H SNYDER [illegible]

IBM® Customer Engineering

Manual of Instruction

TRANSISTOR

Theory Illustrated

Issued to:

Department or
Branch Office Telephone Number

Address City State

Home Address City State

If this manual is mislaid, please notify the above address.

Printed in U.S.A.
Form 223-6794-0

Contents

DIODE THEORY

TRANSISTOR THEORY

Foreword

BECAUSE of the complexity of transistor theory and the newness of the subject matter, two transistor theory manuals have been prepared. These are:

1. Transistor Theory and Application, Form 223–6783
2. Transistor Theory Illustrated, Form 223–6794 (this manual)

Although both manuals cover transistor theory, they do not cover the same material in the same way. In other words, each manual has a character of its own. Both should be read because they complement each other. Although some information is common to both manuals, this is not true of all information. For instance, material covered in the theory and application manual and not covered here include transistor physics, the grounded emitter circuit, the grounded collector circuit, and specific component circuits. On the other hand, this manual covers some concepts not covered in the other.

The purpose of this manual is to teach transistor theory by the use of illustrations. The illustrations were developed by first carefully analyzing the theoretical concepts and then converting this information to drawing form. These drawings were made to tell much of the story, and should be studied carefully for content. Words were then added to describe in detail these simplified drawings.

The use of drawings to describe transistor theory is used here because this technique found very favorable acceptance by Customer Engineering classes held in Poughkeepsie. These classes found that even difficult concepts were easily understood by the use of simplified drawings. Students also found that they could prepare similar drawings of their own to analyze a question not specifically covered in the text.

Because this manual does not include a transistor physics section, a few definitions of terms that are not specifically defined in the text are presented here.

Conduction band describes the region an electron enters when it leaves orbit; i.e., an electron not attached to an atom is in the conduction band.

Couple is an electron-hole pair generated by energy. The specific energy of concern here is ambient heat. Normal room temperature of 70° F is sufficient to produce couples.

Covalent bonding is the co-sharing of electrons by atoms. This sharing of electrons bonds the atoms together into a crystal structure.

Electron flow is conduction band current. This means that electrons that are "free of orbit" can move or flow through the crystal lattice.

Hole is a broken covalent bond. A hole exists wherever a germanium atom has only three electrons in its valence band instead of four.

Hole flow is the movement of a hole from atom to atom. Hole movement is the result of a shift in location of an orbiting electron. For instance, when an electron in an atom adjacent to a hole is attracted into the hole, it not only fills the hole but also leaves behind a hole in the location it left. Note also that hole flow exists at the valence band level. In other words, the electron that moves into the hole location, to fill it, does not have to enter the conduction band first.

Majority carriers are the current carriers of most abundance in a material. They are electrons in N-type germanium and holes in P-type germanium.

Minority carriers are the current carriers of least abundance in a material. They are electrons in P-type germanium and holes in N-type germanium.

Valence band is the outer band of electrons orbiting about an atom.

Transistor Theory Illustrated

TRANSISTOR theory may be learned more readily than otherwise by first having an understanding of a two-element device, or diode. Therefore, this manual presents first the theory of diodes, as introduction to the theory of the three-element transistors.

DIODE THEORY

Construction

A germanium diode is a rectifier. In an electrical circuit it acts as a low resistance to current flow in one direction, and as a high resistance to current flow in the opposite direction. It is constructed by various methods, although the point contact and the alloyed junction types are the most popular.

The two general types of alloyed junction diodes are the PN and the NP. The PN is made by alloying to a small N-type germanium base a dot of indium (a tri-valent impurity). The alloying process consists of controlling the oven temperature, so that the indium becomes molten and diffuses evenly into the N-type germanium base. The indium joins into the crystalline structure in an area fairly well defined as seen in Figure 1. Within the diffused area, the material has now become P-type in nature, because it is more populated by P-type atoms (indium) than by the N-type atoms (antimony). In other words, within this region both N- and P-type atoms exist, but the concentration of the P-type predominates.

The NP junction diode is made by alloying to a small P-type germanium base a dot of antimony (Figure 2). The antimony atoms diffuse into the P-type base and their population in this region is much greater than that of the P-type atoms. Therefore, the diffused region exhibits an N-type character.

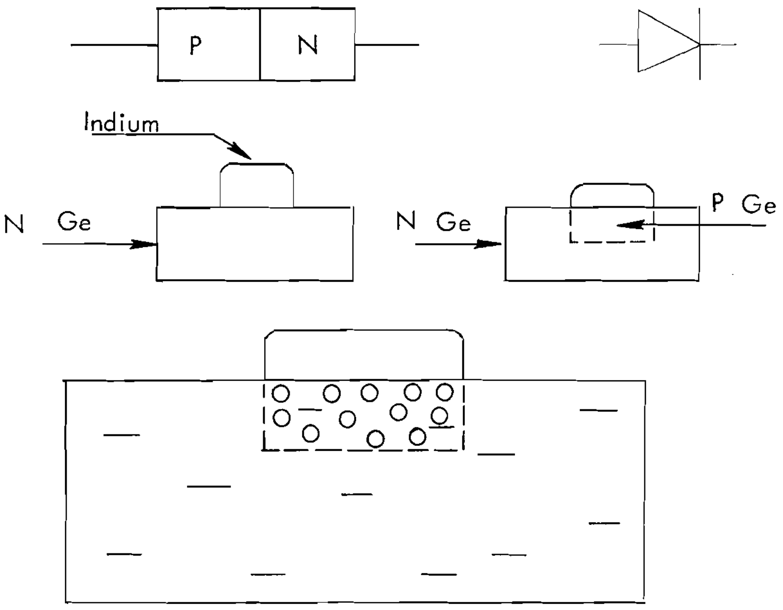


Figure 1. PN Diode Symbols and Alloy Process

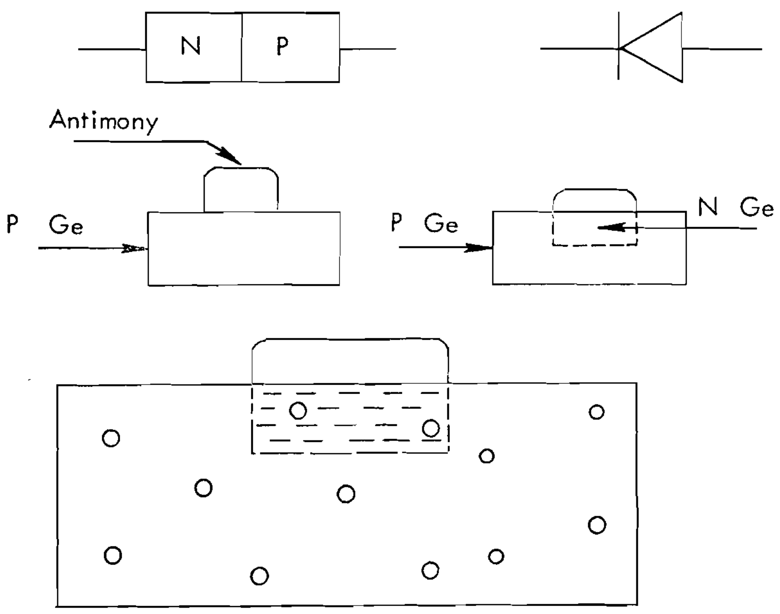


Figure 2. NP Diode Symbols and Alloy Process

The rectifying property of an alloyed junction diode is wholly controlled at the junction of the N and P regions. This junction is an atomic junction; i.e., all atoms are interconnected by covalent bonding. If the junction were not an atomic junction, then it would not have rectifying properties. Thus, highly polished pieces of N and P germanium cannot be clamped together under pressure to form a diode. For, no matter how accurate the polished surfaces are ground, the contact surface has large hills and valleys, and no atomic bond exists.

The construction of a typical point-contact diode is shown in Figure 3. It consists of a fine indium-coated wire (cat whisker) which is formed so that it exhibits a pressure contact on an N germanium base. A high current (usually a capacitor discharge) is then passed through the assembly, which develops a weld at the junction of the wire and the base. At this junction, indium atoms diffuse into the N base and a P region is formed.

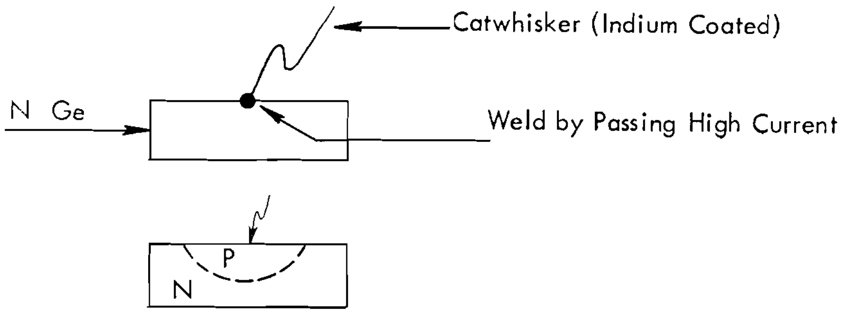


Figure 3. Point-Contact Diode

A comparison of the point contact vs. the junction diode points up the following:

1. The junction is more rugged.
2. The point contact is generally confined to small currents.
3. The electrical characteristics of the two diodes differ.

Because experiments have shown that the alloy junction transistor is more stable and has better over-all circuit gain characteristics than does the point contact transistor, a comprehensive study of the theory of only the junction diode follows.

Formation of Barrier

At the completion of the alloying process, atomic activity takes place at the junction (Figure 4) as follows:

1. A donor electron leaves an antimony atom, crosses the barrier, and unites with a hole generated by an indium atom; i.e., the donor electron finds the broken covalent bond (hole) in the crystal and starts orbiting about this germanium atom (Figure 4a upper). This joining or recombining of a hole and an electron ends the life of both; i.e., after combining, neither the donor electron nor the hole exists. Of course, this annihilation of the donor electron and the hole leaves behind a positive ion in the N region and a negative ion in the P region.
2. A donor electron can cross the barrier and fill the "empty energy level" of an indium atom (Figure 4a lower). Such a transfer produces a positive ion in the N region and a negative ion in the P region.
3. A hole from the P region can cross the barrier and be filled by a donor electron. As in 1 and 2 above, a positive and a negative ion result.

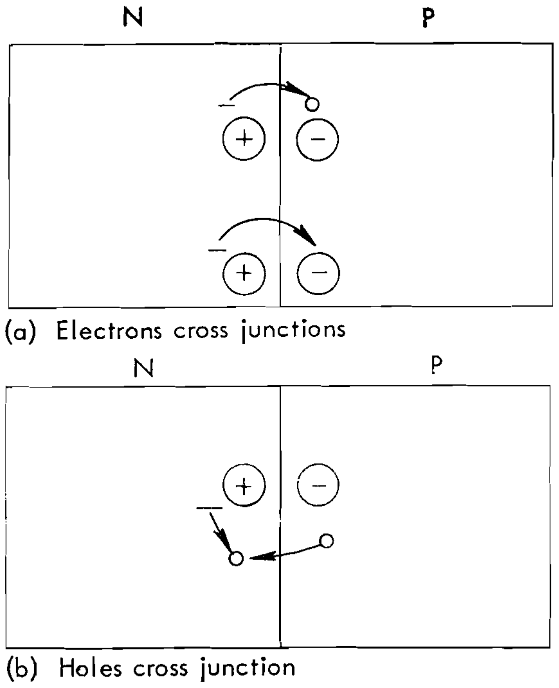


Figure 4. Majority Carrier Transfer across Junction

Stated differently, majority carriers from the N and P regions cross the barrier and recombine. Each recombination leaves at the barrier a positive ion in the N region and a negative ion in the P region. Because this barrier action is the controlling factor in diode action, it is worthwhile to pause here and review this action before going on.

Barrier Potential and Depletion Region

The transfer of majority carriers across the barrier continues until the barrier appears as shown in Figure 5. Notice that an ion barrier has formed and exhibits a small charge called the "barrier potential." Of course, the net charge in the crystal is still zero, but the charge distribution is now such that a potential gradient exists in the crystal at the barrier region. This potential cannot be measured by connecting a voltmeter across the crystal, but its approximate value can be determined by using the diode in a circuit and making a voltage-vs.-current plot of the diode's characteristics. Generally speaking, this value can be considered to be approximately 0.1 volt or less.

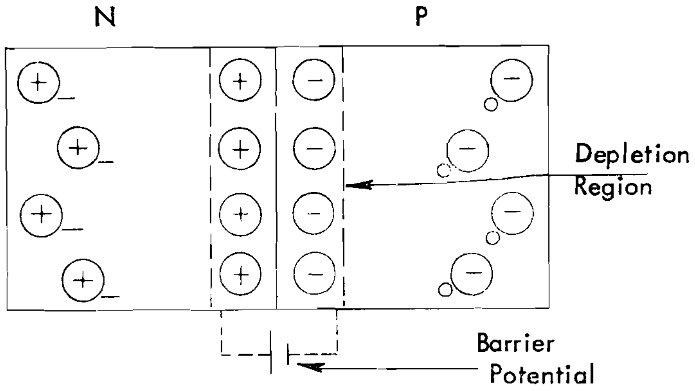


Figure 5. Natural Barrier and Resulting Depletion Region and Barrier Potential

Figure 5 also shows that the ion region (enclosed by dashed lines) is called the "depletion region." A close study shows that majority carriers do not exist in this region. In other words, the region is "depleted" of majority carriers. Notice that on either side of the depletion region the impurity atoms in both the N and P regions are shown counterbalanced by majority carriers. Thus, the diode has a neutral charge distribution except at the barrier.

Figure 6 illustrates the action that takes place at the barrier. Alphabetic notations A, B, and C in the figure identify specific action as follows:

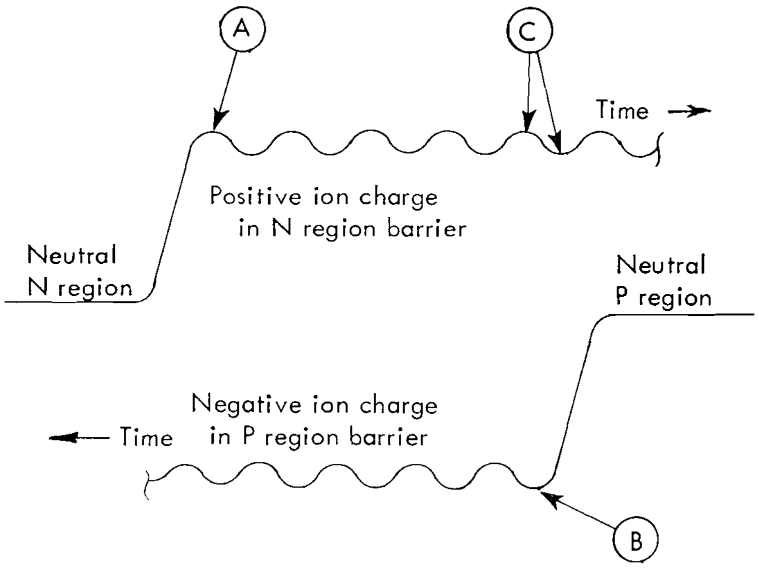

A
C
Time
Positive ion charge in N region barrier
Neutral N region
Neutral P region
Negative ion charge in P region barrier
Time
B

Figure 6. Barrier Activity Oscillates about a Mean

A. A positive ion charge builds up at the barrier until it is sufficiently large to prevent a further transfer of holes from the P region to the N region.

B. A negative ion charge builds up at the barrier until it is sufficiently large to prevent a further transfer of electrons from the N region to the P region.

C. Oscillation of the barrier charge exists about a "mean" charge owing to the barrier activity that always exists. For one thing, some majority carriers enter the depletion region with sufficient energy to cross the barrier. In so doing, the barrier charge becomes sufficiently strong to start attracting back these excessive carriers that sneak across. Also, couples that take place in the barrier region are constantly wandering back and forth across the barrier, causing the barrier charge to oscillate.

Majority carriers adjacent to the depletion region are affected by the barrier charge and tend to shift toward the barrier (Figure 7).

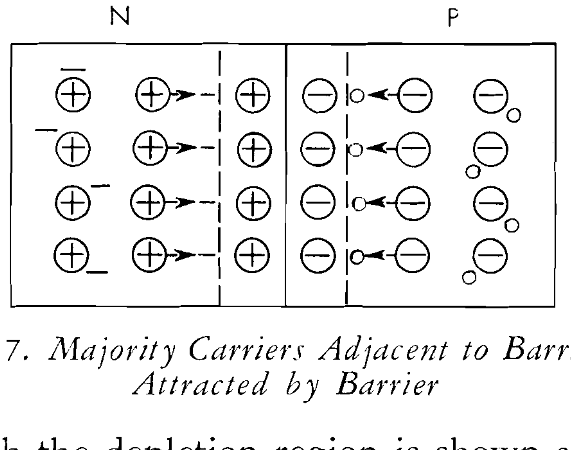


Figure 7. Majority Carriers Adjacent to Barrier Are Attracted by Barrier

Although the depletion region is shown as a sharply defined region in Figure 5, it is in reality a graded region as shown in Figure 8. The maximum electrostatic charge exists at the junction and the charge decreases as the distance from the junction increases. More specifically, the electrostatic charge varies inversely with the distance from the junction.

