Book 2

Basic Semiconductor Devices for Electrical Engineers

Professor C.R. Viswanathan

Electrical and Computer Engineering Department

University of California at Los Angeles

Distinguished Professor Emeritus

Chapter 1

Introductory Solid State Physics

Introduction

An understanding of concepts in semiconductor physics and devices requires an elementary familiarity with principles and applications of quantum mechanics. Up to the end of nineteenth century all the investigations in physics were conducted using Newton's Laws of motion and this branch of physics was called classical physics. The physicists at that time held the opinion that all physical phenomena can be explained using classical physics. However, as more and more sophisticated experimental techniques were developed and experiments on atomic size particles were studied, interesting and unexpected results which could not be interpreted using classical physics were observed. Physicists were looking for new physical theories to explain the observed experimental results.

To be specific, classical physics was not able to explain the observations in the following instances:

1) Inability to explain the model of the atom.
2) Inability to explain why the light emitted by atoms in an electric discharge tube contains sharp spectral lines characteristic of each element.
3) Inability to provide a theory for the observed properties in thermal radiation i.e., heat energy radiated by a hot body.
4) Inability to explain the experimental results obtained in photoelectric emission of electrons from solids.

Early physicists Planck (thermal radiation), Einstein (photoelectric emission), Bohr (model of the atom) and few others made some hypothetical and bold assumptions to make their models predict the experimental results. There was no theoretical basis on which all their assumptions could be justified and unified. In the 1920s, a revolutionary and amazing observation was made by De Broglie that particles also behave like waves. Einstein in 1905 had formulated that light energy behaves like particles called photons to explain the photoelectric results. De Broglie argued that energy and matter are two fundamental entities. The results of the photoelectric experiments demonstrated that light energy behaves also like particles and therefore matter also should exhibit wave properties. He predicted that a particle with a momentum of magnitude $\boldsymbol{p}$ will behave like a plane wave of wavelength, $\boldsymbol{\lambda}$ equal to

$$\lambda = \frac{h}{p} \tag{1.1}$$

where $\boldsymbol{h}$ is Planck's constant and numerically equal to

$$h = 6.625\ x\ 10^{-34} \text{ Joules - sec.}$$

This led to the conclusion that both particles and energy (light) exhibited dual properties of behaving as particles and as waves.

Since Planck's constant h is extremely small, the De Broglie wavelength is correspondingly small for particles of large mass and momentum and therefore, the wave properties are not noticeable in our daily life. However, subatomic particles such as electrons have such a small mass and their momentum is small to make the wavelength large enough to observe interference and diffraction effects in the laboratory.

A particle moving with a momentum $\vec{\boldsymbol{p}}$ is represented by a plane wave of amplitude $\Psi\ (\vec{r}, t)$ given by:

$$\Psi\ (\vec{r}, t) = A\ e^{i(\vec{k}\cdot\vec{r} - \omega\ t)} = A\ e^{i\ (\frac{\vec{p}\cdot\vec{r}}{\hbar} - \omega\ t\ )} \tag{1.2}$$

where $\vec{\boldsymbol{k}}$ is the propagation vector of magnitude k equal to $\frac{2\ \pi}{\lambda}$, $\hbar = \frac{h}{2\ \pi}$, $\omega = 2\ \pi$ times the frequency of the wave and $\vec{r}$, the radius vector is equal to

$$\vec{r} = \vec{a}_x x + \vec{a}_y\ y + \vec{a}_z\ z$$

and $\vec{\boldsymbol{a}}_{\boldsymbol{x}}, \vec{\boldsymbol{a}}_{\boldsymbol{y}}$ and $\vec{\boldsymbol{a}}_{\boldsymbol{z}}$ are the unit vectors along the three coordinate axes. The radius vector denotes the position where the amplitude is evaluated. The wavelength of the plane wave according to equation (1.1), is given by

$$\lambda = \frac{h}{|p|} = \frac{2\ \pi}{|k|} \tag{1.3}$$

$\Psi\ (\vec{r}, t)$ is also called the ***wave-function*** or ***state function***. The expression for the plane wave given in equation (1.2) shows that the amplitude of the plane wave varies sinusoidally with time $\boldsymbol{t}$ and position $\boldsymbol{x}$. For illustrating the plane wave properties we will consider a plane wave travelling from negative infinity to positive infinity. Figure (1.1a) shows the amplitude of the plane wave at as a sinusoidal function of time $\boldsymbol{t}$ at some value $\boldsymbol{x = x_1}$. Figure (1.1b) shows the amplitude varying as a sinusoidal function of position $\boldsymbol{x}$ at some time $\boldsymbol{t = t_1}$.

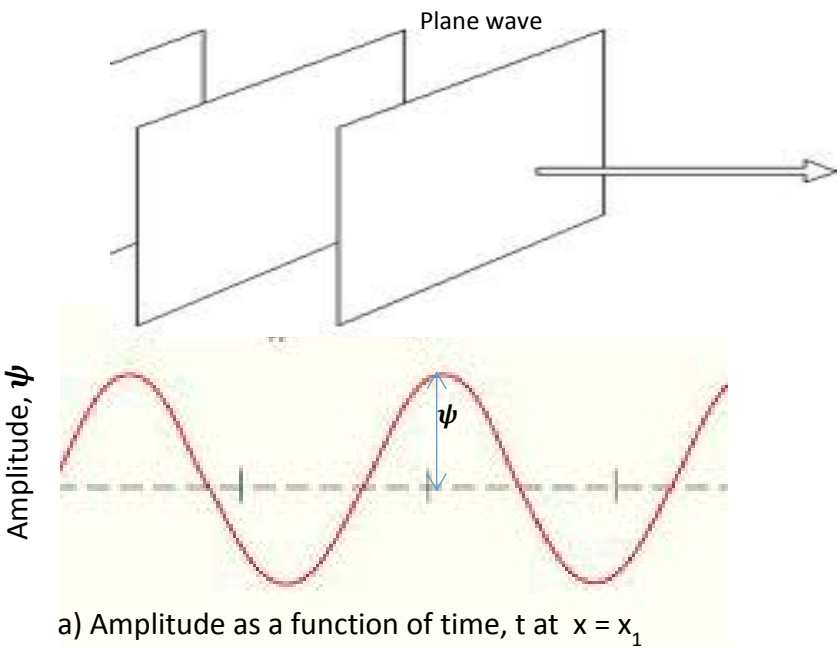


Figure (1.1): (a) Plane wave travelling from negative infinity to positive infinity, a) the amplitude of the plane wave as a function of time, t at some value $x = x_1$, (b) the amplitude of the plane wave as a function of x at some time $t = t_1$

The dynamic behavior of a physical system such as a single or a collection of particles can be deduced from Ψ, the wave-function of the system. In quantum mechanics, the wave-function is obtained by solving an equation called ***Schrödinger equation***. Another consequence of applying quantum mechanical principles to electrons is that the momentum and position of the particle say an electron, cannot be measured or known accurately at the same time to any precision that the measuring equipment is capable of. Either the momentum or the position alone can be measured to any accuracy the measuring equipment is capable of; but both of them cannot be measured at the same time to the accuracy or precision that each property can be measured separately. This is called the ***Heisenberg Uncertainly Principle.*** That is, the more precisely position or momentum is measured, the other can be measured only less precisely. The physics underlying the uncertainty principle is that the act of

measurement of one property disturbs the physical state of the system making measurement of the other property less accurate. In quantum mechanics, there are other examples of pairs of physical properties that obey the Heisenberg Uncertainty Principle. For example, ***energy and time*** is another pair that obeys the Uncertainty Principle.

