APPLICATION NOTE 18

INTRODUCTION TO SOLID STATE DEVICES

by

Bernard M. Oliver

1. CONDUCTORS, SEMICONDUCTORS, INSULATORS AND ENERGY GAPS.

In a crystalline solid, the atoms forming each crystal are held together in a rigid lattice by "chemical bonds". These bonds consist of electron pairs which are "shared" by adjacent atoms. Electrons in bonds are in a lower energy state than would be the case if the bond were broken -- i.e., it requires work to pull apart the electrons in a bond, and hence to separate the atoms so bonded. When a liquid freezes it gives off heat. To melt it again this same heat must be added. The energy represented by this much heat is just the energy needed to break all the chemical bonds (between atoms in an element, between molecules in a compound if the compound does not decompose on melting).

In a metal, the bonds are formed by certain of the outer electrons, but other electrons even less tightly bound to their nuclei do not participate in the bonds, and in the solid state are not even bound at all, but are free to roam throughout the solid. Thus each atom of a typical metal casts loose one or more electrons to form a sort of "electron gas" free to flow thru the crystal lattice. It is this charged gas which conducts electricity by drifting thru the lattice when a field is applied.

The gas as a whole cannot escape the metal for it would leave it oppositely and terrifically charged and the enormous electrostatic forces would quickly pull the electrons back. It is difficult for even one electron to leave since the remaining charges arrange themselves to look like a positron as far below the surface as the escaping electron is above the surface. The electron must overcome the attraction of this image force to escape, and thus do an amount of work known as the work function. At high temperatures occasional electrons are energetic enough to make their getaway and we have thermionic emission.

In a semi-conductor all the outer electrons of each atom are involved in chemical bonds. A free electron is the exception rather than the rule. The elemental semi-conductors: carbon (as diamond), silicon, and germanium all belong to the IVth column of the periodic table. (Other semi-conductors involve compounds of elements from the IInd and IVth columns, such as cadmium sulphide. These will not be considered, though their behaviour is quite similar.) There are thus four electrons in the outer shell and each is shared with an adjacent atom. While the actual structure is tetrahedral, i.e., the four atoms surrounding a given atom lie on the corners of a tetrahedron, nothing essential will be lost and a lot of clarity will be gained if we imagine the structure to lie all in one plane as shown in Fig. 1.

HEWLETT-PACKARD COMPANY
1501 PAGE MILL ROAD
PALO ALTO, CALIFORNIA, U.S.A.
CABLE: "HEWPACK" TEL. (415) 326-7000
54 ROUTE DES ACACIAS
GENEVA, SWITZERLAND
CABLE: "HEWPACKSA" TEL. (022) 42.81.50

PRINTED IN U.S.A.

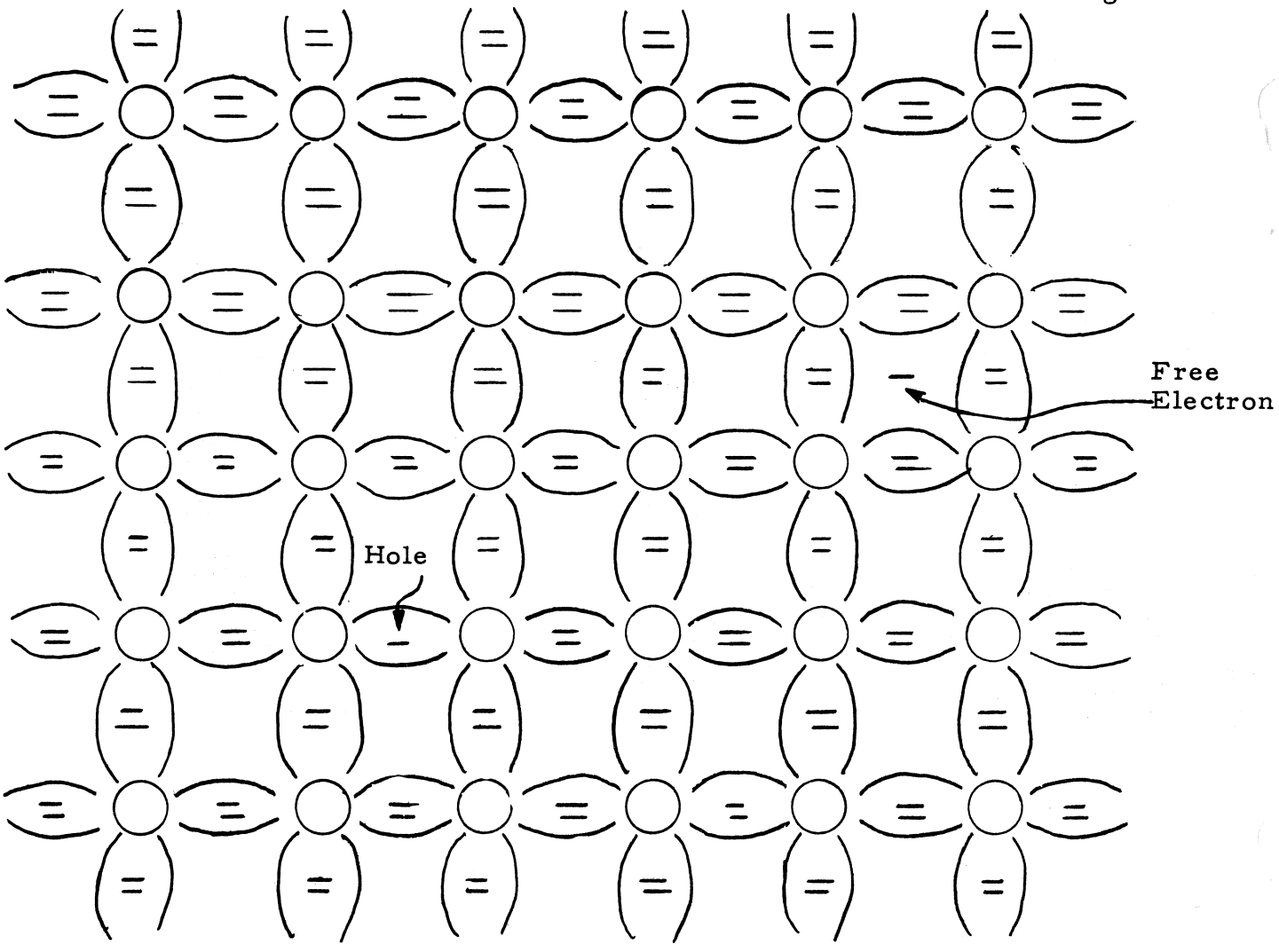


Fig. 1

Here the circles represent the nuclei and inner electrons of each atom, the parenthesis, the bonds, and the - signs, the outer electrons, two of which are as a rule shared in each bond. In pure Si or Ge at absolute zero each bond would have two electrons, there would be no free electrons, and the crystal would not conduct electricity at all.

Now the energy an electron has is the sum of its potential and kinetic energies. An electron on a negatively charged body has energy even at rest. If it escapes the body the field will accelerate it and convert this potential energy to kinetic energy. Similarly in a solid, there are potential and kinetic energies for each electron - energies of position and velocity, if you like. The electrons in the valence bonds have low energies of position. They are all

down in wells of various depths. Just as it would require energy to pull Pluto out of the solar system, or to separate two binary stars, so it requires energy to free an electron from a bond. By the same token a free electron is in a relatively high energy state. In addition to being free, an electron can have unlimited kinetic energy, but to be free it must have at least a certain total energy.

Thus in a pure crystal the situation is as follows: There is a range of total energies which electrons in valence bonds may have. This range is called the valence band (of energy)*. There is an open range of energies above a certain minimum energy which free electrons may have. This range is called the conduction band because free electrons can conduct electricity. We now come to one of the facts germane to germanium (and to all solids). Between the top of the valence band (i.e., the most energy a bound electron can have) and the bottom of the conduction band (i.e., the least energy a free electron can have) there may be a gap (i.e., a range of energies which no electron may have). The energy levels for semiconductors are about as shown in Fig. 2. In such a solid an electron, in order to escape a valence bond and become a carrier of electricity, must acquire at least enough energy to cross the gap. What is more, this cannot be done by easy stages so to speak. That is, an electron cannot acquire part of the requisite energy now and wait for the rest later, for this would imply that it could exist temporarily in the forbidden no-man-land.

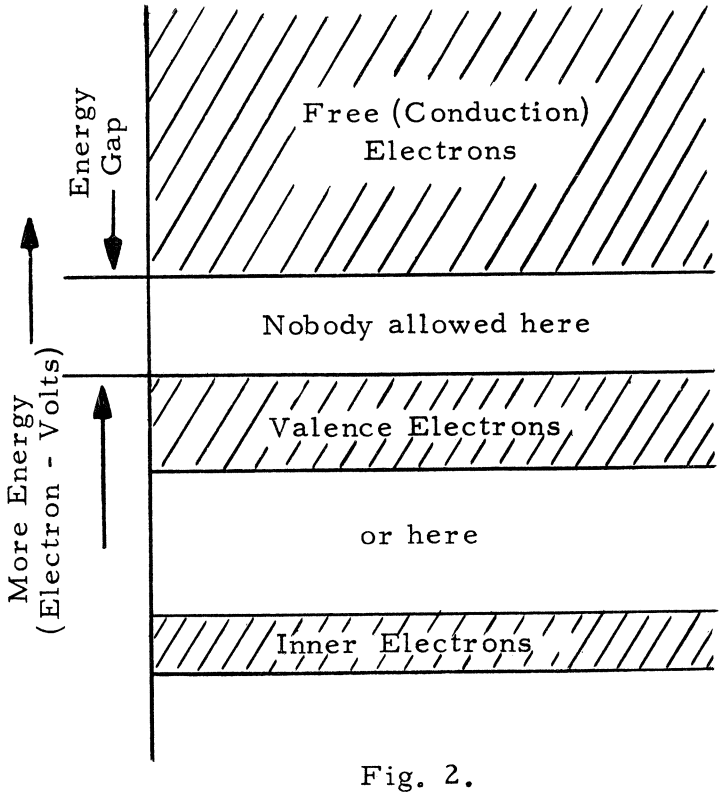


Fig. 2.

