# PHYSICAL OPERATION OF THE P-N JUNCTION, DIODES AND TRANSISTORS©

**Jon Orloff 27 January, 2002**

## *I. Introduction*

The purpose of this text is to give you an idea of how a bipolar junction transistor (BJT) works. In order to do this we will first explain how a solid state diode works, since BJTs and diodes operate on the same basic principles. The operation of a field effect transistor (FET) is a great deal more complicated than that of a BJT, but to begin to understand its operation it is also necessary to understand the concepts we introduce here.

We will explain the BJT in terms of the materials it is made of and how electric currents flow through these materials. To do this we need to first understand: (1) the materials (mainly silicon (Si); (2) what the carriers of the electric current are; (3) what the equations are that govern the behavior of the carriers. These will lead us to an understanding of how current flows in a diode or BJT. We will find that the current-voltage relationship is exponential rather than the simple linear relationship of Ohm's Law, I = V/R, and this is responsible for the interesting and useful properties of diodes and transistors.

## *II. Materials*

We are used to thinking of electric currents flowing through metal wires or resistors (or in the case of AC currents, through capacitors). The equations describing this are not terribly complicated because for the most part the relationship between current and voltage is essentially linear. Diodes and transistors are not made of metal (except for the leads used to connect them to other circuit elements and for certain metal-semiconductor devices) but of Si or other similar materials. Si has very different electric properties than metals and is known as a semiconductor. Conduction is due not just to electrons, as in a metal, but also to another charged particle which is called a hole **(Holes) (Si conduction)**. A hole is the *absence* of an electron, but it can be characterized as a positively charged particle and calculations can be made based on that assumption that are extremely useful and reliable.

One of the things that makes Si such a useful material for electronics is that its conductivity can be changed by many orders of magnitude by changing its chemical composition slightly. This process is called doping and consists of adding a small quantity of an impurity to the Si in amounts ranging from about 1 part in $10^8$ to about 1 part in $10^5$. The impurities add electrons or holes to the Si. (Electrons are denoted by the symbol e and holes are denoted by the symbol h). The density of free electrons - electrons that can move through the Si crystal lattice as opposed to being attached to an atom - is called $n_o$ and the density of holes is called $p_o$ **(Holes).** The units are $cm^{-3}$. In pure Si $n_o = p_o$, because to create a free electron you remove an electron from the chemical bond between the atoms, and that automatically leaves a hole behind. At room temperature (about 293K) the density of holes $p_o$ and the density of electrons $n_o$ is equal: $n_o = p_o = n_i \approx 10^{10}$ $cm^{-3}$. The subscript "i" stands for *intrinsic*, the name given to pure Si. Si that has been doped with impurity atoms is called *extrinsic*. $n_i$ is a function of the temperature: $n_i = n_i(T)$ (the higher the temperature, the larger the number of pairs). The values of $n_o$ and $p_o$ can be varied over a wide range. **(Electron-hole pairs)**

## *III. Si "impurities": Donors and Acceptors*

Si has four valence electrons (See Figures 1 and 2 below). If an atom that has five valence electrons instead of four is put into the Si crystal lattice, it will bind to its four nearest neighbors and have one electron left over (see Figure 3a, below). In the case of the elements arsenic (As) and phosphorous (P), which both have five valence electrons, the fifth electron is extremely weakly bound to the atom when the atom is part of the crystal lattice. Near room temperature the probability for the $5^{th}$ electron to leave the atom is close to unity. In this case there will be a positively charged As or P ion fixed in the lattice (it doesn't move) and an "extra", donated, electron. Now, if one As or P donor atom is added per million Si atoms, there will be roughly an extra $5 \times 10^{16}$ electrons $cm^{-3}$, which is some $10^6$ more electrons than there would be if the Si were pure. If the density of donor atoms is $N_d$ $cm^{-3}$ the density of donated electrons will be almost exactly equal to $N_d$ also, since virtually all the donor atoms are ionized (for the dopants commonly used to make transistors (As, P) this is a very good assumption). If $N_d >> n_i$, then the number of

electrons will be $n_o = N_d$ almost exactly. As you can imagine, this "doping" of the Si with an As or P "impurity" will have a significant effect on the conductivity of the Si.

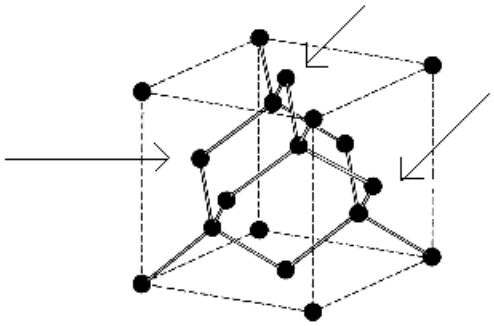


Figure 1. The diamond-lattice structure of Si. Note each Si atom (represented by a black sphere) is connected to its four nearest neighbors. The overall structure is a face-centered cube - note that each face of the cube has an atom centered in it (arrows point to some of these).

It is easier to see what is happening in the lattice if it is represented in a two dimensional form, as shown in Figure 2.

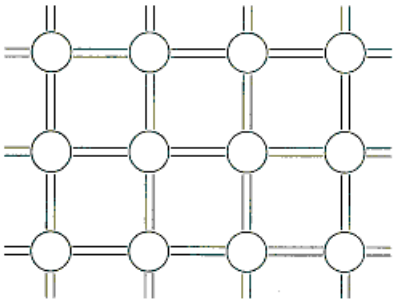
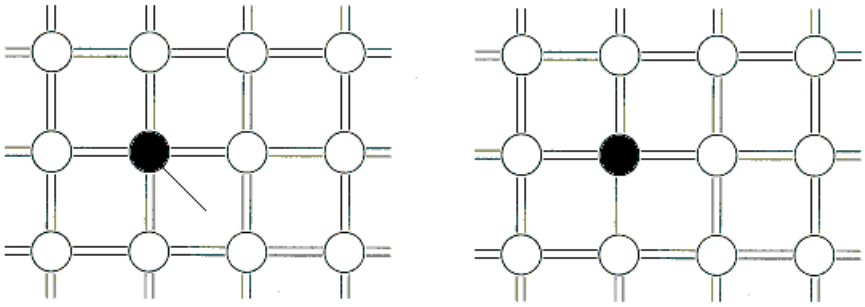


Figure 2. A schematic representation of the Si lattice. Each double line represents a valence bond between two Si atoms **(Single crystals)**.

If instead of As or P we were to add boron (B), which has only 3 valence electrons, then there will be only 3 full valence bonds established between the B atom and its nearest Si neighbors (see Figure 3b, below). In this case the absence of one electron is equivalent to the B bringing a hole into the crystal. There is a very high probability that the B will "accept" an electron and "donate" its hole to the Si lattice, creating a negatively charged B ion. The density of acceptor atoms is $N_a$ $cm^{-3}$ and, as with the doors, if $N_a >> n_i$, the density of holes will be $p_o = N_a$ almost exactly, since essentially all the donor atoms are ionized. As with a donor such as P, a B acceptor will greatly affect the Si conductivity even when the B is added in proportions as small as one part per million.



a b

Figure 3. a) Si lattice with an As atom (dark circle) replacing one Si atom. The diagonal line represents the fifth valence electron that is very weakly bound: essentially all As atoms are positive ions at room temperature. b) Si lattice with a B atom (dark circle) replacing one Si atom. Note that there are only three B valence electrons here. B easily accepts a free electron from elsewhere in the lattice and becomes negatively ionized. Essentially all B atoms are ionized at room temperature.

If both donors and acceptors are put into the Si, the net concentration, or density of impurity atoms, is

$$N_D = N_d - N_a \quad \text{or} \quad N_A = N_a - N_d \tag{1}$$

($N_D$ and $N_A$ are positive quantities). If we put both donors and acceptors into the Si, electrical neutrality of the silicon requires

$$n_o + N_a = p_o + N_d \tag{2}$$

since all the donor and acceptor atoms are ionized. This last equation tells us that the number of electrons plus the number of negatively ionized acceptors must equal the number of holes plus the number of positively ionized acceptors. Another way of writing this is

$$n_o - p_o = N_d - N_{a.} = N_D \quad \text{or} \quad p_o - n_o = N_a - N_d = N_A \tag{2'}$$

Since the conductivity of Si depends on the number of electrical carriers, we must find a way of calculating the number of carriers when donors or acceptors are added. That is, if we start with pure Si and add As atoms in the amount $N_d$ per $cm^3$, then $n_o$ will almost exactly equal to $N_d$ since, in general, $N_d$ will be much larger than the number of free electrons produced due to the breaking of Si bonds, $n_o = n_i = 1.1 \times 10^{10}$ $cm^{-3}$ at room temperature. But, what will be the value of $p_o$? This is not a simple question. $N_a$ and $N_d$ are under our control - we can control the doping levels. But we have only one equation for the two unknowns $n_o$ and $p_o$. We need a second equation in order to find both $n_o$ and $p_o$.

If Si is pure we know that the number of holes and electrons per $cm^3$ will be the same, for every time an electron breaks loose from a valence bond it leaves behind a hole. How does this happen? To break a bond requires about 1.1 eV (eV = electron-volt. 1 eV = 1.6 x 10-19 joules) of energy. This energy can be supplied in many ways: heat, light (photons), radio waves (rf), acoustical waves (phonons) etc. Thus there are many generation mechanisms for electron-hole

pairs. In addition, there may be "cross-talk": an incoming light photon might not have sufficient energy to break a bond, but if thermal jostling adds some energy at the same time, then the bond can be broken. Thus the generation mechanisms will depend on temperature. If the rate of generation of electron-hole pairs per $cm^3$ at temperature T is called g(T), then we can write

$$g(T) = g_{thermal}(T) + g_{light}(T) + g_{rf}(T) + \Sigma_i\, g_i(T) \qquad (3)$$

where the sum over i represents all the processes that can produce electron-hole pairs. The units of g are $sec^{-1}\ cm^{-3}$ (rate per $cm^3$).

Now, if the Si crystal is in equilibrium, the number of electron-hole pairs remains constant **(Equilibrium).** This implies the rate of recombination is equal to the rate of generation. If the rate of recombination is called r, then r = g. Since g is a function of temperature, r must be also be a function of temperature, since equilibrium can be established at any temperature below the melting point, and so we have r(T) = g(T). Recombination means a hole and an electron meet so the electron drops back into a bond: the free electron and the free hole disappear. It is certainly plausible that recombination can be influenced by the different means of delivering energy to the crystal, so, as we did for generation we write

$$r(T) = r_{thermal}(T) + r_{light}(T) + r_{rf}(T) + \Sigma_i\, r_i(T) \qquad (4)$$

The question of importance, is, as we shall soon see, does r(T) = g(T) for every process? For the crystal as a whole we must have, in equilibrium, that the rate of generation of electron-hole pairs is equal to the rate of recombination of electron-hole pairs (equilibrium means, at least, no net current flow). It turns out that the rates of generation and recombination must be equal for each process - this is called the Principle of Detailed Balance. The reason is that if there were not a detailed balance the Si crystal could absorb energy from the outside, a process that creates electron-hole pairs. If this was not balanced by a recombination of electron hole pairs, a process

that gives up energy the crystal would heat up indefinitely, which would be a violation of the Second Law of thermodynamics.

If the generation rate is not too large it will be independent of $p_o$ and $n_o$, because the number of atoms that can provide electron-hole pairs will be vastly larger than the number of electron-hole pairs and g = g(T). On the other hand, the recombination rate depends on $n_o$ and $p_o$, for if recombination is to take place there must be *at least* one hole or one electron available. Therefore, if the crystal is in thermal equilibrium with an electron density $n_o$ and a hole density $p_o$, the recombination rate *must* depend on $n_o$ and $p_o$ as well as T: $r = r(T,n_o,p_o)$. To find the dependence we use essentially the method given in "Microelectronic Devices and Circuits", C. Fonstad, McGraw Hill (1994). Assume that $n_o$ and $p_o$ are always small compared to the density of atoms N, and expand $r(T,n_o,p_o)$ in a Taylor's series in $n_o$ and $p_o$:

$$r(T, n_o, p_o) = C1 + C2\frac{n_o}{N} + C3\frac{p_o}{N} + C4\frac{n_o^2}{N^2} + C5\frac{p_o^2}{N^2} + C6\frac{n_o p_o}{N^2} + C7\frac{n_o^3}{N^3} + C8\frac{p_o^3}{N^3} + C9\frac{n_o^2 p_o}{N^3} + C10\frac{n_o p_o^2}{N^3} + \ldots. \quad (5)$$

Now, since there must always be at least one $n_o$ and one $p_o$ in each term of Equation 5 for recombination to make any sense, we see that C1, C2, C3, C4, C5, C7 and C8 must vanish, and so the first non-vanishing term is $C6\frac{n_o p_o}{N^2}$. The next non-vanishing term is $C9\frac{n_o^2 p_o}{N^3}$. Since $N \sim 10^{22}$ and $n_o$ or $p_o$ will never be much larger than $10^{18}$ in any realistic semiconductor device, we see that the only significant term is the one in which $n_o p_o$ appears to the first power; this term will be at least $10^6$ times larger than the next term. Therefore we find that to an excellent approximation

$$r_i(T,n_o,p_o) = C(T)\ n_o p_o \quad (6)$$

where all the constants have been absorbed into the function of temperature C(T).

Consider the case of intrinsic Si where the breaking of bonds leads to $n_o$ electrons and $p_o$ holes. The process of electron hole creation/recombination can be analyzed as if it were a chemical reaction using the Law of Mass Action. The Law of Mass Action is derived from thermodynamics and relates the number of constituents and products of chemical reactions, such as $2H_2O \leftrightarrow 2H_2 + O_2$ . The Law of Mass Action shows that for the situation of [bonds] $\leftrightarrow$ [electron-hole pairs], the process depends only on temperature. This is very important and so a detailed argument to justify it may be found at the following link **(Mass action).**

Suppose that there is a chemical reaction in which $\nu_i$ molecules of reactants combine (or disassociate) to produce a product. The number of reactants will change as molecules are used up or created. The concentration of any chemical is $n_i/n$, where $n_i$ is the number of moles of material i and $n = \sum_i n_i$. The Law of Mass Action says that

$$\prod_i \left( \frac{n_i}{n} \right)^{\nu_i} = f(T) \qquad (7)$$

the product of the concentrations $\frac{n_i}{n}$ raised to the power $\nu_i$, is dependent only on the

temperature if the pressure is constant (which is certainly the case in a Si crystal, for example). Consider the case of water (steam) disassociating to form hydrogen and oxygen, $2H_2O \rightarrow 2H_2 + O_2$. If the reaction takes place in a chamber under constant pressure, $n_i/n$ is the concentration of each product ($H_2$ and $O_2$) or reactant ($H_2O$) and the $\nu_i$ are +2 for hydrogen, +1 for oxygen and -2 for water, respectively. The negative signs indicate that the reactants (the $H_2O$ molecules) are used up to form the products (**Mass action**). Now, for the case of Si where a Si bond breaks and forms an electron-hole pair, [bond] $\rightarrow$ [ electron-hole pair] the $\nu_i$ = -1 for [Si bond] and +1 for the electrons and for the holes. The concentration of electrons and holes is $\frac{n_o}{N}$ and $\frac{p_o}{N}$, where N is the number density of electrons plus holes plus Si bonds. respectively. The Law of Mass Action

says the product of $\left(\frac{n_o}{N}\right)^1$ by $\left(\frac{p_o}{N}\right)^1$ by $\left(\frac{\text{number of bonds}}{N}\right)^{-1}$ is a function of temperature only,

or $n_o p_o = f(T) =$ constant when T = constant. Therefore, we have found that the product $n_o p_o$ is a constant at constant temperature. Since the generation rate must equal the recombination rate, $g = r = C(T)\, n_o p_o$ and since $n_o p_o$ is a constant at constant temperature, we have found a second equation relating $n_o$ and $p_o$. Now, if we have some doped Si, then the product $n_o p_o$ is constant independent of the doping level (within the approximation made following Equation 5), so if we know $n_o p_o$ for any case at all, we know it for all cases. In fact there is a very simple case where $n_o p_o$ is known: if the Si is undoped $n_o = p_o = n_i$. Therefore we find that the number we are looking for is $n_o p_o = n_i^2 \approx 1.21 \times 10^{20}$ $cm^{-3}$ (at room temperature). This enables us to find $n_o$ and $p_o$ when the Si is doped: when the net doping is donor- like, i.e.,
$N_d > N_a$, $n_o - p_o = N_D$ and with $n_o p_o = n_i^2$ we have

$$n_o = \frac{N_D}{2}\left(1 + \sqrt{1 + \frac{4n_i^2}{N_D^2}}\right) \tag{8}$$

When the net doping is acceptor-like, i.e. $N_a > N_d$, $p_o - n_o = N_A$ and

$$p_o = \frac{N_A}{2}\left(1 + \sqrt{1 + \frac{4n_i^2}{N_A^2}}\right) \tag{9}$$

Often Si is doped only with either donors or acceptors. In these cases $N_D$ can be replaced by $N_d$ in Equation (8) or $N_A$ by $N_a$ in Equation (9). As an example, suppose Si is doped with acceptors (B) at a density $N_a = 5 \times 10^{17}$ $cm^{-3}$. Then $p_o = 5 \times 10^{17}$ $cm^{-3}$ ( $\frac{4n_i^2}{N_a} \approx \frac{4 \times 10^{20}}{25 \times 10^{34}} = 1.6 \times 10^{-15}$ is negligible). The corresponding value of $n_o$ is very small: $n_o = \frac{n_i^2}{p_o} \approx \frac{10^{20}}{5 \times 10^{17}} = 200$. The

larger of $n_o$ and $p_o$ is called the majority carrier, while the other is called the minority carrier. When doped with a donor impurity Si is called n-type and electrons are the majority carriers. When doped with an acceptor impurity Si is called p-type and holes are the majority carriers.