Mathematically the uncertainty principle says that the product of the uncertainty in simultaneous measurement of position and momentum of a particle has to be higher than a minimum value of the order of **Planck's constant** $\boldsymbol{h}$, $\Delta P_x \, \Delta x \geq h$. The uncertainly principle is not observable in the size and mass of particles in our daily life where the uncertainties of the properties are too small to observe because of the small value of Planck's constant. When we deal with atomic particles such as electrons the uncertainty effects become significant.

When the wave-function is determined by solving ***Schrödinger equation*** for an electron that is confined to a narrow region of space along the x-axis, say between x = 0 and $x = L_x$, the result shows that the x - component of momentum can have only discrete values given by

$$P_x = n_x \frac{h}{L_x}$$

where $\boldsymbol{n_x}$ is called a ***quantum number***, and is an integer $(n = \pm 1, \pm 2, \pm 3, .. etc.)$. Thus the x - component of momentum, $\boldsymbol{P_x}$, can have only discrete values and are said to be discretized or quantized and hence $\boldsymbol{n_x}$ is called a *quantum number*. If the electron is constrained to move in a restricted region in the x - y plane, then 2 quantum numbers $\boldsymbol{n_x}$ and $\boldsymbol{n_y}$ will be needed to quantize the x and y components of the momentum. In the case when the electron is confined to a small three dimensional region, the electron state will be specified by three quantum numbers $\boldsymbol{n_x}$ and $\boldsymbol{n_y}$ and $\boldsymbol{n_z}$, as discussed in the next section.

Free Electron Model

The electric current in a conducting solid such as a metal is explained in solid state physics using the Free Electron Model. According to this model, each atom of the solid contributes one (or two in some cases) electron to form a sea of electrons in the solid and these electrons are free to roam around in the solid instead of being attached to the parent atom. If an electric field is applied in the solid, the electrons will move due to the force of the electric field and each electron will carry a charge $-q$ coulomb thus giving rise to the flow of an electric current. Thus free electrons in the solid are treated quantum mechanically as equivalent to electrons in a box of the same dimensions as the solid. We will now derive some properties of the free electrons in a conducting solid by using ***free electron model***.

Let us now consider an electron of mass m that is confined to move within a three dimensional box. Let the box be rectangular with dimensions L_x, L_y and L_z. As shown in Figure (1.2), we will choose the origin of our Cartesian coordinate system at one corner and the x, y and z axes to be along the three edges of the box.

We assume that the potential energy to be constant and that it does not vary with position inside the box. Since force is equal to the gradient of the potential energy, there is no force acting on the

electron because the gradient is zero. The electron is said to be in a *force free region*. Let the potential energy be taken as *zero*. A particle moving in force-free region has a constant momentum and is therefore represented by a plane wave as discussed earlier. The wave-function for the particle is therefore given by

$$\Psi\ (\vec{r},t) = A\ e^{i(\vec{k}\cdot\vec{r}-\omega\ t)} = A\ e^{i\ (\vec{k}\ \cdot\vec{r}\ )}e^{-i\omega t} =\ \psi(\vec{r})e^{-i\omega t} \tag{1.4}$$

where

$$\psi(\vec{r}) = A\ e^{i\ (\vec{k}\ \cdot\vec{r}\ )}$$

$\boldsymbol{\psi}(\vec{\boldsymbol{r}})$ can be written as, in rectangular coordinates,

$$\psi(\vec{r}) = \psi\ (x,y,z) = A_x\ e^{ik_x x}A_y\ e^{ik_y y}A_z\ e^{ik_z z}$$

where

$$A =\ A_x A_y A_z$$

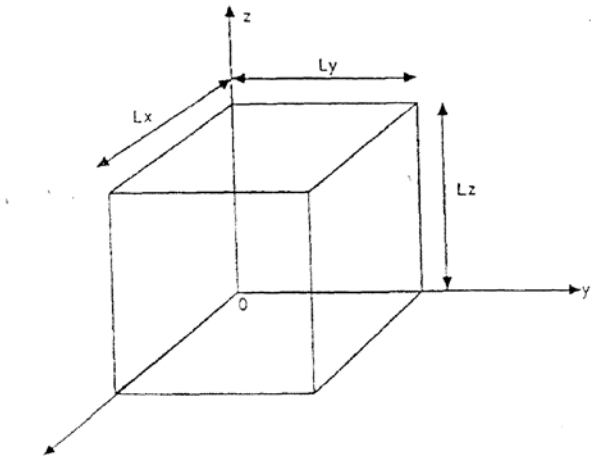


Figure (1.2): Electron in a box

We can also write

$$\psi\ (x,y,z) = \psi_x(x)\psi_y(y)\psi_z(z) \tag{1.5}$$

Where

$$\psi_x\ (x) = A_x\ e^{ik_x x} \tag{1.6}$$

$$\psi_y\ (y) = A_y\ e^{ik_y y} \tag{1.7}$$

$$\psi_z\ (z) = A_z\ e^{ik_z z} \tag{1.8}$$

The boundary condition on the wave-function is that the wave-function is the same in amplitude and phase when x is incremented by L_x, y is incremented by L_y and z is incremented by L_z i.e.

$$\psi\,(x,y,z) = \psi\,(x+L_x,y,z\,) = \psi\left(x,y+L_y,z\,\right) = \psi\,(x,y,z+L_z) \tag{1.9}$$

This boundary condition is called the ***Periodic Boundary Condition***.

The above boundary condition requires

$$\psi_x\,(x) = \psi_x\,(x+L_x\,) \tag{1.10}$$

$$\psi_y\,(y) = \psi_y\left(y+L_y\,\right) \tag{1.11}$$

and

$$\psi_z\,(z) = \psi_z\,(z+L_z\,) \tag{1.12}$$

By applying the boundary condition given in Equation (1.6) to Equation (1.10), we obtain

$$A_x\,e^{ik_xx} = A_x\,e^{ik_x(x+L_x)} \tag{1.13}$$

This implies that

$$k_xL_x = 2\,\pi\,n_x \tag{1.14}$$

where n_x is a positive or negative integer. Therefore,

$$k_x = \frac{2\,\pi}{L_x}\,n_x \tag{1.15}$$

and

$$\psi_x\,(x) = A_x\,e^{i\frac{2\,\pi\,n_x}{L_x}x} \tag{1.16}$$

The x-component of the momentum, equal to $\hbar\boldsymbol{k_x}$, is therefore quantized with values,

$$p_x = 0, \pm\frac{h}{L_x}, \pm\frac{2h}{L_x}, \pm\frac{3h}{L_x}, \pm\frac{4h}{L_x}\,\cdots etc. \tag{1.17}$$

Thus the x-component of the momentum is quantized with the same quantum number $\boldsymbol{n_x}$.

The kinetic energy is related to the momentum. The kinetic energy due to the motion of the particle along the x-axis is also quantized. Let us denote the kinetic energy due to motion along the x-axis as E_1.

$$E_1 = \frac{p_x^2}{2\,m} = \frac{\hbar^2k_x^2}{2m} = \frac{\hbar^2n_x^2}{2\,m\,L_x^2} \tag{1.18}$$

Thus $\boldsymbol{p_x}$, $\boldsymbol{k_x}$ and $\boldsymbol{E_1}$ are all quantized with the same quantum number $\boldsymbol{n_x}$.