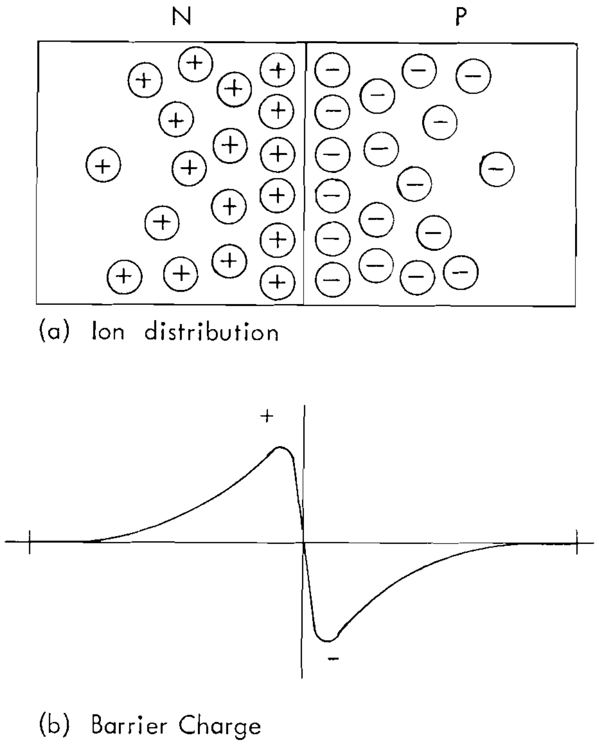


Figure 8. Electrostatic Charge or Potential Hill of an NP Junction

In most alloyed junction diodes, the impurity concentration of the N region is not equal to the impurity concentration of the P region. Thus, the region having the lowest impurity concentration has the widest depletion region, as shown in Figure 9. Although this phenomenon is not significant in diode action, it is of major importance in the study of transistors. In some of the drawings that follow, both regions are drawn with equal concentrations of doping for purposes of simplicity rather than relevancy.

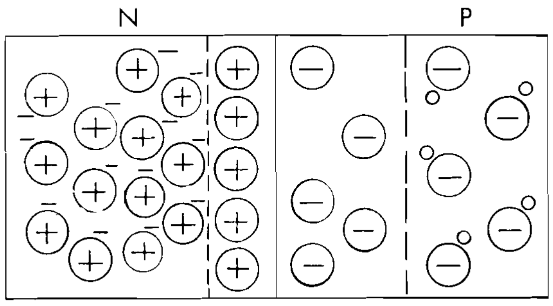


Figure 9. Depletion Width Is Proportional to Concentration

Reverse Bias

Figure 10 shows the normal distribution of charges in a diode before it is connected to a circuit containing a battery source.

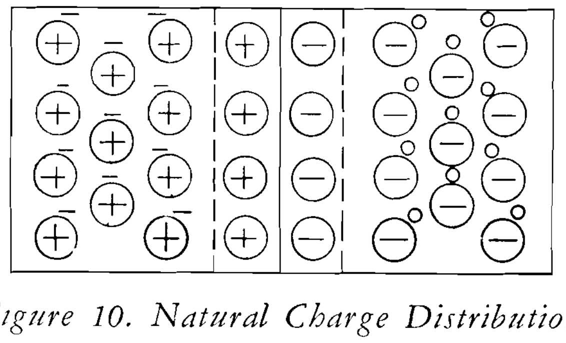


Figure 10. Natural Charge Distribution

A diode exhibits a high resistance by connecting a positive polarity to the N region (reverse bias) and it exhibits a low resistance by connecting a positive polarity to the P region (forward bias). A helpful mnemonic for bias polarity is:

1. Forward bias connects together likes, i.e., the N polarity to the N region and P polarity to the P region.
2. Reverse bias connects together unlikes, i.e., the N polarity to the P region, and the P polarity to the N region.

Figure 11 shows what happens internally to the diode when it is connected in a reverse bias. Donor electrons are attracted by the positive potential and holes are attracted by the negative potential. Thus, charges in both regions are "drawn away" from the junction and the depletion region width increases. This action is similar to a capacitor charge and takes place the instant the battery is connected, after which a steady state condition exists with a wider than normal depletion region. Because the non-depleted N and P regions are neutral, majority carriers in these regions move toward the battery connections until the depletion region is large enough to exert a potential pull equal to the battery. In other words, the capacitive effect ceases when the barrier potential is approximately equal to the battery potential.

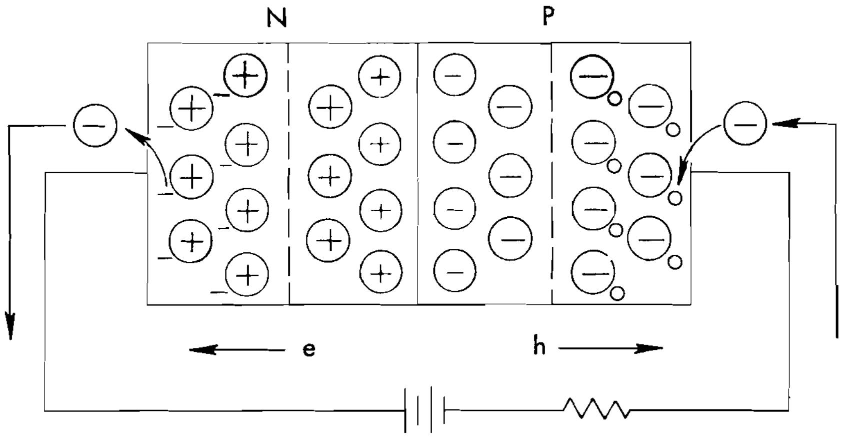


Figure 11. Charge Distribution with Reverse Bias Applied

Although the reverse-bias connection caused majority carriers to move away from the junction, do not reach the erroneous conclusion that a steady-state current does not flow in the external circuit. A small current (generally in μa) does flow as shown in Figure 12. This current is the result of couples (hole and electron pairs) that take place in the barrier region. Couples at the barrier cause minority carriers to exist as follows:

1. Electrons in the P barrier region
2. Holes in the N barrier region

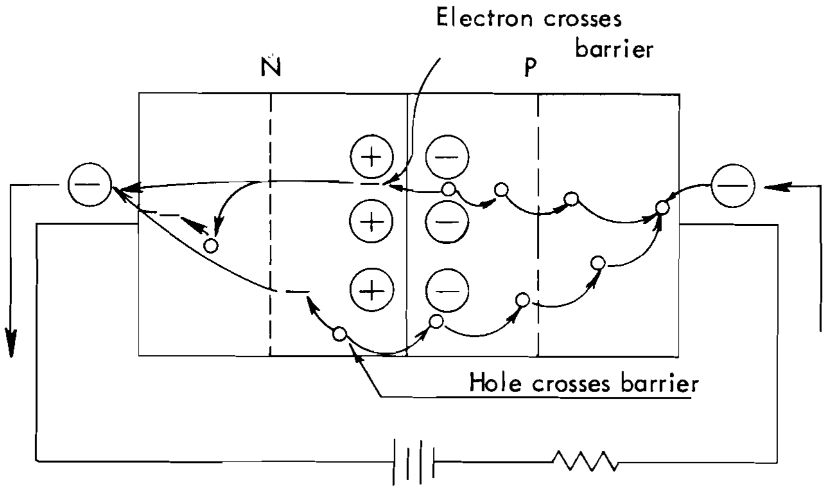


Figure 12. Reverse-Bias Carrier Flow

These minority carriers are forward-biased by the barrier potential; i.e., electrons in the P region are attracted across the junction by the positive ion region and holes in the N region are attracted across the junction by the negative ion region. Minority carriers that cross the junction become majority carriers and are attracted by the battery. The transfer of minority carriers across the junction results in a current flow in the external circuit called "minority carrier current" or "reverse current."

Because minority carrier current flow is an important concept, let us try another approach to understanding it. Study Figure 12 again. The upper current path is obtained as follows:

1. A couple takes place in the P barrier region (because of heat).
2. The electron is attracted by the positive ion region and crosses the junction. It is attracted toward the battery, although it may recombine with a hole on its journey. If a recombination does occur, the electron given up by the couple in the N region is attracted by the battery. Therefore, because an electron crossed the junction to the N region, one electron reaches the positive battery terminal.
3. The hole (caused by the couple) is attracted by the negative battery terminal. It crosses the P region and its life is ended when the excess electron, at the negative battery terminal, flows through the external circuit and recombines with it.

Of course, the above analysis is an oversimplification of carrier action, but it does show the over-all effect. For instance, the electron that crossed the junction is not the same electron that flows in the external circuit. Actually, carrier flow in the non-depleted N region can be compared to the flow in a water pipe; i.e., by putting in some water at one end of a pipe, water is caused to flow out of the other. Likewise, by entering an excess electron in the N region at one end, one electron leaves at the other (same analogy as copper wire).

In Figure 12 the explanation of the lower current path is identical to that for the upper current path, except that, in this case, the hole crosses the junction to the P region, instead of the electron's crossing the junction to the N region.

The over-all effect of a reverse-biased diode can now be stated as follows:

1. The back resistance is due to the barrier resistance; i.e., majority carriers cannot cross the barrier.
2. This back resistance is large and in the order of 100k to 5 megohms, depending on the diode type.
3. The non-depleted N and P regions have resistance values of only a fraction of an ohm and their resistive effects are negligible.

Forward Bias

Figure 13 shows what happens internally to the diode when it is connected in a forward bias. Electrons are repelled by the negative potential and are forced toward the junction. Holes are repelled by the positive potential and are forced toward the junction. Thus, charges in both regions travel to the junction and the depletion region is reduced to zero. Note that in this drawing, and many to follow, ions in the non-depleted regions are not illustrated because the drawing is simpler to understand without them.

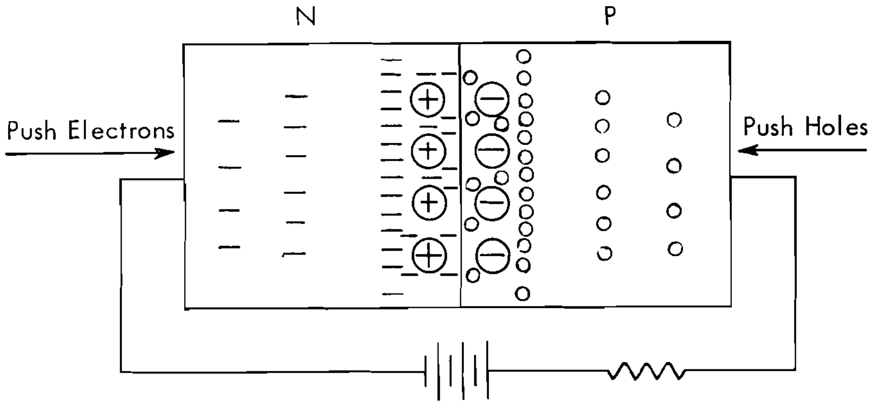


Figure 13. Forward Bias Drives Majority Carriers to the Barrier

Figure 14 is the result of reducing Figure 13 to its simplest form. It is obtained by crossing out one electron for each positive ion (cancel one positive and negative charge) in the N-region barrier and one hole for each negative ion in the P-region barrier. Therefore, forward bias has caused majority carriers in both regions to reach the junction in numbers far greater than the ion population of this region. It is now obvious that forward bias removed the barrier and majority carriers from both regions are attracted across the junction. Once across the junction, they become minority carriers (electrons in the P region, and holes in the N region) and are attracted by the battery.

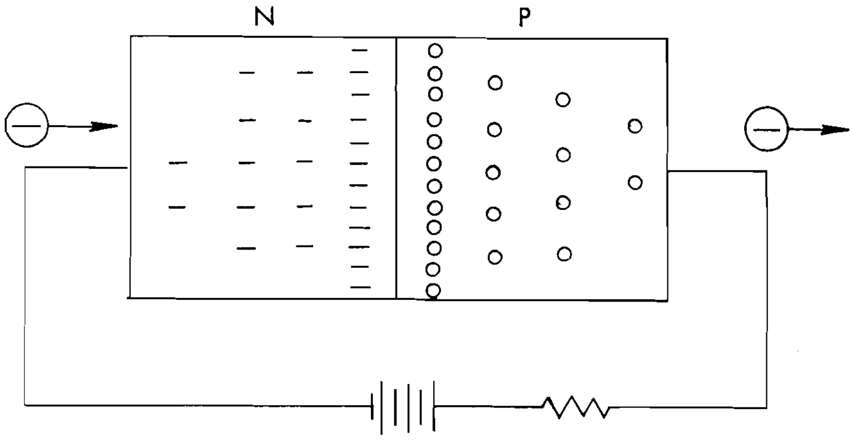


Figure 14. Forward Bias Reduces the Depletion Region to Zero

Figure 15 illustrates the current flow that results when majority carriers from the N region cross the barrier. As seen, electrons cross the barrier, travel through the P region, and flow through the load to the positive potential. To maintain an electrical balance, electrons from the negative potential are returned to the N region.

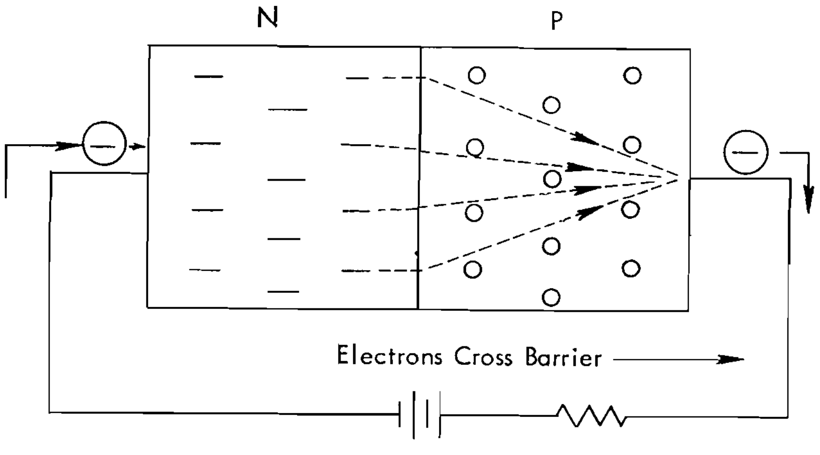


Figure 15. Forward Bias Causes Conduction Band Current (Electrons) to Flow

Figure 16 illustrates the current flow that results when majority carriers from the P region cross the barrier. As seen, holes cross the barrier, travel through the N region, and are filled by electrons from the negative potential. Of course, when holes in the P region move toward the barrier, they leave behind a negative

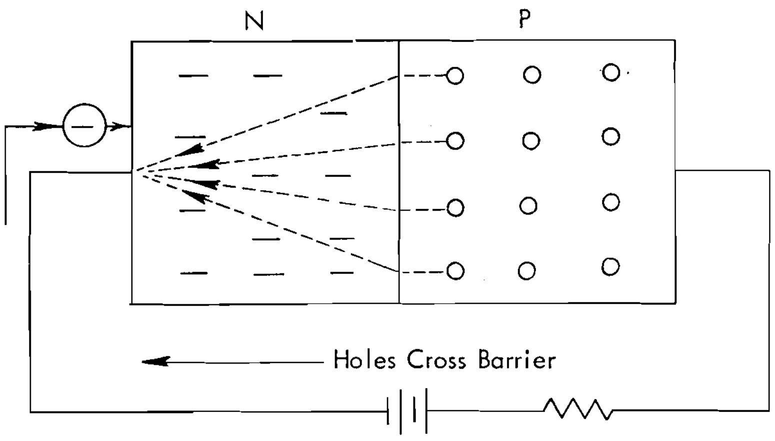


Figure 16. Forward Bias Causes Valence Band Current (Holes to Flow)

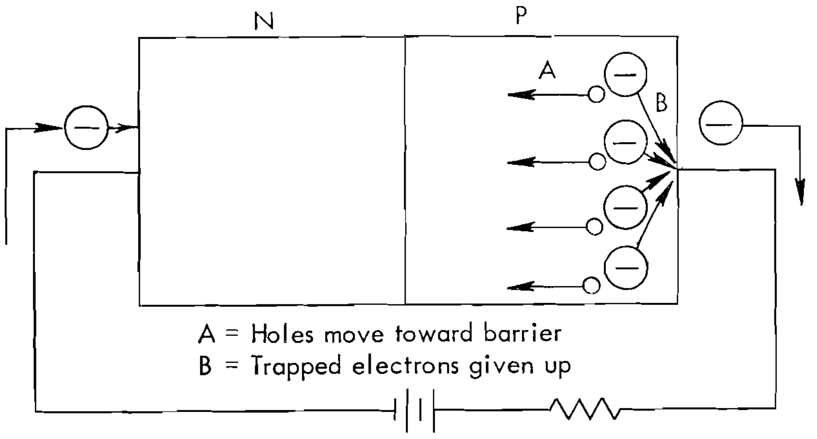


Figure 17. Continuous Hole Generation Exists at the Crystal Surface

ion region, which is locked into the crystal structure (Figure 17).

These negative ions (now not neutralized by holes) are acted on by the positive potential which uncovers them; i.e., the electrons trapped by the impurity atoms are not tightly bound and the positive potential exhibits a force that removes them from their trapped locations. These freed electrons flow through the load to the positive potential, which brings the source back to normal. The uncovered impurity atoms again generate new holes, which are attracted toward the barrier, and the current flow process continues.

Besides electron flow and hole flow, a third type of current exists, called recombination current (Figure 18). The recombination process (an electron drops into a hole) is the result when:

1. A majority carrier (electron) combines with a minority carrier (hole) in the N region.
2. A majority carrier (hole) combines with a minority carrier (electron) in the P region.

In either case, the recombination leaves behind a positive ion in the N region and a negative ion in the P region. To return the N and P regions to neutral, the negative ion gives up the trapped electron to the positive potential and the negative potential delivers one electron to the N region. The uncovered ion in the P region again generates a new hole and the current flow process continues.

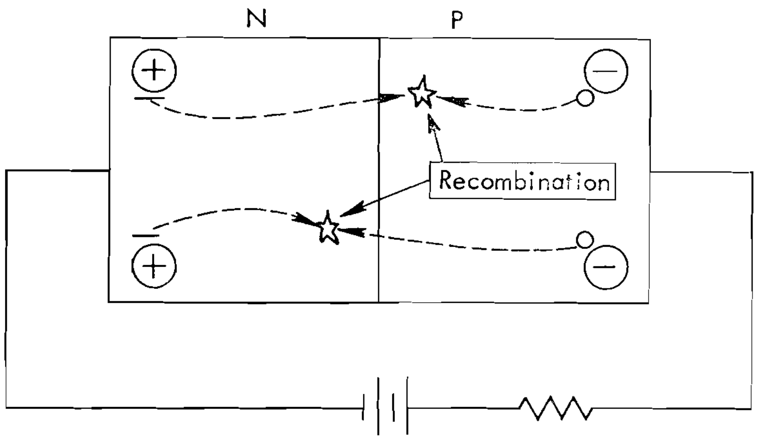


Figure 18. Some Majority Carriers Recombine

Characteristic Curve

The electrical characteristics of a germanium diode are shown in Figure 19. The voltage applied to the diode is plotted on the X (horizontal) axis and the current that flows is plotted on the Y (vertical) axis. The first quadrant plots forward-bias current and the third quadrant plots reverse-bias current. A further study of this curve reveals the following:

1. Only a small forward-bias voltage is required to cause a large current to flow. This also means that the forward resistance is small.
2. A large voltage variation in the reverse-bias direction has little effect on current flow. Back current, you recall, flows because of the generation of couples in the barrier region. These couples are produced by thermal activity (junction temperature) and not by the value of reverse bias. The low value of current tells us that the back resistance is large.
3. An increase in reverse bias voltage produces a breakdown point at approximately 20 to 40 volts, depending on the diode construction. At breakdown, the diode exhibits a small back resistance and a large current flows.

