is an excess of the group V element, electrons will be left over after the covalent bonds are formed so the crystal will be P-type.

Other covalent bond structures can give rise to semiconductor characteristics. Selenium and tellurium in group VI exhibit semiconductor properties, as do compounds like lead sulfide and lead telluride. Semiconductors have also been made from combinations of three or more elements. The valence structure of these elements is not similar to that of the group IV crystals, but is quite complex. Little is known about them at the present time which accounts for the fact that they are not too widely used.

Characteistics of Materials. Although there are a large number of semiconductor materials, relatively few of them are presently being used. The major problem with most materials is the difficulty in producing crystals of adequate purity and perfection. This is more or less a problem of techniques and can probably be solved for a particular material if it shows enough potential.

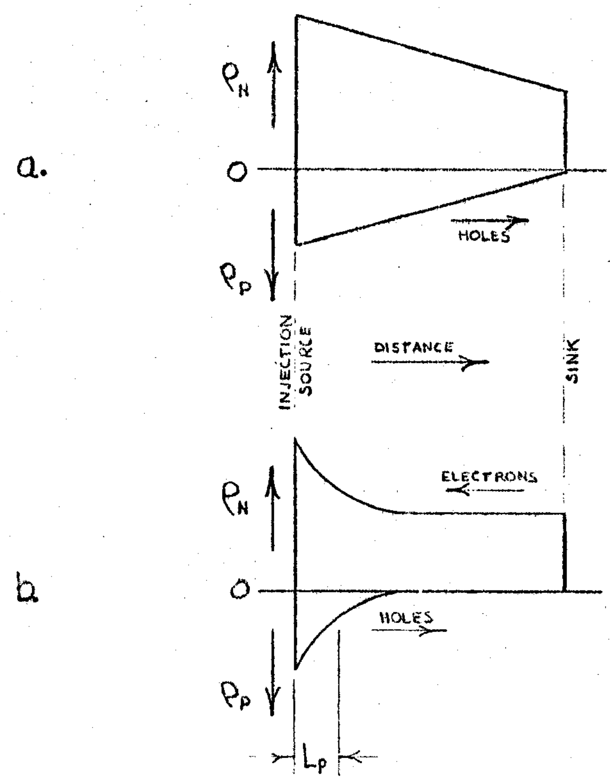


Figure 2.13. Distribution of Current Carriers through a Crystal for a Constant Diffusion Current (a) when there is Negligible Recombination and (b) when there is Appreciable Recombination.

The usefulness of a semiconductor material is ultimately detemined by its physical characteristics. The most important of these are the energy required to ionize electrons from the covalent bonds, which determines the maximum operating temperature, and the mobility of the current carriers. These quantities are given for several materials in table 2.2.

Ionization Energy. The ionization energy is the minimum amount of energy that will free an electron from the covalent bonds. This quantity is zero for a metalic conductor, several electron volts for an insulator, and about one electron volt for a semiconductor. For most applications a high ionization energy is desirable to limit the number of thermally generated hole-electron pairs, since the desirable properties of the semiconductor generally depend on the existence of a majority carrier. If tables 2.1 and 2.2 are compared, a definite trend is evidenced: Semiconductors composed of elements with a low atomic number will have high ionization energies while those composed of elements with high atomic numbers will have low ionization energies. This can be explained by considering the atomic structure. Elements with high atomic numbers will have a greater number of electrons shielding the positive charge of the nucleus. Therefore, the outer shell electrons will be more loosely bound. This is carried over in the formation of covalent bonds.

With elements of higher atomic number, the covalent bonds are sometimes formed in the inner electron shells (secondary velances). This will explain why lead and some forms of tin are metalic conductors, showing no semiconductor properties.

The above discussion is not meant to imply that semiconductors having low ionization energies are useless. In some applications a low ionization energy is required. For example, infared radiation can be detected when it produces ionizations in the covalent bonds of a semiconductor, thus increasing its conductivity. The energy of these radiations is in the order of 0.3 electron volt so a semiconductor that has an ionization energy lower than this value must be used. In these applications, the random thermal ionizations are usually reduced by cooling the crystal to liquid nitrogen temperatures (-196°C).

Mobility. It is desirable that the carrier mobility in a semiconductor be as high as possible since both drift velocity and diffusion rate are dependent on this quantity. Furthermore, if a semiconductor is to be used in a diode, transistor, or any other device where holes and electrons carry current simultaneously, both the hole and the electron mobilities should be high. Table 2.2 shows that there is a wide varience in mobilities for different materials. Some speculation can be made to explain these variations on the basis of atomic structure. Certain trends can be established by examining table 2.2:

Table 2.1. List of Selected Elements from the Periodic Chart.

GROUP				
II	III	IV	V	VI
4 BERYLIUM	5 BORON	6 CARBON	7 NITROGEN	8 OXYGEN
12 MAGNESIUM	13 ALUMINUM	14 SILICON	15 PHOSPHOROUS	16 SULPHUR
30 ZINC	31 GALLIUM	32 GERMANIUM	33 ARSENIC	34 SELENIUM
48 CADMIUM	49 INDIUM	50 TIN	51 ANTIMONY	52 TELLURIUM
80 MERCURY	81 THALLIUM	82 LEAD	83 BISMUTH	84 POLONIUM

Table 2.2. Characteristics of Some Semiconductor Materials.

GROUP	MATERIAL	IONIZATION ENERGY (eV)	MOBILITY (cm^2/volt·sec)	
			HOLES	ELECTRONS
IV	Carbon	6.0	1200	1800
	Silicon Carbide	3.0	10	50
	Silicon	1.1	500	1200
	Germanium	0.7	1900	3800
	Tin	0.08	1000	2000
III, V	Aluminum Antimonide	1.6	300	1200
	Gallium Arsenide	1.4	250	4000
	Indium Phosphide	1.25	650	3400
	Gallium Antimonide	0.8	850	4000
	Indium Arsenide	0.4	200	30,000
	Indium Antimonide	0.18	1250	77,000
II, VI	Zinc Oxide	3.2	180	200
	Cadmium Sulfide	2.4	——	210
	Mercury Selenide	0.16	——	15,000

(1) mobility will generally be higher in crystals containing atoms of greater atomic weight, (2) electron mobility will be higher when there is a difference in the atomic size of the atoms making up the crystal, and (3) the hole mobility will be less in crystals composed of a mixture of atoms.

The first trend, an increase in mobility with crystals of higher atomic weight, can probably be explained by considering the thermal vibrations of the atoms which produce scattering of the current carriers moving through the crystal. The amplitude of these vibrations will be inversely porportional to the atomic weight, but will also depend on the strength of the bonding between atoms which explains departures from this tendency for crystals with very high and very low binding energies.

The second trend, higher electron mobility when there is a difference in the atomic size of the atoms making up the crystal, is probably caused by a greater free space between atoms. This will be particularly true for the larger atoms.

The third trend, decreased hole mobility in crystals composed of a mixture of atoms, is probably the result of local strains produced in the covalent bonds when one atom donates more electrons than the other. These nonuniformities will impede motion of the holes through the crystal.

Crystal Quality. For semiconductor purposes, single crystals are needed: That is, the covalent bonds must be continuous throughout the crystal. Minor defects are produced when atoms are missing from the crystal lattice or when impurities are present. The minority carrier lifetime is extremely sensitive to the number of defects present since recombinations can take place at imperfections in the crystal structure. This represents a major problem with most materials.

Practical Materials. At the present time only germanium and silicon are used in the production of transistors. Their ionization energy is high enough to permit operation at reasonably high temperatures (to about 100°C for germanium and 200°C for silicon); and the carrier mobilities are acceptably high. Furthermore, techniques have been developed to produce the high purity crystals necessary in this application. The technology of germanium is somewhat more advanced than silicon with the result that better crystals can be grown from germanium.

Minority carrier lifetime is not so critical in a junction diode; hence, crystals of lower quality can be used. Germanium and silicon are by far the most popular material for the construction of diodes, but diodes with a maximum operating temperature of 500°C have been made from silicon carbide. The performance of these diodes is somewhat poor because of the low carrier mobility in silicon carbide crystals. Gallium arsenide has been used in the construction of tunnel diodes where the minority carrier lifetime is of no significance whatsoever.

Carbon (diamond) is a promising material for very high temperature operation since it has a high ionization energy and respectable carrier mobilities. However, the difficulties encountered in purifying the material and growing high quality crystals have yet to be overcome.

DEPARTMENT OF WEAPONS TRAINING
LOWRY AIR FORCE BASE
COLORADO

T R A N S I S T O R S

IN SERVICE TRAINING COURSE

CHAPTER 3

THE PN JUNCTION

4 JANUARY 1961

FOR INSTRUCTIONAL PURPOSES ONLY

CHAPTER 3

Author: A1C Robert J Widlar

THE PN JUNCTION

Abstract - The characteristics of the PN junction diode and several other practical PN junction devices are investigated. The formation of an electric field in an insulating region between the P and N type materials is shown to explain the rectifying properties of such a junction. The effects of temperature on diode characteristics are discussed as are reverse breakdown mechanisms. The dependence of diode capacitance on reverse bias voltage is shown. The operation of various photoconductive and photovotaic cells is described, and the tunnel diode is given an elementary explanation. Both rectifying and ohmic metal-to-semiconductor contacts are mentioned.

GENERAL

The concepts of conduction in solids developed in the last chapter are a powerful tool for the understanding of semiconductor devices. In this chapter, they will be applied to the PN junction which has already been mentioned. Devices employing a single PN junction are capable of performing many useful functions that were not brought out in chapter 1. These functions include: voltage regulation (zener diode), frequency control (voltage variable capacitor), detection of radiant energy (photodiode), direct conversion of solar energy into electricity (solar cell), and the amplification of microwave signals (tunnel diode). An explanation of these devices will more completely describe the behaviour of a PN junction.

Since transistors and many other semiconductor devices employ multiple PN junctions, a study of the PN junction alone can explain certain phenomena which may be equally applicable to these more complex devices but more difficult to isolate in them. This chapter will build a foundation for the coming discussion of transistors.

THE PN JUNCTION

If a slab of P-type semiconductor is brought into intimate contact with a slab of N-type material, a PN junction will be formed. (This is not a practical method of making such a junction because of the discontinuous crystal structure that will exist at the interface of the two materials; but, for the present it will be assumed that in bringing the two slabs together, they are fused into a single crystal.) When the two materials are brought into electrical contact, holes will diffuse from the P-type material into the N-type material; and electrons will diffuse from the N-type into the P-type material. This is to be expected because these carriers, driven by thermal forces, will try to equalize their distribution throughout the crystal.

This diffusion will soon be brought to a halt by the unequal charge distribution set up by the displaced carriers. That is, if a free electron diffuses across the junction into the P-type region, it will leave behind an unneutralized donor ion which has a net positive charge. Moreover, after it diffuses into the P region, it will eventually recombine with a hole and create an unneutralized acceptor ion having a net negative charge. Similarly, if a hole diffuses across the junction, it will leave behind a negative acceptor ion. Crossing the junction, it will recombine with an electron and produce a positive donor ion. These unneutralized impurity ions will produce an electric field across the junction. As can be seen from the illustration in figure 3.1, this field will oppose diffusion of current carriers across the junction. This electric field is called a barrier.

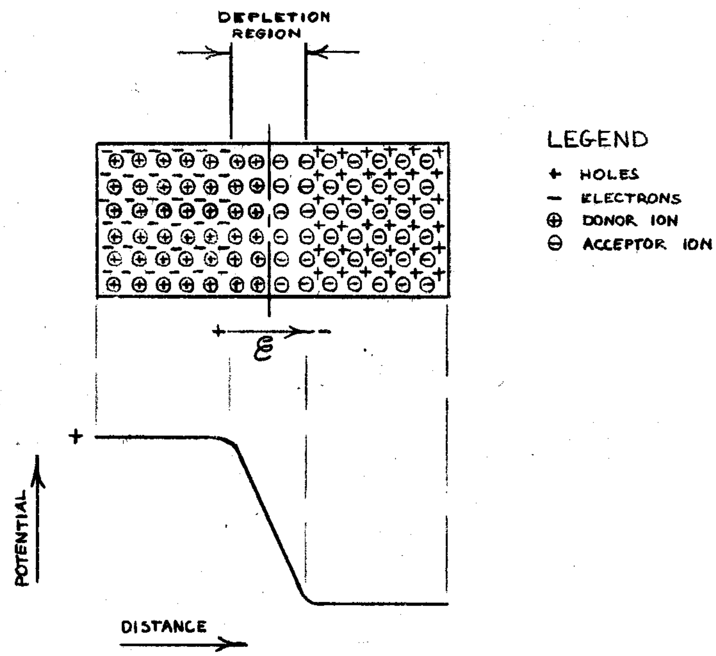


Figure 3.1. Formation of Barrier Potential in a PN Junction from the Diffusion and Recombination of Current Carriers which Exposes Unneutralized Impurity Ions in the Vicinity of the Junction.

The potential difference created across the junction by the unneutralized impurity ions is frequently refered to as the barrier potential. Although small, this potential (in the order of 0.3 volts for germanium and 0.5 volts for silicon) will set up a relatively strong electric field because it is confined to a narrow region near the junction. This region, called the depletion region will normally contain no current carriers. Any current carriers present within the depletion region will be swept away immediately by the electric field.

The barrier potential cannot be measured by ordinary means. A contact potential will be generated at any metalic contact on the crystal. If two probes are placed on the semiconductor, one on each side of the junction, the sum of the barrier and contact potentials will be zero so there will be no potential difference between the probes. A further explanation of this phenomenon involves a treatment of metal-semiconductor contacts which will be covered at the end of this chapter.

Forward Characteristics. If a small external potential is applied to a PN junction, positive to the P-type region and negative to the N-type region, it will produce an electric field in the depletion region opposing the barrier field. This will reduce the barrier field and permit diffusion of high energy current carriers across the junction. For applied potentials less than the barrier potential, the electric field within the crystal will be confined almost entirely to the depletion region because the number of current carriers present there is very much less than in the rest of the crystal, making it a high resistivity region. Under these conditions, current is established through the crystal primarily by thermal diffusion which is opposed by any existing barrier field. Reductions in the barrier potential permit lower energy carriers to diffuse across the junction in addition to the high energy carriers, thus increasing the current.

As current through the crystal increases, the concentration of current carriers in the depletion region also increases. The resistivity of this region will then approach that of the rest of the crystal, and a nearly uniform electric field will be established across the length of the crystal by the applied voltage. When the barrier is reduced to zero, even the lowest energy carriers will be able to cross the junction so further increases in current will come from an increased accelerating field. The potential variation with distance through the crystal is plotted in figure 3.2 for various values of applied voltage to illustrate these points.

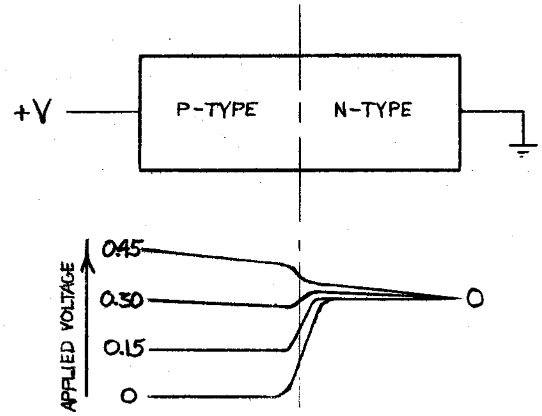


Figure 3.2. Variation of Potential with Distance across a PN Junction for Various Values of Forward Bias.

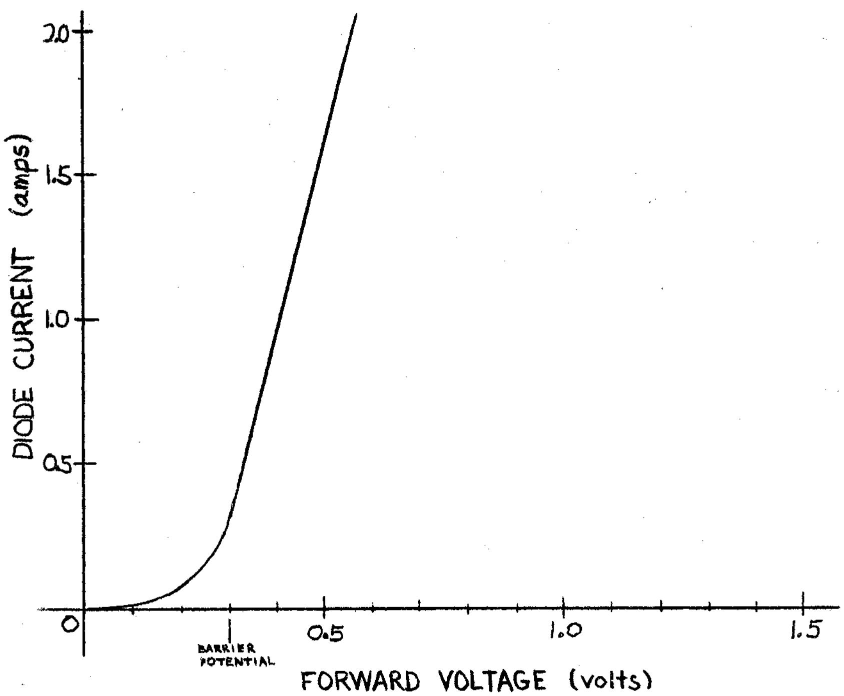


Figure 3.3. Forward Volt-Ampere Characteristics of a Junction Diode.

With this information, it is possible to explain the forward characteristics of a junction diode which are given in figure 3.3. For an applied voltage less than the barrier potential, the current does not increase linearly with voltage. If, for example, the voltage is doubled, the current will be more than doubled. This happens because reducing the barrier by one half will only allow a relatively small number of high energy carriers to diffuse across the junction, but doubling the voltage and eliminating the barrier will permit all the carriers to cross. After the barrier is eliminated, the current carriers are accelerated through the crystal by the relatively strong drift forces of the electric field in addition to the weaker diffusion forces. The velocity of the current carriers, and therefore the current, will depend on the applied voltage. In this region the current increases linearly with voltage.

Reverse Characteristics. If a reverse bias is applied to a PN junction, the barrier potential will be increased by an amount equal to the applied voltage. The entire reverse voltage will be dropped across the depletion region because the absence of current carriers produces a very high resistivity, or insulating, region. There is then no electric field acting on the current carriers in the P and N regions, and the barrier was already high enough to stop the diffusion of majority carriers. Therefore, there will be no current established through the diode.

This is an ideal condition, but the presence of minority carriers prevents its realization. There are holes present in the N-type region and free electrons present in the P-type region from the thermal generation of hole-electron pairs. Any of these minority carriers reaching the junction will be swept across by the barrier field. With no voltage applied to the diode, this reverse current is balanced by the diffusion of high energy carriers across the junction. Increasing the barrier height will reduce the number of high-energy carriers diffusing across the junction in the forward direction while the number of carriers crossing the junction in the reverse direction remains unchanged. Hence, a reverse current will be established.

When a reverse voltage of about one volt is reached, diffusion of even the highest energy carriers is stopped. From this point on, the reverse current is independent of voltage. The magnitude of the current will be determined by the diffusion rate of minority carriers to the junction. Once the carriers reach the junction, they will be swept across regardless of the junction potential.

The reverse characteristics of a junction diode are plotted in figure 3.4. The reverse current is found to increase until the diffusion of high energy carriers across the junction is stopped. Then it reaches a steady value determined only by the diffusion of minority carriers to the junction. This constant value of current is called the reverse saturation current.

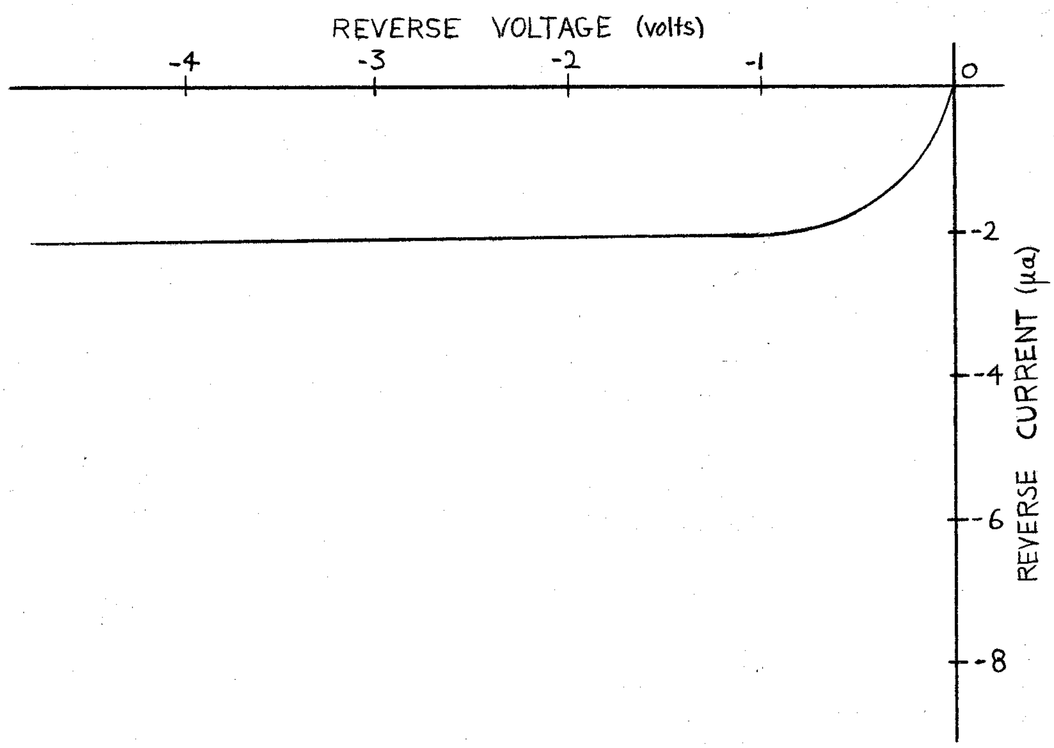


Figure 3.4. Typical Low Voltage Reverse Characteristics of a Junction Diode.

In practical diodes the reverse current will be found to increase slightly with voltage. This increase is caused by leakage on the surface of the semiconductor which contains water and other contaminants that conduct electricity. This leakage path will appear as a high resistance connected in parallel with the diode.

TEMPERATURE EFFECTS

Since the minority carriers in the P and N type regions are created by the thermal generation of hole-electron pairs, the reverse saturation current of a junction diode is dependent on temperature. As temperature is increased, the number of thermal generations increases quite rapidly, increasing the minority carrier concentrations in both the P and N type regions. Consequently, a greater number of carriers diffuse to the junction, thus increasing the reverse current.

The reverse saturation current of a germanium junction diode is plotted as a function of temperature in figure 3.5. At temperatures below 20°C, this current is indeed small; but it increases so rapidly above about 90°C that the diode becomes useless.

Because of the stronger covalent bonding of silicon, higher temperatures must be reached before an appreciable number of hole-electron pairs will be generated. This increases the maximum operating temperature of silicon devices. The plot of reverse saturation current *versus* temperature for a silicon diode has much the same shape as the curve in figure 3.5, except that the maximum operating temperature falls at approximately 200°C.

The forward characteristics of a junction diode are not too greatly affected by temperature. At higher temperatures, there is, in general, a small increase in the low voltage conductance and a small decrease in the conductance at higher voltages. This is shown in figure 3.6.

Increasing the temperature will increase the thermal energy of the majority carriers. Hence, when the barrier potential is reduced a given amount by an applied voltage, more carriers will be able to diffuse across at higher temperatures. After the barrier is completely eliminated, conduction through the diode is dependent on the resistivity of the semiconductor material in the body of the diode, so the conductance will decrease at higher temperatures due to the lowered carrier mobility.

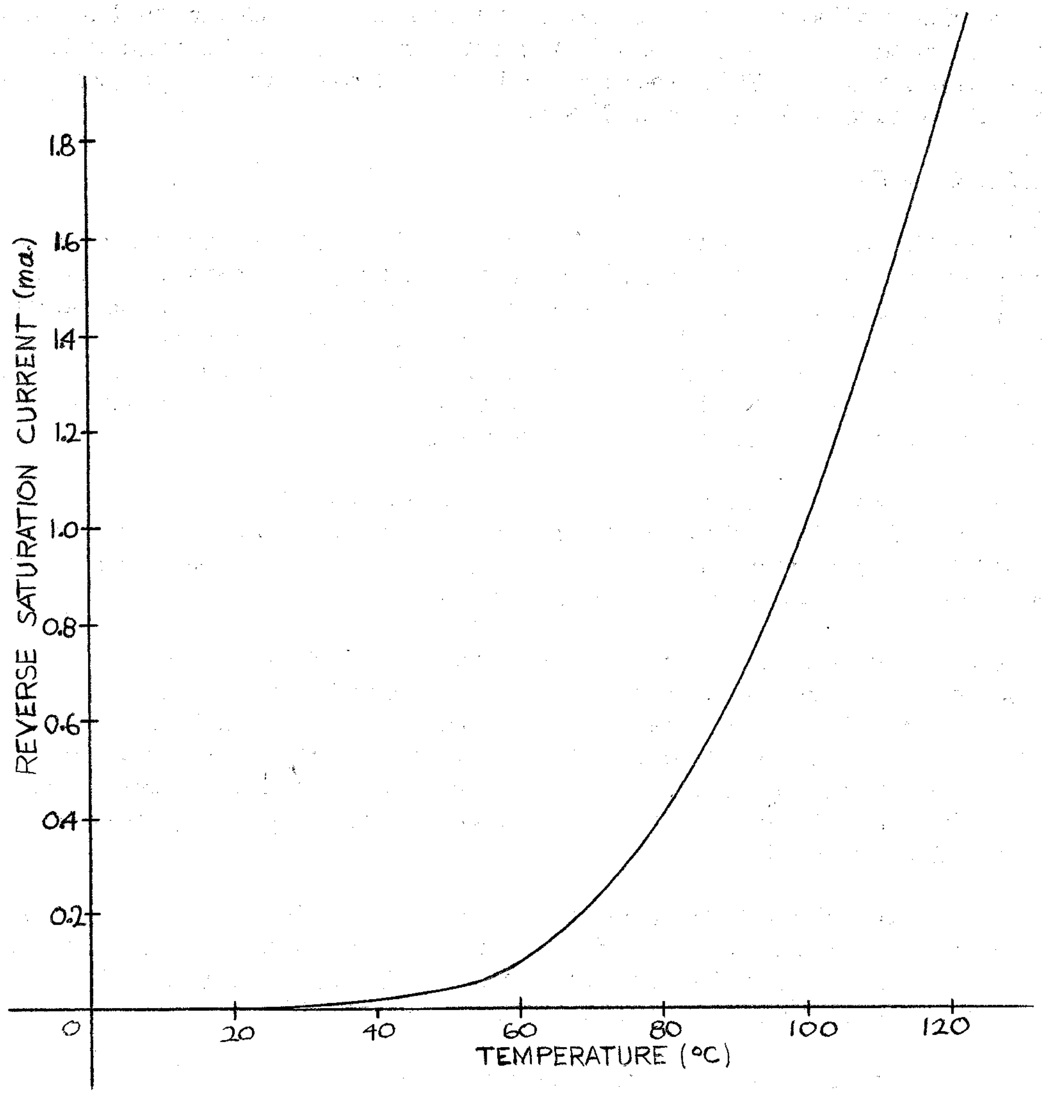


Figure 3.5. Plot of Reverse Saturation Current Versus Temperature for a Germanium Junction Diode.

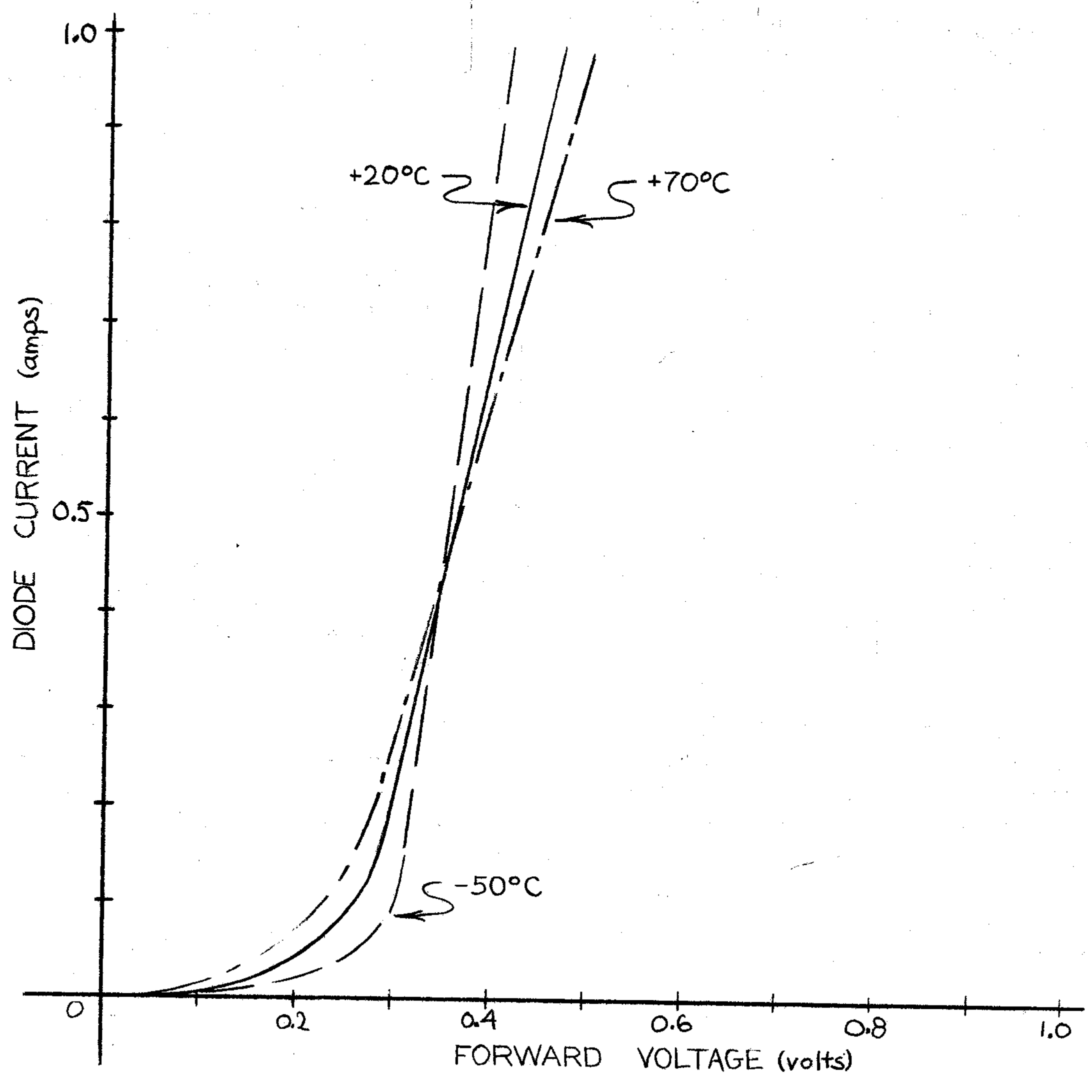


Figure 3.6. **Change in the** Forward Characteristics of a Germanium Junction Diode with Temperature.

REVERSE BREAKDOWN PHENOMENA

As the reverse voltage on a diode is increased, a point will eventually be reached where the diode begins to conduct heavily. This phenomena is known as reverse breakdown. There are three possible reverse - breakdown mechanisms: Thermal, avalanche, and zener breakdown. Whichever mechanism occurs first will temporarily destroy the reverse-blocking characteristic of the diode.

Thermal breakdown generally occurs in point contact diodes. In these diodes, the junction is formed near a point contact on the surface of the crystal. There are a large number of defects present near the surface, so point contact diodes are usually characterized by a high reverse current. At higher reverse voltages, this current gives rise to an appreciable power loss which is confined to a small volume near the point contact. This power dissipation will cause excessive heating of the junction, increasing the number of thermally generated current carriers and, therefore, the reverse current. This action is cumulative. After a certain voltage is reached (peak inverse voltage), the reverse current and the power dissipation will increase so rapidly that the reverse voltage will fall off with increasing current, producing the characteristics shown in figure 3.7. In the case of thermal breakdown, the diode will not be damaged if the power dissipation is kept low enough so that the junction does not melt.

Junction diodes usually undergo avalanche breakdown. As the reverse voltage is increased, the minority carriers that are swept across the junction by the barrier field are accelerated enough to excite electrons from the covalent bonds when they collide with the atoms of the crystal. These electrons will ionize other electrons giving rise to a cumulative action. This is similar in many ways to the breakdown of a gas diode. Avalanche breakdown is characterized by a sharp increase in reverse current at a nearly constant voltage as is shown in figure 3.8.

Zener breakdown, also common to junction diodes, occurs when the electric field across the junction becomes strong enough to rupture the covalent bonds. Even though the reverse voltage might be relatively small, the entire voltage is dropped across the narrow depletion region. This could create a intense field that is strong enough to ionize electrons directly from the covalent bonds. Zener breakdown is also characterized by the constant voltage characteristic shown in figure 3.8.

It is difficult to tell whether a particular diode undergoes avalanche or zener breakdown because of the great similurity of these two mechanisms. However, it is generally felt that the breakdown of junction diodes is caused by avalanche or in some cases a combination of avalanche and zener breakdown.

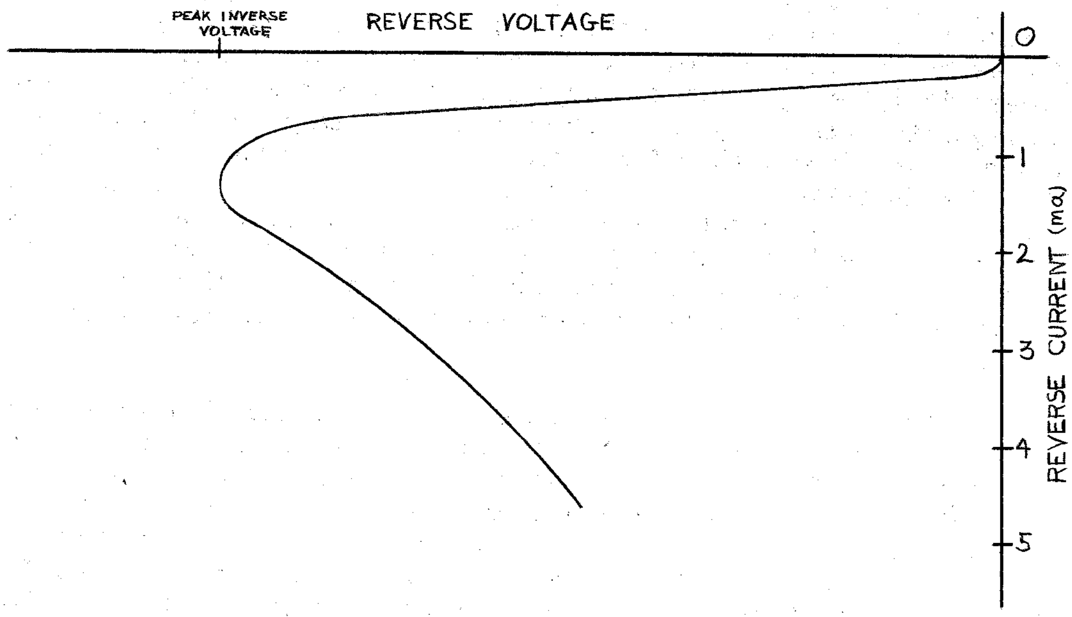


Figure 3.7. Thermal Breakdown Characteristics of a Point Contact Diode.

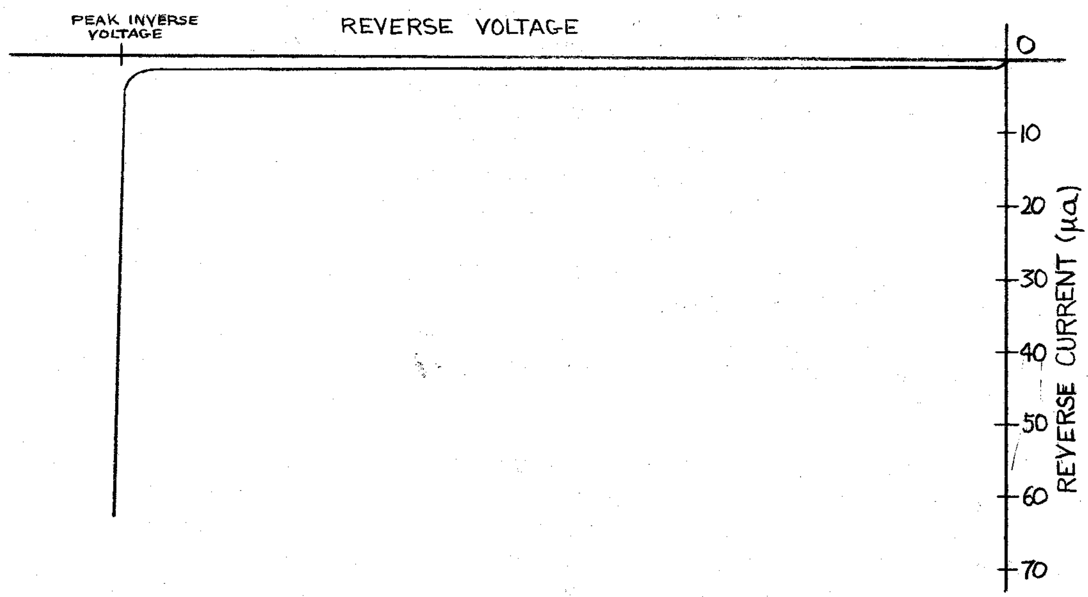


Figure 3.8. Avalanche or Zener Breakdown Characteristics of a Junction Diode.

Zener Diodes. A junction can be used as a voltage regulator by taking advantage of its constant voltage reverse breakdown. With avalanche or zener breakdown, the diode will not be damaged unless the power dissipated generates enough heat to physically alter the junction. In regulator circuits, it is frequently necessary to dissipate considerable power in the diode; therefore, provisions must be made to remove heat from the junction. This can be accomplished by the normal processes of radiation from an encapsulated diode; or, more efficiently, by soldering one end of the diode directly to a copper base and attaching this to a heat sink. Smaller units using the former method will dissipate about 150 mw while larger units mounted on a heat sink can handle up to about 50 watts.

Although their circuit applications are similar, zener diodes provide a greater flexability than gas tube regulators since zener diodes are available with reverse breakdown voltages anywhere between 0.5 and 400 volts. The breakdown voltage is determined by the doping of the P and N regions of the diode during manufacture. The degree of doping will determine the width of the depletion region and, therefore, the intensity of the electric field across the junction for a given reverse voltage (the entire reverse voltage is dropped across the depletion region). If the depletion region is made thinner, the field intensity will be greater; and the breakdown voltage will be lower.

If a diode is made from heavily doped P-type material and lightly doped N-type material, the depletion region will extend primarily into the N region. This happens because an equal number of donor and acceptor ions will be exposed in the formation of the barrier: When a hole diffuses across the junction, it will leave behind an unneutralized acceptor ion; and when it recombines with an electron in the N region, it will create an unneutralized donor ion, etc.. Therefore, the depletion region must extend farther into the lightly doped N region than into the heavily doped P region to unneutralize the same number of impurity atoms on both sides of the junction. In this case, then, the width of the depletion region and, consequently, the reverse breakdown voltage can be controlled during manufacture by the degree of doping in the N-type region. (This is one possible method. It is also possible to alter doping in the P-region or in both regions and achieve similar results.)

The name, zener diode, is somewhat misleading since either avalanche or zener breakdown could take place in these devices. It is generally felt that zener breakdown occurs in highly doped diodes with narrow junctions and a high electric field intensity for a given reverse voltage. Avalanche breakdown occurs in diodes with wider junctions where the accelerating field (potential rise over the mean free path of the carriers) is greater.

JUNCTION CAPACITANCE

A reverse biased PN junction will behave like a small capacitance. The depletion region, being devoid of current carriers, acts as an insulator between the conductive P and N regions, thus forming a capacitor. Electrically, this capacitance will appear to shunt the rectifying junction so it can limit the highest operating frequency of the diode. At high frequencies a voltage applied across the diode will be able to pass through the low reactance of the junction capacitance even though the diode is reverse biased.

The magnitude of the junction capacitance will depend upon the dielectric constant of the material from which the diode is made, the width of the depletion region, and the area of the function. The dielectric constant will be fixed by the choice of material (usually germanium or silicon); and the doping, which determines the width of the depletion region is adjusted to give the diode a low forward resistance while maintaining an acceptably high reverse breakdown voltage. Hence, reducing junction capacitance for high frequency operation is usually accomplished by reducing the junction area which also limits the maximum forward current.

In most low frequency applications, the junction capacitance (in the order of several micromicroforads) can be neglected because of its high reactance. Therefore, the junction area can be made quite large to give an increased current capacity.

VARIABLE CAPACTIANCE DIODES

The junction capacitance of a diode will be a function of the reverse voltage. The depletion region will become wider for increased reverse voltage because more of the immobile impurity atoms must be exposed to support the increased potential across the junction (the holes and electrons are pulled away from the junction by the reverse bias). This increases the width of the insulating region between the conductive P and N regions and, therefore, reduces the capacitance.

This effect is optimized in the variable capacitance diode. An abrupt transition between the P and N type materials is used to produce a maximum variation of capacitance with voltage. (A gradual transition would create a region near the junction which contained practically no impurities. This fixed insulating region would reduce the junction capacitance and also the change in junction capacitance with voltage). It is also necessary to reduce the series resistance of the diode body to give a high Q capacitance.

The characteristics of a typical variable capacitance diode are given in figure 3.9. As expected, a decreasing capacitance is shown for increasing reverse voltage.

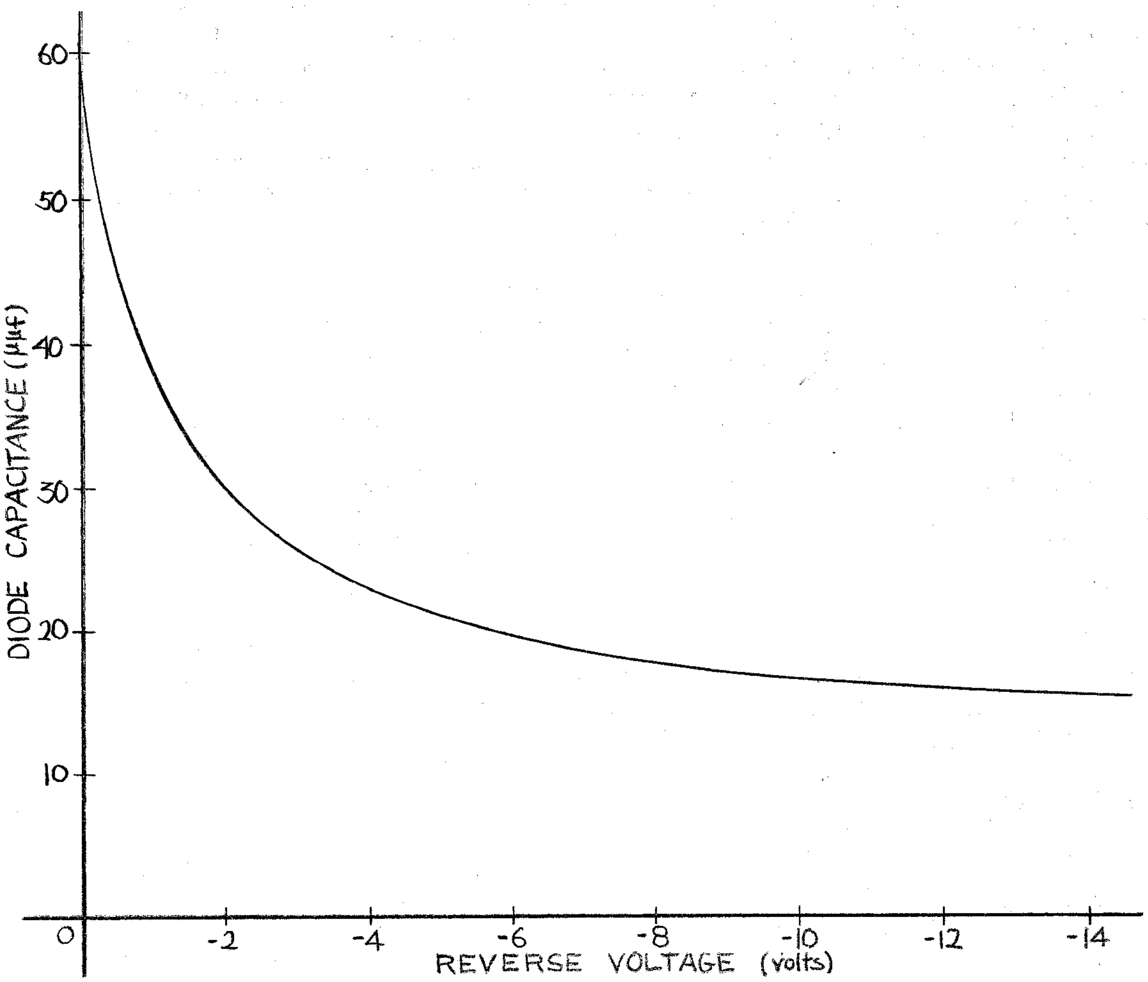


Figure 3.9. Variation of Diode Capacitance with Reverse Voltage for a Variable Capacitance Diode.

PHOTOELECTRIC EFFECTS

Light and other forms of radiant energy can ionize electrons from the covalent bonds of a semiconductor crystal. This will happen only if the photons making up the radiation have sufficient energy. Since the energy of the photons is directly porportional to the frequenty of the radiations, there will be a certain threshold frequency, for any particular material below which ionizations cannot be produced regardless of the radiation intensity. Radiant energy below the threshold frequency will pass through the semiconductor crystal without; but for frequencies very much above the threshold frequency, practically all the radiations are absorbed in producing ionizations. At higher frequencies, the radiations will not penetrate too far into the crystal so the ionizations will take place very close to the surface where recombination is likely to occur. Hence, for a given semiconductor material, there is a range of frequencies that will produce useful ionizations within the body of the semiconductor.

The threshold frequency for germanium and silicon is near the high end of the infrared spectrum, and the photoelectric response extends through visable light into the ultraviolet region. Other semiconductor materials are available (e.g. lead sulfide, lead selonide, indium arsenide, and indium antimonide) which have threshold frequencies farther down into infrared with useful response extending up to visable red.

Photoconductive Cells. Photoconductivity is the decrease in the resistance of a material caused by the increased number of current carriers made available by the absorption of radiant energy. High resistivity semiconductor materials will exhibit photoconductive properties. If a voltage is applied across a pure semiconductor in the absence of light, a small current, called the dark current, will be established due to the presence of uncontrolled amounts of impurities. However, when radiant energy in the proper frequency range is absorbed, electrons will be ionized from the covalent bonds causing an increase in current. This increase will be porportional to the light intensity as this will determine the number of current carriers liberated.

Because it is difficult to produce high resistivity semiconductor materials in practice, the dark current of this type of photoconductive device is quite high so a relatively high light flux must fall on the semiconductor to produce a noticable change in current. A PN junction can be used as a photocell in much the same way; only the dark current will be much less because of the high reverse resistance of such a junction.

If light falls on a back biased PN junction, current carriers generated in the depletion region will be swept away by the reverse voltage field, producing a reverse current. As before, the current will be porportional to the amount of light falling on the junction. Light falling on the bulk of the semiconductor material will be relatively ineffective in producing current because the electric field is confined entirely to the junction region.

Photovoltaic Cells. A PN junction can also be used as a self-generating photocell, or photovoltaic cell. A photovoltaic cell does not require an external voltage source: It will generate its own voltage. Its sensitivity is not as high as that of a photoconductive cell, but the self-generating feature is frequently more desirable than high sensitivity. Furthermore, since photovoltaic cells can convert sunlight directly into electrical energy, they can be used as a power source for portable equipment.

Even though a potential exists across the depletion region of a PN junction diode, it does not appear at the external terminals of the device. As already mentioned, the barrier potential is canceled by the combined contact potentials of the metal-semiconductor contacts to the diode. However, if any one of these potentials is increased or decreased, a voltage will appear at the diode terminals.

When a luminous flux falls on an unbiased PN junction, hole-electron pairs will be generated and swept out of the depletion region by the barrier field. This situation is shown in figure 3.10. The displacement of these charges will reduce the barrier potential making it less than the combined contact potentials; therefore, a voltage will appear at the external terminals of the diode. Furthermore, if a load resistance is placed across the diode, a current will be established, being supplied by the continuous generation of current carriers within the barrier field.

The short circuit current of a photovoltaic cell will depend on the number of current carriers generated in the barrier field so it will be directly porportional to the illumination intensity. The open circuit output voltage will be almost constant over a wide range of illumination intensities.

An output voltage is produced when the barrier potential is lowered by the displacement of current carriers which were generated in the barrier field. But lowering the barrier will also permit the forward diffusion of majority carriers across the junction which produces an opposing displacement of charged carriers. Hence, for a given light flux, the barrier potential will adjust itself until the photoelectric current produced by the generation of hole-electron pairs in the barrier field is balanced by the forward diffusion current across the lowered barrier. Since the number of current carriers with enough thermal energy to surmount the barrier increases quite rapidly as the barrier is lowered, a relatively large increase in illumination will necessitate only a small adjustment in barrier potential for a balanced condition to be reached. Therefore, small changes in illumination intensity will not produce significant changes in the no-load output voltage.

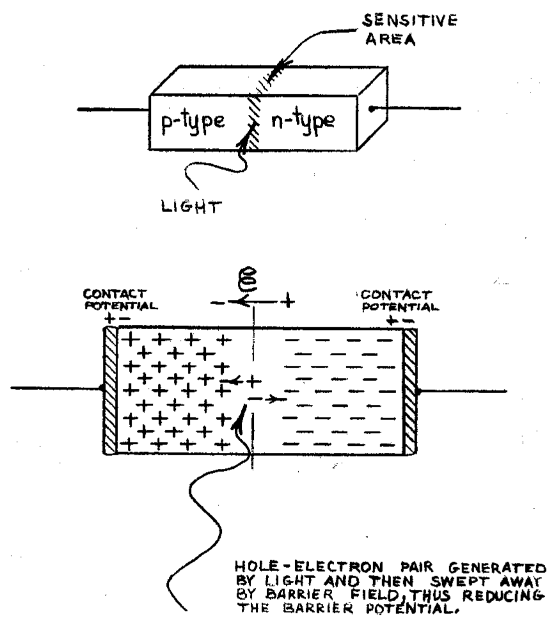


Figure 3.10. Operation of a Photovoltaic Cell.

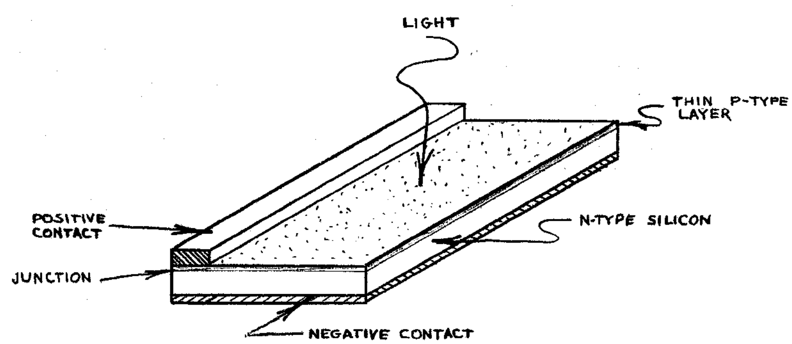


Figure 3.11. Construction of a Solar Cell.

The construction of a high efficiency photovoltaic cell is shown in figure 3.11. P-type impurities are diffused into a N-type crystal to produce a large area junction near the surface of the crystal. The photoelectric generations are confined to the junction region by using a material that absorbs most of the light near the surface, so the P-type layer must be very thin for high efficiency. Because the P region is so thin, the area of the positive contact must be as large as possible, without abscuring too much junction area, to reduce the internal resistance of the cell. A gradual transition between the P and N type materials is used to increase the width of the depletion region and, therefore, the active volume in which there is an electric field.

At the present time, silicon is used as the semiconductor material in solar cells because its photoelectric response is in the correct frequency range for operation on direct sunlight. Furthermore, larger barrier potentials are created in silicon diodes, as compared to germanium, so higher output voltages are possible (the selection of materials is a more rigid physical problem than might be thought because there is a direct relation between the strength of the covalent bonds, the photoelectric response, and the barrier potential). Silicon solar cells have been made that will produce an open circuit output voltage of 0.5 volt and a short circuit of 10 ma per square centimeter of active junction area when exposed to bright sunlight.

TUNNEL DIODES

The tunnel diode is a two terminal device that can be used as an amplifier, an oscillator or a switch. The tunnel diode will perform these functions by virtue of the fact that it exhibits negative resistance over certain ranges of operation. The electrical characteristics of a tunnel diode are shown in figure 3.12. Over a range of forward bias, the diode current decreases with increasing voltage. This is opposite to the behaviour of a normal (positive) resistance: Hence, the name, negative resistance.

A negative resistance will produce a power gain in a circuit, rather than a power loss. This is not in opposition to the conservation of energy; the negative resistance must be supplied power from another source, so it merely converts electrical energy of one form (usually d.c.) into electrical energy of another form. Normally, a negative resistance amplifies by canceling the loss of a positive resistance. However, there is a limit on the maximum gain that can be realized because if the total circuit resistance becomes negative, the circuit will become unstable and oscillate. More will be said about the applications of negative resistance devices in chapter 6.