Example

Let us now calculate k_x , p_x and E_1 for an electron in a state $n_x = 1$ in a box of dimension 10 Å × 20 Å × 30 Å along the x, y and z axes respectively. Remembering 1 Å is $10^{-10}m$,

$$k_x = \frac{2\,\pi}{L_x}\; n_x = \frac{2\,\pi \times 1}{10 \times 10^{-10}} = \frac{2\,\pi}{10^{-9}} = 6.28 \times 10^{9}\, m^{-1}$$

$$p_x = \hbar k_x = \frac{6.626\; x\; 10^{-34}}{2\,\pi} \times 6.628 \times 10^{9} = 6.26 \times 10^{-25}\, K\, gm - {}^{m}/_{S}$$

$$E_1 = \frac{p_x^2}{2\,m} = \frac{p_x^2}{2\; x\; 9.11 \times 10^{-31}} = 2.41 \times 10^{-19}\, Joules = 1.50\, eV$$

$\boldsymbol{k_x}$, $\boldsymbol{k_y}$ and $\boldsymbol{k_z}$ are related to x, y and z components of momentum through the De Broglie relation. By using the boundary condition for k_y and k_z , it can be shown $k_y = \frac{2\,\pi\, n_y}{L_y}$ and $k_z = \frac{2\,\pi\, n_z}{L_z}$ where $\boldsymbol{n_y}$ and $\boldsymbol{n_z}$ are quantum numbers similar to $\boldsymbol{n_x}$.

The kinetic energy due to motion along the y and z axes are denoted by $\boldsymbol{E_2}$ and $\boldsymbol{E_3}$

$$E_2 = \frac{\hbar^2 k_y^2}{2m} = \frac{\hbar^2 n_y^2}{2mL_y^2} \tag{1.19}$$

and

$$E_3 = \frac{\hbar^2 k_z^2}{2m} = \frac{\hbar^2 n_z^2}{2mL_z^2} \tag{1.20}$$

In Figure (1.3), $\boldsymbol{E_1}$ is plotted as a function of $\boldsymbol{k_x}$. Since $\boldsymbol{k_x}$ is quantized, $\boldsymbol{E_1}$ is also quantized. Since the adjacent values of $\boldsymbol{k_x}\ and\ \boldsymbol{E_1}$ lie so close to each other, the plot looks like a continuous curve. We say $\boldsymbol{k_x}\ and\ \boldsymbol{E}_1$ are ***qusai-continuous***. The total wave-function is obtained as a product function of $\boldsymbol{\psi_x}$, $\boldsymbol{\psi_y}\ and\ \boldsymbol{\psi_z}$ and is given by

$$\psi\,(x, y, z) = \psi_x\,(x)\,\psi_y\,(y)\psi_z\,(z) \tag{1.21}$$

$$= A_x\,A_y\,A_z\,e^{i(k_x x + k_y y + k_z z)} \tag{1.22}$$

$$= A_x\,A_y\,A_z\,e^{i\vec{k}\,\cdot\vec{r}} \tag{1.23}$$

where $\vec{\boldsymbol{k}}$ is called the propagation vector and is given by

$$\vec{k} = \vec{a}_x k_x + \vec{a}_y k_y + \vec{a}_z k_z \tag{1.24}$$

$$= \vec{a}_x \frac{2\,\pi\, n_x}{L_x} + \vec{a}_y \frac{2\,\pi\, n_y}{L_y} + \vec{a}_z \frac{2\,\pi\, n_z}{L_z} \tag{1.25}$$

where $\vec{r}$ = radius vector, and $\vec{a}_x$, $\vec{a}_y$ and $\vec{a}_z$ are the unit vectors along the x, y and z axes respectively. E', the total kinetic energy is equal to

$$E' = E_1 + E_2 + E_3 = \frac{\hbar^2 k^2}{2m} = \frac{\hbar^2}{2m}\left[\frac{n_x^2}{L_x^2} + \frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}\right] \quad (1.26)$$

where k is the magnitude of the propagation vector $\vec{k}$. The number n_x, n_y and n_z are called ***quantum numbers***, and once you assign a particular set of three integers to these three quantum numbers, you have specified the momentum states for the particle i.e., the value of momentum the particle will have in that state. We therefore denote the wave-function by the subscripts n_x, n_y and n_z.

$$\psi_{n_x n_y n_z} = A_x\, A_y\, A_z e^{i\,2\,\pi\,(\frac{n_x x}{L_x}+\frac{n_y y}{L_y}+\frac{n_z z}{L_z})} \quad (1.27)$$

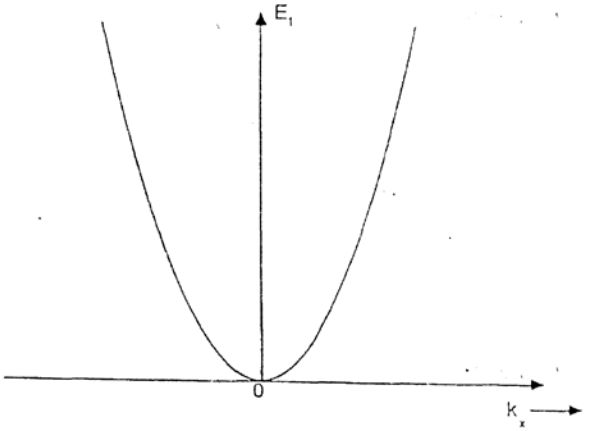


Figure (1.3): Energy E_1 vs. k_x

The state $n_x = 1$, $n_y = 2$ and $n_z = -1$, represents a state in which the particle is moving with a momentum $\vec{p}$ equal to

$$\vec{p} = 2\pi\hbar\left[\frac{\vec{a}_x}{L_x} + \frac{2\vec{a}_y}{L_y} - \frac{\vec{a}_z}{L_z}\right] = h\left[\frac{\vec{a}_x}{L_x} + \frac{2\vec{a}_y}{L_y} - \frac{\vec{a}_z}{L_z}\right] \quad (1.28)$$

Since n_x, n_y and n_z determine the momentum state, and can be only integers, we see that the p_x value changes by $\frac{h}{L_x}$ from one state to the next and similarly the p_y and p_z values change by $\frac{h}{L_y}$ and $\frac{h}{L_z}$ respectively.

It should not be surprising that adjacent values of p_x, differ by $\frac{h}{L_x}$. The maximum uncertainty in the x-component of position is $\Delta x = L_x$ since the particle can be anywhere in the box. Therefore the uncertainty in the x-component of momentum is required from the *Uncertainty Principle* to satisfy the condition

$$\Delta p_x \geq \frac{h}{L_x}$$

The adjacent values differing by $\frac{\boldsymbol{h}}{\boldsymbol{L_x}}$ are consistent with the Uncertainty Principle.

Consider a three dimensional space (imaginary of course!) in which the three axes are $\boldsymbol{p_x}, \boldsymbol{p_y}$ and $\boldsymbol{p_z}$ as shown in Figure (1.4). Each set of integers for $\boldsymbol{n_x}, \boldsymbol{n_y}$ and $\boldsymbol{n_z}$ generates a point in this space (called the momentum space) and each point represents a particular momentum state. Such points in momentum space are called representative points or **phase points**.