If the energy gap is not too great some electrons will occasionally acquire enough thermal energy to cross the gap. However, the distribution of thermal energies at room temperature is such that for each 50 millivolts** increase in gap the number of thermally excited free electrons decreases by 1/e. The energy gaps, free electrons/cm^3, and resistivities for some Group IV elements are given on the following page.

* Inner shell electrons lie in bands having even lower energies.

** Really milli-electron-volts. An electron-volt is the energy an electron acquires in accelerating thru a 1 volt potential difference.

Material	Energy Gap	Free Electrons/cm^3	Ohm - cm
Diamond	6 to 7 volts	10^{-35}	10^{46}
Silicon	1.1 volts	7×10^{10}	64,000
Germanium	0.72 volt	2.5×10^{13}	47
Tin	0	10^{23}	10^{-5}

Thus we see that whether a solid is a conductor, a semi-conductor, or an insulator depends upon its energy gap. In the metals the energy gap is zero, in fact the conduction and valence bands overlap. As a result there are tremendous numbers of free electrons. In a semi-conductor the energy gap is a few tenths of a volt and thermal energies are able to "ionize" a few bonds (about one in 10^9 in Ge, one in 10^{12} in Si) and thereby supply enough charge carriers to permit appreciable conduction. In insulators the gap is several volts and thermal energies just never get that high.

2. HOLES AND ELECTRONS.

When an electron acquires enough energy to jump out of a valence bond - to tear free from its moorings - it is free to drift under the influence of an electric field and thus conduct electricity, as we have said. What is not so obvious is that the bond vacated by the electron - a half-full bond with only one electron rather than two - is also free to roam around the crystal. The way this happens is that an electron from an adjacent full bond moves over. Then another electron jumps into the bond thus vacated and so on. It is as if a man left his seat in a theatre. Then the man in the next row back took his seat, the man next to him moved over and so on. In this way the empty seat can wander all over the house without anyone moving more than one seat. Thus when an electron leaves its seat, the empty seat is apt to wander off too.

The empty seat - the half occupied bond - is called a hole. When a hole exists in a bond between Ge or Si atoms the nuclear charge of the adjacent atoms is not completely neutralized and a net positive charge equal and opposite to the charge of one electron is present. Thus moving a hole in one direction produces the same current as moving an electron in the opposite direction.

Oddly enough, holes have about the same mobility as the free electrons, that is they drift about as fast under an applied field, but in the opposite direction. (This causes current in the same direction.) They also wander around to about the same extent under the influence of thermal agitation. As a result

each ionized bond produces <u>two</u> carriers of electricity: one electron and one hole, both about equally effective.

The motion of holes and electrons due to thermal agitation is the classical Brownian Motion or "random walk", in which the net displacement of a particle from its starting point is equally likely to be in any direction and of an amount which increases as the square root of time. If a field is present a uniform drift is superimposed on the random motion.

3. PHOTOCONDUCTION.

If a quantum of light has more energy than the energy gap it will be absorbed by an electron in a valence bond. The extra energy thus acquired sets the electron free and forms a hole. The energy in a photon is inversely proportional to the wavelength. Thus if light shorter than a critical wavelength falls on a thin wafer of silicon or germanium, hole - electron pairs are formed, and the conductivity increases.

In germanium and silicon the critical wavelengths are in the infra red at about 1.7 and 1.1 microns respectively. However the thermal generation of hole electron pairs is high in Germanium and considerable light is required to double the carrier density and thus double the current over its dark-current value. Silicon is much better in this respect. Diamond has virtually no dark current, but the critical wavelength is at about .18 microns (1800 A°) which is in the ultra-violet.

If an impulse of light is applied to a photo conductor, hole-electron pairs are formed instantly and photo conduction begins abruptly. Conduction then dies off as the holes and electrons either recombine or reach the ends of the cell. Thus the impulse response of a photoconductor resembles a decaying exponential (Fig. 3), and the cell behaves as if it had an RC time constant. The shorter the carrier lifetime, the shorter the time constant will be. However the initial response will be the same. Thus the shorter the lifetime the less the area, and the d.c. sensitivity is reduced. Actually, photoconductors have time constants in the millisecond to second range, and the total charge passed as a result of an impulse of light is many times that corresponding to the number of hole-electron pairs generated. The reason for this is a kind of regeneration involving "trapped" holes, but this action need not be discussed here.

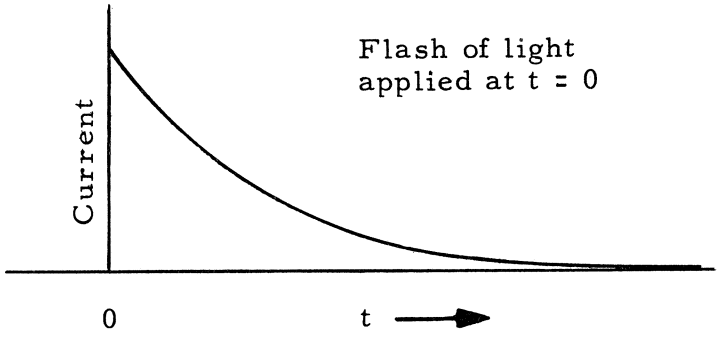


Fig. 3.

4. DOPING.

As a crystal of silicon or germanium is growing, the atoms of Si or Ge lose energy at the (cooler) crystal surface, the electrons form their bonds with the atoms already attached and the new atoms take their place in the lattice. If impurities are present in the melt, an occassional impurity atom will replace a Si or Ge atom and the resulting crystal will contain a sprinkling of these impurity atoms in the lattice.

If the impurity is a Group V element such as Phosphorus, Arsenic, or Antimony, it will have five valence electrons rather than four. Thus after the four covalent bonds have been formed with the adjacent Si or Ge atoms, there will be one electron left over, for each Group V atom in the lattice. Corresponding to this extra electron there is an extra + charge in the nucleus of the Group V atom. However, in the crystal this extra electron is only very lightly bound by this charge and at room temperature easily acquires enough thermal energy to break free. Thus the impurity centers are almost completely ionized at room temperature and each Group V atom donates a free electron to the crystal. The Group V elements are thus called donors and the crystal formed with them is called n-type material since it has an excess of negative carriers.

Group III elements, such as Boron, Aluminum, Gallium, and Indium have only three electrons. Thus when they freeze into the lattice one bond per impurity atoms will have only one electron - i.e., it will contain a hole. Corresponding to this missing electron there is one less positive charge on the nucleus of the impurity atom. When an electron from elsewhere fills the hole (and the hole wanders off) the region around the Group III impurity is left with a net negative charge. In spite of the attraction this charge exerts on the (positive) hole, the thermal agitation at room temperature is great enough to carry away the hole and the impurity atom ends up accepting an electron into the hole it came with. For this reason Group III elements are called acceptors and since the material they produce has an excess of holes (positive carriers) it is called p-type material.

The addition of controlled amounts of donors or acceptors is called doping.

A silicon crystal contains about 5×10^{22} atoms/cm^3. In pure silicon at room temperature there are about 7×10^{10} free electrons and 7×10^{10} holes per cm^3. Thus in undoped silicon there are

$$\frac{1.4 \times 10^{11}}{5 \times 10^{22}}$$

or 2.8×10^{-12} carriers per atom. If now the silicon is doped with only one impurity atom in 10^9 silicon atoms the conductivity will be increased about 300 fold. Doping is thus a process involving very minute amounts of added impurities, and therefore does not materially affect such things as rates of thermal hole-electron pair generation or hole and electron mobilities which still depend on the properties of the overwhelmingly more abundant silicon and germanium atoms, and on the lattice.

5. MASS ACTION AND CHARGE CONSERVATION.

Since every free electron born creates a hole, the birth rate for electrons and holes must be the same. Similarly, since the only way they disappear is by combining one-for-one the death rates must be the same. For the population to be stable the birth and death rates must be equal. Thus, in equilibrium, the rate of generation equals the rate of recombination. Since the rate of generation is independent of the population while the rate of recombination increases with population the population adjusts itself until the rates are equal.

Let

r = rate of generation and recombination

p = density of holes

n = density of electrons

τ_p = average lifetime of holes

τ_n = ditto for electrons

n_i = density of either carrier in pure material

Each density must be the product of the birth rate and lifetime. Thus,

$$\left.\begin{aligned} p &= \tau_p r \\ n &= \tau_n r \end{aligned}\right\} \qquad (1)$$

We note in passing that the ratio of lifetimes is the same as the ratio of concentrations. Now the lifetime of either carrier is inversely proportional to the concentration of the other, so we can write $\tau_p = \frac{K}{n}$ or $\tau_n = \frac{K}{p}$. Substituting, we find from either equation,

$$pn = Kr \qquad (2)$$

Now K and r are independent of the degree of doping, and in pure material $p = n = n_i$. Thus $n_i^2 = Kr$ and we have

$$pn = n_i^2 \qquad (3)$$

This is the <u>mass action</u> law. Increasing the concentration of either carrier depresses that of the other, the product remaining the same as for pure material.

When a donor has given up its electron, there is a bound positive charge left (in the nucleus). Likewise when an acceptor has accepted its electron into a bond the nucleus is short one positive charge compared with the bound electrons around it, so in effect there is a bound negative charge centered at the nucleus.

Thus if,

N_a = density of acceptors

N_d = density of donors

the sample will carry no net charge if

$$p - n = N_a - N_d \tag{4}$$

Equations (3) and (4) specify both p and n.

We find,

$$\left.\begin{aligned} p &= \sqrt{n_i^2 + \left(\frac{N_a-N_d}{2}\right)^2} + \left(\frac{N_a-N_d}{2}\right) \\ n &= \sqrt{n_i^2 + \left(\frac{N_a-N_d}{2}\right)^2} - \left(\frac{N_a-N_d}{2}\right) \end{aligned}\right\} \tag{5}$$

$$p + n = \sqrt{4n_i^2 + (N_a-N_d)^2} \tag{6}$$

It is important to notice in these equations that it is only the <u>excess</u> of one doping agent over the other that is significant. This is a fact of great practical importance, for were this not so it would be almost impossible to make p-n junctions or transistors.