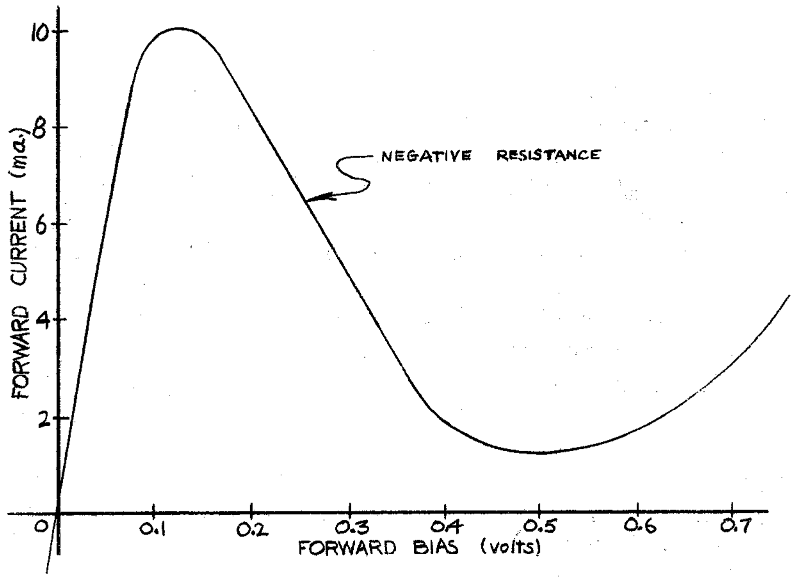


Figure 3.12. Electrical Characteristics of a Typical Gallium Arsenide Tunnel Diode.

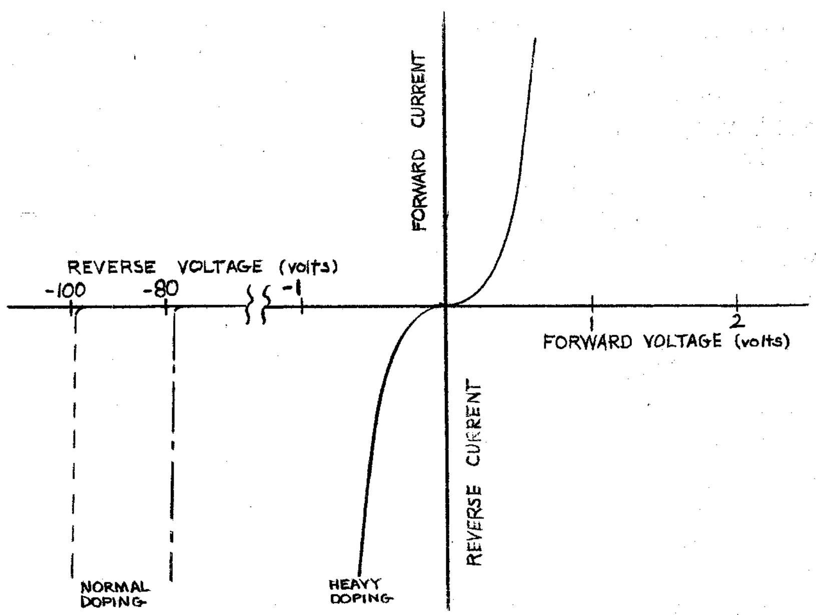


Figure 3.13. Effect of Doping on the Reverse Breakdown Voltage of a PN Junction.

The construction of a tunnel diode is similar to that of conventional diodes, except that the impurity concentration in the P and N regions is more than 1000 times greater. One consequence of such heavy doping has already been suggested in the treatment of zener diodes; that is, the depletion region will be very thin which causes an increased barrier field intensity for a given reverse bias and a lower breakdown voltage. The effect of doping on the reverse breakdown voltage is illustrated in figure 3.13. In the tunnel diode, the doping is carried to an extreme. In fact, the depletion region is so narrow that the normal barrier potential (no bias applied) is enough to induce zener breakdown. The tunnel diode, therefore, conducts heavily even for small reverse voltages.

The ultra thin depletion region (in the order of a few molecules thick) also permits a phenomenon known as quantum-mechanical tunneling, wherein the free electrons in the N type region can cross the junction to fill vacancies in the covalent bonds of the P-type material without being affected by the barrier field. This will only occur over a certain range of forward bias as is illustrated in figure 3.14. At higher forward biases, the conduction within a tunnel diode will take place by the diffusion of current carriers through a reduced barrier field as in a conventional diode.

METAL TO SEMICONDUCTOR CONTACTS

Metal to semiconductor contacts are an important part of every semiconductor device since they are required to make external circuit connections. There are, basically, two types of metal to semiconductor contacts: Rectifying and ohmic. A rectifying contact behaves much like a P-N junction in that it will only pass a current in one direction while an ohmic contact is insensitive to the direction of current and is used primarily to provide a low resistance contact to a semiconductor. Rectifying contacts will be discussed first since ohmic contacts are obtained by degenerating the performance of a rectifying contact.

Rectifying Contacts. When a metal is brought into contact with a P type semiconductor, free electrons from the metal will diffuse into the semiconductor to fill vacancies in the covalent bonds. This process will set up a charge unbalance, the semiconductor becoming negative with respect to the metal, and will continue until the potential difference established is sufficient to stop the diffusion of free electrons. If the semiconductor does not have too many imperfections near the surface where this contact is made, a barrier field will be set up in the semiconductor forming a depletion, or insulating, region as was the case with a PN junction. This depletion region will extend only into the semiconductor because of the vastly greater number of current carriers in the metal. If a voltage is applied to this contact in such a direction as to increase this barrier, there will be no current; but if the applied voltage reduces the barrier, diffusion will resume and a current will be established.

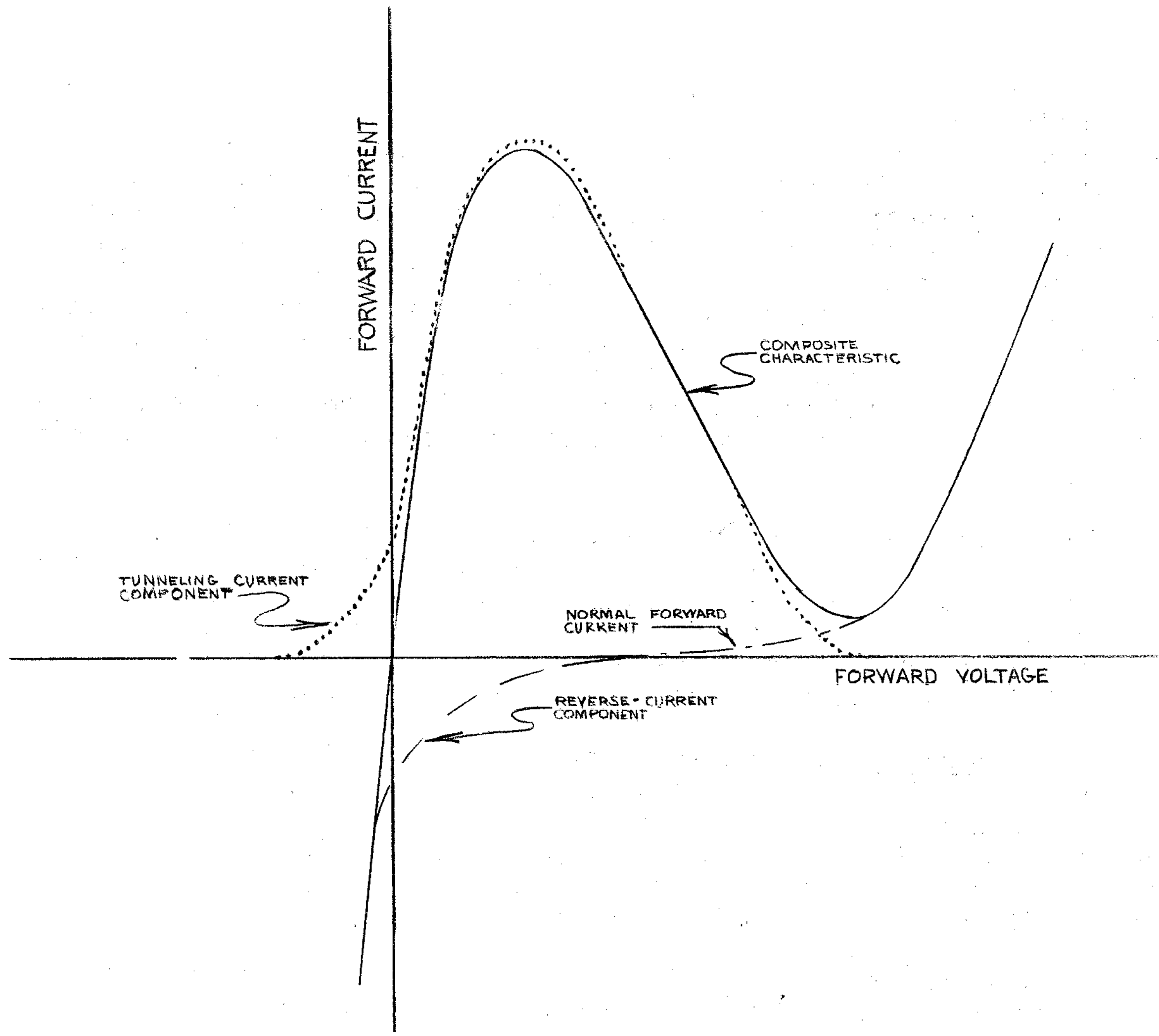


Figure 3.14. Resolving Tunnel Diode Characteristics into Individual Components for Purposes of Analysis.

Similar results can be realized by bringing a N-type semiconductor into contact with a metal. The free electrons in the semiconductor have a higher energy than those in the metal. When these materials are joined, electrons from the semiconductor will diffuse into the metal, becoming high energy free electrons in this material; however, the free electrons in the metal cannot diffuse into the semiconductor because their energy is too low. Therefore, electrons will diffuse into the metal from the semiconductor until a potential difference is set up that will halt the diffusion. Again, if the semiconductor near the contact surface is in good shape, a depletion region will be formed and the contact will be rectifying. A reverse bias will increase the barrier height so no current will flow; but a forward bias will lower the barrier, permitting diffusion.

There are other techniques available for forming rectifying contacts. If, for example, a metal containing a P-type impurity (eg. aluminum wire) is properly fused to a N-type crystal, a thin P-type region will be created around the contact; and a conventional PN junction will be formed. The same thing can be done in making a metalic contact on a P-type semiconductor by including N-type impurities in the metal.

Ohmic Contacts. Perhaps the simplest way to produce an ohmic contact on a semiconductor is to create defects in the crystal near the contact surface. This will destroy the rectifying properties of the contact by converting the depletion region into a region of high resistivity. However, because of this high resistivity, the contact area must be large if the contact resistance is to be low.

A practical method for producing small area, low resistance contacts is to dope the semiconductor very heavily near the contact surface. If the impurity concentration is made high enough, the semiconductor properties will be destroyed and an ohmic contact formed. This type of contact can be easily made by fusing a metal that is alloyed with suitable impurities to the crystal, thereby forming an ohmic P^{++} P or N^{++} N junction. Sucessful contacts can also be made by soldering on the contact using doped solders.

No matter how a contact is made between two dissimilar regions, there will be an initial diffusion of current carriers across the interface between the two materials, just as was the case with the rectifying contacts. This will set up a contact potential between the two materials, but the region in which a field is established is somehow distorted so that the contact does not exhibit rectifying properties.

DEPARTMENT OF WEAPONS TRAINING
LOWRY AIR FORCE BASE
COLORADO

T R A N S I S T O R S

IN SERVICE TRAINING COURSE

DEPARTMENT OF THE AIR FORCE · UNITED STATES OF AMERICA · MCMXLVII

CHAPTER 4

THE JUNCTION TRANSISTOR

AIR TRAINING COMMAND · UNITED STATES AIR FORCE · MCMXLIII

20 July 1961

FOR INSTRUCTIONAL PURPOSES ONLY

CHAPTER 4

Author: A1C Robert J. Widlar

THE JUNCTION TRANSISTOR

__Abstract__ - The operation of the junction transistor is explained in detail considering the junction barrier, minority carrier injection, space charge neutralization and diffusion phenomenon. Furthermore, the physical and electrical parameters affecting transistor performance are given considerable attention. The inherent temperature sensitivities of the transistor are pointed out, and the factors contributing to a general deterioration of performance at high frequencies are analyzed. Modifications of the basic transistor that will greatly extend the maximum operating frequency are described. Finally, transistors that exhibit thyratron-like characteristics are discussed along with other single and multiple junction devices intended for special applications.

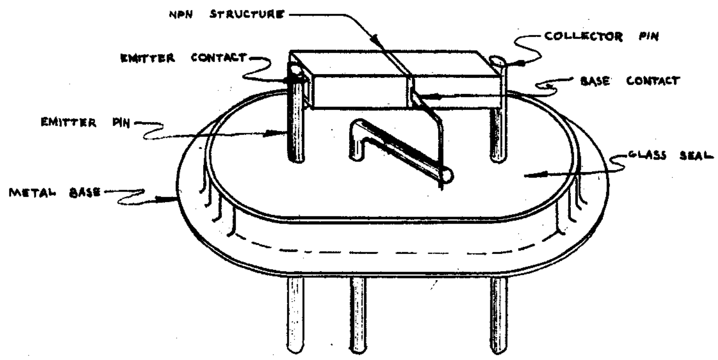


Figure 4.1. Construction of a Grown Junction Transistor.

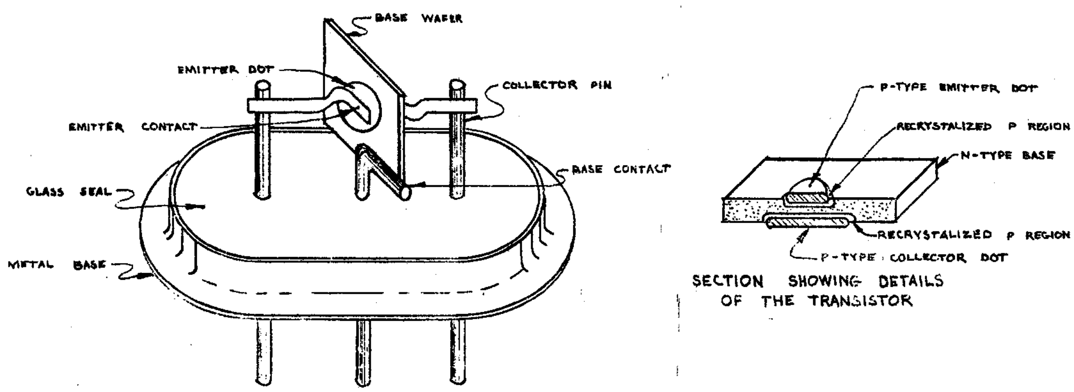


Figure 4.2. Construction of an Alloy Junction Transistor.

INTRODUCTION

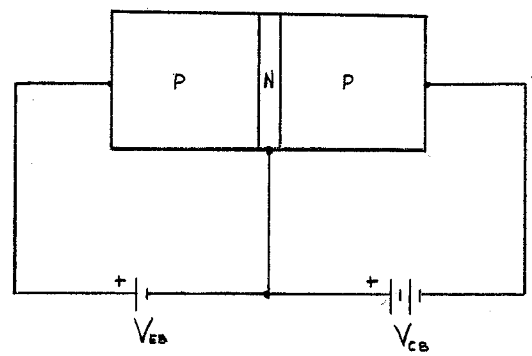
In chapter 1 an elementary description of the junction transistor was given. The NPN and PNP structure of the transistor was shown, and the paths of the current carriers were illustrated. Here the potential barriers existing at the junctions will be taken into consideration, and the mechanisms responsible for the transport of current carriers across the base will be looked into more closely. Furthermore, the design factors affecting performance (current gain, linearity, maximum ratings, temperature sensitivity, and frequency response) will be discussed. Finally, special designs used to improve characteristics for particular applications will be covered. This latter category will include the drift, tetrode, field effect, hook, and unijunction transistors.

THE JUNCTION TRANSISTOR

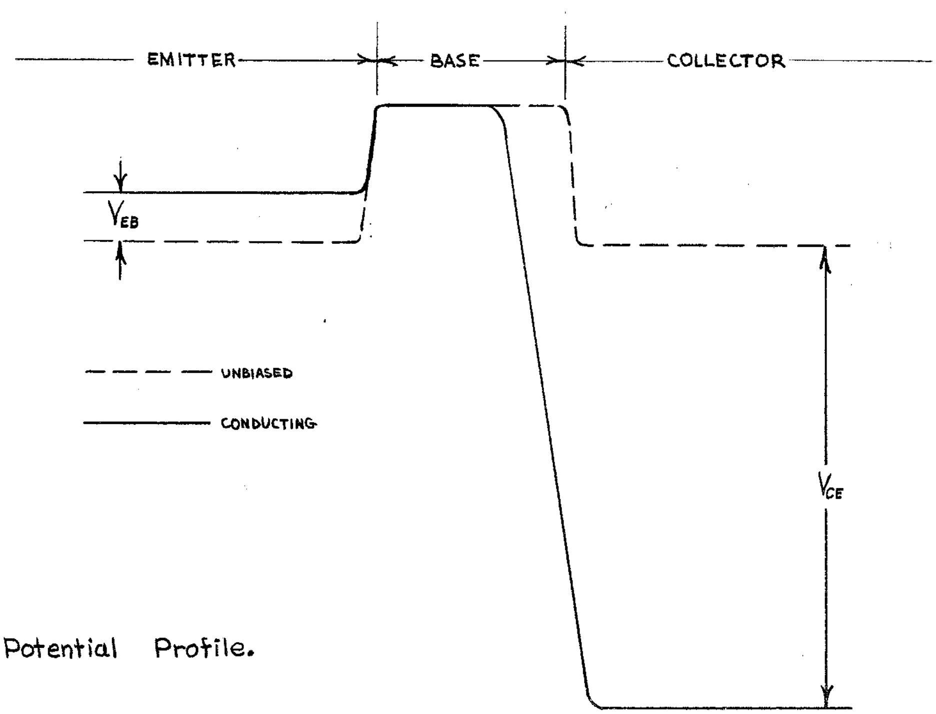
Two possible configurations used in the fabrication of junction transistors are illustrated in figures 4.1 and 4.2. These will be discussed briefly so that the nature of the device can be appreciated. However, a more detailed description of manufacturing techniques will be given in the next chapter.

Figure 4.1 illustrates a grown junction transistor. A NPN transistor of this type is made by dipping a seed crystal into molten germanium which has been moderately doped with a N-type impurity. The seed is withdrawn slowly from the melt; and the doped germanium crystalized on the seed, producing a N-type crystal. After the crystal has grown for some time, sufficient P-type impurity is added to the melt to overcome the N-type impurity; and a thin P-type layer having a high resistivity is grown. Finally, an excess of N-type impurity is dumped into the melt, again reversing the impurity type, so a low resistivity N-type portion is grown. This procedure gives the required NPN structure. Using this technique, good quality transistors having base widths in the order of 0.001 inch can be made.

An alloy junction transistor is shown in figure 4.2. To fabricate this type of transistor, two dots of a P-type impurity having a low melting point are placed on opposite sides of a thin N-type germanium wafer which has a high resistivity. This assembly is then heated, melting the dots which take some of the germanium into solution. When the assembly is cooled, the dissolved germanium, which is now heavily doped with the P-type impurity, recrystalizes on the wafer around the dots. This procedure yields two heavily doped P-type regions separated by a thin layer of lightly doped N-type material. In practice, the penetration of the dots into the wafer can be controled to produce base widths of 0.001-0.003 inch.



a. Biasing Arrangement.



b. Potential Profile.

Figure 4.3. Variation of Potential with Distance through the Base Region for Both an Unbiased and a Conducting Transistor.

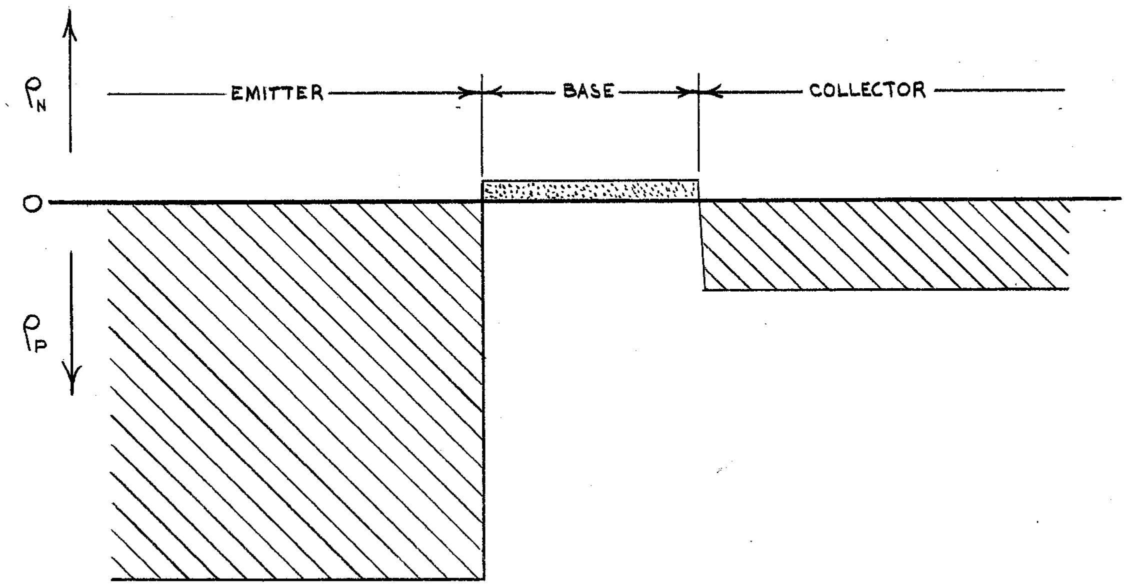
Barrier Formation. As with a junction diode, when the PN junctions of a transistor are formed, there is an initial diffusion of current carriers across the junction, which produces a charge unbalance. This charge unbalance gives rise to a barrier potential and its associated barrier field that is confined to a narrow depletion region. This barrier opposes further diffusion and establishes an equilibrium condition. Therefore, in an unbiased transistor, barriers will be formed at the emitter and collector junctions producing the potential profile shown in figure 4.3b (dashed line).

Transistor Operation. When a reverse bias is applied to the collector junction as shown in figure 4.3a, the barrier height at the collector junction is increased by an amount equal to the applied voltage. The entire collector voltage is dropped across the depletion region of the collector junction, so there is no electric field acting on the current carriers in the base and in the collector. Therefore, no current will flow across the collector junction, except for a small saturation current caused by the diffusion of thermally generated minority carriers to the junction.

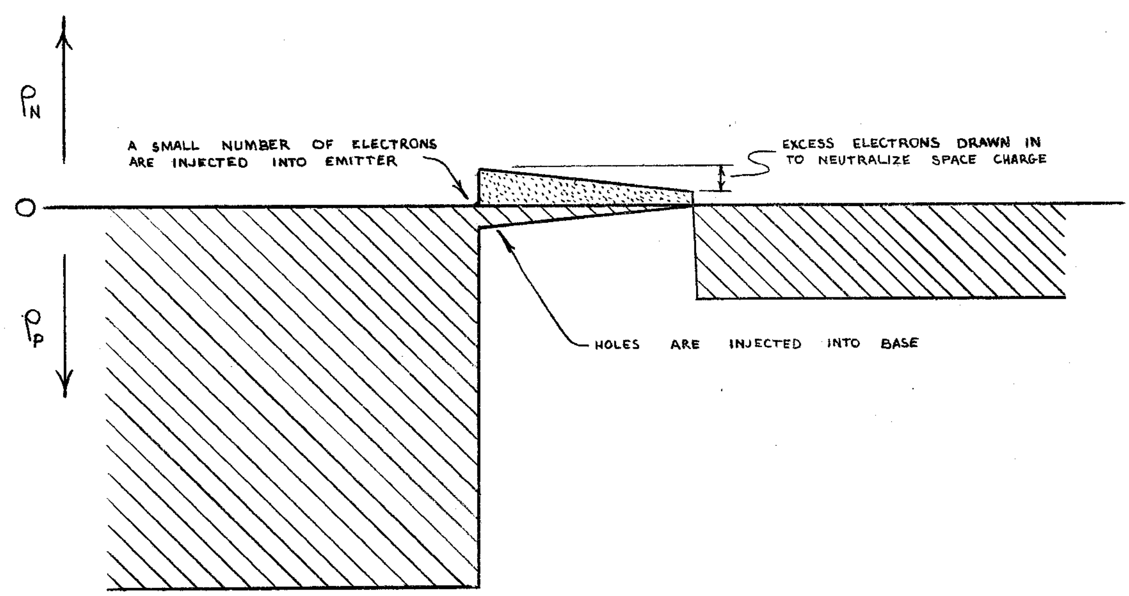
If now a small forward bias is applied to the emitter junction, this barrier will be lowered, permitting the diffusion of holes into the base. It should be emphasized that the holes are not accelerated into the base by the forward bias, but instead the barrier is lowered permitting the thermal diffusion of holes through a reduced barrier field. Furthermore, after the holes are injected into the base, they are not acted upon by an electric field but continue to diffuse until they reach the collector junction. When the holes arrive at the collector junction, they enter the barrier field which sweeps them across the junction into the collector (the drift forces of the electric field are considerably stronger than the diffusion forces).

Space Charge Neutralization. When the emitter junction is forward biased, holes are injected into the base. These excess positive charges create an electric field in the base region. However, the mobile electrons in the base are acted upon by this field and are attracted to the positive charges. Electrons are then drawn in from the base terminal to neutralize the electric unbalance created by the injected holes. This process, which has already been explained in chapter 2, is called space charge neutralization. The hole and electron densities are plotted in figure 4.4 as a function of distance in the vicinity of the base to illustrate this phenomenon.

One result of space charge neutralization is the elinination of the repulsive forces acting on the injected holes. Therefore, the holes do not diffuse through the base by mutual repulsion, instead they are transported across the base by thermal diffusion supported by the higher concentration of holes near the injection source (emitter junction).



a. Nonconducting Transistor.



b. Conducting Transistor.

Figure 4.4. Plot of the Hole (ρ_P) and Electron (ρ_N) Densities with Distance through the Base for a PNP Transistor in Both the Conducting and Nonconducting States.

Current Gain. Before going further into the operation of the transistor, two parameters need to be defined: the common base current gain (α_B) and the common emitter current gain (α_E). The common base d.c. current gain is defined as the ratio of the collector current to the emitter current (figure 4.5a). It has been shown that the collector current will be less than but very nearly equal to the emitter current for a junction transistor so α_B will approach unity. In practice, values of α_B ranging from 0.900-0.995 are common.

The common emitter d.c. current gain is defined as the ratio of the collector current to the base current (figure 4.5b). Since the base current is small in comparison to the collector current, α_E will be considerably greater than one. Values of α_E corresponding to those given for α_B range from 10-200. (Note: the common emitter current gain is sometimes given as the beta (β) current gain.)

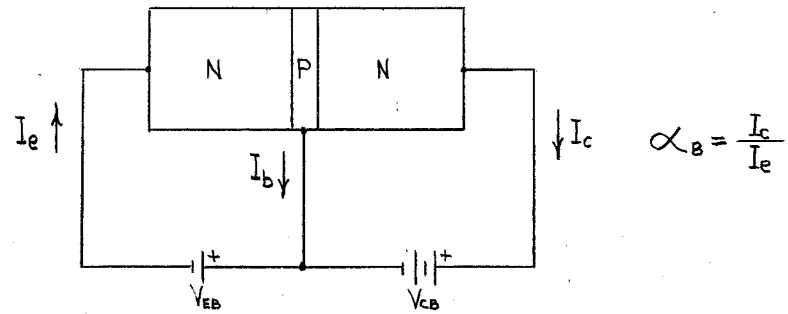
TRANSISTOR PERFORMANCE

The current gain of a transistor can be determined by considering three terms: the injection efficiency, the transport efficiency, and the collector efficiency. The physical phenomena affecting these terms will be investigated for a PNP transistor, although analagous reasoning can be applied to a NPN unit.

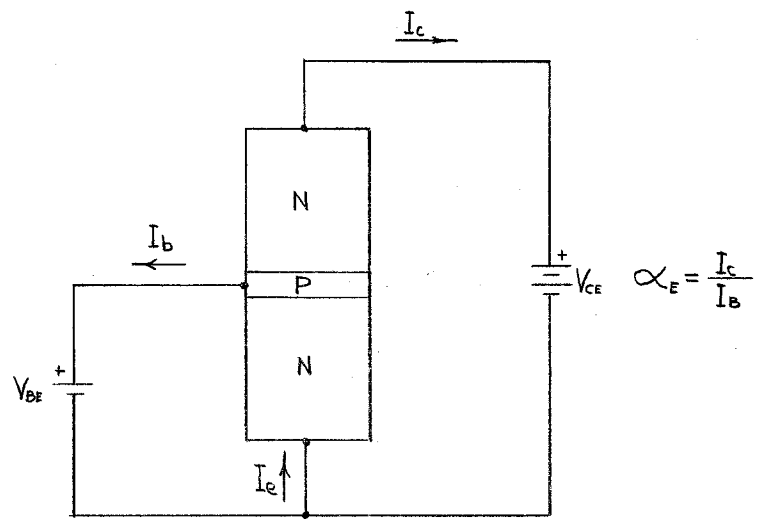
Injection Efficiency. When the emitter junction is forward biased, holes from the emitter are injected into the base; and electrons from the base are injected into the emitter. These hole and electron currents constitute the total emitter current. It is desirable to make the electron current across the emitter junction as small as possible because it produces a base and an emitter current but it does not contribute to the collector current. The electron current, then, reduces both α_B and α_E.

The injection efficiency is defined as the ratio of the hole current across the emitter junction to the total emitter current. High injection efficiencies (near unity) can be realized by suppressing the electron current across the emitter junction. One method that can be used to accomplish this is to make the emitter of a heavily doped, low resistivity material and make the base of a lightly doped, high resistivity material. Hence, when the emitter junction is forward biased, a larger number of holes will cross the junction by virtue of the fact that there is a much higher concentration of holes on the emitter side than there is electrons on the collector side.

The injection efficiency is also affected by the base width and the minority carrier lifetime in the emitter. When the emitter junction is forward biased, the hole concentration on the base side of the emitter-base junction and the electron concentration on the



a. Comon Base Circuit.



b. Comon Emitter Circuit.

Figure 4.5. Circuits Used in the Determination of the Common Base and Common Emitter Current Gain of a Transistor.

emitter side of the emitter-base junction will be determined by the resistivities of the emitter and base materials respectively and by the forward bias. However, the currents resulting from these injected carriers will be determined by the factors mentioned below.

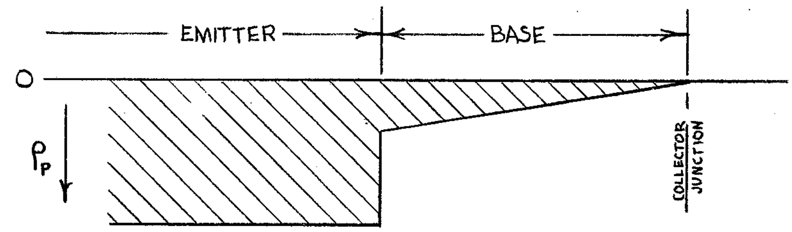
The effect of base width on the hole current across the emitter junction is illustrated in figure 4.6. Since all the holes reaching the collector junction will be swept into the collector by the reverse bias, the hole concentration at this point will be near zero; and the variation of hole concentration with distance through the base will be as shown in the figure. The diffusion current produced by these holes will be proportional to the abruptness of this variation. It should be evident from the figure, then, that the hole current through the base will be inversely proportional to the base width.

Since there is no sink for electrons in the emitter as there was for holes in the base (the collector junction), the point at which the electron concentration in the emitter reaches a neglegible value will be determined by the minority carrier lifetime in the emitter. That is, if the electrons injected into the emitter recombine rapidly, the fall off in electron concentration will be abrupt. This is illustrated in figure 4.7. Because the electron diffusion current will depend on how fast this concentration falls off, it will be strongly affected by the minority carrier lifetime. Actually, the diffusion current is inversely proportional to the diffusion length which has been defined as the average distance that a minority carrier will diffuse before recombination occurs.

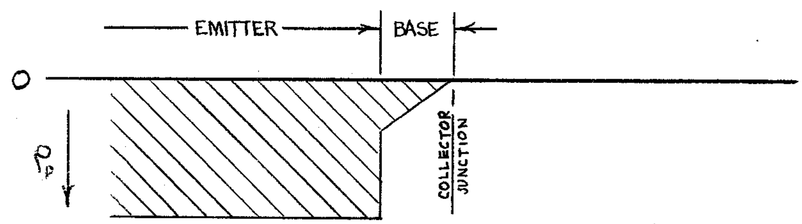
To summarize, a high injection efficiency can be realized by producing a transistor which has a narrow base having a high resistivity and a low resistivity emitter with a long minority carrier lifetime. In practice, it is not too difficult to realize a hole current 1000 times as great as the electron current, which gives an injection efficiency of 0.999.

<u>Transport Efficiency.</u> The next term to be considered is the transport efficiency, defined as the ratio of the hole current across the collector junction to the hole current across the emitter junction. The transport efficiency will be a function of the number of holes lost in the base region by recombination. There are two significant sources of recombination: volume recombination and surface recombination.

The volume recombination, which is the loss of current carriers within the body of the base region, is best described for this purpose by the diffusion length. If most of the holes crossing the emitter junction are to reach the base, the base width must be small in comparison to the diffusion length. Typical values of diffusion length are from 0.05 to 0.005 inch. Hence, with a 0.001-0.002 inch

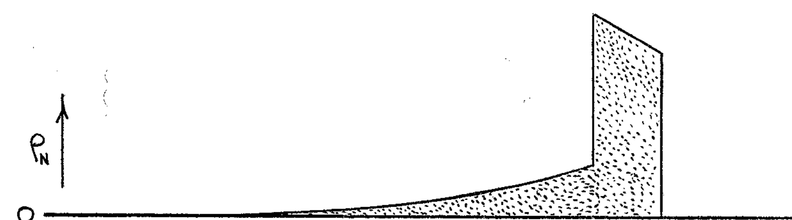


a. Wide Base.

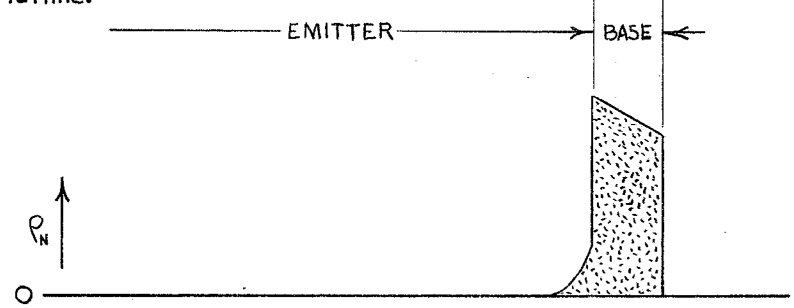


b. Narrow Base.

Figure 4.6. Plot of the Hole Concentration through the Base of a PNP Transistor for a Given Forward Bias, to Show the Effect of Base Width on Hole Diffusion Current.



a. Long Lifetime.



b. Short Lifetime.

Figure 4.7 Plot of Electron Concentration in the Emitter of a PNP Transistor for a Given Forward Bias, to Show the Effect of Minority Carrier Lifetime on Electron Diffusion Current.

thick base, volume recombination in the base region will be small. The number of carriers lost by volume recombination increases quite rapidly with increasing base width and with decreasing diffusion length. For a diffusion length of 0.05 inch and a base width of 0.005 inch, approximately 0.02 percent of the carriers are lost by volume recombination in traversing the base; and for a diffusion length of 0.005 inch and a base width of 0.002 inch; about 8 percent of the carriers are lost.

The second source of recombination, surface recombination, takes place when injected minority carriers diffuse to the surface of the base, become trapped, and eventually recombine. The surface recombination losses will depend on how much base-surface area is in the diffusion path of the holes crossing the base. It seems likely, then, that this term is unimportant in a grown junction transistor (figure 4.1) which has very little base-surface area. In an alloy junction transistor (figure 4.2) which has a large base-surface area, surface recombination is usually more important than volume recombination.

If the minority carriers diffusing to the surface of the base recombine slowly, the minority carrier concentration at the surface will be relatively high, inhibiting further diffusion. The rate of diffusion to the surface and, therefore, the number of carriers lost by surface recombination, will then depend on the recombination rate of the carriers trapped at the surface, or the <u>surface recombination velocity</u>. The recombination rate is determined by the condition of crystal surface: rough, contaminated surfaces give high recombination velocities while clean, etched surfaces that are free of defects give low recombination velocities.

<u>Collector Efficiency</u>. The last term used in the determination of current gain is the collector efficiency. The <u>collector efficiency</u> is defined, for a PNP transistor, as the ratio of the hole current across the collector junction to the total current across the junction. The total collector current can become greater than the hole current alone if carrier multiplication occurs within the collector junction or if a significant number of thermally generated electrons in the collector diffuse to the collector junction.

Avalanche multiplication can take place within the collector junction if the reverse bias on this junction is made sufficiently high. At high reverse voltages, it is possible for the holes crossing the junction to be accelerated enough to produce ionizations when they collide with atoms of the crystal lattice, producing additional current carriers. This can occur at reverse voltages below the breakdown voltage of the junction (self sustaining multiplication) and will produce common base current gains (α_B) greater than one. More will be said about avalanche multiplication in the discussion on reverse breakdown phenomenon.

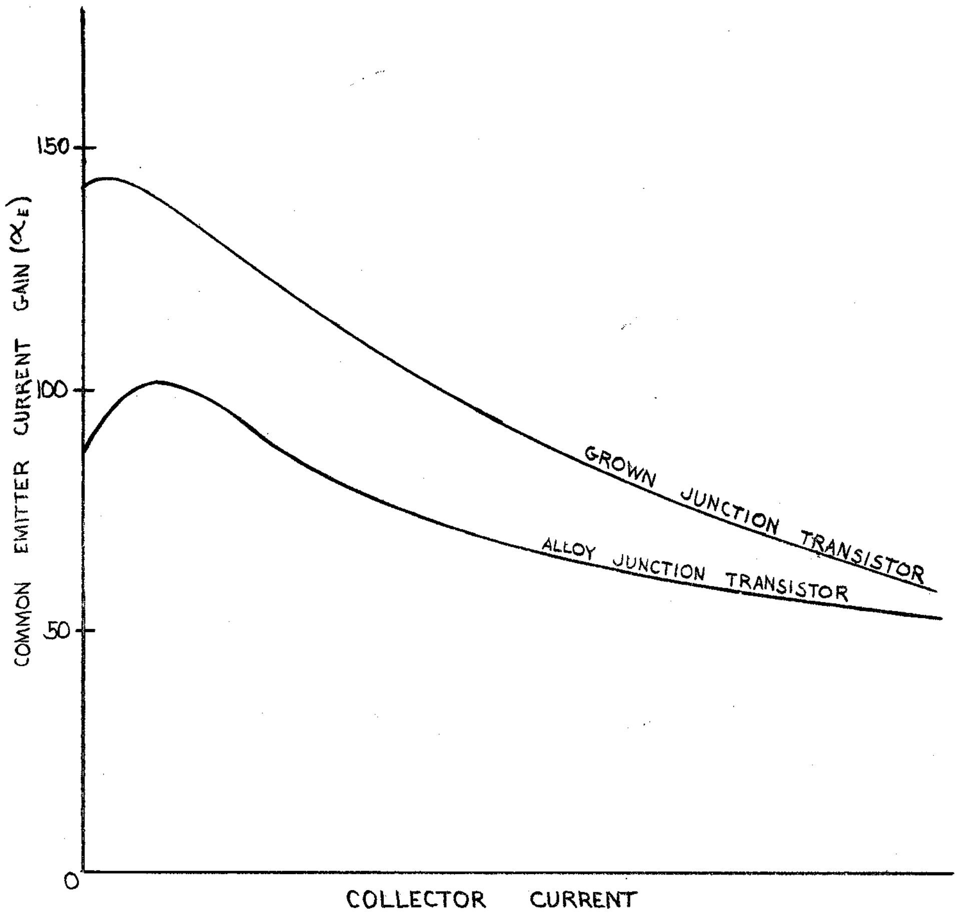


Figure 4.8. Variation of Common Emitter Current Gain with Collector Current for Grown and Alloy Junction Transistors.

Under normal circumstances, the flow of holes into the collector of a conducting PNP transistor does not affect the diffusion of thermally generated electrons to the collector junction, so the collector current is equal to the hole current across the collector junction plus the small reverse saturation current of the reverse biased junction. Therefore, the collector current is essentially equal to the hole current; and the collector efficiency is unity. However, if a high resistivity collector material is used (which means that the hole concentration is relatively low and that the thermally generated electron concentration is relatively high), the flow of holes into the collector will produce a significant increase in the hole concentration near the collector junction and will give rise to an electric field near the junction. This electric field will act on the thermally generated electrons in the collector speeding them toward the junction, thereby increasing the collector current. If the resistivity of the collector material is sufficiently high, this effect will produce an appreciable increase in collector current (over the hole current alone) and increase the collector efficiency above unity.

Improving current gain by increasing the resistivity of the collector material is generally undesirable. The reasons for this are given. First, it is possible to realize common base current gains slightly greater than one by this method. Second, if the collector current does become greater than the emitter current (α_B greater than one), the direction of the base current reverses; and the transistor becomes unstable. In this condition, the transistor is useful only in special applications which require a common base current gain considerably greater than one. And last, when a high resistivity collector is used, the current gain will increase rapidly with temperature as the number of thermally generated electrons in the collector increases. Hence, α_B will become greater than one; and the transistor will become unstable above a certain temperature.

High Current Operation. In the treatment of transistor operation, thus far, it has been assumed that the collector current was low enough that the injected minority carriers did not appreciably alter the majority carrier distribution in the base. At higher current levels the injected minority carriers will produce a significant increase in the majority carrier concentration in the base near the emitter junction because of space charge neutralization. This causes the variation in current gain with collector current illustrated in figure 4.8.

The collector current in a PNP transistor is increased by increasing the concentration of injected holes near the emitter junction. This provides a sharper variation of hole density with distance through the base and increases the diffusion current to the

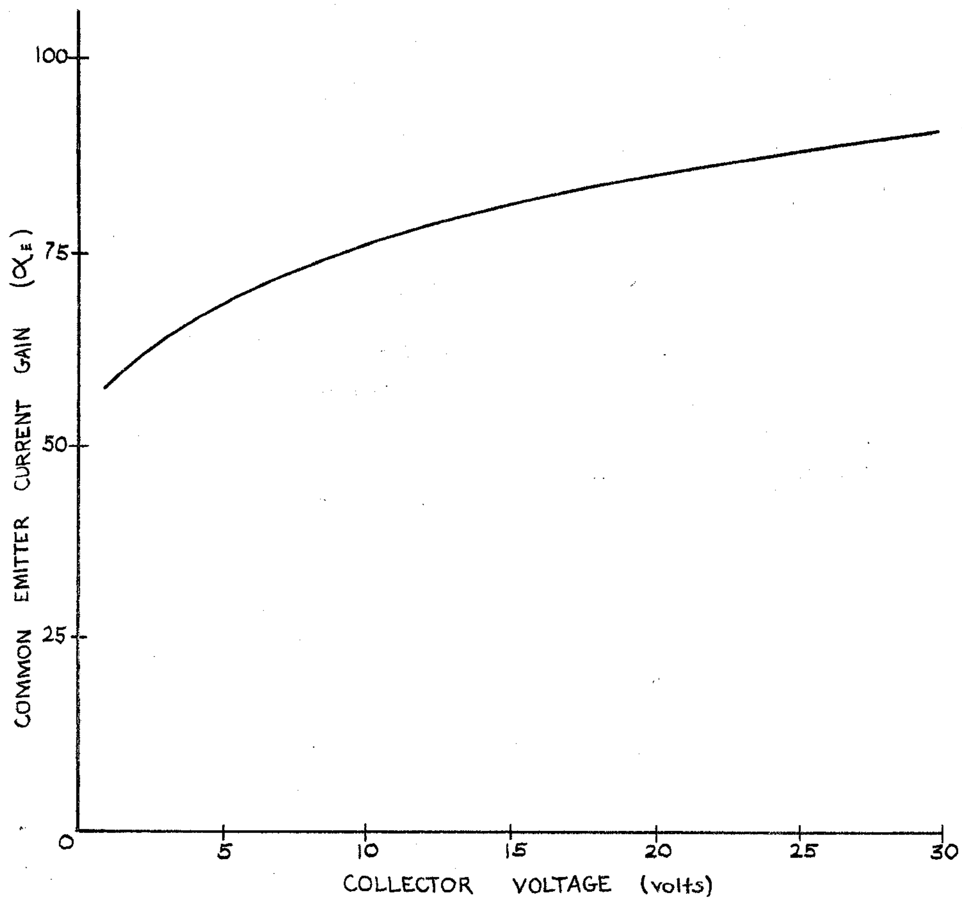


Figure 4.9. Variation of Common Emitter Current Gain with Collector Voltage.

collector junction. These injected carriers create an unbalanced charge distribution which is neutralized by excess electrons drawn in at the base terminals. Hence, the density of free electrons in the base will be higher near the emitter junction, as shown in figure 4.4. This does not reduce the electric field in the base to zero. A small residual field must remain after space charge neutralization to support the uneven distribution of electrons since they will tend to diffuse and equalize their distribution. For small collector currents, this field is also small so it has been neglected up to this point. However, as the collector current is increased, the distribution of current carriers in the base becomes more uneven necessitating an increase in this supporting field. It can be seen from figure 4.4 that in order to support a higher concentration of electrons at the emitter, the field must be positive toward the emitter and negative toward the collector. This field is in the correct direction to speed the holes across the base.

The increase in minority carrier velocity across the base increases the current gain because the change of recombination is lessened. This is shown as the initial increase in current gain in figure 4.8. This initial rise is more pronounced in an alloy junction transistor because the surface recombination term is quite important in the determination of its current gain. For good quality transistors, volume recombination is small so there is only a slight increase with the grown junction transistor which has negligible surface recombination.

After the initial increase, the current gain is found to decrease steadily. This is caused by a reduction in injection efficiency resulting from the increased electron concentration in the base.

As the collector current is increased to high levels, the injected-hole concentration in the base becomes markedly greater. This requires that the electron density increase far above its normal value in order to maintain space charge neutrality. This increased electron density near the emitter junction also increases the number of electrons injected into the emitter. Because these injected electrons will contribute only to the emitter and the base currents, not the collector current, the current gain of the transistor will fall off.

The variation of gain with collector current is of great importance in the application of transistors, particularly with power transistors which must supply large currents. The most practical method of maintaining current gain at high current levels is to dope the emitter as heavily as possible so that a large increase of the majority carrier density in the base can be tolerated before the injection efficiency drops to an unacceptable value. This is one reason why power transistors are usually the alloy junction type. The recrystalized semiconductor material near the emitter dot is normally saturated with the impurity contained in the dot so the emitter has a very low resistivity.

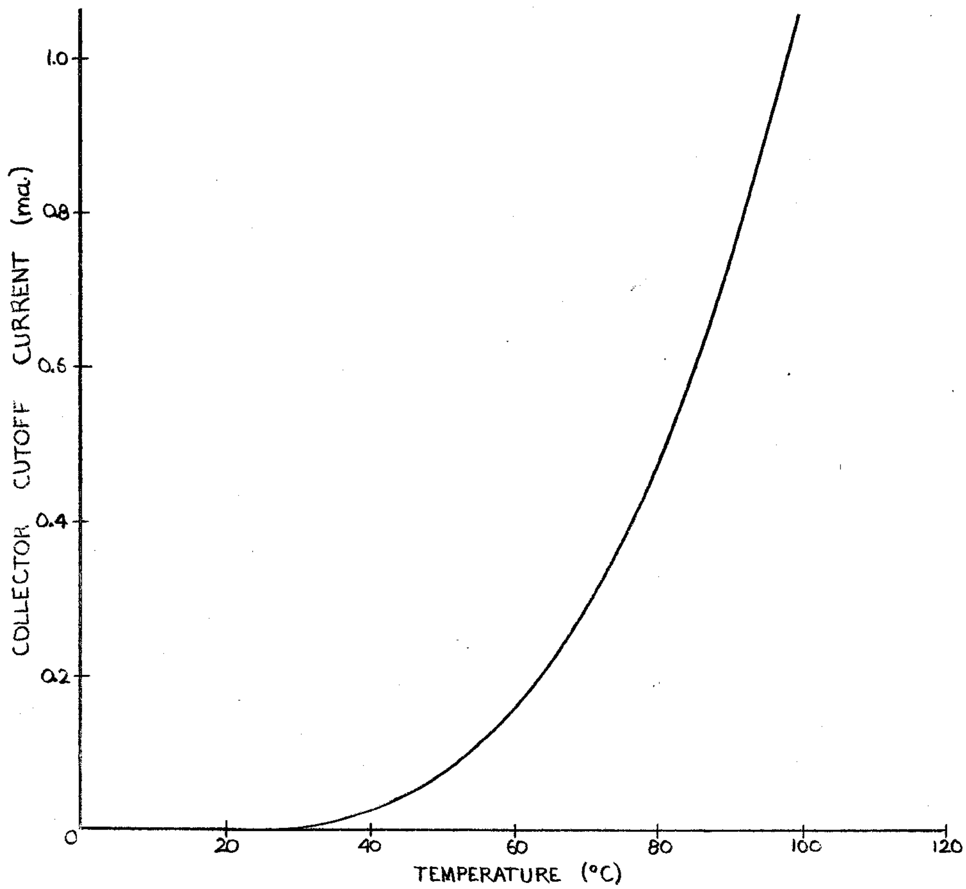


Figure 4.10. Plot of Collector Cutoff Current versus Temperature for a Germanium Transistor.

Variation of Current Gain With Collector Voltage. The resistivity of the collector material is usually considerably lower than the resistivity of the base material. Therefore, the depletion region of the collector junction must extend primarily into the base in order to expose an equal number of immobile impurity atoms on either side of the junction. Furthermore, when collector voltage is applied, the depletion region will become wider and will penetrate further into the base, narrowing the effective base width. This will increase both the emitter and transport efficiencies and will therefore increase the current gain of the transistor. Base width modulation by the collector voltage is illustrated in figure 4.3b, and the effect of collector voltage on current gain is plotted in figure 4.9.

TEMPERATURE DEPENDENCE

Since the properties of semiconductor materials vary widely with temperature, the performance of semiconductor devices, including the transistor, can be expected to be temperature sensitive. The temperature dependence of semiconductor materials has already been discussed at some length in Chapter 2. It has been shown that the thermal generation of hole-electron pairs increases rapidly with temperature and will eventually cause a doped semiconductor to loose its characteristic properties as the number of thermally generated carriers becomes appreciable compared to the number of majority carriers in the material. Furthermore, it was pointed out that the mobility of the carriers decreases with increasing temperature because of the increased thermal vibrations of the crystal lattice. Finally, it was explained that the thermal energy of the current carriers increases with temperature which strongly affects diffusion phenomena. It turns out that the diffusion current for a given density gradient increases with temperature as does the diffusion length.

Collector Cutoff Current. Since the collector junction is reverse biased, the thermally generated minority carriers present in the base and the collector will diffuse to the collector junction and will be swept across the junction. These carriers will then produce a collector current, even though the emitter junction is not forward biased. This uncontrolled collector current is called the collector cutoff current.

Since the number of thermally generated minority carriers will increase rapidly with temperature, so will the collector cutoff current. This is illustrated in figure 4.10 which is a plot of collector cutoff current as a function of temperature. At normal operating temperatures, this current is quite small, if not negligible; but at higher temperatures, it will exceed the normal operating current of the device. When this happens, the emitter junction will loose its control of the collector current, and the transistor becomes useless as an amplifier.

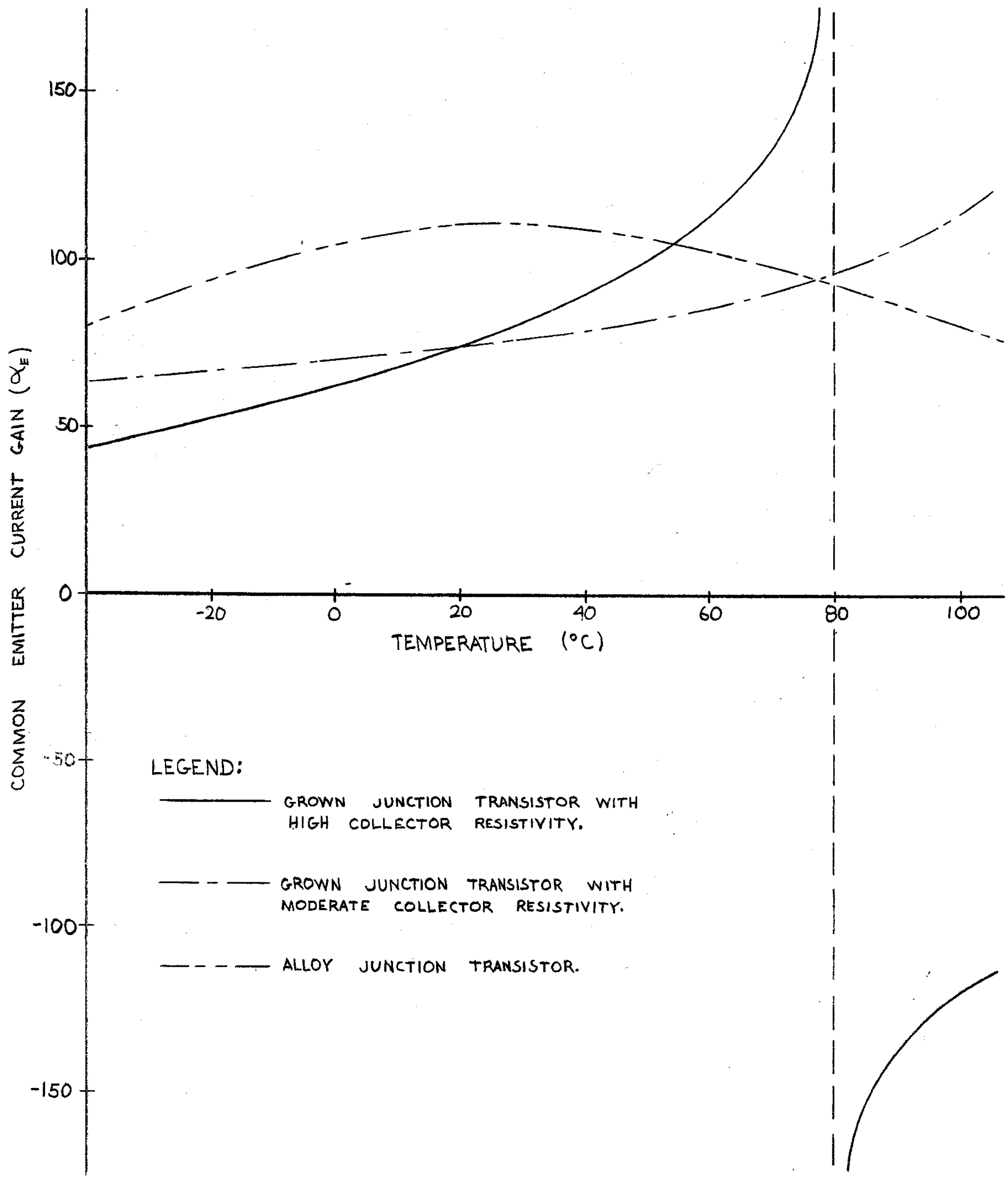


Figure 4.11. Variation of Common Emitter Current Gain with Temperature for Different Transistors.