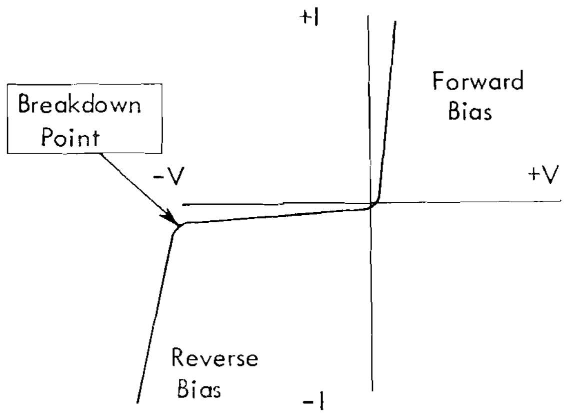


Figure 19. Germanium Diode Characteristic Curve

Avalanche Breakdown

The phenomenon of breakdown is caused by either an avalanche or Zener effect. An explanation of each of these effects follows.

Figure 20 illustrates the internal generation of current carriers which result from avalanche breakdown. The sequence of action is as follows:

1. A couple is generated in the barrier region in the normal manner.
2. The electron, freed by the couple, travels toward the positive ion region. The strong breakdown potential causes the free electron to gain sufficient speed so that, when it strikes an atom, it dislodges

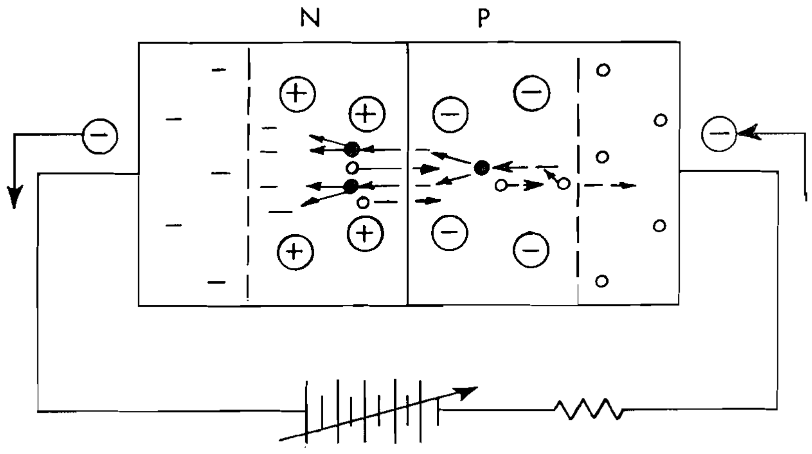


Figure 20. High Potential Accelerates Free Electrons Which Ionize Ge Atoms to Start Avalanche

an electron from orbit. Thus, an atomic collision at the breakdown potential creates a free electron and a hole.

3. Now, two free electrons and two holes exist, one caused by a couple and one caused by a collision. The two free electrons gain sufficient speed to dislodge two additional electrons from orbit. Thus, an avalanche or multiplication process takes place.
4. The holes move to the negative source and recombine with electrons delivered by the battery.

NOTE: Atomic collisions take place all the time, but they have no avalanche effect until the applied potential is strong enough to cause ionization.

Zener Breakdown

Zener breakdown results when the barrier potential is large enough to suck electrons out of orbit. This is similar to the "high-field emission" effect studied in vacuum tube theory. Breakdown depends on developing a large charge whose potential gradient is concentrated in a very small area. Visualize breakdown as the same type of action as the discharge of a condenser through a small gap.

Zener breakdown is a function of the barrier charge, so study Figures 21a, b, and c to see what happens at the barrier. Notice that an increase in bias produces a corresponding increase in the barrier charge. Up to the breakdown voltage, an increase in bias has little effect. But at breakdown, the positive ion region has a strong enough charge to remove from orbit the electron trapped by the impurity atom in the P barrier region. In other words, at breakdown it is the negative ion region that breaks down; negative ions release their trapped electrons (Figure 22).

Figure 22 is a drawing of breakdown, as is Figure 21c, except that in Figure 22 the barrier region is drawn expanded so that breakdown currents can be shown. The lower current path is explained as follows:

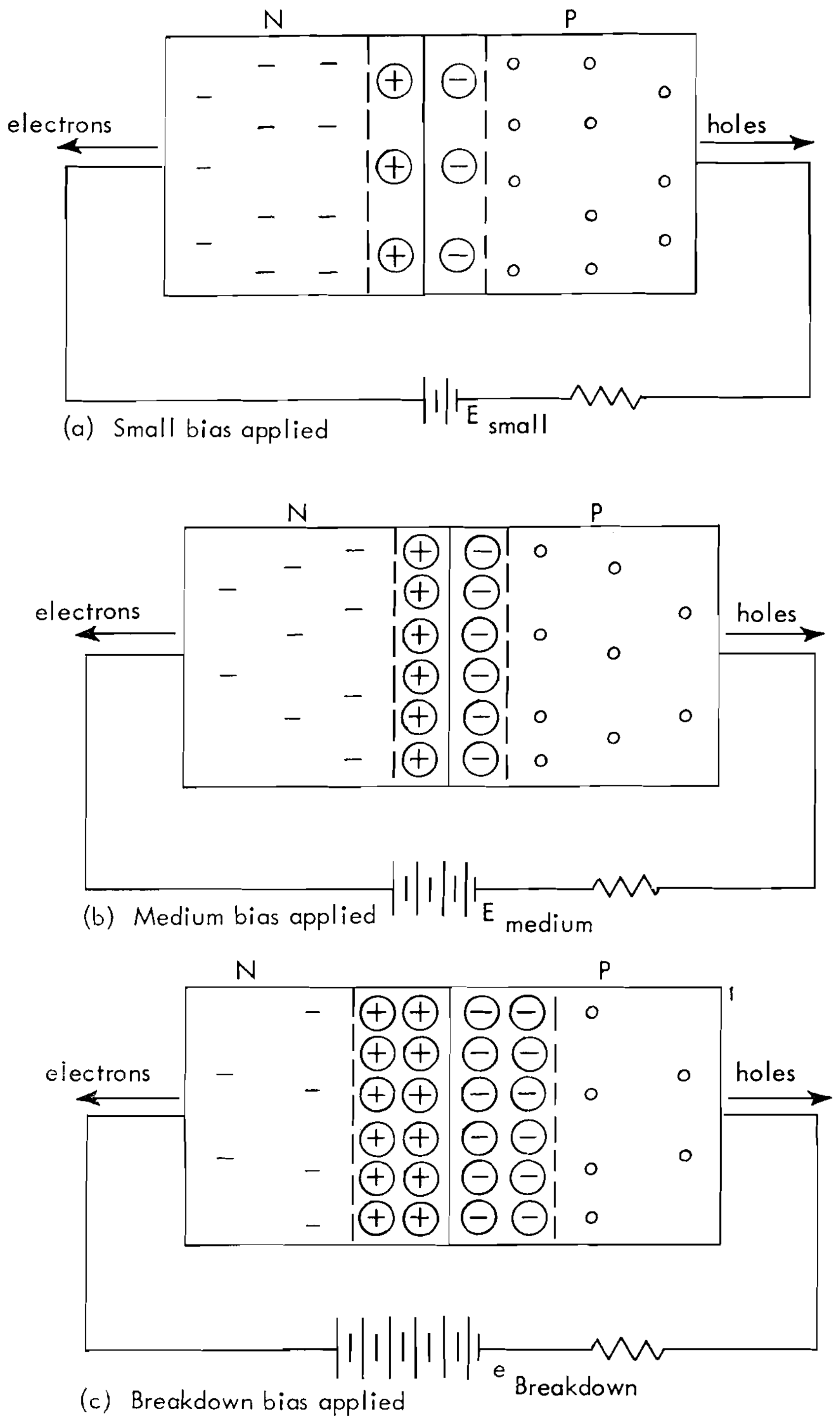


Figure 21. Increasing the Bias Increases the Potential Gradient Existing at the Junction

1. The trapped electron is withdrawn from orbit, crosses the barrier, and is collected by the positive potential.
2. The impurity atom generates a new hole which migrates to the negative terminal and recombines with an electron given up by the supply.
3. The process repeats.

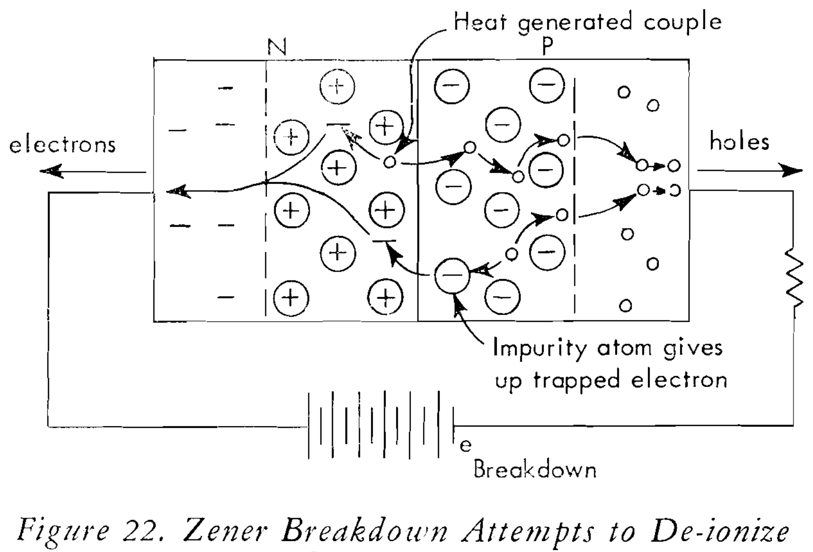


Figure 22. Zener Breakdown Attempts to De-ionize the P Junction

The upper current path is as follows:

1. Couples take place in the normal manner (because of heat).
2. The current generated by the lower current path increases the junction temperature which increases couples.
3. Without a current-limiting device in the external circuit, this cycle (increase of current, increase of heat, increase of current), continues until the physical properties of the diode are destroyed by heat. Excessive junction heat causes the impurity atoms to become mobile. They migrate to new crystal locations and the junction is destroyed.

Two types of breakdowns have been discussed, namely, avalanche and Zener. Under what conditions, then, may one or the other exist? Zener breakdown, of course, requires a high concentration of charge; therefore, the doping concentration must be high. Although Zener breakdown is theoretically possible, tests seem to indicate that avalanche breakdown is generally reached before Zener breakdown is realized. Of course, future semi-conductor developments may indicate opposite results.

Figure 23 illustrates the barrier potential curve (potential hill) for Figures 21a, b, and c. It is a convenient way of showing that increased bias increases the barrier potential and, more significantly, that this charge is wholly concentrated at the barrier. Thus, diode resistance is really barrier resistance.

The slope of the potential hill curve (the line connecting the positive and negative peaks) indicates the concentration of doping existing in the N and P regions. This line is almost vertical in Figure 23, indicating that the diode has a high impurity density. The curve of Figure 8 is not steep, indicating that the concentration of impurities is not large.

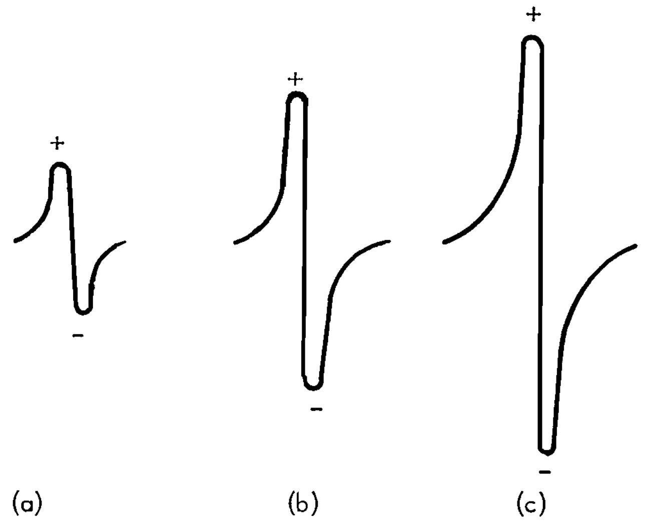


Figure 23. Barrier Potential Hills for (a) Small Bias, (b) Medium Bias, and (c) Breakdown Bias

TRANSISTOR THEORY

A TRANSISTOR is a semi-conductor device having three or more elements, although most transistors today are three-element devices. It has properties that make it a good voltage amplifier and a good current amplifier. It is, therefore, used effectively in small and large-scale calculators to replace tube circuitry. Because of its small size, reliability, long life, ruggedness, good power-handling ability, and low power requirements, it is especially applicable to large-scale calculators. It has lower power requirements than a tube, because it has no filament to heat. Of course, the lack of filament heating reduces or eliminates air conditioning requirements; they become none at all or very little. If you get the feeling that this device is something special, go to the head of the class, for its present potential is great and its future potential appears boundless.

Figure 24 (parts a, b, and c) permits comparing a tube schematic with the transistor form. In Figure 25 the physical appearance of the transistor is shown.

Facts of significance at this time include the following:

1. The NPN and the PNP are two types of three-element transistors made.
2. Only three-element transistors are discussed here because transistors having more than three elements are at present limited in production and use.
3. Each tube element has a transistor equivalent:
   Cathode = emitter
   Grid = base
   Plate = collector
4. In actual circuits, the elements are not labeled *E, B,* and *C* as shown. Identification is made by drawing an arrow on the emitter lead.
5. The arrow (on the emitter lead) points in the direction of conventional current flow (positive to negative).
6. The case is approximately 3/8″ in diameter by 1/4″ high (new type) and 5/16″ high by 3/16″ thick (old type).

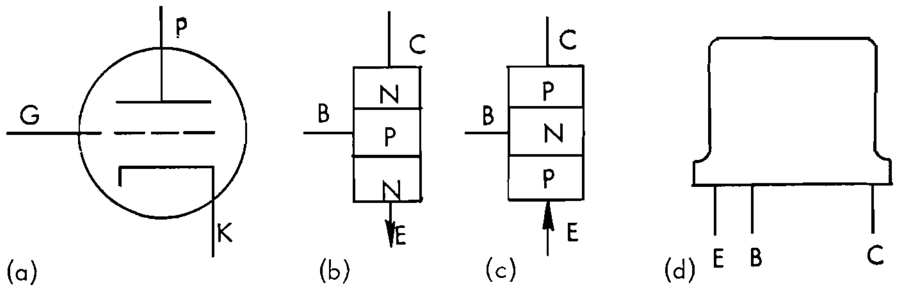


Figure 24. (a) Triode, (b) NPN Transistor, (c) PNP Transistor, (d) Transistor Appearance

Alloyed-Junction Construction

Figure 25 is a cross-section view of a PNP alloyed junction transistor. The dimensions shown are those of a high quality transistor, the type 01 used in the IBM 608 Calculator. The NPN equivalent is similar, except that the collector is .015″ and the emitter-to-collector base thickness is .0006″. The alloying process is delicate because transistor operation is dependent on the emitter-to-collector base thickness and the parallelism of the two junctions. As can be imagined, extremely close control of the oven temperature, the length of time in the oven, and the thickness of the base wafer are required. In other words, transistors are difficult to manufacture at this time. New techniques being developed indicate that the future manufacturing outlook is bright.

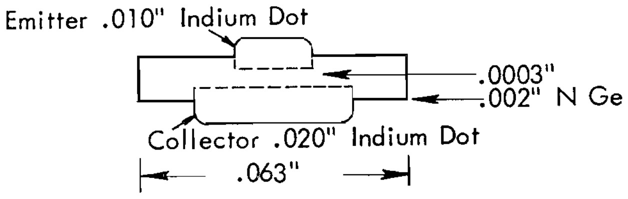


Figure 25. PNP Alloyed-Junction Geometry

Once the alloying process is complete, the emitter, base, and collector elements require wire terminations so that the device can be connected to a circuit. These connections are shown in Figure 26 and are accomplished as follows:

1. A gold-plated base tab is fused to the base.
2. Wire supports (labeled *E, B,* and *C*) are secured (not shown in Figure 26) in a bonding agent called a mount. The mount is an insulating material such as glass.
3. The *B* wire support is connected to the base tab.
4. A fine wire is connected to the emitter and the *E* wire support. In a like manner, the collector is wired to the *C* wire support.

The transistor assembly is made rugged by protecting it with an outer metal case. The case is hermetically sealed to protect the transistor against moisture. This is necessary because of the small size of the elements; moisture would short them. Local moisture is absorbed by a drying agent which is also encapsuled.

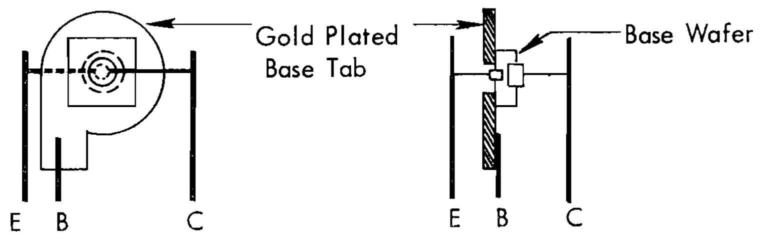


Figure 26. Details of Transistor Assembly

Static Condition

An NPN transistor and the natural barriers that form are shown in Figure 27. Note particularly that the depletion region is wider in the base than in the emitter or collector region. This is so because the emitter and collector are doped more heavily than the base. This doping ratio is roughly 20-100 to one. At this time it is not apparent why, but one should know that the amount of doping in the base is important to transistor operation. Certain advantages and disadvantages are realized if the doping is either high or low.