When $\frac{N_a-N_d}{2n_i} \gg 1$ as is ordinarily true

we find

$$\begin{aligned} p &\approx N_a-N_d \\ n &\approx \frac{n_i^2}{N_a-N_d} \end{aligned} \tag{7}$$

Thus under appreciable doping the <u>majority carrier density</u> approaches the excess density of doping centers, while the minority carrier density is depressed by the mass action law.

The above equations have assumed $N_a-N_d > 1$. If $N_a-N_d < 1$, interchange p and n, and the subscripts a and d.

Equation (4) and therefore (5), (6), and (7) are only true everywhere for a homogeneous sample. They are not true for example near a p-n junction as we shall see. (They <u>are</u> true however a short distance from the junction.) Equation (3) is true for any sample in thermal equilibrium.

6. THE p-n JUNCTION.

If the doping of a melt is abruptly changed from excess acceptor doping to excess donor doping as the crystal is being grown, the crystal will be p-type material up to a certain plane and n-type beyond. This plane is called a p-n junction. Remembering that the excess acceptors in the p-type material and the excess donors in the n-type number usually less than one atom in a million it seems incredible that a place where almost-nothing changes could have such startling properties. Yet this is the case.

Let us analyze qualitatively what happens at an abrupt p-n junction. Let us assume that instead of being grown as a single crystal the junction were created by bringing together a piece of p-type and a piece of n-type material both with surfaces so plane and so clean and se perfectly aligned that they bonded together on contact. At the instant of contact the carrier densities would be as shown in Figure 5(a). (This figure assumes the p-type material to be more heavily doped.) Immediately upon contact, diffusion will set in.

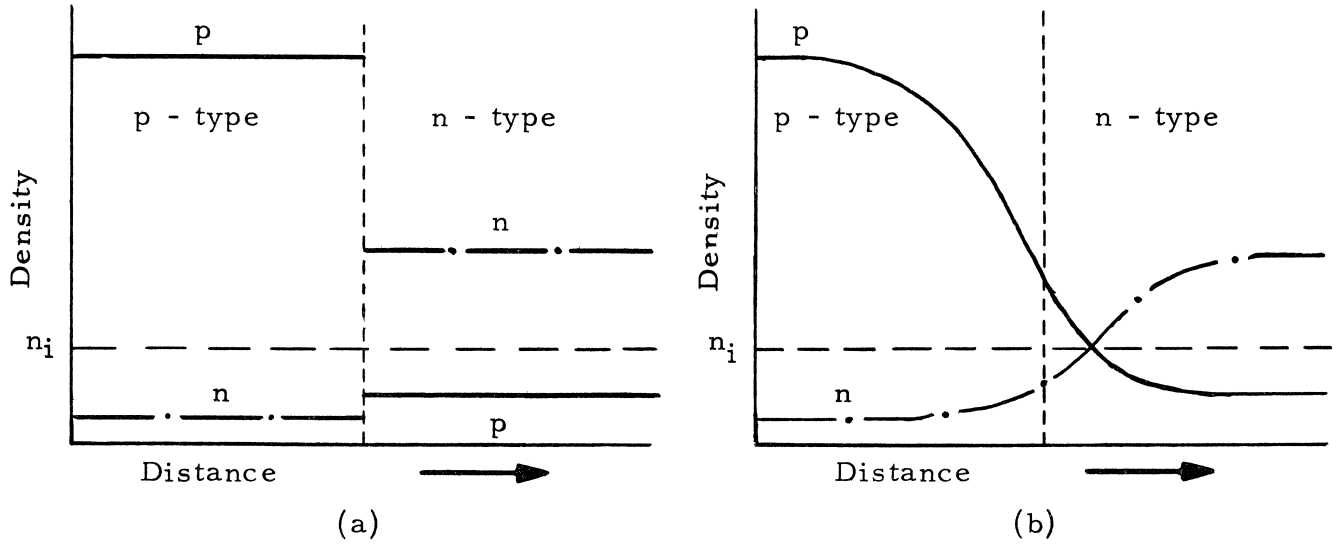


Fig. 5.

Quicker than you can say Shockley, the high concentration of holes in the p-type material will cause holes to spill over into the n-type material, and the high electron density in the n-type material will cause extra electrons to spill into the p-type material. If this diffusion continued unchecked the p and n densities would level out to new values uniform throughout the specimen and corresponding to the average doping. A p-n junction would then have no remarkable properties. What actually happens is that the diffusion creates a net space near the junction. This space charge creates an electric field across the junction which restrains the majority carriers of each side from diffusing without limit into the other side.

The way this field comes about is shown in Fig. 6. The total space charge consists of three components:

(1) A positive space charge, q_p, due to the holes.

(2) A negative space charge, $-q_n$, due to the free electrons.

(3) A space charge $q(N_d-N_a)$ due to the bound charges at the impurity centers. ($-q$ is the charge of an electron.)

The carrier densities shown in Fig. 5a are such as to produce charge neutrality everywhere. When diffusion decreases p and increases n to the left of the junction a net negative space charge builds up on this side. Likewise when n decreases and p increases to the right of the junction a positive space charge builds up there. These charges distributions are shown in Fig. 6a and correspond to the density distributions shown in Fig. 5b.

Between these two layers of space charge there is an electric field. This field is the integral with respect to distance of the space charge. It is shown as negative in Fig. 6b which means that it is directed to the left. In equilibrium, this field just balances the diffusion everywhere. The tendency of electrons and holes to diffuse from regions of high concentration into regions of low concentration is everywhere offset by the superposed drift from the electric field. The space charge layer is ordinarily limited to a distance of a few tenths of an mil or less on either side of the junction. Outside the space charge layer the carrier

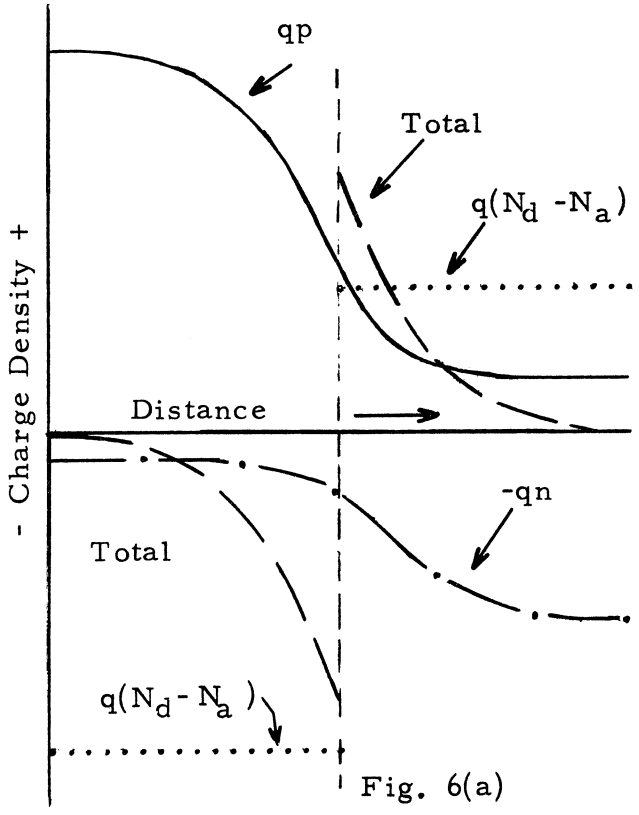


Fig. 6(a)

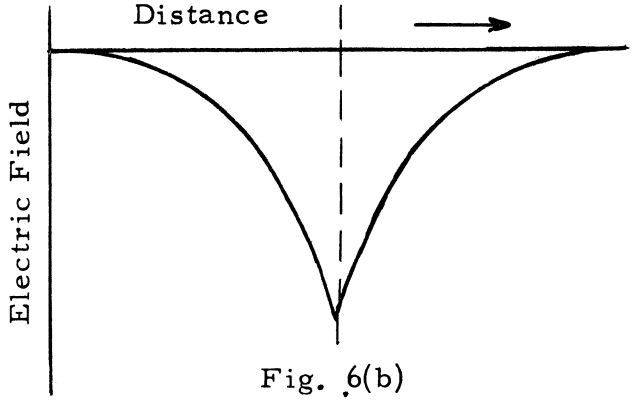


Fig. 6(b)

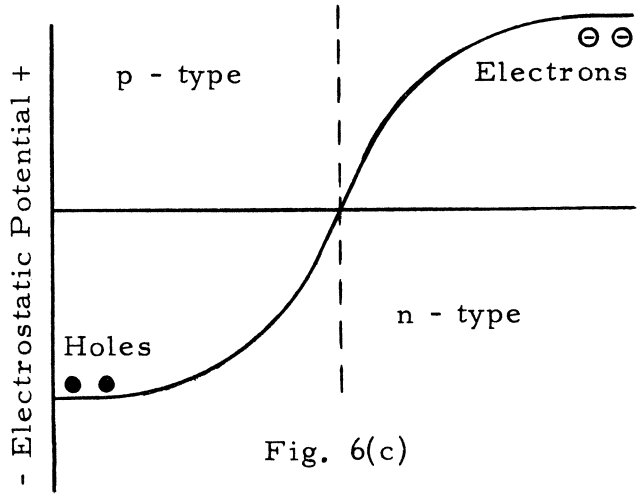
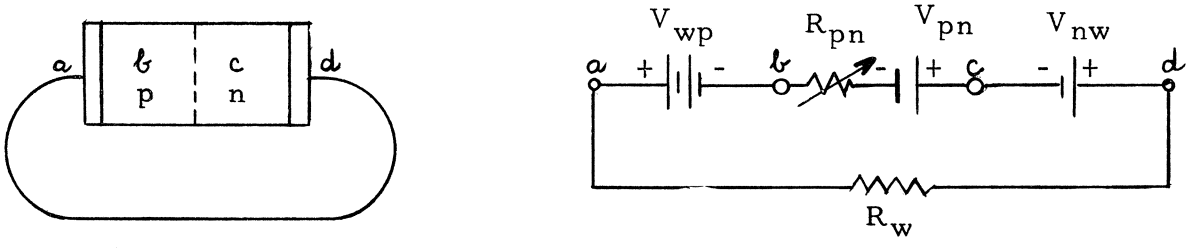


Fig. 6(c)

densities are the same as they would be in homogeneous material of the same doping.