Reduction of Emitter Junction Resistance. For a given forward bias voltage, only a certain number of majority carriers have enough thermal energy to diffuse across the lowered barrier of the emitter junction. However, the average energy of these carriers will increase with temperature so a greater number can cross the junction. This gives an increase in current with no increase in voltage, or a reduction in the emitter junction resistance. This takes place over the entire temperature range of the transistor, and it therefore has a pronounced effect on its biasing.

Effect of Temperature on Current Gain. Figure 4.11 shows the variation of current gain with temperature for several types of transistors. The causes of these changes will be discussed in relation to the emitter, transport, and collector efficiencies.

The emitter efficiency is not greatly affected by temperature. Temperature affects the injection efficiency of NPN and PNP transistors differently because the mobility of holes is reduced more than the mobility of electrons at higher temperatures. The net effect is that at elevated temperatures there is a slight increase in the injection efficiency of NPN transistors while that of PNP transistors remains essentially constant.

In a grown junction transistor, the transport efficiency generally increases with temperature. The higher thermal energy of the current carriers speeds diffusion through the base, which lessens the chance of recombination. However, the transport efficiency in an alloy junction transistor might well deteriorate at higher temperatures because of increased surface recombination. The variations in the current gain of an alloy junction transistor shown in figure 4.11 are probably caused by changes in the transport efficiency. At first, the current gain increases with temperature because of the more rapid diffusion through the base; but at higher temperatures, surface recombination predominates, reducing the current gain.

If the collector is made of highly doped semiconductor material, the collector efficiency will be very close to unity over the entire temperature range of the transistor. This is true for an alloy junction transistor. However, in a grown junction unit, the resistivity of the collector is frequently not much higher than that of the base. In this case, as was mentioned before, the holes drawn into the collector (of a PNP transistor) can produce a significant increase in the hole concentration near the collector junction. This will give rise to an electric field which will act on the thermally generated electrons in the collector speeding them toward the junction. These electrons will be swept across the collector junction into the base and will contribute to the collector current, but not to the emitter current. The number of thermally generated electrons in the collector will increase with temperature and so will the collector current. Therefore, the collector efficiency, and also the current gain, will increase with temperature.

At a sufficiently high temperature, the collector current can become greater than the emitter current, whereupon the direction of the base current will reverse. This is the cause for the instability of the current gain shown in figure 4.11 (solid line).

In general, the current gain of a grown junction transistor will increase with temperature since both the transport and the collector efficiencies increase. Furthermore, if the collector resistivity is comparable to that of the base, the current gain can become unstable at elevated temperatures. The current gain of an alloy junction transistor, however, can vary in almost any fashion depending on the relative importance of the surface and volume recombination terms. The curve given in figure 4.11 is only one possible example.

<u>Conclusions.</u> Practically all the temperature effects mentioned act in such a way as to increase collector current at higher temperatures. As the collector current increases so does the power dissipation at the collector junction. This produces further heating and could lead to a cumulative condition known as thermal runaway, which would ultimately destroy the unit. Therefore, transistor circuits require some method of bias stabilization on compensation to maintain a reasonably constant collector current over the operating temperature range.

COLLECTOR BREAKDOWN VOLTAGE

The reverse biased collector junction of a transistor is subject to avalanche or zener breakdown as was the case with the junction diode. A third type of breakdown known as collector punch through is also possible. On the other hand, the emitter junction is normally forward biased so its reverse breakdown voltage is of little concern.

The base of a transistor is a thin layer of high resistivity material, while the collector resistivity is usually very much lower. Therefore, the depletion region of the collector junction will extend primarily into the base; and it will become wider as collector voltage is increased. At high collector voltages, it is possible for the depletion region to penetrate through the base to the emitter junction, thereby providing a direct conducting path between the emitter and the collector. This is known as <u>collector punch through</u>. As with the other breakdown mechanisms, collector punch through will not injure the transistor unless the power dissipation becomes great enough to cause thermal damage.

The breakdown mechanism in alloy junction transistors is almost always collector punch through. However, in grown junction units the collector resistivity may not be much greater than the base resistivity. In this case, the depletion region will penetrate into both the collector and the base regions; and the device will usually undergo avalanche breakdown before punch through occurs.

Zener breakdown is rarely found in a transistor. The high resistivity of the base material makes for a wide depletion region so that intense fields are not built up across the junction. Zener breakdown only occurs in PN junctions where both materials are heavily doped.

HIGH FREQUENCY OPERATION

The transistor has some rather severe high frequency limitations. Alloy junction transistors are restricted to operation below a few hundred kilocycles, and grown junction units will not operate above a few megacycles. (However, special constructions can be used which will function at frequencies higher than 500 megacycles.) The reasons for these limitations and some of the techniques used to improve high frequency performance will be investigated here.

Transit Time. Since the transport of minority carriers through the base takes place by rather slow diffusion processes, it will take an appreciable length of time for a signal applied to the emitter junction to reach the collector junction. When a signal is applied to the input of a transistor, it will modulate the number of minority carriers injected into the base and alter the minority carrier density in the base above and below the no-signal value. If the input frequency is made sufficiently high, the amplitude (or phase) of this signal will vary with distance through the base because of the finite transit time. Figure 4.12, which is a plot of minority carrier density with distance through the base, illustrates this point.

In the time it takes for the signal to move through the base, the random diffusion forces will act to equalize these density variations as shown in the figure. Hence, the amplitude of the signal reaching the collector is considerably lower than that at the emitter junction. This results in a low current gain at higher frequencies.

Diffusion Capacitance. If a voltage step is applied to the input of a transistor, the collector current will not rise immediately because of the finite transit time across the base. Furthermore, the collector current step will not have a sharp leading edge because of the diffusion that takes place during the transport process. These effects are shown in figure 4.13. Attention will now be focussed on the input circuit to show another phenomenon associated with transit time.

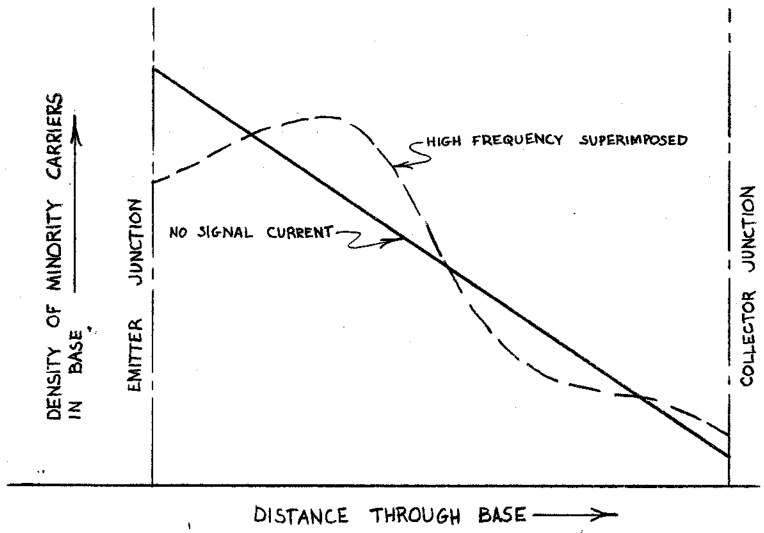


Figure 4.12. Illustrating the Reduction in A-C Current Gain at High Frequencies by Diffusion Which Equalizes Density Variations, a Consequence of the Finite Transit Time across the Base.

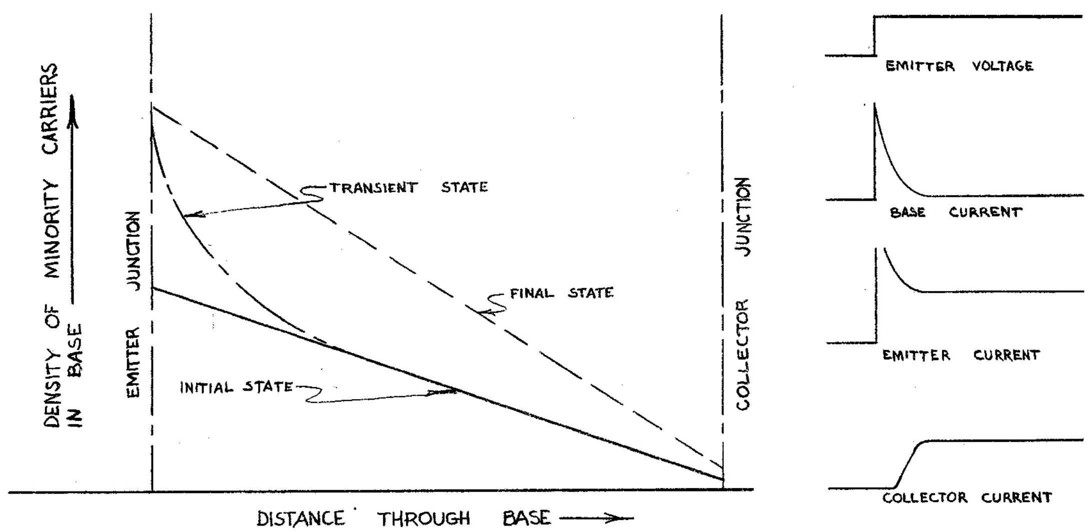


Figure 4.13. Effect of a Sudden Increase in Input Voltage on Minority Carrier Distribution in Base. Emitter, Base, and Collector Current Waveshapes also Given.

When the voltage step is applied to the emitter junction, the minority carrier concentration in the base will increase, or the base will become charged. To charge the base region, an excess of minority carriers must be injected into the base over the emitter junction. Moreover, an equal number of majority carriers must be drawn in at the base terminal to preserve space charge neutrality. These excess charges will not contribute to the collector current but will be stored in the base until the bias on the emitter junction is again decreased. Hence, the emitter-base terminals of the transistor will appear as if they were shunted by a capacitance. This capacitive effect, since it is associated with the diffusion of minority carriers through the base, is referred to as diffusion capacitance.

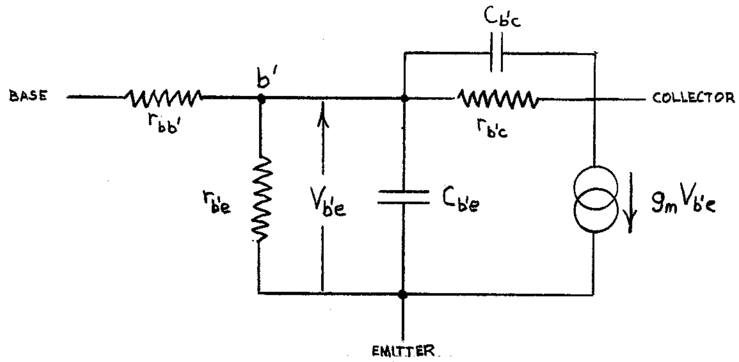
The voltage step applied to the input of the transistor will produce a transient current in the emitter and the base leads which charges the diffusion capacitance. In the transient state, the emitter current will be greater than the final emitter current because the diffusion gradient near the emitter junction (which determines the emitter current for a given forward bias) is greater than the gradient produced by the final current. This is shown in figure 4.13. The initial base current will also be considerably greater than its final value because majority carriers must flow in at the base terminal to preserve space charge neutrality.

It can also be shown, in a similar manner, that a reverse transient current will flow in the base and emitter to discharge the base when the emitter junction bias is decreased.

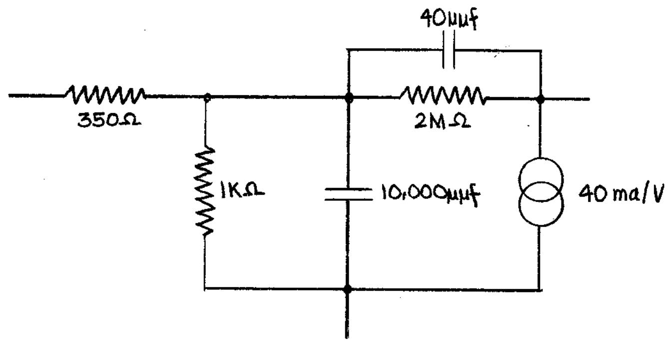
The diffusion capacitance will lower the input impedance of the transistor at high frequencies. The signal source driving the transistor must produce a current that alternately charges and discharges the base rather than producing a collector current. This represents a loss of input signal and lowers the effective amplification of the the device.

Base Spreading Resistance. The base spreading resistance is the resistance of the material between the base contact and the active area of the emitter junction. This resistance appears electrically in series with the base terminal.

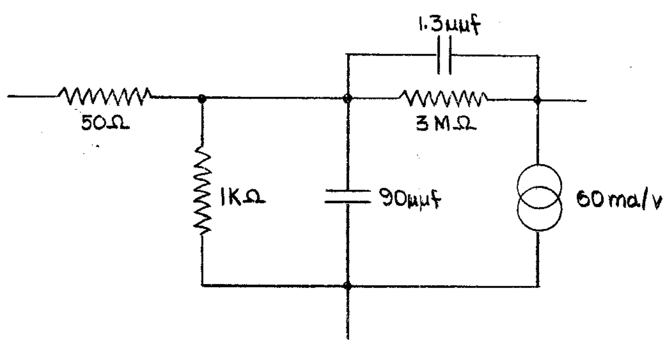
Since the base is made of high resistivity material and is normally very thin, the base resistance can be quite large (several hundred ohms). Hence, a signal applied between the emitter and the base terminals will be attenuated by the base resistance before it reaches the emitter junction. Furthermore, since the emitter junction impedance decreases at high frequencies because of the diffusion capacitance, the signal reaching it will be even smaller at high frequencies.



a. General Equivalent Circuit.



b. Equivalent Circuit of a Low Frequency Alloy Junction Transistor.



c. Equivalent Circuit of a High Frequency Drift Transistor.

Figure 4.14. Equivalent Circuit Useful in Evaluating High Frequency Performance of a Transistor. Specific Examples are Included.

Collector Junction Capacitance. The collector depletion region forms an insulator between the conducting base and collector regions, so the reverse biased collector junction will behave capacitively. This collector junction capacitance is considerably less than the diffusion capacitance; but it shunts the high impedance collector junction, while the diffusion capacitance shunts the low impedance emitter junction. Hence, at high frequencies, the collector capacitance will frequently cause a greater loss of overall gain by shunting the output signal than will the diffusion capacitance by shunting the input signal. This capacitance also provides a feedback path between the collector and the base so it can produce degeneration or instability at higher frequencies when the common emitter configuration is used.

An Equivalent Circuit. The effect of the base resistance, the diffusion capacitance, and the collector capacitance on the high frequency performance of a transistor can be evaluated with the aid of the equivalent circuit shown in figure 4.14a. The elements of this equivalent circuit can be identified as: base resistance ($r_{bb'}$), emitter junction resistance ($r_{b'e}$), diffusion capacitance ($C_{b'e}$), collector junction resistance ($r_{b'c}$), collector capacitance ($C_{b'c}$), and the equivalent-circuit current generator ($g_m V_{b'e}$); the terminal b' is the internal base terminal (near the active portion of the emitter junction), g_m is the transfer conductance (a quantity relating the emitter junction voltage $V_{b'e}$ to the collector current), and $V_{b'e}$ is the voltage on the internal base terminal (emitter junction).

This equivalent circuit does not take into consideration the delay and reduction in current gain caused by transit time across the base so this circuit is not valid where this is the limiting factor of high frequency performance. For most transistors, though, this circuit is adequate over the useful frequency range of the device.

Figure 4.14 shows that the portion of the input voltage appearing at the emitter junction (internal base terminal) will depend on the base resistance, the emitter resistance, the diffusion capacitance, and the signal frequency. The base resistance is generally less than the emitter resistance so at low frequencies practically all of the applied voltage will appear at the internal base and will be effective in producing an output. At high frequencies the reactance of the diffusion capacitance will drop, resulting in an additional attenuation of the input signal. It is therefore desirable to minimize the base resistance and the diffusion capacitance.

Some of the output signal appearing on the collector is fed back to the base through the collector capacitance. Since the collector voltage is out of phase with the base voltage (for a resistive load) this constitutes negative feedback and reduces the useful gain of the transistor. If the transistor is working into an inductive load, this feedback can become positive, possibly introducing instability.

The equivalent circuits of a low frequency alloy junction transistor and a high frequency drift transistor are given in figure 4.14. The maximum useful frequency of the former is in the order of 500 kc while that of the latter is about 100mc. The drift transistor has a lower base resistance, diffusion capacitance, and collector capacitance which permits operation at higher frequencies.

Improving High Frequency Performance. Some of the physical parameters of an alloy or grown junction transistor can be altered to raise the upper frequency limit. These are reducing the base width, using a low resistivity base material, and reducing the cross-sectional area of the device. Of course, this is not an inclusive list; but it represents some of the readily availiable means.

A reduction in base width will decrease transit time and diffusion capacitance as well as increase the transport and injection efficiencies. The reason for the decrease in transit time should be obvious, and the decrease in diffusion capacitance results because a narrower base will store less charge. The increase in transport and injection efficiencies have been explained previously. A thin base does produce one undesirable result: the base spreading resistance is increased because of the narrower current path between the active area of the emitter junction and the base contact. Nonetheless, this is more than compensated for, particularly at high frequencies, by the reduction in transit time and diffusion capacitance.

Reducing base width is a very effective method of improving transistor performance; however, with alloy and grown junction transistors, base widths less than 0.001 inch are difficult to produce which limits the usefulness of the method.

The use of a low resistivity base material will reduce both the base resistance and the injection efficiency. But if a transistor already has a high injection efficiency, a lower resistivity base can give a significant improvement in high frequency performance at a small sacrifice in low frequency gain. This technique is particularly useful in the design of video amplifiers where a flat frequency characteristic is more important than high gain.

Since the collector capacitance is directly proportional to the collector junction area, this area should be as small as possible for optimum high frequency performance. A smaller cross-sectional area can also reduce the base resistance (depending on the geometry of the transistor) by reducing the average distance between the base terminal and the active portion of the emitter junction. This is particularly true for grown junction transistors. On the other hand, a reduction in the cross-sectional area will also reduce the current capacity and maximum power dissipation of the device. Furthermore, if the transistor is made too small it will be fragile and difficult to manufacture.

SPECIAL TRANSISTORS

Thus far the discussion on transistors has been relatively general. The phenomena described applies to the majority of transistors, even though only the grown and the alloy junction transistors were mentioned. In this section, special designs which have been employed to improve the performance of the basic transistor will be covered. Furthermore, other PN junction devices - which operate quite differently from the basic transistor will be explained.

The Tetrode Transistor. The tetrode transistor differs from an ordinary grown junction transistor in that two separate base contacts are used. These contacts are located on opposite sides of the base as shown in figure 4.15. In operation, the emitter junction is forward biased by one of these base terminals (B_1) and reverse biased by the other (B_2). Therefore, minority carriers will be injected into the base near the forward terminal only. This reduces the average distance from the active area of the emitter junction to the B_1 terminal and the base spreading resistance when B_1 is used as the input terminal.

A cross-base current will flow between the two base terminals since there is a potential difference between them (for a NPN transistor, B_1 is about 0.5 volts and B_2 about -1.5 volts with respect to the emitter). However, the cross-base current will be small because the base is made of high resistivity material.

The electric field set up by the cross-base bias will draw the minority carriers, diffusing across the base, toward the B_1 terminal. Thus, a number of these carriers, depending on the bias, will reach the base terminal and not the collector junction. This will decrease the low frequency current gain; but it is not a serious limitation since, at high frequencies, the reduction in current gain is more than compensated for by the reduction in base spreading resistance. This is proven in figure 4.16 which gives plots of the low frequency current gain, base spreading resistance, and 150 megacycle power gain as a function of cross-base current.

Another advantage of the tetrode transistor is that the B_2 terminal can be used to electrically control the gain of the device. This provides a convenient method of applying automatic gain control to a r-f amplifier. Furthermore, since the cross base bias will affect the low and high frequency gains differently, it can be used

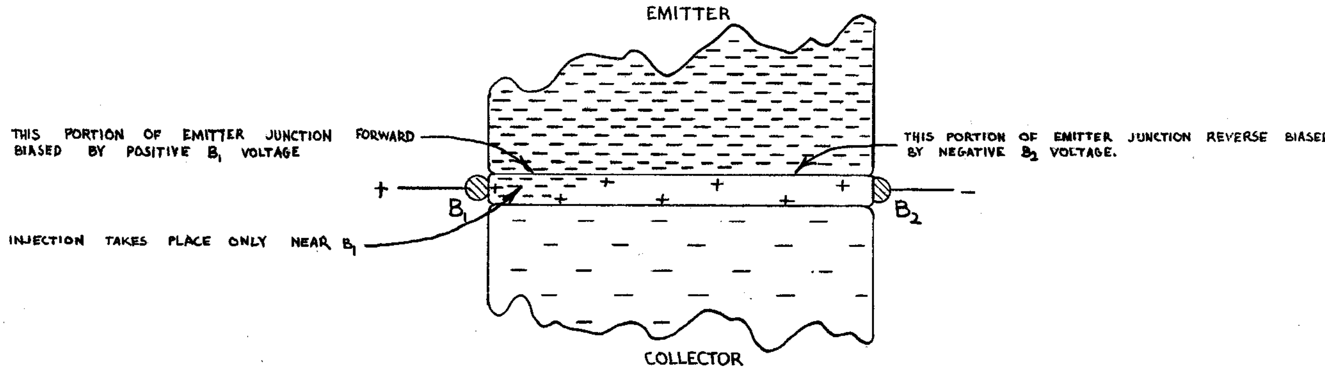


Figure 4.15. Path of Minority Carriers through the Base of a Tetrode Transistor Showing Reduction in Base Spreading Resistance.

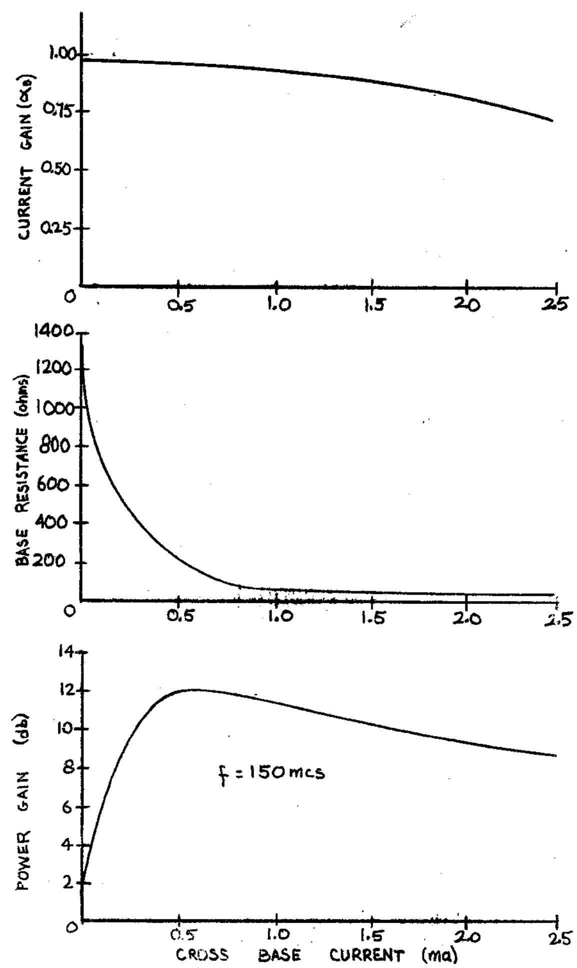


Figure 4.16. Effect of Cross Base Bias on the Characteristics of a Tetrode Transistor.

to produce a reasonably flat frequency response in video amplifiers by reducing the low frequency gain to equal the high frequency gain.

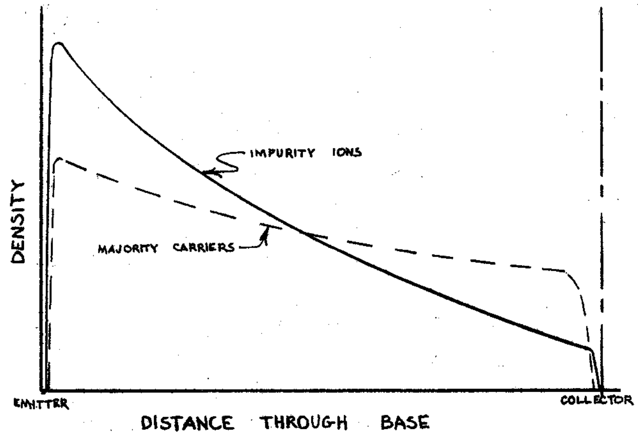
<u>The Drift Transistor.</u> It was stressed previously that the transport.of minority carriers across the base is accomplished by the relatively slow process of diffusion with little or no aid from an electric field. In the drift transistor, the transport is greatly accelerated by establishing a built in electric field across the base in such a direction as to speed the carriers from emitter to collector. The electric field is created during fabrication by doping the base material rather heavily near the emitter junction and tapering off the impurity concentration with distance toward the collector junction. The majority carriers will tend to equalize their distribution because of thermal diffusion, but the impurity atoms are fixed in the crystal lattice. Therefore, there will be an initial diffusion of majority carriers toward the collector junction. This will continue until the charge unbalance sets up an electric field which opposes further diffusion. An equilibrium is reached with an excess of impurity ions near the emitter junction and an excess of majority carriers near the collector junction. This is illustrated in figure 4-17.

The electric field created in the base, opposing the diffusion of majority carriers from emitter to collector, will be in the correct direction to accelerate the minority carriers through the base since they have an opposite charge.

Considering, as an example, a NPN drift transistor, the acceptor concentration in the P-type base will be high near the emitter junction but low near the collector junction. The higher concentration of holes near the emitter will cause diffusion toward the collector. When equilibrium is reached, there will be an excess of holes near the collector and an excess of negative acceptor ions near the emitter. Hence, an electric field will be established in the base, positive toward the collector and negative toward the emitter. When electrons are injected into the base, they will be accelerated to the collector junction by this field.

In addition to reducing the transit time, the drift field materially reduces diffusion capacitance. Since the minority carriers move faster through the base, a lesser number is required to produce a given collector current. Therefore, less charge will be stored in the base. This is shown in figure 4.18.

Another advantage of the drift transistor is that the base resistance is reduced by the high impurity concentration near the emitter junction which gives a low resistivity path for the base current. Although this might seem to reduce the injection efficiency, it must be remembered that the minority carriers injected



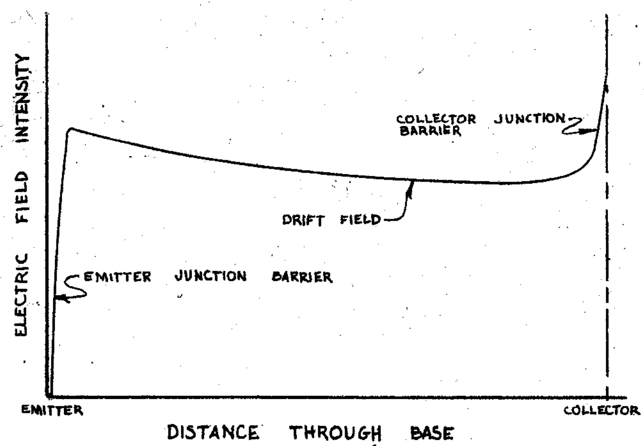


Figure 4.17 Illustrating Formation of a Built in Electric Field in the Base of a Drift Transistor, Having an Uneven Impurity Distribution, by the Initial Diffusion of Majority Carriers.

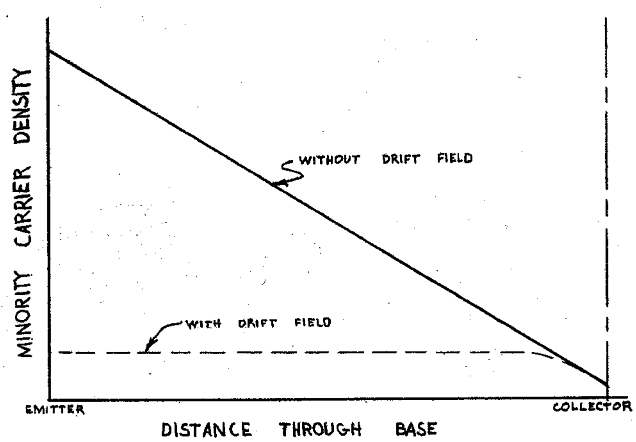


Figure 4.18 Comparison of the Minority Carrier Concentrations in the Base of a Transistor Required to Produce a Given Collector Current with and without the Drift Field.

into the base are acted upon by the drift field and are therefore more effective in producing an emitter current than are the minority carriers injected into the emitter which must diffuse away from the junction. Therefore, even if an equal number of carriers cross the junction in either direction, the injection efficiency can still be high.

The graded base resistivity of the drift transistor also assists in lowering the collector capacitance since the high resistivity material near the collector junction makes for a wide depletion region.

The merit of the drift transistor can be seen by comparing the equivalent circuits given in figure 4.14.

At high current levels where the concentration of injected carriers becomes greater than the majority carrier concentration, the performance of the drift transistor will deteriorate; and it will operate as a diffusion transistor. This happens because of the distortion of the drift field by the injected carriers. It is on indirect result of space charge neutralization.

The graded base resistivity of the drift transistor is not as difficult to produce as might be thought. This type of resistivity profile is the natural outcome of many fabrication processes as will be seen in the next chapter.

<u>The Intrinsic Transistor.</u> In the intrinsic transistor, a layer of undoped (intrinsic) semiconductor is sandwiched between the collector and the base. This gives a PN1P (or NP1N) structure as shown in figure 4.19. Since this layer contains no current carriers, the collector depletion region will extend through the intrinsic material from the collector to the base. This produces a wide depletion layer, greatly reducing the collector capacitance.

The intrinsic layer can be made quite thick, yet it will not appreciably increase the transit time: a strong electric field, produced by the collector voltage, exists in this region; this field speeds the carriers through the intrinsic layer much more rapidly than the diffusion forces take them through the base.

To appreciate the advantage of the intrinsic transistor, it must be remembered that the collector junction capacitance provides a feedback path from collector to base. This greatly reduces the gain of a transistor operating as a wide band video amplifier and requires that neutralization be provided for high frequency tuned amplifiers (common emitter). In the case of a video amplifier very little can be done to compensate for this feedback; and neutralization of tuned amplifiers cannot always be accomplished conveniently, nor is it entirely effective because of base spreading

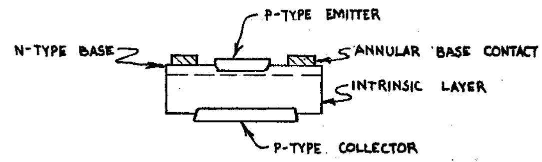


Figure 4.19. Construction of an Intrinsic Transistor, Emitter and Collector Alloyed onto Partially Doped N-Type Base.

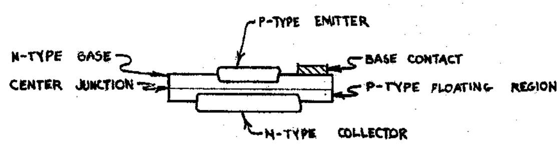


Figure 4.20. Construction of a Hook Junction Transistor, Emitter and Collector Alloyed onto a Grown PN Junction.

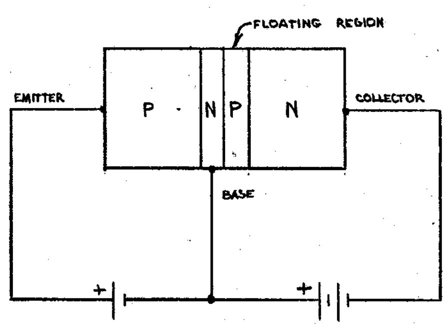


Figure 4.21. Bias Circuit for a Hook Junction Transistor.

resistance. Therefore, a reduction in collector capacitance is of considerable importance.

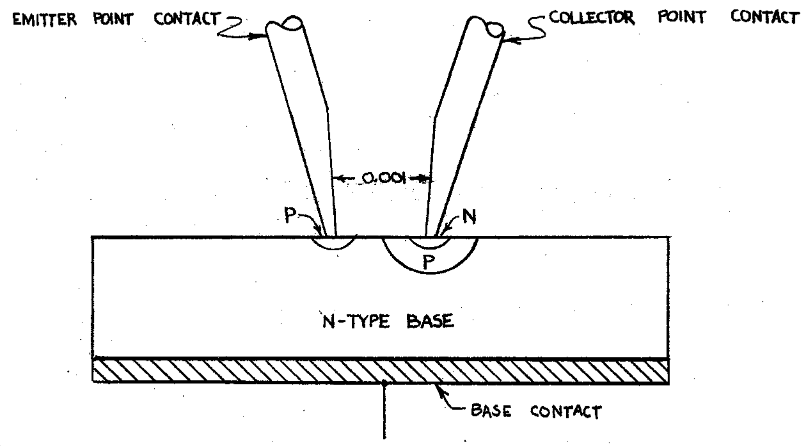
At the present time it is difficult to obtain intrinsic material of sufficient purity, so this principle cannot be fully utilized. Even a low concentration of impurities in the intrinsic layer will restrict the width of the depletion region and result in poor performance. Nonetheless, some of the advantages of the intrinsic transistor can be exploited by using graded junctions, i.e. junctions where the transition from N to P-type is gradual. This will widen the depletion region somewhat, reducing collector capacitance and increasing the collector breakdown voltage.

The Hook Junction Transistor. Departing now from high frequency transistors, a junction transistor that exhibits a common base current gain considerably greater than one will be described.

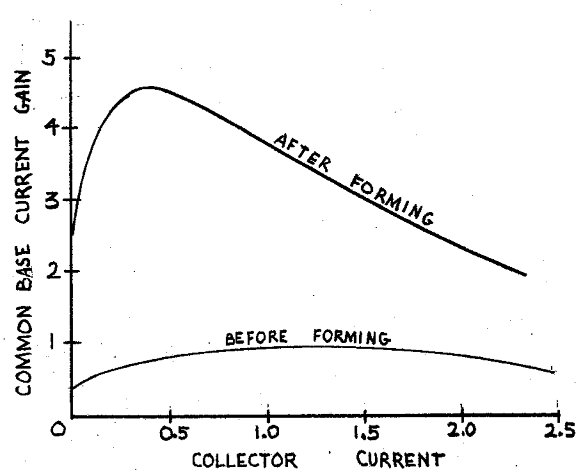
Ordinarily, in a two-junction transistor, the common base current gain will approach unity, values of 0.7 to 0.99 being common. In the avalanche transistor which is a junction transistor operating with the collector junction near avalanche breakdown, the common base current gain will be about 5 because of carrier multiplication in the collector depletion region. The hook junction is capable of giving common base current gains in the order of 100.

The hook junction transistor employs a PNPN (or NPNP) structure as shown in figure 4.20. Connection is made to the emitter, base, and collector; but the P region between the base and collector is left electrically floating. Furthermore, both center layers are thin in comparison to the diffusion length in the material, and the resistivity of these layers is considerably higher than that of either the emitter or the collector.

The biasing arrangement for a hook junction transistor is shown in figure 4.21. The center junction is reverse biased; the emitter and base function as usual. When the emitter junction is forward biased, holes will be injected into the base and will diffuse to the reverse biased center junction. The holes will then be swept into the floating region. These holes will not be free to diffuse into the collector since the collector junction is not forward biased and a barrier to holes still exists. However, the collector junction will become forward biased as the hole concentration in the floating region increases, since the excess holes make the floating region positive with respect to the collector. This forward bias will enable some of the holes to diffuse into the collector, and it will also permit electrons from the collector to diffuse into the floating region. (Note: these electrons will not eliminate the forward bias on the collector junction; enough holes will be retained in the floating region to maintain a forward bias condition.)



a. Possible Structure Produced by Forming Process.



b. Current Gain Before and After Forming.

Figure 4.22. The Point Contact Transistor.

These electrons will diffuse across the floating region to the center junction and will be swept into the base. Hence, a collector current will be produced both by the holes leaving the base and by the electrons swept into the base. It remains to be shown that the electron current is very much greater than the hole current, which would indicate that the collector current is much greater than the emitter current.

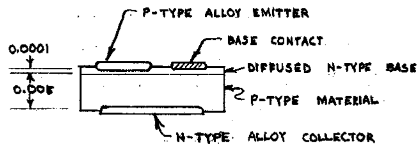
Since the concentration of electrons in the collector is considerably greater than the concentration of holes in the floating region, the number of electrons crossing the collector junction will greatly exceed the number of holes. Furthermore, since the width of the floating region is less than the diffusion length for holes in the collector, the diffusion current produced by a given number of electrons will be greater than the current produced by an equal number of holes, (see section on injection efficiency). It can be seen, then, that the total collector current will be much larger than the original emitter current producing it.

Although some of the electron current reaching the base from the collector do produce an emitter current, the increase in emitter current will be small. The diffusion of electrons across the emitter junction is inhibited by the remaining barrier even though this junction is forward biased. The electrons are preferentially swept out the base terminal because no barrier exists along this path.

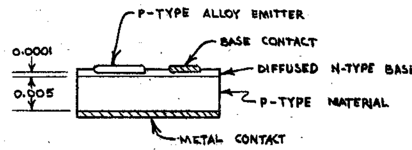
The hook junction transistor has not seen much application. The biggest reason for this is probably the difficulty encountered in producing such a device. It is not easy to obtain reproducable results in two junction units on a production basis, and the problems associated with a multiple junction transistor can be expected to be considerably greater.

The Point Contact Transistor. Although the point contact transistor was the first to appear, its operation never has been well understood. A point contact transistor is made by bringing two fine, pointed wires into contact with the surface of a N-type semiconductor wafer, the spacing between the wires being in the order of 0.001 inch. The contact between these wires and the semiconductor is a rectifying junction, and it is supposed that the emitter injects minority carriers into the base which diffuse to the collector. This explanation is very similar to that for a junction transistor, except that one irregularity exists: the common base current gain of a point contact transistor is usually greater than one. This is not in agreement with junction transistor theory.

The explanation most frequently offered to explain the high current gain of a point contact transistor is that a hook collector is created during a forming process. The forming process usually consists of heating the collector point to a very high temperature



a. PNPN Triode.



b. PNPM Triode

Figure 4.23. Construction of the PNPN and PNPM Thyratron Transistors.

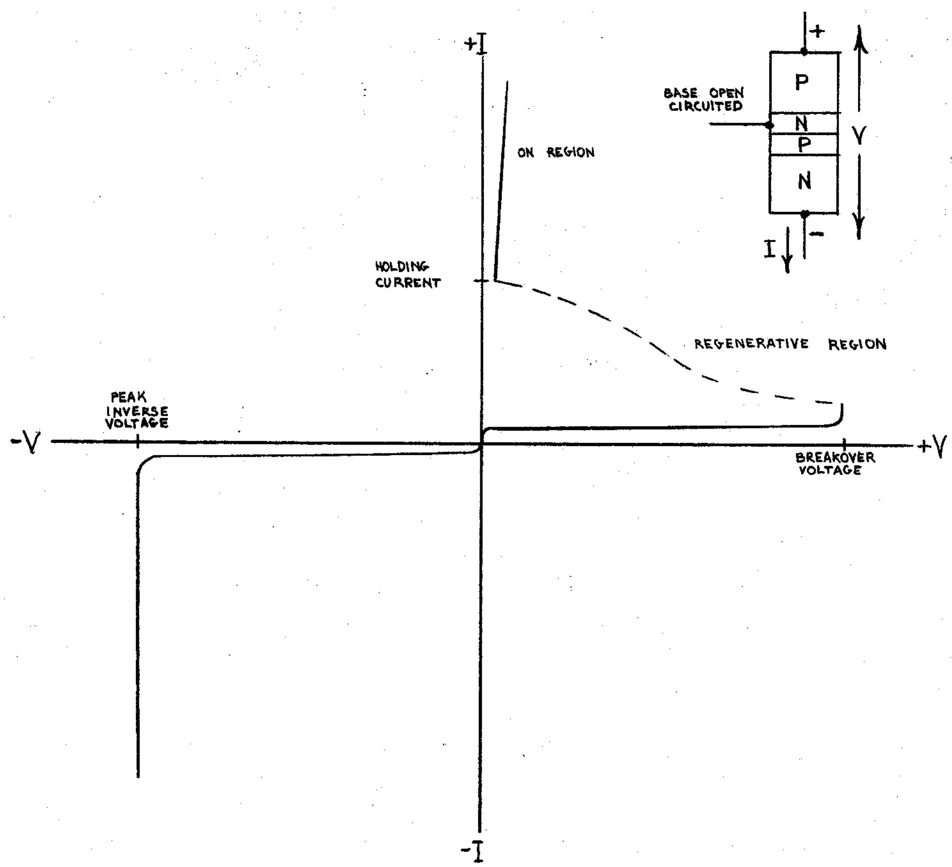


Figure 4.24. Breakdown Characteristics of a Thyratron Transistor, Base Open Circuited.

for a short period of time. It is believed that this causes the conversion of the N-type base material near the collector junction into a hook structure as shown in figure 4.22. This seems plausible considering the current gains obtained before and after forming (figure 4.22) and considering that the impurities contained in the contact points greatly affect the results obtained.

At the present time the point contact transistor is considered obsolete. Shock and vibration can cause large changes im the electrical characteristics since it is difficult to fix the position of the contact points. Furthermore, the wide variation in the performance of individual units of the same type makes direct replacement of defective units impossible. Generally speaking, far superior performance can be obtained with grown junction transistors.

<u>The Thyratron Transistor.</u> Solid state devices have been built that display characteristics similar to those of a gas thyratron. These devices can be switched from a high resistance state to a low resistance state by a relatively small trigger signal. In addition, the solid state thyratron can also be turned off by the control signal.

One type of thyratron transistor has a PNPN structure similar to that of a hook junction transistor. This is shown in figure 4.23a. A thin, high resistivity P-type base is diffused into a N-type crystal which also has a high resistivity. A N-type collector is alloyed to the P-type crystal, and a P-type emitter is alloyed to the diffused layer. Hence, both the emitter and the collector have a low resistivity. An ohmic contact is included on the base which is made very thin to give a fast switching action. The floating P region is made relatively thick, but its width (0.005 inch) is still small in comparison to the diffusion length of minority carriers in the material.

The operation of the thyratron transistor can best be described by considering its breakdown characteristics when a voltage is applied between emitter and collector, reverse biasing the center junction, and the base is left open circuited.

For voltages less than the breakdown voltage, the transistor will not conduct because of the blocking action of the center junction. This condition is represented by the off region in figure 4.24. However, when the breakover voltage is reached, the current will increase abruptly due to avalanche multiplication of the small reverse current through the reverse biased center junction. The holes and electrons generated in the center junction by avalanche multiplication will be swept into the central P and N regions by the barrier field. These carriers cannot diffuse across the collector

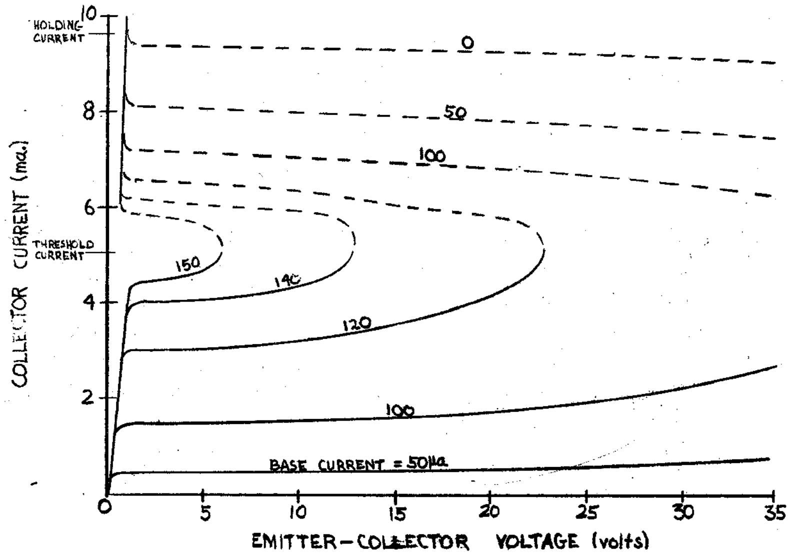


Figure 4.25. Turn On Characteristics of a Thyratron Transistor.

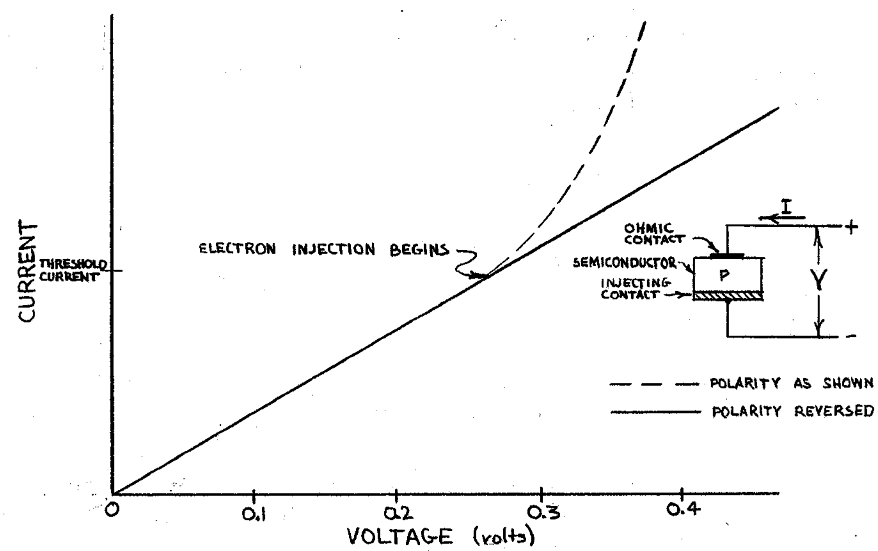


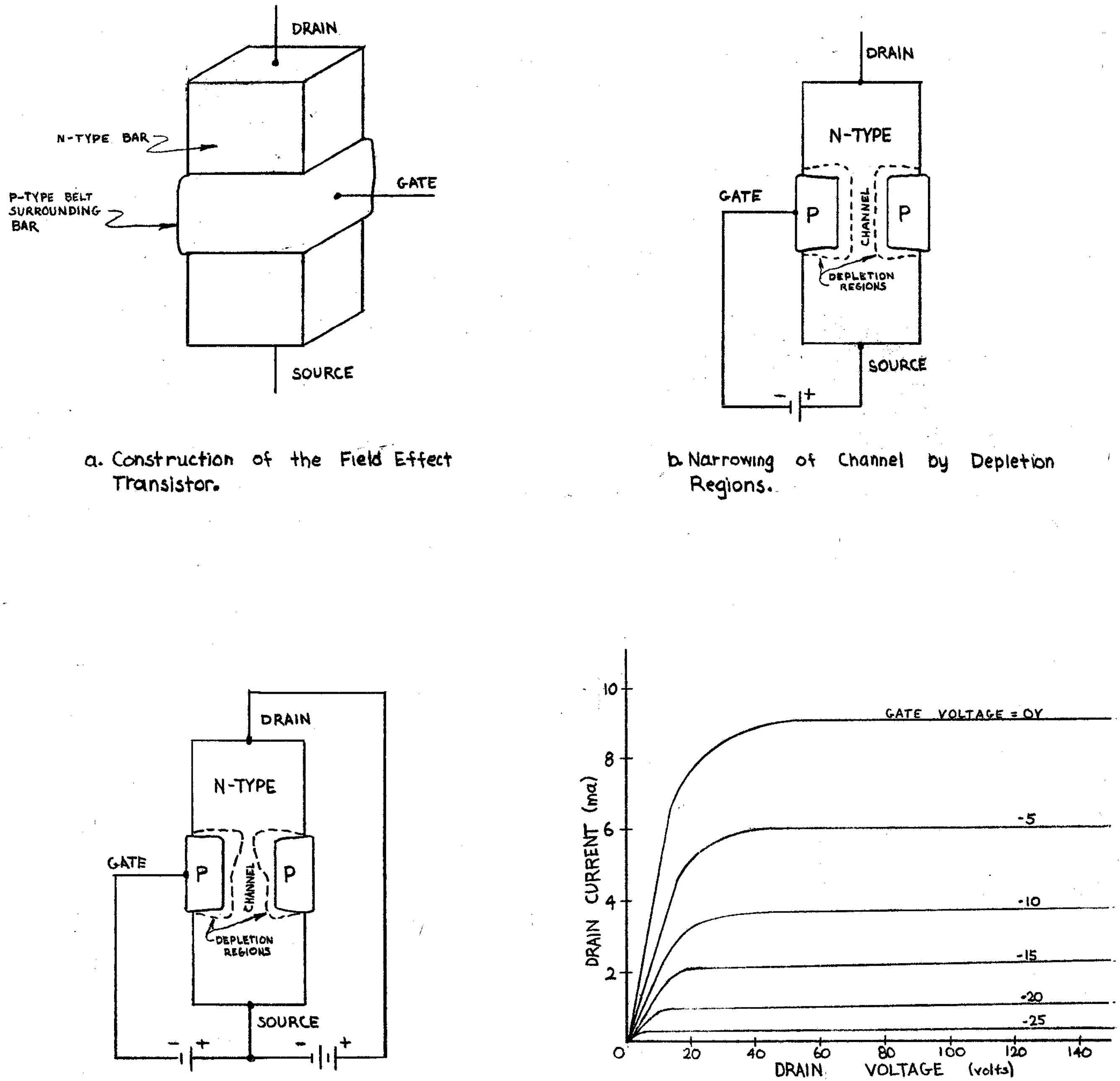
Figure 4.26. Volt-Ampere Curve for the Injecting Contact on a PNPM Thyratron Transistor for Both Directions of Current Flow

and emitter junctions because of the barriers existing there so they will set up a charge unbalance which reduces these barriers and forward biases both the collector and emitter junctions. Hence, the emitter will inject holes into the base which diffuse to the center junction. Likewise, the collector will inject electrons into the floating N region which will also diffuse to the center junction. This makes available at the center junction a far greater number of current carriers than are required to sustain the avalanche breakdown. Therefore, the voltage across this junction decreases with increasing current. This situation is represented by the regenerative region in figure 4.24. The voltage will continue to fall until the drop across the transistor is equal to that required to forward bias the emitter and collector junctions and make up for the ohmic losses in the material. In this condition, the device exhibits a very low resistance.

The thyratron transistor will remain in this conducting state until the current through it falls below the holding current. When this happens, the number of excess carriers in the central P and N region will be insufficient to forward bias the emitter and collector junctions so the transistor returns to the nonconducting state.

When a PNPN triode is in the off state, a small current does flow. This is the saturation current of the reverse biased center junction. This small current will produce a charge unbalance in the central P and N regions which forward biases the emitter and collector junctions, permitting the injection of carriers. However, these injected carriers will become trapped in recombination centers in the central P and N regions (i.e. crystal imperfections such as missing atoms in the lattice). As long as the carriers are not injected faster than the trapped carriers recombine, very few will reach the center junction; and the device will not switch into the conduction state. After the current reaches a certain level, the recombination centers become saturated. Then the injected carriers will diffuse to the center junction producing a drop in its resistance, switching the transistor into the on state.

If the emitter-collector voltage is less than the breakover voltage, the PNPN transistor can be switched on by applying a trigger signal to the base that forward biases the emitter junction. Holes will then be injected into the base and will diffuse to the center junction, producing a collector current. If a large enough trigger is applied, the collector current will increase above the threshold value; and the collector will inject enough carriers into the floating region to saturate the recombination centers so the transistor will switch on.



a. Construction of the Field Effect Transistor.

b. Narrowing of Channel by Depletion Regions.

c. Distortion of Depletion Region by Drain Voltage.

d. Characteristic Curves.

Figure 4.27. The Field Effect Transistor.

Figure 4.25 gives the turn on characteristics of a thyratron transistor. The characteristics are similar to those of an ordinary junction transistor as long as the collector current is below the threshold level. When collector injection starts, the transistor becomes regenerative and switches on.

If a reverse bias is applies to the emitter-base junction, the injection of holes by the emitter will be stopped. When this happens, holes will no longer be swept across the center junction into the floating region; and the forward bias on the collector junction will drop. Collector injection will then fall off, and the transistor will become nonconducting.

The PNPM transistor show in figure 4.23b differs from the PNPN transistor in that a metal-semiconductor contact is used as the collector. As long as the collector current is below a certain value, the metal contact is ohmic; and the device operates as a PNP transistor. However, when the current through the metalic contact is increased above a certain level, it injects electrons into the floating P region as would a N-type collector. When this happens, the PNPM triode will switch into a low resistance state.

The threshold current of the PNPM transistor is determined by the characteristics of the injecting metal contact, not by the presence of recombination centers in the floating P region. When a current is reached where the metal contact begins injecting, the transistor switches on.

The volt-ampere characteristics of the injecting contact are given in figure 4.26. The contact has a low resistance to current in either direction. However, at high currents, when the semiconductor is positive with respect to the contact, the contact is no longer ohmic; but it displays a sharp drop in resistance with increasing current. This is caused by the injection of electrons into the semiconductor from the metal.

The Field Effect Transistor. The construction of the field effect transistor is shown in figure 4.27a. A small, high resistivity N-type bar is surrounded by a belt of low resistivity P-type material. Ohmic contacts are made to both ends of the bar and to the P-type belt.

When this PN junction is reverse biased as shown in figure 4.27b, the depletion region will extend primarily into the bar because of its high resistivity. Therefore, the bias on this junction can control the resistance of the bar by modulating the width of the conducting channel between the depletion regions. Hence, if a voltage is placed across the bar as shown in figure 4.27c, the resulting current can be controlled by the junction bias.