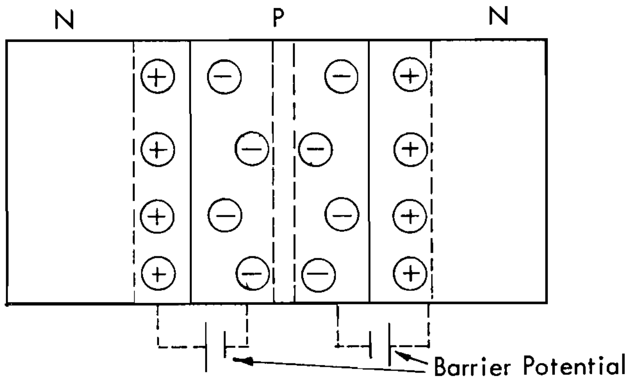


Figure 27. Natural Base-to-Emitter and Base-to-Collector Barriers

Reverse Bias

Figure 28 shows a reverse-biased NPN transistor and the current flow that exists. The *E*-to-*B* diode and the *C*-to-*B* diode are both reverse biased because the battery polarity connects to unlike elements. Reverse current flows from base to emitter and base to collector. The *B*-to-*E* current is called I_{ebo} for current flow, emitter to base, with the collector open-circuited. The *B*-to-*C* current is called I_{cbo} for current flow, collector to base, with the emitter open-circuited. I_{ebo} and I_{cbo} are generally used in the reduced form of I_{eo} and I_{co}. I_{co} and I_{eo} are small currents in the order of 2-60 μa for high-frequency transistors. Although this current may seem small, I_{co} is an important consideration in circuit design because it flows in the output circuit, which is generally a high impedance.

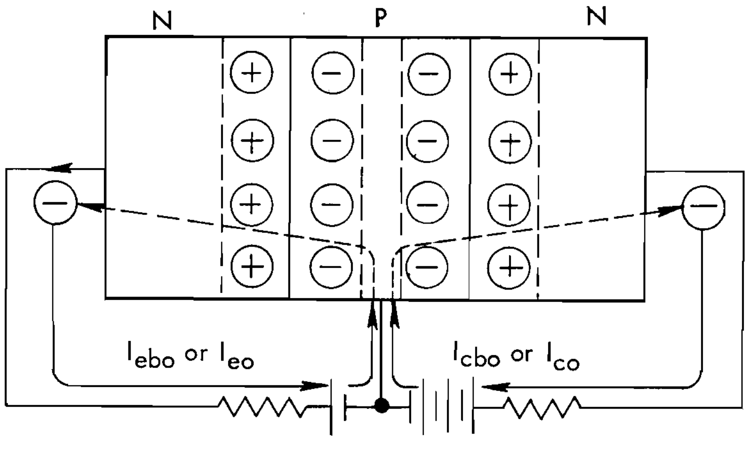


Figure 28. Reverse-Biased Transistor and Back Currents

Forward Bias

The NPN transistor circuit of Figure 29 is forward biased and is identical to Figure 28 except that the E-to-B battery is reversed. Forward bias drives majority carriers to the barrier region in abundance. Majority carriers reduce the depletion region to zero, and in addition set up a barrier potential which "aids" majority carriers across the barrier (Figure 30).

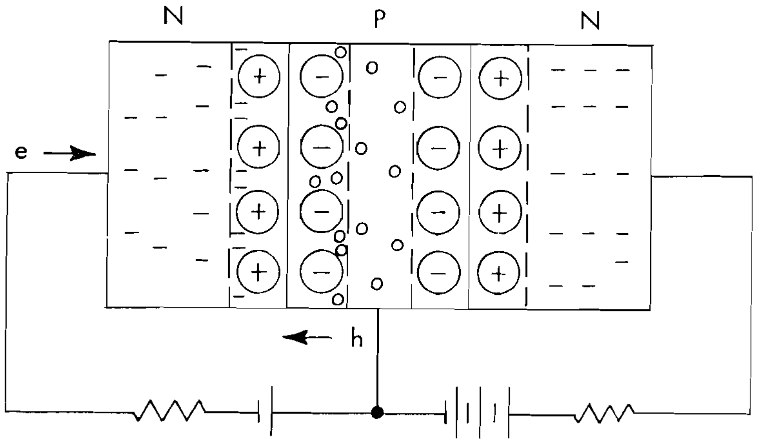


Figure 29. Forward Bias Drives Majority Carriers to the Barrier

Figure 30 is the equivalent of Figure 29 after the depletion region is reduced to zero; that is, by cancelling negative ions with holes and positive ions with electrons, the E-to-B region appears as shown in Figure 30. Thus, it is obvious that electrons are attracted into the base region and holes are attracted into the emitter region.

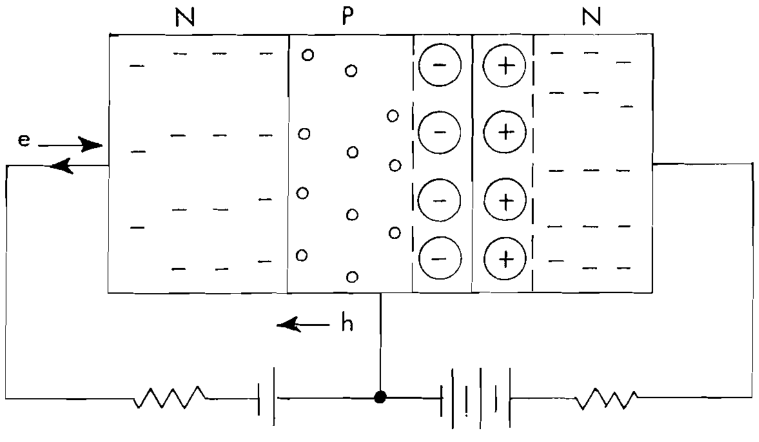


Figure 30. Forward Bias Reduces the Depletion Region to Zero

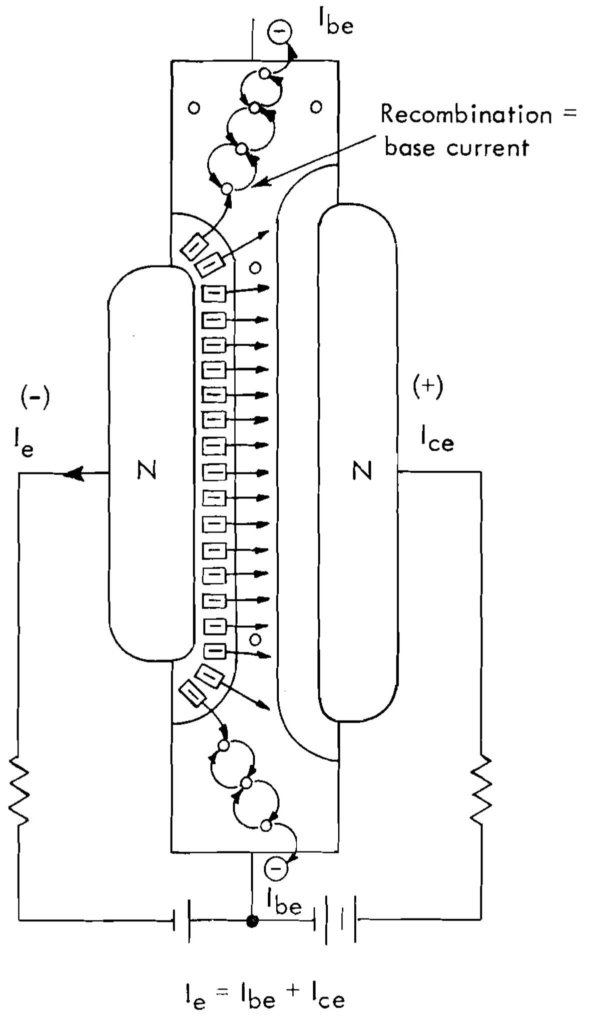
General Operation

It is now of advantage to describe in general terms how a transistor works. Basically, a driving source (external circuit) controls the emitter-to-base bias, which in turn controls a current flow from the emitter to the collector. Bias control works as follows:

1. Reverse bias prevents current flow from the emitter to collector (output current).
2. Forward bias permits emitter-to-collector current flow.
3. The degree of forward bias (how much) controls the amount of emitter-to-collector current.
4. The collector-to-base is always reverse-biased.

Although transistors do not act identically to tubes, certain similarities exist. For instance, in tube circuits, a signal is fed to the grid or cathode to control current through the tube, and in a transistor circuit a signal is fed to the base or emitter to control current through the transistor.

Further anaylsis of a transistor's electrical characteristics will be clear, if its physical properties are again studied. Remember that the emitter junction has a large surface area (compared to the base width) and the collector junction has an even larger surface area. These two large surface areas are extremely close to one another. This is similar to two capacitor plates spaced close together. With this in mind, study Figure 31, which is a cross-sectional view of an NPN transistor. The outer areas containing the N notations are considered as part of the external circuit; i.e., they contain the non-alloyed bulk of the emitter and collector dots and their ohmic value is practically zero. The actual emitter is the alloyed region shown containing an electron source. The actual collector is the alloyed region shown, similar to the emitter region except that it is larger. The base, of course, is the region between the



$I_e = I_{be} + I_{ce}$

Figure 31. Cutaway View of Transistor Geometry

emitter and the collector which is drawn as a rectangle containing "holes."

Again study Figure 31, only this time try to picture what happens electrically while not forgetting what the physical properties are. This study should reveal the following:

1. The *E*-to-*B* is forward biased.
2. The *C*-to-*B* is reverse biased as it always is in transistor circuits.
3. Current entering the emitter is called I_e.
4. I_e flows into the base region where it divides into I_{be} (base-to-emitter current) and I_{ce} (collector-to-emitter current).
5. Electrons entering the base find the most direct route to a positive potential by traveling to the collector region.
6. Because the collector is larger than the emitter, many of the electrons leaving the periphery of the emitter still reach the collector.
7. Most base current occurs because electrons emitted from the emitter periphery are not directed toward the collector. (See the emitter geometry.) These electrons find the base potential a more direct return than the collector potential.

Minority Carriers

Because the whole concept of transistor action deals with dumping minority carriers into the base and then acting on these carriers, let us investigate this action more closely. First, a clearer picture of the emitting source is needed (as shown in Figure 32).

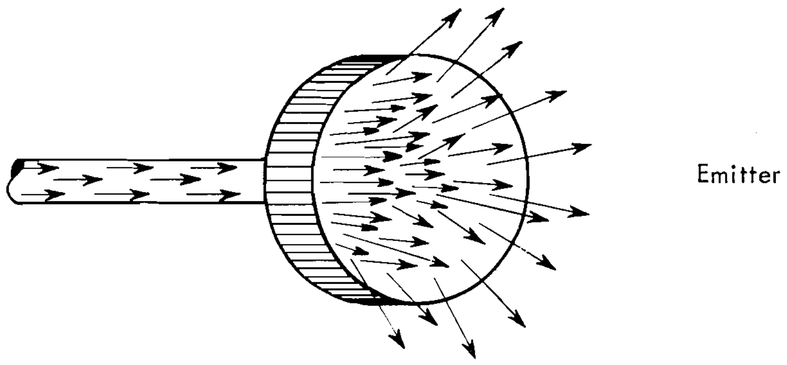


Figure 32. Each Surface Location Is an Emitting Point Source

Here we are looking into the emitter surface from the base side. Notice how each location on the surface acts as an "emitting point source." In other words, if a rectangular graph were laid across this surface, each intersection would represent an emitting source. This emitting action is similar to the water spray leaving a shower nozzle.

A three-dimensional view of the emitter and collector geometry is shown in Figure 33. Notice especially that most minority carriers reach the collector. This is so because the collector is made larger than the emitter and is spaced very close to it (approximately .0003″ to .0006″). Therefore, very few emitted carriers can escape this direct path to the collector region. Nevertheless, some do. Some carriers do not reach the collector primarily because of the geometry of the emitter periphery. (See Figure 36.) Carriers leaving the emitter periphery enter the base at angles almost perpendicular to the emitter surface, which is not perpendicular to the collector plane. Therefore, these carriers can migrate to either the collector region or the base surface. This action is represented in Figure 33 by long arrows (carriers not reaching the collector).

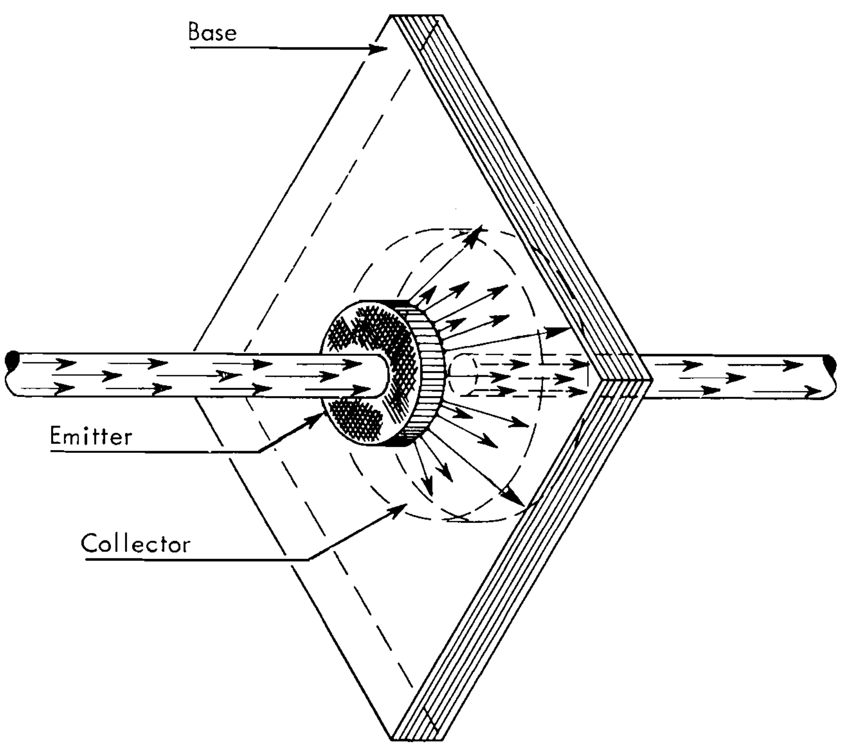


Figure 33. Three-Dimensional View of Emitting and Collecting Regions

Many people find it helpful to compare transistor operation to tube operation. In some areas the operation is similar while in others it is not. Of course, it is mostly the "not" areas that require explanation. Minority carrier flow through the base is a "not" area; that is, this action is *not* similar to current flow in a tube. Current flow in a tube, you recall, requires the emitting element (cathode or filament) to emit free electrons into a vacuum, after which the potentials of the grid and plate act on these carriers. The point here is that in tubes the free electrons move from one point to another, because of the attraction or repulsion of a potential acting on them. This is not true of minority carrier current in the base.

Minority carriers entering the base are not influenced by a potential because none exists in the base region; the base region is a neutral region. Of course, the *B*-to-*E* and *B*-to-*C* barriers exist, but only at the junction regions, and they do not extend any appreciable distance into the base region.

Diffusion Current

If potentials are not acting on minority carriers in the base, what is? Diffusion is. Diffusion results whenever like charges collect. Like charges repel, so they try to get away from one another. This "getting away" is a spreading out or diffusion process. This action, shown in Figure 34, is similar to gas diffusion.

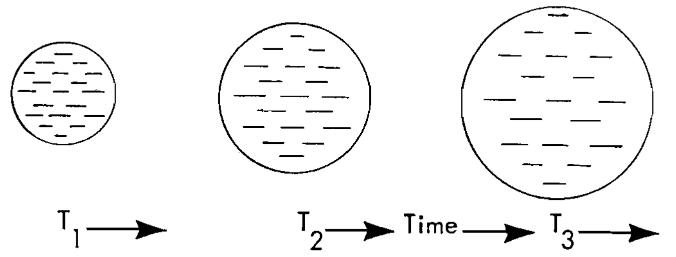


Figure 34. Like Charges Diffuse

Actually, diffusion is only part of the picture. The other part is the path taken by a minority carrier while traveling to the collector. The ideal path would, of course, be straight across, but a minority carrier may instead follow a random path as shown in Figure 35. A random path results when a minority carrier comes under the influence of charges in the base; that is, the minority carrier is deflected by a collision with an atom or by a concentration of charges existing in a location it is entering.

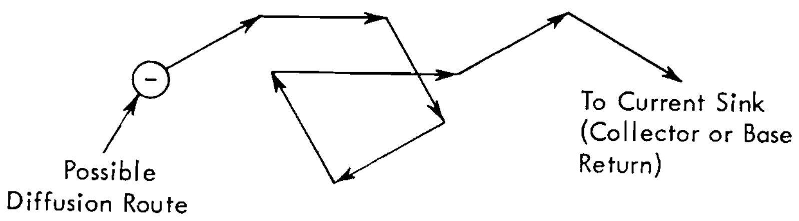


Figure 35. Diffusion Current Travels in a Random Manner

By now, it may appear that diffusion current is a rather haphazard action, that minority carriers "float" over to the collector. Actually this is not so. Diffusion current also has a direction and force component because of emitter action; the emitter is continuing to supply the base with minority carriers which force those previously emitted away from the emitter. This action causes a diffusion gradient to exist in the base as shown in Figure 36.

Current Sinks

Minority carrier transit in the base is shown in Figure 36. Because a clear understanding of this action will be helpful later, take the time to study this drawing carefully. See how a high concentration of minority carriers exists at the emitter and how the concentration decreases as carriers approach the collector. Actually, minority carriers in the base are searching for a return to a source or "current sink." The collector, of course, is a good sink because minority carriers reaching the collector become majority carriers and are swept through the collector region. The surface of the base region adjacent to the emitter is also a good sink because surface germanium atoms have incomplete covalent bonds. These surface atoms bond with atoms on three sides only and, therefore, each has a hole location. In other words, when the crystal surface is reached, there are no more atoms and the lattice is no longer diamond shaped.