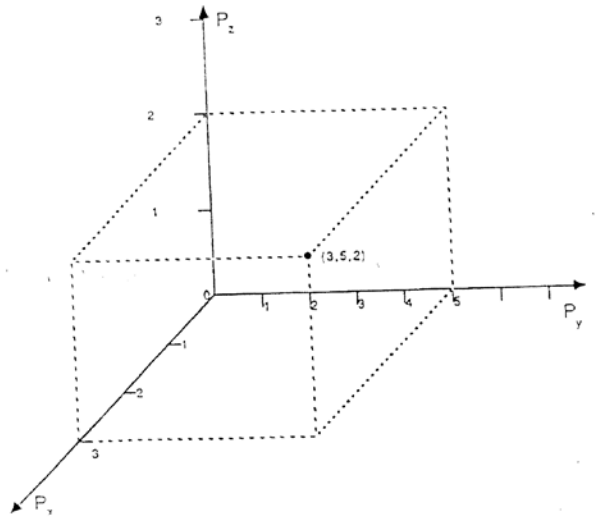


Figure (1.4): Momentum space with p_x, p_y and p_z axes

The momentum space comprises an infinite number of phase points each separated from its neighbor by $\frac{\boldsymbol{h}}{\boldsymbol{L_x}}$ along the $\boldsymbol{p_x}$ axis or $\frac{\boldsymbol{h}}{\boldsymbol{L_y}}$ along the $\boldsymbol{p_y}$ axis or $\frac{\boldsymbol{h}}{\boldsymbol{L_z}}$ along the $\boldsymbol{p_z}$ axis. The origin of this space is located at $\overline{\boldsymbol{p}} = \boldsymbol{0}$, and hence represents the state with no kinetic energy. The vector connecting the origin of the momentum space to a point with coordinate numbers $\boldsymbol{n_x}, \boldsymbol{n_y}$ and $\boldsymbol{n_z}$ represents the momentum vector of the state with x-component $\boldsymbol{p_x} = \frac{\boldsymbol{h}}{\boldsymbol{L_x}}\boldsymbol{n_x}$, y-component $\boldsymbol{p_y} = \frac{\boldsymbol{h}}{\boldsymbol{L_y}}\boldsymbol{n_y}$, and z -component $\boldsymbol{p_z} = \frac{\boldsymbol{h}}{\boldsymbol{L_z}}\boldsymbol{n_z}$.

COORDINATE NUMBERS	
A	n_x, n_y, n_z
B	$n_x + 1, n_y, n_z$
C	$n_x + 1, n_y + 1, n_z$
D	$n_x, n_y + 1, n_z$
E	$n_x, n_y + 1, n_z + 1$
F	$n_x + 1, n_y + 1, n_z + 1$
G	$n_x, n_y, n_z + 1$
H	$n_x + 1, n_y, n_z + 1$

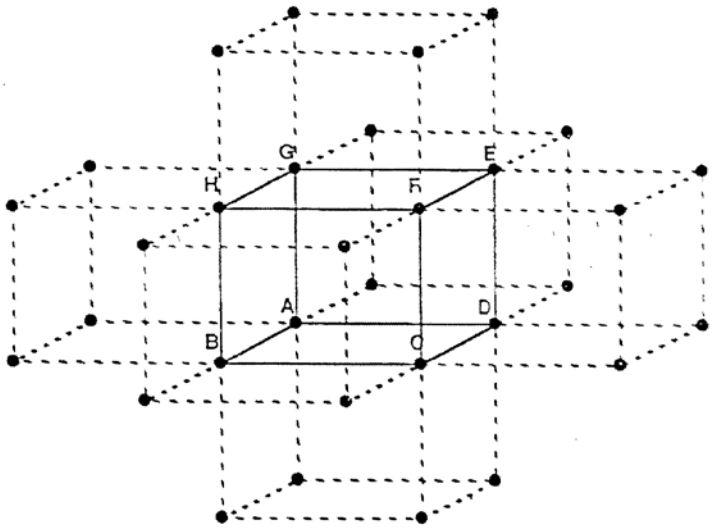


Figure (1.5): ABCDEFGH represents the elementary volume in the momentum space containing one momentum state.

We obtained three quantum numbers because we restricted the motion of the particle to be within the small box of volume $\boldsymbol{L_x L_y L_z}$. If we constrain the particle to only a small region in one dimension, say along the x axis, we get one quantum number $\boldsymbol{n_x}$ and the momentum along the x axis is quantized. Similarly, if we restrict the particle to a small area in the xy plane, we will get two quantum numbers $\boldsymbol{n_x}$ and $\boldsymbol{n_y}$ and the momentum along the x and y axes will be quantized. If we restrict the particle to a three dimensional space, three quantum numbers $\boldsymbol{n_x}, \boldsymbol{n_y}$ and $\boldsymbol{n_z}$ will be needed to specify the momentum state.

Let us consider a specific state characterized by a particular set of quantum numbers $\boldsymbol{n_x}, \boldsymbol{n_y}$ and $\boldsymbol{n_z}$. This state is represented by a point A in the momentum space with coordinates $\frac{\boldsymbol{h}}{\boldsymbol{L_x}}\boldsymbol{n_x}, \frac{\boldsymbol{h}}{\boldsymbol{L_y}}\boldsymbol{n_y}$ and $\frac{\boldsymbol{h}}{\boldsymbol{L_z}}\boldsymbol{n_z}$ as shown in Figure(1.5). The rectangular volume contained in the momentum space within coordinate numbers (n_x, n_y, n_z), $(n_x + 1, n_y, n_z)$, $(n_x, n_y + 1, n_z)$, $(n_x, n_y, n_z + 1)$, $(n_x + 1, n_y + 1, n_z)$, $(n_x, n_y + 1, n_z + 1)$, $(n_x + 1, n_y + 1, n_z + 1)$ are the corner has eight adjacent states, one at each of its corners, and no states inside the rectangular volume. The rectangular volume has sides $\Delta\boldsymbol{p_x} = \frac{\boldsymbol{h}}{\boldsymbol{L_x}}, \Delta\boldsymbol{p_y} = \frac{\boldsymbol{h}}{\boldsymbol{L_y}}$ and $\Delta\boldsymbol{p_z} = \frac{\boldsymbol{h}}{\boldsymbol{L_z}}$, and has a volume equal to $\frac{\mathbf{h^3}}{\mathbf{L_x L_y L_z}}$. Since each corner is shared by eight adjacent rectangular volumes, we can think that each rectangular volume equal to $\frac{\boldsymbol{h^3}}{\boldsymbol{L_x L_y L_z}}$ contains only one momentum state. The density of momentum states in the momentum space is the reciprocal of this volume. Refer to Figure (1.5).

According to quantum mechanics, the electron state is completely specified by four quantum numbers which are the three quantum numbers such as $\boldsymbol{n_x}, \boldsymbol{n_y}$ and $\boldsymbol{n_z}$ obtained by constraining the particle to a small three dimensional volume and a fourth quantum number called the ***spin quantum number***. The spin quantum number can either be $\frac{1}{2}$ or $-\frac{1}{2}$ for electrons. Thus for a specific set of integers

for $\boldsymbol{n_x, n_y}$ and $\boldsymbol{n_z}$, we have 2 quantum states one with numbers $n_{x,} n_y$ and n_z and $\frac{1}{2}$, and the other with $n_{x,} n_y$ and n_z and $-\frac{1}{2}$.