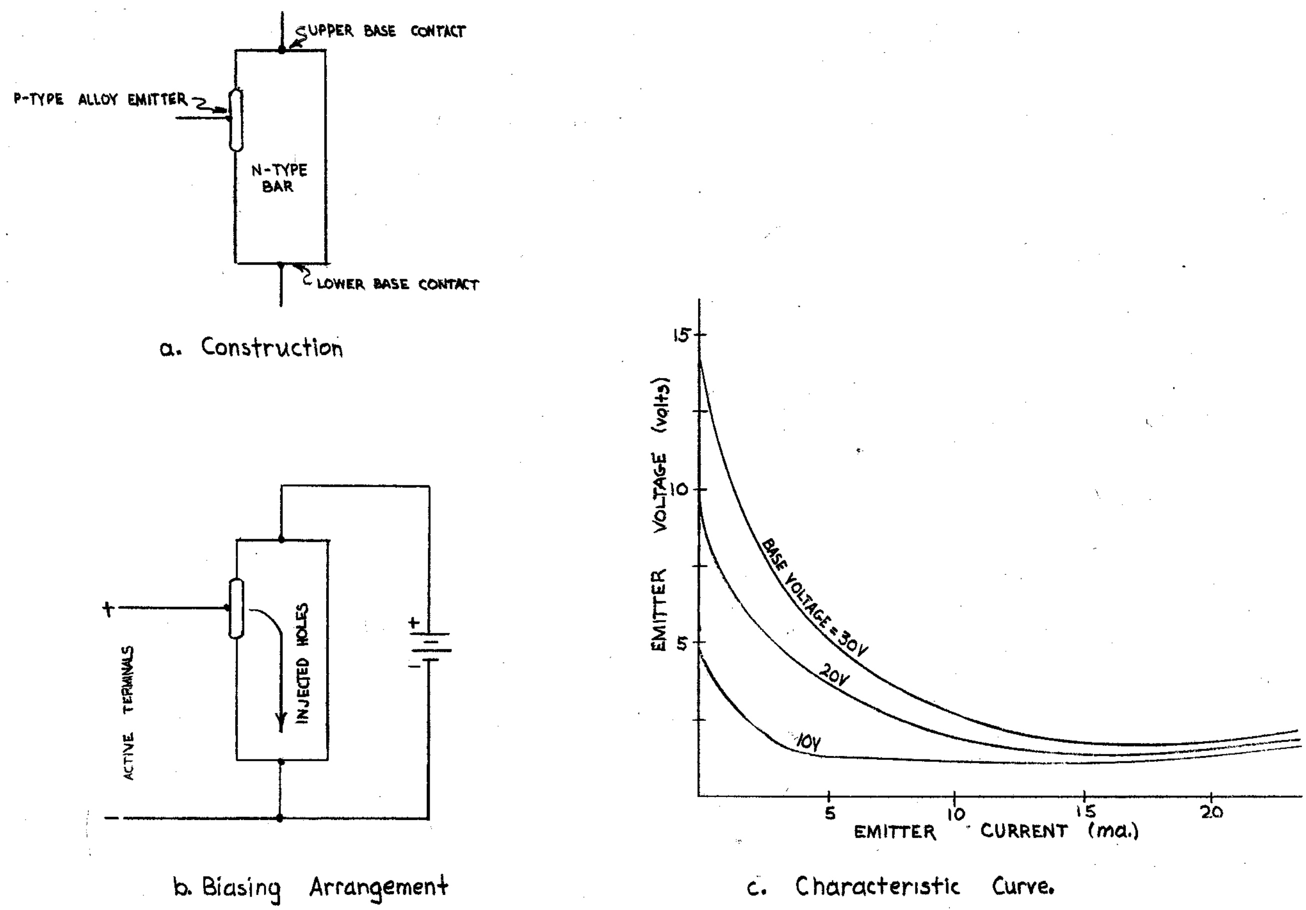


Figure 4.28. The Double Base Diode, Diagrams and Characteristics.

The electrical characteristic of the field effect transistor are shown in figure 4.27d. It can be seen that increasing the reverse bias on the gate will lower the drain current by decreasing the channel width.

At low drain voltages, the drain current increases linearly with voltage. This is to be expected because of the resistive nature of the bar. However, as the drain voltage increases, the drain current approaches a constant value. This happens because the voltage dropped in the channel will produce an additional reverse bias on the gate junction, decreasing the channel width. This action tends to keep the drain current constant. Since the channel becomes more positive toward the drain terminal, widening of the depletion region will take place near the drain, producing the distortion shown in figure 4.27c.

The electrical characteristic of the field effect transistor are similar to those of a vacuum tube. The input impedance is high (50,000-200,000 ohms) because the gate junction is reverse biased. Furthermore, the curves shown in figure 4.27d resemble those of a vacuum pentode.

The Double Base Diode. The double base diode is a regenerative switch that has found application in relaxation oscillator circuits. It is made by alloying a P-type emitter to a high resistivity N-type bar which has ohmic contacts attached to both ends. This configuration is shown in figure 4.28a.

In operation, a voltage is applied between the two base terminals, positive on the upper base as shown in figure 4.28b. This voltage will be distributed evenly along the length of the bar, so that portion of the bar near the emitter junction will be at some voltage less than the upper base voltage. Therefore, when the active terminals are shorted together (emitter to lower base), the emitter junction will be reverse biased. If now a positive voltage is applied to the emitter, the reverse bias will be reduced; and if the emitter voltage becomes greater than the voltage on the bar near the junction, the emitter will become forward biased and inject holes into the bar. These holes will be swept toward the lower base by the interbase electric field, lowering the resistivity in the region below the emitter. When this happens, the voltage distribution along the bar is altered, the lower portion of the bar becoming less positive because of its reduced resistance. Hence, the forward bias on the emitter is increased; and more carriers are injected. This regenerative action continues so the current will increase until the emitter voltage drops to a low value. This produces the negative resistance emitter characteristics shown in figure 4.28c.

If the emitter voltage is reduced while the device is in the on state, the number of injected carriers will decrease; and the resistance of the region below the emitter will increase. This will decrease the forward bias on the emitter so another regenerative cycle will begin, and the device will turn off.

DEPARTMENT OF WEAPONS TRAINING
LOWRY AIR FORCE BASE
COLORADO

T R A N S I S T O R S

IN SERVICE TRAINING COURSE

DEPARTMENT OF THE AIR FORCE · UNITED STATES OF AMERICA · MCMXLVII

CHAPTER 5

THE FABRICATION OF DIODES AND TRANSISTORS

AIR TRAINING COMMAND · UNITED STATES AIR FORCE · MCMXLIII

2 AUGUST 1961

FOR INSTRUCTIONAL PURPOSES ONLY

CHAPTER 5

Author: A1C Robert J. Widlar

THE FABRICATION OF DIODES AND TRANSISTORS

__Abstract__ - The preparation of germanium and silicon single crystals is described from the chemical purification and zone refining to the crystal growing and doping. Examples of both the techniques and the equipment used are given. Next, the formation of PN junction by direct growing, alloying, and diffusion methods is explained. Special emphasis is placed on the fabrication of practical diodes and transistors. The attachment of ohmic and rectifying metal-semiconductor contacts is also covered. Soldered, bonded, and deposited contacts are included. Lastly, the surface treatment and encapsulation of completed devices is briefly considered.

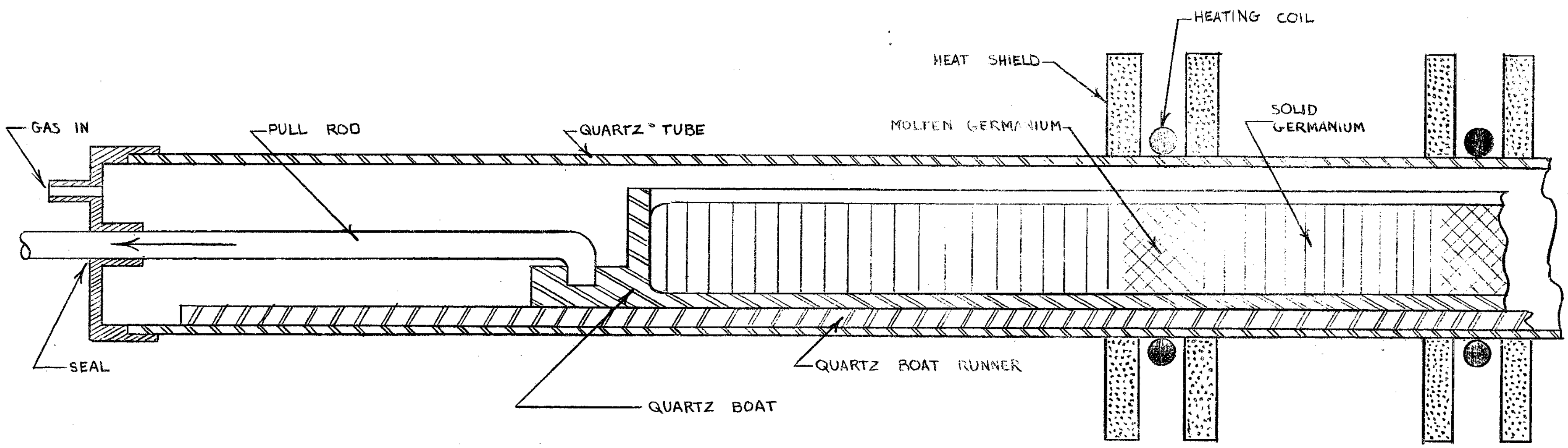


Figure 5.1. Germanium Zone Refiner.

INTRODUCTION

The purpose of this chapter is not so much to describe the exact processes used in manufacturing semiconductor devices, but to show the feasibility of producing extremely pure materials and seemingly complex structures on a production scale. In view of this, the chapter will bring out some of the more important processes used in preparing and purifying semiconductor materials. Moreover, the growing and doping of high quality crystals will be covered, and some of the techniques currently employed in fabricating PN junction devices will be explained. This will include direct growing, alloying, and diffusion techniques as applied to diodes and two junction transistors.

At the present time, the only semiconductor materials used to any appreciable extent are germanium and silicon. Of these, the technology of germanium is fairly well developed, with silicon lagging behind somewhat because of handling difficulties associated with its high melting point. This chapter will be confined to these two materials. Other compounds, i.e. silicon carbide, gallium arsenide, and indium antimonide, have limited application; but it is difficult to obtain good quality crystals so they are of minor importance and will not be covered here.

PREPARATION AND PURIFICATION

Before a semiconductor is suitable for use in a diode or transistor, the impurity concentration must be reduced to less than one part per hundred million . Prior to the transistor, this degree of purity could only be realized using chemical methods on a laboratory scale. However, intensive research has resulted in new techniques which can be applied to the mass production of extremely pure semiconductor materials.

Germanium. Germanium can be purified chemically by the fractional distillation of the volatile liquid, germanium tetrachloride (boiling point: 83°C). Germanium tetrachloride decomposes on contact with water into insoluble germanium dioxide, and the dioxide can be reduced to elemental germanium by passing hydrogen gas over it in a reduction furnace operating at 650°C. Using this process, germanium of extraordinary purity can be obtained. However, before it can be used in semiconductor devices, it must be refined further.

Most of the impurities remaining in germanium after chemical purification are more soluble in molten germanium than in the solid. Therefore, when molten germanium is cooled, these impurities will tend to stay in the melt as the material freezes. This suggests a method for further purification. This is, if a rod of molten germanium is progressively solidified down its length, that portion which solidified first will have a lower concentration of impurities.

ELEMENT	IMPURITY TYPE	SEGREGATION COEFFICIENT		MELTING POINT	BOILING POINT
		Germanium	Silicon	°C	°C
Boron	P	20	0.9	2300	2550
Aluminum	P	0.1	0.004	660	1800
Gallium	P	0.1	0.01	29.8	1600
Indium	P	0.001	0.0004	155	1450
Phosphorus	N	0.12	0.35	44	280
Arsenic	N	0.03	0.3	615	615
Antimony	N	0.003	0.04	630	1380
Zinc	P	0.01	—	420	907
Copper	P	0.00001	—	1083	2300
Gold	P	0.00001	0.00003	1063	2600
Iron	P	0.000001	—	1535	3000

Table 5.1. Physical Constants of Impurities Frequently Encountered in Semiconductor Work.

This portion of the rod could be cut off and the process repeated until a material of the desired purity was obtained.

One disadvantage of this system is that the rod must be removed from the furnace after each successive freezing to crop off the contaminated end. This is complicated by the fact that germanium must be melted in an inert atmosphere because it reacts with oxygen at these temperatures to form germanium dioxide. Hence, this process will be time consuming since the furnace must be purged to remove oxygen before heat is again applied. This difficulty is overcome by the zone refining process.

Zone Refining. At the present time, zone refining is used in the purification of practically all electronic grade germanium. In this process, a narrow molten zone is passed down the length of a germanium bar. The impurities tend to stay in the melt so they too are carried down the bar. An advantage of this system is that many passes can be made without removing the bar from the furnace.

A drawing of a zone refiner is given in figure 5.1. The germanium is contained in a quartz boat which is inserted into a quartz tube. The tube is sealed and purged with nitrogen, and then filled with an inert gas. Gas flow is maintained throughout the process to prevent the seepage of oxygen into the system. Molten zones are created in the germanium by radiation from heating coils. These zones are confined by the use of heat shields between the coils. The regions between the coils is left open to facilitate cooling of the germanium between zones.

After the molten zones have been established, the boat is slowly pulled down the tube. The germainum passing under the coils is then melted, and that passing out from the coils is cooled by radiation and refreezes. Therefore, the molten zones move down the bar, carrying along the impurities. The process can be recycled by removing heater power and returning the boat two its original position. After several passes, the impurities will be concentrated at one end of the ingot. Normally, about 90 percent of the bar will have adequate purity.

The number of passes required is determined by the nature of the impurities. The greater their solubility in molten germanium as compared with the solid, the more effective will be the process. This characteristic is described for a given impurity by the segregation coefficient which is the ratio of its solubility in the solid germanium to its solubility in the liquid. The segregation coefficient of several impurities frequently encountered in semiconductor work are given in table 5-1.

Fortunately, true chemical purity is not always required. Impurities that do not produce current carriers are of little

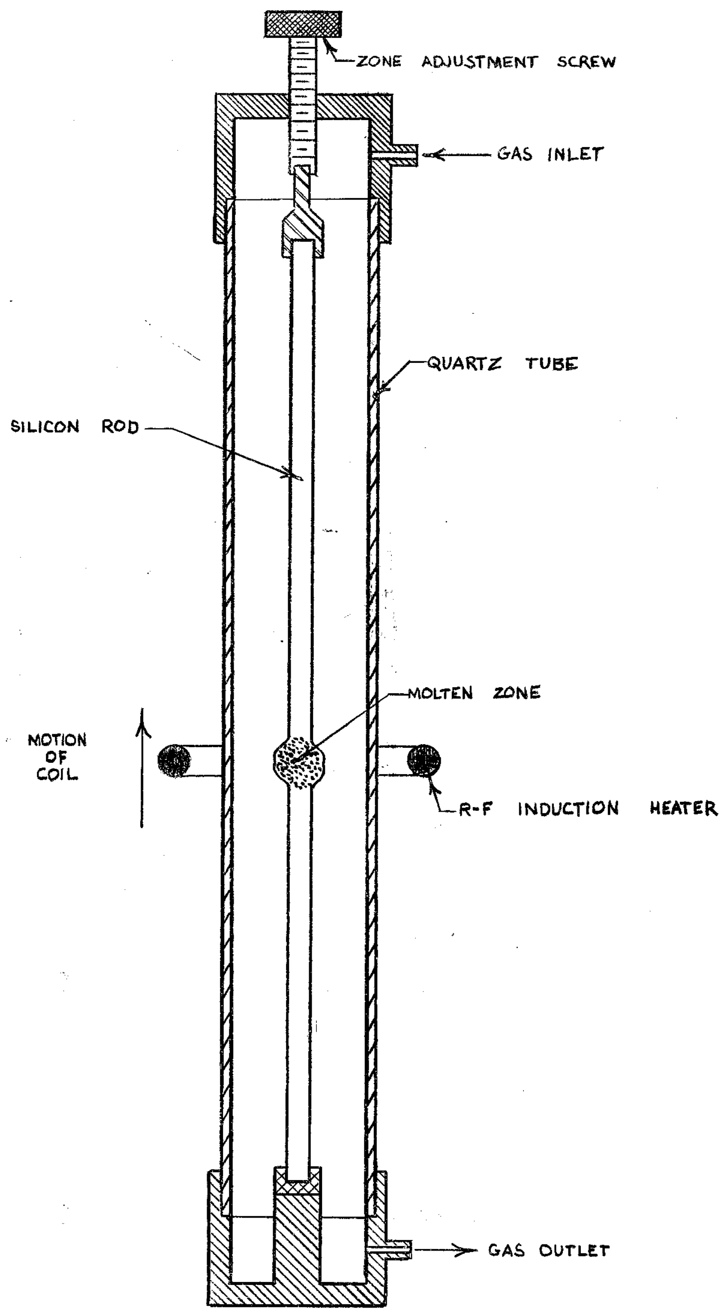


Figure 5.2. Silicon, Floating-Zone Refiner.

consequence as long as their concentration is low enough so that they do not produce excessive distortions in the crystal structure. It is believed that with zone refining some impurities such as boron, which has a greater solubility in solid germanium, are not removed but are converted to their respective oxides. In this state they do not produce current carriers since the outer shell electrons are tied up in producing a single, oxide molecule.

The purity of a semiconductor sample can easily be determined by measuring its resistivity and comparing it with the theoretical intrinsic resistivity of the material. With zone refining, near intrinsic resistivities can be obtained by making a large number of passes.

Silicon. Chemical purification of silicon can be accomplished, as with germanium, by fractional distillation of volatile silicon tetrachloride. The pure silicon tetrachloride is then reduced to elemental silicon by exposure to metalic zinc at high temperatures. The reaction product, zinc chloride, is volatile at these temperatures and is drawn off. The excess zinc can be removed in subsequent operations by evaporation since its boiling point is lower than that of silicon.

Further purification of silicon is difficult so the chemically pure silicon is frequently used "as is". Zone refining has been tried, but it does not give any significant improvement. This is partially due to the high segregation coefficient of most impurities in silicon and the contamination of silicon by its container.

High purity quartz is about the only satisfactory material that can be used to contain molten silicon. The reaction between quartz (silicon dioxide) and molten silicon to produce silicon monoxide proceeds slowly. However, molten silicon tends to take on impurities from the quartz; and it will also stick to quartz upon cooling causing frequent breakage of containers.

A process that has been used to produce high quality silicon is floating zone refining. The apparatus is shown in figure 5.2. A silicon rod is vertically supported at both ends within a sealed quartz tube. A single molten zone is established in the rod by a RF induction heater. The molten zone is contained by surface tension so there is no container problem. In operation, the heating coil is raised slowly, carrying the molten zone and the impurities to the upper end of the bar.

Only one zone can be maintained in the bar at one time so a large number of passes are required to obtain the required purity. Boron, which is frequently found in silicon, cannot be removed in any reasonable number of passes because of its high segregation coefficient. However, it can be removed chemically by passing hydrogen gas saturated with water vapor through the apparatus during operation.

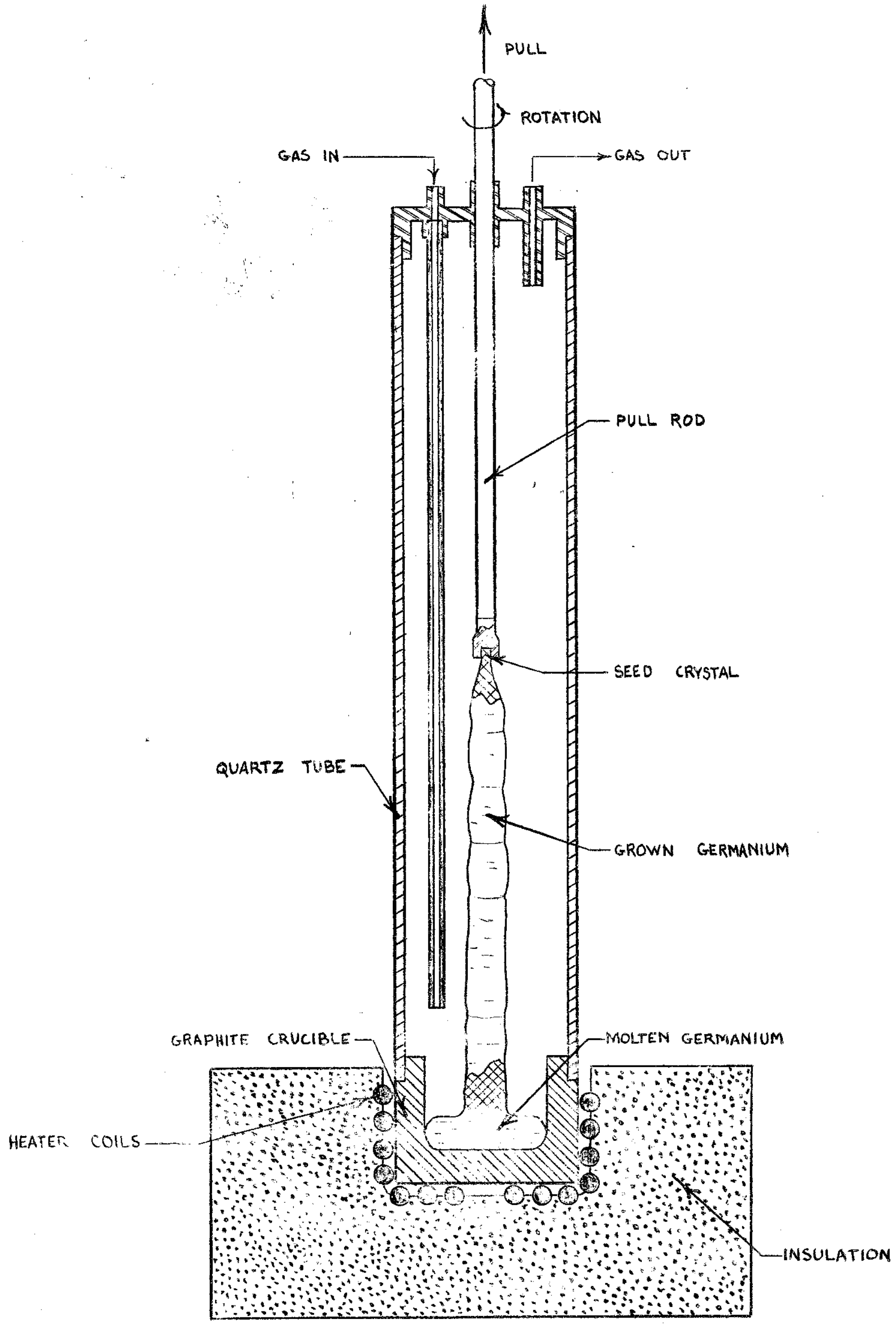


Figure 5.3. Germanium Crystal Puller.

CRYSTAL GROWING

Normally, when a molten semiconductor is cooled, freezing will start at several points; and a number of separate crystals will be formed. This polycrystaline material is useless for the fabrication of diodes and transistors because extensive discontinuities exist at the interface between crystals. However, if the melt is seeded with a single crystal and cooled in such a way that freezing proceeds around the seed, the atoms of the molten material will build onto the seed crystal and produce a single crystal. This process is referred to as crystal growing.

Germanium crystals are usually grown in a crystal puller like that shown in figure 5.3. Similar equipment can be used for growing silicon crystals, except that the graphite crucible must be replaced by quartz.

In operation, a seed crystal is lowered into contact with molten germanium which is heated just above its melting point (960°C). That portion of the melt in contact with the seed will be cooled by conduction to the seed holder so freezing will begin. As the seed is withdrawn slowly (a few inches per hour) a rod of single crystal germanium will be grown.

The diameter of the grown crystal depends on both the melt temperature and the pull rate so these quantities must be carefully controlled. The seed crystal is rotated during the pulling operation to promote uniform growth.

Silicon single crystals can also be grown using the floating zone apparatus illustrated in figure 5.2. Although it has its disadvantages, this system eliminates the container problem associated with silicon. To produce a single crystal, the lower support is seeded before the silicon rod is inserted. Then, as the floating zone is raised, silicon will refreeze on the seed producing the desired single crystal.

Since defects in a semiconductor crystal introduce recombination centers, the quality of a grown crystal can be determined from measurements of the minority carrier lifetime. The lifetime can be measured by creating hole electron pairs in the crystal, i.e., with a pulse of light, and measuring how long these carriers contribute to conduction through the crystal. If, for example, an intrinsic crystal is exposed to light its conductivity will increase. After the light source is removed, conductivity will fall off exponentially because of carrier recombination. The time required for the conductivity change to decrease to one half its value is an indication of the minority carrier lifetime.

Doping. A semiconductor can be doped by the addition of certain impurities to the melt during the crystal pulling operation. The impurity atoms will grow into the crystal taking the place of germanium or silicon atoms in the crystal structure. Boron, aluminum, gallium, and indium will produce P-type crystals while phosphorus, arsenic, and antimony will yield N-type crystals. Other impurities can be used for doping, but they generally produce complex electron structures in forming covalent bonds. This gives rise to recombination centers in the crystal.

The concentration of the impurity in the grown crystal will normally be lower than that in the melt because of impurity segregation. The segregation coefficients given in table 5.1 are for slowly grown crystals. Increasing the growth rate will also increase the segregation coefficient. The amount of increase will depend on the element: impurities with low segregation coefficients are generally more sensitive to growth rate. If the crystal is grown rapidly enough, impurity segregation will not take place so the impurity concentration in the crystal will equal that in the melt.

The best impurity for a particular application is determined by many factors, including segregation coefficient, the maximum solubility of the impurity in the crystal, and its boiling point. For heavy doping, an impurity with a high segregation coefficient and high solubility must be used. For light doping, an impurity with a low segregation coefficient is desirable since it would be difficult to measure the extremely small quantity of impurity required if one with a high segregation coefficient were used. The boiling point is also improtant. If a volatile element such as phosphorus is added to the melt, most of it will be lost by evaporation. This can cause contamination of the crystal puller which could affect subsequent operations.

JUNCTION FORMATION

There are several different methods available for making PN junctions using various doping techniques. For example, a crystal can be directly doped during growth to produce a PN junction; or impurities can somehow be added to a grown crystal producing a junction as with the alloy or diffusion processes. Frequently, as a result of these processes, a particular section of a semiconductor will contain both donor and acceptor impurities. When this is the case, the carrier type will be determined by whichever impurity is present in excess. If donor impurities are present in a crystal containing an excess of acceptor impurities, they will not contribute electron current carriers to the crystal. Instead, the free electrons supplied by the donor atoms will be used to complete the covalent bonds of some acceptor atoms. Thus, the presence of donor impurities in a P-type crystal will reduce the effective number of acceptor impurities by providing electrons which fill vacancies in the covalent bonds.

The presence of two impurity types in a single crystal will not produce any harmful effects as long as the total impurity concentration is kept sufficiently low. If, however, the total impurity concentration becomes too high both the carrier mobility and the monority carrier lifetime of the crystal will be reduced by the localized fields of the impurity atoms.

Direct Doping. A PN junction can be made by directly doping the melt during crystal growth. In this process the melt is doped with a single impurity and a length of crystal grown. Then an impurity of opposite type is added to the melt in sufficient quantity to reverse the carrier type in the grown crystal. This produces a PN junction at the point of impurity reversal. Several junctions can be formed on a single bar by alternately reversing the impurity type, but the quality of the junctions will be progressively degraded due to the buildup of compensating impurities. This problem can be overcome by placing two crucibles side by side in the crystal growing apparatus. The melt in one crucible is doped with P-type impurity and that in the other with a N-type impurity. The seed crystal is dipped into the first crucible, and a length of P-type crystal is grown. It is then withdrawn and dipped into the other crucible, so a length of N-type material can be grown. This process can be repeated, and a large number of junctions grown.

Direct doping can also be used in the production of transistors. Again the carrier type in the grown crystal is controlled by the addition of impurities to the melt during growth. The collector is grown first and is lightly doped to reduce the amount of impurity required to produce a reversal when the base is grown. This is necessary because the performance of a transistor is greatly affected by the minority carrier lifetime in the base. The emitter, which is grown last, is heavily doped to give good injection efficiency.

After the emitter is grown, the melt is left heavily doped so the process cannot be repeated. Again reversing the impurity type would give an excessive impurity concentration and would deteriorate crystal quality. This shortcoming can be overcome by using the two (or three) crucible system.

A graded junction is usually formed with direct doping because of the slowness of the growth process. Furthermore, a wide range of resistivities can be obtained without complication since this is done merely by the addition of appropriate amounts of impurities. Finally, a large number of devices can be made from a single ingot by slicing out a section containing the junction and dicing it into individual units.

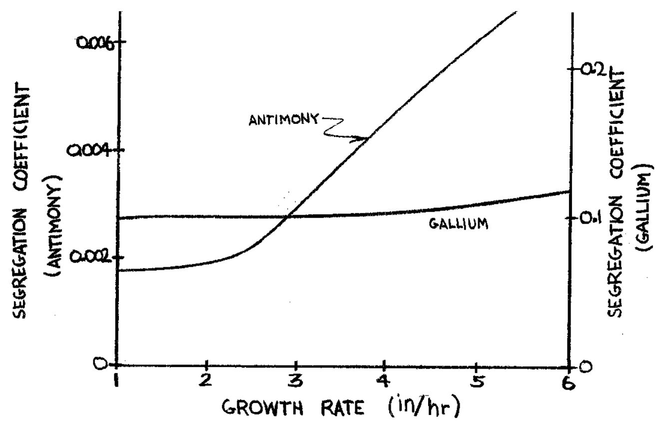
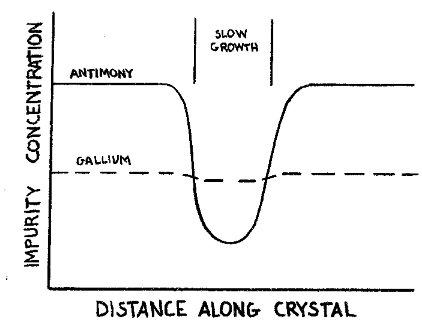
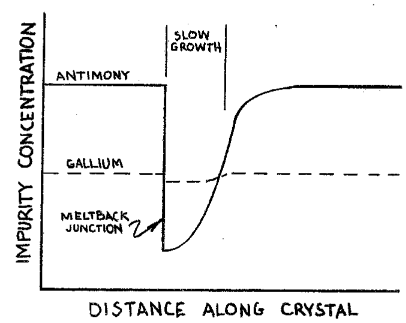


Figure 5.4. Variation of Segregation Coefficient with Growth Rate for Antimony and Galhum (in Germanium).



a. Without Meltback



b. With Meltback.

Figure 5.5. Impurity Concentrations in a Rate-Grown Transistor.

Rate Growing. In the rate growing process, both donor and acceptor impurities are present in the melt. One impurity is made to dominate over the other in the grown crystal by changing the growth rate.

As shown in figure 5.4, the segregation coefficient of antimony increases markedly with crystal growth rate, while that of gallium is practically constant over the range shown. If both these impurities are present in a germanium melt, their relative concentrations in the grown crystal will depend on the growth rate. If the melt is properly doped, an impurity reversal can be obtained by changing the growth rate such that rapid growth will produce an N-type crystal and slow growth will produce a P-type crystal.

This method is particularly applicable to the formation of transistor junctions. An N-type crystal is produced by rapid growth, but the growth rate is intermittently slowed to give a P-type base as shown in figure 5.5a. Using this system many sets of junctions can be produced in a single ingot by periodically slowing the growth rate.

It is advantageous to produce the base during slow growth because the base region is most sensitive to crystal quality. Slow growth reduces both the total impurity concentration and the likelihood of grown-in imperfections, such as missing atoms.

The usual method employed to slow the growth rate in transistor production is to increase the temperature of the melt. This causes the germanium to freeze less rapidly. If the melt temperature is increased enough so that a portion of the grown crystal is remelted, an abrupt junction will be produced as shown in figure 5.5b. This happens because the gradual junction produced as the temperature increased is melted off. The second junction will be formed as the melt cools. Impurity reversal will take place gradually and a graded junction will be produced.

The advantages of the rate grown transistor employing meltback should be noted. A graded impurity distribution is grown into the base giving an accelerating drift field, and the collector junction is graded which results in a lowered collector capacitance and a higher maximum collector voltage.

With rate growing a great degree of control over the base width and resistivity profiles is possible, but it is difficult to realize the low emitter resistivities necessary for high injection efficiencies at high current levels. This is fixed by the segregation properties of the impurities used. For this reason, rate growing is used primarily for the production of low power, medium-frequency transistors.

Rate growing can also be used for making junction diodes. In this case the growth rate is usually altered by changing the rate of

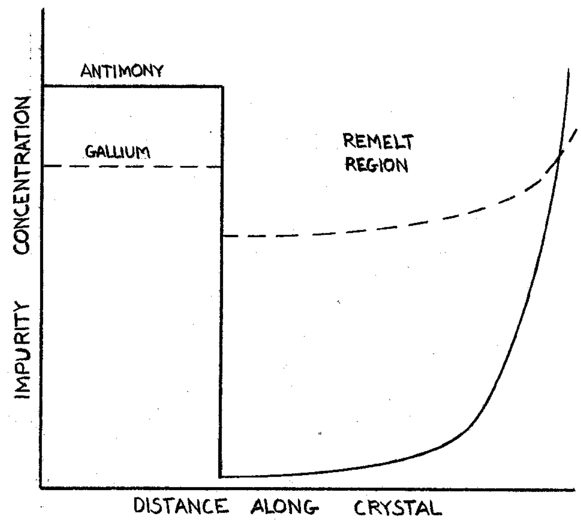


Figure 5.6. Formation of a Remelt Junction.

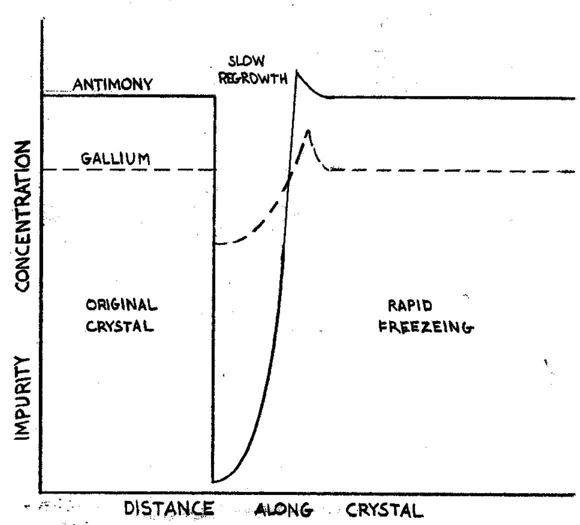


Figure 5.7. Formation of a Remelt Transistor.

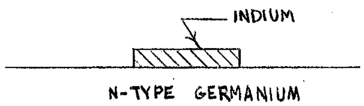
withdrawl although temperature control may be employed. The latter method usually results in alternate junctions having somewhat different characteristics. This is especially true if any amount of meltback takes place. Again a disadvantage of rate growing is the limited resistivity range. The low resistivities necessary for high conductance diodes cannot be realized. However, the great degree of control possible over junction characteristics, particularly in producing graded junctions, makes the system attractive for some applications requiring high reverse voltage or low junction capacitance.

Remelt Junctions. The remelt operation is performed on a good quality germanium crystal that has been doped with both gallium and antimony, but with antimony in excess. One end of this bar is heated causing it to melt. After about half the bar has been melted, the heat source is removed; and the bar is allowed to refreeze slowly. As the slow regrowth proceeds, an impurity reversal takes place because much less antimony will go into the crystal at the slow growth rate. Thus, a PN junction is produced at the furthest penetration of the molten zone as shown in figure 5.6.

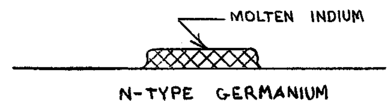
After the regrowth proceeds for some distance, the impurity concentration in the melt will increase greatly since all the impurities do not grow into the crystal. This is especially true for antimony. Therefore, near the end of the bar, which is last to freeze, the antimony will again dominate and a second junction will be produced. There will be a considerable distance between these two junctions so the structure is not suitable for use in transistors. However, if the second junction is cropped off, the remelt junction can be used in the fabrication of diodes.

Remelt transistors can be made by attaching one end of the double-doped bar to a heat sink while the other is heated strongly. After about half the bar has melted, the heat is removed. At first the bar will refreeze slowly, but rapid freezing will soon take over because of conduction to the heat sink. Impurity segregation takes place in the slow regrowth region (about 0.001 inch wide) so an abrupt junction is produced as shown in figure 5.7. However, as rapid freezing takes over, the impurities do not segregate. Instead, they grow into the crystal in their original concentration. This produces a second junction close to the first. As shown in the figure, this junction will tend to be graded.

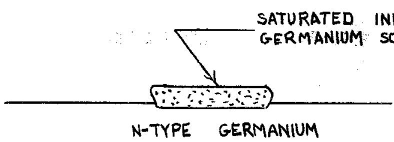
Reasonably narrow junctions can be obtained with reproducable results using this process. Furthermore, a graded base with its associated accelerating field is produced. Because this is essentially a rate-growing process, similar limitations on the range of resistivities obtainable are present. Hence, the remelt technique is used in practice for low power transistors.



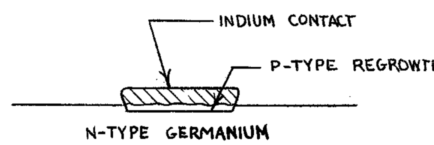
a. Indium dot is placed on N-type germanium and heated.



b. Indium melts (155°C).

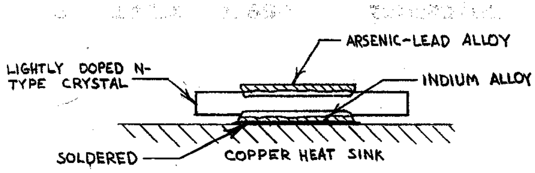


c. Indium dissolves germanium.

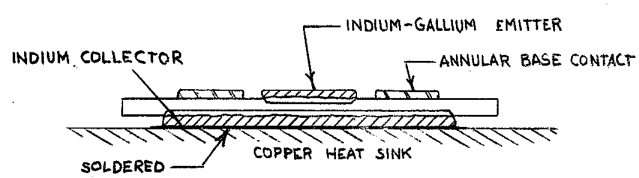


d. On cooling, dopped germanium recrystalizes.

Figure 5.8. Steps in the Formation of an Alloy Junction.



a. High Efficiency PNN+ Diode.



b. Alloy-Junction power transistor.

Figure 5.9. Examples of Alloy-Junction Devices.

Alloying. The formation of an alloy junction is illustrated in figure 5.8. A low melting point P-type impurity such as indium is placed on a lightly doped N-type germanium wafer. The assembly is heated in an oven containing an inert atmosphere. The indium will melt and take some of the germanium into solution. The assembly is then cooled, reducing the solubility of germanium in the molten indium. This causes the dissolved germanium to recrystalize on the N-type wafer. In recrystalizing the germanium will take along a considerable amount of the P-type impurity so a junction will be formed between the recrystalized germanium and the N-type wafer.

The depth of the alloy junction is determined by the maxium temperature used in the process. The solubility of germanium in molten indium increases with temperature. Thus, at higher temperatures more germanium will be dissolved and the penetration will be deeper. Time of exposure at this temperature has little effect on the shape of the junction.

It is necessary for the molten indium to wet the crystal surface for successful alloying. An ohmic contact will be formed between the indium and the crystal in any unwet areas. A soldering flux is usually applied to the crystal surface to remove surface oxides and promote wetting. Only that area beneath the indium is fluxed so that the indium does not spread when heated.

PN junctions can be alloyed to a P-type crystal using N-type alloys. Arsenic and antimony have been used, but these materials are brittle and do not wet germanium easily. Other materials such as lead can be added to the alloy to improve the mechanical properties, and with proper fluxing the alloy can be made to wet the germanium. Nonetheless, better results are usually obtained with indium so P-type alloy junctions are prefered for most applications.

The low resistivity of the recrystalized region makes the alloy junction well suited for high conductance diode rectifiers. The abrupt junction, characteristic of the alloy process, is somewhat of a disadvantage in applications requiring high inverse voltages.

In one type of power diode (figure 5.9a) an acceptor impurity (indium) is alloyed to one side of a high resistivity N-type wafer, producing a PN junction. A donor impurity (arsenic) is alloyed to the opposite side of the wafer producing a NN^+ junction. Rectification takes place at the PN junction, which has a low forward resistance because of the very low resistivity of the recrystalized P region. Current flow across the junction is largely due to the injection of holes from the P region. The high resistivity base crystal serves to widen the depletion region, thereby permitting high inverse voltages. The second alloy junction provides an ohmic contact to the base crystal and lowers the forward resistance by shortening the current path through the high resistivity material.

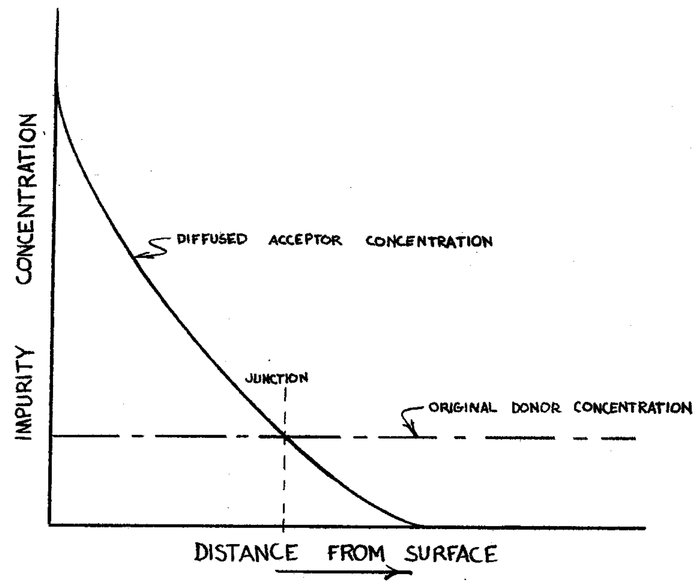


Figure 5.10. Formation of PN Junction by Vapor Diffusion of Acceptor Impurity into Dopped Crystal.

Alloy junction diodes are generally used in low frequency power applications requiring very low forward resistance. When low junction capacitance or high reverse voltage is the important consideration, the graded junctions produced by other methods are more desirable.

An alloy junction transistor is made by alloying junctions to opposite sides of a thin high resistivity wafer (figure 5.9b). The collector junction is usually made about three times larger than the emitter junction to improve collection of carriers crossing the base. In practice, base widths of 0.001 to 0.003 inch can be realized, but the degree of control is not too great. Most alloy junction transistors are the PNP type because of the superior alloying properties of indium.

Alloy junction transistors are used in low level audio amplifiers, audio frequency power amplifiers, and low frequency switching circuits. The design has not found much application at high frequencies. The alloy process is well suited to the fabrication of power transistors for the following reasons:

(1) The emitter can be heavily doped to give good injection efficiency at high current levels. Gallium, which has a higher solubility than indium in solid germanium, can be added to the emitter alloy to give these low resistivities.

(2) With proper techniques it is possible to make uniform, large-area junctions.

(3) Efficient heat transfer from the collector junction can be accomplished by soldering the collector directly to a heat sink. The thermal resistance of germanium is considerably higher than that of indium, and using the alloy process the only germanium in the heat path is the thin recrystalized region.

An additional advantage of the alloy process is that it can be performed by automatic machinery which lowers the device cost considerably.

Alloying techniques have been used on silicon with varying degrees of success. Problems are encountered in finding suitable alloys. Most tend to be very brittle. Furthermore, wetting of silicon with the alloy is hampered by the formation of silicon dioxide on the crystal surface. Alloy junction silicon transistors have not seen much use, but alloy emitters are widely used on silicon diffused base transistors. The impurity is usually vacuum deposited on the crystal surface and alloyed by subsequent heating.

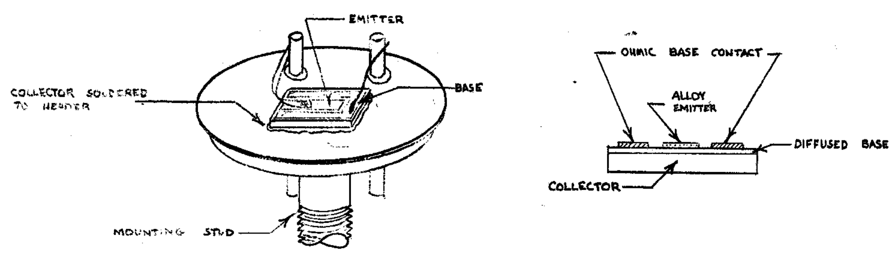


Figure 5.11. Construction of a Diffused Base Power Transistor Including a Section Showing Junctions.

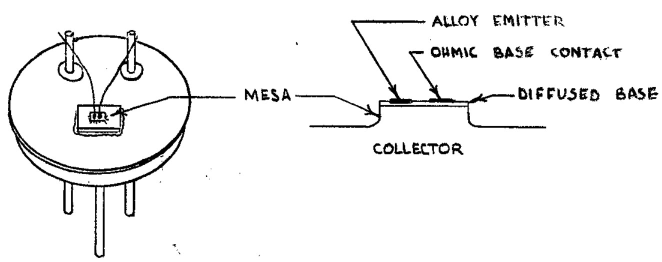


Figure 5.12. Construction of a Diffused Base Mesa Transistor.

Diffusion. Silicon and germanium can be doped by exposing a crystal to the vapors of an appropriate impurity at temperatures below the melting point of the crystal. Activated by thermal agitation, the impurity atoms bombard the crystal, some of them fixing themselves into the crystal structure by displacing the parent atoms. This will not just occur on the surface. The impurities can pentrate some distance into the crystal, but the number reaching the interior will fall off rapidly with distance. A comparatively long time is required for the impurities to diffuse into the crystal which makes the process easily controlled.

Diffusion doping is carried out in an evacuated furnace. The crystal and a volatile impurity are inserted into the furnace. On heating, some of the impurity will evaporate into the chamber and diffuse into the crystal. The depth of pentration will depend on the time of exposure, the temperature, and the impurity used (smaller impurity atoms diffuse more rapidly).

A PN junction can be formed on a doped crystal using diffusion doping. For example, if a donor impurity is diffused into a P-type crystal, an impurity reversal will take place near the surface of the crystal; and a juntion will be formed. The diffused-impurity concentration will fall off exponentially with distance from the surface so a graded junction will be produced. This is shown in figure 5.10. The impurity distribution and junction depth can be closely controlled.

At the present time the most important use of diffusion techniques is in the production of high performance transistors. A diffused base transistor is shown in figure 5.11. Acceptor impurities are diffused into the top of a thin N-type wafer, producing a junction about 0.0005 inch below the surface. Two thin metalic strips are deposited on the diffused P layer. The assembly is then heated. One of the metal strips is a P-type impurity so it alloys into the P-type base forming an ohmic contact. The other strip is an N-type impurity so it forms an alloy emitter on the base. The crystal is then soldered to the transistor case which serves as a heat sink and a collector terminal.

The advantages of this construction should be noted: The collector junction is graded giving a high collection breakdown voltage and low junction capacitance. The impurity distribution in the base is graded giving rise to an accelerating drift field. The graded base extends all the way to the surface so the injected carriers are accelerated away from the surface, and surface recombination is greatly reduced. The base is extremely thin. Base widths of 0.0001 inch can be realized in production. An alloy emitter is used. This gives high injection efficiencies at high

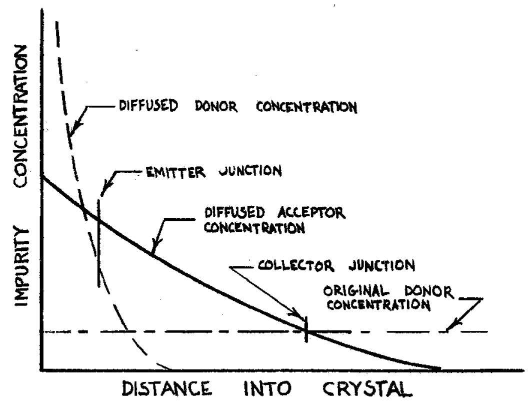


Figure 5.13. Plot of the Impurity Concentrations in a Double Diffused Transistor.

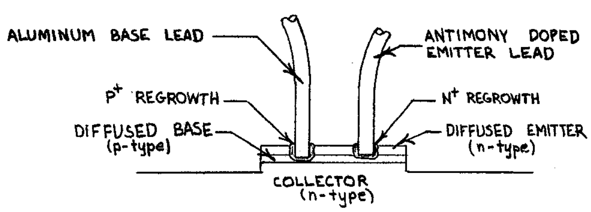


Figure 5.14. Construction of a Double Diffused Mesa Transistor Showing Compression Bonded Emitter and Base Leads.

current levels. A short thermal path is provided between the collector and the heat sink. In addition, the diffusion process is applicable to silicon. The collector junction of silicon transistors can be operated at a higher temperature, greatly increasing maximum power dissipation. Large area planar junctions can be produced without complication. The entire operation can be controlled closely and yields transistors having uniform characteristics.

Another diffused base transistor is shown in figure 5.12. It is similar to the one already described except that special techniques along with the "mesa" construction are employed to greatly reduce size.

In the mesa transistor the emitter and the base contact are deposited on the diffused base. Typically, both are a few ten thousandths of an inch wide. The excess material surrounding these contacts is etched away leaving a small mesa protruding from the base crystal. This greatly reduces junction area. The emitter and base leads are fine gold wires that are compression bonded to the deposited metal.

Significant results are obtainable with the mesa construction. Transistors capable of delivering several watts at frequencies above 250 megacycles have been made. High power mesa transistors that operate at kilowatt power levels at frequencies in the order of 10 megacycles are being developed. And an experimental mesa transistor that operates above 1000 megacycles has been built. In this transistor, the area of the mesa is less than the cross-sectional area of a human hair.

<u>Double Diffusion.</u> If a fast-diffusing acceptor impurity and a high-solubility donor impurity are simultaneously diffused into a P-type crystal, a NPN structure will be produced. This is illustrated in figure 5.13. The acceptor impurity will diffuse furthest into the crystal, but the donor will be present in higher concentration near the surface. As the acceptor concentration falls off with distance, the impurity type will revert to the original N-type. Small base widths can be obtained with this method. Since the process is governed entirely by the physical laws of diffusion, the results are easily reproduced.

Double diffusion techniques are particularly applicable to the fabrication of mesa transistors. The usual method of lead attachment is shown in figure 5.14. Connection is made to the base through the emitter. The bonding process used to make this connection is similar to alloying. Thus, when the aluminum wire is bonded to the mesa, it is surrounded by a P^+ regrowth. A PP^+ ohmic contact is made to the base, but the PN junction formed with the emitter prevents shorting of the base lead to the emitter. Similarly, the antimony doped emitter lead is surrounded by a N^+ regrowth that makes an ohmic contact with the emitter, but will form a PN junction if it penetrates through to the base.

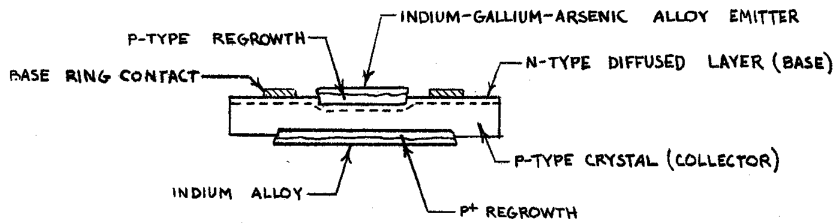


Figure 5.15. The Alloy-Diffused Transistor.

In power transistors of this type, it is desirable to diffuse a high-solubility donor impurity into the bottom of the wafer. This gives a triple diffused N+PNN+ structure. The highly doped region diffused into the collector reduces the thermal resistance between the collector junction and the heat sink, thereby facilitating the removal of heat.

Alloy Diffused Junctions. Multiple junctions can also be formed in a single operation using the alloy-diffusion process. An indium-gallium-arsenic alloy is placed on a P-type crystal and heated. When the alloy melts, arsenic will diffuse into the P-type crystal below the molten alloy and produce an impurity reversal since it diffuses considerably faster than either gallium or indium. When the assembly is cooled, gallium which has the highest maximum content in solid germanium will dominate in the recrystalized region. Hence, a P-type alloy emitter surrounded by a thin N-type layer will be formed. This is illustrated in figure 5.15.

Arsenic is usually evaporated into the furnace during this process to produce a low resistivity N-type layer on the crystal surface. This serves two purposes. First, a low resistance contact can be made to the base on this surface. Secondly, a strong drift field is created near the surface by the impurity gradient. This field turns the injected carriers away from the surface and reduces surface recombination. An indium-alloy contact is also fused to the P-type crystal. This contact does not serve as a collector, but instead it provides a low resistance path for the removal of heat from the collector junction.

Alloy-diffused transistors have found application at high frequencies and as low frequency power amplifiers where collector voltage and operating frequency requirements cannot be satisfied by alloy junction transistors. As with other diffusion transistors, the alloy-diffused transistor has a narrow base, graded base resistivity, and graded collector junction. The design is not adaptable to making very small devices, as in the mesa construction. It is, however, more economical since fewer operations are required in manufacture. The major frequency limitation of this transistor is the large area of the collector junction, but the design is useful at frequencies as high as 100 megacycles.

METAL-SEMICONDUCTOR CONTACTS

When a metal is attached to a semiconductor, either a rectifying or an ohmic contact will be formed. The type will depend on the metal used, the conductivity type of the semiconductor and the surface condition of the semiconductor where contact is made. Many metals will form a rectifying contact on a N-type semiconductor, but few will produce rectifying contacts on P-type material. Any rectifying contact will be degraded if mechanical defects or contaminants are present near the surface.

Metal-semiconductor contacts can also be classified as injecting or noninjecting. Rectifying contacts will inject minority carriers into the semiconductor while ohmic contacts will not. This distinction points out the necessity of strictly ohmic contacts in some applications. For example, if an injecting contact were made to the base of a transistor, injected minority carriers would be swept across the collector junction. This would increase the base current and reduce the current gain of the device. Injecting contacts have already been mentioned in connection with the PNPM transistor in chapter 4.