The electric field at the junction means there is a potential difference between the p and n sides of the junction. Since the potential difference is the negative of the integral of the electric field, the n-type material is positive with respect to the p-type as shown in Fig. 6(a). With low doping on both sides, the potential difference will be small. With very high dopings it can approach the energy gap of the material. With usual doping concentrations it runs about .3 volts for germanium and about .5 volts for silicon.

In view of this potential difference and the finite impedance of the device it begins to look as if the p-n junction is a battery from which we should be able to draw power. If this were possible the device would defy the second law of thermodynamics, in that we would have a power source in a system in thermal equilibrium. While p-n junctions are remarkable they are not that good. Suppose we close a circuit around a p-n junction with a piece of resistance wire as shown in Fig. 7a. No current flows.



(a) (b)

Fig. 7.

The reason is to be found in Fig. 7b which shows the equivalent circuit of Fig. 7a. At the p-n junction there is a junction (or contact) potential V_{pn}. At the junction of the n-type material and the wire there is a contact potential difference V_{nw}, and finally at the junction of the wire with the p-type material there is a contact potential V_{wp}. The sum of these potential differences around the loop is zero. Alas.

Referring again to Fig. 6c, the potential difference there depicted constitutes a sort of hill which holes in the p-side must climb to get over to the n-side. We may think of the holes as little marbles on the floor at the bottom of a slide, and bouncing around with various kinetic energies. Occassionally a marble gets enough kinetic energy to roll up the slide clear to the top. Likewise the electrons may be thought of as little balloons under the slide at the top. Occassionally one of these balloons gets enough energy to slip down the under side of the slide to the bottom, in spite of its buoyancy. There is thus

a forward component of current, I_e , due to energetic majority carriers from both sides fighting their way across the junction against the field there. The distribution of energies is such that the number that make it falls off exponentially with the height of the hill, i.e., with the total potential drop across the junction. The total potential across the junction is the contact potential less any external forward bias voltage V. Thus we should expect I_e to increase exponentially with forward bias. It turns out this is true, and

$$I_e = I_o e^{\frac{qV}{kT}}$$

where

-q = charge on the electron

k = Boltzmann's constant

T = absolute temperature

V = applied voltage (positive for forward bias)

We shall say more about I_o in a moment.

In addition to the above forward component of current, there is a reverse component of current, I_s, due to minority carriers from both sides which diffuse to the junction and get swept across by the field. Since the minority carrier densities outside the space charge region are independent of applied voltage, I_s is also.

Thus we have two components of current one voltage dependent, the other not. The total current is their difference. Thus the total current is

$$I = I_e - I_s = I_o e^{\frac{qV}{kT}} - I_s.$$

With no applied voltage (V = 0), the total current I must be zero, so $I_o = I_s$ and we find

$$I = I_s \left[e^{\frac{qV}{kT}} - 1 \right] \qquad (8)$$

For reverse bias, V is negative, and the first term of (8) quickly becomes negligible. Thus the reverse current saturates to a constant value, I_s, (whence the subscript.)

At room temperature $\frac{q}{kT} = 38.6/\text{volt}$ so

$$I = I_s (e^{38.6V} - 1) = I_s (10^{17V} - 1) \qquad (8a)$$

Under forward bias the -1 soon becomes negligible and from then on the current increases by a factor e with every additional 26 millivolts of forward bias - goes up 10 to 1 every 60 millivolts. It can be shown that this is as sharp a rectification characteristic as is theoretically possible at the temperature T.

Differeniating (8) we find

$$\frac{dI}{dV} = \frac{q}{kT} e^{\frac{qV}{kT}} = \frac{q}{kT} (I + I_s) \tag{9}$$

Thus the (dynamic) resistance, $R = \frac{dV}{dI}$, of the junction is $R = \frac{1}{\frac{q}{kT}(I + I_s)}$

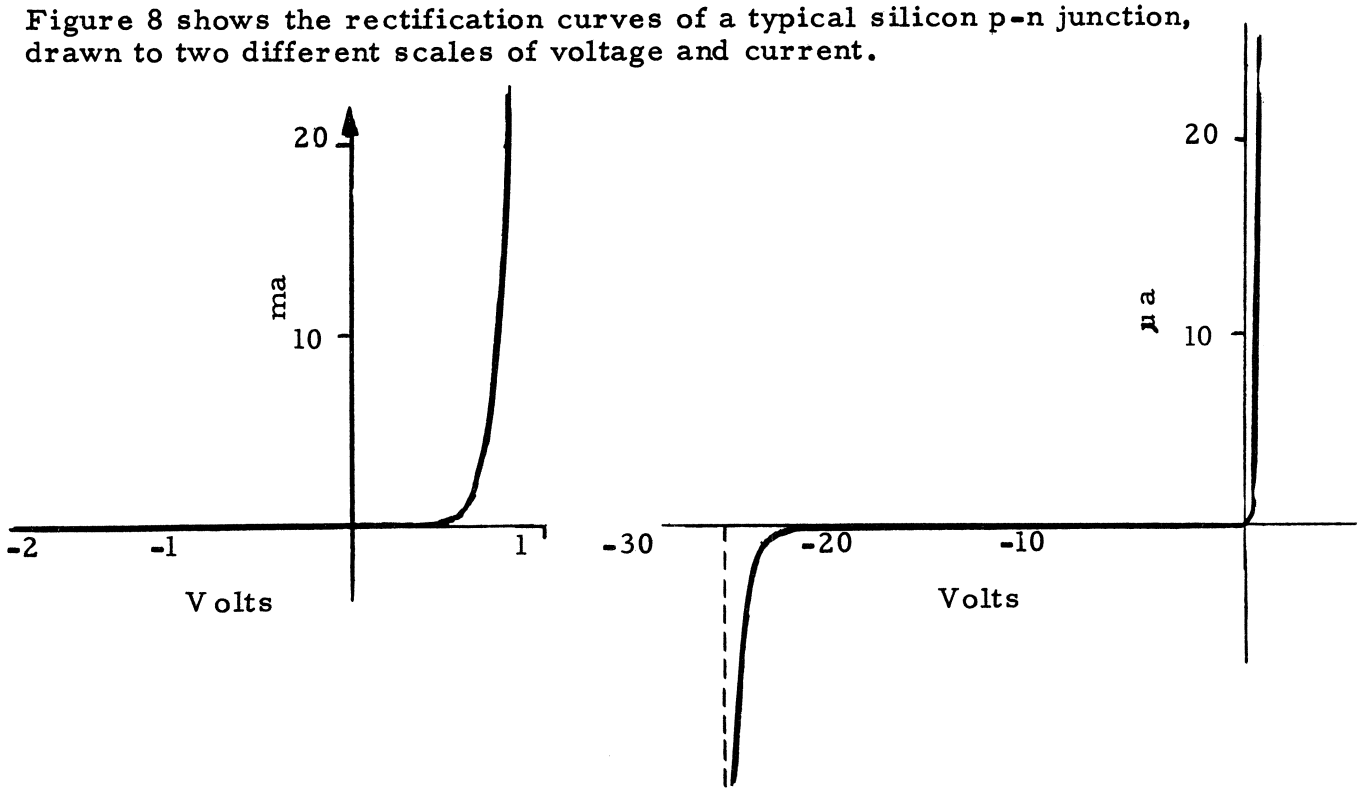
If we express current in milliampere, this becomes

$$R = \frac{26}{I + I_s} \text{ ohms.}$$

For forward currents of 0.1 milliamps or more I_s may ordinarily be neglected.

As stated earlier I_s depends on the minority carrier densities in the p and n type materials. These densities are $n_p \approx \frac{n_i^2}{Na}$ and $p_n \approx \frac{n_i^2}{Nd}$. I_s therefore varies as n_i^2 and since n_i varies exponentially with temperature, I_s will do likewise. As a result the reverse saturation current of an ideal p-n junction should double about every 8°C for germanium and every 5°C for silicon. This is found to be the case with germanium. In silicon at ordinary temperatures the ideal saturation current is so low that the reverse current is dominated by surface leakages and other effects which vary less rapidly with temperature.

Figure 8 shows the rectification curves of a typical silicon p-n junction, drawn to two different scales of voltage and current.



(a) Fig. 8. (b)

At sufficiently high reverse voltages, the reverse current begins to increase rapidly with voltage. Originally this reverse breakdown was thought to be a sort of field emission - a tearing of electrons out of their bonds - as a result of the high electric field in the junction, and was called Zener breakdown after Clarence Zener who studied such effects. It is now believed to be an avalanche action similar to breakdown in gases. Electrons swept across the junction from the p-side are able to acquire enough energy in one "mean free path" to ionize other bonds thus creating new hole-electron pairs and the action becomes regenerative.

7. TRANSIENT EFFECTS IN p-n JUNCTIONS.

In figure 9 the hole and electron densities as a function of distance are shown for three conditions of bias. Curves (f) are for forward bias, curves (o) are for no bias, and curves (r) are for reverse bias.