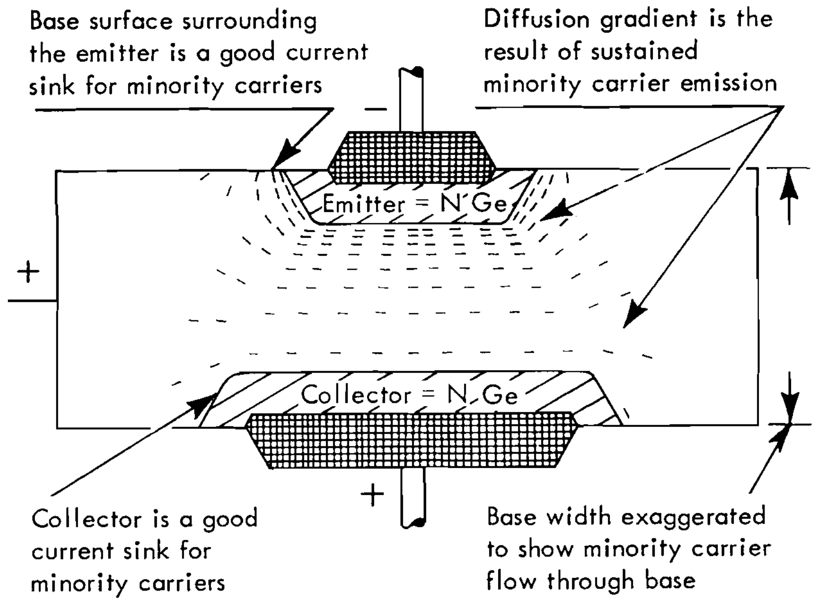


Figure 36. Minority Carrier Flow through Base Region

This surface structure makes the surface resistance of germanium much less than the resistance of the bulk material. For this reason, practically all base current originates when minority carriers reach the base surface adjacent to the emitter and recombine.

Base Recombination

Recombination is a difficult concept for many to understand clearly, so base surface recombination will be analyzed closely. The sequence of activity is as follows:

1. Hole locations exist on the surface because of the incomplete covalent bonding of germanium atoms.
2. The surface looks like a positive location to minority carrier electrons in the base.
3. Minority carriers reach the surface and recombine (attach themselves to germanium atom locations).
4. Once surface recombination takes place, the region has lost its neutrality and is acted on by the positive potential applied to the base.

5. Surface current flows to return the recombination region back to a neutral state.
6. The amount of surface current that flows is determined by how rapidly charges can move across the surface. This rate of flow is called "surface velocity."
7. A high surface velocity restores the recombination region back to normal fast; that is, it "cleans out" the base-surface current sink rapidly so that the sink can again attract minority carriers.
8. The rate of recombination is proportional to surface velocity.
9. Surface velocity should be kept as low as possible so that base current is held to a minimum. We will find later that transistors having the best gain characteristics are those in which the percentage of minority carriers reaching the collector is high and the percentage reaching the base is low.
10. The surface is contaminated by gas atoms from the atmosphere joining into the surface structure. This contamination usually results in increased surface velocity and is not desirable. Therefore, the surface is chemically treated in the manufacturing process and the transistor is sealed for protection from the atmosphere. Broken seals reduce the lifetime of a transistor through surface contamination, so treat them carefully.

It should now be clear that surface recombination is a dominant factor determining base current. Bulk recombination (recombination other than surface) also exists, but the quantity is small and can be disregarded.

A close look at the emitter should also reveal that minority carriers from the periphery set up a sort of minority carrier cloud, which tends to focus toward the collector those carriers emitted from the inner emitter area. Notice also that a narrow *B*-to-*C* width results in a less pronounced "diffusion gradient" and an increased arrival rate of minority carriers at the collector.

Behavior

Generally speaking, it is desirable for a transistor to:

1. Voltage-amplify the input signal or current-amplify the input signal.
2. Produce an output signal with minimum distortion.

At this time, consider the distortion caused by transistors. Voltage and current amplification is covered in detail later, in the "Circuits" section.

Signal Distortion

Figure 37 shows an input signal and a resulting output signal. As you can see, the output signal is not a faithful reproduction of the input signal. The specific characteristics that distort the output signal are labeled *A, B, C,* and *D.* Because each of these characteristics requires lengthy explanation, they are individually covered in detail later, and only a brief explanation is given here.

A is *turn-on delay.* This delay results because carriers leaving the emitter become minority carriers which take a finite period of time to cross the base region. This time interval is called "transit time." The point of significance here is that even though the emitter is passing a signal, the collector circuit does not recognize this signal until emitted carriers reach the collector. When emitted carriers enter the collector region, others leave the collector region and flow through the external load. Think of this majority carrier current in the collector as you would a copper wire; for each carrier entering the collector, one leaves to enter the external circuit. Thus, current flow through the collector region does not delay the output signal.

B is *turn-on transition.* This phenomenon results because emitted carriers travel to the collector by random routes and because individual carriers travel through the base at different velocities. Therefore, all of the first carriers emitted do not arrive at the collector at the same time. Those that travel the most direct route at the fastest velocity arrive first, while those of slow speed which travel the least direct route arrive last. In any case, the non-uniform arrival rate means that the leading edge of the signal is distorted.

C is *turn-off delay.* It is due to transit time through the base. Thus, when the input signal is cut off, the output signal does not fall until the last increment of emitted carriers starts arriving at the collector.

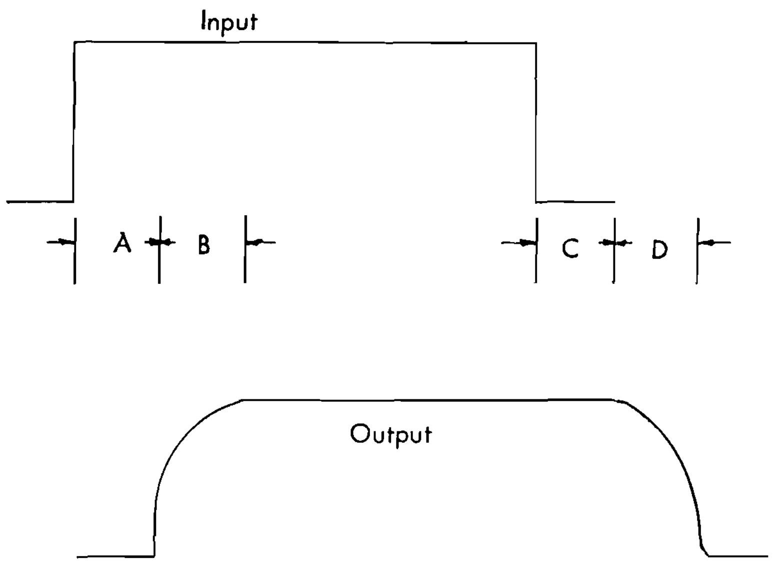


Figure 37. Output Signal Is Distorted by Transistor Characteristics of Delay and Transition

D is *turn-off transition.* It is identical in nature to turn-on transition except that it takes place on the trailing edge of the signal.

Delay

Figure 38 illustrates the transit time of minority carriers through the base, turn-on delay, and turn-off delay. A time base, T_0 through T_5, is drawn so that the input signal, output signal, and minority carriers in the base can be referenced to one another. Assume that the transistor is reverse-biased on the down level of the input signal, and forward-biased on the up level of the input signal, in the explanation that follows.

1. At time T_0, the signal is down and no input current flows.
2. Between T_0 and T_1 the input signal rises and electrons enter the base and become minority carriers.
3. Between T_1 and T_2 the input signal falls. A study of transistor T_2 shows that electrons are no longer entering the base (because the input signal is down). Those previously emitted, during the up level of the input signal, continue to travel to the collector.

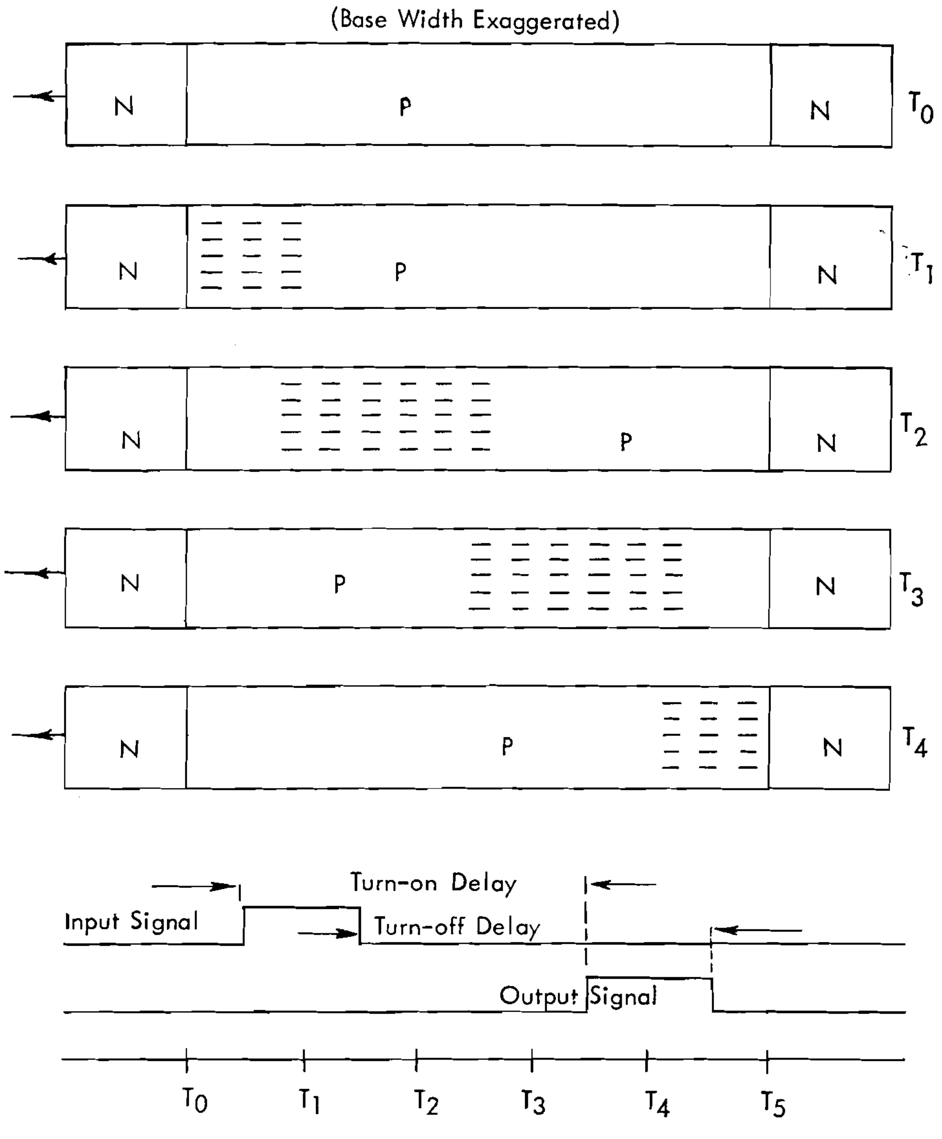


Figure 38. Turn-On and Turn-Off Delay Are Caused by Base Transit Time

4. Between T_3 and T_4 the output signal rises because the first carriers emitted finally arrive at the collector.
5. The output signal falls between T_4 and T_5 when the last carriers are collected.

Although transit time through the base is the major factor of turn-on delay, the base-to-emitter capacitance is also a contributing factor. This capacitance effect is explained as follows:

1. When the input signal is down, the *B*-to-*E* barrier is charged to the value of reverse bias and a depletion region exists.
2. When the input signal rises, majority carriers first fill the depletion region (a charge effect). After the depletion region is reduced to zero, majority carriers enter the base. Thus, the time required to reduce the depletion region to zero is part of turn-on delay.

NOTE: In Figure 38 it was assumed that all carriers emitted took the same transit time through the base. Actually this is not so, and was so shown only for purposes of simplicity.

Transition

Figure 39 diagrams possible routes taken by electrons emitted into the base. A time reference, T_1, T_2, T_3 and T_4, is shown so that velocities can be compared. The lettered notations have the following significance:

A. On its journey to the collector, this electron collides with an atom and is deflected from its original path.

B. This electron travels the most direct route.

C. The time notations show that the electron following this route is of a higher velocity than is the electron following route *D*.

E. Electrons take various routes to the collector. These routes are influenced by:

1. The "initial" emitted direction, i.e., the direction of travel of the electron as it entered the base.
2. The diffusion process in the base; i.e., carriers traveling through the base tend to spread out.

Transit time is also influenced by the non-uniform base-to-collector width (not shown in Figure 39). This non-uniformity results because the *B*-to-*E* and *B*-to-*C* junctions are not straight lines as shown, but are of a slightly ragged definition.

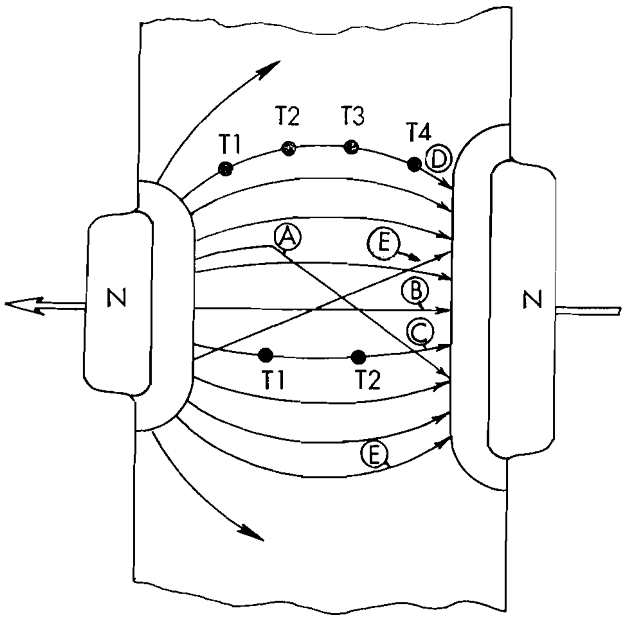


Figure 39. Possible Routes and Speed of Minority Carriers in the Base

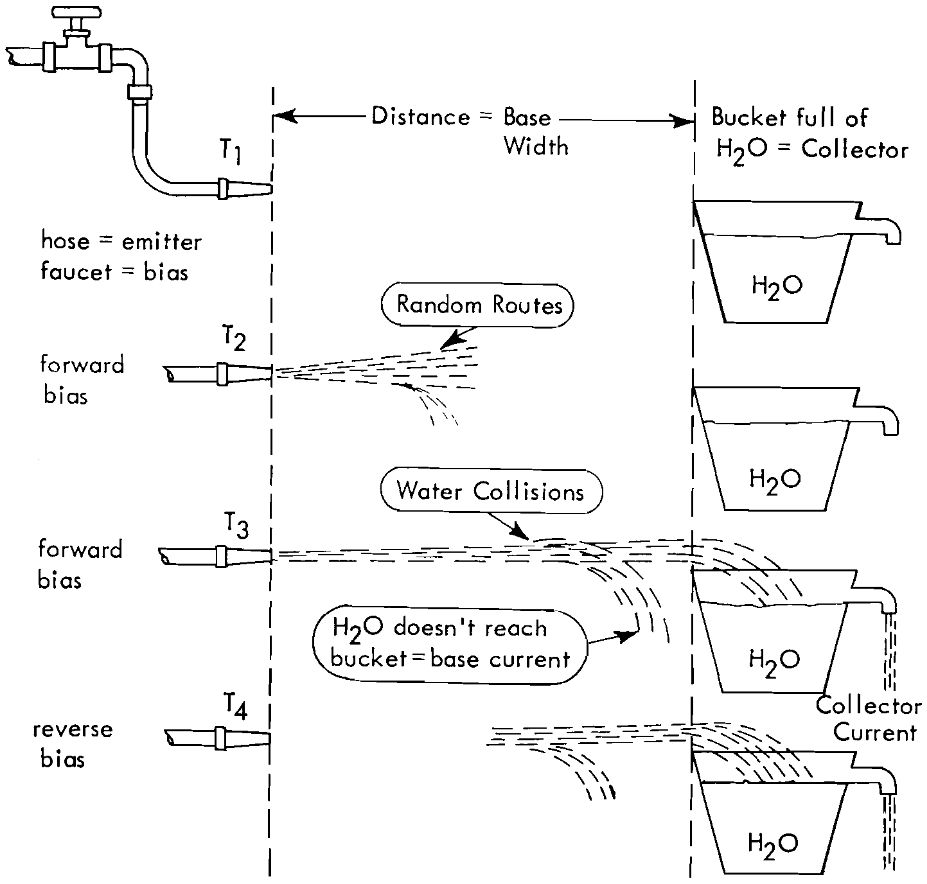


Figure 40. Water Analogy of Transistor Action

It should now be apparent that the non-uniform carrier transit time through the base is the result of several factors, and is not a simple, cut and dried concept. It is an important consideration because it is this phenomenon which distorts the leading and trailing edges of a signal. In other words, it distorts the "high frequency" component of a signal. One should keep in mind, of course, that various transistor types have different frequency response parameters, so that while transit time will result in distortion of a 50 kc signal for one transistor type, another will pass a clean one-megacycle signal.

Water Analogy

A water analogy of transistor action is shown in Figure 40. It is presented here so that many of the concepts previously described can be better visualized and consolidated.

Figure 40 is divided into time periods, T_1, T_2, T_3 and T_4, in which the following action takes place:

T_1. The faucet is off and no water flows from the hose. This is equivalent to a reverse-biased transistor.

T_2. The faucet is turned on (forward bias) and water travels part way to the bucket (transit time through the base = turn-on delay). Notice how the hose nozzle causes water molecules to take random routes.