Soldering. Almost any soldered contact will be ohmic. Soldering introduces defects into the semiconductor from thermal stresses set up during the operation. Hence, essentially ohmic contacts will be formed even when the metals used normally produce rectifying contacts. Appropriate dopants are usually added to the solder when the requirements for a noninjecting contact are demanding. These dopants are selected so that NN^{++} or PP^{++} junctions are produced on the semiconductor in the soldering operation.

Bonding. A bonded contact is formed by fuzing a metal containing certain impurities to the semiconductor. This is usually accomplished by the simultaneous application of heat and pressure. During the process, a liquid solution of the metal and the semiconductor is created at the interface of the two materials. Upon cooling, doped semiconductor is redoposited around the contact. The junction formed is similar in many respects to an alloy junction, and sometimes no distinction is made between the two processes. The primary differences are the time required for junction formation and the method of applying heat. The time required for bonding is appreciably less. Heat is applied by passing a high current through the contact or by conduction through the metal. Wetting of the contact area can usually be accomplished by the application of sufficient pressure.

By the choice of suitable contact materials, bonded contacts can be made either ohmic or rectifying. For example, diodes have been made by bonding an aluminum wire to N-type silicon. Aluminum is an acceptor impurity so the regrowth around the wire is P-type and a PN junction is formed. The same contact applied to P-type silicon would provide an ohmic PP^{+} junction. Similarly, if a copper wire containing phosphorous were bonded to P-type germanium, a rectifying contact would be produced. Phosphorous would dominate in the regrowth since it has a lower melting point, higher segregation coefficient, and higher maximum solubility than copper. Hence, the regrowth will be N-type.

Bonded contacts are frequently employed as the rectifying junction of medium power,diode rectifiers. The method is less complicated than alloying, particularly for silicon,because wetting of the contact area is difficult when conventional alloying techniques are used.

Another important application of bonding is the making of small area ohmic contacts to transistors. For instance, bonding is used to attach the base lead to grown junction transistors and, as shown in the previous section, to attach the emitter and base leads to double diffused transistors. Compression bonding is usually employed for these small area contacts. Sufficient pressure is applied to the lead to melt the semiconductor. Heat is not applied from an external soruce.

Vapor Deposition. A metal can be evaporated in an evacuated chamber and,deposited on a semiconductor. The presence of surface oxides on the semiconductor degrades the performance of these contacts, but an alloy junction (rectifying or ohmic) can be formed by heating the deposited contact. Therefore, this method is essentially on extention of alloying techniques.

One advantage of vapor deposition is that intricate contact pattern can be reproduced on a semiconductor surface. A contact is deposited over the entire surface. The desired pattern is then registered on the surface using photographic masking techniques. Subsequent etching will remove undersired portions of the contact. Very small patterns (down to the resolution power of a microscope) can be successfully reproduced using this system. The emitter and base contacts of diffused base mesa transistors are applied by vapor deposition.

Electroplating. Good quality rectifying contacts can be electroplated directly onto a semiconductor. The success of this process depends on the cleaning of the crystal surface prior to plating by electrolytic etching. The contacts are then electrodeposited without removing the crystal from the electrolyte, which prevents possible contamination.

The surface barrier transistor is an unusual design that uses electrodeposited contacts for the emitter and collector. The manufacturing process involves the electrolytic etching of depressions into opposite sides of a N-type germanium wafer. This is accomplished by directing small jets of an electrolyte against opposite faces of the wafer and passing a current from the electrolyte into the germanium. The system is shown in figure 5.17. After etching, metalic ions in the electrolyte are plated into the wells by reversing the etching current.

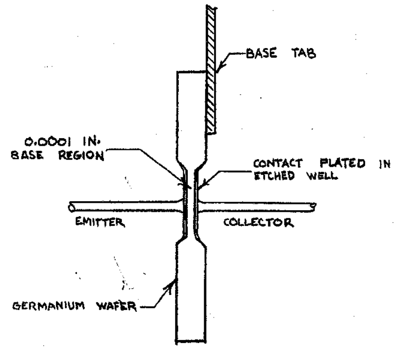


Figure 5.16. Construction of the Surface-Barrier Transistor.

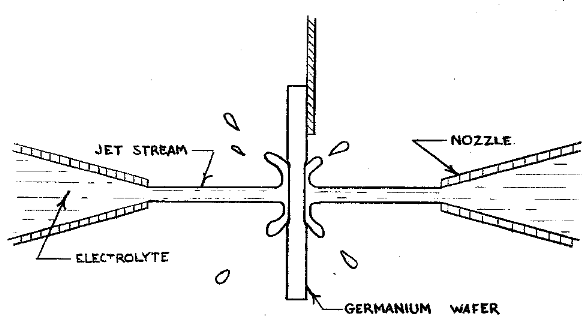


Figure 5.17. Set up for Etching and Plating Operations on a Surface-Barrier Transistor.

An automatic control over base width is inherent in the etching process. During etching, depletion regions are formed near the surface of the crystal in contact with the solution. When the depletion regions on each side of the crystal meet, etching will stop and current will drop off. Thus, flat bottomed wells separated by about 0.0001 inch are formed.

SURFACE TREATMENT AND ENCAPSULATION

The condition of the crystal surface near a junction of a diode or transistor greatly affects the performance of the device. It is desirable to have the surface free from contaminants, defects, and mechanical strain. This is necessary to reduce surface recombination of minority carriers and to eliminate conducting paths across the junction.

The usual method used to clean the surface is chemical or electrolytic etching. Etching removes the damaged portions of the crystal and produces a clean, mirror-like finish.

Even when the surface is properly cleaned, it has been found that the absorption of water vapor and air affects device performance in an unpredictable fashion. These contaminants can cause large changes in characteristics, particularly the reverse current of a back-biased junction or the current gain of a transistor. Therefore, it is usually necessary to outgas the completed units in a vacuum. This should be done at a high temperature, but the temperature is limited by the device. After outgassing, the diodes and transistors must somehow be protected from the atmosphere. The most effective protection is to hermetically seal the device in an inert atmosphere. Coating with some protective compound such as silicone grease and encapsulation in plastic has been shown effective for noncritical applications.

The encapsulation used must also provide for the dissipation of heat from the transistor. In low power units no special precautions are required since adequate cooling can be provided by conduction through the leads. Medium power transistors should be mounted directly to a metal case. This is also true for high power transistors except that the case must be designed so that it can be fastened to a heat sink.

DEPARTMENT OF WEAPONS TRAINING
LOWRY AIR FORCE BASE
COLORADO

TRANSISTORS

IN SERVICE TRAINING COURSE

CHAPTER 6

TRANSISTOR CIRCUITS

17 October 1961

FOR INSTRUCTIONAL PURPOSES ONLY

CHAPTER 6

TRANSISTOR CIRCUITS

Author: A1C Robert J. Widlar

Abstract: The basic differences between common emitter, common base, and common collector amplifiers are explained and justified. Biasing circuits are given, and the effects of temperature on transistor circuits are shown. Temperature compensation and bias stabilization techniques are covered. Small-signal, direct-coupled, and power amplifiers are described; practical circuits and coupling methods are given. Furthermore, the special problems associated with high-frequency operation are considered, and typical tuned and untuned (video) amplifiers are discussed. Both sine-wave and non-sinusoidal oscillators are treated. The latter category includes blocking oscillators, sawtooth generators, and multivibrators. Finally, the operation of the transistor as a switching device, rather than as an amplifier, is analyzed.

INTRODUCTION

The previous chapters have been devoted to the physics of semiconductor devices. This chapter will discuss their use in practical circuits. The transistor will receive the greatest attention because of its obvious importance. Nonetheless, other devices will be mentioned in conjunction with transistor circuits.

Frequently in the analysis of transistor circuits, transistors are compared with vacuum tubes. Although the transistor, like the vacuum tube, is a three terminal amplifying device, it does have many peculiarities. Among these are:

(1) The transistor is not strictly a voltage amplifier. The input impedance is often quite low so an appreciable current must be supplied by the driving source.

(2) The voltages and impedances encountered with transistors are considerably lower.

(3) The transistor has a far greater temperature sensitivity than the vacuum tube. Hence, a considerable portion of its circuitry may be devoted to temperature compensation.

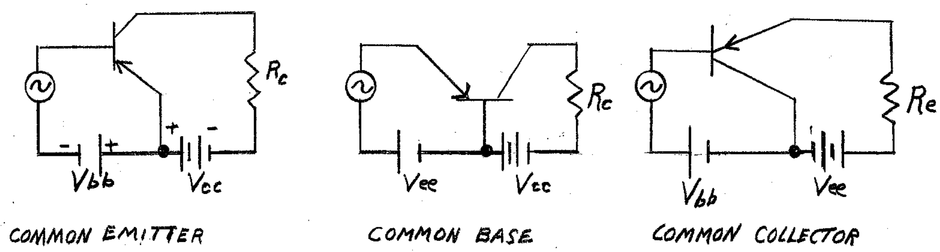
(4) The transistor has more internal feedback which means that the input characteristics are affected by the load impedance and the output characteristics are affected by generator impedance.

These and other differences limit the value of vacuum tube-transistor analogies. Therefore, they will be avoided in this chapter.

BASIC AMPLIFIER CIRCUITS

Before going into transistor circuitry, the basic amplifier circuits will be discussed. These are the common base, common emitter, and common collector configurations shown in Figure 6.1. In each of these circuits biasing potentials are provided to reverse bias the collector-base junction (about 20 volts) and forward bias the emitter-base junction (about 0.5 volt). In all cases amplification takes place because the input signal varies the forward bias on the emitter-base junction which causes variations in the collector current.

<u>Common Emitter</u>. Perhaps the most straightforward of these circuits is the common emitter amplifier. The input signal is applied between the emitter and base; it varies the forward bias on the emitter junction. The resulting variations in collector current cause an output voltage to be developed across the load



Circuit	Input Resistance	Output Resistance	Voltage Gain	Current Gain
Common Emitter	1000	50k	200	50
Common Base	100	2 meg	200	0.98
Common Collector	50k	1000	0.99	50

Figure 6.1. Basic Amplifier Circuits and Approximate Characteristics.

resistance. Only the base current flows in the input circuit, and it is small compared to the collector current so a current gain will be realized. The voltage gain of the common emitter amplifier can be high because small changes of input voltage across the forward biased emitter junction can produce large changes in collector current. The changing collector current can then produce a comparatively large voltage swing across the load if its impedance is sufficiently high. Current gains of 50 and voltage gains of 200 are not uncommon in practice.

As can be seen from Figure 6.2, the collector current of a common emitter stage is primarily determined by the base current and not too greatly affected by collector voltage. The collector current is affected somewhat because increasing collector voltage widens the collector depletion region and decreases base width. This increases the current gain and, consequently, the collector current. This effect is not too great so the output resistance ($R_o = \frac{\Delta V_c}{\Delta I_c}$) of a common emitter amplifier is moderately high. On the other hand, the base current changes rather rapidly with base voltage. Hence, the input resistance ($R_{IN} = \frac{\Delta V_b}{\Delta I_b}$) is moderately low.

Common Base. In the common base configuration the input signal is applied to the emitter and the output is taken from the collector. The base is the common terminal. An a.c. signal applied to the input will vary the emitter-base voltage and, therefore, the collector current. The changes in collector current will be slightly less than the corresponding changes in emitter current - differing, of course, by the base current. Hence, the current gain will be slightly less than one. If the changing collector current passes through a high impedance load, relatively large variations in collector voltage will result. It follows, then, that the voltage gain of the common base amplifier can be large. Current gains of 0.98 and voltage gains greater than 200 are not unusual with the configuration.

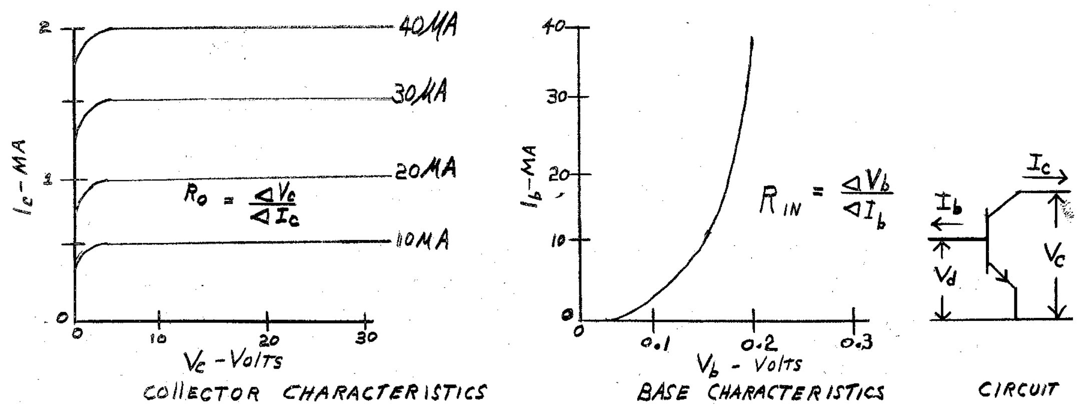


Figure 6.2. Common Emitter

The input resistance ($R_{IN} = \frac{\Delta V_c}{\Delta I_c}$) of the common base amplifier is considerably less than that of a common emitter amplifier as can be seen from Figure 6.3. The change in emitter current for a given change of emitter-base voltage is greater than the change in base current. The emitter current is approximately equal to the base current times the common emitter current gain so the input resistance is about equal to the common emitter input resistance divided by this current gain.

The effect of base width on common base current gain is minimal. Although significant changes in common emitter current gain are produced by varying collector voltage, they are reflected as comparatively small changes in common base current gain. Since the collector current is affected less by collector voltage in the common base configuration, the output resistance ($R_0 = \frac{\Delta V_c}{\Delta I_c}$) is much higher. It turns out that the common base output resistance is about equal to the common emitter output resistance times the common emitter current gain. In other words, since the effect of collector voltage on the common base characteristics is so small, the output resistance is nearly equal to the resistance of the reverse biased collector junction alone.

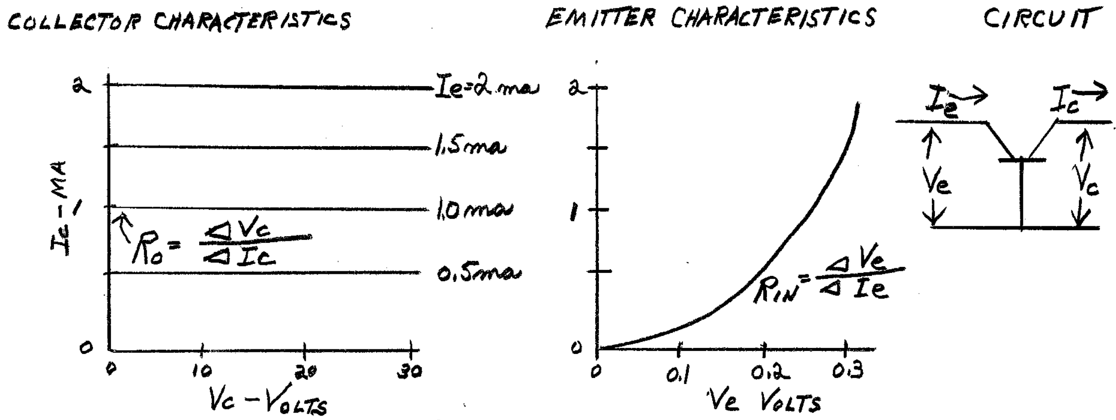


Figure 6.3. Common Base

<u>Common Collector</u>. With the common collector configuration, the input signal is applied to the base, and the load is connected to the emitter. As before, the amplifying action takes place when the input signal varies the forward bias on the emitter-base junction. A changing base current will vary the emitter current and produce an a.c. voltage across the load. The emitter current will be considerably greater than the base current so the circuit will have a rather high current gain - slightly greater than that of a common emitter amplifier. The voltage gain, however, will be somewhat less than unity.

Considering the circuit shown in Figure 6.4, if the base current is increased by an increased input voltage, the emitter current will also rise. This will raise the voltage drop across the load resistance. In addition, the larger base current will also increase the emitter-base voltage. Now from Figure 6.4 it can be seen that the increase in input voltage required to produce this effect will be equal to the voltage change across the load plus the increase in emitter-base voltage. Hence, the change in output voltage will be less than the change in input voltage, differing by the change in emitter-base voltage. Since the emitter-base voltage will be small in comparison to the output voltage, the voltage gain will be only slightly less than one. Current gains of 50 and voltage gains of 0.99 are practical with this circuit.

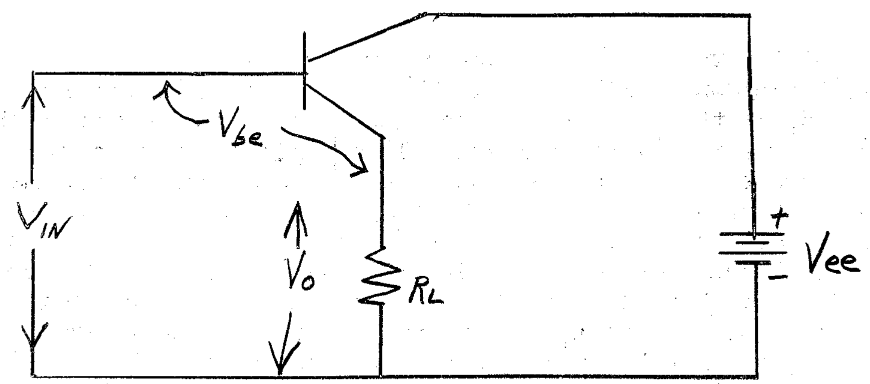


Figure 6.4. Common Collector Amplifier

The input resistance ($R_{IN} = \frac{\Delta V_b}{\Delta I_b}$) of the common collector amplifier is approximately equal to the product of the current gain and the load resistance. The output voltage is nearly equal to the input voltage, so for a given a.c. input voltage, the load current and, consequently, the base current will depend on the load resistance. As the load resistance is decreased, the base current must increase. For very low load resistances the input resistance will be equal to that of a common emitter amplifier, while for high load resistances the input impedance may be several megohms. The latter case is illustrated in Figure 6.5.

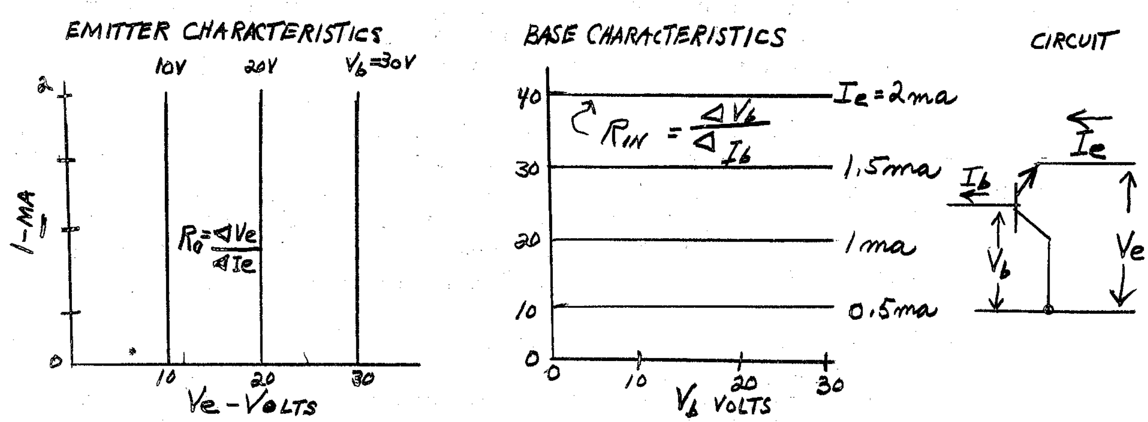


Figure 6.5. Common Collector

Similarly, the output resistance ($R_o = \frac{\Delta V_e}{\Delta I_e}$) is about equal to the source resistance divided by the current gain. If the base is fed by a constant voltage source (low impedance) a change in emitter voltage will be felt entirely across the emitter base junction. Hence, small changes in emitter voltage will produce large changes in emitter current so the output impedance will be low. This is shown in Figure 6.5. On the other hand, if the base is fed from a constant current source (high impedance) changes in emitter voltage will not produce any great changes in emitter current. In this case, the output resistance is equal to that of a common emitter amplifier; and the output characteristic curves will be similar to those of a common emitter stage. For intermediate values of source resistance, a change in emitter voltage will be felt almost undiminished on the base through the forward biased emitter junction. This change in base voltage will vary the base current in inverse proportion to the generator resistance. The resulting variations in emitter current will, in turn, be proportional to the current gain. Further analysis will show that the output resistance is roughly equal to the source resistance divided by the current gain.

This interaction between the input and output circuits is not confined to the common collector configuration. It is present to a lesser degree in both the common emitter and common base circuits. This phenomenon is the result of internal feedback within the transistor. It will be discussed further in a later section.

Conclusions. The common emitter amplifier exhibits both a voltage and a current gain; it has the highest power gain of all three configurations. As a result, the common emitter amplifier is used where high gain per stage is required. In addition, the moderately low input impedance and the moderately high output impedance of the common emitter amplifier do not create too much of an impedance mismatch when several stages are cascaded without impedance matching devices. These characteristics favor the use of the common emitter in most general purpose applications.

The power gain of the common base amplifier is between that of the common emitter and common collector amplifiers. The current gain of this type approaches unity, but the voltage gain can be quite high. The low input impedance and high output impedance of a common base stage necessitates impedance matching if the stages are to be cascaded but are sometimes useful in special cases.

Near the maximum operating frequencies of a transistor, the gain of the common base amplifier is about equal to that of the common emitter amplifier. However, the common base stage has the additional advantage that neutralization is not required. Consequently, the common base configuration is most frequently used in high frequency amplifier circuits.

The common collector amplifier has the lowest power gain of the three types. The current gain is high, but the voltage gain is less than one. The common collector amplifier is essentially a common emitter stage with 100 percent negative voltage feedback. Therefore, the signal distortion will be small even at high levels if it is driven by a low impedance source.

The common collector configuration finds most frequent application as a high impedance input stage, a driver where low output impedance is required, and as a power amplifier where large voltage swings are needed with a minimum of distortion.

BIASING METHODS

The biasing of transistor circuits is not as simple as that of vacuum tube circuits. In most cases, some form of bias stabilization or temperature compensation is required. The problem with transistors is that the collector current tends to increase rapidly with temperature. There are two principle causes of this temperature instability: the increase of the reverse saturation current of the collector junction with temperature and the decrease in emitter junction impedance with increasing temperature. The current gain frequently increases with temperature and adds to the problem.

Reverse saturation current is the current that flows across the reverse biased collector junction. It is produced by thermally generated minority carriers that are swept across the junction so it will increase with temperature. Figure 6.6 shows the increase in reverse saturation current with temperature for typical germanium and silicon transistors.

The saturation current is only troublesome at high temperatures. For example, if the germanium transistor in Figure 6.6 is biased to a collector current of 1 ma, it can be seen that the saturation current will contribute little to the total collector current until a temperature of about 50°C is reached. However, at higher temperatures the saturation current becomes an appreciable portion of the collector current. When this happens, the bias must be decreased with increasing temperature in order to maintain the collector current constant. This compensation is effective until the saturation current becomes equal to or greater than the quiescent collector current. At these higher temperatures, the input circuit loses control of the collector current since the saturation current is independent of the emitter junction bias. Therefore the transistor ceases to function as an amplifier. This places an upper limit of about 90°C on germanium transistors. The lower saturation current of silicon transistors permits operation up to about 175°C.

With the common emitter configuration, the effect of the reverse saturation current is even more serious. If the base is open circuited, the minority carriers swept across the collector junction will cause a charge unbalance in the base and will forward bias the emitter junction. Hence, additional carriers will be injected into the base and will contribute to the collector current. The collector current will then be

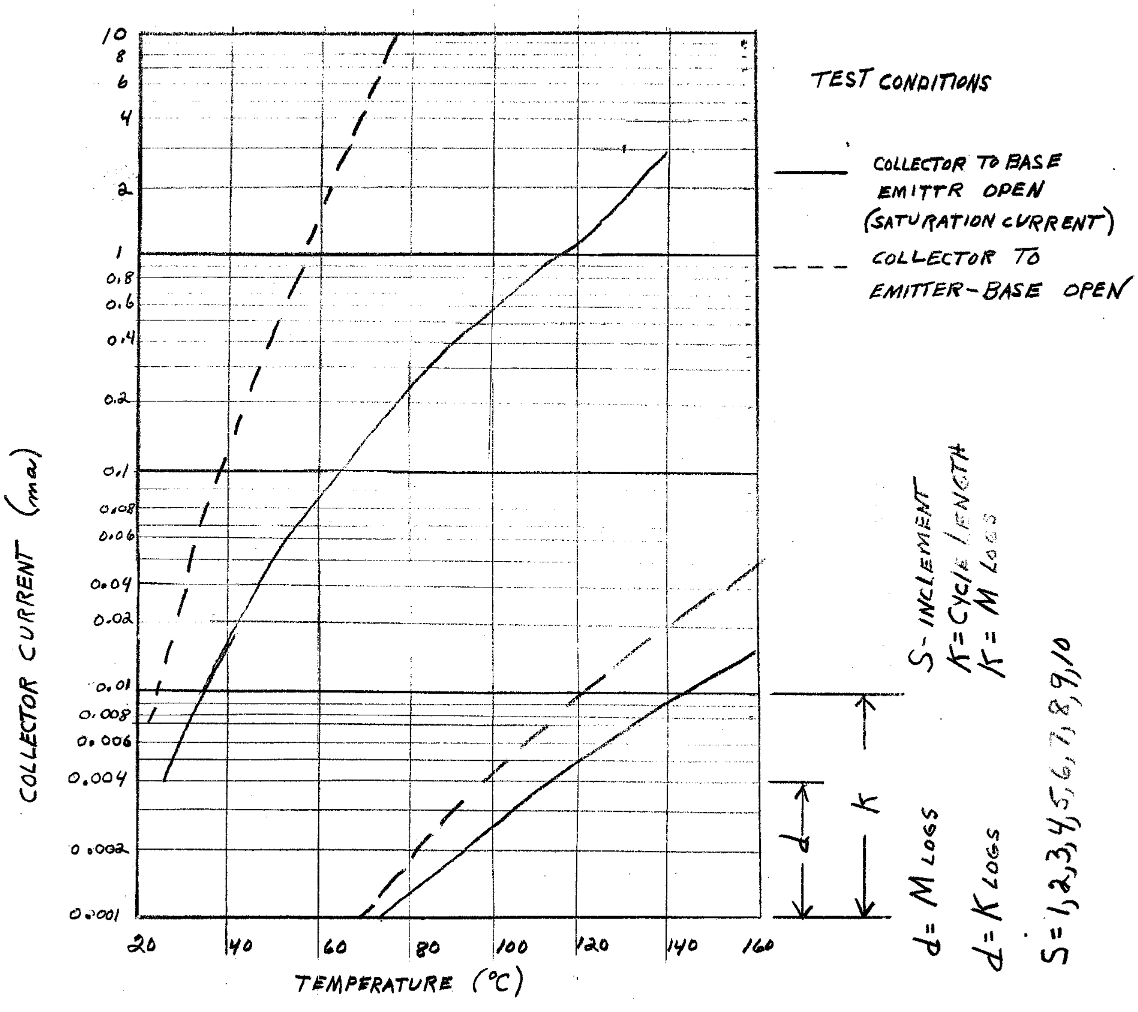


Figure 6.6. Plot of the Collector Cutoff Current for Silicon and Germanium Transistors as a Function of Junction Temperature.

equal to the saturation current times the current gain. Since the current gain is quite small at very low collector currents, the collector cutoff current will be slightly greater than the saturation current at low temperatures. As temperature increases, the saturation current will increase sharply, and the current gain will increase with the collector current. Therefore, the collector cutoff current will increase more rapidly than the saturation current alone. This is shown in Figure 6.6.

If the base is shorted to the emitter, the collector cutoff current will be only slightly larger than the reverse saturation current. Most of the excess charge in the base will then be drained off through the base terminal so the forward bias on the emitter junction will be reduced. This will significantly reduce the collector current.

Constant Voltage Bias. Figure 6.7 illustrates a possible biasing circuit. A constant d.c. bias voltage is applied to the emitter-base junction. This method has the advantage that there is a low d.c. resistance between the emitter and the base. Therefore, the excess charge produced in the base by the reverse saturation current will be drained off and multiplication of the saturation current by the emitter junction will be minimized.

Nonetheless, this method is entirely unsatisfactory. The collector current will increase steadily with temperature over the entire operating range of the transistor.

The forward bias lowers the emitter junction barrier so that current carriers can diffuse across the junction. As temperature is increased, thermal motion is more rapid so a greater number of carriers can diffuse across the junction if the barrier height remains the same. The result is that the emitter current and, consequently, the collector current will increase steadily with temperature for a given emitter-base bias. This phenomenon manifests itself over the entire temperature range of the device. Hence, to maintain a constant collector current, the forward bias voltage must be decreased with increasing temperature. This is shown in Figure 6.7. It can be seen from the figure that the bias must be decreased linearly with increasing temperature until the reverse saturation current becomes appreciable. Then the bias must be decreased more rapidly to compensate for the increased saturation current in addition to the changing emitter junction resistance.

Temperature sensitive elements are available that will drop the biasing voltage as temperature increases. These will be discussed later.

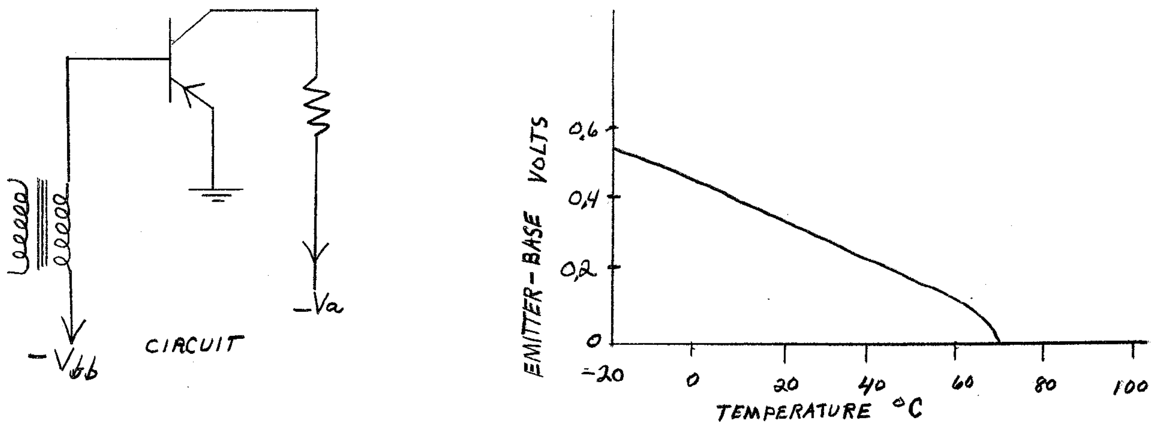


Figure 6.7. Emitter-Base Bias Voltage Required for Constant Collector Current with Increasing Temperature.

<u>Constant Current Bias</u>. A transistor can also be biased by supplying a constant current to the base. This could be accomplished by forward biasing the base through a large resistance from a high voltage (for example, the collector supply voltage). Such a circuit is shown in Figure 6.8. The emitter-base resistance of the transistor will be very much smaller than the biasing resistor so changes in the emitter-base resistance will have little effect on base current.

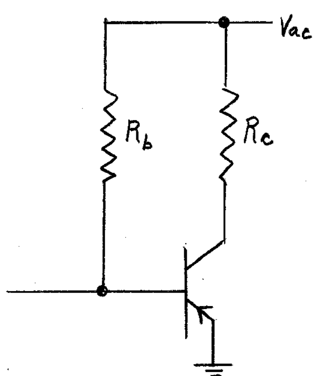
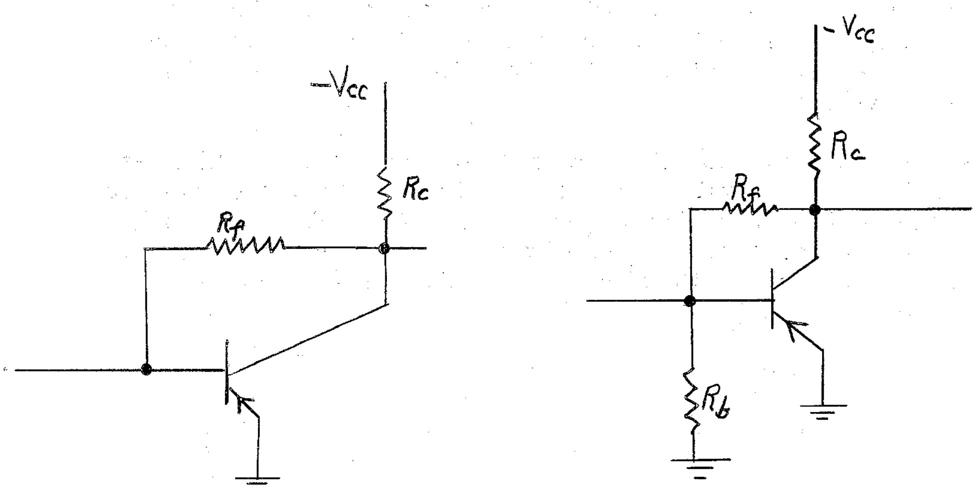


Figure 6.8. Constant Current Bias.

Although it is much better than constant voltage bias, constant current bias is still unsatisfactory for two reasons. First, the d.c. current gain of a transistor usually increases with temperature. Consequently, if the base current is constant, the collector current will increase with temperature. Second, the collector current will increase uncontrolled as the collector saturation current rises. Since the base current is constant, the saturation current must cross the emitter junction. As a result, the total increase in collector current produced by the saturation current of the collector junction will be equal to the saturation current times the current gain. This causes instability even at low temperatures.

Collector Feedback Bias. It should be obvious that some form of bias stabilization is required to keep the collector current constant over any appreciable temperature range. One method of bias stabilization is collector feedback bias. As shown in Figure 6.9a, the base current is supplied through a resistor connected to the collector. Therefore, if the collector current increases, the collector voltage will drop, reducing the bias current proportionally. This feedback action tends to maintain the collector current constant.

Another circuit employing collector feedback bias is shown in Figure 6.9b. This circuit has the advantage that the external emitter-base resistance is reduced. The saturation current of the collector junction will pass through the external resistance rather than cross the emitter junction. Thus, multiplication of the saturation current by the emitter junction is minimized. This resistance does, however, lower the input resistance of the amplifier which might be detrimental in some cases.



a.

Figure 6.9. Examples of Collector Feedback Biasing.

A characteristic common to both these circuits is that negative feedback is introduced by the collector-base resistance. This will lower the stage gain, but it will reduce distortion and stabilize the gain. The feedback can be eliminated, if desired, by tapping the feedback resistance at a suitable point and connecting a large capacitor between the tap and ground. This will prevent feedback of the a.c. signal but will not affect the d.c. feedback.

Collector feedback bias is only effective when there is a large d.c. resistance in the collector circuit. If the collector load were a transformer primary winding with a low d.c. resistance, increases in the collector current would not change the collector voltage so the stabilizing action could not take place.

Emitter Feedback Bias. Another method of bias stabilization is illustrated in Figure 6.10. This is emitter feedback bias. As shown in Figure 6.10a, the base is forward biased from a constant voltage source; a resistor is included in the emitter lead. The voltage developed across the emitter resistor opposes the forward bias applied to the base. Thus, if the emitter current increases, the voltage drop on the emitter resistor will increase thereby reducing the bias on the emitter-base junction. This action stabilizes the emitter current against changes. The collector current will also be stabilized since it is very nearly equal to the emitter current.

One of the advantages of emitter feedback bias is that the base is fed from a constant voltage source so very little of the collector saturation current will cross the emitter junction. In addition, the circuit can be used when the collector load has a low d.c. resistance since stabilization is provided by the emitter resistance.

It is desirable to make the emitter resistance as high as possible to realize stable operation. However, the voltage drop across the emitter resistance does reduce the effective supply voltage and causes a loss of supply voltage. For this reason, there are practical limitations on the size of the emitter resistance.

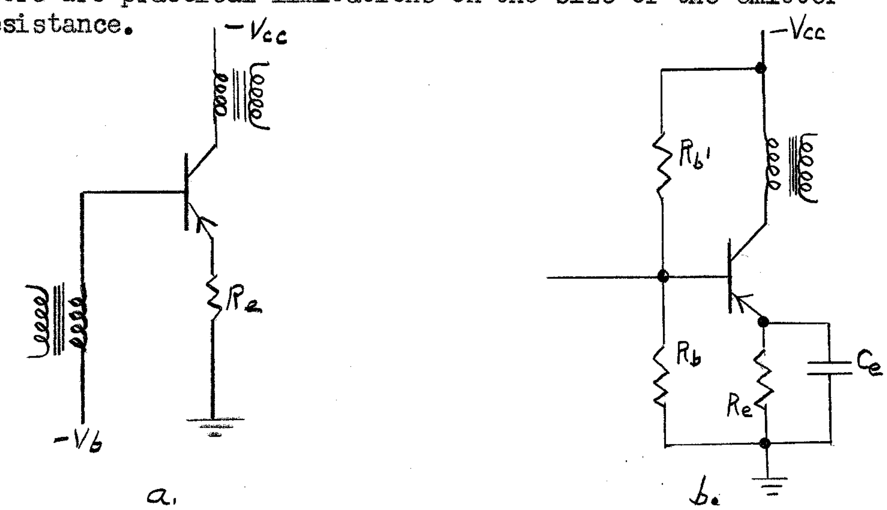


Figure 6.10. Examples of Emitter Feedback Bias.

The base bias voltage can also be supplied from the collector supply through a voltage divider. This is shown in Figure 6.10b. The resistance of this divider should be as low as possible - without excessive shunting of the input signal - to approximate a constant voltage source. If the base is fed from a high resistance divider (nearly constant current) the desirable effects of the emitter resistor will be nullified because the base current will remain nearly constant regardless of the emitter voltage drop.

Emitter feedback biasing will produce degeneration of the a.c. signal, unless the emitter resistance is by-passed as shown in Figure 6.10b. The by-pass capacitor prevents a.c. voltages from appearing across the emitter resistor.

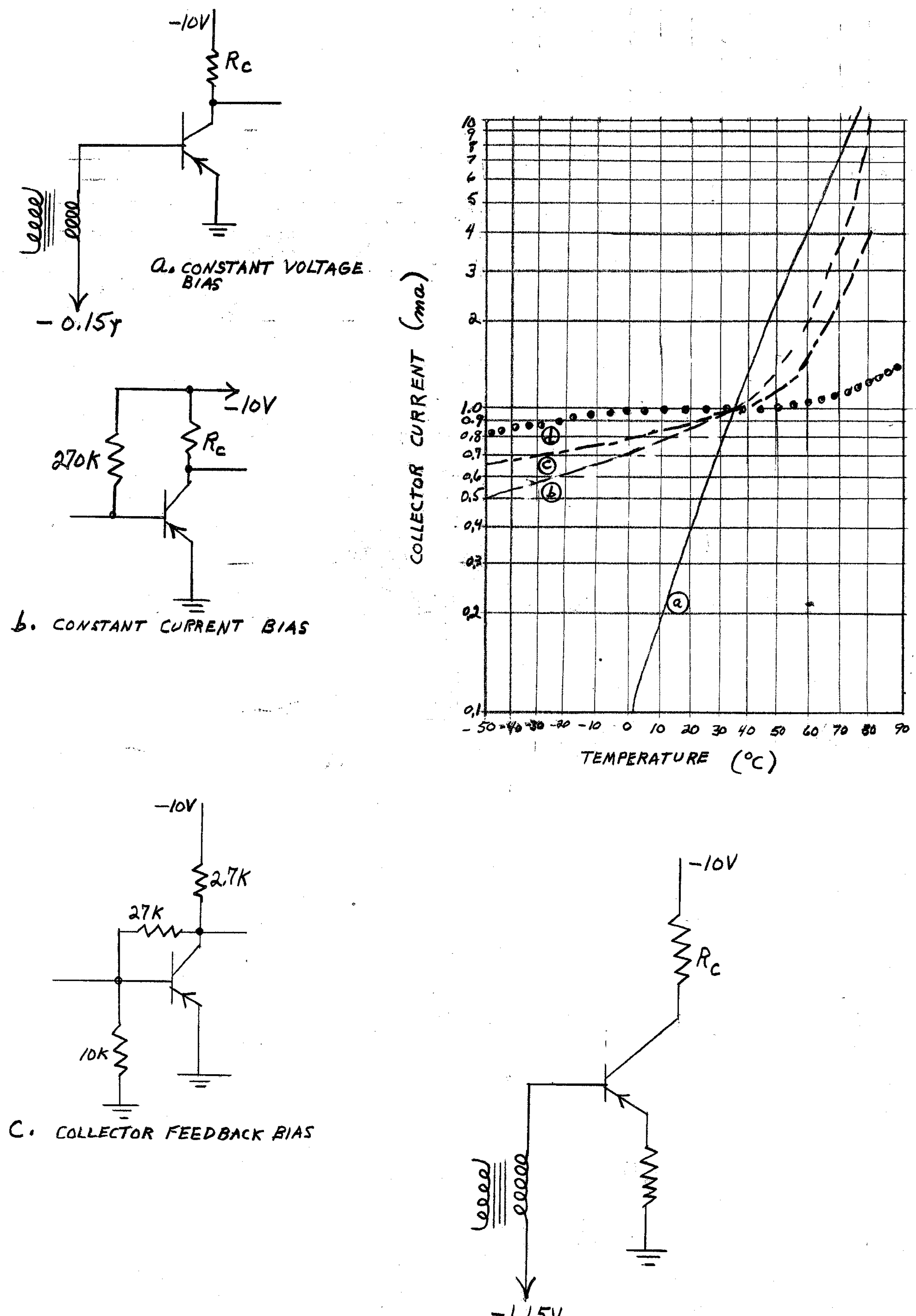
Conclusions. The biasing circuits discussed thus far are compared in Figure 6.11. The curves are for germanium transistors. In each case, the bias was adjusted to give 1 ma collector current at room temperature.

Constant voltage bias is unsatisfactory since the collector current is very unstable even over a restricted temperature range.

With constant current bias the collector current is reasonably constant at low temperatures, but rises rapidly at junction temperatures above 40°C. Nonetheless, even if a circuit is to be operated at a nearly constant temperature, constant current bias will not give completely satisfactory results. The collector current is directly proportional to the common emitter current gain, and this can vary greatly with age and from transistor to transistor. Thus, constant current bias is suitable only for experimental work.

Collector feedback bias provides adequate stabilization for most commercial applications, but its use is restricted to circuits where the collector load has a relatively high d.c. resistance. It cannot be used in transformer coupled stages. For this reason, collector feedback bias is usually limited to resistance coupled, audio-frequency amplifiers. With collector feedback bias, the collector current can be made nearly independent of the transistor characteristics. This is done by using a low resistance divider to feed the base from the collector.

Emitter feedback bias is about the best general-purpose biasing method. The stability of this method is evidenced in Figure 6.11. As with collector feedback bias, the collector current can be made nearly independent of transistor parameters by using a sufficiently high emitter resistance. This is highly desirable for interchangeability of transistors. Emitter feedback bias is not too frequently used in power amplifiers since the emitter resistance can cause an excessive loss of power. Other biasing techniques are generally used in power stages. These will be discussed in the next section.



d. Emitter Feedback Bias

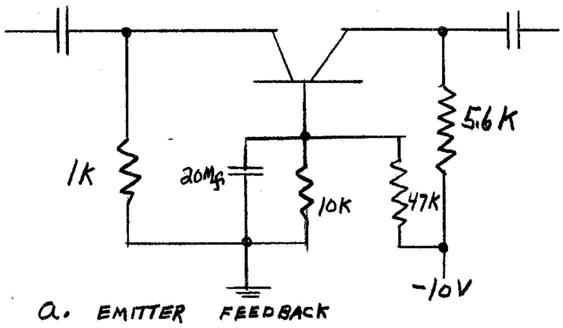
Figure 6.11. Variation of Collector Current with Temperature for Typical Germanium Transistor Using Different Biasing Circuits.

The biasing of common base and common collector stages is similar to that of common emitter stages. In Figure 6.12a, the biasing used is identical to the emitter feedback bias described previously. The base is fed from a voltage divider connected to the collector supply, and a resistor is included in the emitter lead. If the collector current should rise, the voltage drop across the emitter resistor will increase and lower the bias current. The only difference with this circuit is that the input signal is fed to the emitter, and the base is by-passed to ground.

Another common base circuit is given in Figure 6.12b. Here, the base is fed from a voltage divider connected to the collector. Hence, when the collector current increases, the collector voltage will drop and will lower the forward bias on the base. This is identical to the collector feedback bias circuit for a common emitter stage.

Common collector biasing circuits are given in Figure 6.13. The arrangement in Figure 6.13a is the same as the collector feedback circuit given in Figure 6.9a. The circuit in Figure 6.13b employs a combination of collector and emitter feedback. When the emitter current increases, the increased voltage drop across the 4.7K resistor lowers the voltage supplied to the voltage divider supplying the base as with collector feedback; and the increased voltage drop across the 1K resistor opposes the forward bias as with emitter feedback.

Thermistor Compensation. The thermistor is a resistor with a large, negative temperature coefficient. That is, its resistance will decrease quite noticeably as temperature increases. If a thermistor is used as part of a voltage divider supplying bias to a transistor, as shown in Figure 6.14a, the bias will be reduced as temperature increases. This compensates for the tendency of the collector current to increase with temperature. It can be seen from the plot in Figure 6.11 that this compensation is not exact. The temperature coefficient of the thermistor is not matched to that of the transistor. Hence, over part of the temperature range the compensation is too great while over other parts it is not enough. If the thermistor were not shunted with a resistor (Figure 6.14a), the compensation at low temperatures would be far too great. At low temperatures when the resistance of the thermistor is high, R_2 will determine the voltage division ratio of the bias network and there will be practically no compensation. As temperature increases, the resistance of the thermistor becomes comparable to that of R_2 and the bias voltage decreases with increasing temperature. At medium temperatures this compensation is too great until the collector saturation current becomes appreciable. Finally, the rapidly increasing saturation current causes a net increase of collector current in spite of the thermistor.



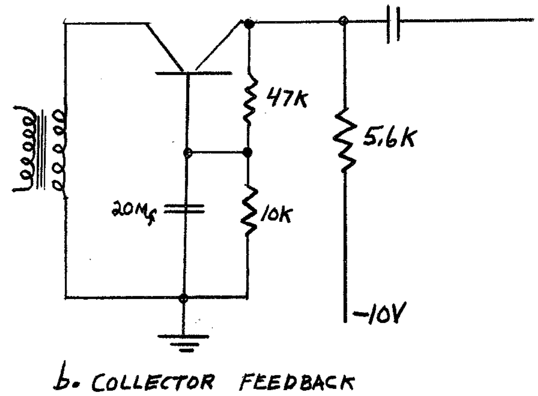
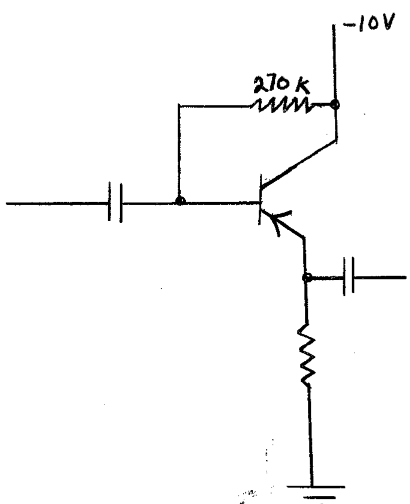


Figure 6.12. Common Base Biasing Circuits



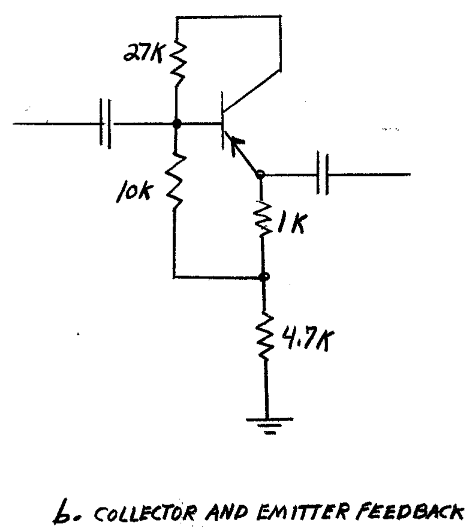


Figure 6.13. Common Collector Biasing Circuits.

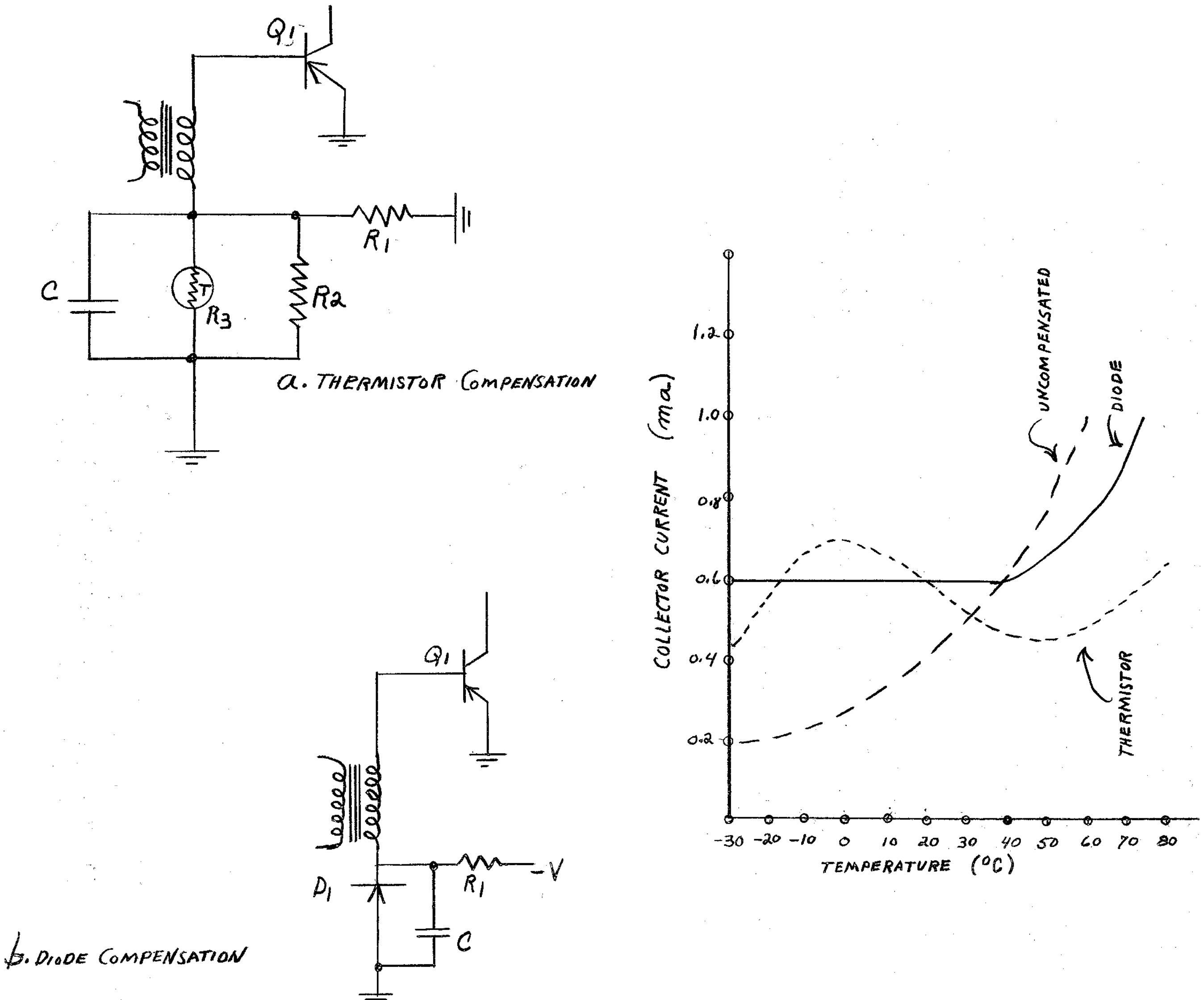


Figure 6.14. Negative Temperature Coefficient Thermistor and Semiconductor Diode Compensation.

Recently, positive temperature coefficient, silicon thermistors have been made with thermal characteristics closely resembling those of a silicon transistor. When these elements are used in conjunction with feedback circuits, almost exact compensation is realized. An example is shown in Figure 6.15.

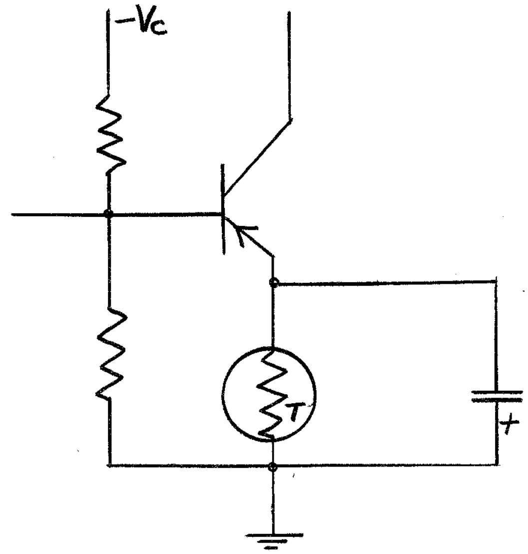


Figure 6.15. Example of Temperature Correction Using a Positive Temperature Coefficient Thermistor in Conjunction with Feedback.

Diode Compensation. An ordinary junction diode can be used to compensate transistor bias. In Figure 6.14b, the transistor bias voltage is developed across the forward-biased diode (D_1). The forward resistance of the diode varies with temperature in the same manner as does the impedance of the emitter-base junction of the transistor. If the current through the diode is large compared to the base current of the transistor, the diode alone will determine the bias voltage. This system will provide almost perfect compensation as long as the saturation current is small. This is evidenced by the graph. If the transistor is to be operated at higher temperatures, some auxiliary means of compensation will be necessary.

When temperature compensation is used, it is important to locate the temperature sensitive element close to the compensated transistor. This is necessary because the transistor may become noticeably hotter than its surroundings, particularly if power is being dissipated in the unit.