## *IV. The motion of Electrons and Holes in Si*

In order to find the relation between voltage and current in Si it is necessary to understand how the electrons and holes behave when a voltage is placed across the Si crystal, resulting in an electric field in the crystal, and also how they behave if the spatial distribution of electrons and holes isn't uniform. This is actually a very interesting and important situation: electric current flows not only because of an impressed electric field but also by the diffusion of charge carrying particles from regions of high density to regions of low density. Diffusion is critically important for the operation of diodes and transistors.

First we consider what happens when a Si crystal with a uniform distribution of electrons and holes has a voltage placed on it and then consider the problem of motion when the distribution is not uniform.

### *IV A. Uniform Distribution.*

When a charged particle with mass m is placed in an electric field $\mathbf{E}$ it feels a force $q\mathbf{E}$ and undergoes an acceleration a = qE/m. An electron in free space in a field $E = 1$ VM$^{-1}$, considered as a classical particle with mass m = 9 x 10$^{-31}$ kg, would accelerate at the rate of $1.6 \times 10^{-19} \div 9 \times 10^{-31} = 1.78 \times 10^{11}$ M sec$^{-2}$. It would attain a speed of 1500 M sec$^{-1}$ in about 8 nsec and in this time it would travel a distance of about 6 micrometers. However, in a solid such as a Si crystal an electron (or a hole, for that matter) is constantly colliding with the atoms in the crystal and changing velocity. Experiments show that the speed attained by an electron or a hole in Si is proportional to the electric field according to $\mathbf{v}_e = \mu_e E$ and $\mathbf{v}_h = \mu_h E$, respectively, at least until a fairly high velocity is reached where the rate of collisions limits the velocities. $v_e$ and $v_h$ are called *drift* velocities of the particles. The constants $\mu_e$ **and** $\mu_h$ are called the *mobilities* of the electron and the hole, respectively **(Mobility)**. Their values in Si are

$\mu_e$ = -1500 $cm^2\ V^{-1}\ sec^{-1}$ and $\mu_h$ = 600 $cm^2\ V^{-1}\ sec^{-1}$. The negative value for $\mu_e$ reflects the fact that because of its negative charge the electron moves in a direction opposite to the direction of the *E* field **(Drift velocity).**

If the average velocity v = $\mu E$ of charged particles is multiplied by the charge density of the particles (which has units C $cm^{-3}$) the result is the current density J (units A $cm^{-2}$) **(Current density)**.

Current density is a more useful quantity than current for describing the properties of semiconductors. Its meaning is as follows. The usual way Ohm's Law is written is I = V/R (the unit of resistance is the ohm, whose symbol is $\Omega$). This is all right for wires and resistors, but when dealing with Si there is a problem in that the size and shape of the crystal may vary. Therefore it is more convenient to write Ohm's Law in a form that is independent of the shape and size of the crystal, and we use the concept of resistivity (also symbolized by $\rho$ - the meaning of $\rho$ depends on the context) which is defined as the resistance of a piece of material multiplied by its cross sectional area and divided by its length: the units are $\rho \sim \Omega$-cm. The utility is easily seen by calculating the resistance of a piece of Si with $\rho = 1\ \Omega$-cm that is L = 100 micrometers long, W = 5 micrometers wide and T = 2 micrometers thick. Since 1 micrometer = $10^{-4}$ cm, we have $R = \frac{\rho \times L}{W \times T} = \frac{1 \times 100 \times 10^{-4}}{5 \times 10^{-4} \times 2 \times 10^{-4}} = 1 \times 10^5\ \Omega$. Since resistivity has units of $\Omega$-cm its inverse has units of ($\Omega$-cm)$^{-1}$, which is called the conductivity $\sigma$. Writing Ohm's Law as I = GV where G is the conductance, we easily see that $G = \frac{\sigma \times W \times T}{L}$. Then

$I = \frac{\sigma \times W \times T}{L} \times V$ or, dividing by the area W x T to get the current density J, we have

$J = \sigma \times \frac{V}{L}$ or J = $\sigma E$, where the electric field is the voltage across the crystal divided by its length. Since we have seen that the current density is equal to the charge density multiplied by the average velocity of the charges, we find the current density for holes and electrons to be

$J_h^{drift} = qp_o \mu_h E$ and $J_e^{drift} = qn_o \mu_e E$. Note that both current densities have the same sign; that is because the negative sign of the electron charge cancels the negative sense of the mobility $\mu_e$. The superscript "drift" signifies this current density is due to the motion of the charges in an electric field **(Resistivity of Si)**.

*IV-B. Currents due to non-uniformity in carrier densities*: *diffusion currents*

If there is a non-uniform distribution of charges in the Si crystal, then the charges will tend to redistribute themselves through *diffusion* (**Diffusion)**.

The flux of holes or electrons away from a region of higher than average density will be given by

$$\begin{aligned} F_h &= -D_h \frac{\partial p_o}{\partial x} \\ F_e &= -D_e \frac{\partial n_o}{\partial x} \end{aligned} \tag{10}$$

respectively. The corresponding current densities are found by multiplying the fluxes by ±q:

$$\begin{aligned} J_e^{diff} &= qD_e \frac{\partial n_o}{\partial x} \\ J_h^{diff} &= -qD_h \frac{\partial p_o}{\partial x} \end{aligned} \tag{11}$$

where the superscript "diff" indicates the current density is due to the diffusion of particles (partial derivatives are used because $p_o$ and $n_o$ may be functions of time as well as position). Note that unlike drift current, diffusion currents have opposite sign for holes and electrons because the driving force - the gradient of the density - is independent of the sign of the particles.

*IV-C. The Total Current Densities and Gauss' Law*

The total current densities are given by the sum of the drift and diffusion current densities:

$$J_e^{tot} = J_e^{drift} + J_e^{diff} \quad , \quad J_h^{tot} = J_h^{drift} + J_h^{diff} \tag{12}$$

These expressions contain the carrier densities, their derivatives and the electric field, which are all functions of position and possibly time.

Two other relationships are be needed, to find the current densities, Gauss' Law and the equation of continuity. Gauss' Law links the electric field to the charge density $\rho(x) = \rho$ and, in one dimension, it is given by

$$\epsilon \frac{\partial E}{\partial x} = \rho \tag{13}$$

where $\epsilon$ is the dielectric constant, in this case the dielectric constant of Si ($\epsilon_{Si}$ - we are assuming the dielectric constant $\epsilon_{Si}$ for Si is constant when extra carriers are introduced; $\epsilon_{Si} = 11.8\ \epsilon_o$ where $\epsilon_o = 8.85 \times 10^{-12}$ F/M, or farads per meter is the dielectric constant of vacuum). The other physical law that applies to diffusing particles that we must take into account is the equation of continuity. The equation of continuity tells us that matter is conserved: the net flux of particles out of a volume in space must equal the change in time of the density in that region. This is expressed in one dimension by

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial J(x,t)}{\partial x} = 0 \tag{14}$$

## *V. Recapitulation*

The point of this is to understand how a BJT works, by which we mean understanding the current-voltage relationships of the transistor. We begin by studying a p-n junction, often called

a diode, which is an abrupt juntion between p and n doped Si. To understand how the diode works we need to understand how the current is carried in the Si material.

So far we have discussed the types of carriers (electrons and holes) and some of their properties, and the two kinds of currents (drift and diffusion). We next have to learn about what happens when (1) Si is doped either p or n and (2) electrons or holes, are "injected" into these materials (a p-n junction is made from two kinds of Si (p-type and n-type) in intimate contact and for current to flow in a p-n junction it first has to be gotten into the junction; this process is called injection). After we study this we will consider what happens when n-type and p-type Si are brought into contact to form what is called the p-n junction: it is of very great importance in modern electronics and is the heart of the diode and the transistor.

## *VI. The Lifetime of Minority Carriers*

It is important to understand the concept of carrier lifetimes in order to understand how the p-n junction works. The basic idea of lifetime can be understood as follows. Suppose Si is n-type and is doped with donors so $N_d = 10^{18}$. Since essentially all the donors are ionized the density of electrons will be $n_o = N_d = 10^{18}$. The density of the positive carriers, the holes, is given by $p_o = \frac{n_i^2}{n_o} = \frac{(1.08 \times 10^{10})^2}{10^{18}} \approx 100\,\text{cm}^{-3}$ (at room temperature). This small number is the equilibrium value of the hole density. If more holes are generated, for example by shining light on the Si to create additional electron-hole pairs, or by injecting holes across a boundary from p-type Si as happens when current flows through a p-n junction, then these new holes will rapidly recombine with the large number of electrons already present. If light is shone on the Si then additional electron-hole pairs are created and if the creation, or generation rate is small (small is defined below, see page 16) then the additional electrons generated make little difference, but the excess holes generated may easily outnumber the equilibrium value of the hole density, possibly by a very large amount. When the light is turned off the excess holes disappear as they recombine with their more populous neighbors, the electrons. This process causes the hole number density to decay with time while the number density of the electrons is hardly affected since it is so much

larger. The decay time is of great importance because it can be related to how far the holes move before their number decays from its initial value to a smaller value.

In equilibrium the generation rate $g_o$ of electron-hole pairs equals the recombination rate $r_o$ , where the subscript o indicates the equilibrium rate. As we saw earlier, $r_o = C(T)\, n_o p_o$. If additional electron-hole pairs are created then the generation rate will exceed the recombination rate.

**(Note on notation**: if we call the *total* number of electrons and holes n and p, respectively, then $n > n_o$ and $p > p_o$ when extra electron-hole pairs are being generated. The difference between n and $n_o$ or p and $p_o$ is the number of *excess* particles, indicated by a prime: $n' = n - n_o$ ; $p' = p - p_o$ ).

In equilibrium, obviously $\frac{\partial n(x,t)}{\partial t} = \frac{\partial p(x,t)}{\partial t} = 0$ and $n = n_o$, $p = p_o$. When electron-hole pairs are being generated then

$$\frac{\partial n(x,t)}{\partial t} = \frac{\partial p(x,t)}{\partial t} = g(x,t) - r(t) \qquad (15)$$

Since the equilibrium rate of generation is $g_o$, $g = g_o + g$, where $g$ may be a function of time and position (if it is due to light, the light intensity may change with time, etc., so $g = g(x, t)$). The recombination rate depends on n and p, so $r = C(T)\, n(t)p(t)$. Since $g_o = r_o$ in equilibrium and since $r_o = C(T) n_o p_o$, $g(t) = C(T)\, n_o p_o + g(x,t)$. Then the rate of change of the density of carriers is given by

$$\begin{aligned} \frac{\partial n(x,t)}{\partial t} = \frac{\partial p(x,t)}{\partial t} &= g(x,t) + C(T)\, n_o p_o - C(T)\, n(t)p(t) \\ &= g(x,t) - C(T)\left( n(t)p(t) - n_o p_o \right) \end{aligned} \qquad (16)$$

This equation can be simplified by defining $n'(x,t) = n(x,t) - n_o$ and $p'(x,t) = p(x,t) - p_o$ . As defined above, n' and p' are functions of position and time; recall that they are the number densities of the *excess* electrons and holes created by the generation process. Since an electron-hole pair creates a free hole for every free electron, n' = p'. Since $\frac{\partial n_o}{\partial t} = \frac{\partial p_o}{\partial t} = 0$ Equation

(15) becomes

$$\frac{\partial n'(x,t)}{\partial t} = \frac{\partial p'(x,t)}{\partial t} = g(x,t) - C(T)\left( (n_o + n')(p_o + p') - n_o p_o \right) \qquad (17)$$

Multiplying out the terms in the square bracket we end up with

$$\frac{\partial n'(x,t)}{\partial t} = \frac{\partial p'(x,t)}{\partial t} = g(x,t) - C(T)\left( (n_o + p_o)n' + n'^2 \right) \qquad (17')$$

A problem is immediately evident - for the general case we have a non-linear partial differential equation. However, the equation can be linearized if $n' << (n_o + p_o)$, for then we have $\frac{\partial n'(x,t)}{\partial t} = \frac{\partial p'(x,t)}{\partial t} \approx g(x,t) - C(T)\,(n_o + p_o)\,n'$. Suppose that $g(x,t)$ is non-zero for time $t < 0$ and that $g(x,t) = 0$ for $t \geq 0$. Then for $t \geq 0$ we have

$$\frac{\partial n'(x,t)}{\partial t} \approx -C(T)\,(n_o + p_o)\,n'(x,t) \qquad (18)$$

which is easily solved: let $n'(x,t) = N(x)\,\nu(t)$. Equation 18 becomes

$$\begin{aligned} \frac{\partial N(x)\,\nu(t)}{\partial t} &\approx -C(T)\left(n_o + p_o\right) N(x)\,\nu(t) \rightarrow \\ N(x)\,\frac{d\nu(t)}{dt} &\approx -C(T)\,(n_o + p_o)\,N(x)\,\nu(t) \end{aligned} \qquad (18a)$$

which becomes

$$\frac{d\nu(t)}{dt} \approx -C(T)\,(n_o + p_o)\,\nu(t)$$

The quantity $C(T)\,(n_o + p_o)$ has units of $sec^{-1}$ so we replace it by $\frac{1}{\tau}$ and also replace $\nu(t)$ by $n'(t)$ to obtain

$$\frac{dn'}{n'} = -\frac{dt}{\tau} \quad \text{and} \quad n'(t) = n'(0)\, e^{-t/\tau_o} \tag{19}$$

This tells us the number density of the excess carriers falls exponentially with time, with a time constant $\tau$ (which is typically of the order of a fraction of a micro-second). This gives us the meaning of a "small" rate of generation of electron-hole pairs referred to above - the excess number density must be $<< n_o + p_o$ **(Non-linear equation)**.

In general we can evaluate the behavior of the p-n junction quite satisfactorily by using the assumption of a small generation rate of electron-hole pairs. Since what we will be interested in is primarily the flow of electrons and holes across the boundary of the junction, a process called injection, we will be looking at the case of low-level injection (the exact meaning of this will be defined below).

## *VII. The Equations for the Motion of Electrons and Holes*

We can now give a complete description of the behavior of the electrons and holes in doped Si in one dimension. We have the following numbered equations , where p = p(x,t), n = n(x,t), J = J(x,t) and *E* = *E*(x,t): first, the two equations giving the total electron and hole current densities in terms of the drift and diffusion currents,

I) $$J_e^{tot}(x,t) = q\, n\, \mu_e\, E + q\, D_e \frac{\partial n}{\partial x}$$

II) $$J_p^{tot}(x,t) = q\, p\, \mu_h\, E - q\, D_h \frac{\partial p}{\partial x}$$

Then we have Gauss' Law and the Continuity Equations,

III) $$\epsilon \frac{\partial E}{\partial x} = q\,[p + N_d(x) - n - N_a(x)]$$

IV) $$\frac{\partial J_e(x,t)}{\partial x} = q \frac{\partial n}{\partial t}$$

V) $\frac{\partial J_h(x,t)}{\partial x} = -q \frac{\partial p}{\partial t}$

(The charge for electrons is -q, remember). Finally, we have the two equations giving us the rate of change of n' and p' due to creation of (excess) electron-hole pairs, VI ) and VII)

$$\frac{\partial n'}{\partial t} = \frac{\partial p'}{\partial t} = g(x,t) - C(T)\,[(n_o + p_o)n' + n'^2] \quad .$$

(The equations for $\frac{\partial n'}{\partial t}$ and $\frac{\partial p'}{\partial t}$ are the same). Note that the generating function *g* has been written as a function of position and time, $g = g(x,t)$, to be general. These last two equations for the time variation of *n* and *p* can be combined with the continuity equations to give

IV') $\frac{\partial n'}{\partial t} = g(x,t) - C(T)\,[(n_o + p_o)n' + n'^2] + \frac{1}{q}\frac{\partial J_e}{\partial x}$ and

V') $\frac{\partial p'}{\partial t} = g(x,t) - C(T)\,[(n_o + p_o)n' + n'^2] - \frac{1}{q}\frac{\partial J_h}{\partial x}$

Here, whenever a derivative with respect to time is taken we can replace p by p' and n by n'. These five coupled, non-linear, partial differential equations cannot be solved in general. But, for the conditions that exist in practical bi-polar transistors and in diodes, simplifying assumptions can be applied that result in equations that can be solved and that give useful information.