In the momentum space, in a rectangular volume equal to $\frac{h^3}{L_x L_y L_z}$ there are two electron states. Another way of saying this is that each electron state is contained within a volume $\frac{h^3}{2\, L_x L_y L_z} = \frac{h^3}{2V}$ of the momentum space where the volume V of the box is $\boldsymbol{L_x L_y L_z}$. We can write that the number of states in unit volume of the momentum space is equal to $\frac{2V}{h^3}$ and this is called the ***density of electron states*** *in* ***momentum space***. If we take the potential energy to be zero (by suitably choosing the zero reference for the energy), then the total energy E, is equal to the kinetic energy. If we describe a sphere of radius $\boldsymbol{p}$ in the momentum space with the origin at the center of the sphere, all states lying on the surface of the sphere will have the same energy. This leads us to conclude that the constant energy surface, i.e., a surface generated by connecting all states with the same energy, is a spherical in the momentum space.

Suppose we want to determine the number of electron states having momentum components between p_x and $p_x + dp_x$, p_y and $p_y + dp_y$, p_z and $p_z + dp_z$. We consider an elementary rectangular volume of sides $dp_x dp_y dp_z$ with its corner at the coordinates $p_x\ p_y\ and\ p_z$ in the momentum space as shown in Figure (1.6), and count the number of states within this volume. Instead of counting, we can simply multiply the volume of the elementary rectangular momentum space by the density of states $\frac{2V}{h^3}$ to obtain the number of states.

The magnitude of the momentum $\boldsymbol{p}$, is given by

$$p = \sqrt{p_x^2 + p_y^2 + p_z^2}$$

Usually, we are more interested in finding the number of states having the magnitude of momentum between p and$d\ p + dp$. In the momentum space if we describe two spheres around the origin, one with the radius equal to p and the other with radius equal to $p + dp$ as shown in Figure (1.6), all the states in the interspace between the surfaces of these two spheres correspond to momentum magnitude between p and $p + dp$. The volume of the interspace between the two spheres is equal to $\boldsymbol{4\pi p^2 dp}$. If we multiply this volume by the density of states $\left(\frac{2V}{h^3}\right)$ in the momentum space, we obtain $Z(p)dp$ the number of states with the magnitude of momentum in the range between p and $p + dp$. The number of states having magnitude of momentum between p and $p + dp$ as

$$Z(p)dp = \frac{2V}{h^3} 4\pi p^2 dp$$

$$= \frac{8\,\pi V}{h^3} p^2 dp \qquad (1.29)$$

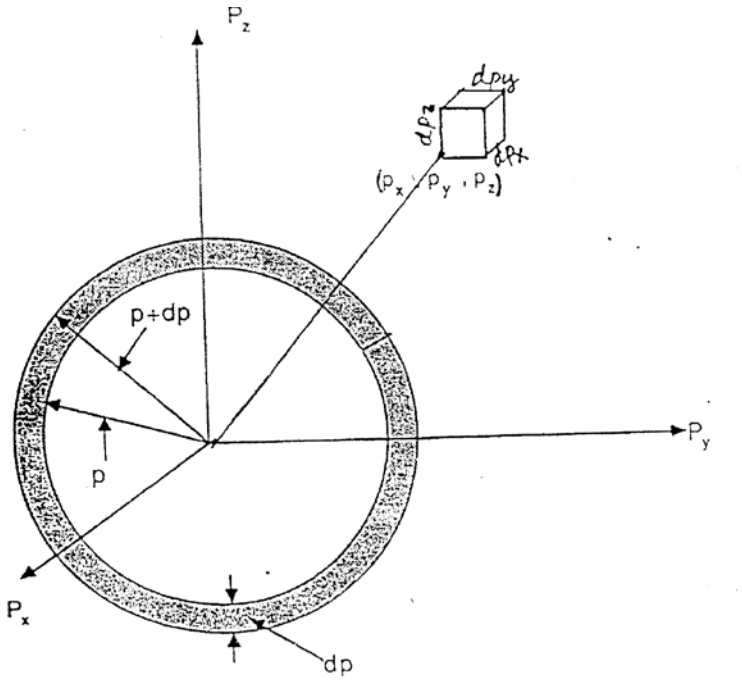


Figure (1.6): Constant momentum (magnitude) and constant energy surface

Fermi-Dirac Statistics

When we have a single electron, this electron will occupy the state with the lowest energy. If we have a very large number of electrons in the box, all these electrons cannot occupy the same lowest energy state. According to a famous principle in physics called ***Pauli exclusion principle***, only one electron can occupy a given quantum state i.e., a state specified by the assignment of four quantum numbers. We start therefore filling the states, starting from the lowest energy state, until we have exhausted all the electrons. The last electron fills the state with the highest energy among all occupied states. This kind of distribution of electrons among the states is called the ***Fermi-Dirac statistics or distribution.***

According to Fermi-Dirac statistics, the ***probability*** that a given state is occupied by an electron is given by the following function:

$$f(E) = \frac{1}{e^{\frac{E-E_F}{kT}}+1} \tag{1.30}$$

This function given by Equation (1.30) is called the ***Fermi function*** and the parameter $\boldsymbol{E_F}$ is called the ***Fermi energy***. This function is plotted in Figure (1.7).

At absolute zero temperature, $\boldsymbol{f(E)}$ is 0 for $E > E_F$ and is 1 for $E < E_F$. This means that at $T = 0^0 K$, states with energy less than E_F are occupied and states with higher energy are empty. E_F represents the maximum value of the energy of occupied states. The Fermi energy acts as a

threshold energy separating the occupied states and the vacant states. The Fermi function exhibits a discontinuity at $T = 0^0 K$ in going from a value of 1 to 0 at $E = E_F$.

As the temperature is raised, the electrons in states close to the Fermi energy get excited by thermal energy to states with energy higher than the Fermi energy. This leads to a gradual variation with energy of the Fermi function from 1 to 0 around E_F. Since the thermal energy is of the order of kT which is very much smaller than E_F, the region of transition from 1 to 0 is very small compared with E_F as shown in Figure (1.7). Even at higher temperature, states with energy less than the Fermi energy by several kT, have a probability of 1 for occupation by an electron. Similarly, states with energy greater than the Fermi energy by several kT have a probability of nearly 0 for occupation by an electron. The probability of a state with energy equal to E_F, being occupied by an electron is $\frac{1}{2}$.

Determination of Fermi energy

We denote the number of electrons having the magnitude of momentum between p and $p + dp$ as dN_p. dN_p is obtained by multiplying Equation (1.29) and Equation (1.30).

$$dN_p = Z(p)dp\, f(E) = \frac{8\pi V}{h^3} \frac{1}{e^{\frac{E-E_F}{kT}}+1} p^2 dp \tag{1.31}$$

Let us now determine the distribution of electrons as a function of energy E. We want to find the number of electrons having energy between E and $E + dE$. Let us denote this number by dN_E. In Figure (1.6) the surfaces of the two spheres with radius p and $p + dp$ are constant energy surfaces one with energy E and the other with $E + dE$. Therefore dN_E is the same as dN_p, but dN_E needs to be expressed in terms of E instead of in terms of p. Since E, the total energy equal to the kinetic energy, when the potential energy is taken as zero,

$$E = \frac{p^2}{2m}$$

Differentiating,

$$dE = \frac{2p\, dp}{2m} = \frac{p\, dp}{m} \tag{1.32}$$

$$p\, dp = m\, dE \tag{1.33}$$

and

$$p = \sqrt{2\, m\, E} \tag{1.34}$$

Substituting these in Equation (1.31), we obtain

$$dN_E = dN_p = \frac{8\pi V}{h^3} \frac{1}{e^{\frac{E-E_F}{kT}}+1} p^2 dp = \frac{4\pi V(2m)^{\frac{3}{2}}}{h^3} \frac{E^{\frac{1}{2}}}{e^{\frac{E-E_F}{kT}}+1} dE \tag{1.35}$$