SMALL SIGNAL AUDIO AMPLIFIERS

The transistor is basically a non-linear device: the input and output impedances, as well as the current gain, vary with collector current. However, when the signal level is low, these changes will be small so some definite magnitude can be assigned to the impedances and the gain. Then, the transistor can be considered to be a linear amplifier. In this section it will be assumed that this condition is satisfied. Furthermore, only low-frequency operation will be considered so the deleterious effects of junction capacitance and transit time can be neglected.

Single-stage amplifiers -- common emitter, common base, and common collector -- were discussed previously. This section will be primarily concerned with methods of coupling, or cascading, these basic amplifiers to realize higher gains than might be possible with a single stage.

Although specifying the voltage and current gain of an amplifier is frequently useful, the determination of these values becomes somewhat complex with cascaded amplifiers. These two terms can be combined into one -- the power gain which is the product of the voltage and current gains. The power gain gives an absolute measure of amplifier performance. The input to an amplifier, be it from a high-impedance crystal microphone or a low-impedance pickup, is at a certain power level. This signal must be amplified to supply power to a load whether it is a set of headphones or a large motor.

From the preceding discussion it follows that it is desirable to produce maximum power gain per stage, the voltage or current gain alone being of minor significance. To do this, it is necessary to match impedances through the amplifier. Any mismatch will result in a lower gain per stage.

Transformer Coupling. From an electrical viewpoint, the simplest way to couple two stages is with a transformer. The output of one stage is fed to the input of the following, as shown in Figure 6.16. The transformer provides isolation of d.c. voltages and impedance matching.

Transformer coupling has the advantage that very little d.c. power is wasted: the d.c. resistance of the transformer primary is small compared to the load impedance. As compared with resistive coupling, equal results can be obtained with half the supply voltage. Nonetheless, a transformer does have certain disadvantages: for example, comparatively high cost, limited bandwidth, and large phase shifts at high frequencies. Size is not a particular disadvantage because miniature transformers can be fabricated to operate at the low power levels of transistors.

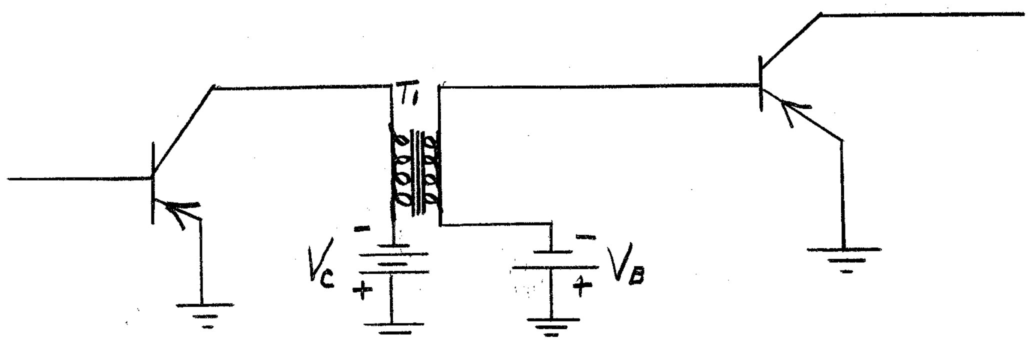


Figure 6.16. Example of Transformer Coupling.

A two-stage, transformer-coupled audio amplifier is shown in Figure 6.17. Input and output transformers, as well as an inter-stage transformer, are used. The input transformer must match the source to the 1000 impedance (1000-Ω) input of the common emitter amplifier. The interstage transformer must match the low impedance input of the second stage to the medium output impedance (15,000-Ω) of the first stage. The third transformer matches the load to the second transistor output.

It should be noted that maximum power gain for each stage can be realized because each transistor can be made to work into an optimum load impedance by appropriate adjustment of the transformer turns ratio. For this reason, the common emitter configuration is most frequently used because it is capable of a higher gain per stage since it has both current and voltage gain.

In Figure 6.17 two methods of bias feed are shown. Both stages use emitter feedback resistors, and the base bias is supplied from a voltage divider. In the first stage the bias is applied directly to the transistor; the transformer is connected to the base <u>via</u> a coupling capacitor which prevents shunting the d.c. bias. In the second stage the transformer is connected directly to the base, and the bias voltage is fed through the transformer primary from the voltage divider. A bypass capacitor is required across the voltage divider to prevent loss of the input signal across the divider resistance. Generally speaking, the latter method is more satisfactory because the voltage divider resistance does not shunt the input signal; however, the two circuits are equivalent as far as biasing is concerned.

If the emitter resistance is not bypassed, the output current flowing through the emitter would produce a voltage across this

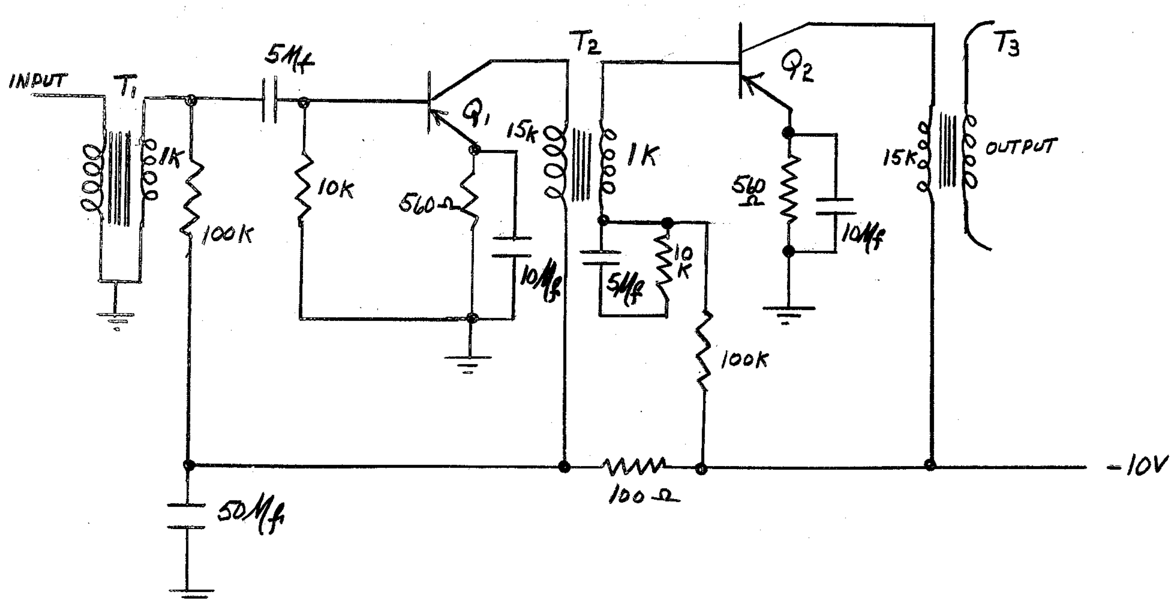


Figure 6.17. Two-Stage, Transformer Coupled Amplifier Illustrating Methods of Bias Feed.

resistance which would oppose the input signal being applied between the base and emitter. Also, the signal developed across the resistor would not appear in the output, giving an additional loss.

Because of the low d.c. resistance of the transformer primary it is necessary to employ effective bias stabilization with transformer coupled amplifiers. The increased collector current caused by increased temperature will not materially reduce d.c. collector voltage, so the power dissipation of the transistor will increase as the square of the collector current. The increased power dissipation will increase heating and collector current, causing ultimate destruction of the transistor if precautions are not taken.

A decoupling network is shown in the collector supply between the two stages. Although this is not necessary for a two-stage amplifier it is usually required when there are three or more stages. This network attenuates any low-frequency being fed back through the collector supply. Lack of adequate decoupling usually results in instability and low-frequency oscillations (motor boating).

R-C COUPLED AMPLIFIERS

Another method of cascading several stages is resistance-capacitance coupling. This scheme is illustrated in Figure 6.18. Direct current voltage is fed to the collector of the first stage (Q_1) through a resistor (R_c). Variations in the collector current of Q_1 develop a signal voltage across the collector resistor. This signal is coupled to the base of the following stage (Q_2) through a capacitor (C_c), and bias is supplied to this second stage through the resistor R_b.

It should be recognized that the purpose of the coupling network is to deliver the signal power, generated in the collector circuit of Q_1, to the base of Q_2, therefore it is necessary that the shunt resistance of R_c and R_b be high compared to the input impedance of Q_2 to prevent loss of power in these elements. Also, the capacitance reactance of C_c should be small compared to the input impedance of Q_2 at the lowest frequency to be amplified.

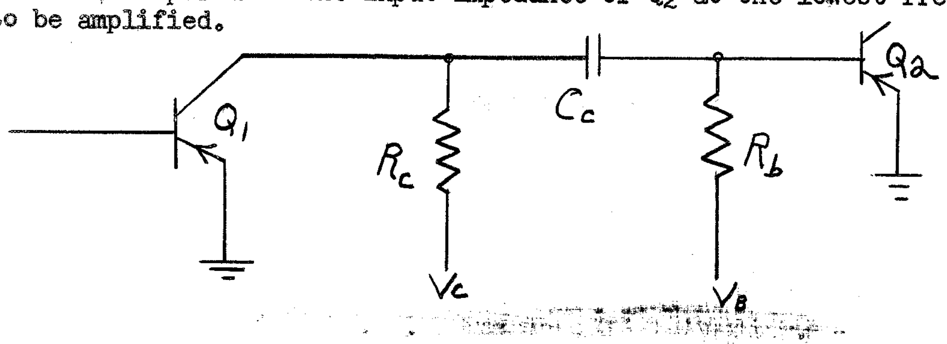


Figure 6.18. Example of R-C Coupling.

There is a limit on the maximum value of collector resistance (R_c). Because the d.c. power must be supplied to Q_1 through this element, too high a resistance will limit collector current and, therefore, the maximum current swing of the transistor.

Figure 6.19 illustrates some of the limitations of the r-c coupled amplifier. The loadlines of both an r-c coupled amplifier and a transformer coupled amplifier, both operating with a collector supply voltage of 5v and quiescent current of 0.3 ma, are plotted. Referring to the loadlines of the r-c amplifier, the d.c. loadline is determined by the collector resistor. This loadline determines the static conditions of the transistor as all possible combinations of static collector voltage and current lie on this line. The impedance presented to the collector is considerably lower for an a.c. signal. The a.c. load impedance consists of the parallel combination of the collector resistor and the base resistor and input impedance of the following stage. This results in a separate, lower-impedance, a.c. load line. This load line must pass through the operating point as shown in the figure. It can be seen that the voltage swing is restricted, but this in itself is not too serious because the voltage swing requirements for the input of the following transistor are small.

The bias stabilization requirements of an r.c. coupled amplifier are not as stringent as with transformer coupling because of the comparatively large d.c. resistance in the collector circuit. Thermally generated increases in collector current will also decrease collector voltage; hence, the collector dissipation will not increase too rapidly with collector current. In fact, if the collector resistance is sufficiently high, the collector dissipation will decrease with increasing d.c. collector current. All this aside, thermally induced changes of collector current will shift the operating point which is undesirable so some degree of stabilization is still required.

An example of a two-stage, r-c coupled amplifier is given in Figure 6.20. Collector feedback bias stabilization is employed on the first stage. The voltage divider supplying bias to the base of the first transistor is fed from the transistor collector, so an increased collector current will lower the forward bias by decreasing the collector voltage. The second stage uses emitter feedback bias stabilization. A voltage divider from the negative collector supply forward biases the second transistor, and a resistor in the emitter circuit provides bias stabilization by decreasing base to emitter voltage when the collector current increases. The emitter resistor is bypassed with a 10μf capacitor to prevent degeneration at signal frequencies.

The input signal is coupled into the amplifier through a 5μf capacitor, amplified by the first stage, and coupled into the second stage through another 5μf capacitor. This capacitor couples the

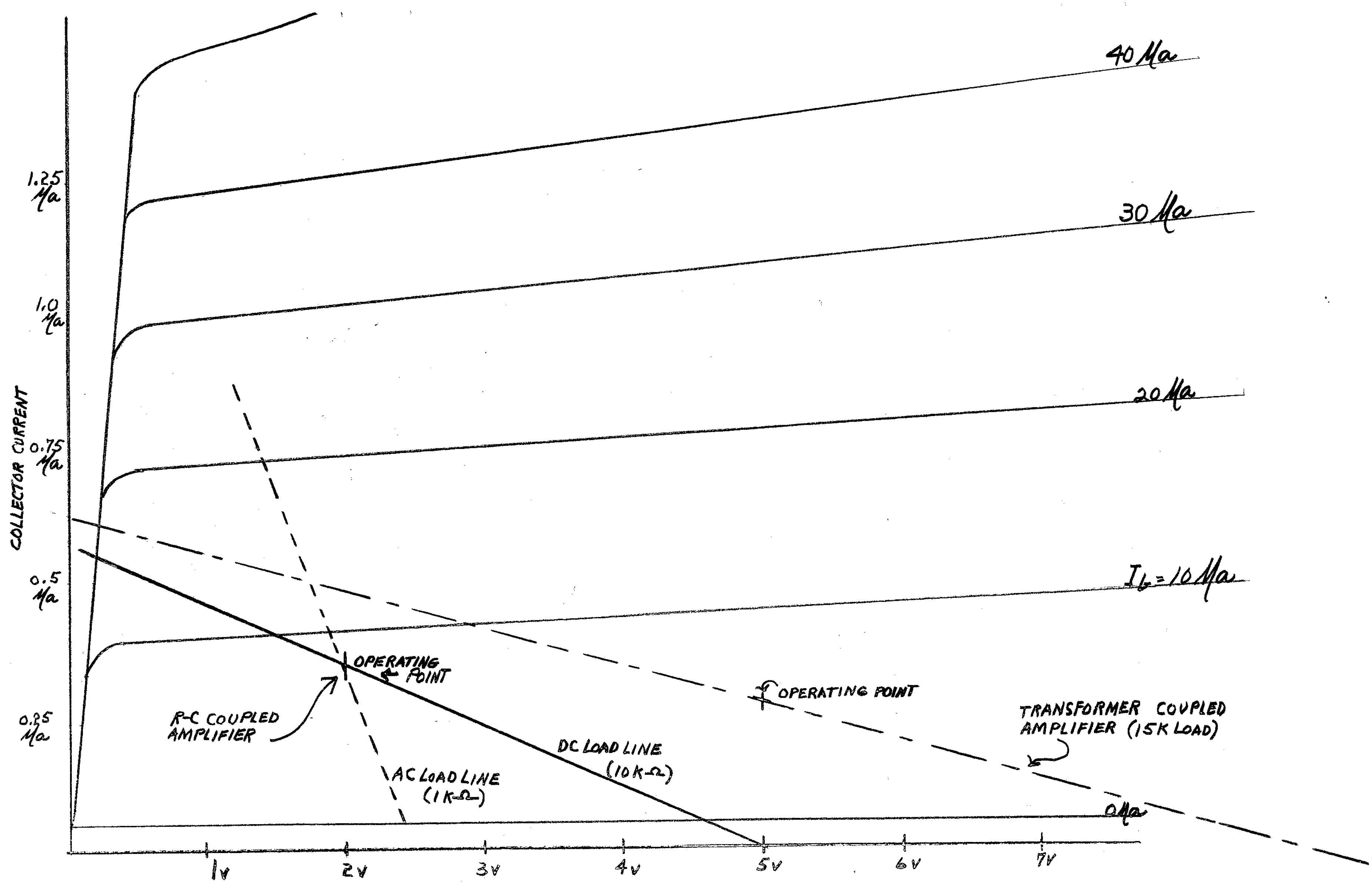


Figure 6.19. Comparison of Transformer and R-C Coupled Amplifiers.

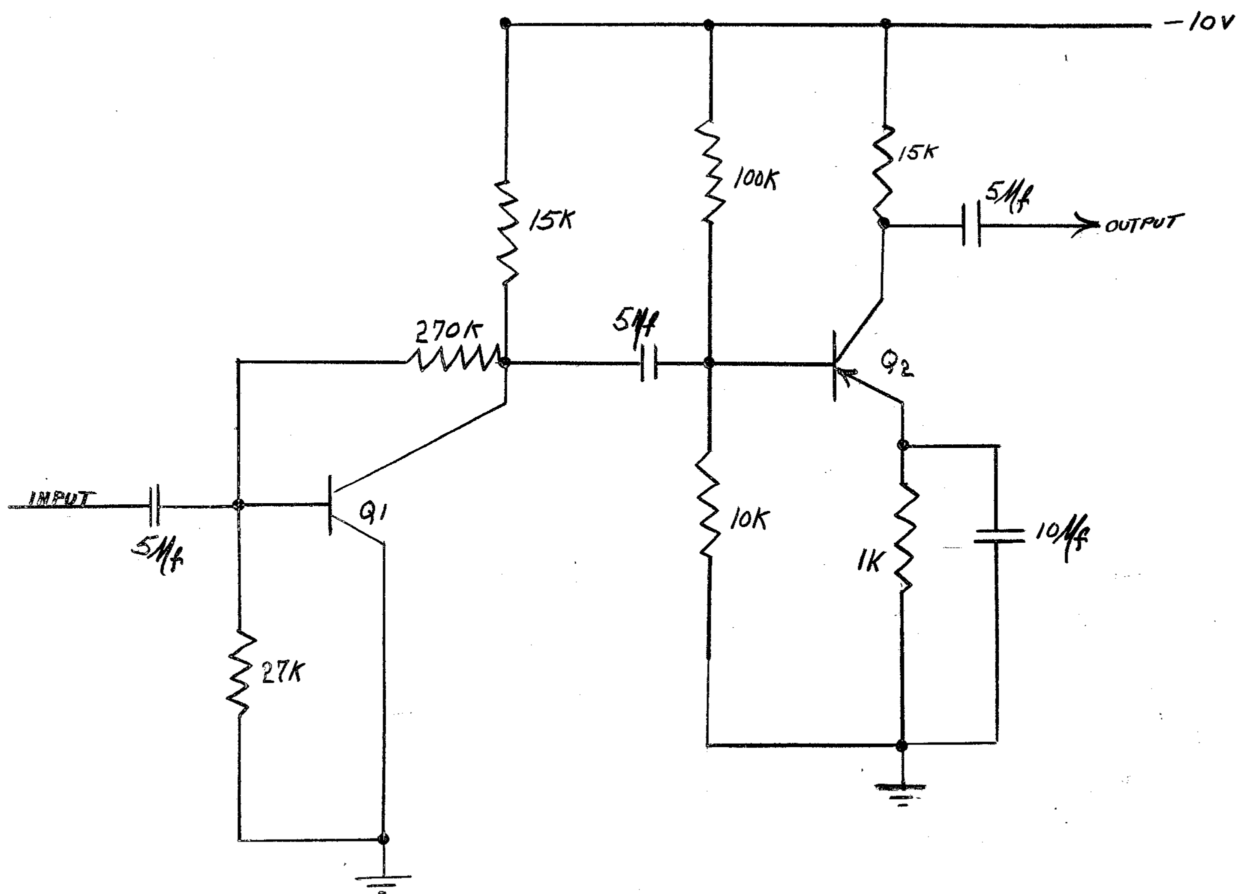


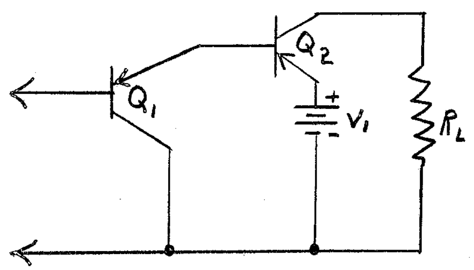
Figure 6.20. Two-Stage, Resistance-Capacitance Coupled Amplifier.

15000 Ω impedance of the collector circuit to the 1000 Ω input impedance of the following transistor. Because of this mismatch the maximum gain of Q_1 is not realized. The second transistor amplifies the signal further, and it is finally coupled out from the collector circuit of this stage through still another coupling capacitor. The gain of the second stage will depend upon the impedance of the load to be driven by the amplifier. If the load impedance is very much different from the collector impedance of Q_2 the power gain of the stage will be reduced. Also, as shown in Figure 6.19, the voltage swing of the amplifier will be restricted by low impedance loads.

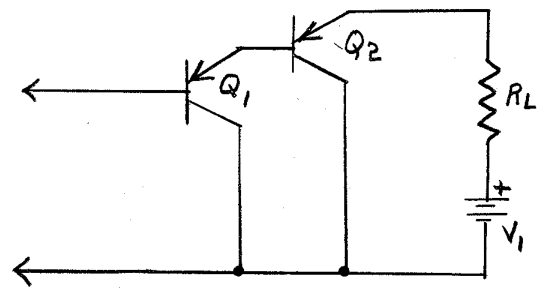
Resistance-capacitance coupling is generally used where wide-band frequency response is needed or where the amplifier is to be employed within a negative feedback loop (because of the lower phase shift per stage of r-c coupled amplifiers as compared to transformer coupled amplifiers). Also, in some cases, r-c coupling is less expensive than transformer coupling. However, the realizable gain per stage is lower with r-c coupling and it is only practical between stages with input and output impedances in the same order of magnitude because large impedance mismatches produce excessive loss of gain. Moreover, r-c coupling cannot be used on high-level stages because the nonlinear input characteristics of the transistor cause charging of the coupling capacitor and a shift in operating point -- but this will be discussed further along with high-level amplifiers.

Direct Coupled Amplifiers. The characteristics of the transistor and also the availability of PNP and NPN types makes this device well suited for direct coupled amplifiers. Several configurations for direct-coupling transistor stages are shown in Figure 6.21. Figure 6.21a shows a common collector stage coupled directly to a common emitter stage. The emitter current of Q_1 passes through the base of Q_2 with the input resistance of Q_2 acting as the load for Q_1 and the voltage V_1 acting as the collector supply. In Figure 6.21b a two-stage common collector amplifier is shown. Again the output current of the first stage is the input current of the second and the voltage source (V_1) has been connected to act as a collector supply for both stages. Figures 6.21 (c through e) illustrate the use of both NPN and PNP transistors. Figure 6.21e is of particular interest because it clearly shows that there is no intrinsic limit on the number of stages that can be cascaded. It can be seen from the figure that the collector circuit of the intermediate stages is in series with the input circuit of the following stage and a voltage source. The transistors and voltage sources must be connected so there is a path for current flow around this loop in such a direction as to forward bias the base-emitter junctions and reverse bias the collector junctions.

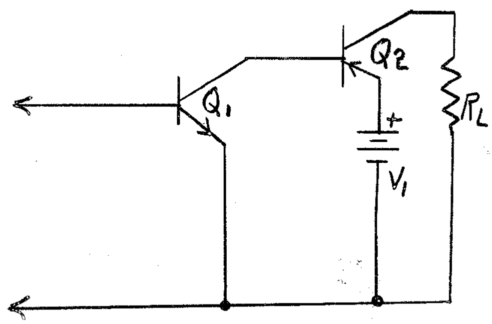
One thing that should be immediately apparent in these circuits is that temperature stability can become quite a problem. An increase in temperature will cause increases in collector current in all the stages, but the collector current increase of the first stage will be amplified by successive stages which will result in intolerable increases in the final stages, if precautions are not taken. Therefore, the circuits, as shown in Figure 6.21 are not practical.



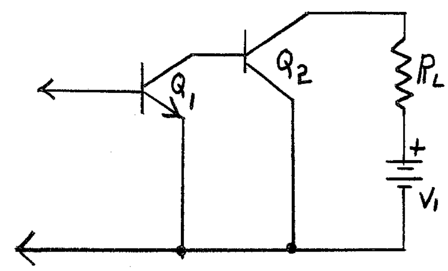
a. Common Collector, Common Emitter



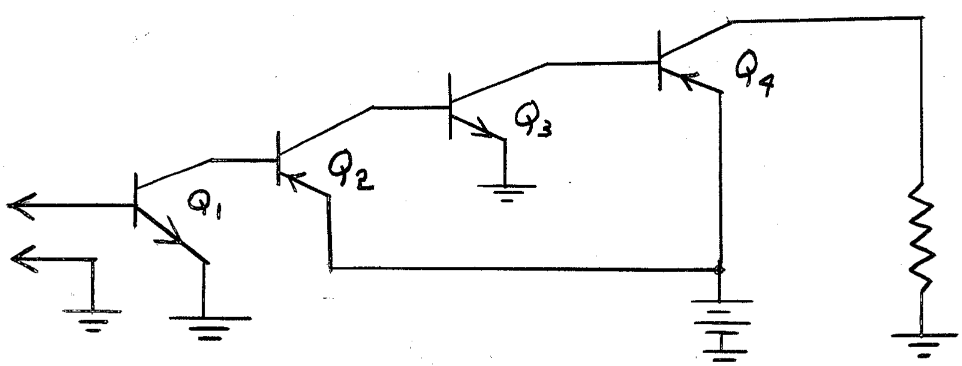
b. Two Stage Common Collector



c. Two Stage Common Emitter



d. Common Emitter, Common Collector



e. Cascaded Common Emitter Amplifier

Figure 6.21. Elementary, Direct-Coupled Amplifiers.

Figure 6.22 illustrates a practical two-stage common emitter, direct-coupled amplifier. Diode compensation is employed on the first stage. The voltage across D_1 will vary inversely with temperature, reducing the bias on Q_1 at higher temperatures. The input signal is developed across R_3 which delivers the bias voltage to Q_1. The resistor R_2 produces the bias on the compensating diode from the collector supply. A resistor (R_1) is placed on the input of Q_2 (base to emitter) to prevent forward biasing of Q_2 by its own reverse saturation current when Q_1 is cut off. The diode (D_2) is a zener diode. The voltage regulating qualities of this device are used to prevent signal variations of the collector-supply voltage. This diode takes the place of a bypass capacitor which would not be effective at very low frequencies. However, a capacitor could also be employed for additional bypassing at high frequencies.

Although temperature compensation alone is usually adequate for two-stage amplifiers, some form of feedback must be employed if more stages are used or if a wide range of temperatures is to be encountered. Figure 6.23 indicates how this might be accomplished. The circuit of Figure 6.23 uses diode compensation, three local feedback loops, and overall feedback. The purpose of the feedback is not only to improve d.c. stability but also to decrease nonlinear distortion.

The first stage employs diode compensation as did the previous circuit; however, the emitter resistor (R_3) also provides bias stabilization. In addition, the output voltage (across R_L) is fed back to the input stage <u>via</u>. R_7. As the conduction of any of the transistors increases, regardless of the cause, the voltage across R_L will increase. A portion of this increase will be coupled back across R_3 through R_7, decreasing bias on the first stage. This degeneration is also present for the signal so the overall gain is reduced and distortion decreased. The second transistor (Q_2) is connected in the common-collector configuration. The load of this stage (R_5, R_6, and the input impedance of Q_3) serve as an emitter resistor and provides local feedback on the stage. The third stage (Q_3) also incorporates an emitter resistor for local feedback. The input resistor (R_5) is connected between the base and emitter across R_6 so that the voltage variations across R_6 might be felt in the input circuit of Q_3. In addition to providing bias stabilization, the local feedback loops also improve the transistor input characteristics, thereby reducing distortion in the previous stages resulting from driving a nonlinear impedance.

In this circuit the zener diode (D_2) serves to keep signal variations out of the supply voltage.

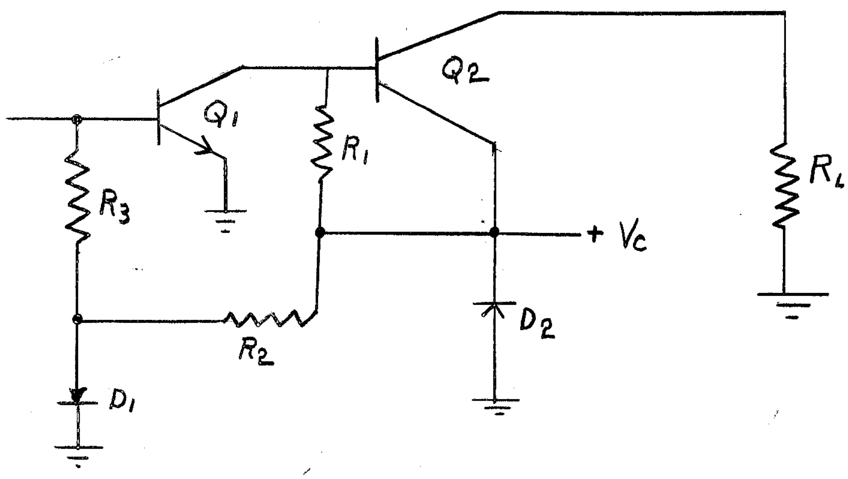


Figure 6.22. Temperature Compensated D-C Amplifier.

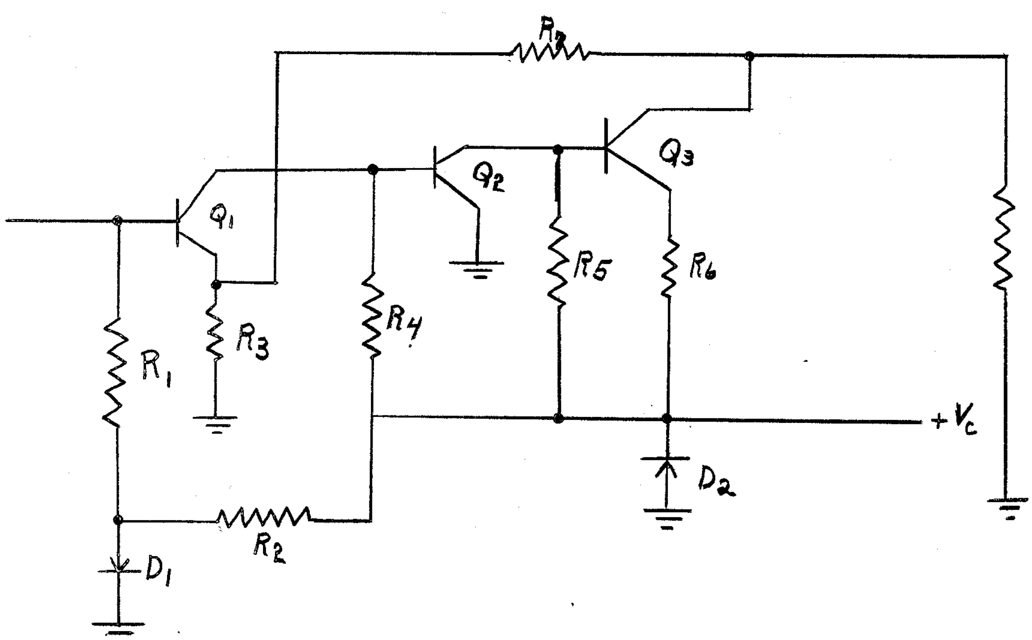


Figure 6.23. Temperature Compensated D-C Amplifier with Feedback to Improve Linearity and Stability.

The use of direct coupled amplifiers along with NPN and PNP transistors can be extended further to make many useful and interesting circuits. One of the most outstanding of these is the complementary-symmetry amplifier shown in Figure 6.24. This amplifier is most frequently used as a power amplifier, but its operation will be introduced here.

The input signal is fed to both Q_1 and Q_2, but these transistors have opposite polarity: one is an NPN and the other PNP. On the positive half cycles of the input signal, Q_1 will conduct while Q_2 is cut off. As the collector current of Q_1 passes through the base of Q_3, this transistor will also conduct, reproducing the positive half cycle of the input signal across R_L. On the negative half of the input signal Q_2 becomes forward biased and conducts, and Q_1 cuts off. This causes Q_4 to conduct and impresses some portion of Y_2 across the load, reproducing the negative half cycle. When the input signal is zero, no current passes through R_L so the output is also zero.

The circuit as shown in Figure 6.24 is somewhat simplified. Some form of bias stabilization is required. Also it is generally desirable to employ negative feedback to equalize the gain of the upper and lower channels. However, the techniques discussed previously can be applied here so these points will not be discussed further.

In establishing some relative merit for direct coupling, it can be said that it has many advantages over both transformer and r-c coupling. It is frequently used in low frequency signal amplifiers and also in wide band amplifiers using large amounts of negative feedback. The absence of any reactive components (capacitors and inductors) gives the d.c. amplifier a favorable position as regards low-frequency phase shift. The elimination of coupling components may permit reductions in equipment size -- particularly in the case of high-power circuits. However, some of the limitations of r-c coupling are also present with direct coupling. In particular, the direct coupling of stages does not permit impedance matching so maximum gain cannot be realized. Nonetheless, higher gains than with r-c coupling are possible because none of the signal is lost in a collector resistor. Moreover, the absence of the collector resistance eliminates the inefficient use of d.c. power as in r-c amplifiers.

POWER AMPLIFIERS

Thus far low-level amplifiers have been considered. The signal levels involved have been assumed to be low enough so that operation of the transistor could be considered linear. If this condition is satisfied, some definite value can be assigned to the transistor parameters, such as input impedance, output impedance, and current

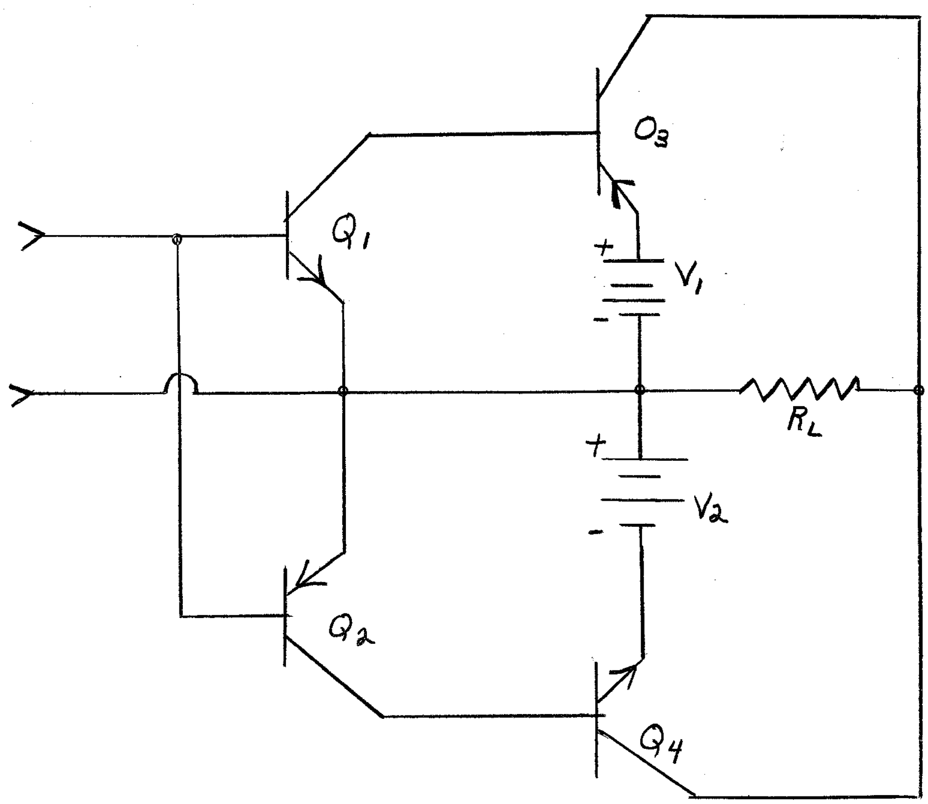


Figure 6.24. Single-Ended, Push-Pull Class B Amplifier.

gain. Moreover, the nonlinear distortion produced in the amplifier could be neglected. When the signal amplitude is increased, as in a power amplifier, this is no longer true: the input impedance will vary with the instantaneous input voltage, and the current gain will vary with instantaneous input voltage, and the current gain will vary with instantaneous collector current. Because of these nonlinearities, the amplifier will produce distortion and care must be taken in the design of circuitry to keep the distortion at a minimum while producing a large output signal.

The Output Circuit. Before going into the drive requirements of power transistors and the reduction of nonlinear distortion, the generation of power in the output circuit will be related to the maximum ratings of the transistor.

As was mentioned, the output impedance of a transistor does not determine the optimum load impedance for a power amplifier. Instead, it is the maximum ratings of the device that must be considered. In order to produce a large power output, it is necessary to have a large voltage and current swing, so the collector reverse breakdown voltage and the maximum collector current will be limiting factors. Moreover, in the operation of the amplifier, power will be dissipated into heat, so the maximum permissible power dissipation must be considered.

Amplifier Circuits. The discussion of low-frequency power amplifiers will be limited to two types: the single-ended, class A amplifier and the push-pull, class B amplifier. Other types (push-pull class A and class AB) will require only simple modification of the given information.

A single-ended, class A amplifier is shown in Figure 6.25a. The transistor is biased to some quiescent collector current, and the input signal varies the collector current about this point. For maximum output power the collector current varies between zero and twice the quiescent value while the collector voltage varies between zero and twice the supply voltage. The collector characteristics of a transistor are shown in Figure 6.25b. The loadlines (R_1, R_2, and R_3) represent three different values of collector load for the transistor in Figure 6.25a. R_1 is the optimum load resistance because it allows the maximum voltage and current swings. The loadline, R_2, represents the result of too low a load impedance: the full current swing can be realized, but the voltage swing is restricted. The loadline, R_3, illustrates the effect of too high a load impedance: the full voltage swing can be realized, but the current swing is restricted.

It can be seen from Figure 6.25b, that with an optimum load impedance, the available power output at a given supply voltage (V_c) will be determined by the quiescent collector current: the larger

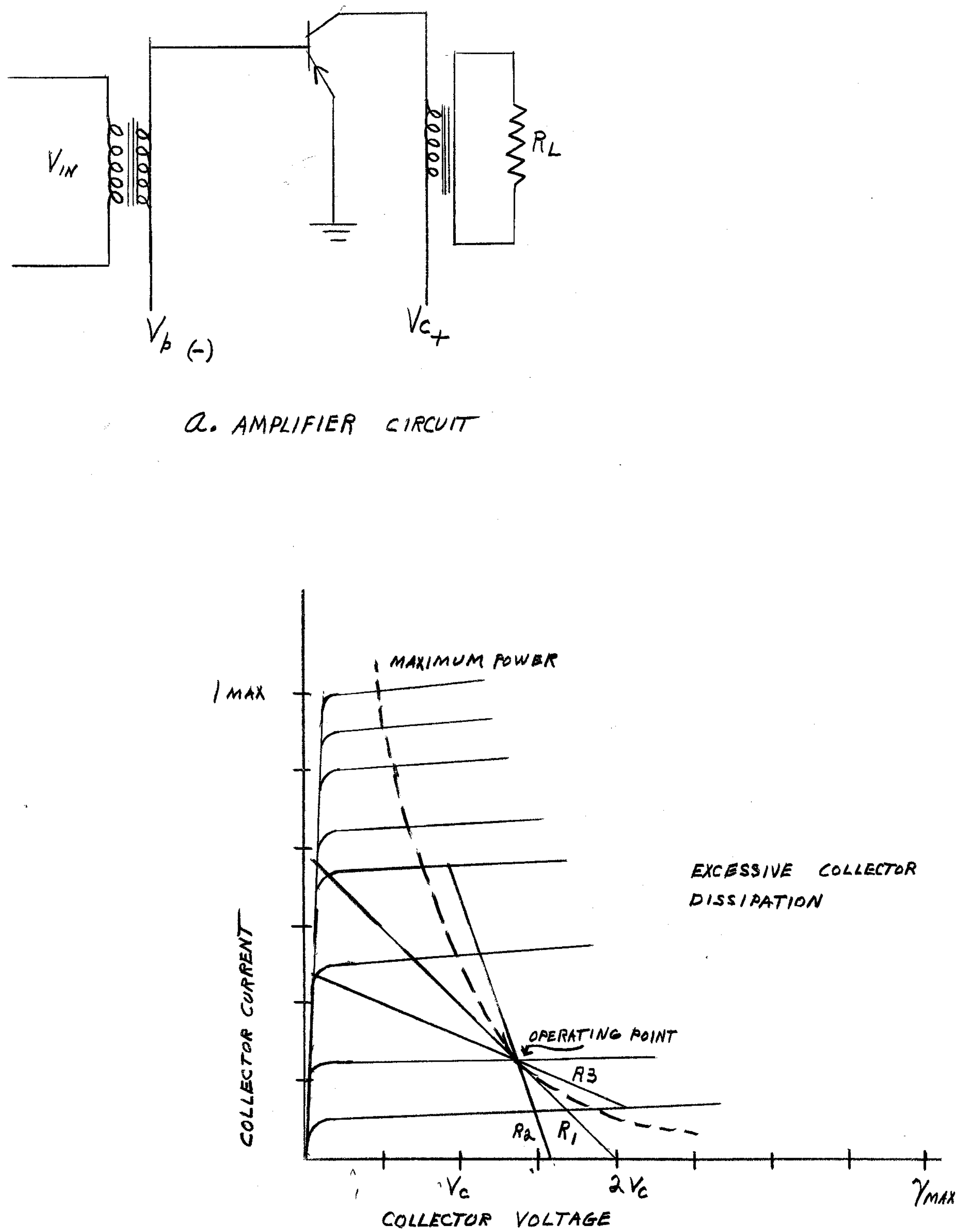


Figure 6.25. Single-Ended, Class A, Power Amplifier.

the quiescent collector current, the larger the current swing. However, the magnitude of the quiescent current is limited by the maximum permissible power dissipation of the transistor. The maximum theoretical power output of a class A amplifier is equal to one-half the zero-signal power dissipation, which indicates an overall collector circuit efficiency of 50%. This theoretical figure can only be approached in practice. Efficiencies in the order of 40-45% can be realized with transistor circuits.

One disadvantage of the class A power amplifier is that large amounts of d.c. power are consumed, even under no-signal conditions. This is particularly undesirable with amplifiers operating from batteries. The push-pull class B amplifier, illustrated in Figure 6.26b, overcomes this limitation. Power is drawn from the d.c. circuit only when there is an output signal. With reference to the circuit, it is evident that neither transistor is conducting under zero-signal conditions, as no forward bias is provided. However, when a signal is applied, the transistors will conduct on alternate half cycles. Hence, each transistor amplifies half the input signal and remains nonconducting during the other half.

The loadline for the push-pull amplifier can be represented using the composite characteristic curves shown in Figure 6.26a. To draw these curves it is necessary to align the two individual curves so that the collector voltages correspond at the supply voltage, as shown in the figure. The loadline must then pass through zero collector current at this point of correspondence. With the push-pull class B amplifier, the output power increases with decreasing load resistance: the maximum voltage swing will be very nearly equal to twice the supply voltage regardless of the load resistance, so a smaller load will give a larger current swing and a higher power output. The minimum load resistance, and therefore the maximum power output, is limited by the peak current or maximum power ratings of the transistor - whichever would be exceeded first.

The theoretical, maximum efficiency of a class B amplifier is 78.5%. Ideally, the efficiency is independent of load resistance because, as the power output is increased or decreased by changing the load resistance, the input power will change proportionally. This condition is closely approximated in transistor amplifiers, and efficiencies greater than 65% are not difficult to realize over a wide range of load impedances. The maximum, average-power dissipation of the output transistors does not occur at maximum power output with class B amplifiers. It occurs at somewhat less than full output where the output efficiency falls to approximately 50%. This must be taken into consideration with the permissible dissipation of the transistor.

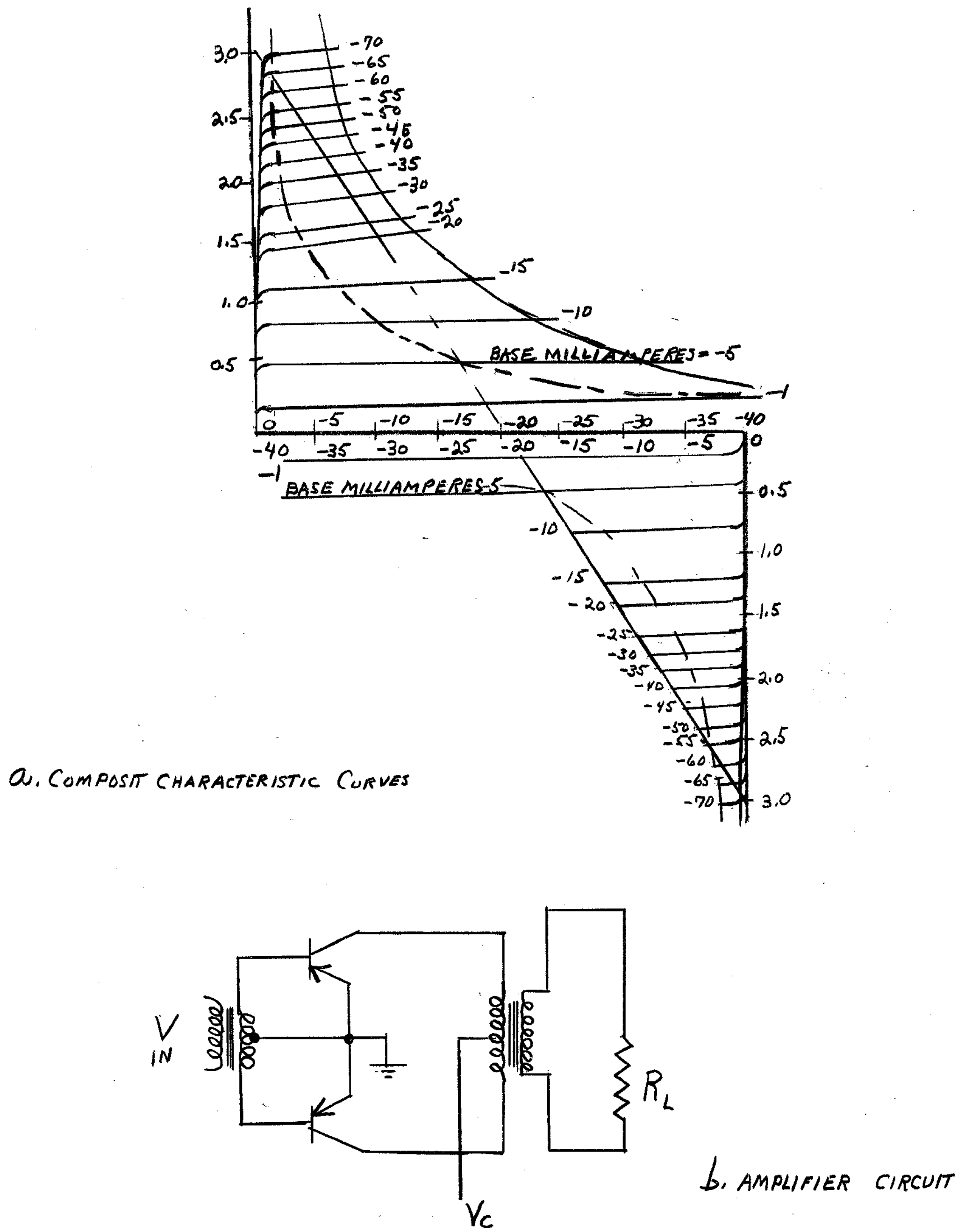


Figure 6.26. Push-Pull, Class B Power Amplifier.

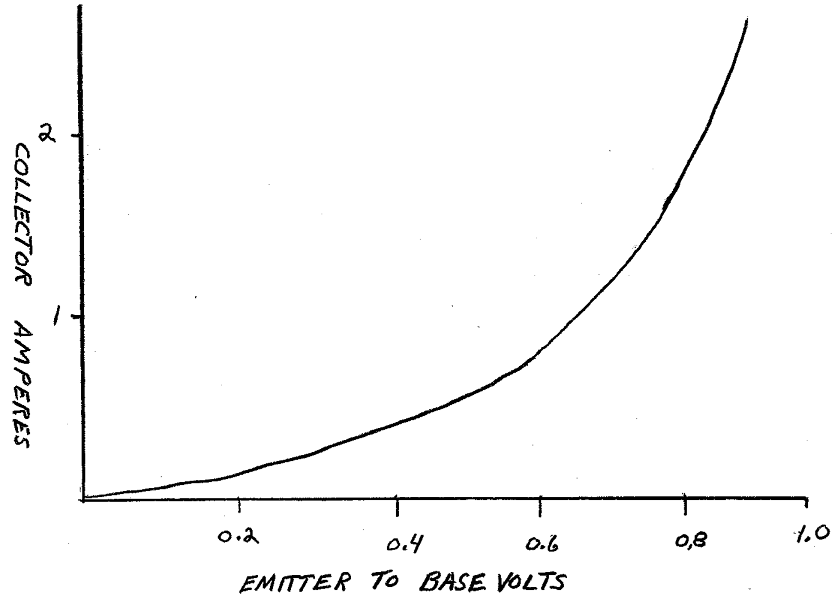
In comparing the output circuits of the class A and class B amplifiers, it can be said that the letter is more conservative from the standpoint of power drain, both because of its higher efficiency and because of its low standby current. The advantages of the class A amplifier arise from the economy of using only one transistor and also from the less stringent control of transistors required to reduce distortion to a given level. It should be obvious that the characteristics of the output transistors of a class B amplifier must be closely matched to produce an acceptably low distortion level. This task is difficult, and often some form of negative feedback is required.

It should be noted that the previous discussion covered only common-emitter amplifiers. This is not meant to imply that other types are not used; however, as far as the output circuit is concerned, common-emitter and common-collector amplifiers are identical. The difference lies in the input circuit. The output characteristics of the common-base circuit are slightly different, but the same considerations apply. Again, with the common-base circuit, the main difference lies in the drive circuits.

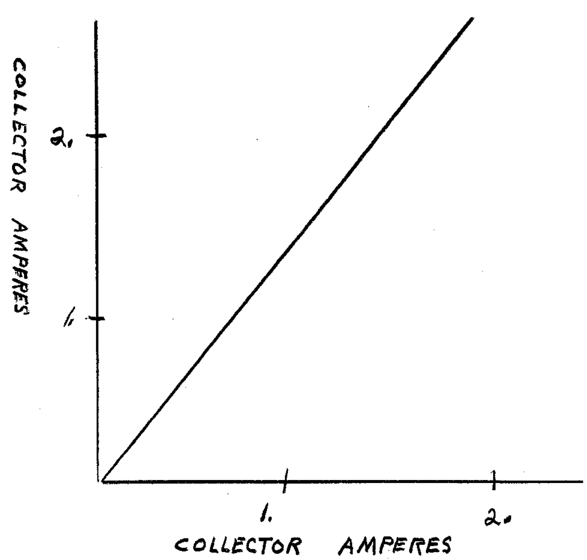
Drive Requirements. It has already been stated that no particular valve can be assigned to such parameters as the input impedance and current because of variations with signal level. Nevertheless, under certain conditions these parameters can be assigned a nominal value if operation is sufficiently linear.

In coupling the driver to the output stage the objective is not to transfer maximum power; but instead, the driver should be matched to the output stage in such a way as to produce minimum distortion. This point is clearly illustrated in Figure 6.27 which gives the transfer characteristics of a common-base amplifier. Figure 6.27a shows that the voltage transfer characteristics of a common-base amplifier are extremely nonlinear: that is, if a sine wave voltage were applied to the input of a common-base amplifier, the collector current would not be a sine wave -- an appreciable amount of distortion would be generated. However, if a current sine wave were applied to the input, it can be seen from Figure 6.27b that the collector current would be a sine wave; and very little distortion would be generated. Hence, it is desirable to drive the common-base amplifier from a high-impedance source.

This behaviour can be explained with the help of the simplified equivalent circuit of Figure 6.28. The reason for the nonlinearity in the voltage transfer characteristics is the variation in the input impedance with instantaneous signal voltage; therefore the input current will not vary linearly with input voltage. The current transfer characteristics, on the other hand, are linear because the output current is equal to the product of the input current and the alpha current gain (α) which is always very nearly equal



a. VOLTAGE TRANSFER CHARACTERISTICS



b. CURRENT TRANSFER CHARACTERISTICS

Figure 6.27. Transfer Characteristics for a Common-Base Amplifier.

to one. In Figure 6.28, the driver is represented by a constant voltage source with its internal impedance in series (R_g). If this internal impednace is high compared to the low input impedance of the transistor, the current flowing into the emitter will depend on the source voltage and the internal impedance: the fluctuations of input impedance in the transistor will produce very small changes in the total impedance seen by the voltage source, so the input current will have the same waveshape as the source voltage. If, however, the driver impedance is low compared to the input impedance, fluctuations of the latter quantity will produce appreciable changes in the total impedance seen by the voltage source; and the input current will depend not only on the source voltage and internal impedance, but also on the input impedance which is not constant. In this case the input current does not have the same waveshape as the source voltage, and distortion is produced.

This argument shows that it is undesirable to match the driver impedance to the input impedance of the output stage. Instead, it is desirable to have the driver impedance very much higher than the input impedance. This makes for inefficient transfer of power between the driver and output stage. Moreover, the low (unity) current gain of the common-base amplifier puts excessive current demands on the driver. For these reasons, the common-base configuration is not used frequently as a linear power amplifier.

In the case of a common-emitter amplifier, both the voltage and current transfer characteristics are nonlinear to some extent - although the nonlinearity is not as great as with the common-base configuration. The base input characteristics in Figure 6.29a, however, do resemble those of the common base connection: the input impedance decreases with increasing input voltage. A given change in base voltage will be more effective in producing a change in base current at higher input voltages. Nevertheless, this effect is not so pronounced in the voltage transfer characteristics because of the fall-off of common-emitter (α_E) current gain at high current levels. This fall-off explains the nonlinearity in the current-transfer plot (Figure 6.29c).

Because of the nonlinearity in both the voltage and current transfer characteristics, it cannot be stated whether a high or low impedance driver will produce the lowest distortion. Nevertheless, a comparison of the two plots will show that the voltage characteristic has an increase in slope at high collector currents while the current characteristic has a decrease in slope at high collector currents. It would seem, therefore, that some intermediate value of driver impedance would produce the lowest distortion. This is indeed the case, and because of this it is possible to have a reasonable impedance match between the driver and the output stage consistent with low distortion. The common-emitter amplifier will therefore,

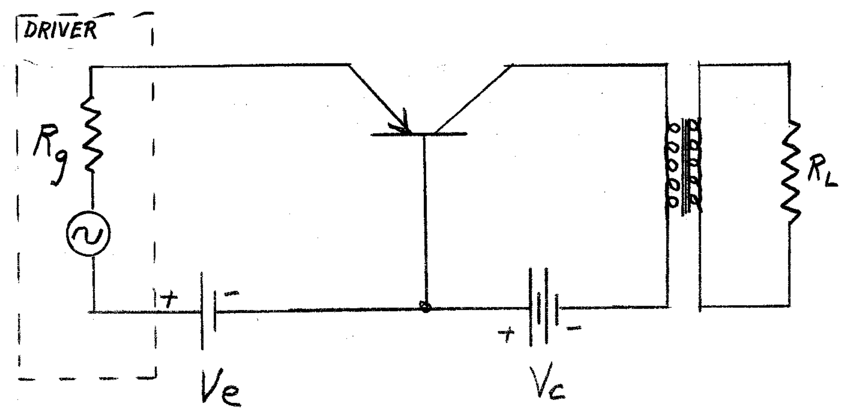


Figure 6.28. Simplified Equivalent Circuit for Common-Base Amplifier and Driver.