## *VIII. The behavior of Electrons and Holes in Typical p-n Junctions and BJTs.*

By "typical" we mean the sort of semiconducting materials and devices you would ordinarily run into when building circuits: signal diodes and BJTs. For these devices the equations I - V can be greatly simplified. The simplifications come from the way the devices are actually made: the p and n regions are uniformly doped and there are sharp boundaries between the n and p doped regions. In addition, and very importantly, under ordinary operation the number density of excess holes and electrons is small compared to the density of the majority carriers, which can be expressed as $\{n', p'\} << n_o + p_o$.

We will now find ways of simplifying our equations, but it is useful to put things in perspective at this point. We find that in working devices the minority drift currents can be ignored. This simply means that, for example, in p-type Si where $p_o >> n_o$, $J_h^{drift} >> J_e^{drift}$ (essentially by the factor $p_o/n_o$). It is also found that minority diffusion currents are much larger than minority drift currents, and cannot be ignored. In fact the minority diffusion currents are extremely important for calculating the current flowing through a p-n junction.

In these p-n devices we find that the doping is essentially uniform throughout the p and n regions of Si (obviously the doping isn't uniform everywhere or there would be no p-n junction). This means that $\frac{\partial p_o}{\partial x} = \frac{\partial n_o}{\partial x} = 0$. Since $n = n_o + n'$ and $p = p_o + p'$ it follows that

$\frac{\partial n'}{\partial x} = \frac{\partial n}{\partial x}$ and $\frac{\partial p'}{\partial x} = \frac{\partial p}{\partial x}$. In addition, the doping is time independent so that $\frac{\partial n'}{\partial t} = \frac{\partial n}{\partial t}$ and $\frac{\partial p'}{\partial t} = \frac{\partial p}{\partial t}$.

In Gauss' Law, $\epsilon \frac{\partial E}{\partial x} = q\,[p + N_d(x) - n - N_a(x)]$, where the net charge density $\rho$ is written in terms of the doping concentrations $N_d$ and $N_a$ because the sum of the holes and the positive ions created when donor atoms give up their electrons to the crystal must be equal to the number of electrons plus the number of negatively ionized acceptors created when the acceptors give up their holes: $p_o + N_d = n_o + N_a$. If the doping was not uniform $N_a$ and $N_d$ would be functions of position. Since $n_o = N_d$ and $p_o = N_a$ to a high degree of accuracy, we can simplify Gauss' Law $\epsilon \frac{\partial E(x,t)}{\partial x} = q\,[p' - n'] = \rho_{excess}$ where the net charge density is now only a function of the excess charges.

Another approximation that greatly simplifies the equations has to do with the time variation of the charge density. It turns out that charges redistribute themselves very quickly in a semiconductor, and if you look at any volume greater than a few thousand nanometers cubed (~

10 nm radius), on the average it is electrically neutral. This allows us to ignore the time derivatives in the equations above **(Relaxation time)**. If there is a region in the Si where there is an imbalance of excess charge, the resulting charge density will create an electric field that will tend to move the charges to reduce the charge imbalance - to smooth out the charge density distribution. We already have seen how the current density is related to the electric field and the charge density variation in equation number I) above (for electrons, assumed to be the majority carriers): $J_e^{tot} = \sigma E + q D_e \frac{\partial n'}{\partial x}$, where we set $\sigma = q n_o \mu_e$. We can use equation IV') from above, $\frac{\partial n'}{\partial t} = g(x,t) - C(T)\,[(n_o + p_o)n' + n'^2] + \frac{1}{q}\frac{\partial J_e}{\partial x}$, to find the time variation of n. To do this we take advantage of the fact that $n' << (n_o + p_o)$ to linearize the equation by dropping the term in $n'^2$. We take the derivative of J and divide by q to get

$$\frac{\sigma}{\epsilon}[p' - n'] + D_e \frac{\partial^2 n'}{\partial x^2} = \frac{\partial n'}{\partial t} - g(x,t) + C(T)(n_o + p_o)\, n' \qquad (20a)$$

where $n = \frac{\rho}{q}$ and $\frac{dE}{dx} = \frac{(p' - n')}{\epsilon}$. The term $C(T)(n_o+p_o)$ in the equation multiplying n' has units of $sec^{-1}$ and from before we call it $\tau^{-1}$; also set $\frac{\epsilon}{\sigma} = \tau_e$ (**Relaxation time**). This simplifies the equation somewhat to

$$\frac{\partial n'}{\partial t} = D_e \frac{\partial^2 n'}{\partial x^2} + \frac{p' - n'}{\tau_\epsilon} + g(x,t) - \frac{n'}{\tau} \qquad (20b)$$

The equation for p' is similar:

$$\frac{\partial p'}{\partial t} = -D_e \frac{\partial^2 p'}{\partial x^2} - \frac{p' - n'}{\tau_\epsilon} + g(x,t) - \frac{p'}{\tau} \qquad (20c)$$

We can now show why time variations of n' and p' are unimportant. Take a simple case as an example. Suppose that n' and p' don't depend on position and that $g(t) = 0$. If we subtract

Equation 20b from Equation 20c we get $\frac{\partial(p'-n')}{\partial t} = -2\frac{p'-n'}{\tau_\epsilon} - \frac{p'-n'}{\tau} \approx -2\frac{p'-n'}{\tau_\epsilon}$, since

$\frac{1}{\tau_\epsilon} >> \frac{1}{\tau}$ ($\tau \sim 10^{-6}$ sec). The solution to the differential equation for (p' - n'),

$(p' - n') = (p'- n')_{t=0}\, e^{-\frac{2t}{\tau_\epsilon}}$, says that p' - n' falls exponentially with a time constant $\tau_\epsilon$,

which we know **(Relaxation time)** to be ~ $10^{-13}$ sec. This doesn't mean that the charge imbalances are reduced through recombination, but rather that the charges move in such a way that, on the average, there is no net charge density. If the charges are not mobile, as with the ions, then there can be a charge imbalance that is static, but if the charges are mobile, then imbalances cancel out in short time periods or over short distances of the order of $\sqrt{D\tau_\epsilon} = L_{Debye}$, where D is the diffusion constant ($L_{Debye}$ is called the Debye length, after Pieter Debye). Since $\tau_\epsilon$ ~ $10^{-13}$ sec and D ~ 50 cm, the distance $L_{Debye} \sim 10^{-6}$ cm = 10 nm (roughly 20 - 30 times the diameter of a Si atom).

This is an important result because, since the time constant is so short, we can ignore time variations in n' and p' caused by outside influences whose variation rate is long compared to this time constant. That is, we can ignore the effect of electric fields $E(t) = E_o\, e^{j\omega t}$ on n' and p' when $\omega \ll \frac{1}{\tau_\epsilon}$. Therefore we need only concern ourselves with the spatially varying parts of the equations governing n' and p': for electrons we deal only with the second order differential equation

$$D_e \frac{d^2n'}{dx^2} - \frac{n'}{\tau} = -g(x) \qquad (21a)$$

and for holes we have

$$D_h \frac{d^2 p'}{dx^2} - \frac{p'}{\tau} = - g(x) \qquad (21b)$$

The product of the diffusion constant $D_h$ or $D_e$ with $\tau$ has units of $cm^2$. In the case of holes $D_h\tau = L_h^2$ and in the case of electrons $D_e\tau = L_e^2$. The lengths L are the distances minority carriers (*n* in p-type Si or *p* in n-type Si) will diffuse before their density has declined by the factor 1/e. $L_e$ and $L_h$ are of the order of micrometers (1 micrometer = $10^{-4}$ cm). This is easily seen for the case $g = 0$, for then Equations 21a and 21b have solutions of the form $n'(x) = A \exp^{x/L_e} + B \exp^{-x/L_e}$ and $p'(x) = A \exp^{x/L_h} + B \exp^{-x/L_h}$.

In doped Si only the majority carrier drift current and the minority carrier diffusion currents are important. The reason is that minority drift currents are many orders of magnitude less than majority drift currents, and since the density of the majority carriers is essentially constant, only the minority carriers will have significant (and often quite large) diffusion currents. The minority carrier density will change because the carriers are injected at one point and, as they move around and recombine with majority carriers and their numbers drop, their density decreases. We solve the differential equation for the minority carrier density and find the minority diffusion current by taking the spatial derivative of the carrier density. If we assume the total current density $J^{TOT}$ is known, a reasonable assumption since we presumably know the current flowing through the Si and the dimensions of the Si, then we can find the majority current density $J_{maj}^{TOT}$ by subtracting the minority diffusion current density $J_{min}^{diff}$ from the total current density - we ignore the minority drift current, remember. We can then find majority drift current $J_{maj}^{drift}$ by subtracting the majority diffusion current $J_{maj}^{diff}$ from the total majority current. How do we find $J_{maj}^{diff}$? We assume that the gradients of the minority excess charge density and the majority excess charge density are roughly equal because the Si crystal as a whole has no net charge over distances large compared to the Debye length. If the charge density evens itself out, then the

currents of the two carriers should be roughly equal. The majority diffusion current is found by using the minority carrier gradient instead of the majority carrier gradient. To see this, suppose for definiteness that the majority particles are holes and the minority particles are electrons. Then $J_{maj}^{diff} = J_h^{diff} = -qD_h \frac{dp'}{dx}$ and $J_{min}^{diff} = J_e^{diff} = qD_e \frac{dn'}{dx}$. With what is called the quasi-static approximation $\frac{\partial p'}{\partial x} \approx \frac{\partial n'}{\partial x}$ we have $J_{maj}^{diff} = -\frac{D_h}{D_e} J_{min}^{diff}$. The electric field in the crystal can then be found by using $J_{majority}^{drift} = \sigma E$.

To obtain numerical solutions to the differential equations we need to impose boundary conditions. What are the boundary conditions we might impose on differential equations I) and II)? In considering current flow through the p-n junction we have to worry about where the current comes from and where it goes to. The p-n junction diode or a BJT will have metallic contacts to the outside world and we assume that the excess carrier densities vanish at these contacts (this is certainly reasonable for holes, since the electron density in a metal is ~ $10^{22}$ $cm^{-3}$; it is reasonable for electrons too because the excess electrons will rapidly be removed through the low resistance metal contact ). If electrons or holes are being injected into the Si crystal across a p-n junction, then we must find a way of specifying the density of carriers at the boundary. Thus, in general we will be specifying the minority carrier densities n(x) or p(x) at specific locations x. Since we are interested in finding the minority carrier densities and since the diffusion currents are proportional to the gradients of these densities, if there is no current flow (i.e., the Si isn't connected to anything) then another boundary condition would be that the gradient $\frac{dn}{dx}$ or $\frac{dp}{dx}$ is zero at the end of the Si. We have provided some examples **(Density distribution)**

## *IX. The p-n junction and some of its properties*

Now that we have a set of equations to determine the flow of current through a semiconductor, we are in a position to determine the characteristics of a p-n junction. A p-n

junction consists of silicon that has been doped differently in two regions that have a fairly sharp boundary. In a thought experiment you can think of a piece of n-type Si and a piece of p-type Si joined together to make a single crystal, the p-n junction being their common border. Since this isn't possible in real life, we can instead think of a single piece of Si that has been doped differently on either side of a sharp boundary. This is possible to do in real life and in fact, it's how semiconductor devices are actually made.

We will consider the simplest approximation of a p-n junction: an infinitely thin boundary between p and n-type Si, but to begin we won't impose any conditions on the way the Si is doped. A key feature of the way non-uniformly doped Si behaves is that the holes and electrons will tend to diffuse away from the regions of high density. As this happens an electric field will begin to develop as the positive and negative charges move around, leaving behind negative ions (atoms with valence 3 that give up a hole) and positive ions (valence 5 atoms that give up an electron). The electric field will tend to push the holes and electrons back to where they came from (in the opposite direction of diffusion) and diffusion currents and drift currents balance each other. The electric field implies there will be a spatially varying potential inside the Si. To be specific, suppose the Si is p-type and that $N_a = N_a(x)$, as shown in Figure 4

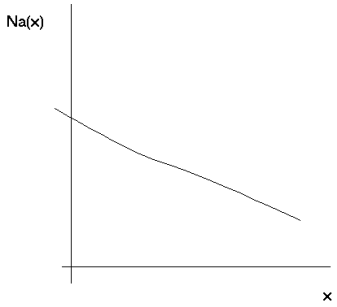



Figure 4. The variation of acceptor impurity density with position in non-uniformly doped Si. Consider our equations governing current flow and electric fields in the case of thermal equilibrium, where there is generation or recombination taking place.

I) $J_e^{tot} = J_e^{diff} + J_e^{drif} = q\, n_o\, \mu_e\, E + q\, D_e \frac{\partial n}{\partial x}$

II) $J_p^{tot} = J_h^{diff} + J_h^{drif} = q\, p_o\, \mu_h\, E - q\, D_h \frac{\partial p}{\partial x}$

III) $\epsilon \frac{\partial E}{\partial x} = q\,[p + N_d(x) - n - N_a(x)]$

IV) $\frac{\partial J_e}{\partial x} = q \frac{\partial n}{\partial t}$

V) $\frac{\partial J_h}{\partial x} = -q \frac{\partial p}{\partial t}$

Since the Si is in thermal equilibrium there can be no current flow, so $J_e^{tot} = J_h^{tot} = 0$. Also, in equilibrium there can be no time variation of n or p, so the time derivatives in Equations IV) and V) must be zero and in fact IV) and V) are merely identities, and $n = n_o$, $p = p_o$. The first three equations are then

I') $q\, n_o(x)\, \mu_e\, E + q\, D_e \frac{dn_o(x)}{dx} = 0$

II') $q\, p_o(x)\, \mu_h\, E - q\, D_h \frac{dp_o(x)}{dx} = 0$

III') $\epsilon \frac{dE(x)}{dx} = q\,[p_o + N_d(x) - n_o - N_a(x)]$

where the partial derivatives have been replaced by total derivatives since there is no time dependence. A profitable way to proceed is to write the electric field as $E(x) = -\frac{d\phi}{dx}$ so that

Equation II') becomes

$$-q\, p_o(x)\, \mu_h \frac{d\phi}{dx} - q\, D_h \frac{\partial p_o(x)}{\partial x} = 0 \qquad (22)$$

or, on cancelling the charge q,

$$p_o(x)\ \mu_h \frac{d\phi}{dx} = -D_h \frac{dp_o(x)}{dx} \qquad (23)$$

To simplify this we return to the diffusion equation **(Diffusion)** where the diffusion constant D was stated to be $D = \frac{l\bar{v}}{3}$, where $l$ is the mean distance between collisions and $\bar{v}$ is the mean speed.

The mobility is $\mu = \frac{q\tau}{m}$ where $\tau$ is the mean time between collisions **(Mobility)**. If we take the ratio of the diffusion constant to the mobility we have $\frac{D}{\mu} = \frac{\frac{l\bar{v}}{3}}{\frac{q\tau}{m}} = \frac{\frac{l\bar{v}}{3}}{\frac{q\tau}{m}} = \frac{ml\bar{v}}{3q\tau}$.

Since the average speed is the average distance traveled between collisions divided by the average time between collisions, $\bar{v} = \frac{l}{\tau}$, we end up with $\frac{D}{\mu} = \frac{m\bar{v^2}}{3q}$. The kinetic theory of gases **(Maxwell-Boltzmann distribution)** tells us that the average (thermal) kinetic energy of a particle is $\frac{1}{2} m\bar{v^2} = \frac{3}{2}kT$, therefore $\frac{D}{\mu} = \frac{kT}{q}$. This is called the Einstein relationship; it is true for holes and electrons: $\frac{D_h}{\mu_h} = \frac{D_e}{\mu_e} = \frac{kT}{q}$. Thus we can re-write Equation (23),

$p_o(x) \frac{d\phi}{dx} = -\frac{D_h}{\mu_h} \frac{dp_o(x)}{dx}$ as

$$\frac{q}{kT}\frac{d\phi}{dx} = -\frac{1}{p_o(x)}\frac{d\,p_o(x)}{dx} \tag{24}$$

When this is integrated we get an equation relating the internal potential of the Si to the density of holes:

$$\ln\left(\frac{p_o(x)}{p_o}\right) = -\frac{q}{kT}(\phi(x) - \phi_o) \tag{25}$$

or

$$p_o(x) = p_o\, e^{-\frac{q}{kT}(\phi(x) - \phi_o)} \tag{26}$$

$\phi_o$ is the potential reference point where $p_o(x) = p_o$. For convenience we take $\phi(x) = 0$ where $p_o(x) = n_i$, the intrinsic density of holes and electrons. Then $p_o(x) = n_i\, e^{-q\phi/kT}$. It is straightforward to show that $n_o(x) = n_i\, e^{q\phi/kT}$, with $n_o(x) = n_i$ when $\phi(x) = 0$.