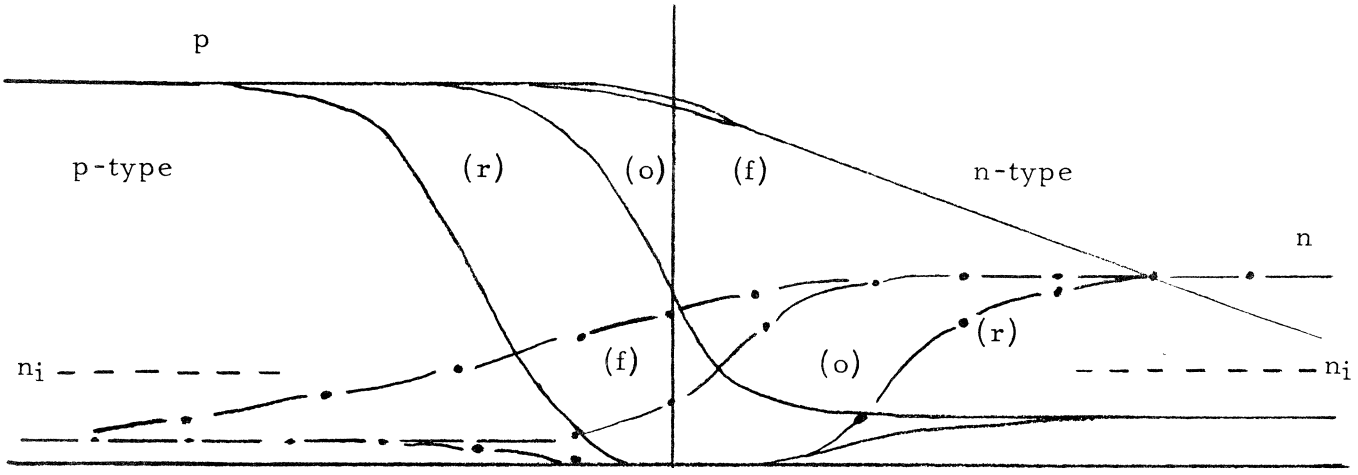


Fig. 9

Under reverse bias (r), the densities of both carriers in the vicinity of the junction are lower than in the unbiased case (o), and this depletion of carriers extends over a greater distance. As a result the space charge layers (from the bound charges) are stronger and thicker than with no bias. Thus the electric field across the junction is stronger and thicker under reverse bias. The integral of this field - the potential difference across the junction - is correspondingly greater, as it must be to include the applied voltage.

As the reverse voltage is varied the thickness and strength of the space charge layers will vary. Thus there is charging and discharging of the opposite faces of the junction, much as in the plates of a condenser, and indeed the junction exhibits a capacitance across it. However the capacitance varies with the applied voltage. The reason for this is that unlike a condenser, in which the plate separation is fixed, we have in the junction a capacitor in which the effective plate separation increases with voltage (as the depth of each layer increases).

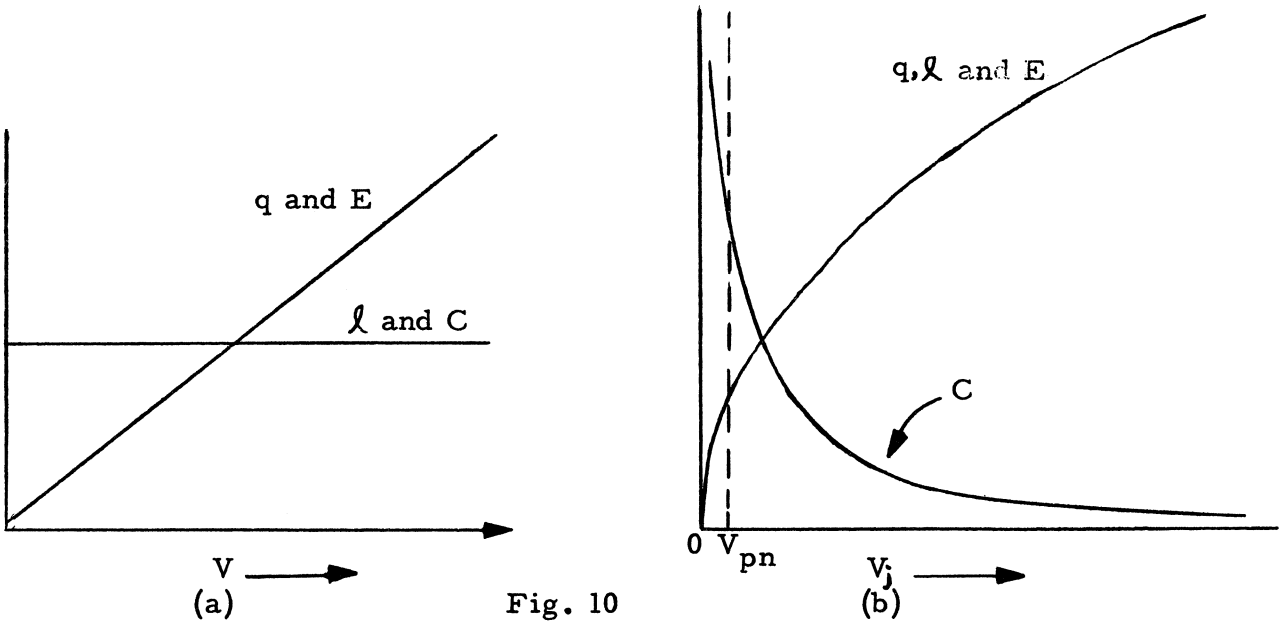


Fig. 10

Figure 10a shows the variation of the total charge, q, the field strength, E, the length of the field, ℓ, and the capacitance, C, with applied voltage in a condenser. Fig. 10b shows the same things for a junction. In the latter case q, ℓ, and E vary as $V_j^{\frac{1}{2}}$ while C varies as $V_j^{-\frac{1}{2}}$. Thus E x ℓ and $\frac{q}{C}$ both vary as V_j as must be true. In these relations V_j is the <u>total</u> potential drop across the junction, i.e., V_j = V applied + V_{pn} where V_{pn} is the contact potential.

When the junction is forward biased the densities of holes in the n-type material and of electrons in the p-type material are greatly increased near the junction as shown in curves (f). If the bias is suddenly reversed, these carriers will be pulled back across the junction and the reverse current will not decay to its steady state value until this action, together with normal recombination process, have decreased the carrier densities to those shown by curves (r). As a result under sudden application of reverse bias, after forward conduction has been taking place, a pulse of reverse current will occur, as shown in Fig. 11. This effect is often called "carrier storage" because the high p and n carrier densities in the material of opposite types represent a store of carriers which will supply reverse conduction and which must be exhausted before full back resistance can develop.

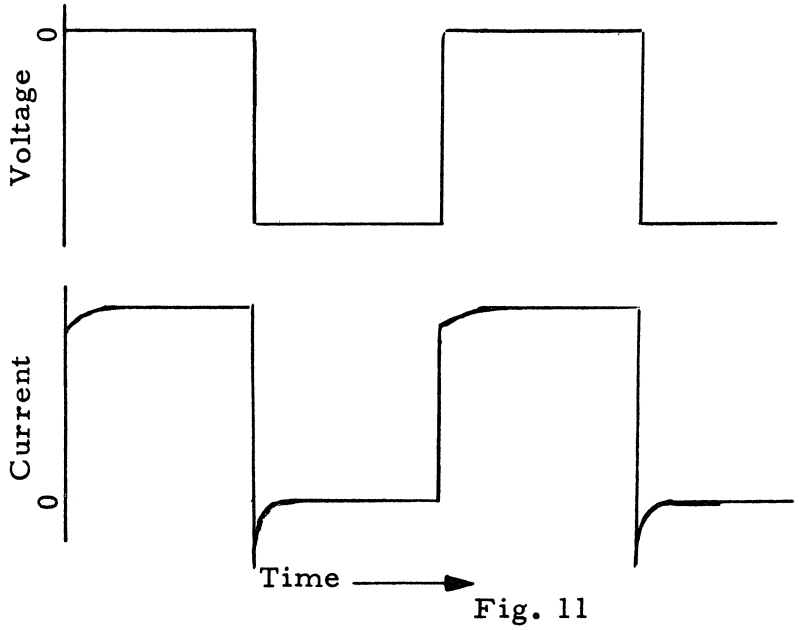


Fig. 11

The effect of junction capacity and carrier storage limit considerably the high frequency performance of p-n

junction diodes. The effects are less the less the junction area. As the area of a junction is decreased, the reverse current should also decrease proportionaly, while the voltage for a given forward current should only increase logarithmically. Thus techniques for producing microscopic true p-n junctions are constantly being sought.

8. THE SOLAR BATTERY.

When light falls on a p-n junction, the hole electron pairs created by the light are swept out of the junction by the field present there. Holes are swept into the p-type side, electrons into the n-type side. This, of course, constitutes a current from the n-side to the p-side. If the junction is short circuited externally, the potential drop across the junction cannot change. Thus all the normal thermal currents in the junction (which add to zero) are unchanged, and the extra photo current will flow through the external circuit from the p-side to the n-side. This current is proportional to the incident light intensity.

On the other hand, if the junction is open circuited, there is no external return for this current and the voltage across the junction must drop in order that the photo current may be returned as forward current in the junction. Referring to Fig. 7b, with $R_w = \infty$, we see that if V_{pn} decreases a potential difference equal to this decrease will appear between points a and d with a being positive with respect to d.

The voltage-current equation for an <u>illuminated</u> junction is

$$I = I_s \left[e^{\frac{qV}{kT}} - 1 \right] - I_L \qquad (11)$$

where I_L is the short circuit photo-current. The curves for various light levels are just the universal diode curve displaced downward by amounts proportional to the light intensity as shown in Fig. 12a.