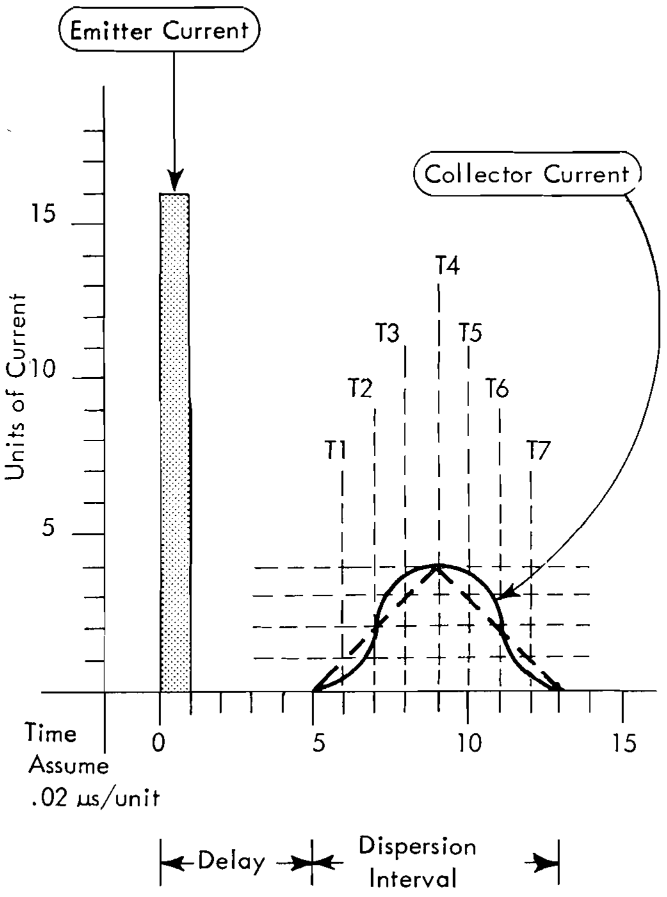
T_3. Water enters the bucket (carriers reach the collector) and simultaneously water flows out of the overflow pipe (carriers are not delayed in the collector region; as carriers enter, others leave). The outer spray caused by the nozzle does not have sufficient energy to reach the bucket (base current).

T_4. The faucet is turned off (reverse bias) but the water in transit continues to flow into the bucket (turn-off delay).

Dispersion Interval

Many electronic applications require square-wave signals for proper operation. The leading edge of such signals consists of odd harmonics of the fundamental square-wave frequency. Generally, the leading edge or trailing edge is reproduced without distortion if the response of the circuit is approximately ten times the fundamental frequency. Thus, the frequency response of transistors is an important parameter. It was shown previously that turn-on and turn-off transition distorted the edges of a square wave. Therefore, a closer investigation of transition time is required. Such an investigation is shown in Figure 41, where the arrival rate of emitted carriers is plotted for a very short burst of emitter current. To be of any value, the increment of time used in Figure 41 must be much less than the turn-on transition time for the transistor used.

Carefully study the information shown in Figure 41. This study should reveal the following:



Time Period	Units of Current
T1	1
T2	2
T3	3
T4	4
T5	3
T6	2
T7	1
Total	= 16
Thus, sum of collected current = emitted current = 16 units	
Base current is neglected for simplicity of calculation	

Figure 41. Plotting of a Dispersion Interval

1. The vertical axis plots units of current.
2. The horizontal axis plots units of time (.02 μs per unit).
3. Emitter current is switched on and off for .02 μs and emits 16 units of current.
4. This current has a transit time through the base of five units of time.
5. The fastest and most direct carriers reach the collector first.
6. The arrival time of carriers following the fastest is shown as a dome-shaped curve, which is approximated for simplicity by two dashed lines.
7. The spread in arrival time is called the "dispersion interval."
8. Units of current reaching the collector, when added, equal the amount of current emitted. Base current is neglected for simplicity.
9. The output signal is not an image of the input signal; it is quite distorted.

Figure 41 contains many details, but the information of key importance is the dispersion interval. Although the length of the dispersion interval varies with transistor types, the fact is that all transistors have a specific dispersion interval. The dispersion interval is a valuable tool because it can be used to plot the output signal resulting from a given input signal. An example of such a plot is shown in Figure 42.

Signal Graph

Figure 42 illustrates a high-frequency emitter signal and the output signal resulting when the transistor used has a dispersion interval of the base shown (five units of time). The following procedure was used to plot this output signal:

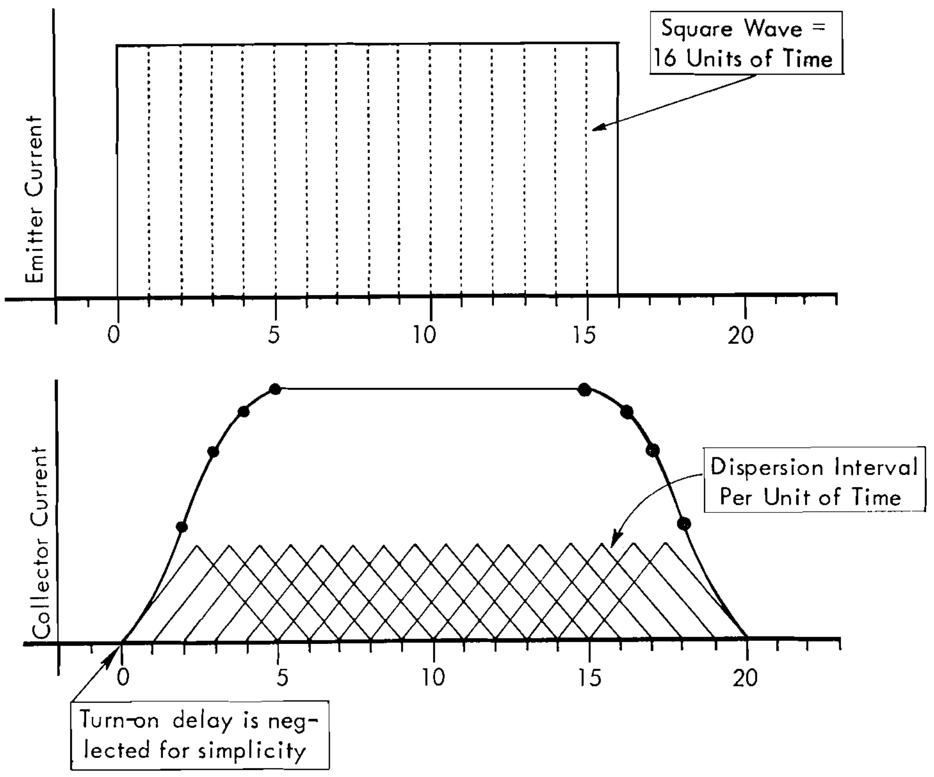


Figure 42. Output Signal Plotted by Summation of the Dispersion Intervals

1. Divide the input signal into increments of time (16 shown).
2. On a base line, draw the dispersion interval for each time increment (16 shown).
3. At each increment of time on the base line, add up the currents flowing, and plot this point above the line.
4. Draw an envelope through the points plotted.

It is obvious that the output signal is a distorted version of the input signal. But this waveform is only true for the conditions shown. We find that by changing the time base (frequency) of the emitter current, the output signal is affected in the following manner:

1. It is greatly distorted when the frequency is increased. For example, if the frequency is 16 times as great, emitter current flows for only one increment of time, and the output signal would look like the dispersion interval shown in Figure 41.
2. It has negligible distortion when the frequency is decreased. For example, if the frequency was only one-sixteenth as great, emitter current would flow for 16 times 16 or 256 units of time, and the leading and trailing edges would be steep when plotted to this time base.

Frequency Response

Previous discussions involving distortion were actually sub-topics of frequency response. Frequency response is an important transistor parameter because it establishes the highest pulse freuqency that can be used effectively in a circuit (Figure 43).

To show the effect of frequency response, a square-wave emitter signal of frequency f, f_1, f_2 and f_3 is used and the output signal is studied (Figure 43). Analysis of the output signal shows that:

1. At a low frequency f the output signal is an image of the input (no distortion exists).
2. At frequency f_1 (which is greater than f) the amplitude of the signal is not affected, but the leading and trailing edges are distorted.
3. At frequency f_2 (which is greater than f_1) the signal is distorted and the amplitude is reduced.
4. At frequency f_3 (which is greater than f_2) the output is practically a steady state output of 5 ma.

NOTE: It should now be clear that distortion first affects the edges of a signal. Further increases in frequency reduce the signal amplitude; the hills are removed and the valleys are filled in. Also notice that the same amount of current reaches the collector, but the "change" of signal is reduced to almost zero.

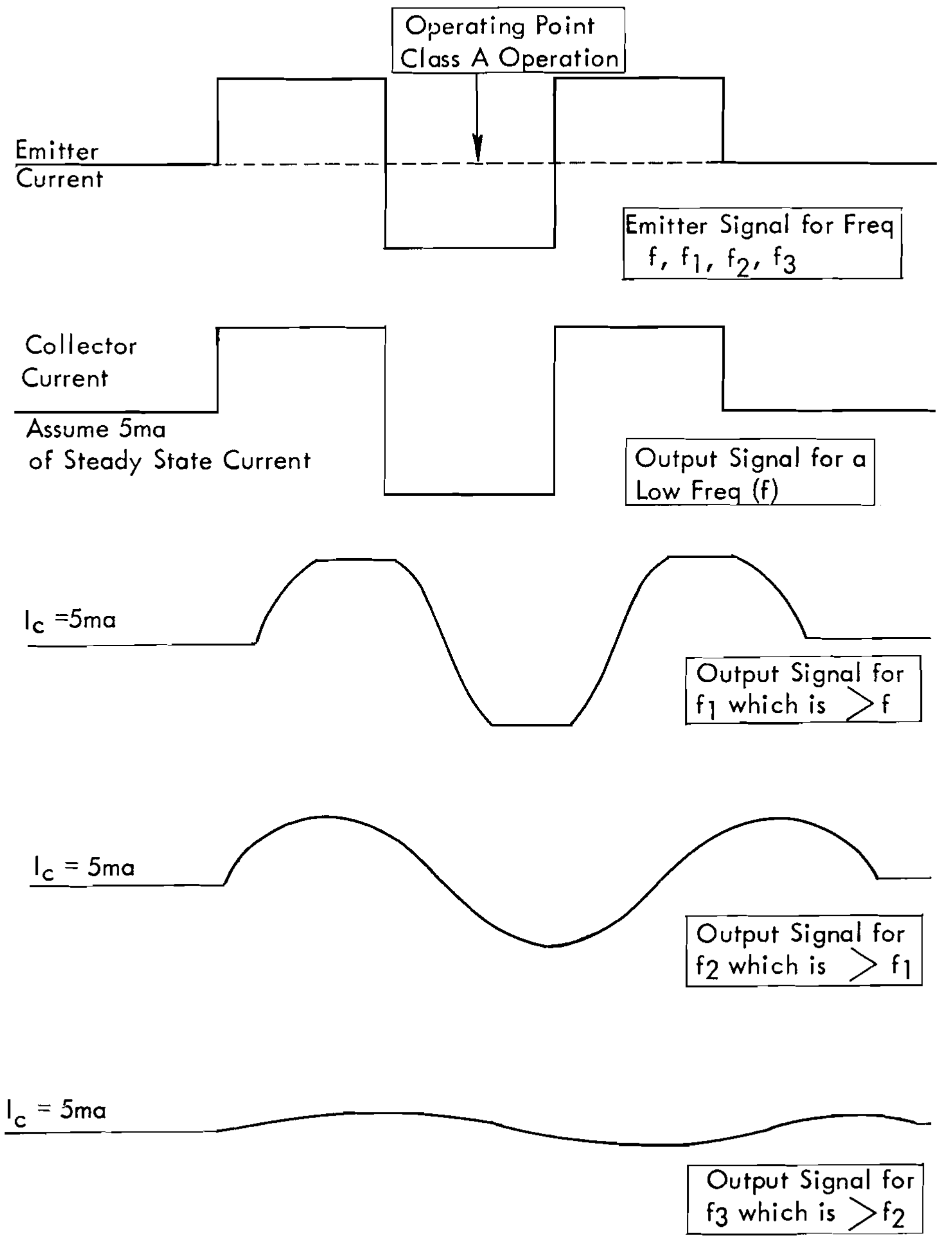


Figure 43. Output Signal Response to Low and High Frequencies

Frequency Cut-off

In order to establish parameters dealing with frequency response, a response curve must first be drawn (Figure 44). Such a curve shows how the output signal is affected when the frequency is increased. Frequency is plotted on the horizontal axis and gain (the amount of output signal) is plotted on the vertical axis. Gain is defined in detail later in the study of the grounded base circuit. The curve shows a uniform response until at some high frequency the gain falls toward zero.

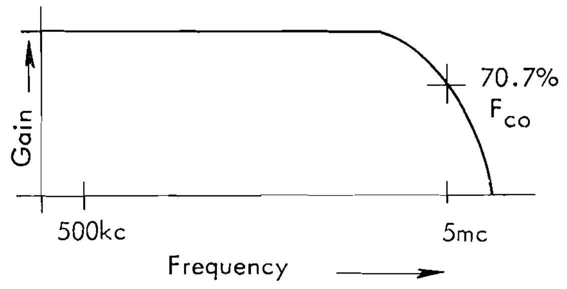


Figure 44. Frequency Response Curve

The transistor frequency response is considered usable until it falls to a one-half power value. This value is .707 of its low frequency response and is referred to as either:

1. f_{co} for frequency cut-off.
2. α_{co} for alpha (gain) cut-off.

The low frequency reference point now finding favor is ten percent of the F_{co} specification. In the past almost any reference was used, depending mostly on the frequencies available from the test source.

It was shown previously that frequency response is dependent on carrier transit through the base region. This dependency on base width is illustrated in Figures 45 and 46. The comparison shows that frequency response increases as the base width decreases.

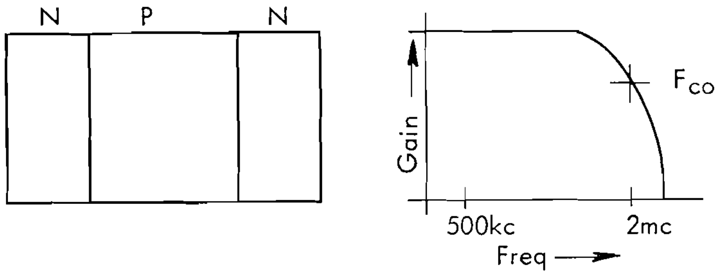


Figure 45. A Wider Base Results in a Lower Frequency Response

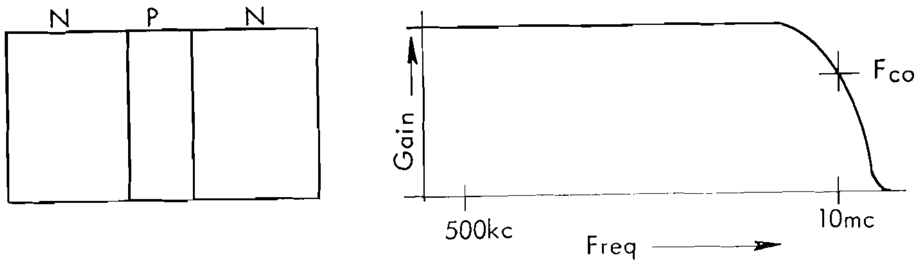


Figure 46. A Narrower Base Results in a Higher Frequency Response

It seems that a high-frequency response is easy to satisfy for transistors by just making the base extremely thin. This is true, but practical limits are placed on this thickness because of the following:

1. The manufacturing cost increases as the base thickness decreases, because a thinner base requires more stringent controls on the materials used and on the manufacturing process.
2. A base that is too thin will "punch through" when used in a circuit. Punch-through is covered in detail later but, briefly, this phenomenon results when the circuit voltage is sufficient to completely ionize the base region or "punch through" from collector to emitter. Under this condition, the transistor has exceeded its limits of control and it acts like a low-resistance device.

Therefore, base thickness is actually a compromise of frequence response, punch-through, and cost.

Punch-Through

Punch-through is unique to transistors and results when the reverse bias supply completely ionizes the base region. A series of drawings is presented here to show the progressive action that causes punch-through and also to show why the transistor loses control and acts like a low resistance device. The first drawing is Figure 47, which shows the distribution of charges for a normal reverse-biased transistor.

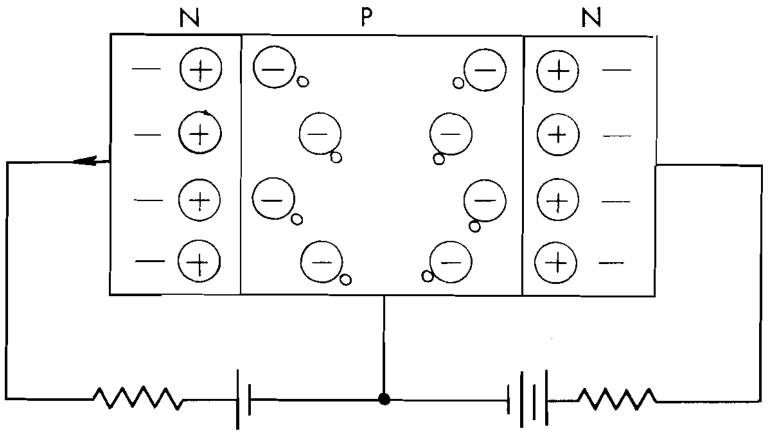


Figure 47. Collector-to-Base Barrier Produced by a Nominal Bias

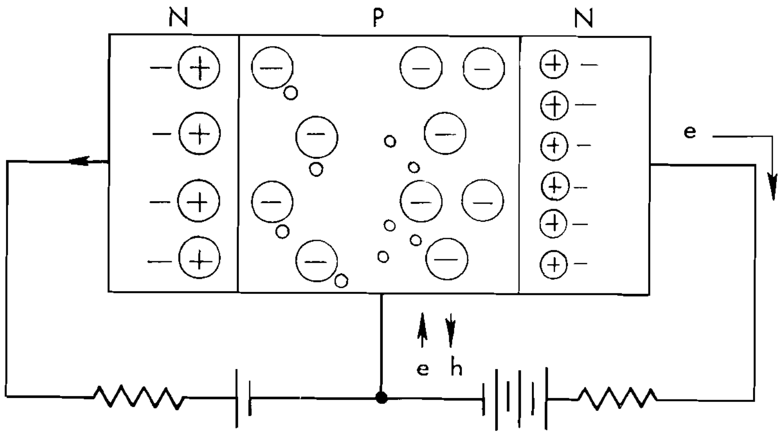


Figure 48. Collector-to-Base Barrier Produced by a Bias Larger than Nominal

Figure 48 shows the charge distribution after the *C*-to-*B* bias is increased. As always, an increase of bias increases the depletion region. Because the concentration of doping in the base is so much less than in the collector, the depletion region is shown extending a considerable distance into the base. Of course, the depletion region is increased because majority carriers are drawn away from the barrier by the increased bias. Because of the physical property of the *B*-to-*C* junction (a large circular area exists) majority carrier holes in the base are acted on as follows:

1. Those near the periphery of this large junction area are attracted to the surface of the base crystal, where some recombination with electrons from the battery source takes place.
2. Those leaving the barrier from the inner area are forced toward the emitter junction.