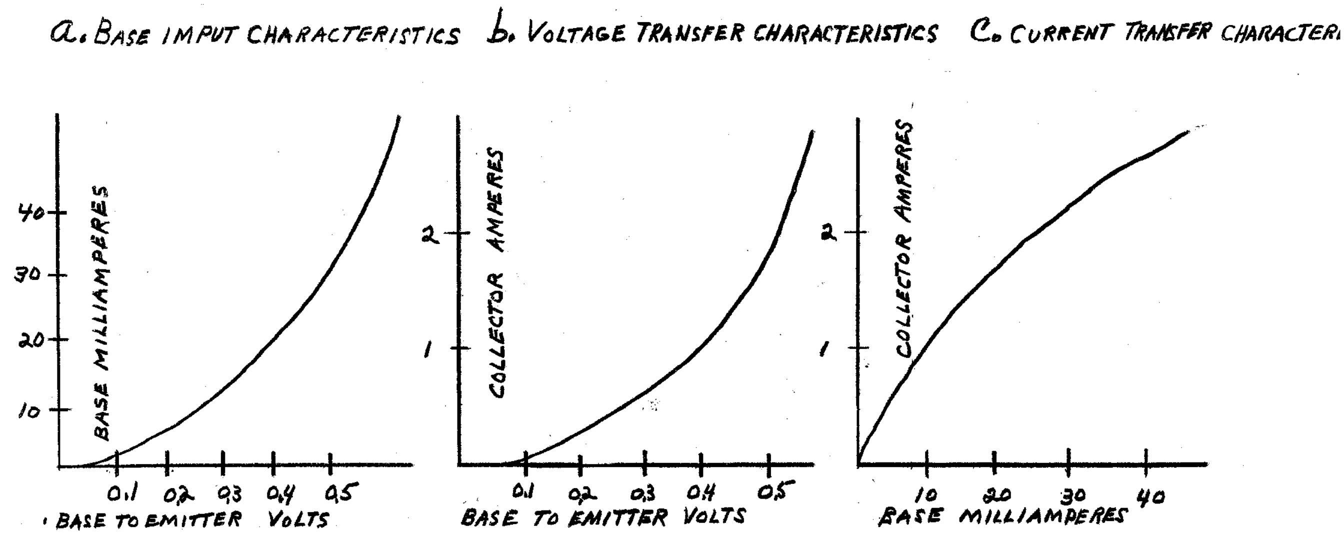


Figure 6.29. Transfer Characteristics for a Common Emitter Amplifier

make efficient use of driver power; moreover, both the driver and the output stage will have a much greater power gain as compared to the common base amplifier. For these reasons the common emitter amplifier is commonly used in power amplifiers.

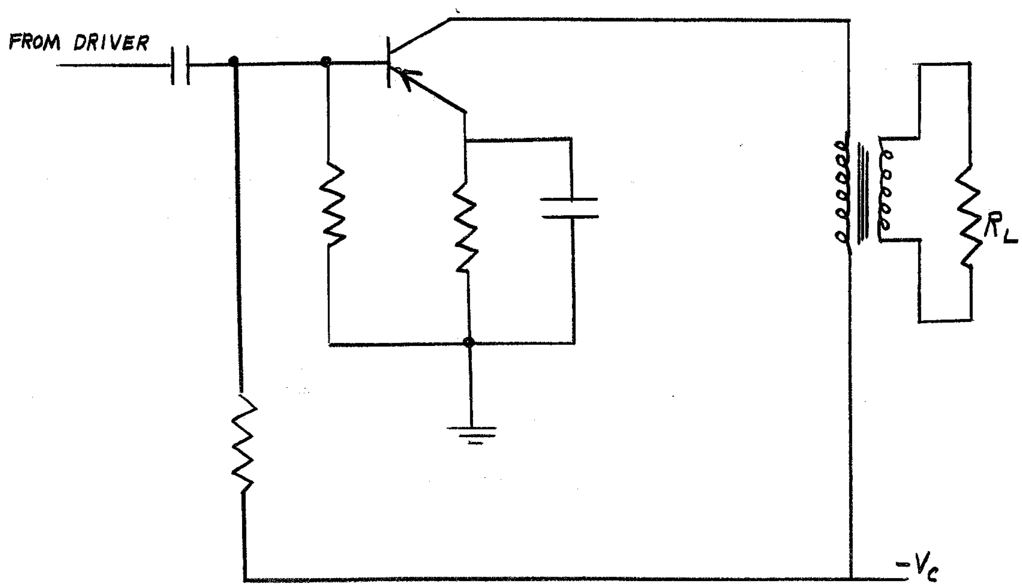
The common-collector power amplifier can be considered to be a common-emitter stage with 100% negative voltage feedback, applied in such a way as to increase the apparent input impedance and linearize the voltage transfer characteristic. It is clear, therefore, that the common-collector amplifier should be driven from a low-impedance source. One disadvantage of this configuration is the large voltage-swing requirement that must be met by the driver: the negative feedback reduces the voltage gain to less than unity. Driven from a constant voltage source, the common-emitter emplifier has a very low distortion; but if it is driven from a constant current source, its operation will be identical to that of a common-emitter amplifier.

The above results can be summarized as follows: the common-base power amplifier should be fed from a constant-current (high-impedance) source, and the common-collector amplifier should be fed from a constant-voltage (low-impedance) source. Moreover, the former will tend to maintain a constant current through a varying load impedance in its output circuit so it can be considered to have a high impedance output, while the latter will try to maintain a constant output voltage across a varying load so it can be considered a low impedance source. Both these configurations have a low power gain when driven properly and put excessive current or voltage requirements on the driver. Therefore, they are not too frequently used in practice.

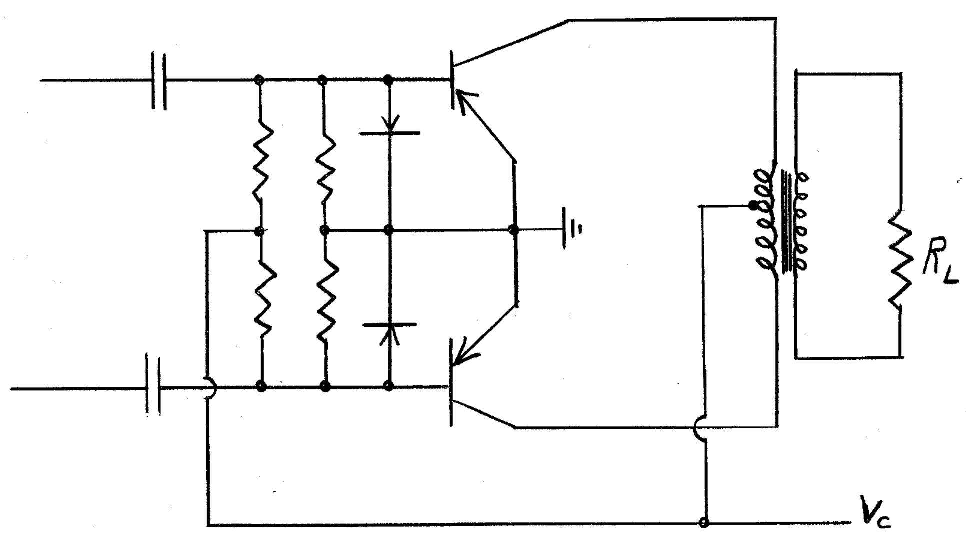
The common-emitter power amplifier has a high power gain. In addition, for minimum distortion, the driver is very nearly matched to the output stage so there is a reasonable efficiency in using the driver power. This configuration is the most practical for use as a power amplifier so the major portion of the circuits given will be confined to the common-emitter circuit.

<u>Driver Coupling</u>. The driver can be r-c coupled to a class A output stage as shown in Figure 6.30a. However, because some significant amount of power must be generated in the driver, this method is frequently too inefficient. Furthermore, the nonlinearity in the transistor input resistance can cause charging of the coupling capacitance at high signal levels, which will shift the operating point of the transistor. Hence, r-c coupling is used only in low-power amplifiers, such as driver stages.

An example of an r-c coupled, class B amplifier is given in Figure 6.30b. Diodes must be included in the input circuit to prevent



a. CLASS A AMPLIFIER



b. CLASS B AMPLIFIER

Figure 6.30. Resistance-Capacitance Coupling.

charging of the coupling capacitors by the half-cycle conduction of the output transistors. A small, forward bias is applied to the output transistors, as with any class B stage, to minimize "cross-over distortion" when the conduction shifts from one transistor to the other. As with class A amplifiers, the use of r-c coupling is quite limited.

The usual method of coupling the driver to the output stage is transformer coupling. A transformer allows a good efficiency in the driver stage as well as giving any desired impedance ratio. A driver transformer also eliminates the necessity of a phase-inverter in push-pull circuits.

A practical class A amplifier is shown in Figure 6.31. The driver stage is transformer coupled to the output transistor. The output transistor is operating with a split load: that is, half the load is in the emitter and half in the collector circuit. This combines the advantages of common collector with common emitter by applying 50% negative feedback. Amplifiers of this type can be built to deliver powers greater than 5 watts using germanium transistors.

A transformer coupled, class B amplifier is shown in Figure 6.32. A class A amplifier is used as a driver. The output stage has a small forward bias to minimize crossover distortion. This bias is temperature compensated with a germanium diode. Negative feedback is applied to the emitter of the driver from the secondary of the output transformer to minimize the distortion caused by inequalities in the output transistors. Class B amplifiers with outputs as high as 200w can be built using available transistors.

Direct coupling can also be used to couple the driver to the output stage, but this method presents the same impedance-mismatch problems as are experienced with r-c coupling. Also, there is the problem of d.c. instability arising from temperature effects. Direct coupled power amplifiers are used, however, where it is desirable to eliminate the size and weight as well as the frequency and phase shift limitations of the transformer. An example of a direct-coupled power amplifier, the class B complementary-symmetry amplifier, has already been given (Figure 6.24). This circuit uses complementary (matched PNP, NPN) power transistors which are difficult to obtain commercially so this circuit has only limited uses. Another circuit, the quasi complementary amplifier, (Figure 6.33) has a direct-coupled, push-pull output stage and uses only PNP power transistors. In this circuit, the transistors (Q_1, Q_2, Q_3, and Q_4) are not conducting under no-signal conditions so there is no voltage across the load. A negative input signal on the base of Q_3 drives this transistor (Q_3) and, consequently, Q_1 into conduction; and the negative half of the signal appears across the load. A positive signal, applied at the

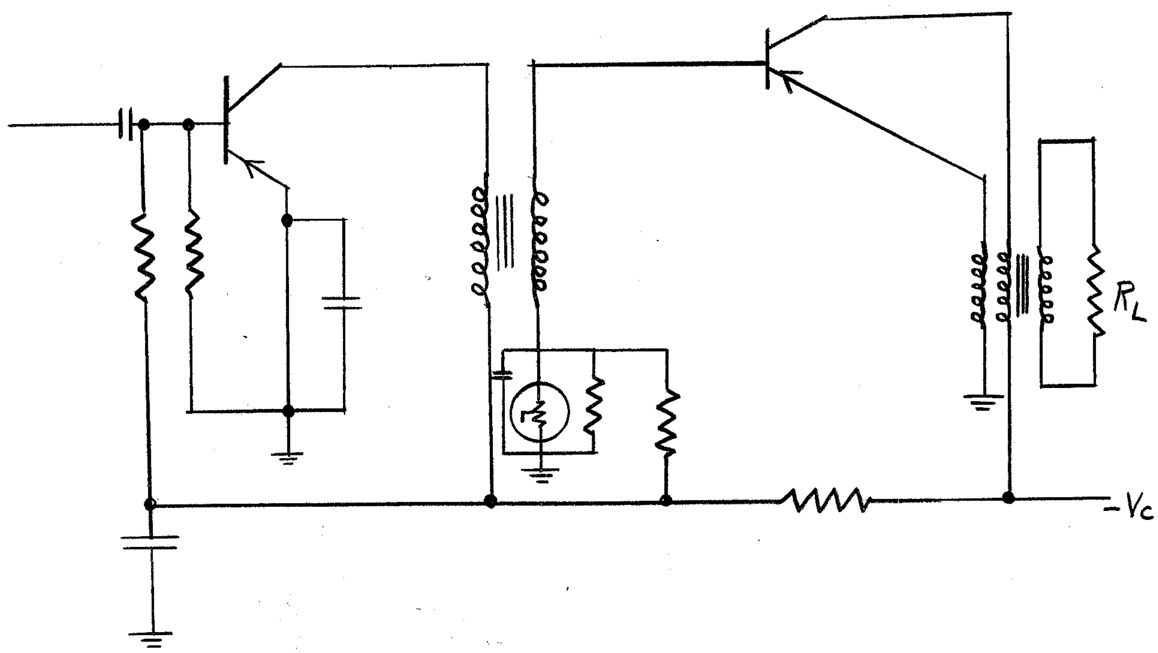


Figure 6.31. Class A Power Amplifier.

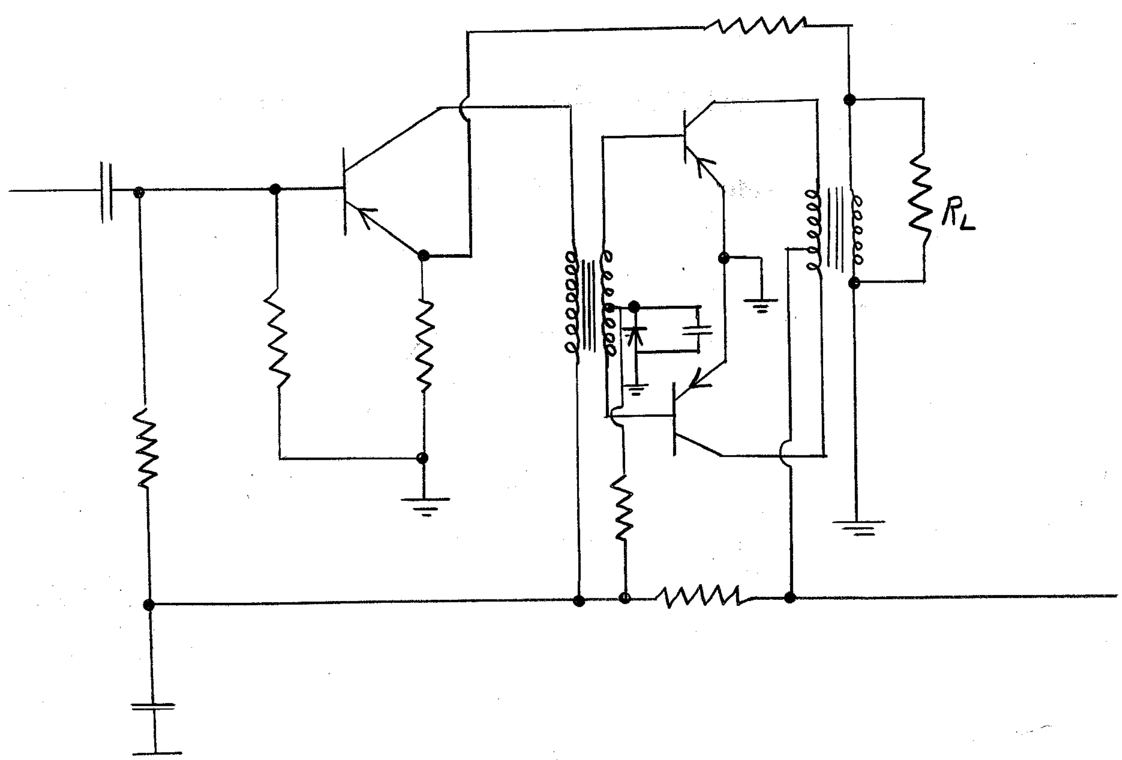


Figure 6.32. Class B Power Amplifier.

same point, drives the NPN transistor, Q_4, into conduction. When Q_4 conducts, Q_2 will also conduct and the positive half of the signal will appear across R_L. The upper (Q_3 and Q_1) and lower (Q_4 and Q_2) channels are not symmetrical, but the application of 100% negative feedback eliminates signal unbalance. The upper channel uses two PNP transistors in the common-collector connection. The input signal is applied to the base of Q_3, and its emitter is connected to the input (base) of Q_1. The emitter of Q_1 goes to the load. The collectors of both Q_3 and Q_1 are brought to signal ground at the negative terminal of V_1. Because of this common collector connection, the voltage gain of the upper channel is very close to one. The lower channel uses a NPN and a PNP transistor cascaded in common emitter connection. The input signal is connected to the base of Q_4, and its collector is direct coupled to the base of Q_2. The collector of Q_2 is connected to the load. The voltage gain of cascaded common-emitter stages is normally very much greater than one, but 100% negative feedback is applied to this channel by connecting the emitter of Q_4 to the output rather than to signal ground. The feedback reduces the voltage gain of this channel to slightly less than one; therefore, the voltage gain of both the upper and lower channels is nearly equal; and they are electrically balanced. Signal is fed to the amplifier by the class A driver, Q_5. If it is assumed that Q_5 is operating at its maximum swing, it can be seen that the bases of Q_3 and Q_4 will alternately be connected to the negative battery (through R_1) when Q_5 is nonconducting and to the positive battery when Q_5 is saturated.

Direct-coupled power amplifiers are particularly applicable in high-fidelity systems. The elimination of the output transformer can reduce cost considerably while equivalent circuit performance is maintained.

Conclusions. There is much to be said for transistor power amplifiers, their high (near theoretical) efficiency being only one advantage. Furthermore, transistor amplifiers can operate directly from standard d.c. systems (6, 12 and 28V), thus eliminating the need for power converters. Also, their small size permits reduction in equipment size. However, they are not without limitations: one is that the beta cut-off frequency of many alloy-junction power transistors is near the high end of the audio frequency band; a second is the restricted temperature range of germanium transistors which imposes severe limits on collector dissipation above about 70°C. Both of these limits have been appreciably extended with silicon transistors employing a diffused-base structure, but at the present time the cost of these units is quite high.

HIGH-FREQUENCY AMPLIFIERS

There are two important limitations to the high-frequency performance of a transistor: the first is the finite transit time for current

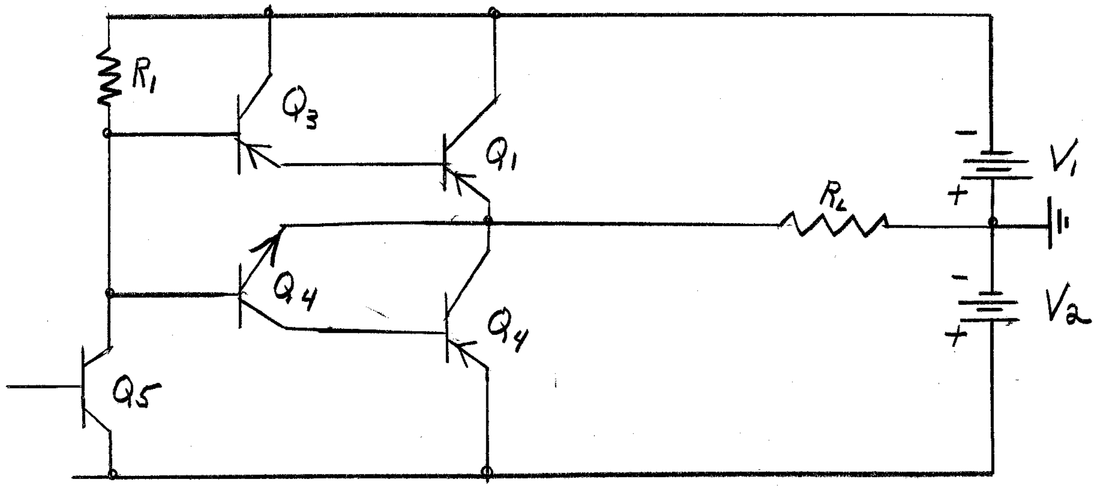


Figure 6.33. Simplified Schematic of the Quasi-Complementary Class B Amplifier.

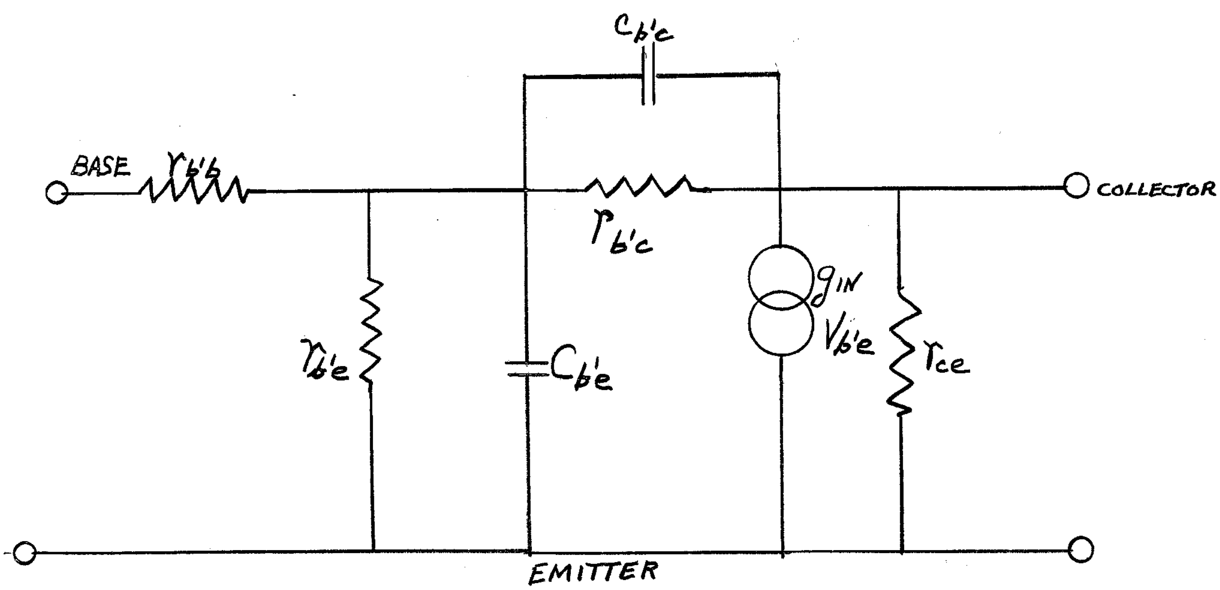


Figure 6.34. Transistor Equivalent Circuit.

carriers crossing the base, and the second is the collector-junction capacitance. The transit time effect is reflected in the alpha cutoff frequency ($f_{\alpha co}$) rating of the device. This is an absolute limitation: that is, there is no way to design around, or compensate, for this frequency restriction. On the other hand, the collector capacitance can be tuned out or compensated for to some extent.

An equivalent circuit can be used to approximate the characteristics of a transistor. These circuits do not describe transistor performance but only approximate it over the useful frequency range for purposes of circuit analysis. Figure 6.34 gives one possible equivalent circuit convenient for analysis of common emitter circuits. The capacitance shunting the input circuit (C_{b^1e}) represents the alpha cutoff frequency limitations, while the collector junction capacitance is shown as C_{b^1c}. It can be seen that C_{b^1c} provides a feedback path between the output and input circuits in addition to shunting the output circuit (C_{b^1e} is very much greater than C_{b^1c}). This feedback path frequently necessitates the use of neutralization in tuned amplifiers.

<u>Video Amplifiers</u>. High-frequency amplifiers will be divided into two groups: video amplifiers which are intended to cover a wide band of frequencies from about 10cps to about 10mcps, and tuned amplifiers which employ resonant circuits and cover a relatively narrow frequency range.

The appearance of transistor video circuits is quite similar to the low-frequency amplifiers already discussed. Common base, common emitter, and common collector stages are possible. Because of the wide band of frequencies to be passed, it is necessary to use either direct or r-c coupling as it is generally impossible to build transformers with the required bandwidth. In order to cascade single-stage amplifiers and realize any gain, then, there must be a reasonable impedance match between the input and output impedances of successive stages since there is no means available for matching. For this reason, video amplifiers almost always use the common-emitter configuration except, possibly, in the input or output stages where there is some special impedance requirement. Hence, the discussion will be confined to this circuit.

In common-emitter amplifiers the beta cutoff frequency is of interest; it is related to the alpha cutoff frequency by

$$f_{\beta co} = f_{\alpha co}(1-\alpha_o)$$

where α_o is the low-frequency alpha current gain. From this equation it can be seen that the beta cutoff frequency is considerably lower than the alpha cutoff frequency. How much lower will depend on the low-frequency alpha current gain; therefore, transistors with exceptionally high current gain are not generally desirable.

Figure 6.35 shows the characteristics of a common-emitter video amplifier that has its upper limit determined by the beta cutoff frequency. Lowering the load resistance below the optimum value reduces gain but does not raise the upper frequency limit. However, the insertion of a compensating inductor will increase the maximum frequency, because the coil raises the effective load impedance at higher frequencies.

As a rule, alloy-junction transistors and some grown-junction transistors suffer from this type of frequency limitation. These transistors do not have an accelerating (drift)field across the base region so the transit time is comparatively large.

Transistors that have their maximum frequency imposed by collector-junction capacitance can be compensated to some extent as shown in Figure 6.36. In this case lowering the load resistance will increase the bandwidth in addition to reducing gain because the shunt impedance of the output capacitance will not be comparable to the load resistance until a higher frequency is reached. On this basis, gain can be traded evenly for bandwidth. Moreover, compensation will extend the maximum frequency in either case because, in this case, the compensating inductor will form a low-Q parallel-resonant circuit with the output capacitance and effectively remove it from the circuit over a limited frequency range.

Transistors having a graded base giving a drift field, such as diffused junction and some grown junction transistors, have their high-frequency performance limited by collector-output capacitance. However, in most cases, the maximum frequency of these devices is considerably greater than those without the drift field.

Another circuit worth mentioning is shown in Figure 6.37. This is a video amplifier using a tetrode transistor. The same considerations apply here as did in the previous discussion. The only unusual feature of this circuit is that a cross-base bias must be supplied to the transistor to electrically reduce the junction area. Bias of the correct polarity is applied as shown in the figure.

As was pointed out, video amplifiers are very much like direct or r-c coupled, low-frequency amplifiers except that high-frequency transistors must be used. There are, nevertheless, other points that should be considered. First, the cutoff frequency is raised by increasing the collector voltage. The increased voltage widens the collector depletion region and, therefore, reduces output capacitance. Furthermore, the depletion region extends into the base so base width and, consequently, transit time is reduced. Secondly, both the input and output impedances are considerably lower at high frequencies. This is an important point where stages are to be cascaded.

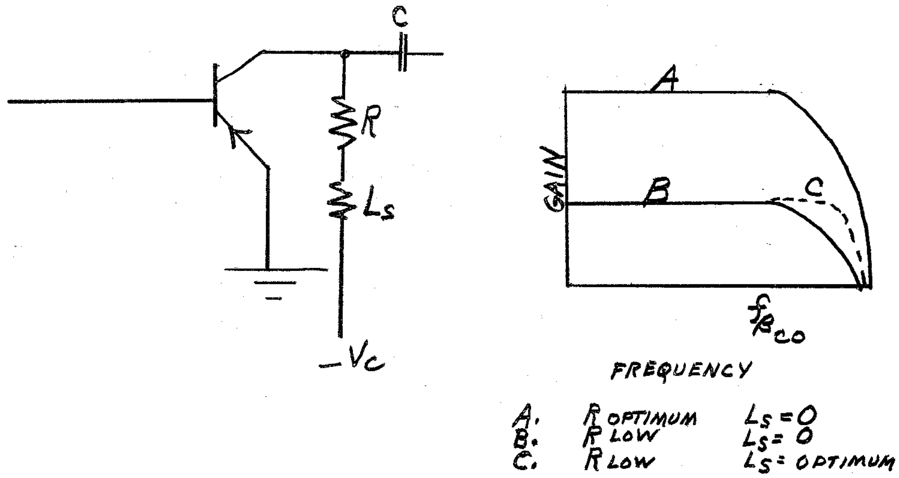


Figure 6.35. Output Circuit of Video Amplifier Limited By Beta-Cutoff Frequency.

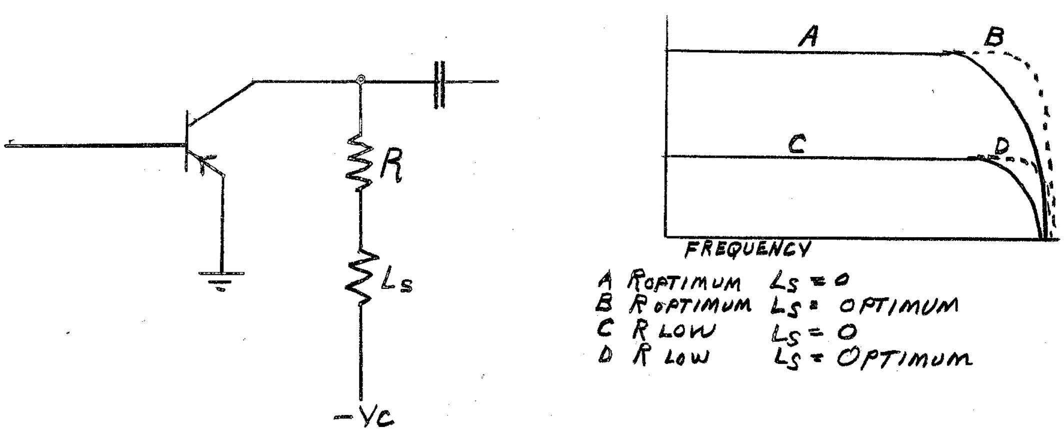


Figure 6.36. Output Circuit of Video Amplifier Limited by Collector-Junction Capacitance.

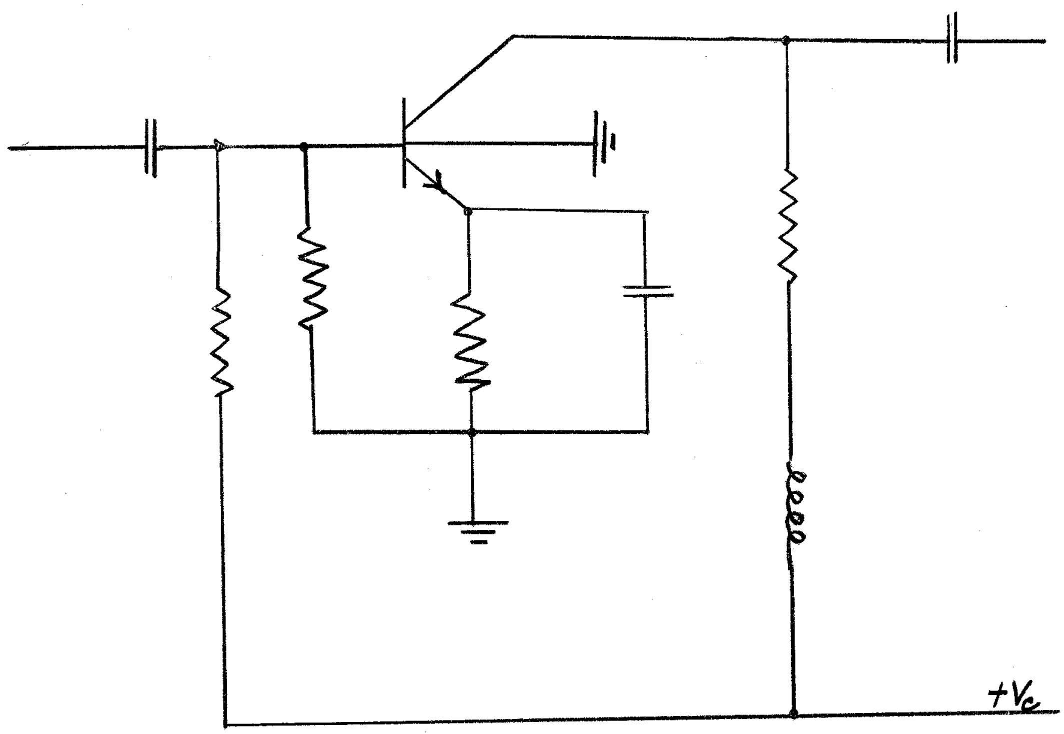


Figure 6.37. Video Amplifier Using a Tetrode Transistor.

Tuned Amplifiers. In a tuned amplifier, the major frequency limitation is the alpha-cutoff frequency since the collector junction capacitance can be made part of the resonant circuit in the collector. This capacitance, nevertheless, does provide a feedback path and in some cases neutralization must be provided for stable operation.

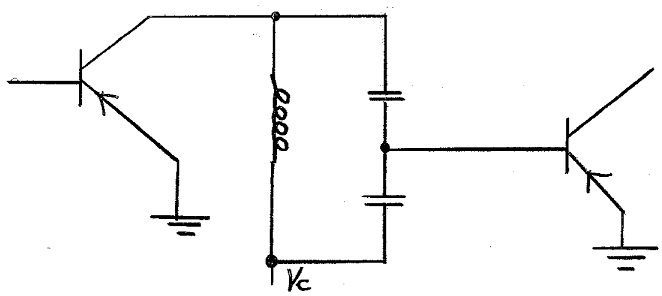
Impedance transformation is no problem with tuned circuits as is shown in Figure 6.38, so in tuned amplifiers impedances are usually matched to realize maximum gain. This, of course, is imperative with the common-base connection because the entire gain arises from the difference between input and output impedances. The matched-impedance power gain of the common emitter configuration is higher than both common collector and common emitter so this connection is generally preferred. However, at frequencies near the alpha-cutoff of the transistor, the difference between the common emitter and common base connections becomes small; and the latter is frequently used in the interest of circuit stability.

With present day transistors, the collector-junction capacitance is small enough so that a common emitter amplifier can be operated without neutralization. But in this case the stage cannot be designed for maximum gain and still be stable. Neutralization permits operation with maximum gain; however, circuit adjustment will be more critical and transistor interchangeability will be poorer. Nevertheless, one neutralized stage is roughly equivalent to two unneutralized stages.

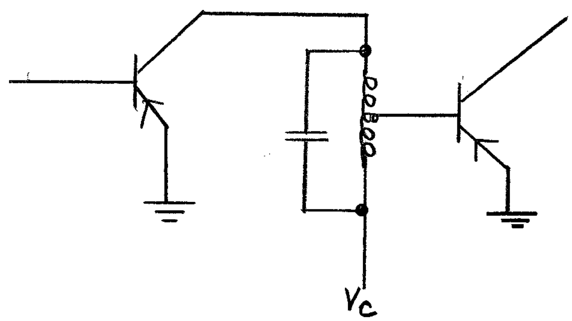
Figure 6.39 illustrates the difference between neutralized and unneutralized r-f amplifiers. In the neutralized stage, double-tuned transformers are used. Moreover, the low-impedance input (base) of the transistor as well as the collector are tapped down on the tuned circuits to prevent loading which would reduce effective Q. The primary winding in the collector circuit is tapped; and a signal that is 180° out of phase with the collector voltage is fed back to the base through the neutralizing capacitor (C_n). This signal cancels that which is fed back to the base through the collector-junction capacitance. Because of variations in parameters from unit to unit of the same type, this circuit may become unstable with a particular transistor requiring readjustment of the neutralizing capacitance.

In the unneutralized stage (Figure 6.39b), single-tuned transformers are used; and the transformers are not tapped for optimum performance. This reduces stage gain, but at the same time it permits operation without neutralization. This, however, requires careful design to achieve a certain degree of stability with a minimum loss of gain.

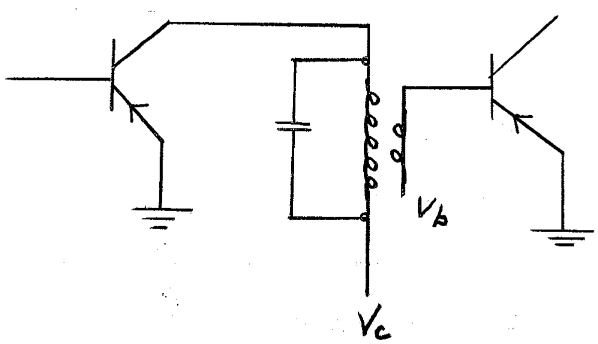
As was mentioned before, the common-base connection is used at frequencies near the alpha cutoff frequency of the transistor



a. TAPPED CAPACITOR

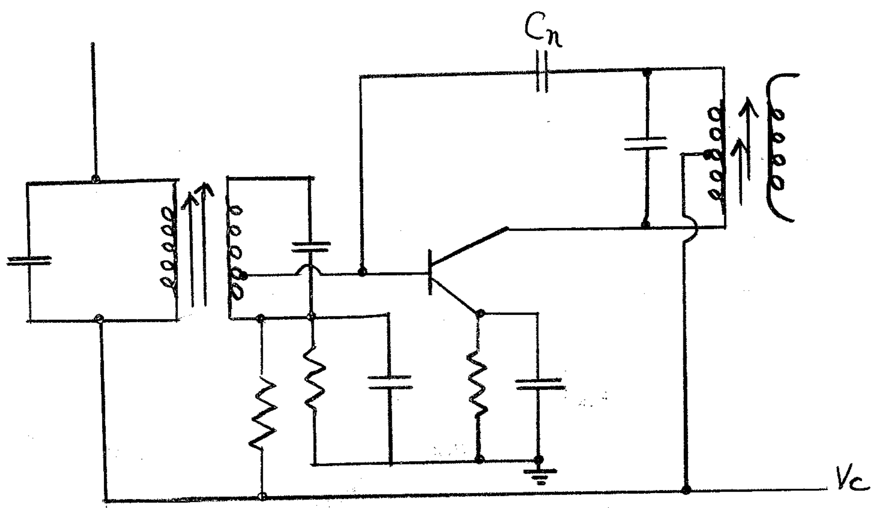


b. TAPPED INDUCTOR

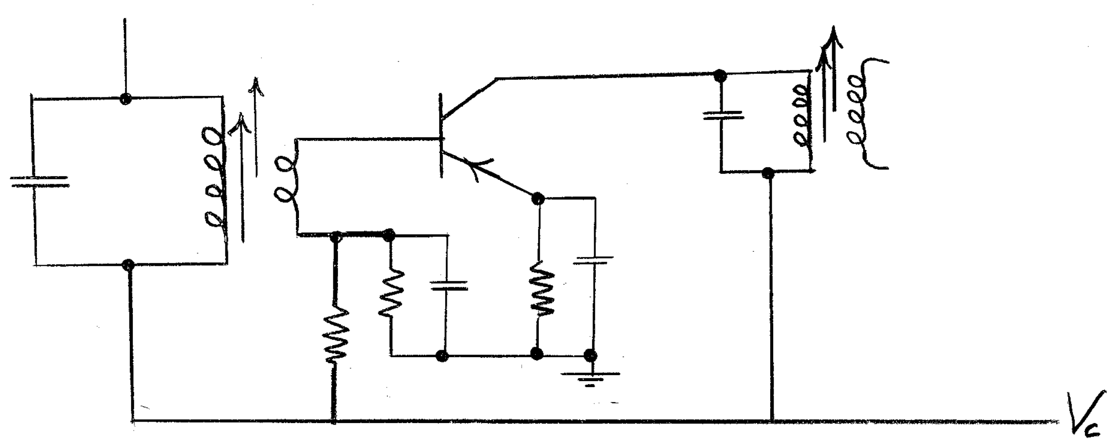


C. SINGLE-TUNED TRANSFORMER

Figure 6.38. Interstage Impedance Matching for Tuned Amplifier.



a. NEUTRALIZED AMPLIFIER



b. UNNEUTRALIZED AMPLIFIER

Figure 6.39. Tuned R. F. Amplifiers.

where the difference between common-emitter and common-base power gain becomes small. A common-base V.H.F. amplifier is shown in Figure 6.40. The circuit is straightforward, and it infrequently requires neutralization. Careful design should be used to conserve all possible gain because the transistor is operating close to its maximum frequency, and the available gain is low.

In most r-f receiving equipment, it is desirable to incorporate some form of automatic gain control (a.g.c.). The easiest way of accomplishing this is to vary the transistor bias, but there are many effects that must be considered. If the collector current is reduced below some optimum value, the power gain will be reduced. This can be accomplished by changing the d.c. bias. However, a change in bias will also effect the input and output impedances of the device. If the circuit is adjusted so that all the impedances are matched for maximum gain, this impedance shift will also contribute to the gain reduction which will add to the a.g.c. control of the stage. Care must be taken so that the change in impedance does not cause instability.

Conclusions. The previous section showed that high-frequency operation is more dependent on transistor characteristics than it is on circuitry. The circuit used for a particular application does not change greatly as the frequency is advanced. However, at higher frequencies, more care must be taken in insuring d.c. stability of the transistor. Steps taken to increase the operating frequency of a transistor usually decrease the maximum power dissipation so, even in low-power circuits, transistors are operating near this limit.

OSCILLATORS

An oscillator, as referred to here, is an electronic device that generates an a.c. signal of some kind. Oscillators will be divided into two general categories: sinusoidal oscillators which generate a single-frequency sine wave and nonsinusoidal oscillators, such as blocking oscillators, multivibrators, and sawtooth generators, which generate some particular nonsinusoidal waveform.

Sine-Wave Oscillators. An oscillator contains an amplifier and a positive feedback path of some sort. A sinusoidal oscillator must also have some frequency-sensitive element - such as a resonant tank, a quartz crystal, or a phase shift network which determines the frequency of oscillation. In order for the oscillator to function, the amplifier must be capable of supplying the power lost in the load and various circuit elements as well as its own input power.

Several oscillators using a resonant tank as a frequency-determining element are shown in Figure 6.41. The first two use an additional winding on the resonant tank to supply the feedback which

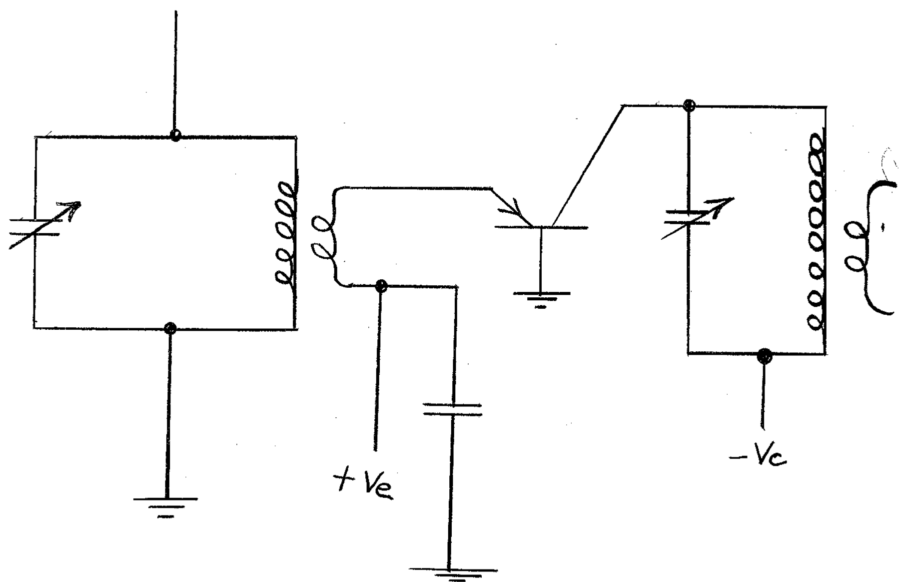


Figure 6.40. Common-Base V.H.F. Amplifier.

maintains oscillation. The remaining circuits use a tapped (either capacitive or inductive) tank to obtain the correct feedback between the input and output.

Figure 6.41a shows a common-emitter amplifier with a resonant tank in the output circuit. Oscillation is maintained by coupling part of this signal back to the transistor input. Figure 6.41b shows a similar circuit except that the transistor is connected in the common-base configuration. In Figure 6.41c that portion of the tank between the emitter and ground acts as the emitter load for the transistor while that portion between the emitter and base supplies the transistor drive. Figure 6.41d is identical to Figure 6.41c except that a capacitive tap is used on the tank. The circuits in Figure 6.41e and f are identical except that a capacitive tap is used on the former and an inductive tap is used on the latter. That portion of the output voltage across the lower portion of the tank appears between the emitter and the base to supply transistor drive.

The a.c. performance of all these circuits is the same. The only reason that one circuit would be preferred over another would be to simplify the d.c. circuitry or to put one particular transistor terminal at a.c. ground (e.g. whichever terminal is connected to the transistor case).

The circuits shown in Figure 6.41 are not self starting. When power is applied to the circuits no current will flow because the transistors are cut off. A small emitter to base bias must be applied to start operation. Figure 6.42 illustrates a method of supplying a starting bias to the circuit of Figure 6.41b. The voltage dropped across R_2 appears between the emitter and base of the transistor to provide forward bias. If the bleeder current through R_1 is made greater than the normal base current, this bias will remain after oscillation has started; and it can be adjusted for class A operation of the oscillator. If the bleeder current is of the same magnitude as the base current which is in the opposite direction, there will be very little voltage dropped across R_2; and the operation will be essentially class B. If the bleeder current is made still smaller and if R_2 is sufficiently large, the base current will charge the capacitor and the emitter-base junction will become reverse biased, permitting current flow only at the peak of the input cycle. Even higher values of R_1 and R_2 will cause blocking or "squeegeeing" which will result in intermittent operation or pulsing of the oscillator.

At frequencies above the alpha cutoff frequency a transistor still has some gain. This gain is low so the transistor is not too useful as an amplifier, but it can still function as an oscillator. Generally speaking, a transistor can be made to oscillate considerably above its cutoff frequency.

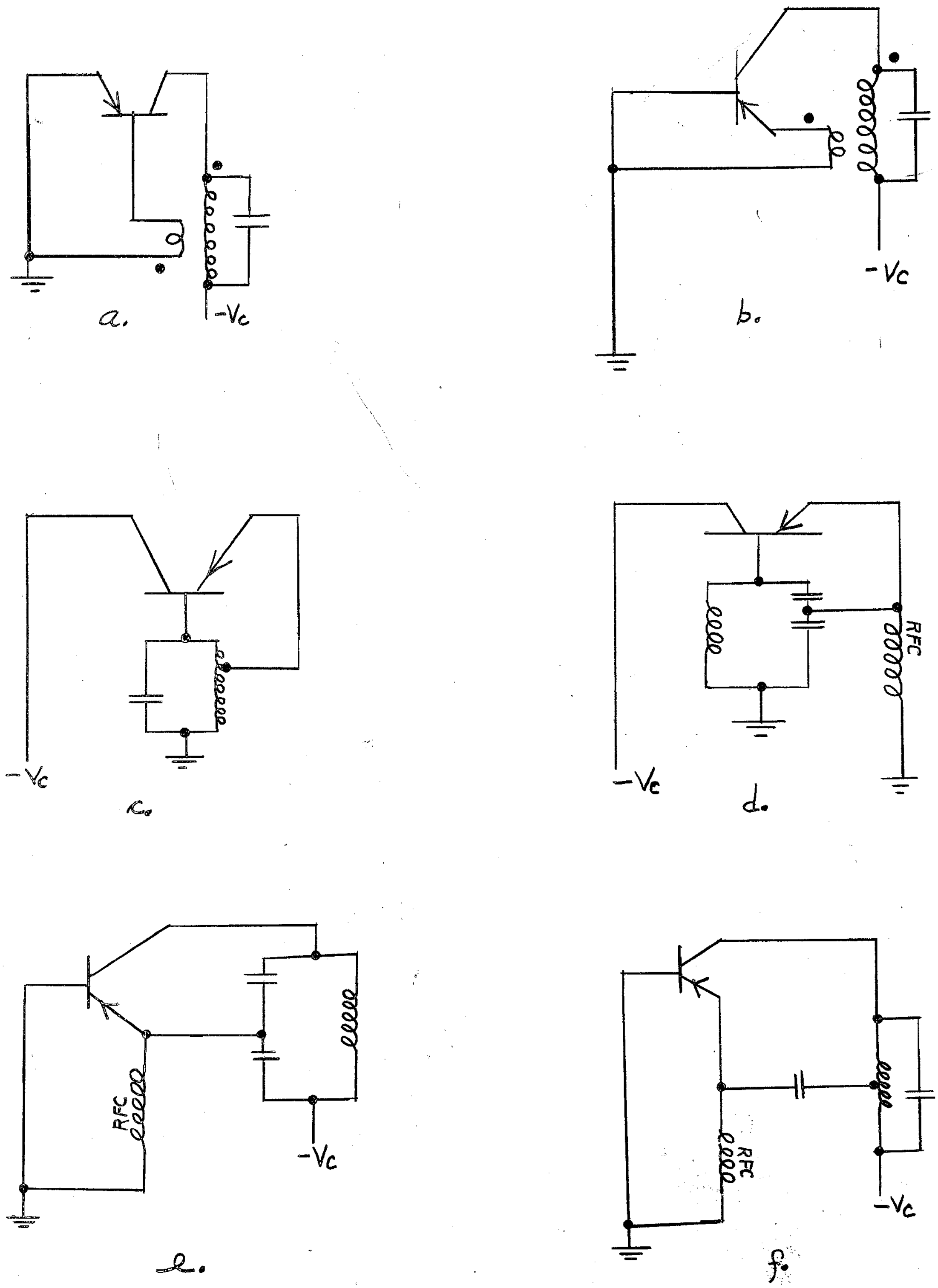


Figure 6.41. Several Transistor Oscillator Circuits.

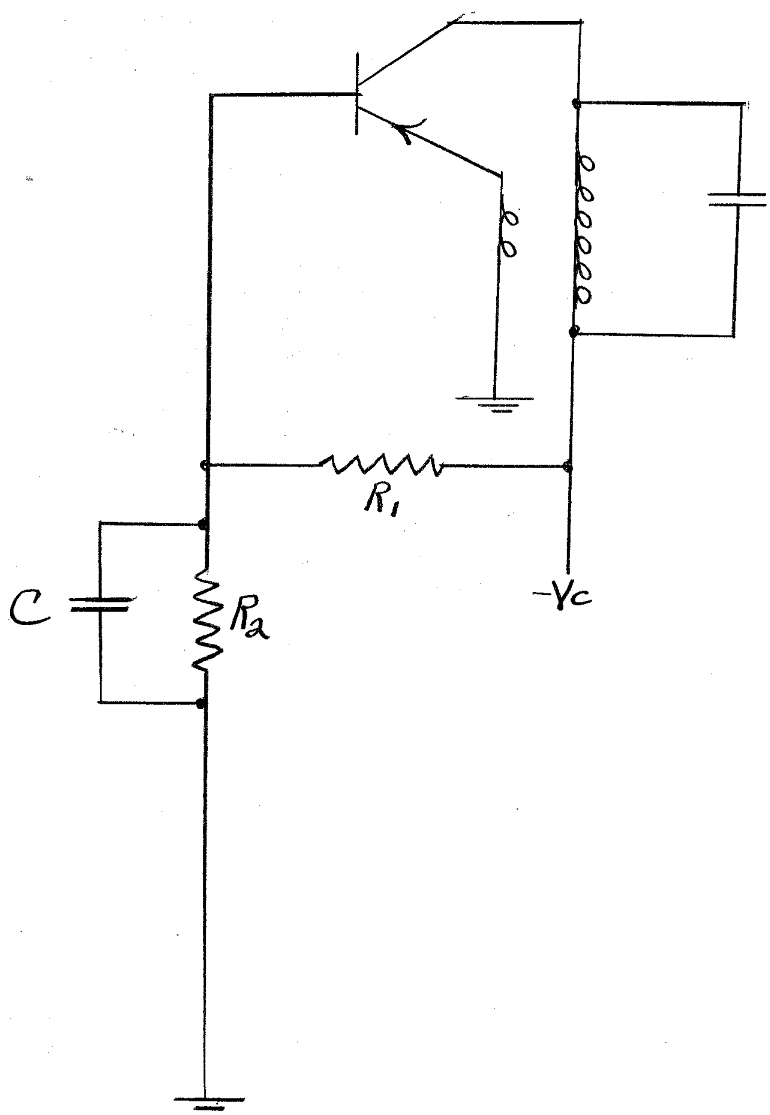


Figure 6.42. Oscillator Biasing Circuit.

Even at moderately high frequencies the phase shift within a transistor due to transit time effects becomes appreciable. This phase shift alters the frequency of oscillation slightly. Because the transit time will vary with temperature, collector voltage, etc., frequency instability may be encountered with transistors operating at higher frequencies.

There are no special requirements for high-frequency operation. However, because available power gain is usually low, particular care should be taken to match impedances and minimize circuit losses. Near the frequency limit of the transistor the emitter input may appear inductive. It is advisable to tune this out. One method is shown in Figure 6.43.

Sine-wave oscillators can be built using only resistance-capacitance networks. An example is the phase-shift oscillators in Figure 6.44. In the circuit of Figure 6.44a, the output of Q_1 is fed through a r-c network which shifts the phase 180°. The 180° phase shift will occur at only one frequency so the oscillations will be sinusoidal as long as the feedback voltage does not become so great as to overdrive Q_1. The resistance of R_1 should be low compared to the input impedance of the transistor so that changes in the latter quantity do not affect the oscillation frequency. The emitter resistor, R_4, is left unbypassed to reduce distortion in Q_1 and to increase the apparent input impedance of the transistor.

Because of the low impedance of the r-c network needed to obtain frequency stability, the output circuit may become so heavily loaded that oscillation is not possible. Hence, a common collector stage is sometimes added to drive the low impedance network as shown in Figure 6.44b. This stage has no phase shift so it does not affect the operating frequency. Q_2 does reduce the collector load on Q_1 and so it increases the gain of this stage.

The phase-shift oscillator is only useful at low frequencies where the gain of the transistors is high. The feedback circuit is lossy and does not provide impedance matching so it will not operate unless there is enough gain available.