We now look at what happens when the p-n junction is abrupt, as shown in Figure 5, below. The meaning of Figure 5 is that $N_D = N_d$ for $x > 0$ and $N_D = -N_a$ for $x < 0$. For $x << 0$, $p_o = N_a$ and , so the potential is $\phi_p = -\frac{kT}{q}\ln\left(\frac{N_a}{n_i}\right)$. For $x >> 0$, $n_o = N_d$ and the potential is $\phi_n = \frac{kT}{q}\ln\left(\frac{N_d}{n_i}\right)$.

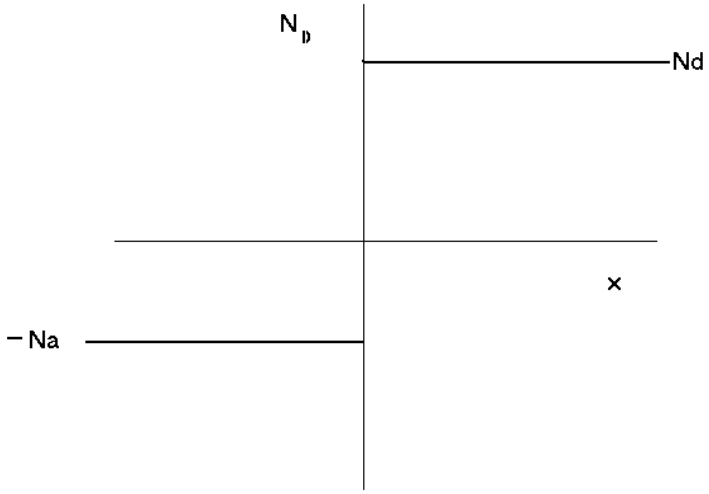



Figure 5. The doping profile of Si doped uniformly in the regions $x < 0$ and $x > 0$.

If we imagine that the p-type and n-type Si are instantaneously created at time $t = 0$, then immediately afterward, just to the right of $x = 0$ at $x = +\delta$, $n_o >> p_o$, and just to the left of $x = 0$ at $x = -\delta$, $p_o >> n_o$ ,where $\delta$ is a very small distance. Because the electron density is much higher for $x = +\delta$ than it is for $x = -\delta$, electrons would immediately begin diffusing towards the left, as described previously. Similarly, holes would immediately begin diffusing towards the right. The electrons would leave behind positively charged donor ions (the ions are fixed in the crystal lattice and cannot move) while the holes would leave behind negatively charged acceptor ions. The result is a net positive charge for $x > 0$ and a net negative charge for $x < 0$. This charge imbalance will cause an electric field pointing to the left, and the field will tend to drive the positively charged holes back to the left and the negatively charged electrons back to the right. Eventually an equilibrium will be established in which the diffusion and drift currents just cancel each other.

The situation just described can be relatively easily analyzed using the *depletion approximation*. This approximation assumes all the donors and acceptors within a small distance of the p-n junction are ionized, and outside of these small distances none of them are ionized.

Although simple, it is a very good approximation. If we graph the net charge density describing this approximation we get the result shown in Figure 6.

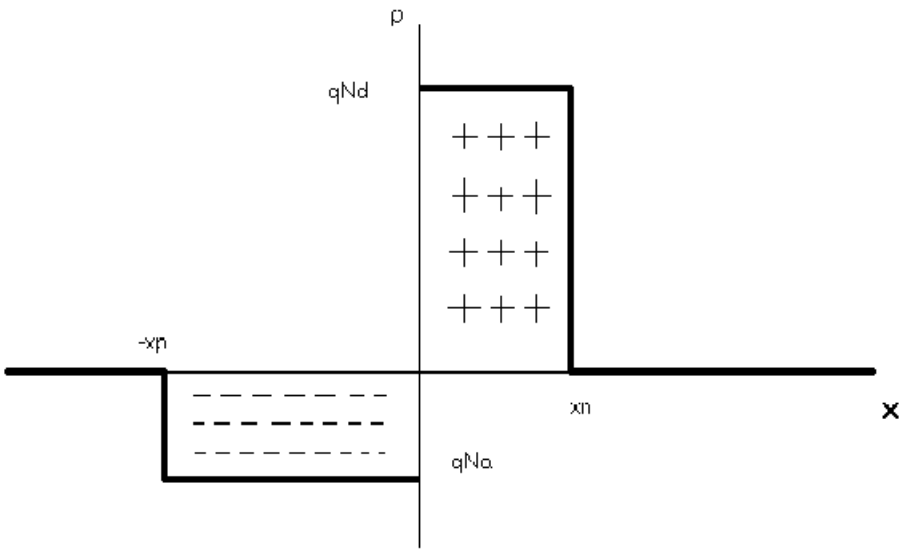



Figure 6. Schematic diagram of the charge distribution of an abrupt p-n junction in the depletion approximation.

In reality, of course, there will not be a sharp cutoff of the charge density at $-x_p$ and $+x_n$, but the approximation is very good. Free holes and electrons exist only for $x < -x_p$ and $x > x_n$ and it is easy to find the fields and the potentials in the Si. First, however, note that the Si as a whole is electrically neutral - the fabrication procedure involves implanting donor and acceptor atoms into Si that is kept electrically grounded. Therefore the total negative charge must balance the total positive charge. Since the cross sectional area of the junction (the area perpendicular to the x-axis) is constant, the volume is proportional to x. Thus the absolute value of the total negative charge is $\mathbf{Q = q\ N_a\ A\ x_p}$ while absolute value of the total positive charge is $\mathbf{Q = q\ N_d\ A\ x_n}$. When these expressions are equated we find an equation for $x_p$ and $x_n$: $\mathbf{N_a\ x_p = N_d\ x_n}$. The charge density as a function of position is given in Table 1 below.

| x | $\rho(x)$ |
|---|---|
| $x < -x_p$ | 0 |
| $-x_p < x < 0$ | $-qN_a$ |
| $0 < x < x_n$ | $+qN_d$ |
| $x_n < x$ | 0 |

Table 1. The variation of charge density with position for the abrupt junction of Figure 6.

To find the potential we must first find the electric field distribution and integrate it. The electric field is found by using Gauss' Law and integrating the charge density with the boundary condition that there should be no electric field for $x < -x_p$ and for $x > x_n$ . We integrate $\epsilon \frac{dE}{dx} = \rho(x)$, where the dielectric constant of Si $\epsilon$ is assumed to be constant, and get the result shown in Table 2:

| x | $E(x)$ |
|---|---|
| $x < -x_p$ | 0 |
| $-x_p < x < 0$ | $-qN_a (x + x_p)/\epsilon$ |
| $0 < x < x_n$ | $+qN_d (x - x_n)/\epsilon$ |
| $x_n < x$ | 0 |

Table 2. The electric field for the abrupt junction of Figure 6.

The electric field distribution is shown in Figure 7 below.

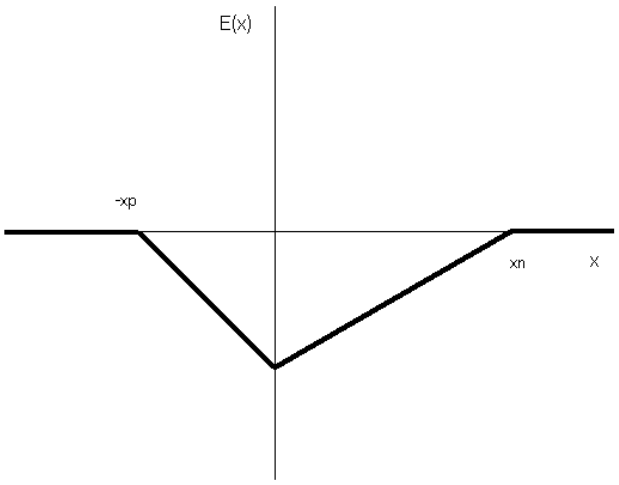



Figure 7. The electric field distribution of an abrupt p-n junction based on the depletion approximation. The field is directed towards the left, from the n-side of the junction towards the p-side.

We now find the potential by integrating $E = -\frac{d\phi}{dx}$, with the boundary conditions that

$\phi = \phi_p$ for $x < -x_p$ and $\phi = \phi_n$ for $x > x_n$. The result is shown in Table 3 and in Figure 8 below.

| x | $\phi(x)$ |
|---|---|
| $x < -x_p$ | $\phi_p$ |
| $-x_p < x < 0$ | $\phi_p + qN_a (x + x_p)^2/2\epsilon$ |
| $0 < x < x_n$ | $\phi_n - qN_d (x - x_n)^2/2\epsilon$ |
| $x_n < x$ | $\phi_n$ |

Table 3. The electrostatic potential in the Si for the abrupt junction of Figure 6.

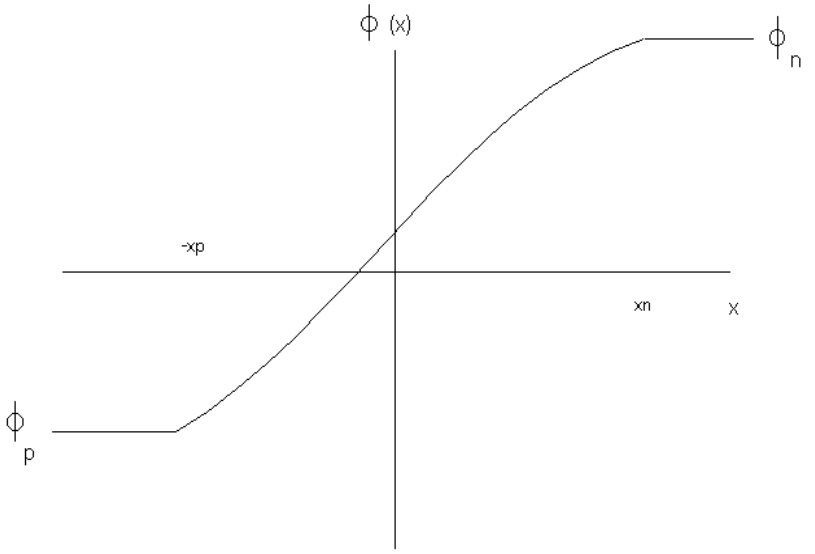



Figure 8. Schematic representation of the electrostatic potential of an abrupt p-n junction based on the depletion approximation.

The equations for the field *E* give no new information about $x_n$ and $x_p$, but the equations for the potential do. If we match the two expressions for the potential for $x < 0$ and for $x > 0$ at x = 0, we get $\phi_p + q N_a \frac{(x_p)^2}{2\epsilon} = \phi_n - q N_d \frac{(x_n)^2}{2\epsilon}$ .With $N_a x_p = N_d x_n$ we find, with a bit of algebra,

$$x_n = \sqrt{\frac{2\epsilon\phi_m}{q} \frac{N_a}{N_d (N_a + N_d)}}$$

$$x_p = \sqrt{\frac{2\epsilon\phi_m}{q} \frac{N_d}{N_a (N_a + N_d)}} \qquad (27)$$

where the "mean" value of $\phi$, $\phi_m$, is defined as

$$\phi_m = \frac{kT}{q} \ln\left(\frac{N_d N_a}{n_i^2}\right) \tag{28}$$

The width of the region which is depleted of free charges - the holes and electrons - is given by

$$w = x_n - (-x_p) = x_n + x_p = \sqrt{\frac{2\epsilon\phi_m}{q}\,\frac{N_a + N_d}{N_a N_d}} \tag{29}$$

The region between $-x_p$ and $x_n$ is sometimes called the space-charge region. The electric field in this region is quite strong - at its peak the field can reach more than $10^5$ V/cm (10 volts per micrometer). The field is large enough that any hole or electron that wanders into the space-charge region is immediately swept out again.

## *X. The properties of a p-n junction when a voltage is applied across it*

Before we begin this section, we address a couple of questions that might occur: can you measure the potential cross the p-n junction; might the junction act like a battery? The answer is no, and the reason is most easily seen by considering a circuit involving a p-n junction as shown in Figure 9:

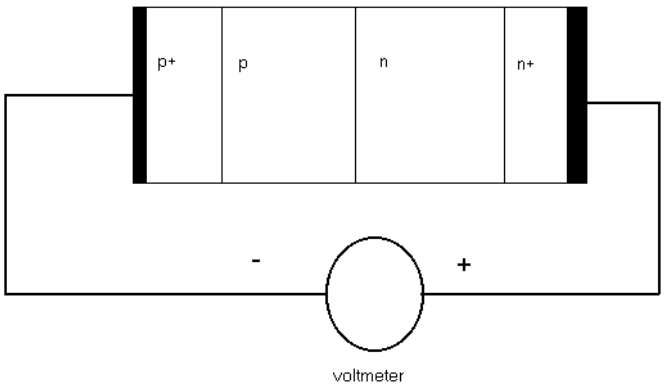



Figure 9. A p-n junction hooked up so as to be able to measure the potential across it.

In Figure 9 the thick black lines at each end of the Si represent Al contacts that connect the junction to the outside world. The regions marked $p^+$ and $n^+$ are Si that is heavily doped to make a good electrical contact with the Al. There is a potential called a contact potential between the metal of the Al leads and the Si relative to the zero or reference potential of pure (intrinsic) Si which is about -0.3 volts (that is, there is a 0.3 volt decrease in potential when going from Si to Al. If we assume the voltmeter contributes nothing to the potential (no potential drop in the meter), we can now draw a potential diagram for the whole system.

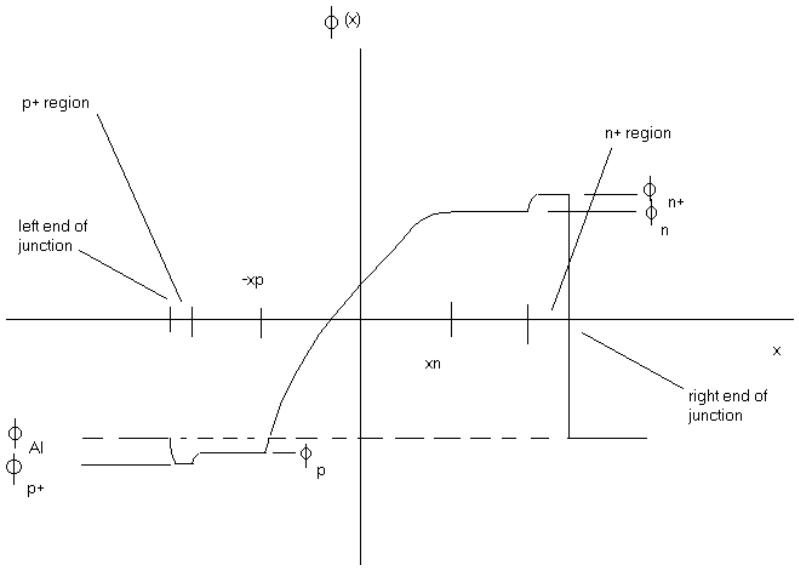



Figure 10. The potential around the p-n junction circuit shown in Figure 9.

As you might expect, since there are no perpetual motion machines, the potential difference all the way around the circuit is zero, i.e., the p-n junction will not behave like a battery. The potential starts out at $\phi_{Al}$ at the left end, drops to $\phi_{p^+}$ at the connection between the Al and the $p^+$ Si, rises to $\phi_p$ where the $p^+$ -Si meets the p-Si, begins to rise again at $x = -x_p$ and reaches $\phi_n$ at $x = x_n$. The potential rises a little more to $\phi_{n^+}$ at the point where the n-Si becomes $n^+$ Si, and the drops back down to $\phi_{Al}$ at the Si - Al interface. Incidentally, at a metal -

Si interface there is a potential change that looks rather like what happens across a p-n junction, except that since there are so many free electrons in a metal there is no depletion in the metal. Rather, for n-type Si there is a very thin (~ 1 nm) sheet of negative charge along the surface of the metal. In the Si the depletion range is of normal dimension. For p-type Si the layer of positive charges due to the holes is almost infinitely thin at the boundary. This is called a Schottky barrier and the phenomenon can be used to make a diode.

Next we consider what happens when a voltage is placed across the p-n junction, as shown in Figure 11.