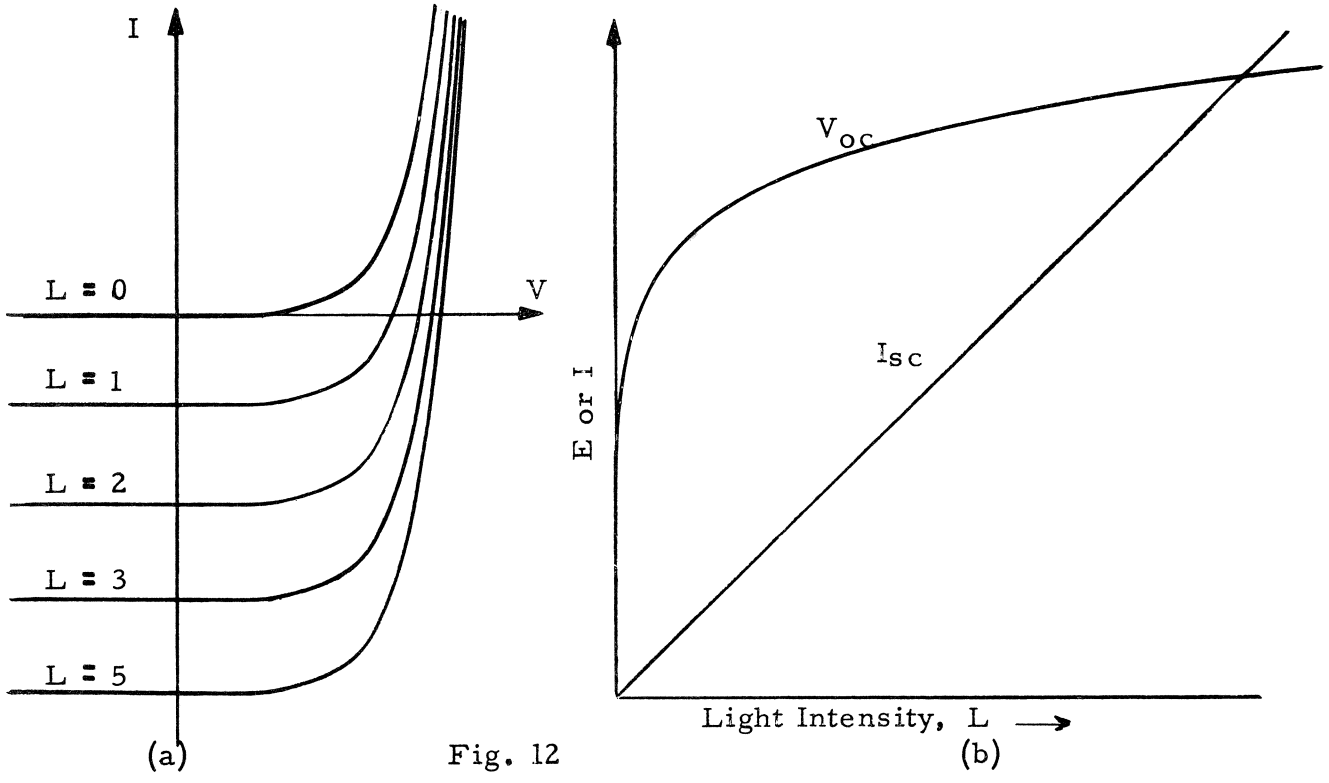


Fig. 12

Since the curves now extend into the fourth quadrant power can be taken from an illuminated p-n junction.

Figure 12b shows the short circuit current (y-axis intercepts of curves of Fig. 12a) and the open circuit voltage (x-axis intercepts of curves of Fig. 12a) both plotted a function of light intensity, L. It will be seen that short circuit current varies linearly with the light, while the open circuit voltage varies logarithmically with $L + L_o$ where L_o is the light intensity required to make $I_L = I_s$. Thus for high light levels, the cell tends to develop a nearly constant voltage, and has an available current proportional to the illumination. In typical cells the output power is a maximum when $V \approx .6\ V_{oc}$ and this maximum power is approximately $.8 V_{oc}\ I_{sc}$.

Solar batteries are made by converting the surface of n-type silicon to p-type by diffusing Boron into the crystal to a depth of several mils. On the back side this surface layer is etched off and ohmic contacts made to both the n-body and p-surface. The result is a large area junction which is illuminated through the thin p-layer, as shown in Fig. 13.

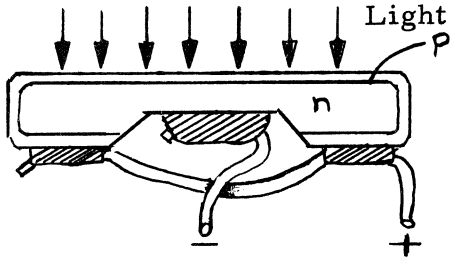


Fig. 13

If the light were monochromatic with a wavelength just shorter than the critical value for which the energy per photon is equal to the energy gap, the cell would be a highly efficient energy converter. With a broad spectrum such as sunlight, much of the energy is at a longer wavelength and produces no photo-current while much is at a shorter wavelength and the excess energy per photon is wasted. These losses together with surface losses and I^2R losses in the thin p layer reduce the actual "solar battery" efficiency to about 12%.

9. JUNCTION TRANSISTORS.

A junction transistor is a pair of p-n junctions back to back, made from a single crystal of germanium or silicon. It may consist of a very thin n-type layer sandwiched between two p-type layers or of a very thin p layer sandwiched between two n layers. The two types are called p-n-p and n-p-n respectively. Their operation is essentially similar except for polarity of applied voltages and direction of current flow. Since these polarities and current directions in the n-p-n type are the same as for a vacuum tube we will choose this type for discussion as being less confusing.

N-p-n transistors are usually made by altering the excess doping in the melt as a single growing crystal is slowly being pulled out. At the start the melt has a slight excess of n-type impurities. Thus the first part of the crystal to grow is lightly doped n-type. This part will be the collector

(plate) of the transistor. Although the light doping makes the collector region have a comparatively high resistivity, the collector junction is normally back biased, so this resistance is negligible compared with the back resistance of the junction. Then p-type doping agent is added to the melt in sufficient quantity to convert the material which is then freezing to p-type. Almost at once a large excess of n-type doping agent is added to the melt, so that only a thin layer of p-type material freezes. This thin layer is the base (grid) of the transistor. The remainder of the crystal is heavily n-doped and is the emitter (cathode) region. The position of the thin p-layer in the grown crystal is then found by electrical measurements and a thin wafer sawed out which include this p-layer. This wafer, or sandwich, is then diced into a hundred or more small squares and connections are made to the three layers. These are transistors. They must yet be sealed from moisture (by encapsulation) to assure long life but already they are operable transistors.

In normal operation, the collector is biased positively with respect to the base. This back biases the base-collector (p-n) junction. If there is no connection to the emitter the only current to flow will be the saturation current, of this junction. This current is called I_{CO} meaning the collector current with no emitter current. If the collector voltage were to go negative with respect to the base, a large current would flow into the collector, for this would represent forward bias of this junction. If we call the current positive when it flows into the collector from the supply (as we do in the case of plate current) the collector characteristic with no emitter current is the universal diode curve inverted by our conventions of sign as shown in Fig. 14.

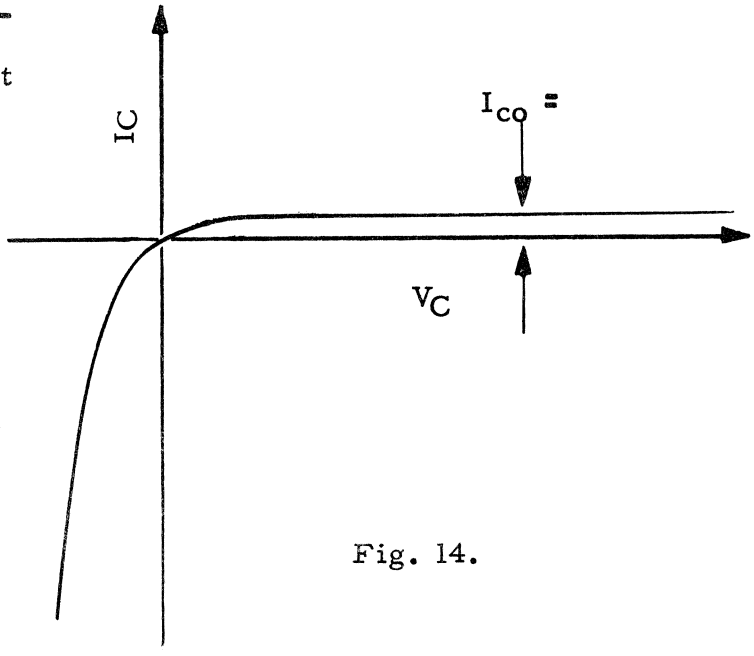


Fig. 14.

Exactly the same characteristic obtains for emitter current versus emitter voltage. If the emitter is negative with respect to the base the emitter-base (n-p) junction is forward biased and forward current is drawn out of the emitter. If no connection is made to the collector, this current is all drawn out of the base.

We now come to the heart of transistor action. <u>If the collector is held positive</u> with respect to the base while the emitter is run negative so that emitter current flows, <u>95 to 99% of the emitter current will be drawn not from the base but from the collector</u>. The transistor is very similar in its characteristics to a very high μ positive grid triode. If the plate (collector) is left unconnected, and the cathode (emitter) is negative with respect to the grid (base) a large cathode current will be drawn. If, however, the plate is held

positive most of the cathode current will be drawn from the plate, and very little from the grid. While the grid still serves to promote emission from the cathode, most of the electrons (n-carriers) emitted by it miss the grid, and passing through the openings of the grid, find themselves in a strong field which accelerates them to the plate. In the transistor, because of the extremely heavy doping of the emitter, most of the emitter current is carried by electrons from the emitter flowing to the base rather than by holes from the base flowing to the emitter. These electrons upon crossing into the base find themselves in a very thin p-layer. Before they can drift along this layer and out the base lead, most of them diffuse across the thin layer to the collector junction where the strong space charges field sweeps them into the collector region.

The fraction of the emitter carriers collected by the collector is called α. More precisely,

$$\alpha = \left.\frac{\Delta I_c}{\Delta I_e}\right|_{V_c = \text{const.}} \tag{12}$$

Typically α will range from 0.9 to 0.98. It may drop a few percent at high currents. But with a positive collector and moderate currents it is quite constant. (The corresponding factor in a positive grid triode is the ratio of plate current increment to cathode current increment.)