Naturally, for each electron from the battery source that recombines with a hole in the base, one electron is withdrawn from the collector. Although this collector action exists, it is not of particular significance in describing punch-through. Rather, it is the majority carrier action in the base region which is the culprit.

In Figure 49, a punch-through potential is applied to the *B*-to-*C* junction and the resulting effect is shown. The important action here is caused by holes which are driven to the emitter junction. This action reduces the negative ion region at the *B*-to-*E* junction to zero, which of course attracts majority carriers in the N region to the barrier. This action actually reduces the *B*-to-*E* depletion region to zero, and is similar to the action which would result if a *B*-to-*E* forward bias were applied.

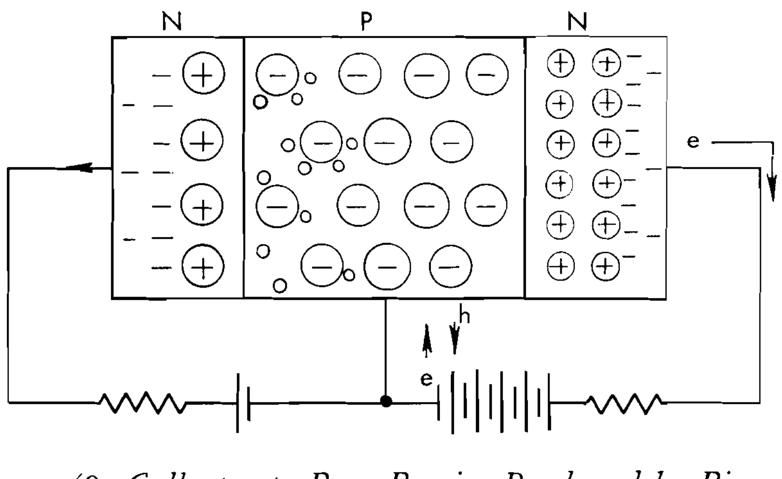


Figure 49. Collector-to-Base Barrier Produced by Bias of Punch-Through Value

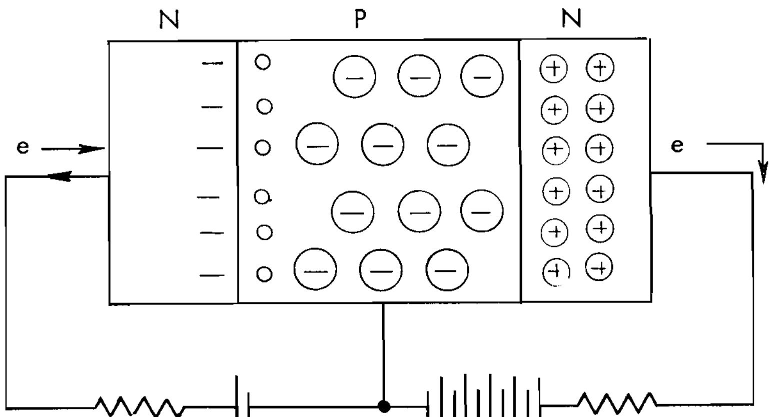


Figure 50. Equivalent Base-to-Emitter Barrier Due to Holes Trapped at the Emitter Junction

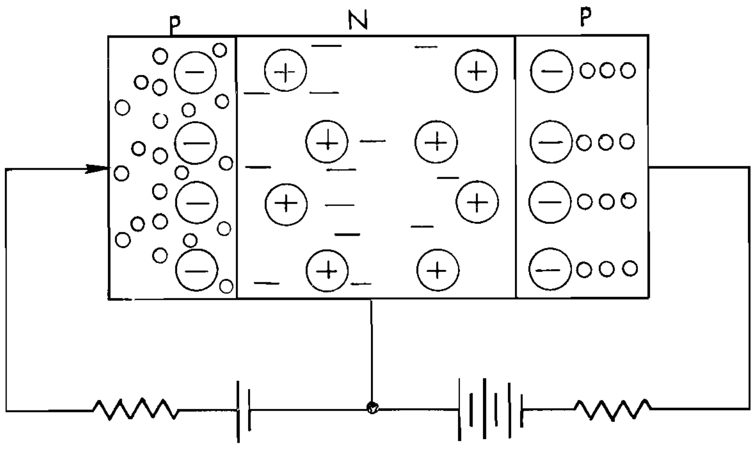


Figure 51. Forward Bias Drives Majority Carriers to the Barrier

Figure 50 is the simplified equivalent of Figure 49. The barrier looks as if majority carriers from both regions would cross the junction, but actually only the majority carriers from the emitter cross. They are, of course, attracted into the base region by the concentration of holes at the barrier. When they enter the base they become minority carriers and diffuse to the collector. Holes, on the other hand, are not driven into the emitter region. They were forced to the emitter originally by a high *B*-to-*C* bias and not by a forward *B*-to-*E* bias. Therefore, these trapped holes are influenced by both the concentration of majority carriers in the emitter and by the negative ion region in the base. Thus, they become trapped charges that permit the emitter to "turn on." Obviously, some recombination will take place, but its action has little effect on the *E*-to-*C* current that flows.

It should now be apparent that punch-through limits the amount of reverse bias that can be applied to a transistor. Also, the punch-through voltage value increases as the base thickness increases and as the concentration of base doping increases. But increasing the base thickness decreases the frequency response and increasing the concentration of doping decreases the current gain of the transistor (covered later). Therefore, high values of punch-through are not generally sought. It is only important that a certain minimum be met. Most transistors used now have a minimum punch-through value of 15–20 volts, although some special types are made to withstand a minimum of 60 volts.

PNP Current Flow

In previous illustrations, the operation of an NPN transistor was shown. Many people find the explanation of electrons being emitted into the base similar to the cathode action in vacuum tube theory and, therefore, easy for them to understand. Yet, they stumble when trying to understand the operation of the PNP in which the emitter contains majority carrier holes. How are holes emitted into the base? Carrier flow in a forward-biased PNP is presented here to answer this question.

Figure 51 shows the effect of forward bias when a PNP transistor is used. Notice particularly the *B*-to-*E* barrier. See how forward bias has driven majority carriers to the barrier and reduced the depletion region to zero. The equivalent barrier (after neutralizing barrier ions with majority carriers) is shown in Figure 52.

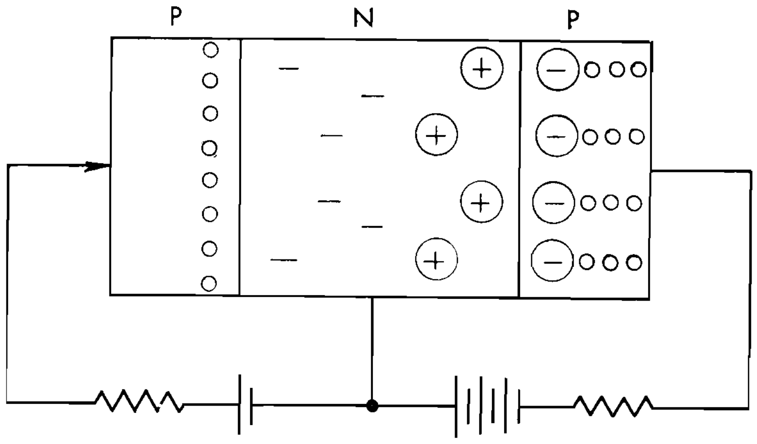


Figure 52. Forward Bias Reduces the Depletion Region to Zero

From Figure 52 it is clear that majority carriers are forward-biased to cross the junction; i.e., electrons from the base enter the emitter (conduction band current), and holes from the emitter enter the base (valence band current). But how does a hole enter the base? In the same way that a hole moves anywhere, that is, when an electron from a neighboring germanium atom swings from orbit about the germanium atom to an orbit in the hole location. Figure 52 also illustrates that the emitter provides the major current source because it is doped more than the base.

In Figure 53 the emitter action is shown. Electrons are shown crossing the junction to fill holes in the emitter. The effect of such a transfer is that holes now appear in the base, so that it can be correctly stated that the emitter emitted holes into the base.

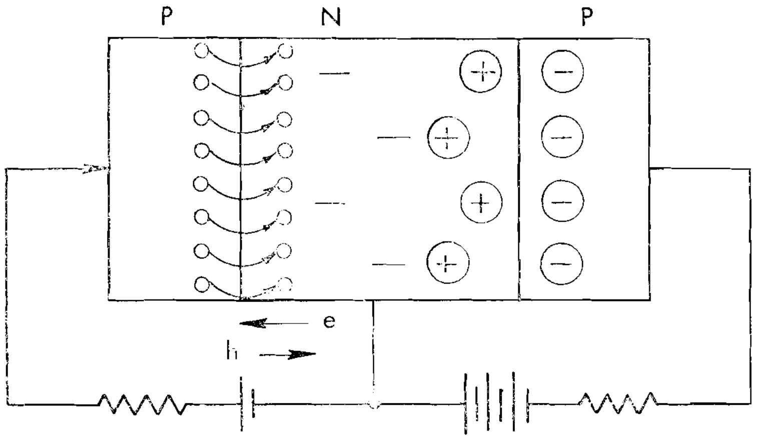


Figure 53. Forward Bias Drives Holes into the Base

Transit of holes through the base is shown in Figure 54. Holes entering the base become minority carriers and travel through the base region by diffusion. The illustration shows this atom-to-atom hole movement through the base and into the collector where the holes again become majority carriers and are strongly influenced by the negative source. Holes in the collector travel to the surface where they recombine with electrons delivered by the source. At the emitter terminal, captive electrons are released by the P-type impurity atoms. These electrons flow to the positive potential of the forward-bias source, and through it to the return of the collector source. Thus, electrons flow into the collector and out of the emitter. The uncovered atoms in the emitter then generate new holes which travel to the base region, and the current cycle continues.

Also notice in Figure 54 that some holes emitted never reach the collector because they recombine with electrons supplied by the negative base source. Thus, as in the NPN, the emitter current divides in the base to become I_{ce} and I_{be}.

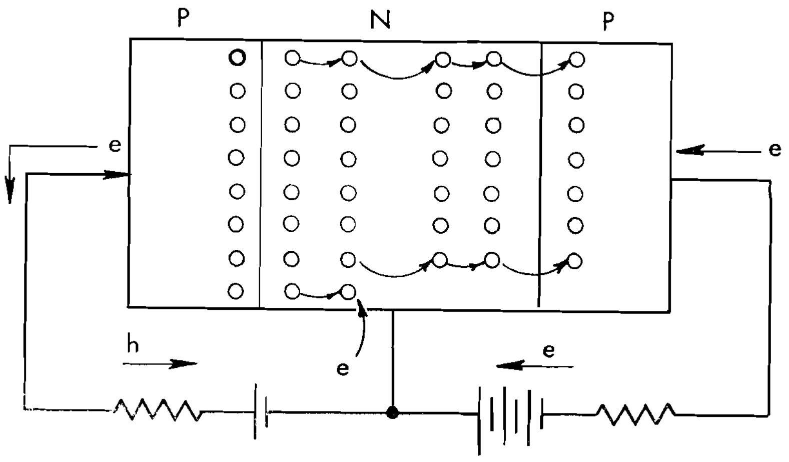


Figure 54. Minority Carrier Holes Diffuse to the Collector

Basic Circuit Configurations

Transistor circuits have three basic circuit configurations that are similar to the three basic tube circuit configurations. These are:

1. Grounded grid amplifier = grounded or commoned base.
2. Grounded cathode amplifier (inverter) = grounded or commoned emitter.
3. Grounded plate amplifier (cathode follower) = grounded or commoned collector.

Each of these circuits has certain characteristics which will be covered in detail in the study of each circuit. Each circuit has certain advantages over another. It is the utilization of these advantages which results in intelligent circuit design.

Grounded Base

The familiar grounded-grid amplifier is shown in Figure 55. In such an amplifier the signal voltage is fed to the cathode and the grid is held fixed or grounded. This circuit produces an output signal which is an amplified in-phase reproduction of the input signal. The transistorized version of this circuit is the grounded base shown in Figure 56. The circuit is so named because the input bias and the output bias are commoned to the base lead.

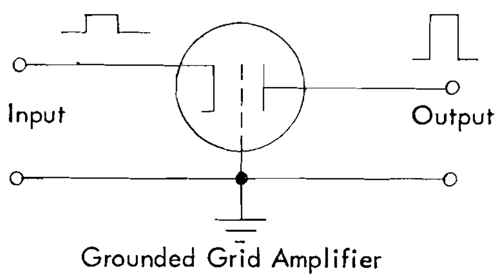


Figure 55. Grounded-Grid Amplifier Produces an In-Phase, Amplified Output Signal

Analysis of Figure 56 shows the following:

1. The B-to-E is forward biased.
2. I_e (emitter current) flows into the base where it divides into I_{be} (base-to-emitter current) and I_{ce} (collector-to-emitter current).
3. The input resistance (R_s) is small and the output resistance (R_l) is large.
4. This circuit produces a current amplification (ratio of output current to input current) of less than one (.98 shown), because some emitter current is lost to the base circuit $(I_{ce} = I_e - I_{be})$.
5. A small input signal (current through R_s) produces a large output signal (current through R_l).

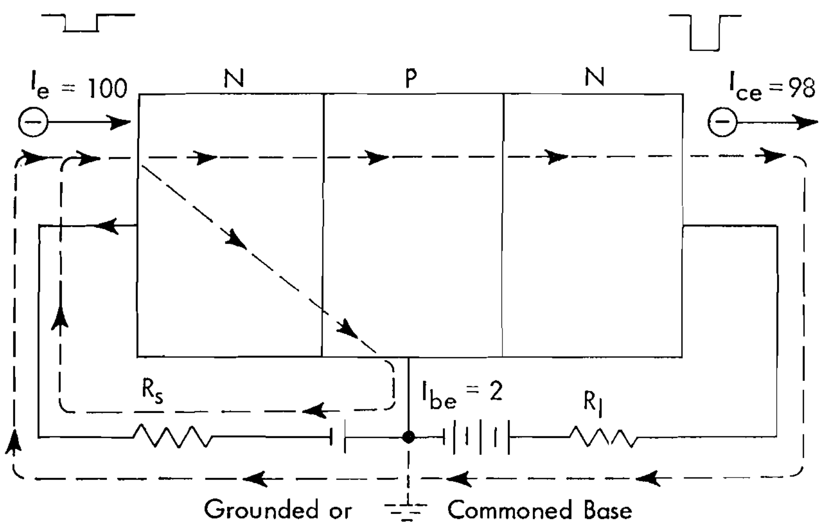


Figure 56. Grounded Base Produces an In-Phase, Amplified Output Signal

6. Because output current and input current are approximately equal, output voltage to input voltage is approximately equal to output resistance to input resistance. This ratio could be 100 to 1 or higher. Thus, this circuit is an excellent voltage amplifier.

Current flowing in the collector circuit is called I_c (Figure 57). It consists of I_{ce} and I_{co} (*B*-to-*C* reverse current). I_{co} is a small current and is a function of junction temperature, not the potential applied. Normal signal levels (up or down) have little effect on I_{co}, which is therefore a fixed amount (a constant). Thus, the output signal (change of signal level) is due to the change in I_{ce} and is not affected by I_{co} which shows little to no change.

A reverse-biased grounded base circuit and the resulting current paths were previously shown in Figure 28.

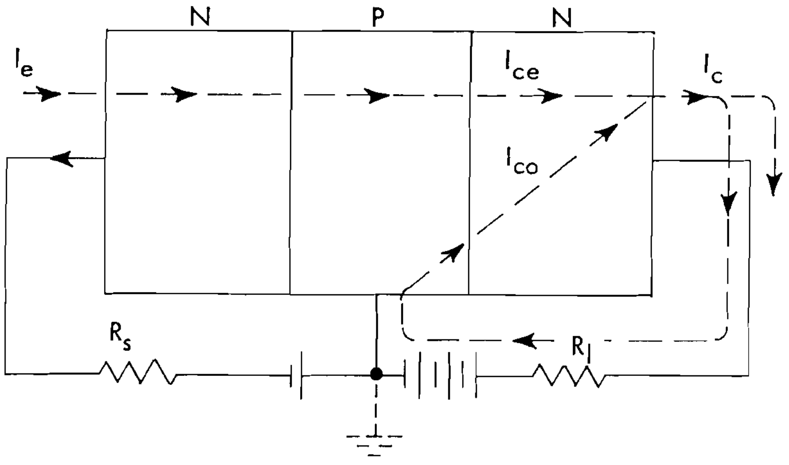


Figure 57. Output Current $I_c = I_{ce} + I_{co}$

Characteristic Curves

The operating behavior of a transistor is best described by its V_cI_c current characteristics shown in Figure 58. These curves are plotted in the following manner:

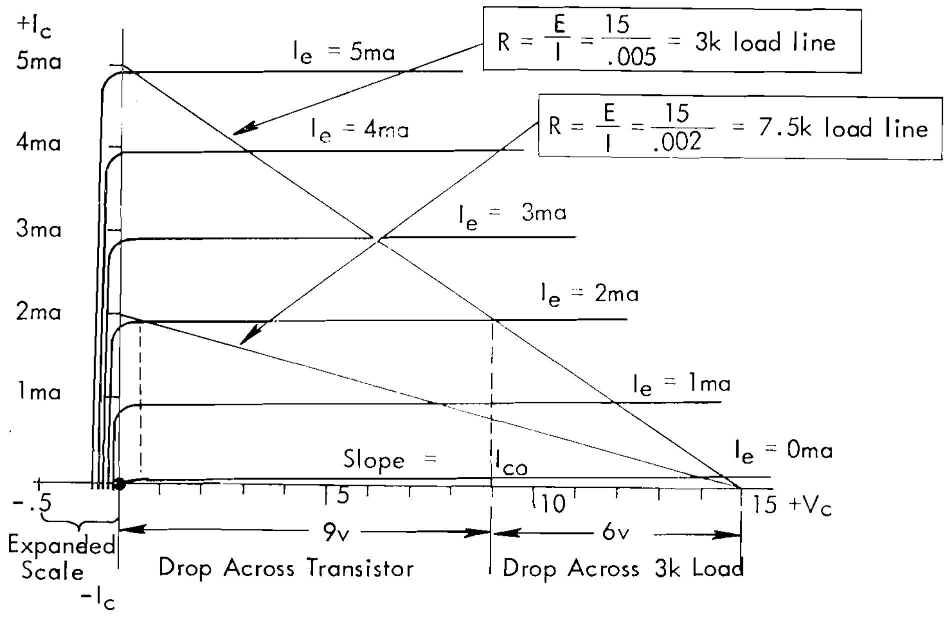


Figure 58. V_cI_c Characteristics for an NPN Ground-Base Circuit

1. A current generator is connected to the emitter and is adjusted so that a fixed emitter current flows.
2. A current measuring device in the collector circuit records I_c for various values of collector voltage (V_c).
3. These values are then plotted on a horizontal (voltage axis) and a vertical (current axis).