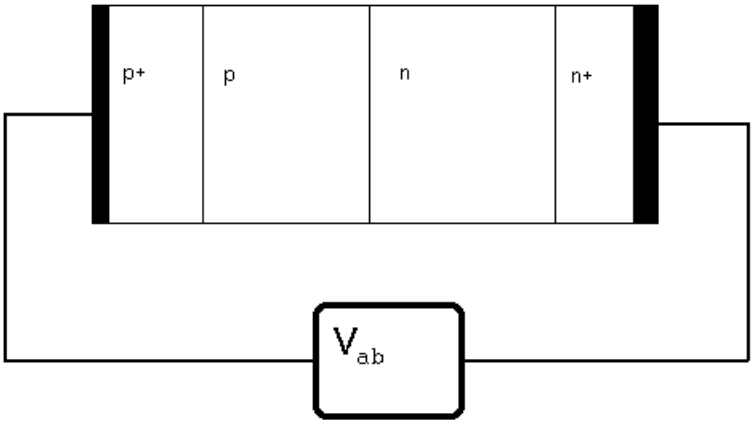



Figure 11. A p-n junction with a potential $V_{ab}$ applied across it.

Since we are only interested in what happens at the junction itself, we will ignore the metal leads and the heavily doped regions. When the additional potential $V_{ab}$ is placed across the junction the width of the depletion or space charge region changes. This has two effects: it affects the current flowing through the junction and it affects the capacitance of the junction. The width changes because the potential $\phi_m \rightarrow \phi_m - V_{ab}$ and the width of the depletion region w behaves as

$$w = x_n + x_p = \sqrt{\frac{2\epsilon\phi_m}{q}\,\frac{N_a + N_d}{N_a N_d}} \rightarrow \sqrt{\frac{2\epsilon(\phi_m - V_{ab})}{q}\,\frac{N_a + N_d}{N_a N_d}} \qquad (30)$$

which means that as $V_{ab}$ increases, w decreases (the sense of $V_{ab}$ is that, as it increases the p-side of the junction becomes more positive relative to the n-side).

*X-A. Depletion capacitance of the p-n junction*

Capacitance is defined as the amount of charge stored per unit voltage, or $C = dQ/dV$. To consider the charge stored in the depletion region we repeat Figure 6:

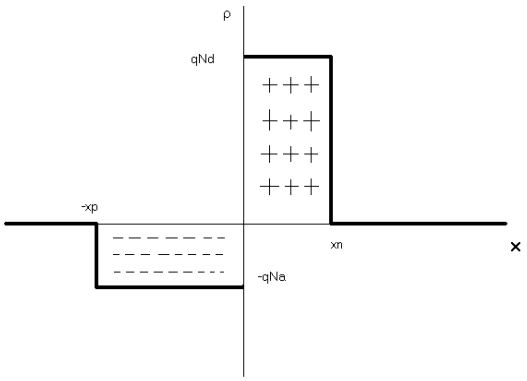



Figure 12 (Figure 6 repeated). Schematic diagram of the charge distribution of an abrupt p-n junction in the depletion approximation.

The charge stored between x = 0 and x = $x_n$ is $Q = qN_dAx_n$ where $N_d$ is the density of positive ions, A is the cross sectional area and q is the electronic charge ($1.6 \times 10^{-19}$ C). With the expression for $x_n$ (Equation 27) and replacing $\phi_m$ by $\phi_m - V_{ab}$ we get

$$x_n = \sqrt{\frac{2\epsilon\phi_m}{q}\,\frac{N_a}{N_d(N_a + N_d)}} \rightarrow \sqrt{\frac{2\epsilon(\phi_m - V_{ab})}{q}\,\frac{N_a}{N_d(N_a + N_d)}} \qquad (31)$$

With this value for $x_n$ we find the stored charge Q to be

$$Q(V_{ab}) = A\sqrt{2q\epsilon(\phi_m - V_{ab})\frac{N_a N_d}{N_a + N_d}} \qquad (32)$$

where we write $Q = Q(V_{ab})$ since Q is a function of $V_{ab}$. It is clear that if we try to express the capacitance through $C = \frac{dQ}{dV}$ we will end up with a non-linear relationship whose meaning will be complicated. However, if the change in $V_{ab}$ is small, we can linearize Equation 32 by expanding $Q(V_{AB})$ in a Taylor's series and keeping only the first term:

$$C(V_{ab}) = \frac{dQ(V_{ab})}{dV_{ab}} \approx A\sqrt{\frac{q\epsilon}{2(\phi_m - V_{ab})}\frac{N_a N_d}{N_a + N_d}} \equiv C_{depletion} \qquad (33)$$

Note that this implies that when the voltage applied across the junction changes with time the charge stored in the junction will change with time so there will be a current. If the voltage changes by a small amount so $V_{ab} = V_{AB} + v(t)$, then the current flowing in the junction will be given by $i = \frac{dQ}{dt} = \frac{dQ}{dV}\frac{dV}{dt} = C\frac{dv(t)}{dt}$, where C is $C_{depletion}$ from Equation 33. With $\phi_m$ = 0.7 V, $V_{ab}$ = 0 and typical values $N_a = 10^{17}$ cm$^{-3}$ , $N_d = 10^{18}$ cm$^{-3}$, q = 1.6 x $10^{-19}$ C, A = $10^{-8}$ m$^2$ (100 micrometers square) and the dielectric constant of Si being $\epsilon \approx 10^{-10}$ F m$^{-1}$, $C_{depletion} \sim 10^{-14}$ F or ~ $10^{-6}$ F m$^{-2}$ ($10^{-10}$ F cm$^{-2}$).

*X-B. Current flow through the junction*

In order to calculate the current flow through the p-n junction we have to solve the second order differential equation for the carriers in the two parts (p and n) of the junction, subject to the appropriate boundary conditions (this last is a very important point). Then we can find the drift currents and, by differentiating the densities, we get the diffusion currents. The boundary

conditions are very important and contain physical information that is the key to the junction behavior.

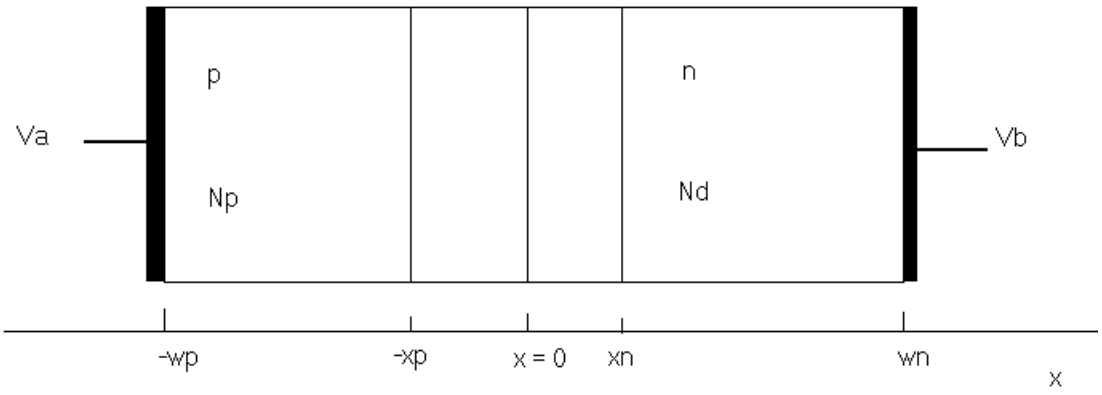



Figure 13. Schematic diagram of a junction diode. The distance $w_p - x_p >> x_p$ and $w_n - x_n >> x_n$. The ends of the diode are at $-w_p$ and $w_n$, where the Al contacts are placed. The voltage across the junction diode is Va - Vb = $V_{ab}$ . The heavily doped regions $p^+$ and $n^+$ near the metal contacts are ignored.

To begin we examine a diagram of the junction, or, as we will now call it, the junction diode (Figure 13). The region between $-x_p$ and $x_n$ is the depletion region. The regions between $-x_p$ and $-w_p$ , and between $x_n$ and $w_n$ are called *quasi-neutral* regions. This means that in any region large compared to the Debye length, there is no net charge - the holes, electrons and ions approximately balance each other out. The only place there is a net charge density is in the depletion region. More precisely, in the quasi-neutral regions we have the relations $n'(x) = n(x) - n_o \approx p'(x) = p(x) - p_o$ where the meaning of "≈" is that $|n'(x) - p'(x)| << n'(x) + p'(x)$. Also, quasi-neutrality requires $\left|\frac{\partial n'(x)}{\partial x} - \frac{\partial p'(x)}{\partial x}\right| << \left|\frac{\partial n'(x)}{\partial x} + \frac{\partial p'(x)}{\partial x}\right|$. These are the equations that define low level injection, which was mentioned previously. These may

seem like drastic requirements, but they aren't, actually. The reason is that any imbalance in charges is rapidly removed by motion of the carriers, as was described in section *VIII*. Only for distances of the order of nanometers or less are there any imbalances. Over macroscopic distances of the order of micrometers, there is essentially no imbalance to be seen: electrostatic forces are very strong and charges move very quickly **(Relaxation time)**.

Before considering the equations and their solutions, we will explain in words how we calculate the current flows. Differential equations 21a and 21b from section *VIII,*

$D_e \frac{d^2n'}{dx^2} - \frac{n'}{\tau} = -g(x)$ for electrons and $D_h \frac{d^2p'}{dx^2} - \frac{p'}{\tau} = -g(x)$ for holes (remember

that $p_o$ and $n_o$ are independent of position), can be solved in the quasi-neutral regions to find the excess electron and hole densities. We know that any charge that is injected into the depletion region moves through it very quickly (the electric field is high so the velocities are high, implying small chance of electron-hole recombination), so we assume there is no loss or recombination taking place there. This means we only have to consider the currents that flow in the quasi-neutral regions. To find these currents we first find the solutions to the differential equations, which means boundary conditions need to be specified at the ends of the quasi-neutral regions: at $x = -w_p$; at $x = -x_p$; at $x = x_n$; and at $x = w_n$. The boundary conditions at $-w_p$ and at $w_n$ are simple: $n' = p' = 0$, since the metal effectively absorbs all electrons and holes that flow into it. Electrons are injected into the diode at $w_n$ and holes are injected in at $-w_p$. Beyond these points the injection is zero, and so we can set $g(x) = 0$ within the quasi-neutral regions (this means the differential equations for the holes and electrons are homogeneous). So, to finish the problem it is necessary to relate the density of holes at $-x_p$ to the density of holes at $x_n$, and the density of electrons at $x_n$ to the density of electrons at $-x_p$. When this is done, we will have the necessary boundary conditions to obtain numerical solutions for the equations for $n'(x)$ and $p'(x)$.

How do we find $p'(x_n)$ and $n'(-x_p)$? We begin by assuming a junction diode in thermal equilibrium with no applied voltage. We then know there is no net motion of the holes and electrons and so the density of holes at $x = -x_p$ is $p_o(-x_p) = N_a$. Similarly, the density of electrons at $x = x_n$ is $n_o(x_n) = N_d$. At the ends of the junction diode we have $p_o(-w_p) = 0$ and $n_o(w_n) = 0$.

Consider the case of electrons moving from the n-doped region through the depletion region into the p-doped region. In the n-doped region the electrons are majority carriers and $n_o >> p_o$. At x = $-x_p$ the electrons are minority carriers and $n_o(-x_p) = \frac{n_i^2}{p_o} = \frac{n_i^2}{N_a}$. If we divide $n_o(-x_p)$ by $n_o(x_n)$ we get

$$\frac{n_o(-x_p)}{n_o(x_n)} = \frac{n_i^2}{N_a N_d} \tag{34}$$

The quantity on the right side of the last equation is familiar. If we look back to Equation 28, $\phi_m = \frac{kT}{q} \ln\left(\frac{N_d N_a}{n_i^2}\right)$, we see that

$$\frac{n_i^2}{N_a N_d} = e^{-\frac{q\phi_m}{kT}} \tag{35}$$

and as a result we end up with the very important relationship

$$n_o(-x_p) = n_o(x_n)\, e^{-\frac{q\phi_m}{kT}} \tag{36}$$

This says that the density of electrons on the p (left) side of the depletion region is smaller than the density on the n (right) side by the factor $e^{-q\phi_m/kT}$. This is called the Boltzmann factor **(Maxwell-Boltzmann distribution)**. Here is what this means. The location x = $-x_p$ is at a potential of $-\phi_m$ relative to the location x = $x_n$. Since the electrons carry a negative charge, as far as they are concerned the location x = $-x_p$ is at an energy U = $+q\phi_m$ relative to the location x = $x_n$. Now, the energy of the particles is governed by the Maxwell-Boltzmann distribution, which means the fraction of the particles that have kinetic energy $U = +q\phi_m >> \overline{U}$ (where $\overline{U}$ is the average kinetic energy), is $e^{-q\phi_m/kT}$. Put another way, there are electrons moving around in the

region $x > x_n$ with all sorts of velocities, and the distribution of velocities is governed by the Maxwell-Boltzmann relation. When these electrons run up against the boundary of the depletion region, only those with kinetic energy greater than $+q\phi_m$ will be able to surmount the potential barrier ("get to the top of the hill") and make it to the p-side of the depletion region (see Figure 14). Exactly the same thing happens with the holes going in the opposite direction from the p-side to the n-side of the junction. As far as the holes are concerned, the location $x = x_n$ is at a positive potential $+\phi_m$ relative to the location $x = x_n$ and so, since the holes carry a positive charge they also see an energy barrier of height $+q\phi_m$. Thus for holes

$p_o(x_n) = p_o(-x_p)\, e^{-q\phi_m/kT}$. The energy barrier looks something like this

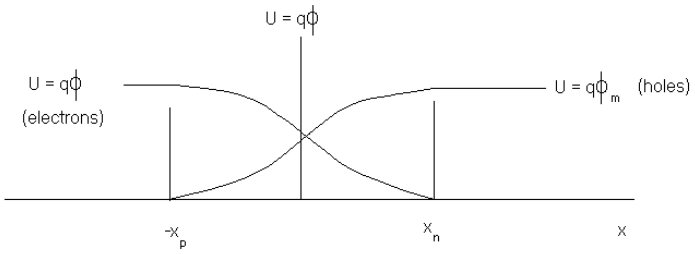



Figure 14. Schematic diagram of the potential energy for holes and electrons as a function of position across the depletion region. The energy of the hole starts at U = 0 at $x = -x_p$ and rises to $U = +q\phi_m$ at $x = x_n$. The potential energy for an electron starts at U = 0 at $x = x_n$ ands rises to $U = +q\phi_m$ at $x = -x_p$.

If we now apply a potential $V_{ab}$ across the junction, $\phi_m \rightarrow \phi_m - V_{ab}$. Holes and electrons are injected into the p-side and n-side of the junction, respectively, from a battery or power supply. Recall that the total charges are denoted by p(x) and n(x). As long as the current is not too large so that the quasi-neutrality approximation remains valid, we can approximate $n(x_n)$

by $n_o(x_n)$ ($n'(x) = n(x) - n_o(x) << n_o(x)$ ) and, since all that has changed is the barrier height, we have

$$n(-x_p) \approx n_o(x_n)\, e^{-\frac{q(\phi_m - V_{ab})}{kT}} \tag{37}$$

The excess charge $n'(-x_p) = n(-x_p) - n_o(-x_p)$ is given by

$$\begin{aligned} n'(-x_p) &= n(x_n)\, e^{-\frac{q(\phi_m - V_{ab})}{kT}} - n_o(-x_p) \\ &\approx n_o(x_n)\, e^{-\frac{q(\phi_m - V_{ab})}{kT}} - n_o(-x_p) \\ &= n_o(x_n)\, e^{-\frac{q\phi_m}{kT}}\, e^{+\frac{qV_{ab}}{kT}} - n_o(-x_p) \\ &= n_o(-x_p)\, e^{+\frac{qV_{ab}}{kT}} - n_o(-x_p) \\ &= n_o(-x_p)\left( e^{+\frac{qV_{ab}}{kT}} - 1 \right) \end{aligned} \tag{38}$$

Finally, using $n_o(-x_p) = \frac{n_i^2}{N_a}$ we end up with

$$n'(-x_p) = \frac{n_i^2}{N_a}\left( e^{\frac{qV_{ab}}{kT}} - 1 \right) \tag{39}$$

and by similar reasoning for the holes, we find, with $p_o(x_n) = \frac{n_i^2}{N_d}$ ,

$$p'(x_n) = \frac{n_i^2}{N_d}\left( e^{\frac{qV_{ab}}{kT}} - 1 \right) \tag{40}$$

Equations 39 and 40, and $n'(-w_p) = 0$ and $p'(w_n) = 0$ are our boundary conditions for the equations $D_e \frac{d^2n'}{dx^2} - \frac{n'}{\tau} = 0$ and $D_h \frac{d^2p'}{dx^2} - \frac{p'}{\tau} = 0$, and with the solution to these equations we can find the currents in the p-n junction when a potential is placed across it.