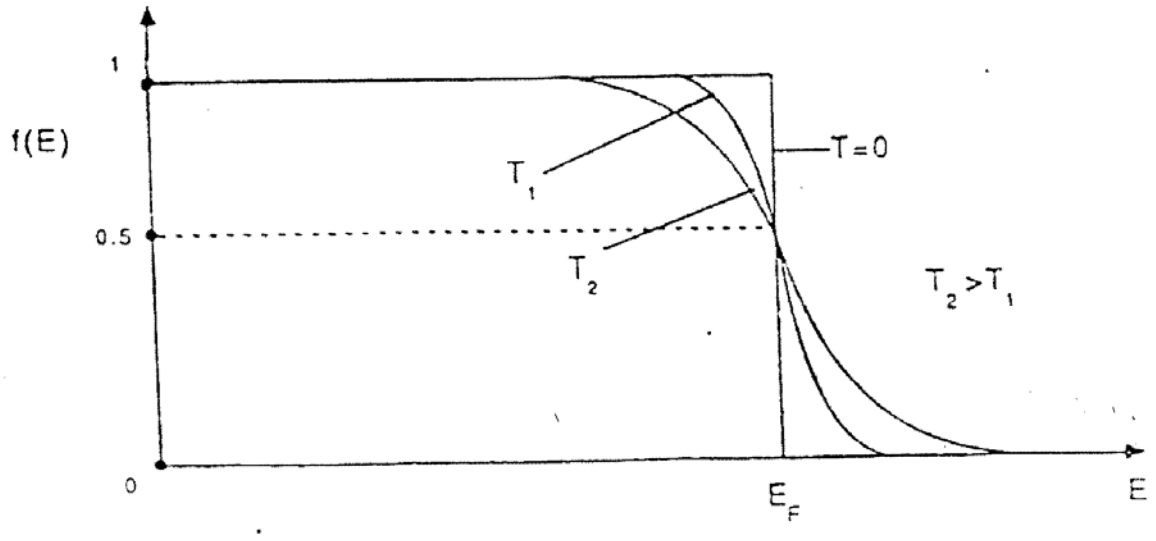


Figure (1.7): Fermi function (also called the Fermi-Dirac distribution function)

This can also be written in the form

$$dN_E = f(E)Z(E)dE \tag{1.36}$$

where $f(E)$ is the Fermi function and $Z(E)$ is given by

$$Z(E) = \frac{4\,\pi\,V(2m)^{\frac{3}{2}}}{h^3}\;\; E^{\frac{1}{2}} \tag{1.37}$$

$\boldsymbol{Z(E)}$ is called the ***density of states*** *(in energy)* and can be interpreted as the number of quantum states in unit energy interval i.e., $\boldsymbol{Z(E)dE}$ is the number of states having energy between E and $E + dE$.

Figure (1.8) shows a plot of the density of state function, $Z(E)$. This figure gives the distribution of states as a function of energy. We see that the density of states is proportional to $E^{\frac{1}{2}}$. We multiply $Z(E)$ by the Fermi function $f(E)$ or divide Equation (1.35) by dE to obtain $\frac{dN_E}{dE}$, the distribution of electrons as a function of energy. $\frac{dN_E}{dE}dE$ represents the number of electrons with energy between E and $E + dE$ and is plotted in Figure (1.9). At $T = 0^0K$, no electron has energy larger than E_F. However, at high temperatures, a few electrons occupy states with larger energy. The higher the temperature, the more electrons there are with higher and higher energy. This process is referred to as thermal excitation of electrons to higher energy states.

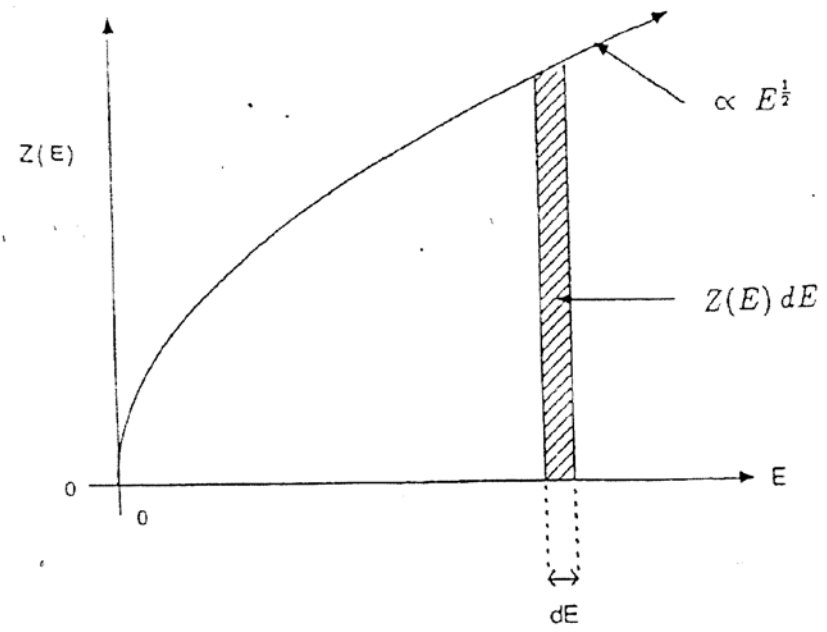


Figure (1.8): Distribution of states as a function of energy

The Fermi energy, $\boldsymbol{E_F}$, can be obtained by a simple treatment as shown below: At absolute zero temperature, the total number of occupied states in the momentum space shown in Figure (1.6) are those which lie within a sphere on whose surface the energy is equal to E_F and this should be equal to N the total number of electrons in the box. A sphere bounded by the constant surface E_F has a radius $\boldsymbol{p_F}$ given by

$$\frac{p_F^2}{2m} = E_F$$

The volume of the sphere is equal to $\frac{4\pi}{3}p_F^3$. Multiplying this by the density of states, $\frac{2V}{h^3}$ we obtain

$$N = \frac{4\pi}{3}\, p_F^3\, \frac{2V}{h^3}$$

$$= \frac{8\pi V}{3h^3} p_F^3$$

$$= \frac{8\pi V}{3h^3}\, (2m)^{\frac{3}{2}}\, E_F^{\frac{3}{2}} \qquad (1.38)$$

We used the relation between p_F and E_F in the above equation. We can now rearrange this equation to express E_F as

$$E_F = \left(\frac{3h^3}{8\pi\,(2m)^{\frac{3}{2}}}\, \frac{N}{V}\right)^{\frac{2}{3}}$$

$$= \left(\frac{3\,h^3}{8\,\pi\,(2m)^{\frac{3}{2}}}\; n \right)^{\frac{2}{3}} \qquad (1.39)$$

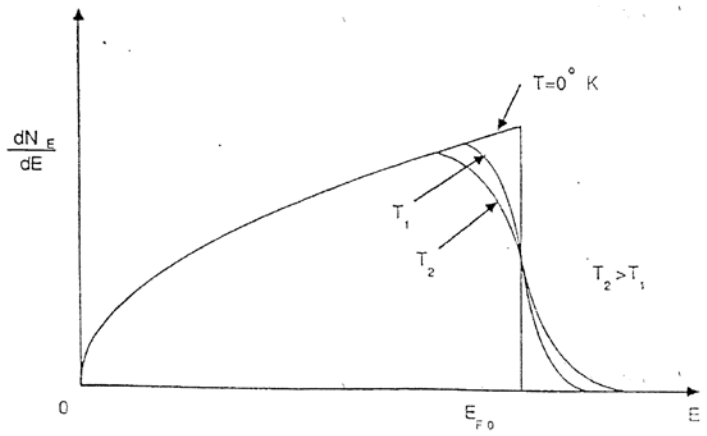


Figure (1.9): Electron distribution as a function of energy

where $n = \frac{N}{V}$ is the number of electrons per unit volume (i.e. density) in the box. *The Fermi energy is dependent only on the density and not on the total number of electrons*. As we will see later on, electrons in metallic solids that are free to roam around in the solid are modeled as electrons in a box of the same dimensions as the piece of solid. We can determine the Fermi energy of the solid, knowing n, the density of electrons. Whether a solid is large or small in size, the Fermi energy is the same independent of size because ***it depends only on the density of electrons***.