The V_cI_c characteristics show that:

1. With zero emitter current, $I_c = I_{co}$.
2. I_c is slightly less than I_e.
3. The value of V_c is relatively unimportant as far as I_c is concerned; i.e., the same collector current flows for a low value of V_c as for a high value of V_c. In such curves, V_c is applied directly to the collector so that only the transistor's inherent characteristics are being graphed. Although, in this case, V_c does not affect I_c, this is not true in a circuit where the collector contains a load resistor. In a loaded circuit, the value of V_c determines the maximum collector current that can flow. This action is called saturation and is discussed later.

These curves are used in the same manner as are the plate family of characteristic curves (E_p vs. I_p) used in tube analysis. Probably the most important use of these curves is associated with a load line graph. For example, when a particular load resistor is plotted (Figure 58), the following information is recognizable:

1. The amount of *IR* drop across the load resistance when a specific input current flows. For example, a 6-volt drop exists across a 3k load when 2 ma of input current flows.

2. The amount of *IR* drop across the transistor when a specific input current flows into a specific load resistor. For example, a 9-volt drop exists across the transistor when 2 ma flows into a 3k load.
3. The swing in output voltage resulting from a specific input current swing. For example, verticals dropped from the 1 ma and 4 ma intersections of the 3k load line show an output voltage swing on the horizontal axis of approximately 9 volts.
4. An operating point is located for a specific class of operation. For instance, for class A operation, the operating point is the midpoint on the load line.
5. The point of saturation is shown. The saturation point of a transistor is the point at which the total reverse-bias voltage is developed across the load resistance, so that additional input currents do not produce a further increase in output current. In Figure 58 the transistor is saturated when more than 5 ma of input current flows into a 3k load, and is also saturated when more than 2 ma of input current flows into a 7.5k load.
6. When a specific input current flows, the output current is reduced to zero by slightly forward biasing the *B*-to-*C* diode to make the collector an emitting source. Thus, when the amount of current leaving the collector equals the amount of current reaching the collector, the collector current equals zero. This is why the emitter current plots are shown falling to zero only after the collector potential is slightly forward-biased.

Load Line

A load line is obtained in one of two ways.

1. When the value of load resistor is known, the reverse-bias voltage is divided by the load resistor to obtain the maximum output current. For example, a 15-volt bias divided by a 3k load gives a 5 ma current flow. A line connecting the 5 ma vertical point with the 15-volt horizontal point is the 3k load line.
2. When the maximum output current is known, the reverse-bias voltage is divided by the output current to obtain the load resistor. For example, a 15-volt bias divided by a 5 ma output current results in a 3k load resistor. In this case, the line is drawn from the values given and the *E/I* solution gives the value of this load line.

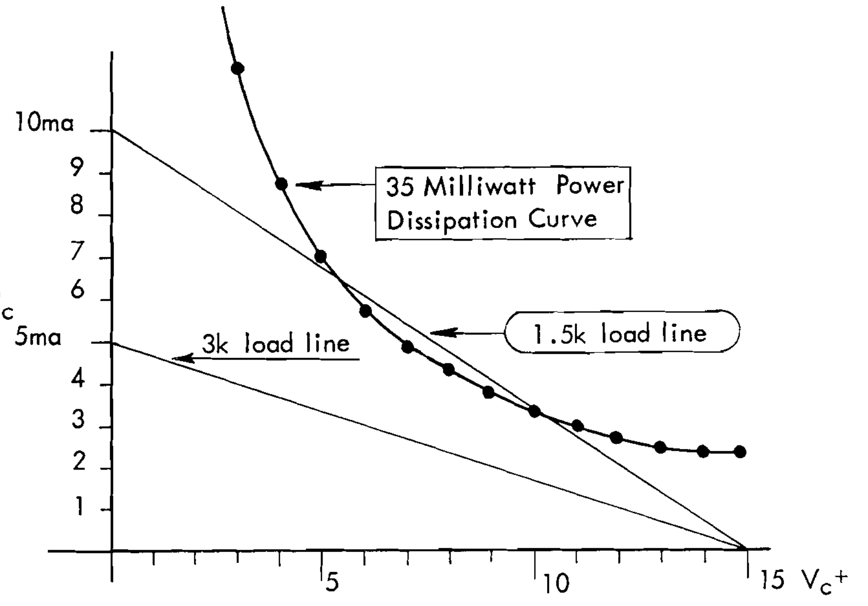
Power Dissipation Curve

The V_cI_c curves are also used to plot a power dissipation curve (Figure 59). This curve is drawn as follows:

1. The power rating is obtained from the transistor specification sheet.
2. The power rating (35 milliwatts, shown in Figure 59) is then divided by each voltage value on the horizontal axis and these values are charted. Each such value obtained tells us the maximum safe operating current for a value of voltage drop across the transistor.
3. The values obtained are located on the V_cI_c curves, and a curve is drawn through all points.

Such a curve locates, on the V_cI_c curves, the safe operating range of the transistor. The area to the left of this curve means that the transistor is operating at less than 35 milliwatts and the area to the right means that 35 milliwatts is exceeded. Exceeding the power rating, of course, will damage the transistor. Now by drawing load lines on the V_cI_c curves, it is easy to see if the power-handling capacity of the transistor is exceeded. For example, in Figure 59, the 3k load line is in the safe region, while the 1.5k load is in the safe region part of the time and in the danger region part of the time. Consequently, the 3k resistor is a safe load while the 1.5k is not.

Because the 1.5k load line exists in both the safe and danger regions, it can be used safely in a switching circuit whose steady state level (up or down) never exists in the danger region. Such a circuit is safely used when the switch time is very rapid; that is, the amount



Vc	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ic	35	17	11.7	8.8	7	5.8	5	4.4	3.9	3.5	3.2	2.9	2.7	2.5	2.3

$P = EI$

$I = \frac{P}{E} = \frac{35mw}{5} = \boxed{7ma}$

Figure 59. Power Dissipation Curve Plotted on V_cI_c Characteristic

of time that the transistor is conducting in the danger region is very short.

Current Gain

For many transistors the current gain parameter is the most important specification. The current gain of a transistor really means how much of the emitted current reaches the collector. This parameter is called alpha (α) and is defined in Figure 60 as a change in I_{ce} divided by a change in I_e when V_c is held constant. To illustrate what this means, let us say that a change in I_e of 100 current carriers results in a change in I_{ce} of 98 current carriers; then, the α would be .98 as shown in Figure 60. Naturally, if the change in I_{ce} had been only 95, then the α would be .95. Actually α is a ratio of output current to input current as measured in a grounded-base circuit. This ratio for a three-element alloyed-junction transistor is always less than one, and is a measure of a transistor's amplifying capabilities.

Current gain of a grounded base is called (alpha)

$\alpha = \frac{\Delta I_{ce}}{\Delta I_e}\Big/ V_c$ Thus if $\left(\begin{matrix} I_e = 100 \\ I_{ce} = 98 \end{matrix}\right)$ then $\alpha = \frac{98}{100} = .98$

Figure 60. Current Gain Formula

It should be noted that α is the current gain measurement of the grounded-base circuit only. We shall see later how this specification is converted to α' to determine the current gain of a grounded emitter or a grounded collector circuit, which have current gains up to 200 to 1.

Power Gain

The basic criterion of a vacuum tube circuit is voltage gain, while in a transistor circuit it is the power gain. Power gain takes into account both the current gain and the voltage gain characteristics. It is this product which is the basic criterion of transistor circuit performance. Power had no meaning in tube circuits because the input power was approximately zero. This is not true of a transistor circuit which always requires some input power; to get output current, input current must also flow.

The power gain formula is defined in Figure 61 as the output power divided by the input power. To see what this really means, again refer to Figure 61. The output power here is I^2R_l and the input power is I^2R_s. Now, because the alpha of this transistor is .98, the output current is approximately equal to the input current. By cancelling out the I^2 quantities, the power gain of a grounded-base circuit is reduced to approximately the ratio of output resistance to input resistance.

A grounded base produces a "Power Gain"

$P = IE = I(IR) = I^2R$

$P_{gain} = \frac{P_{out}}{P_{in}} = \frac{I^2R_l}{I^2R_s}$

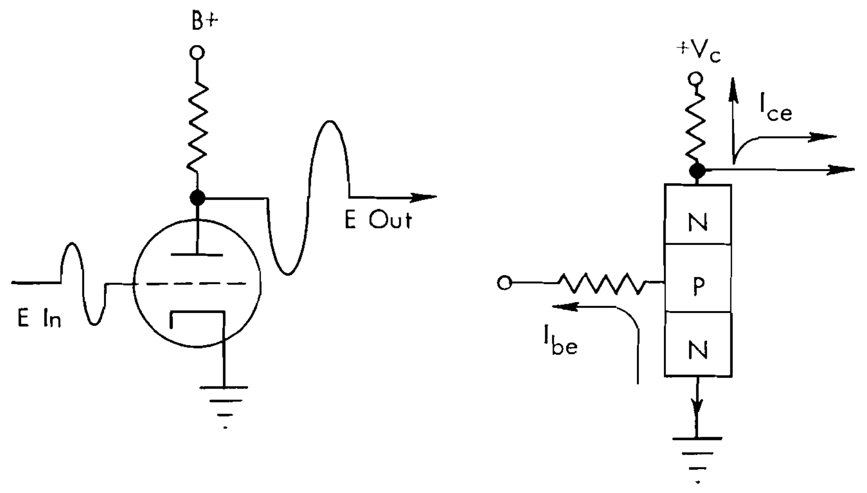
Since $I_e \cong I_{ce}$ then $\frac{I^2R_l}{I^2R_s}$ reduces to $\frac{R_l}{R_s}$

Figure 61. Power Gain Solution for a Grounded Base Circuit

Voltage Function vs. Current Function

Transistors are current-operated devices and tubes are voltage-operated devices (Figure 62). This is true, but what, exactly, do we mean? After all, analysis of a transistor circuit shows that an input signal (voltage) results in an output signal (voltage). Is not this exactly what a tube circuit does? No, they only appear to be the same. Actually, they are quite different, and the information that follows is presented so that the expression "current-operated device" will be clear. Once understood, the V_cI_c characteristics become meaningful.

First of all, again study Figure 62. Here we see that a tube is operated as a function of voltage; a voltage on the grid controls a current flow through the tube. The grid, of course, does not draw current unless it is overdriven and its voltage level is more positive than that of the cathode. Thus, in tube operation, the output signal is controlled by a grid-to-cathode bias (voltage difference) and is so plotted in tube manuals.



Tubes operate as functions
of voltage = f (E)
As E_{in} increases, E_{out} increases
As E_{in} decreases, E_{out} decreases

Transistors operate as functions
of current = f (I)
As I_{be} increases, I_{ce} increases
As I_{be} decreases, I_{ce} decreases

Figure 62. Operation Function of Tube vs. Transistor

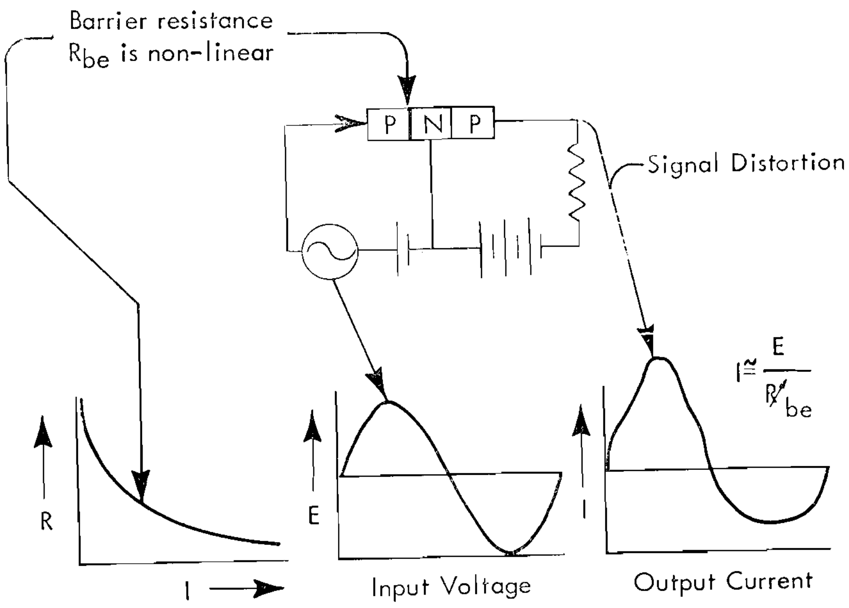


Figure 63. Input Voltage Produces a Distorted Output Signal

Now compare the tube circuit to the transistor circuit. The transistor is shown here as a function of current, because the output current I_{ce} is directly related to how much input current I_{be} (not voltage) is fed into the base. Voltage cannot be used as a reference because of the non-linear resistance of the B-to-E diode; i.e., the diode exhibits proportionately more resistance to low currents than to high currents. For instance, it may have a resistance of 50 ohms when 1 ma flows and only 10 ohms when 5 ma flows. So if voltage, instead of current, were used as the input reference, the output current curves would be non-linear, and of little use. Input current, on the other hand, does produce linear output current curves, and for this reason is used as a reference. If it is not now clear why the transistor is a current-operated device, let us try another approach. First, study a transistor circuit in which the input signal is voltage (Figure 63), and then compare this circuit with one in which the input signal is current (Figure 64). Figure 63 shows that:

1. The only input resistance is the B-to-E diode resistance R_{be}.
2. R_{be} is non-linear as shown in the plot. R is maximum when I is minimum, and R is minimum when I is maximum.
3. A signal voltage is fed to the B-to-E diode.

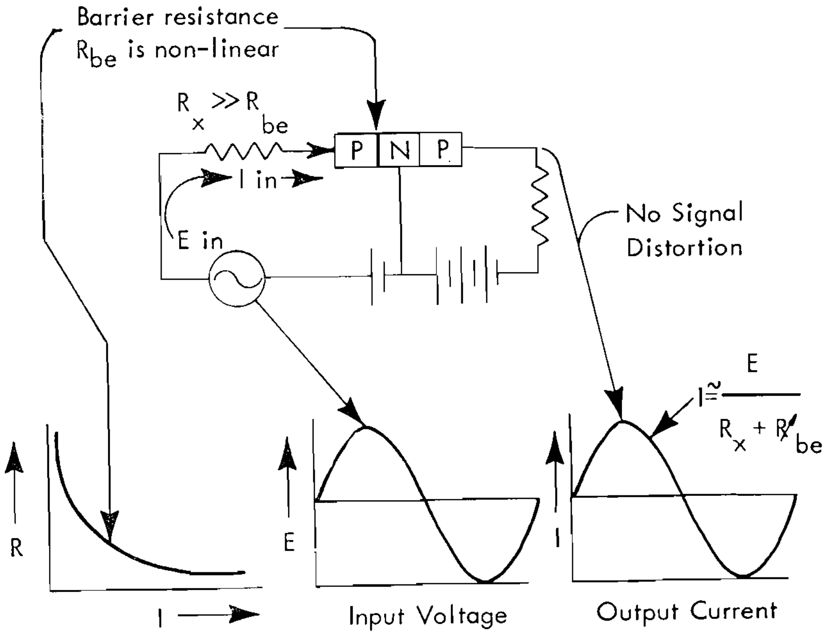


Figure 64. Input Current Produces an Output Signal without Distortion

4. Output current flows because of the input voltage. The input voltage causes carriers to be emitted into the base, after which they diffuse to the collector and constitute output current.
5. Output current is not a true reproduction of the input voltage because the input resistance (B-to-E) is non-linear.
6. Distortion of the output signal results because the high input voltage levels produce current peaks and the low input voltage levels produce current limiting. In other words, I is not proportional to E because R_{be} is a variable.

Distortion of the output signal is eliminated by converting the input signal voltage to an input signal current. This is easily accomplished by inserting a series resistor in the input circuit which is much greater than the nominal value of R_{be} (Figure 64). A value of this series resistor R_x could be 100 to 1000 ohms. The signal voltage, E, is now effectively converted to signal current because the input resistance is primarily R_x; R_{be} has negligible effect. So the transistor is not being fed a voltage level, but rather a specific value of current is fed to the emitter. This is what is meant by E_{in} and I_{in} in Figure 64.

Of course, the distortion-free output obtained is desirable, but it is gained at a price. Actually, to get it, the input losses become greater; a greater input signal voltage is required to get the same output current because of the IR loss through R_x.