We consider first the hole current flowing from the p-side of the junction towards the n-side. The solution to the equation for the minority holes in the n-type Si, $D_h \frac{d^2p'}{dx^2} - \frac{p'}{\tau} = 0$, is $p'(x) = A\, e^{x/L_h} + B\, e^{-x/L_h}$. With an ohmic boundary at $x = w_n$ (see Figure 13, repeated below) $A\, e^{w_n/L_h} + B\, e^{-w_n/L_h} = 0$ and the constant B has the value $B = -A\, e^{2w_n/L_h}$. The density of excess holes is now given by

$$\begin{aligned} p'(x) &= A\left( e^{\frac{x}{L_h}} - e^{\frac{2w_n}{L_h}} e^{-\frac{x}{L_h}} \right) \\ &= Ae^{\frac{w_n}{L_h}}\left( e^{\frac{x - w_n}{L_h}} - e^{-\frac{x - w_n}{L_h}} \right) \\ &= 2Ae^{\frac{w_n}{L_h}} \sinh\left( \frac{x - w_n}{L_h} \right) \end{aligned} \qquad (41)$$

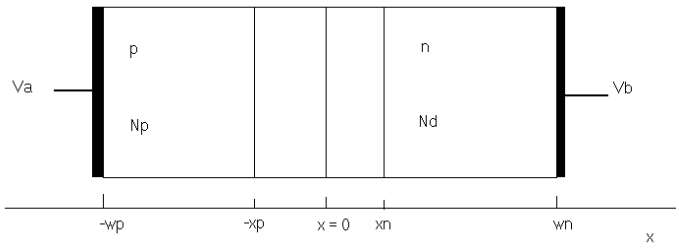



Figure 13 (repeated). Schematic diagram of a junction diode. The distance $w_p - x_p >> x_p$ and $w_n - x_n >> x_n$. The ends of the diode are at $-w_p$ and $w_n$, where the Al contacts are placed. The voltage across the junction diode is $V_a - V_b = V_{ab}$ .

At the right edge of the depletion region we saw that $p'(x_n) = \frac{n_i^2}{N_d}\left(e^{\frac{qV_{ab}}{kT}} - 1\right)$, so we have

$$\begin{aligned} p'(x_n) &= \frac{n_i^2}{N_d}\left(e^{\frac{qV_{ab}}{kT}} - 1\right) \\ &= 2A\, e^{\frac{w_n}{L_h}} \sinh\left(\frac{x_n - w_n}{L_h}\right) \end{aligned} \tag{42}$$

and the constant A is

$$A = \frac{n_i^2}{N_d} \frac{e^{-\frac{w_n}{L_h}}\left(e^{\frac{qV_{ab}}{kT}} - 1\right)}{2\sinh\left(\frac{x_n - w_n}{L_h}\right)} \tag{43}$$

and the density of holes is (see Figure 15)

$$\begin{aligned} p'(x) &= \frac{n_i^2}{N_d} \frac{\left(e^{\frac{qV_{ab}}{kT}} - 1\right)\sinh\left(\frac{x - w_n}{L_h}\right)}{\sinh\left(\frac{x_n - w_n}{L_h}\right)} \\ &= \frac{n_i^2}{N_d}\left(e^{\frac{qV_{ab}}{kT}} - 1\right) \frac{\sinh\left(\frac{w_n - x}{L_h}\right)}{\sinh\left(\frac{w_n - x_n}{L_h}\right)} \end{aligned} \tag{44}$$

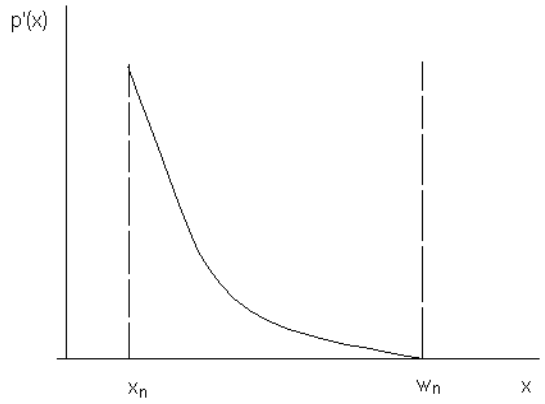



Figure 15. A schematic drawing of the density of excess holes in the n-side of a p-n junction. The distances $x_n$ and $w_n$ are not to scale.

The hole diffusion current density is found by differentiating p'(x):

$$J_h^{diff} = -qD_h \frac{dp'(x)}{dx}$$

$$= q\frac{D_h}{L_h}\frac{n_i^2}{N_d}\left( e^{\frac{qV_{ab}}{kT}} - 1 \right) \frac{\cosh\left( \frac{w_n - x}{L_h} \right)}{\sinh\left( \frac{w_n - x_n}{L_h} \right)} \qquad (45)$$

Note that at $x = w_n$ the current density is $\frac{q\frac{D_h}{L_h}\frac{n_i^2}{N_d}}{\sinh\left(\frac{w_n - x_n}{L_h}\right)}\left(e^{\frac{qV_{ab}}{kT}} - 1\right)$. The pre-exponential factor is of the order of $10^{-15}\ \frac{n_i^2}{N_d}$ A cm$^{-2}$. For $N_d = 10^{17}$ cm$^{-3}$ this is of the order of $10^{-12}$ A cm$^{-2}$.

The electron diffusion current density is calculated in exactly the same way and we find

$$J_e^{diff} = q\frac{D_e}{L_e}\frac{n_i^2}{N_a}\left(e^{\frac{qV_{ab}}{kT}} - 1\right)\frac{\cosh\left(\frac{w_p - x}{L_e}\right)}{\sinh\left(\frac{w_p - x_p}{L_e}\right)} \qquad (46)$$

To get a feel for the size of this, suppose $D_e \sim 50$ cm$^2$ sec$^{-1}$, $L_e = 100$ micrometers $= 10^{-2}$ cm, $N_a = 10^{17}$, $w_p = 100$ micrometers, $V_{ab} = 0.7$ V and T = 290K. Then $J_e^{diff} \approx 1$ A cm$^{-2}$. Clearly the diffusion current is large only where the derivative of the excess charge density is large. If $L_e$ or $L_h$ is small compared to the x-dimensions of the junction then as $x \to w_n$ (or $w_p$) the diffusion current will be small and the total current at the ends of the junction (near the metal contacts) will be dominated by the drift current. One way to picture this is to imagine a large current density of electrons injected into the n-side of the junction at $x = w_n$. These are supplied by a battery, say. As the current density moves towards the left it encounters a current density of holes injected across the space-charge region. As the excess electrons and holes recombine the result is a current density shared between diffusion and drift as shown in Figure 16:

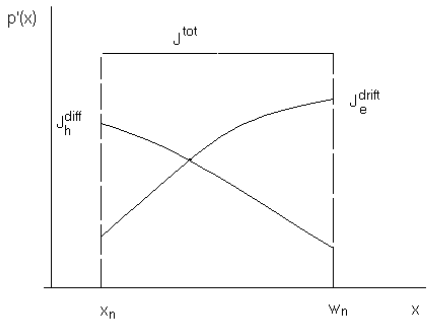



Figure 16. Schematic diagram of the electron drift current $J_e^{drift}$ and the hole diffusion current $J_h^{diff}$ in the n-side of a p-n junction. $J^{TOT} = J_e^{drift} + J_h^{diff}$ (not to scale).

The same thing happens on the p-side of the junction except with the roles of electrons and holes reversed.

So far we have said nothing about the currents through the space-charge region between $-x_p$ and $x_n$. This distance which is ~ 1 micrometer is small compared to the overall dimensions of a diode. The current through it consists of those holes and electrons that have sufficient energy to traverse it. Since the dimensions are small and the particles are moving quickly (they have relatively high energy) there will be little recombination in the depletion region and we can, to an excellent approximation, assume the total current density across the depletion region remains constant except at very low voltages when the width of the depletion region is larger and the particles moving less rapidly. This means the electron total current density at $x = x_n$ (shown in Figure 16) equals the electron diffusion current density at $x = -x_p$, and that the hole diffusion current density at $x = x_n$ equals the hole total current density at $x = -x_p$, as shown in Figure 17. Thus the total current density is $J^{TOT} = J_e^{diff}(-x_p) + J_h^{diff}(x_n)$:

$$J^{TOT} = qn_i^2 \left( \frac{D_h}{N_e L_h \tanh\left(\frac{w_n - x_n}{L_h}\right)} + \frac{D_e}{N_p L_e \tanh\left(\frac{w_p - x_p}{L_e}\right)} \right) \left( e^{\frac{qV_{ab}}{kT}} - 1 \right) \text{ or}$$

$$J^{TOT} = qn_i^2 \left( \frac{D_h}{N_e L_h^*} + \frac{D_e}{N_p L_e^*} \right) \left( e^{\frac{qV_{ab}}{kT}} - 1 \right) \tag{47}$$

where $L_h^* = L_h \tanh\left( \frac{w_n - x_n}{L_h} \right)$ and $L_e^* = L_e \tanh\left( \frac{w_p - x_p}{L_e} \right)$

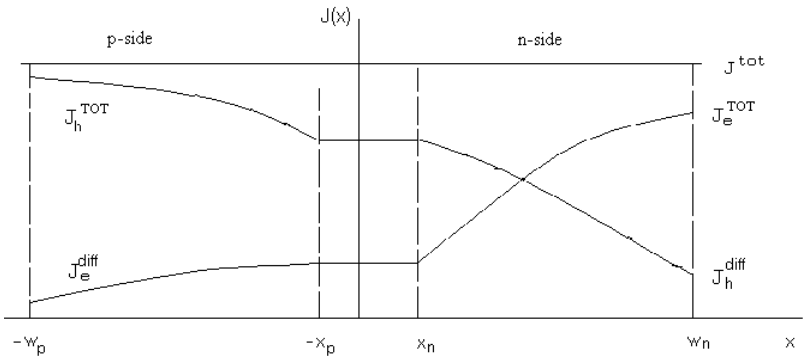



Figure 17. Schematic diagram of the electron and hole drift and diffusion current densities and the total current density through the p-n junction. Note the electron and hole current densities are assumed to be constant across the depletion region in the approximation we work with (ignoring recombination in the depletion region).

The current through the p-n junction diode is then

$$I_D = qAn_i^2 \left( \frac{D_h}{N_e L_h^*} + \frac{D_e}{N_p L_e^*} \right) \left( e^{\frac{qV_{ab}}{kT}} - 1 \right) = I_S \left( e^{\frac{qV_{ab}}{kT}} - 1 \right) \quad (48)$$

which defines the saturation or scale current $I_S$ (A is the cross sectional area of the diode). If the diode is *forward biased* ($V_{ab} > 0$), we have the behavior shown in Figure 18.

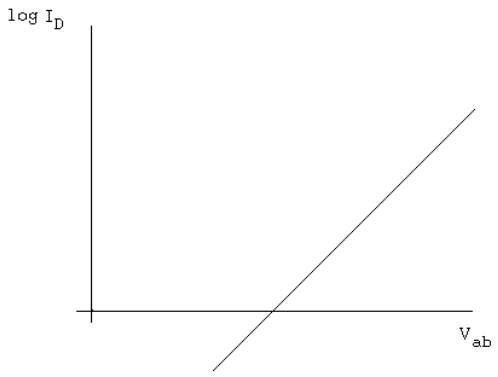



Figure 18. A log plot of the current $I_D$ through an ideal p-n junction diode as a function of applied voltage $V_{ab}$

There are some relatively small corrections to bring this result into closer accord with experiment. First, there is some resistance in the diode and so there will be an IR drop across it. This causes the current to rise less quickly than shown, at higher currents. Second, there is, in actuality, a certain amount of recombination that takes place in the depletion region. The additional recombination (above and beyond what happens in the regions $w_n - x_n$ and $-w_p - (-x_p)$ has to be made up for by additional holes and electrons injected into the junction. This happens

mainly at lower applied voltages and so when $V_{ab}$ is small the current is somewhat higher than what is indicated in Figure 18.

*X-C. Diffusion Capacitance*

In section *X-A* the capacitance across a p-n junction due to the charges in the depletion region was calculated ($C_{depletion}$) . There is an additional source of capacitance in the p-n junction that shows up only when current is flowing, called the diffusion capacitance ($C_{diffusion}$). The origin of this capacitance lies in the excess holes and electrons in the Si in the regions $w_n - x_n$ and $-w_p - (-x_p)$. Since there are adjacent charge densities we should expect some capacitance associated with them, but this capacitance vanishes when the current $I_D \rightarrow 0$. The charge can be found by multiplying n'(x) and p'(x) by q and by the volume $A(w_n - x)$ (on the n-side of the junction). A is the cross sectional area of the junction. The excess hole density is given by Equation 44

$$p'(x) = \frac{n_i^2}{N_d}\left( e^{\frac{qV_{ab}}{kT}} - 1 \right) \frac{\sinh\left( \frac{w_n - x}{L_h} \right)}{\sinh\left( \frac{w_n - x_n}{L_h} \right)} \tag{49}$$

From our quasi-neutrality assumption $n'(x) \approx p'(x)$. The positive and negative charges are then given by

$$Q(x) = qA(w_n - x)\frac{n_i^2}{N_d}\left( e^{\frac{qV_{ab}}{kT}} - 1 \right) \frac{\sinh\left( \frac{w_n - x}{L_h} \right)}{\sinh\left( \frac{w_n - x_n}{L_h} \right)} \tag{50}$$

If we write $V_{ab}$ as a sum of a constant bias voltage $V_{AB}$ plus a small variable voltage $v_{ab}$ : $V_{ab} = V_{AB} + v_{ab}$, then we can differentiate the expression for Q with respect to $v_{ab}$ to get an estimate of $C_{diffusion}$ . We set $x = x_n$ and find $Q = qA(w_n - x_n)\frac{n_i^2}{N_d}\left( e^{\frac{qV_{AB}}{kT}} e^{\frac{qv_{ab}}{kT}} - 1 \right)$ or,

when $v_{ab}$ is sufficiently small, $Q \approx qA(w_n - x_n)\frac{n_i^2}{N_d}\left( e^{\frac{qV_{AB}}{kT}}(1 + \frac{qV_{ab}}{kT}) - 1 \right)$ so we arrive at

$$C_{diffusion} = \frac{dQ}{dv_{ab}} = \frac{q^2A(w_n - x_n)}{kT}\frac{n_i^2}{N_d}\left( e^{\frac{qV_{AB}}{kT}} \right) \quad (51)$$

To get a feel for the size of this capacitance, suppose $w_n$ = 100 micrometers = $10^{-4}$ m, $x_n$ = 0, $N_d = 10^{17}$ $cm^{-3}$, $V_T = \frac{kT}{q} = 25$ mV, A = $10^{-8}$ $m^2$ (100 micrometers square) and $V_{AB}$ = 0.7 V; then $C_{diffusion} \sim 10^{-14}$ F, which is similar to the size of the depletion capacitance calculated earlier.

*X-D. Concept of The Bipolar Junction Transistor*

We can now begin to understand the operation of the BJT, a device that uses two p-n junctions to perform amplification of signals.

The usual notation shows all the currents flowing into the transistor, implying one or more of the currents is negative. We will depart from that and use a more intuitive notation in which, for a PNP transistor, positive current flows into the emitter and out of the base and collector. For the NPN transistor positive current will be shown flowing into the collector and base, and out of the emitter.

The conceptual layout of a BJT is shown in Figure 20, although an actual transistor has a planar structure like that shown in Figure 19.

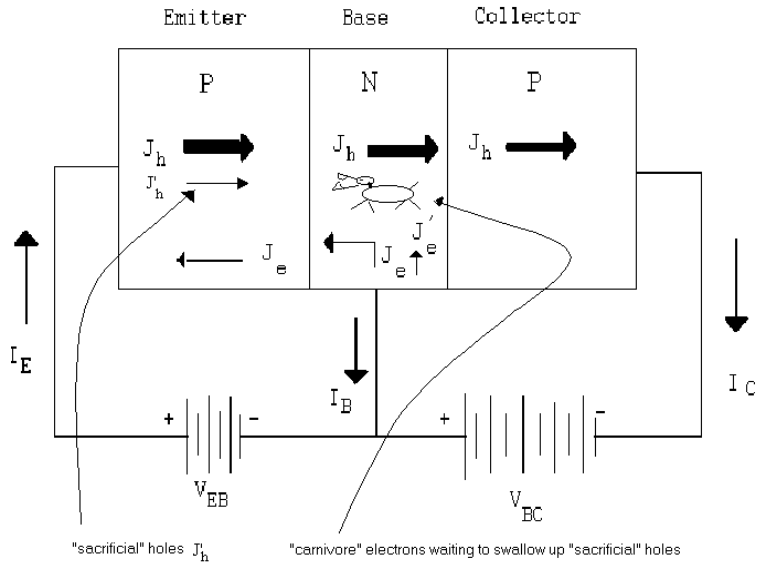



Figure 19. A somewhat whimsical schematic representation of a PNP BJT. The direction of positive current flow is indicated by the arrows, as well as the flow of the holes and electrons in the transistor. Note that there are two p-n junctions, the emitter - base (EB) junction and the base - collector (BC) junction, and that the EB junction is forward biased while the BC junction is reverse biased. Typically $V_{BC}$ is an order of magnitude larger than $V_{EB}$ (9 volts vs. 0.7 volts). A tiny fraction (~ 1%) of the holes ($J'_h$) injected from the emitter into the base are "sacrificed" when they recombine with carnivorous electrons ($J'_e$) in the base lying in wait for them; the rest ($J_h$) flow into the collector to become $I_C$. Most of the emitter current $I_E$ is carried by the emitter holes, a tiny fraction (~ 1%) is carried by electrons injected from the base.