The effect of the emitter current drawn from the collector is to shift the curves of Figure 14 upward bodily by an amount αI_e. Thus we get the typical junction transistor collector family of curves shown in Fig. 15. It is noteworthy that the collector current does not drop to zero until the collector goes negative, i.e., until the forward current of the collector-base junction equals $\alpha I_e + I_{co}$.

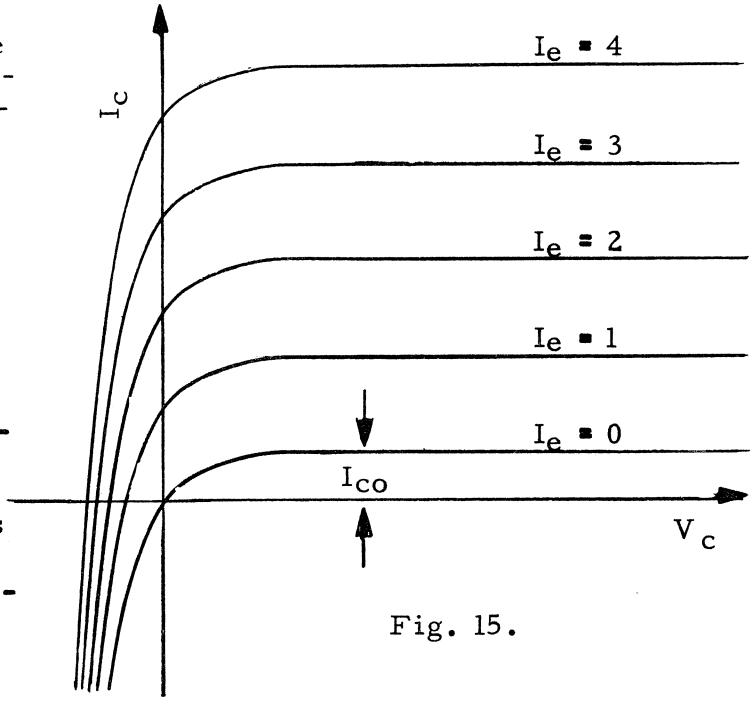


Fig. 15.

For positive collector voltage the collector current saturates at the voltage which makes the exponential term of the diode law (10) negligible, i.e., a few tens of millivolts. This, together with the fact that I_{co} is typically a microampere or less, and that α is constant at low currents, means that transistors can have good gains with only a few microamps collector current (say 10) and a few millivolts collector voltage (say 100), i.e., with only about one microwatt of collector dissipation.

The base current of a transistor is of course the difference between emitter and collector currents. Thus:

$$I_b = I_e - I_c$$

$$= (1 - \alpha) I_e - I_{co} \qquad (13a)$$

$$= \frac{(1 - \alpha) I_c - I_{co}}{\alpha} \qquad (13b)$$

Equations 13a and 13b may be inverted to give

$$I_e = \frac{I_b + I_{co}}{1 - \alpha} \qquad (14a)$$

$$I_c = \frac{\alpha I_b + I_{co}}{1 - \alpha} \qquad (14b)$$

If we neglect I_{co}, or hold it fixed and vary I_b by an incremental amount i_b, the changes in emitter and collector current are

$$i_e = \frac{i_b}{1-\alpha} = \frac{\beta}{\alpha} i_b \qquad (15a)$$

$$i_c = \frac{\alpha}{1-\alpha} i_b = \beta\, i_b \qquad (15b)$$

where

$$\beta = \frac{\alpha}{1-\alpha} .$$

For $\alpha = .98$, $\beta = 49$, in which case the emitter current will change 50 times as much as the base current, and the collector 49 times as much.

10. FREQUENCY LIMITATIONS OF JUNCTION TRANSISTORS.

Two factors limit the frequency response of junction transistors. The first of these is the collector capacitance. Since the collector-base junction is back biased, the collector impedance is high and for high gain, the collector should face a high load impedance. However, the capacitance of the collector to base is in shunt with this load, and limits the gain - bandwidth product just as does the plate capacity of a tube. The only difference is that the collector capacitance (due to space-charging of the junction) is non-linear, as we have seen.

The other factor is the so-called α cutoff of the transistor. As frequency is increased, α decreases from its value at low frequencies. To see why this is we must examine more carefully the diffusion through the base. For a current to flow due to diffusion there must be a gradient in the concentration

of carriers. Then more carriers tend to migrate from the more crowded region into the less crowded region than the other way around. The difference is the net current one way. What is more, current is proportional to the density gradient. Now at the collector side of the base, electrons are being swept out as fast as they arrive so the density there is low. As we move toward the emitter junction the electron density must increase with a slope proportional to collector current as shown in Fig. 16. Thus to change the collector current requires that the base be charged with a number of carriers proportional to the shaded area. At high frequencies most of the emitter current will be lost in this charging and discharging operation and the collector current will change very little, i.e., α will be low.

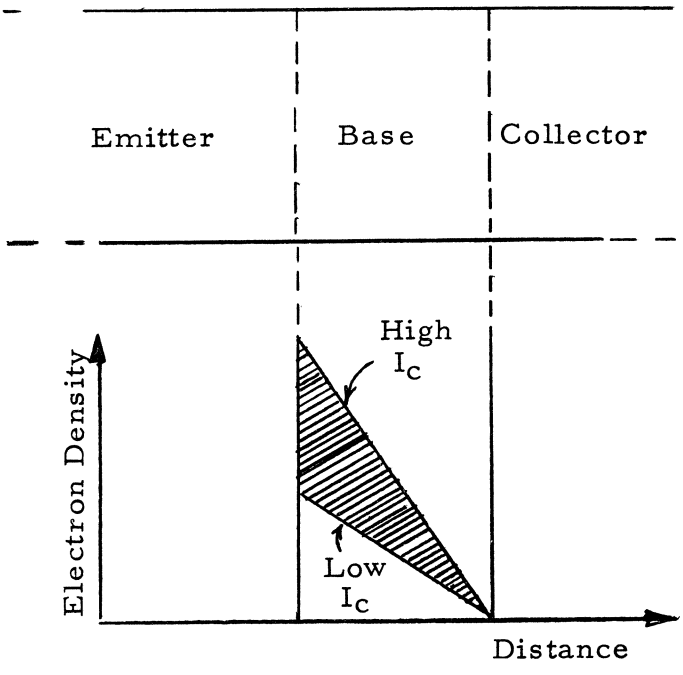


Fig. 16

To a first approximation, then, α behaves as if it had a simple RC high end cutoff. That is

$$\alpha = \alpha_o \frac{\omega_\alpha}{p + \omega_\alpha} \qquad (16)$$

where

α_o = d.c. value of α

ω_α = "Alpha cutoff" - the frequency at which α has decreased 3 db.

Typical α cutoff frequencies are now from 10 mc up to 250 mc or higher. A few years ago 2 mc was considered good. The improvement has come from thinner base layers. If the slopes in Fig. 16 are kept fixed the shaded area will increase as the square of the base layer thickness. Thus ω_o should vary inversely as the square of the base layer thickness and this is found to be the case.

The effect of α cutoff is essentially different from that of collector capacitance in that the latter limits only the gain bandwidth product and places no restrictions on the midband frequency, whereas the α cutoff sets an upper limit to the frequency at which gain may be had at a given bandwidth.

The frequency dependence of β is greater than that of α. Thus

$$\beta = \frac{\alpha}{1-\alpha} \qquad \frac{\alpha_o \frac{\omega_\alpha}{p+\omega_\alpha}}{1-\alpha_o \frac{\omega_\alpha}{p+\omega_\alpha}}$$

$$= \frac{\alpha_o \omega_\alpha}{p + (1-\alpha_o)\omega_\alpha} \qquad (17)$$

and we see that β has decreased 3 db when $\omega = \omega_\beta = (1-\alpha_o)\omega_\alpha$. Thus ω_β may typically be one tenth to one fiftieth of ω_α: values on the order of 100 kc are common.

11. TRANSISTOR EQUIVALENT CIRCUIT.

Probably the simplest accurate equivalent circuit for a transistor is that shown in Fig. 17. In this circuit, r_b, the "base resistance" is an ohmic resistance which arises from the fact that base current must flow through an extremely thin layer of a semiconductor, the actual resistance of which is therefore appreciable - typically on the order of 100 to 2000 ohms. The "emitter resistance", r_c, is the (forward) resistance of the emitter basic junction and is $\frac{26}{I_e}$ ohms. It is therefore nonlinear. The "collector resistance", r_e, is the (back) resistance of the base-collector junction and is typically several megohms. Under usual operating voltages the "collector junction capacitance", C_c, is on the order of 5 to 50 $\mu\mu$fds.

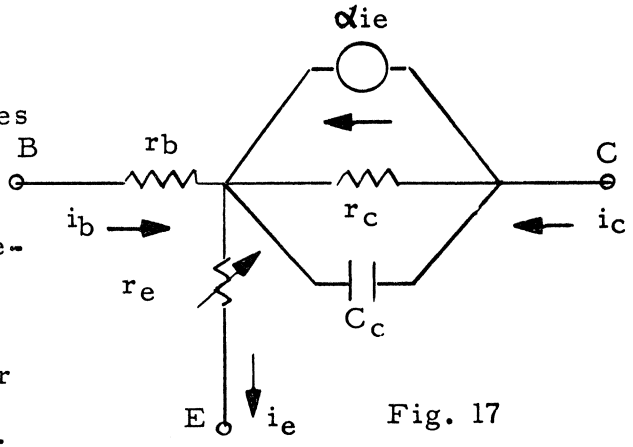


Fig. 17

The current generator around r_c can, of course, be replaced by a voltage generator $\alpha r_c i_e$ in series if desired. Using either equivalent circuit the properties of the different types of stages discussed in the next sections can be deduced.

12. THE GROUNDED (COMMON) EMITTER STAGE.

Just as is true with vacuum tubes, the highest power gain per stage is achieved if the emitter (cathode) is grounded. The input is then applied to the base (grid) and the output is taken from the collector (plate). Fig. 18a shows the configuration of a grounded emitter stage and Fig. 18b its vacuum tube equivalent.