Nonsinusoidal Oscillators. A nonsinusoidal oscillator, as the name implies, is a generator of nonsinusoidal a.c. voltages. This class includes pulse generators, square-wave generators, sawtooth generators, etc. The period, or basic frequency, of these circuits is usually determined by the time constants of R-C and/or R-l circuits rather than by resonant elements.

Blocking Oscillators. The circuit of a transistor blocking oscillator along with pertinent waveforms is shown in Figure 6.45. A pulse transformer is used to obtain feedback from the transistor

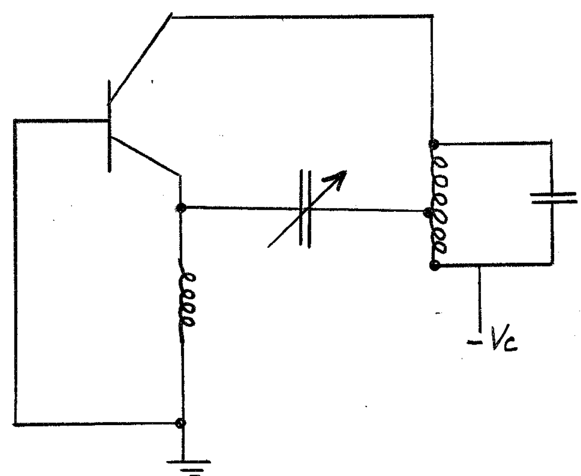


Figure 6.43. High-Frequency Oscillator.

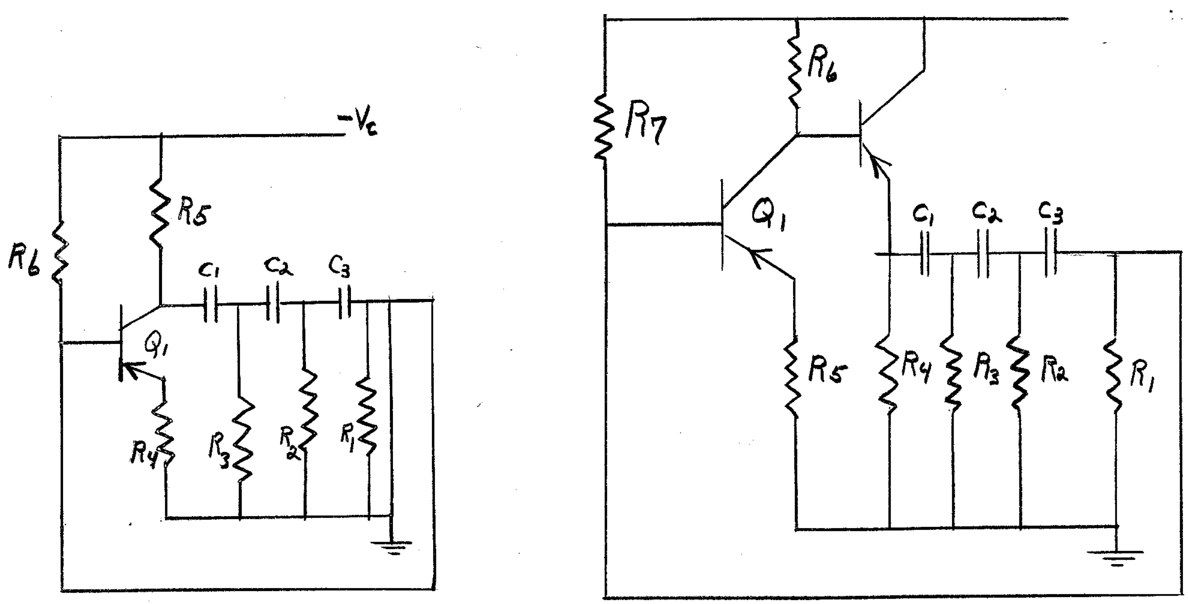


Figure 6.44. Phase-Shift Oscillator Circuits.

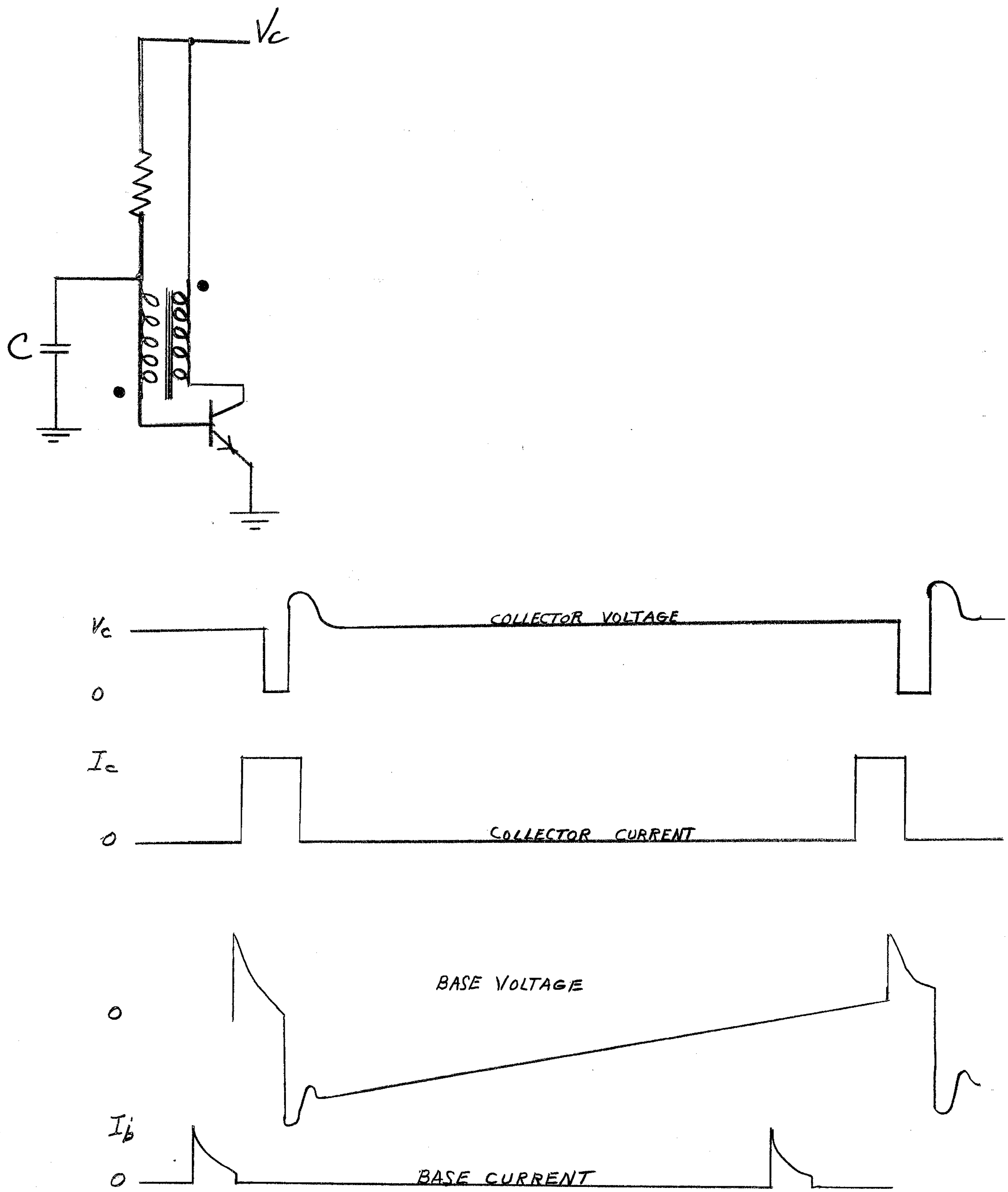


Figure 6.45. Circuit and Waveforms of a Transistor Blocking Oscillator.

output (collector) circuit to the input (base) circuit. A r.c. combination is used to determine the basic operating frequency (repetition rate). Operation is as follows:

When power is applied, the forward bias voltage on the transistor through R will cause conduction. The increase of collector current through the transformer primary will induce a voltage in the secondary in such a direction as to increase the transistor drive. This process is cumulative so the transistor rapidly switches into saturation.

Assuming that the collector current is constant for the duration of the output pulse - although it may either increase or decrease depending on the mode of operation - the phenomenon responsible for maintaining an induced secondary voltage during the "on" portion of the blocking-oscillator operation will be investigated.

After the transistor has been switched to the "on" condition, the large base-drive current will begin to charge the capacitor. When this happens, the base current will decrease because more and more of the secondary voltage is being dropped across the capacitor leaving less available for base drive. This action does not affect the collector current because the transistor is driven far into saturation and the input circuit no longer has control of collector current. The flux setup in the transformer core due to secondary current opposes the flux established by the primary current so the decreasing secondary current causes an increase in net core flux, which maintains the secondary voltage. Hence, the transistor will conduct until the base drive voltage has dropped to the point where the input circuit regains control of the collector current. At this point the decreasing base voltage decreases the collector current which decreases the base voltage further through the action of the transformer. Again the process is cumulative and the transistor quickly switches off.

After conduction has ceased, the charge stored in the capacitor holds the transistor in the nonconducting state until this charge is drained off through the resistor, R. When this happens, the transistor will begin conducting, and the cycle will repeat.

During the "on" portion of the cycle, energy is stored in the magnetic field of the transformer. When the transistor is switched off, this energy is dissipated by the "inductive kickback" voltage as shown in the figure. If there is no damping resistance across one of the windings, this inductive voltage swing will become excessive.

One particularly effective way of reducing this voltage to a negligible value is the use of a diode across one winding. The diode is nonconducting on the useful portion of the output pulse, but it conducts on the backswing rapidly dissipating the energy of the magnetic field in its forward resistance.

Sawtooth Generators. The blocking oscillator can be conveniently used in a sawtooth generator circuit as shown in Figure 6.46. The large emitter current flowing during the conducting portion of the blocking oscillator cycle is used to charge a capacitor. After operation of the blocking oscillator ceases, the capacitor discharges slowly through a large resistor producing a nearly-linear sawtooth. When the capacitor discharges to the point where the emitter becomes positive with respect to the base, the transistor again conducts; and the cycle repeats.

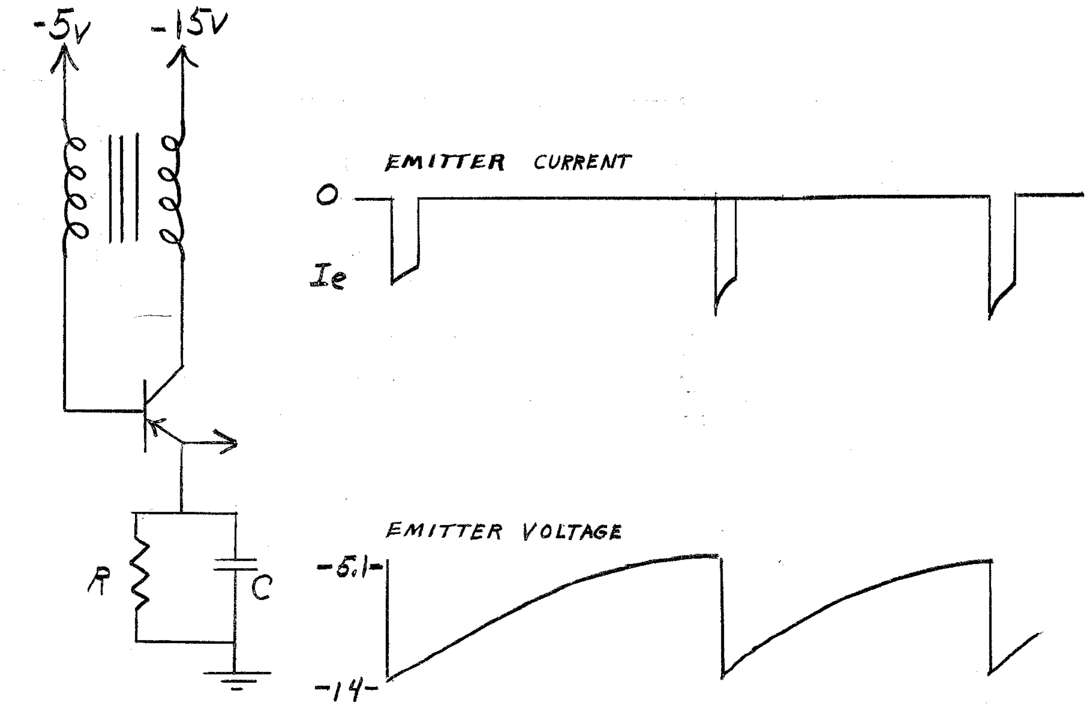
A step-up ratio in the blocking oscillator transformer of about 3 to 1 provides adequate current drive while producing enough secondary voltage to allow the capacitor to charge nearly to the supply voltage. The capacitor discharges until the emitter voltage drops to about 0.1 volt less than the fixed, base supply voltage. When this happens, the transistor will become forward biased; and the circuit will recycle. The sawtooth can be made nearly linear by adjusting the fixed base voltage so that only a small portion of the discharge is used, or by returning the discharging resistor to a positive voltage rather than to ground.

Another scheme used for linearizing the sawtooth is shown in Figure 6.46b. In this circuit the constant-current characteristics of the common-base configuration is used to provide a constant discharge current for the capacitor.

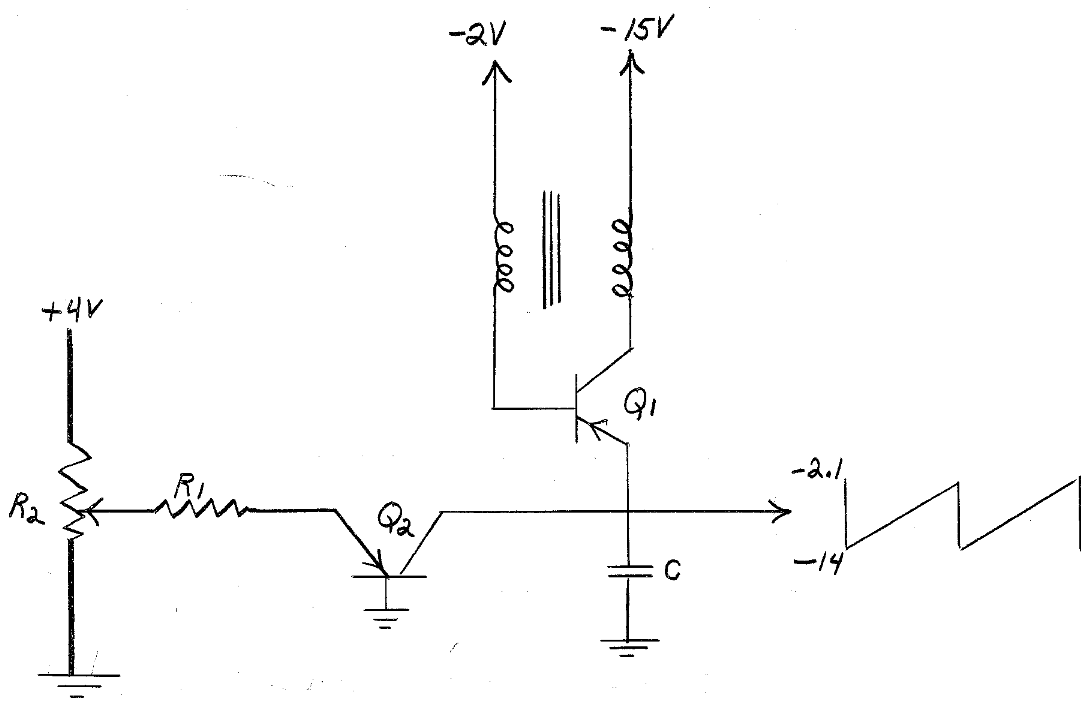
The transistor, Q_2, is forward biased by the application of a positive voltage to the emitter. The current flowing will be practically independent of transistor parameters because the resistor, R_1, is made large in comparison to the emitter-base impedance of Q_2. The resulting collector current will be equal to the emitter current times the d.c. alpha current gain of the transistor which is always very close to unity. Since the collector current of Q_2 is used to discharge the timing capacitor, this discharge will be extremely linear. The period of the output wave can be adjusted with R_2 which varies the emitter current of Q_2; and, therefore, the discharging current of the capacitor.

It is also possible to build a sawtooth generator using negative-resistance devices such as the avalanche transistor or the PNPN diode or transistor; but these schemes are generally less satisfactory because the devices mentioned are not, as a rule, capable of supplying the large capacitor discharge currents. At any rate, this will be discussed further in the section on special devices.

Free-Running Multivibrators. The output of a free running multivibrator is normally an asymmetrical rectangular wave. Such a circuit is shown in Figure 6.47. During operation, the transistors, Q_1 and Q_2, are alternately driven into conduction by a regenerative switching



a. Basic Sawtooth Generator.



b. Linear-Sawtooth Generator.

Figure 6.46. Sawtooth Generator Circuits.

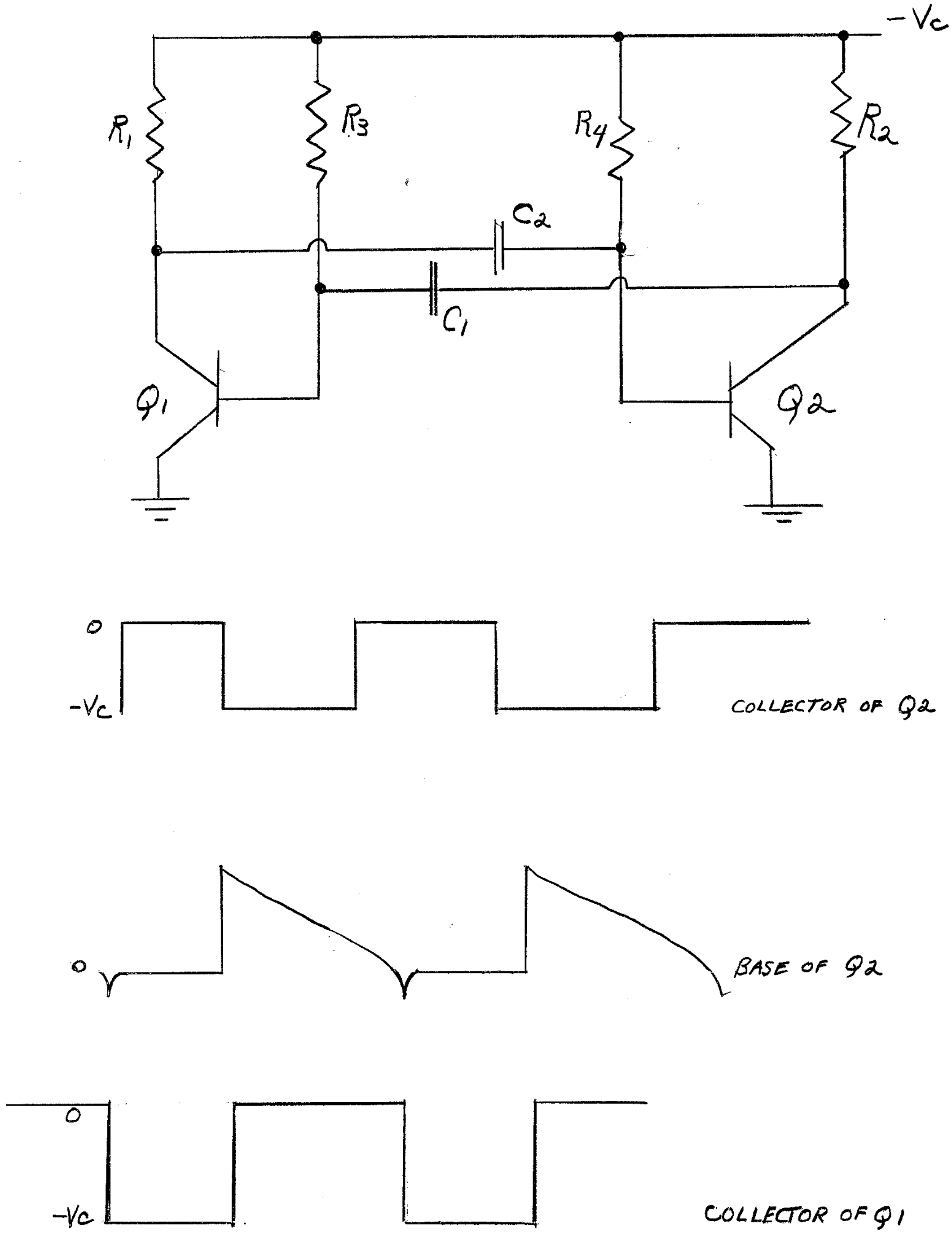


Figure 6.47. Asymmetrical Multivibrator Including Waveshapes.

action. The period and asymmetry of the output wave form is controlled by adjusting the time constants of the R-C circuits in the feedback network.

Assuming that the transistor Q_1 is conducting the action is regenerative so Q_1 rapidly switches into cutoff and Q_2 switches into conduction. This condition will prevail until C_1 discharges to the negative supply voltage through R_3, and Q_1 again begins conduction.

When power is first applied to the circuit, one of the transistors will conduct harder because of a circuit dissimilarity or a random noise voltage and this transistor will switch into conduction. Therefore, it can be seen that the circuit will be self starting.

The conducting time of each transistor can be adjusted independently by controlling the time that the other transistor is cut off. This factor is solely dependent on the R-C time constant in the transistor base circuit. Therefore, even though the components have the same nominal value, the output of this oscillator is not, generally, a symmetrical square wave.

The Monostable Multivibrator. The monostable and the bistable multivibrators cannot be considered oscillators, in any strict sense, because they are not free running. Nevertheless, their similarity to the free-running variety justifies a discussion of these circuits here.

The monostable, or one-shot, multivibrator will be discussed first. This circuit is designed to produce a rectangular output pulse of known duration when excited by a trigger pulse. (The shape of the input pulse as long as its amplitude, is sufficient). Because of the many schemes available for direct-coupling transistors, there are many possible monostable multivibrator circuits. Several are given in Figure 6.48.

In the first circuit a two-stage, direct-coupled amplifier is used. In the stable condition, both transistors are conducting, and the collector voltage on Q_2 is low. A positive input pulse will reduce conduction of Q_1 which, in turn, will reduce conduction of Q_2. The collector voltage of Q_2 will then rise, and this positive-going change will be coupled through the feedback capacitor, C_f, to the base of Q_1 reducing its conduction further. Thus, both transistors are regeneratively switched off. This condition will prevail until C_f charges through R_1, and the base voltage of Q_1 becomes negative which will cause Q_1 and, therefore, Q_2 to conduct. The collector voltage on Q_2 will drop, and this negative-going drop will be coupled back to the base of Q_1 switching both transistors back into conduction until the arrival of another positive input pulse.

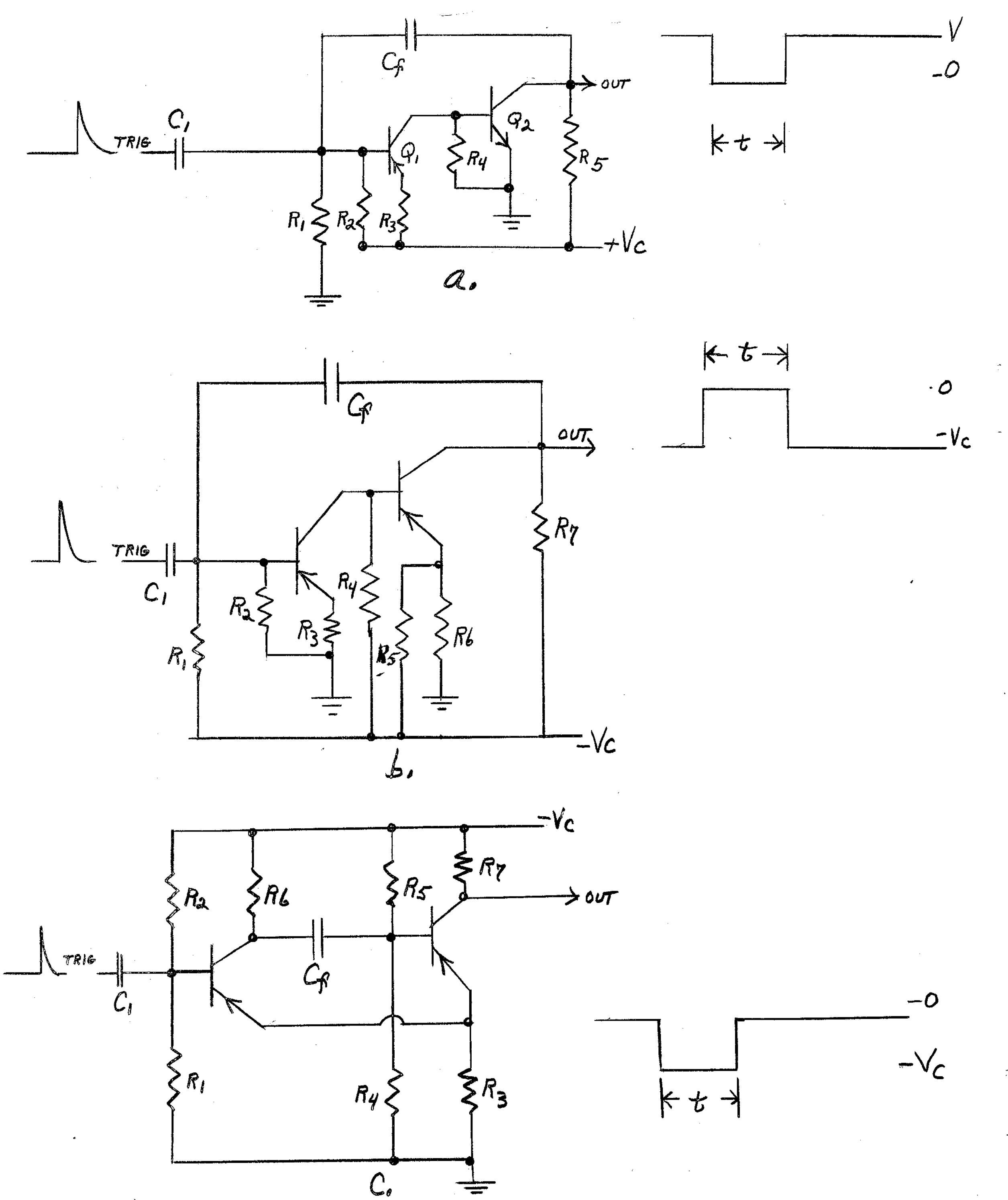


Figure 6.48. One-Shot Multivibrator Circuits.

The "on" state will only be stable if one of the transistors is in saturation. Then a definite trigger amplitude is required for operation, and triggering from random noise voltages is prevented.

The circuit in Figure 6.48b is somewhat different. In the stable state, Q_1 is conducting so that its collector voltage is less negative than the emitter voltage of Q_2. Therefore, Q_2 will be nonconducting. When a positive trigger pulse arrives, Q_1 conducts less so its collector voltage and, consequently, the base voltage of Q_2 becomes more negative. Q_2 will then begin to conduct, and its collector voltage will drop. This positive going change will be felt back on the base of Q_1 through C_f, re-enforcing the input signal and cutting off Q_1. Q_2 then goes into full conduction. The circuit will remain in this unstable state until C_f discharges through R_1 and R_2. The multivibrator will then switch back to the stable state, and wait for another positive input pulse.

The circuit shown in Figure 6.48c is another variation. The bias resistors are adjusted so that Q_2 is conducting and Q_1 is nonconducting. When a trigger is applied, Q_1 will conduct; and a signal will be coupled through C_f cutting off Q_2. The feedback loop is completed by the resistor, R_3, which is common to the emitters of both transistors. When Q_2 cuts off, the voltage across R_3 will drop allowing Q_1 to switch into the conducting state. When C_f charges, the circuit will recycle.

<u>The Bistable Multivibrator</u>. The bistable multivibrator, or "flip flop", will remain in either of two stable states until excited by a trigger pulse. There is no internal timing in this circuit, and output is entirely dependent on the input pulse rate. The schematic of a bistable multivibrator is given in Figure 6.49. It will be shown that the two stable states are Q_1 conducting with Q_2 cut off and Q_1 cut off. Moreover, it will be seen that the multivibrator can be switched between these states with a trigger pulse properly applied to the device.

If Q_1 is conducting, the emitters of both transistors will be at some negative voltage because of the drop in resistor, R_7. Furthermore, the collector of Q_1 will be at nearly the same potential as its emitter. The base of Q_2 will then be at some potential less negative than its emitter because it is being fed through a voltage divider from the collector of Q_1. Therefore, Q_2 will be nonconducting and the collector potential of Q_2 will be nearly equal to the supply voltage. It is this high collector voltage on Q_2 that supplies the conducting bias to Q_1 through the voltage divider. The circuit will remain in this state indefinitely because the charging or discharging of capacitors does not determine the bias levels of the transistors.

This situation can be disturbed by impressing a trigger pulse on the circuit in such a way as to cut off Q_1 or drive Q_2 into

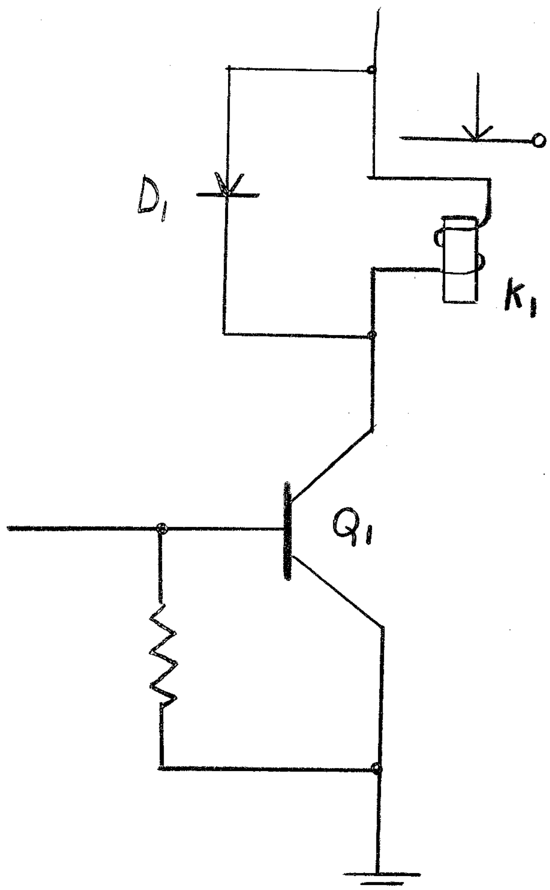


Figure 6.48. Illustrating Use of Protective Diode with Inductive Loads.

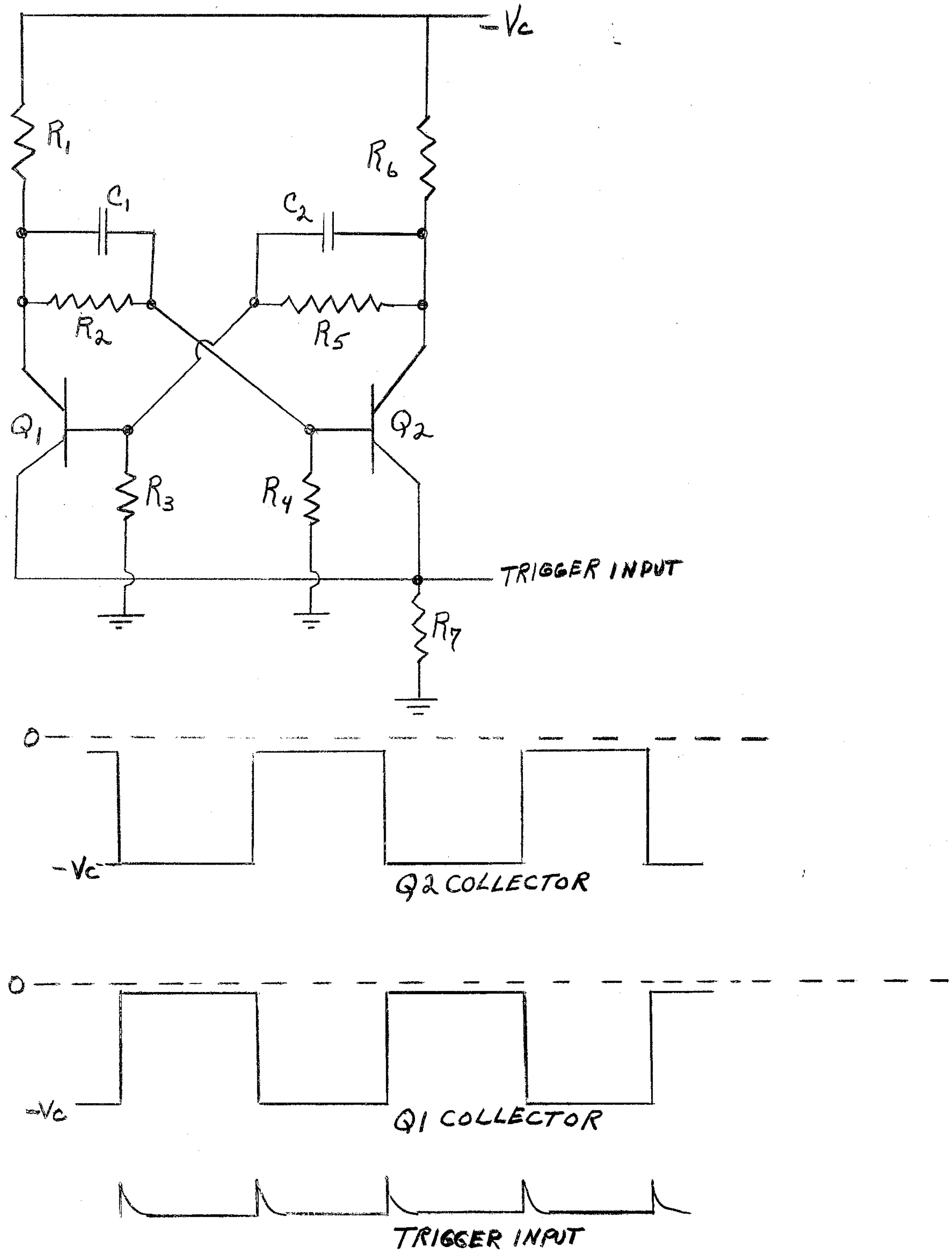


Figure 6.49. Circuit of a Bistable Multivibrator Including Waveshapes.

conduction. One such trigger arrangement is shown in Figure 6.49. Positive trigger pulses are fed in across the common emitter resistor. These positive pulses will not affect Q_1 because it is already conducting. However, the emitter of Q_2 will become sufficiently positive so that it conducts. When this happens, the collector voltage of Q_2 will drop, removing the forward bias on Q_1; and the multivibrator will switch states.

It can be shown in a similar manner that the next positive pulse will switch the "flip flop" back to its original state.

A negative trigger can also be used because it will cut off the conducting transistor causing its collector voltage to rise and drive the other transistor into conduction.

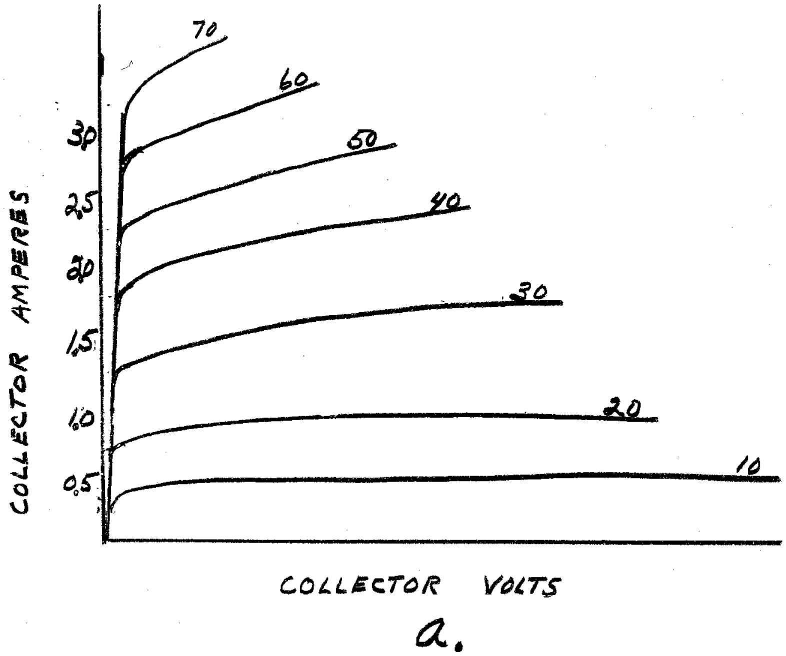
The switching time of the bistable multivibrator is an important consideration as the repetition frequency of the trigger pulses increases. In order to reduce switching time, "speed up" capacitors (C_1 and C_2) are shunted across the base voltage dividers. These capacitors are small so they couple the rapid, switching signal without attenuation while not disturbing the quiescent, d.c. level.

Although the circuit in Figure 6.49 is typical, many variations are possible. For example, the trigger signal may be applied to the transistor base through triggering diodes, or to the collectors in a similar manner.

Moreover, for improved high-frequency performance, special circuits may be employed to prevent saturation of the transistors. This latter detail will be covered in the following section.

Transistor Switches. The transistor makes an efficient switch. When in the conducting state, it presents a very small resistance and can pass large currents with a voltage drop of a few tenths of a volt. In the nonconducting state, it presents a high resistance, and only a small leakage current flows.

When the transistor is saturated, the low resistance of the collector circuit is referred to as the collector saturation resistance. Its value is given by the slope of the collector saturation curves. For the transistor described in Figure 6.46a, this value is in the order of 0.1 ohm. Using an expanded scale for the saturation region shows that the collector-current curves do not merge in the saturation region (Figure 6.46b). This expanded scale more accurately describes the saturation characteristics and is frequently useful in switching circuits where large currents must be passed.



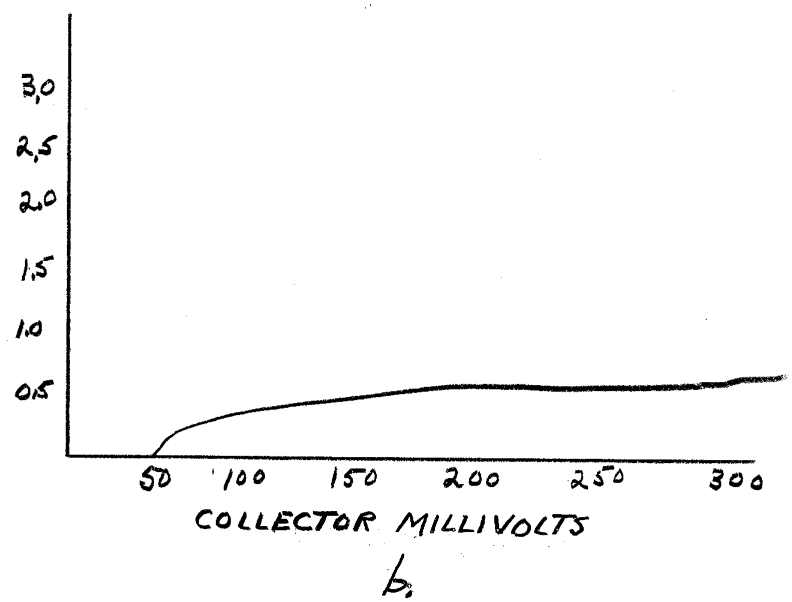


Figure 6.46. Transistor Characteristic Curves Including an Expanded Scale of the Collector Saturation Region.

The "off" current of a transistor can be divided into two components. The first is the result of thermally-generated carriers within the semiconductor. This current is the reverse saturation current of the collector junction. This current will not depend on collector voltage, except at very low voltages when the reverse junction potential is not high enough to sweep across all the carriers or at high reverse potentials where carrier multiplication will occur. The second current component is surface leakage current. This current arises from resistive leakage across the junction due to the presence of contaminating agents on the semiconductor surface. This component, being resistive in nature, does depend on collector voltage. These two components produce the characteristic shown in Figure 6.47.

The variation of collector junction saturation current with temperature can easily be calculated; however, the surface leakage component cannot. Hence, no fixed relationship between cutoff current and temperature can be established; and manufacturers data must be consulted. Nonetheless, in the absence of more accurate data, an approximation can be used: the collector cutoff current will double for every 10°C rise in temperature.

The power dissipation of a transistor in the "on" or "off" state is small. In one case the collector current can be quite large but the collector voltage is small. In the other case the collector voltage is large but the current small. In switching between those states the dissipation will increase; and, even though the time involved is small, this can result in damage to the transistor if it does not have the necessary peak power capacity. This is particularly true when inductive loads are switched off or capacitive loads are switched on. With inductive loads, such as a relay, a diode can be used to absorb the energy in the collapsing magnetic field, thus protecting the transistor from voltage surges and sustained current during switching. This is illustrated in Figure 6.48.

High-Speed Switching. Because of the high frequency limitations of transistors, it is only reasonable to expect that some restriction will be imposed on switching speed. Referring now to Figure 6.49, when an input signal is applied to a transistor, base current flows immediately; but the corresponding increase of collector current is delayed because of the carrier transit time across the base region. After the collector current does begin to rise, some finite time will be required for it to reach its final value both because of the irregular diffusion rate of current carriers crossing the base and because of the collector junction capacitance.

The delay time (t_d) is shown to be the time between the application of the input signal and the point where the collector voltage reaches 10% of its final value. The rise time is the time taken for the collector voltage to go from 10% to 90% of its final value.

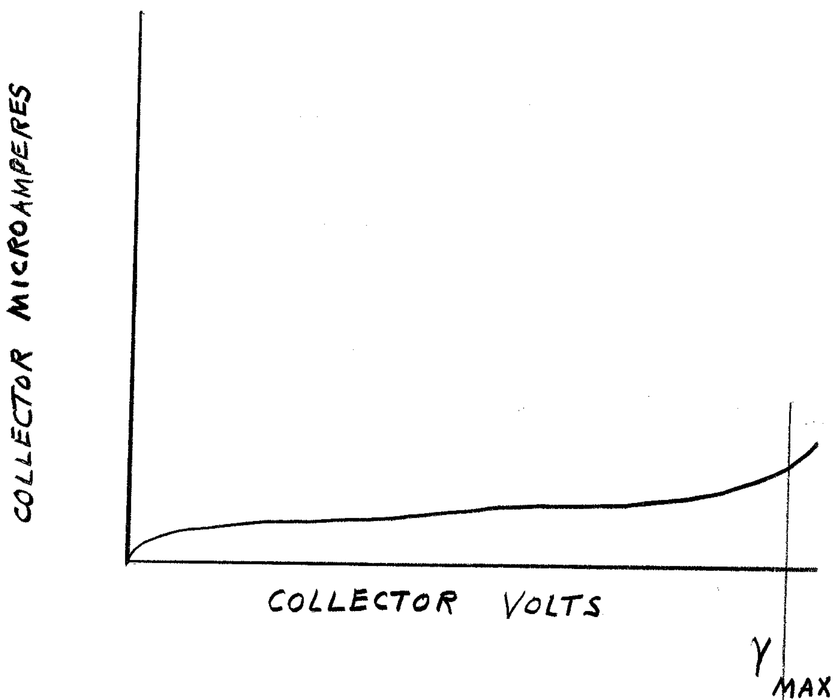
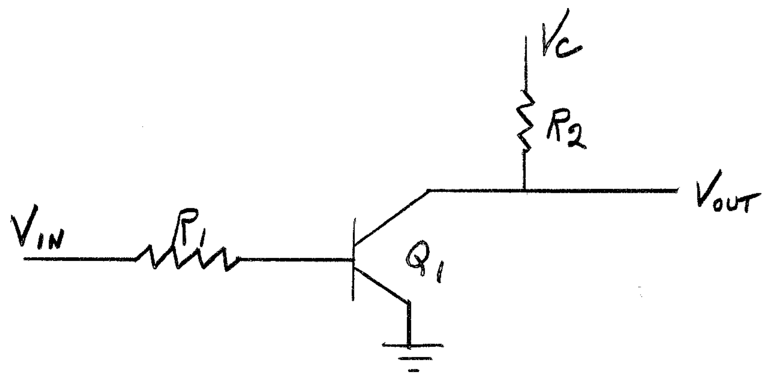


Figure 6.47. Variation of Collector-Cutoff Current with Collector Voltage.



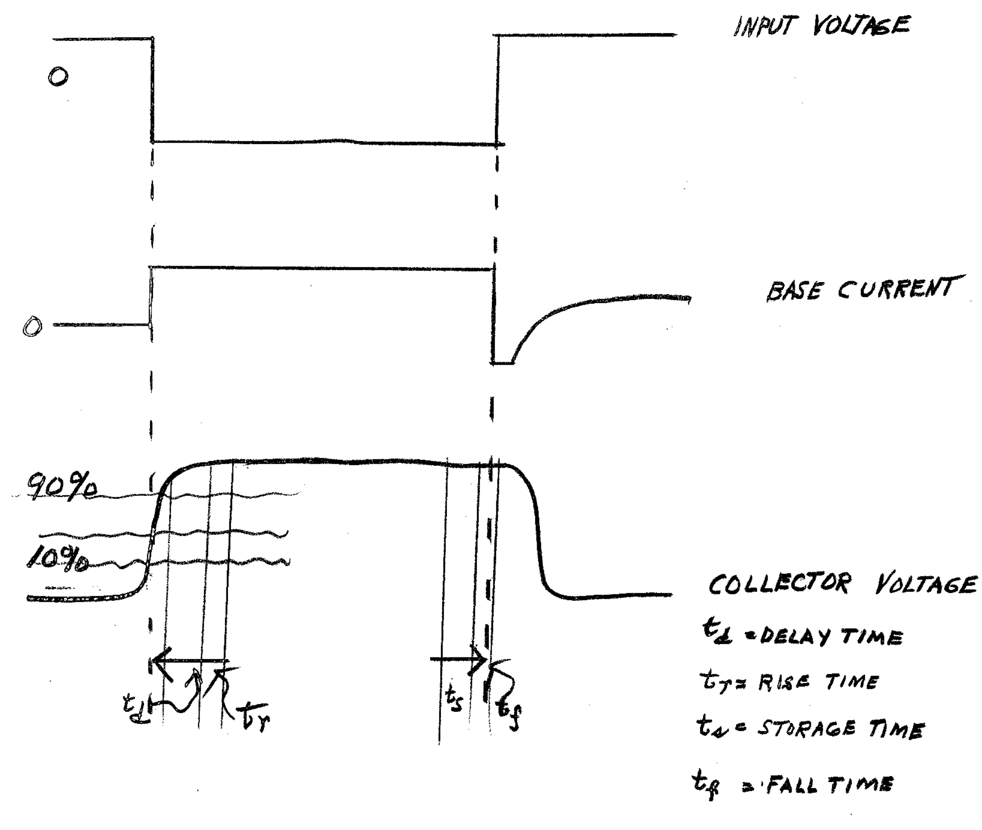


Figure 6.49. Description of Transistor Switching Time.

In this case it is assumed that the transistor is driven into saturation: that is, the input signal is of such a magnitude that the collector voltage drops below the base voltage; and both the emitter and collector junctions become forward biased. When this happens, both the emitter and collector inject minority carriers into the base region. The base will then become charged with excess current carriers as shown in Figure 6.50.

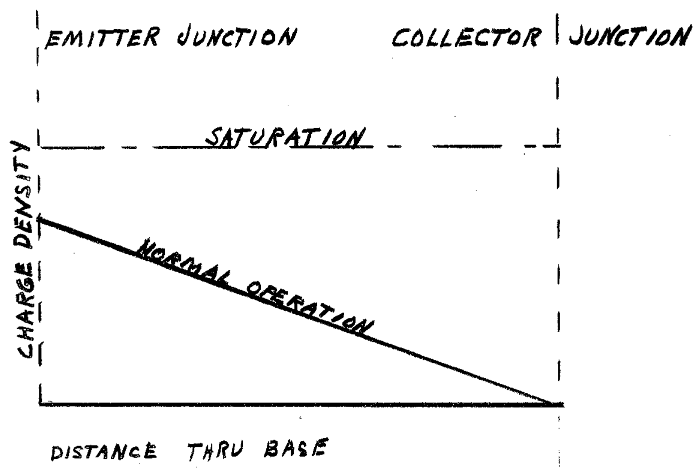


Figure 6.50. Injected Carrier Density in Base Region for Normal and Saturated Operation.

When the turnoff signal arrives, the excess charge must be drained from the base before the collector current can fall which produces a delay in the output waveform. This delay is defined as the storage time (t_s). Because the base is not abruptly cleared of charge and also because of the collector junction capacitance, an additional increment of time (t_f) is required for the transistor to switch off. The fall time is defined in Figure 6.49.

In high-speed switching circuits, the switching time can become appreciable compared to the pulse duration. It is therefore desirable to keep these time lags at a minimum. One method suggests itself almost immediately: if the collector is kept from going into saturation, it is obvious that the storage time can be reduced. The most direct method of accomplishing this is to use a diode clamp to prevent the collector voltage from dropping below a particular value. Such a circuit is illustrated in Figure 6.51a. This circuit suffers from the disadvantage that the collector current in the "on" state can vary over a wide range. When the collector voltage is held at some value by a clamp circuit, the collector current will depend on the base drive current and the alpha current gain. The maximum possible variation of base current drive will depend on circuit design, but the current gain will vary considerably with aging and temperature. A collector current variation of ten to one would not be unusual under these conditions.

A circuit that does not suffer from this limitation is shown in Figure 6.51b. A feedback circuit is used. When the collector voltage drops below some design value, the diode will conduct and prevent any further increase of base current by limiting the voltage at the junction of R_1 and R_2. Hence, when the transistor is in the "on" condition, the collector will be at a higher voltage than the base by the amount of the voltage drop across R_2; and it is impossible for it to go into saturation.

In all the circuits mentioned thus far, the switching time can be reduced by shunting a capacitor across the series, base resistor. This capacitor will provide a large base drive to charge and discharge the base region during switching, but it will not affect the steady-state current values.

Another technique is available for reducing the rise and fall times but at the expense of output amplitude. If the collector is clamped to prevent the collector voltage from reaching the supply voltage when the transistor cuts off, the switching time will be reduced. This is illustrated in Figure 6.52. With the clamp circuit, the collector current will already have started switching by the time the collector voltage begins to change so the slow initial rise does not appear in the voltage waveform. A similar situation is present when the transistor cuts off. The final portion of the charge curve for the collector junction capacitance is eliminated by the clamp circuit thus reducing the fall time.

Figure 6.52 illustrates a typical high speed transistor switch, where D_1 prevents collector saturation, D_2 provides the cutoff clamp of 5 volts below the collector supply voltage, and C_1 gives a switching transient to quickly charge or discharge the base region. The improvement in pulse shape (particularly the reduction of storage time) is evident from the accompanying waveforms.

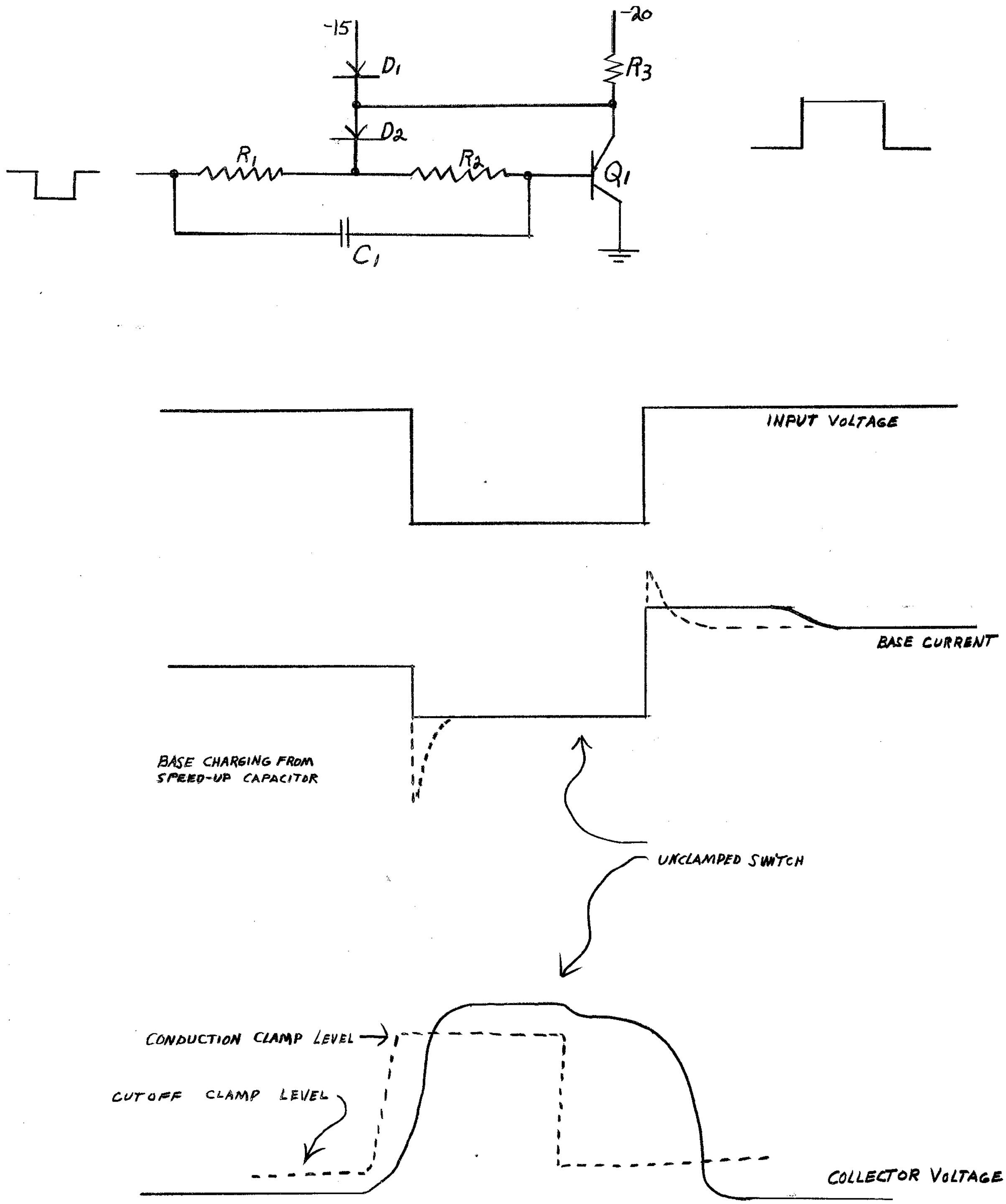


Figure 6.52. High Speed Switching Circuit.