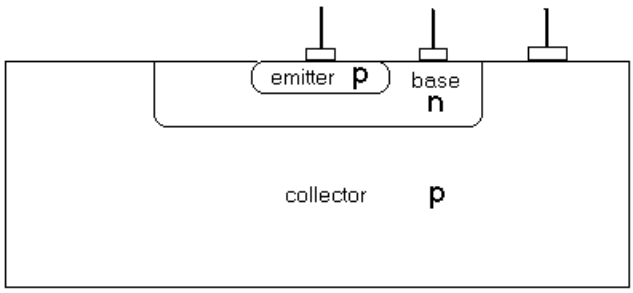



Figure 20. Schematic physical layout of an pnp transistor. Note the emitter is physically small compared to the collector. The thickness of the base region has been exaggerated.

The way the BJT works is that the emitter-base p-n junction is forward biased to about 0.7 volts. A small signal is added to the bias voltage and this has a great effect on the emitter current because of the exponential relationship between voltage and current. The emitter current consists mainly of holes flowing through and from the emitter, with a small current of electrons flowing from the base into the emitter. Most of the hole current from the emitter diffuses through the thin base region and winds up at the base-collector junction, which is reverse biased. Normally, very little current would flow through the base-collector circuit because of the reverse biasing. However, when a hole diffuses across the base from the emitter and ends up at the B-C junction, it sees a strong electric field pointing *into* the collector (note the polarity of the power supplies) and the hole is immediately swept into the collector. Essentially every hole from the emitter that does not recombine with an electron in the base finds its way into the collector. Typically, only about 1% or less of the holes are lost in the base due to recombination, so that $I_C$ is ~ 99% pf the emitter hole current. The base current consists of electrons that partially flow from the base into the emitter and partially recombine in the base with holes from the emitter. The hole current flowing into the collector returns to the emitter through the large power supply $V_{BC}$. A resistor can then be placed in this part of the circuit and a large power dissipated through it, and a substantial power gain can be had with the transistor.

*XI. More Detailed Operation of the BJT and the Ebers-Moll Model*

A widely used model for BJT operation is the Ebers-Moll model. If we examine Figure 20, reproduced again here, we can see and analyze the various current flows in the BJT.

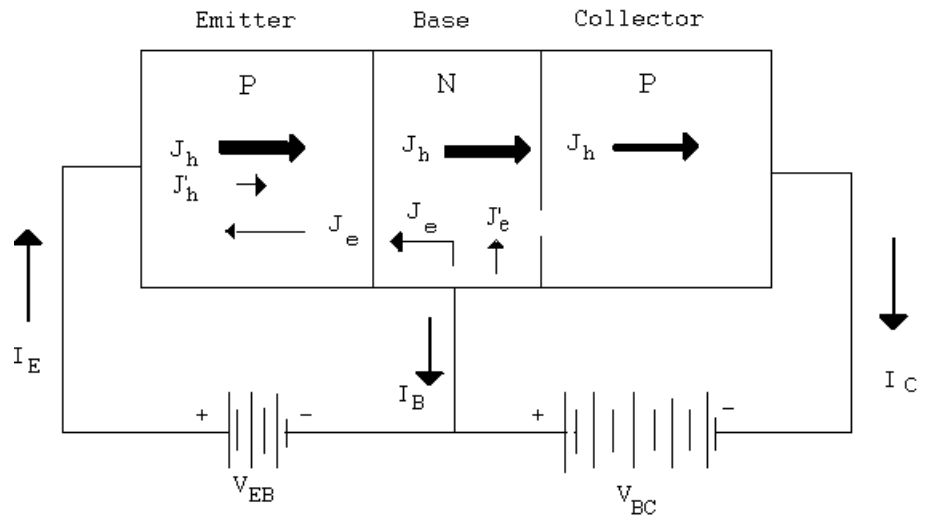



Figure 20 (repeated, in more conventional format). A PNP bipolar junction transistor showing the current carriers in the various parts of the transistor. $J'_e$ represents the base electron current density that recombines with holes from the emitter $J'_h$.

The current $I_E$ flowing into the emitter manifests itself in the emitter as a large current of holes and a small current of electrons, the latter coming from the base. This is because the doping of the emitter is much heavier than the doping of the base, resulting in the excess charge profiles shown schematically in Figure 21 below.

We now want to characterize the transistor by finding the currents flowing in the base, emitter and collector in terms of the voltages $V_{EB}$ and $V_{BC}$. An immediate problem in doing this would appear to be that the currents in the E-B junction and the B-C junction are very non-linear functions of the voltages, at least when the junctions are forward biased. This means there is a question as to the validity of the usual procedure of calculating the effect of each voltage separately and the superimposing the results, which works only if the systems are linear. However, if we keep each junction voltage constant, then the results can be superimposed because for this case, there will be no non-linearities. We proceed by setting $V_{BC} = 0$ and finding the currents. Then we set $V_{EB} = 0$ and do the same thing again.

*XI-A. The Forward Mode*

With $V_{EB} > 0$, the E-B junction is forward biased, and the density of holes at $x = x_n$ is $p'(x_n) = \frac{n_i^2}{N_{dB}}\left(e^{\frac{V_{EB}}{V_T}} - 1\right)$ while the density of electrons at $x = -x_p$ is $n'(-x_p) = \frac{n_i^2}{N_{aE}}\left(e^{\frac{V_{EB}}{V_T}} - 1\right)$, where $N_{dB}$ is the donor density in the base and $N_{aE}$ is the acceptor density in the emitter. The thermal voltage is $V_T = \frac{kT}{q}$, which is about 0.026 V at room temperature. The density of holes falls to zero at the metal contact to the emitter at $x = -w_E$. $p'(w_B)$ is also zero, for reasons we will discuss shortly. The resulting profiles of the excess charge densities is shown in Figure 21.

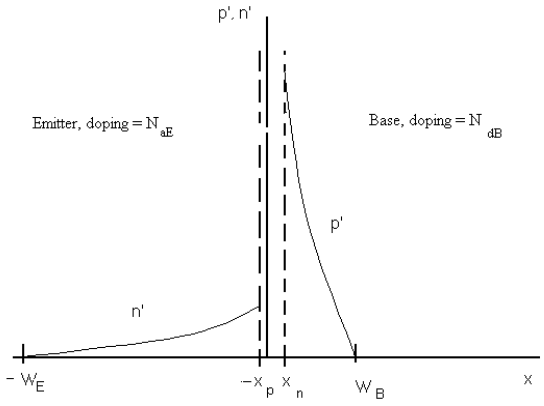



Figure 21. Schematic representation of the excess charge densities n' and p' in the emitter and the base. In an actual transistor $p'(x_n)$ is more like $100\ n'(-x_p)$.

The current density due to electrons in the emitter is much smaller than the current density due to holes in the base:

$$J_e^{diff}(x) = q\frac{D_e}{L_e}\frac{n_i^2}{N_{aE}}\left(e^{\frac{qV_{ab}}{kT}} - 1\right)\frac{\cosh\left(\frac{w_E - x}{L_e}\right)}{\sinh\left(\frac{w_E - x_p}{L_e}\right)}$$, whereas

$$J_h^{diff}(x) = q\frac{D_h}{L_h}\frac{n_i^2}{N_{dB}}\left(e^{\frac{qV_{ab}}{kT}} - 1\right)\frac{\cosh\left(\frac{w_B - x}{L_h}\right)}{\sinh\left(\frac{w_B - x_n}{L_h}\right)}$$. Since the emitter is much more heavily doped than the base ($N_{aE} >> N_{dB}$), $J_h^{diff} >> J_e^{diff}$. This is shown schematically in Figure 17 repeated below.

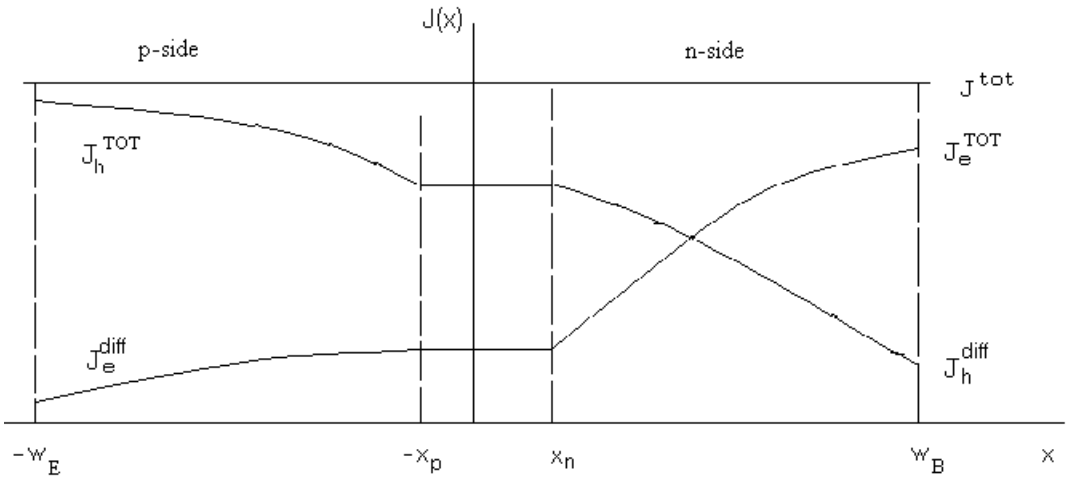



Figure 17 (repeated). The current density profiles through the emitter-base junction of the pnp transistor (emitter on the left, base on the right). The amount of electron current has been greatly exaggerated. Note that $J_e$ and $J_h$ remain constant across the depletion region.

The current in the emitter is denoted $I_{EF}$, which stands for forward emitter current. It consists of two parts, $I_{hF}$ and $I_{eF}$: $I_{EF} = I_{hF} + I_{eF}$, where $I_{hF}$ is the forward hole current and $I_{eF}$ is the forward electron current. Since it is the hole current that is of interest we write this as

$$I_{EF} = I_{hF} ( 1 + \delta_E ) \qquad (52)$$

$\delta_E$ is called the emitter defect: it is the fraction of the emitter current $I_{EF}$ carried by electrons from the base that doesn't eventually contribute to the collector current. $\delta_E$ is easily calculated. At the right side of the depletion zone in the base, $J_h^{diff}(x_n) = J_h^{TOT}(-x_p)$, the total hole current density at the left end of the depletion zone ($-x_p$), because we assume no recombination in the depletion region. The hole current injected into the base is therefore obtained from the diffusion current: $I_{hF} = qA\frac{n_i^2}{N_{dB}}\frac{D_h}{L_{hB}^*}\left(e^{\frac{V_{EB}}{V_T}} - 1\right)$, where $L_{hB}^* = L_h \tanh\left(\frac{w_B - x_n}{L_h}\right)$. Similarly, the electron current density at the right end of the depletion zone $J_e^{TOT}(x_n) = J_e^{diff}(-x_p)$, and so $I_{eF} = qA\frac{n_i^2}{N_{aE}}\frac{D_e}{L_{eE}^*}\left(e^{\frac{V_{EB}}{V_T}} - 1\right)$ with $L_{eE}^* = L_e \tanh\left(\frac{w_E - x_p}{L_e}\right)$. Then from the relation $I_{EF} = I_{hF} + I_{eF} = I_{hF}(1 + \delta_E)$, with a little algebra we find $\delta_E = \frac{D_e}{D_h}\frac{L_{hB}^*}{L_{eE}^*}\frac{N_{dB}}{N_{aE}}$. Since $N_{aE} >> N_{dB}$ and the other terms are of the same (close) order of magnitude, $\delta_E << 1$.

The next question is: what is the current in the collector? The collector current is the current that passes through the high voltage part of the power supply and represents the amplified power. Clearly, since the B-C junction is reverse biased no electron current will be injected from the base into the collector, nor will any hole current be injected from the collector into the base. However, because there is a large electric field with positive sense into the collector, any holes injected into the base that appear at the B-C junction will be immediately swept into the collector (this is the justification for setting $p'(w_B) = 0$). The collector current will therefore be $I_{hF}(x_n)$ less whatever holes are lost as they diffuse across the base towards the collector. We take into account this loss by expressing the forward collector current as

$$I_{CF} = I_{hF}(1 - \delta_B) \qquad (53)$$

where $\delta_B$ represents the fraction of the holes lost to recombination. $\delta_B$ Is called the base defect. Another way of writing $I_{CF}$ is $I_{CF} = I_{hF} - I_{RB}$, where $I_{RB}$ is the current due to recombination. From Equation 53 we see that $I_{RB} = I_{hF}\ \delta_B$, or $\delta_B = \dfrac{I_{RB}}{I_{hF}}$. The recombination current can be

calculated by finding the total charge Q injected into the base and dividing by the mean time for recombination, $\tau_R$. Q is simply the volume integral over the excess charge density qp'(x). The integral is taken from $x = x_n$ to $x = w_B$, but the result is much cleaner if we make the approximation $x_n = 0$:

$$Q = qA \int_0^{w_B} p'(x)\ dx$$

$$= qA \int_0^{w_B} \frac{n_i^2}{N_{dB}} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right) \frac{\sinh\left( \dfrac{w_B - x}{L_h} \right)}{\sinh\left( \dfrac{w_B}{L_h} \right)}\ dx \qquad (54)$$

where the differential volume element dV = A dx (A is the cross sectional area of the base, here assumed to be a constant), and we have set $x_n = 0$ in the lower limit of the integral and in the hyperbolic sine in the denominator of p'(x). Since $x < w_B$ and $w_B << L_h$, we can use $\sinh(u) \to u$ as $u \to 0$ to find Q:

$$Q = qA \frac{n_i^2}{N_{dB}} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right) \int_0^{w_B} \frac{(w_B - x)/L_h}{w_B/L_h}\ dx$$

$$= qA \frac{n_i^2}{N_{dB}} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right) \int_0^{w_B} \left( 1 - \frac{x}{w_B} \right) dx \qquad (55)$$

$$= qA \frac{n_i^2}{N_{dB}} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right) \frac{w_B}{2}$$

We obtain the recombination current by dividing Q by $\tau_R$:

$$I_{RB} = qA \frac{n_i^2}{N_{dB}} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right) \frac{w_B}{2\,\tau_R} \qquad (56)$$

The equation for $\delta_B$ can be simplified by recalling that $I_{hF} = qA \frac{n_i^2}{N_{dB}} \frac{D_h}{L_{hB}^*} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right)$. $L_{hB}^*$ can be reduced to $w_B$ using the fact that $w_B$ is large compared to $x_n$ and the series expansion $\tanh(u) = u - \frac{u^3}{3} + \frac{2u^5}{15} - + \ldots$: $L_h^* = L_{hB} \tanh\left( \frac{w_B - x_n}{L_h} \right) \rightarrow L_h \frac{w_B - x_n}{L_h} \approx w_B$.

Thus $I_{hF} = qA \frac{n_i^2}{N_{dB}} \frac{D_h}{w_B} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right)$. We then find

$$\delta_B = \frac{I_{RB}}{I_{hF}} = \frac{w_B^2}{2 D_h \tau_R} = \frac{w_B^2}{2 L_h^2} \qquad (57)$$

$\delta_B << 1$ since $w_B/L_h << 1$, so the forward collector current is very nearly equal to $I_{hF}$. We can easily relate the forward collector current to the total emitter current $I_{EF}$ as follows: $I_{EF} = I_{eF} + I_{hF} = I_{hF}\left(1 + \delta_E\right)$, and $I_{CF} = I_{hF}\left(1 - \delta_B\right)$. Therefore we have $I_{CF} = \frac{1 - \delta_B}{1 + \delta_E} I_{EF} = \alpha_F I_{EF}$, where the parameter $\alpha_F \equiv \frac{1 - \delta_B}{1 + \delta_E}$. Since $\delta_B$ and $\delta_E$ are both positive quantities $\alpha_F < 1$, although since the defects are both very small, $\alpha_F$ will be very close to unity (~ 0.99, typically).