Example

Let us calculate the Fermi energy of a solid that has 10^{22} cm^{-3} free electrons. These electrons are free in the sense that they can move from one region of the solid to another region. Therefore they can be treated as electrons in a box of the same dimension as the solid.

$$n = 10^{22}\ cm^{-3}$$

$$= 10^{28}\ meter^{-3}$$

$$m = 9.11 \times 10^{-31}\ K\ gm$$

$$h = 6.63 \times 10^{-34}\ Joules - sec$$

Substituting these values in the expression for Fermi energy given in Equation (38) we obtain the value of Fermi energy as

$$E_F = \left(\frac{3 \times 10^{28}}{8 \times \pi \times (2 \times 9.11 \times 10^{-31})^{\frac{3}{2}}} \right)^{\frac{2}{3}} \times (6.63 \times 10^{-34})^2$$

$$= 2.71 \times 10^{-19} \, Joules = 1.69 \, eV$$

According to classical physics, all motions stop at $T = 0^0 K$. But according to quantum mechanics, even at $T = 0^0 K$ electrons move with speeds distributed between 0 and a maximum value v_F, where v_F is the speed of electrons with energy equal to E_F. We can calculate the speed v of the electron with energy (kinetic) E, through the relation $E = \frac{1}{2} m v^2$. The maximum velocity that an electron can have at $T = 0^0 K$ is given by

$$E_F = \frac{1}{2} m v_F^2 \tag{1.40}$$

$$v_F = \sqrt{\frac{2 E_F}{m}} \tag{1.41}$$

We can determine the average speed or momentum or energy of a collection of electrons in a box since we know the distribution functions $\frac{dN_p}{dP}, \frac{dN_E}{dE}$ for the electrons. In the example below we calculate the average speed of electrons in a box at absolute zero temperature.

Example

v, the speed, is equal to $\frac{p}{m}$ and hence the average speed at $T = 0\ ^0$K is given by

$$< v > = < \frac{p}{m} >$$

$$= \frac{\int_0^\infty \frac{p}{m} dN_p}{\int_0^\infty dN_p}$$

$$= \frac{\int_0^\infty \frac{p}{m} Z(p) f(E) dp}{\int_0^\infty Z(p) f(E) dp}$$

To evaluate this integral we take advantage of the property that $f(E)$ is equal to 1 when p is between 0 and p_F and is equal to 0 when p is greater than p_F at absolute zero temperature.

$$< v >= \frac{\int_0^{p_F} \frac{p}{m}\, p^2 dp}{\int_0^{P_F} p^2 dp}$$

$$= \frac{p_F^4}{4m} \times \frac{3}{p_F^3}$$

$$= \frac{3}{4}\,\frac{p_F}{m} = \frac{3}{4}\, v_F$$

Thus we see that the average speed of the electrons is $\frac{3}{4}$ of the maximum speed v_F.

If we want to localize the position of the electron in the box, by representing it as a wave packet, we superimpose a large number of plane waves of suitable amplitude and phase and of different k values with an average value of k_0. The velocity of the electron is the group velocity of the wave packet and is equal to

$$\frac{d\omega}{dk} = \frac{1}{\hbar}\,\frac{dE}{dk} \tag{1.42}$$

determined at $k = k_0$. Referring to Figure (1.10), the velocity of the electron for the state $k = k_0$, is related to the slope of the energy E vs k plot at $k = k_0$. Since we have

$$E = \frac{\hbar^2 k^2}{2m}$$

the velocity of the electron in the state $k = k_0$ is equal to

$$v_{el} = v_g = \frac{1}{\hbar}\left.\frac{dE}{dk}\right|_{k=k_0} = \frac{\hbar\, k_0}{m} \tag{1.43}$$

This is the usual result that velocity is momentum divided by m. We considered the potential energy to be constant inside the box. However, the potential energy inside the solid will be periodically varying. When we assume that the potential energy inside the box is varying periodically, an interestingly different E vs k diagram is obtained and the velocity is not equal to the momentum divided by the mass.

In the treatment of electrons in a box, the potential energy was assumed to be constant and hence there was no force on the electron. Hence the electrons were said to be free. When we apply this model to treat the case of electrons in a metallic solid, we call it the ***free electron model of the solid***. We found that the energy levels are discrete (**quantized**). However, since the dimensions of the solid are large in comparison with atomic distances, the discrete energy levels are very close to each other and hence for all practical purposes can be considered to be continuous. This is particularly true when we consider higher energy states where the quantum numbers are large and the separation in energy between adjacent states expressed as a fraction of the energy of the state becomes extremely small. For this reason, the energy is said to be ***"quasi-continuous"***. For the same reason the corresponding momentum $k_{x,}k_y$ and k_z as well as the magnitude of $|k|$ are quasi-continuous. This model is applicable to describe the behavior of conduction electrons in metallic solids. You will see in the next chapter that, in a metallic solid, each atom of the solid contributes an electron or two to form a sea of electrons that belong to the whole solid and are free to move in the solid like electrons in a box. This model does not

explain why some solids are conductors and some others are not. We have to use a new model called the "Band Theory of Solids" to understand the conducting properties of solids.

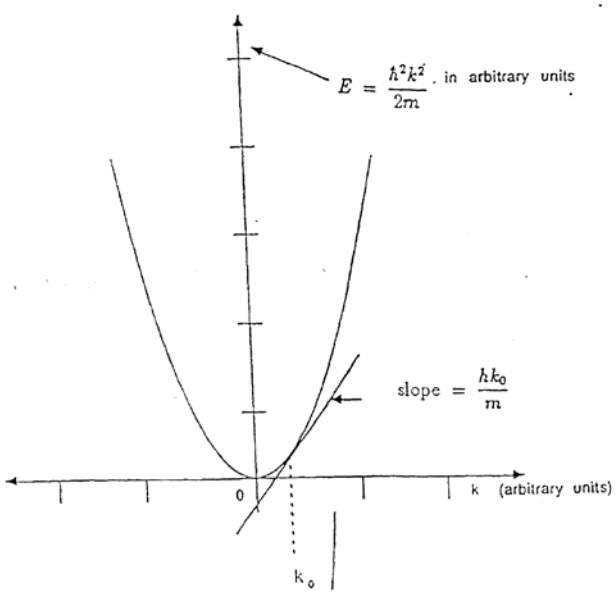


Figure (1.10): E vs k digram

Band Theory of Solids

In a typical solid with free electrons, the electrons are subjected to an internal force due to the positively charged ions of the atoms. The potential energy varies periodically due to the periodic arrangement of the atoms in the solid. This force acting on the electrons is called the ***internal force*** as distinct from an externally applied force such as that due to an electric field arising from the application of a voltage between the two ends of the solid.