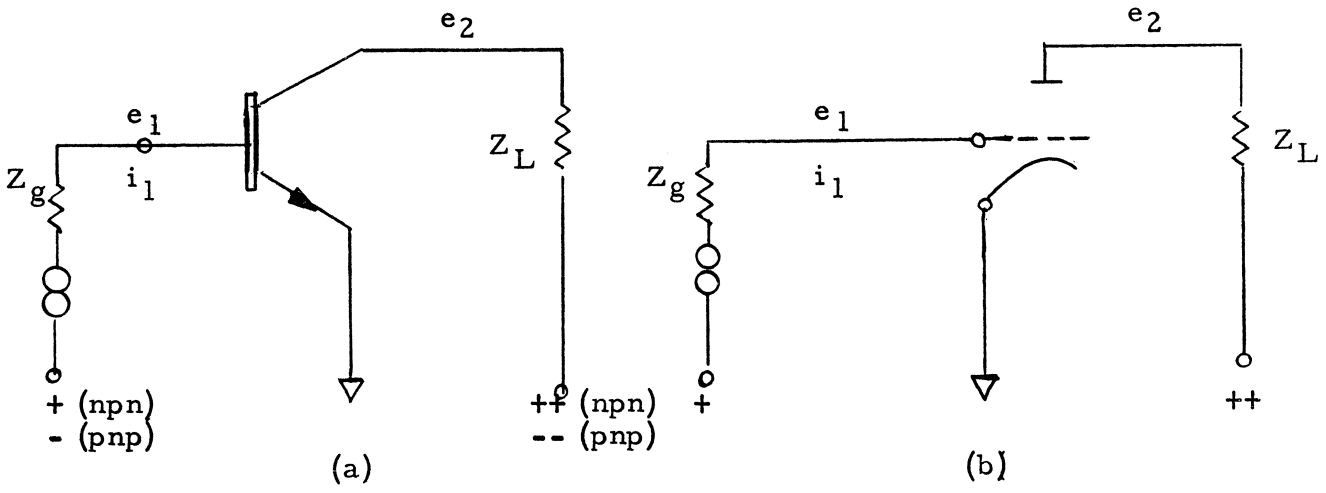


Fig. 18

The equivalent vacuum tube stage employs a high μ triode operated in the positive grid region. Such a stage requires driving power, has finite current gain, has a non-linear input impedance, and has reaction from output to input as does the transistor stage. Both circuits give a polarity reversal between input and output.

Using the equivalent circuit of Fig. 17, the characteristics of the grounded emitter stage can be determined. The approximate formulae, based on the assumptions:

$$r_e \ll r_c (1-\alpha)$$

$$r_b \ll r_c$$

$$r_e \ll Z_L \ll r_c (1-\alpha)$$

are as follows:

Input impedance $\quad \frac{e_1}{i_1} \approx r_b + \frac{r_e}{1-\alpha}$

Output impedance $\quad \frac{e_2}{i_1} \approx r_c (1-\alpha) + r_e \frac{\alpha r_c + Z_g}{r_e + r_b + Z_g}$

Voltage amplification $\quad \frac{e_2}{e_1} \approx - \frac{\alpha Z_L}{r_e + r_b (1-\alpha)}$

Current amplification $\quad \frac{i_2}{i_1} \approx -\beta = \frac{-\alpha}{1-\alpha}$

It will be noticed that the current gain of the grounded emitter stage is greater than unity - typically 20 to 50 times. If a stage has less than unity current gain, any power gain must result from the output impedance being large compared with the input impedance. In a multistage device, since the output of each stage must drive the input of the next stage any such gain will be lost unless impedance matching networks (such as transformers) are used. For video amplifiers this is impractical and each stage must therefore have current gain. For this reason grounded emitter stages are always used for video amplification. The grounded emitter connection also realizes the highest power gain per stage. On the other hand, the current gain varies as β and thus begins to fall off at a frequency ω_β.

13. THE GROUNDED (COMMON) BASE STAGE.

In the grounded base stage the input is applied to the emitter and the output is taken from the collector. The grounded base stage is analogous to the grounded grid amplifier as shown in Fig. 19. As in the grounded grid amplifier the current gain is essentially unity, the input impedance is low, the output impedance is high and multistage amplifiers must employ impedance matching devices as interstages. There is no polarity reversal in the grounded base stage.

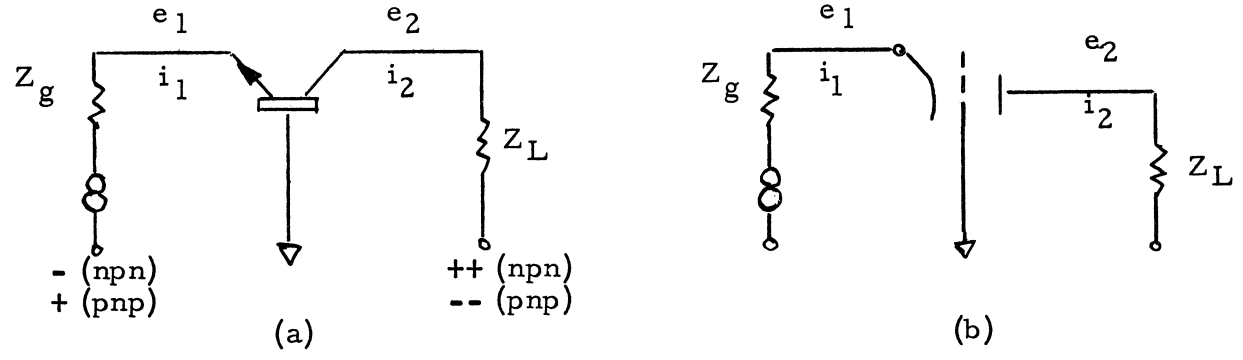


Fig. 19.

The appropriate formula for this type of stage, again based on the same assumptions, are as follows:

Input impedance $\quad \frac{e_1}{i_1} \approx r_e + r_b(1-\alpha)$

Output impedance $\quad \frac{e_2}{i_2} \approx r_c \frac{r_e + r_b(1-\alpha) + Z_g}{r_e + r_b + Z_g} \approx r_c$

Voltage amplification $\quad \frac{e_2}{e_1} \approx \frac{\alpha Z_L}{r_e + r_b(1-\alpha)}$

Current amplification $\quad \approx \alpha$

Compared with the grounded emitter stage we see that

(1) The input impedance is less by the factor $(1-\alpha)$

(2) The output impedance is greater by about the factor $\frac{1}{1-\alpha}$

(3) The voltage amplification is nearly the same

(4) The current amplification is less by the factor $(1-\alpha)$

From the last two statements we see that the power amplification is less by the factor $(1-\alpha)$.

While the power amplification of the grounded base stage is less than that of the grounded emitter stage, it has other advantages which often outweigh this fact. The grounded base stage provides better isolation between input and output circuits. Further the gain depends upon α, not β, so the upper frequency limit for the grounded base stage is $\omega\alpha$ rather than $\omega\beta$.

Thus grounded base circuits are commonly employed in r-f circuits where transformer coupling is easy and where high frequency operation is expected.

14. THE GROUNDED (COMMON) COLLECTOR CIRCUIT.

This stage is the analog of the "cathode-follower" stage in vacuum tube circuits, and in fact is often called the "emitter follower" stage. The configurations are shown in Fig. 20.

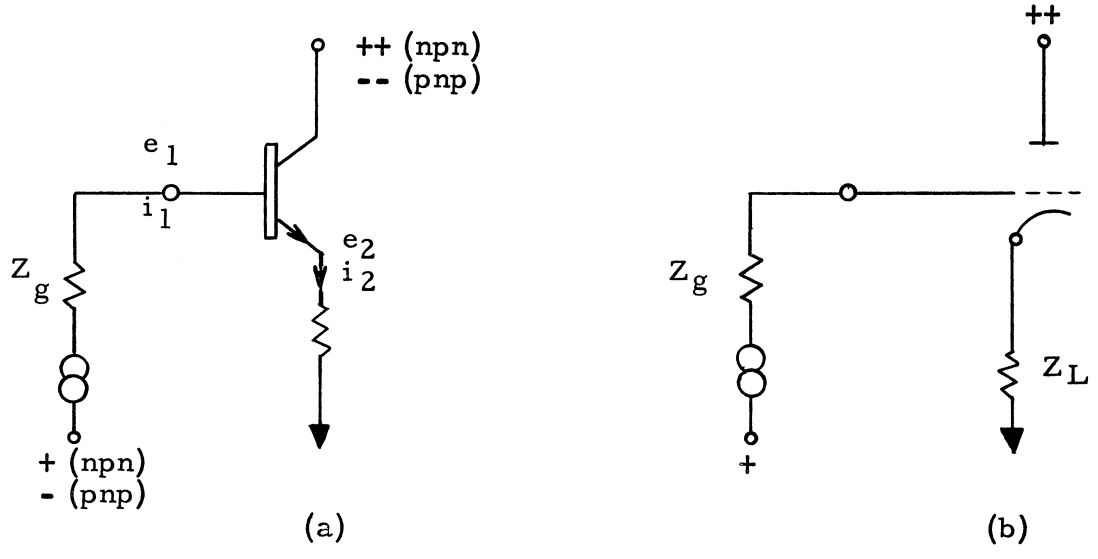


Fig. 20

Under the same assumptions as before, and in addition assuming $|Z_g| << r_c$, the approximate expressions are:

Input impedance	$\frac{e_1}{i_1}$	$\approx$	$r_b + \frac{r_e + Z_L}{1 - \alpha}$
Output impedance	$\frac{e_2}{i_2}$	$\approx$	$r_e + (r_b + Z_g)(1 - \alpha)$
Voltage amplification	$\frac{e_2}{e_1}$	$\approx$	1
Current amplification	$\frac{i_2}{i_1}$	$\approx$	$\frac{1}{1 - \alpha}$

Thus the grounded collector stage (like its cathode follower prototype) provides the highest input impedance and the lowest output impedance of all. It also has the least power gain, but this disadvantage is often offset by impedance considerations.