It is very useful to find the relationship of $I_{CF}$ to the forward base current $I_{BF}$. This is easily done since the base current is just $I_{BF} = I_{eF} + I_{RB}$ and $I_{EF} = I_{eF} + I_{hF} = I_{hF}\left(1 + \delta_E\right)$, so, as we saw above, $I_{eF} = \delta_E I_{hF}$. Since $I_{RB} = \delta_B I_{hF}$, we have $I_{BF} = I_{hF}\left(\delta_E + \delta_B\right)$, which is very much less than $I_{hF}$. Then since $I_{CF} = I_{hF}\left(1 - \delta_B\right)$ we find the ratio of the collector current to the

base current is $\frac{I_{CF}}{I_{BF}} = \frac{1 - \delta_B}{\delta_B + \delta_E} \equiv \beta_F$. $\beta_F$ is called the forward current gain of the transistor, and is a very important parameter. Because $\delta_E$ and $\delta_B$ are both very small, $\beta_F$ is quite large. From the definitions of $\beta_F$ and $\alpha_F$ it is easy to show that $\alpha_F = \frac{\beta_F}{\beta_F + 1}$. For typical transistors $\beta_F$ ranges from 100 - 200.

*XI-B. The Reverse Mode*

In the forward active mode the emitter-base junction is forward biased and the base-collector junction is reverse biased. If the emitter-base junction is reverse biased and the base-collector junction is forward biased, the transistor is said to be in the reverse active or reverse mode. If the transistor were physically symmetric the results would be the same as before, with the roles of the emitter and the collector reversed. Because of the physical asymmetry, however, $\beta_R$ is usually substantially less than $\beta_F$, where $\beta_R$ is the ratio of the emitter current to the base current. The reverse behavior is calculated by setting $V_{EB} = 0$ and setting $V_{CB}$ positive, thus forward biasing the base-collector junction. The method of calculation is exactly the same as for the forward case except instead of an emitter defect $\delta_E$ we define a collector defect $\delta_C$. We then end up with $\beta_R = \frac{I_{ER}}{I_{BR}}$ and $\alpha_R = \frac{1 - \delta_B}{1 + \delta_C} = \frac{\beta_R}{\beta_R + 1}$ .

For the forward case we have $I_{EF} = I_{eF} + I_{hF}$ and $I_{CF} = \alpha_F I_{EF}$. For the reverse case we have $I_{CR} = I_{eR} + I_{hR}$ and $I_{EF} = \alpha_R I_{CF}$. From our expressions for $I_{eF}$ and $I_{hF}$ we have that the forward emitter current is $I_{EF} = qAn_i^2 \left( \frac{D_h}{N_{dB} L_{hB}^*} + \frac{D_e}{N_{aE} L_{eE}^*} \right) \left( e^{\frac{V_{EB}}{V_T}} - 1 \right) = I_{ES} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right)$,

where $I_{ES} = qAn_i^2 \left( \frac{D_h}{N_{dB} L_{hB}^*} + \frac{D_e}{N_{aE} L_{eE}^*} \right)$. By analogy we define $I_{CS}$ as

$I_{CS} = qAn_i^2 \left( \frac{D_h}{N_{dB} L_{hB'}^*} + \frac{D_e}{N_{aC} L_{eC}^*} \right)$, with $N_{aC}$ the collector doping. Then the reverse collector current - the analog of the forward emitter current (sort of a mirror image) - is

$I_{CR} = qAn_i^2 \left( \frac{D_h}{N_{dB} L_{hB'}^*} + \frac{D_e}{N_{aC} L_{eC}^*} \right) \left( e^{\frac{V_{CB}}{V_T}} - 1 \right) = I_{CS} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right)$. Now, however,

$L_{eC}^* = L_e \tanh \left( \frac{w_C - x_p'}{L_e} \right)$ is the length in the collector, with $x_p$' now being the boundary of the base-collector junction depletion region in the *collector*, and $L_{hB'}^* = L_h \tanh \left( \frac{w_B - x_n'}{L_h} \right)$

where $x_n$' is the boundary of the base-collector junction depletion region in the *base*.

The quantities $I_{ES}$ and $I_{CS}$ are called the *saturation* or sometimes, the scale currents. If we now add the expression for the forward and reverse currents together, we get the complete current-voltage relationships for the transistor. For the emitter current

$$I_E = I_{EF} + I_{ER} = I_{ES} \left( e^{\frac{V_{EB}}{V_T}} - 1 \right) - \alpha_R I_{CS} \left( e^{\frac{V_{CB}}{V_T}} - 1 \right) \qquad (58)$$

the minus sign in the second term coming from the fact that the current direction is reversed in the reverse active mode. For the collector current

$$I_C = I_{CF} + I_{CR} = \alpha_F I_{ES} \left( e^{V_{EB}/V_T} - 1 \right) - I_{CS} \left( e^{V_{CB}/V_T} - 1 \right) \qquad (59)$$

(remember that the notation here is that current flows into the emitter and out of the collector). Equations 58 and 59 are the Ebers-Moll model for the transistor. For this model the relationship $\alpha_F I_{ES} = \alpha_R I_{CS}$ holds.

*XI-C. The Forward Current Gain*

You can see why $\beta_F$ is called the current gain as follows. The forward collector current is $I_{CF} = \alpha_F I_{EF} = \alpha_F I_{ES} \left(e^{V_{EB}/V_T} - 1\right)$ and the forward base current is $I_{BF} = \frac{I_{CF}}{\beta_F}$. Suppose that the emitter-base voltage is changed by a small amount, $V_{EB} \rightarrow V_{EB} + \Delta V$. Now, in order to operate the transistor so that $\Delta V$ can change in either the positive or negative sense without turning the transistor off, $V_{EB}$ must be large compared to $\Delta V$ as shown in Figure 22.

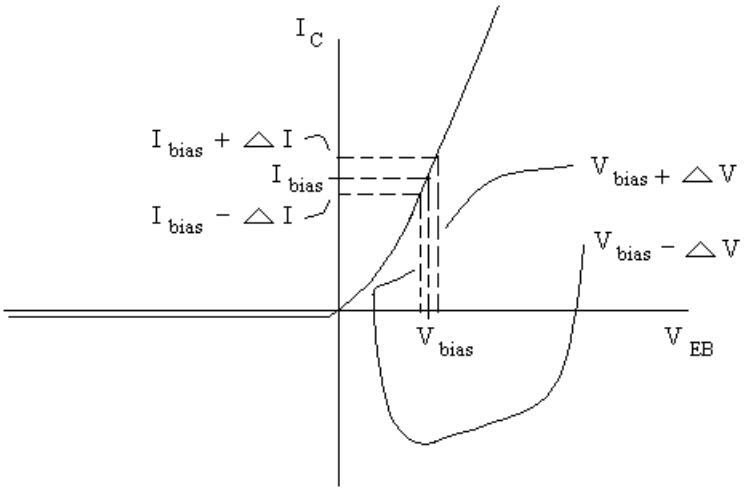



Figure 22. Schematic diagram showing how the transistor is biased. The signal voltage is $\Delta V$, which is small compared to $V_{bias}$. $V_{bias}$ ensures that the transistor is always in the forward mode with collector current $I_{Co} = I_{bias}$. The signal current varies by $\pm \Delta I$ about $I_{Co}$ as the signal voltage varies by $\pm \Delta V$ about $V_{bias}$.

Typically, $V_{EB}$ is set to a constant value of about 0.7 volts, and is called the *bias voltage*. Since $V_T$ is about 1/40 volt the factor $e^{V_{EB}/V_T} \approx 1.5 \times 10^{12}$. Therefore we can say

$e^{V_{EB}/V_T} - 1 = e^{V_{EB}/V_T}$, to an excellent approximation. If $V_{EB}$ changes slightly from $V_{EB}$ to $V_{EB} + \Delta V$ the emitter current changes by an amount

$$\begin{aligned} \Delta I_{EF} &= I_{ES}\left(e^{\frac{V_{EB}+\Delta V}{V_T}} - e^{\frac{V_{EB}}{V_T}}\right) \\ &\approx I_{ES}\, e^{\frac{V_{EB}}{V_T}}\left(1 + \frac{\Delta V}{V_T} - 1\right) \\ &= I_{ES}\, e^{\frac{V_{EB}}{V_T}}\,\frac{\Delta V}{V_T} \end{aligned} \tag{60}$$

Then the collector current will change by an amount

$$\begin{aligned} \Delta I_{CF} &= \alpha_F\, \Delta I_{EF} \\ &= \alpha_F\, I_{ES}\, e^{\frac{V_{EB}}{V_T}}\,\frac{\Delta V}{V_T} \\ &= I_{CS}\, e^{\frac{V_{EB}}{V_T}}\,\frac{\Delta V}{V_T} \end{aligned} \tag{61}$$

Similarly, the base current changes by an amount

$$\begin{aligned} \Delta I_{BF} &= \frac{I_{CS}}{\beta_F}\left(e^{\frac{V_{EB}+\Delta V}{V_T}} - e^{\frac{V_{EB}}{V_T}}\right) \\ &\approx \frac{I_{CS}}{\beta_F}\, e^{\frac{V_{EB}}{V_T}}\left(1 + \frac{\Delta V}{V_T} - 1\right) \\ &= \frac{I_{CS}\, e^{\frac{V_{EB}}{V_T}}}{\beta_F}\,\frac{\Delta V}{V_T} \\ &= \frac{\Delta\, I_{CF}}{\beta_F} \end{aligned} \tag{62}$$

In other words, the change in the collector current due to the change in base current is

$$\Delta I_{CF} = \beta_F \Delta I_{BF} \quad \text{or} \quad \frac{dI_{CF}}{dI_{BF}} = \beta_F \tag{63}$$

A useful way to write the relation between collector current and input voltage is to define $I_{Co}$ as the forward collector current due to the bias voltage $V_{EB} = V_{bias}$: $I_{Co} = I_{CS}\, e^{V_{bias}/V_T}$ ,and to define $g_m = \frac{I_{Co}}{V_T}$. From Equation 61 we then have

$$\Delta I_{CF} = g_m \Delta V \tag{64}$$

which directly relates the change in collector current to the change in input voltage. $g_m$ is called the *transconductance*; it has units of amperes per volt (ohms)$^{-1}$, or Siemens (after the large German manufacturing company, or perhaps its founder). As you can probably guess, when analyzing a transistor circuit one of the first steps will be to relate the current $I_{Co}$ to the bias voltage so the transconductance can be found.

*XI - D. The Diffusion Capacitance*

Finally, we look at one other interesting phenomenon. Previously (section X-A) we saw that the stored charge in the depletion region leads to a capacitance across the p-n junction called the depletion capacitance. There is an additional source of capacitance that shows up only when current is flowing, due to the positive and negative excess charges in the region beyond the depletion region. This is called the diffusion capacitance, and it is non-zero only when the voltage across the junction $V_{ab}$ is non-zero. An estimate of the size of this capacitance can be made by calculating the charge stored in the junction material. The positive charge stored in a conducting p-n junction where $L_h >> w_n - x_n$ is given by the following expression

$$Q(V_{ab}) = qA \int_{x_n}^{w_n} p'(x)\, dx \approx qA \frac{n_i^2}{N_d} \left(e^{V_{ab}/V_T} - 1\right) \frac{w_n - x_n}{2} \tag{65}$$

where the approximation sinh(u) = u is used, and replacing the integral by a triangular approximation to the area under p'(x), since p'(x) is approximately linear ($p'(x) \sim 1 - \frac{x}{w_n}$).

We define $C_{diff} = \frac{d\,Q\,(V_{ab})}{d\,V_{ab}}$ and find (ignoring the 1 in the expression $\left(e^{\frac{V_{ab}}{V_T}} - 1\right)$)

$$C_{diff} \approx q^2 A \frac{n_i^2}{N_d} \frac{e^{V_{ab}/V_T}}{kT} \frac{w_n - x_n}{2} \qquad (66)$$

where the $V_T$ in the denominator has been replaced by $\frac{kT}{q}$.

*XI-E An Interesting Result*

The forward collector current for a PNP transistor is $I_{CF} = I_{hF}\left(1 - \delta_B\right)$, where $I_{hF}$ is the forward hole current injected into the base by the emitter and $\delta_B$ is the base defect that represents the fraction of the hole current lost due to recombination. $\delta_B = \frac{W_B^2}{2L_h}$ where $W_B$ is the width of the base. The width of the base is the region in which recombination can take place. When a hole diffuses from the right side of the depletion region of the emitter-base junction at $x = x_n$ to the left side of the depletion region of the base-collector junction at $x_n'$ , it is immediately swept into the collector by the field at that junction. Therefore the effective width of the base is the distance from $x_n$ to $x_n'$ (see Figure 22). Since the (forward) bias of the emitter base junction remains essentially constant, $x_n$ is constant. But when the base-collector or emitter-collector voltage is increased, the width of the reverse biased base-collector junction depletion region increases. In fact, $x_n' = \sqrt{\frac{2\epsilon_{Si}}{q}\left(\phi_m + V_{BC}\right)\frac{N_{aC}}{N_{aC} + N_{nB}}}$, where $V_{BC} = V_{EC} - V_{EB}$ is the value of the *reverse* bias (that accounts for the + sign before $V_{BC}$ in the radical). This mean that as $V_{EC}$ increases the

effective width of the base decreases and so $I_{CF}$ increases as well. Since the factor $\delta_B$ contains $W_B^2$, $\delta_B$ decreases linearly with $V_{EC}$. This can be expressed concisely as

$I_C = I_{CO}\left(1 + \frac{V_{EC}}{V_A}\right)$. $I_{CO}$ is the collector current established by the forward bias voltage of the emitter-base junction $V_{EB}$. $V_A$ is a constant characteristic of the transistor and which depends on the parameters in $x_n$'; it is called the Early voltage, after Early, an early pioneer in transistor behavior. Since $I_C$ vanishes when $V_{EC} = V_A$, a graph of $I_C$ vs. $V_{EC}$ for various values of $V_{EB}$ will have the tangents to all the curves coincide at $V_{EC} = -V_A$ (see Figure 23).

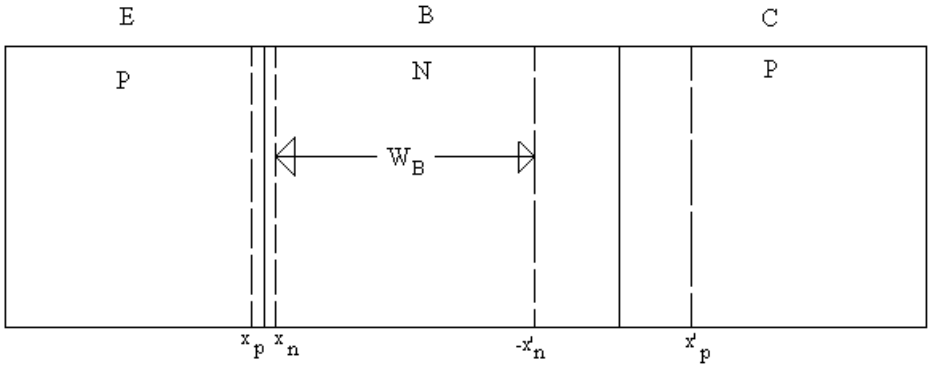



Figure 22. Schematic diagram showing the effective base width $W_B$ in terms of the distance between the right edge of the (essentially fixed) emitter-base depletion region at $x_n$ and the left edge of the (variable width) base-collector depletion region at $-x_n$'. Distances are not to scale.

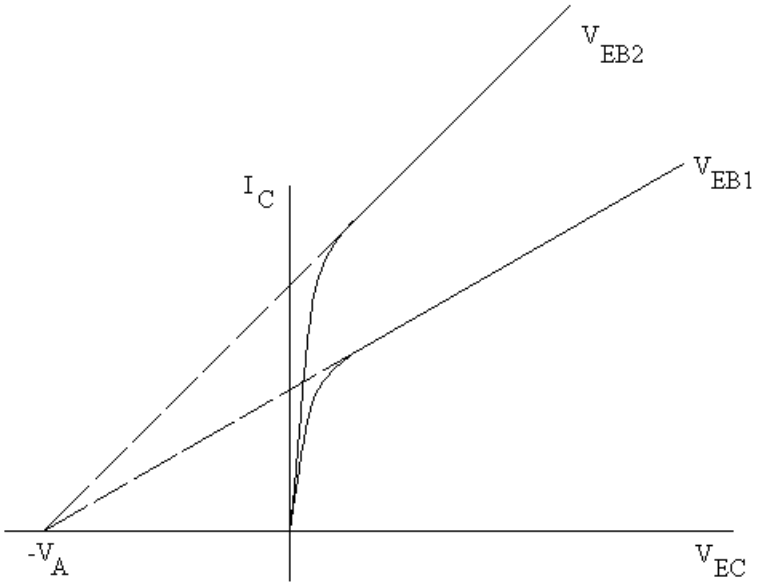



Figure 23. Graphical demonstration of the Early effect: $I_C = I_{CO}\left(1 + V_{EC}/V_A\right)$. The two values of the emitter-base bias voltage $V_{EB1} < V_{EB2}$ establish two values of collector current which are then modified by an increase in the emitter-collector voltage $V_{EC}$.