To illustrate the concept of "allowed" and "forbidden" bands of energy, let us assume a one-dimensional solid as shown in Figure (1.11). The positive ions are assumed to be situated periodically along the x-axis with an inter-atomic distance of 'a' units. The electrons can move only along the x-axis in either direction (i.e., in only one dimension). The potential energy of the electron due to the attractive electrostatic force of the positive ions varies periodically with x as shown in Figure (1.12). Due to the periodic potential energy, the electron states are allowed to have energy values only in certain ranges and not in others. The range of energy values in which the electron has states is called an ***allowed band of energy***. The allowed bands are separated by ***forbidden* (or disallowed) bands** of energy in which the electrons have no states and the forbidden bands are called ***energy gaps,*** (see Figure 1.13).

In the free electron model, the electron energy varies quasi-continuously from $0\ to\ \infty$ as shown in Figure (1.14). In this figure, the kinetic energy, which is the same as the total energy since the potential energy is constant and can be taken as zero by a proper choice of the zero reference for energy, is plotted as a function of k. Motion is possible only along one dimension. k varies quasi-continuously from $-\infty\ to\ +\infty$.

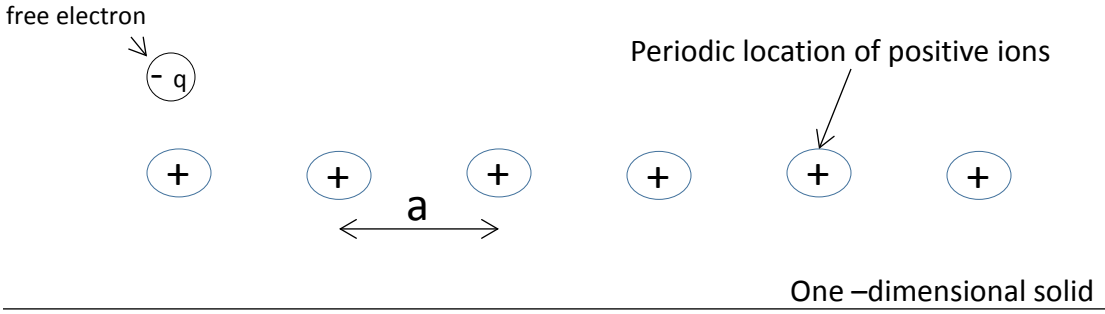


Figure (1.11) Positive ions and free electrons in a one-dimensional solid

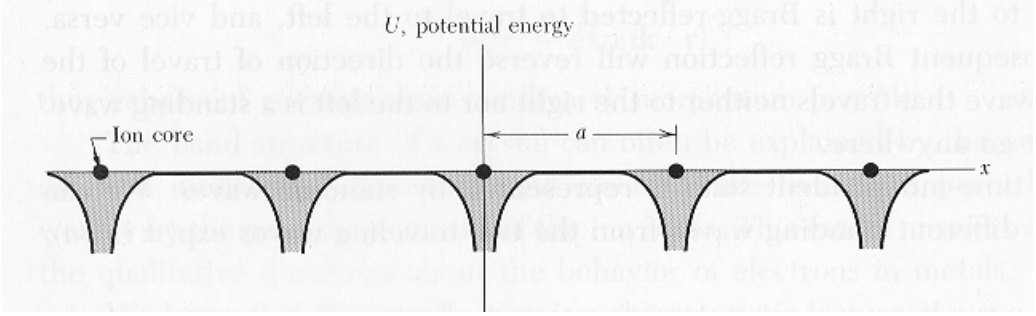


Figure (1.12) Electrostatic potential energy variation with distance

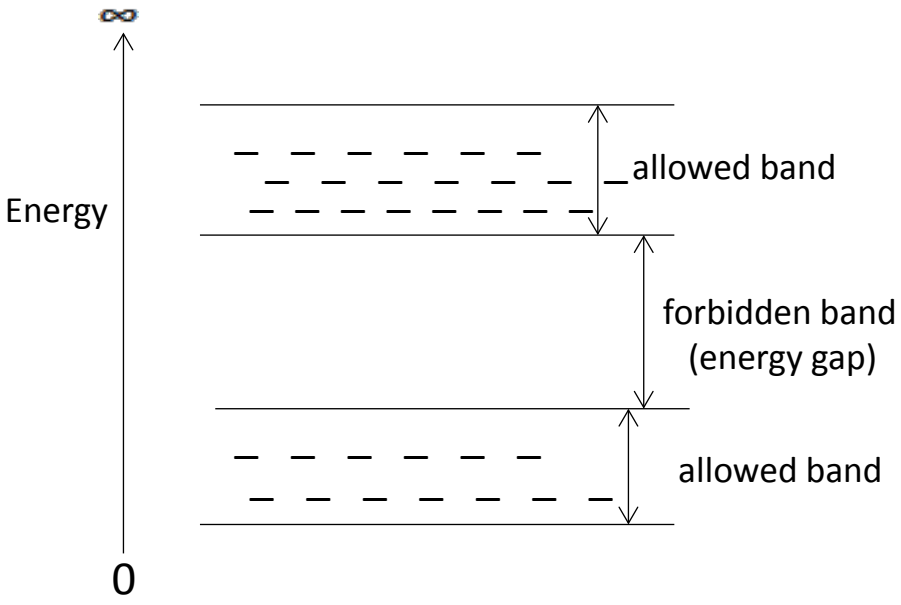


Figure (1.13) Bands of allowed and forbidden energy

Let us assume that the potential energy varies periodically with the periodicity a and is given by

$$U(x) = U(x + a) \tag{1.44}$$

When the existence of the periodic potential energy is treated quantum mechanically, it is shown in solid state physics books that the energy varies with k as shown in Figure (1.15). Notice that the energy of the states is split into bands. The solid line in the Figure (1.15) represents the energy according to the band theory. Let us now examine the essential features of the band theory.

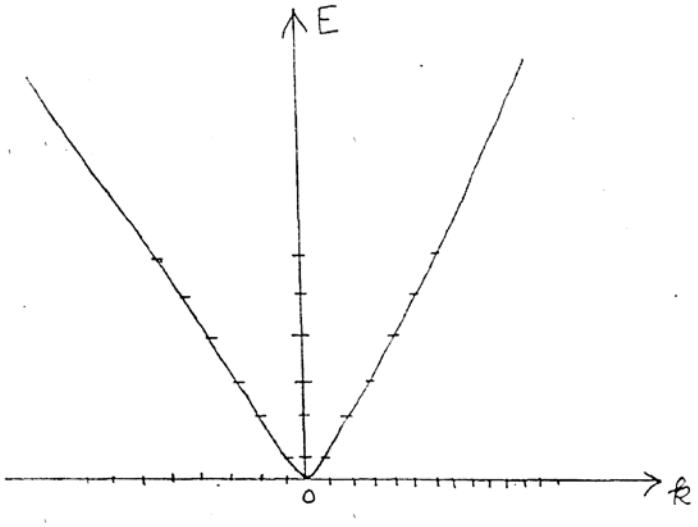


Figure (1.14): $E - k$ diagram for free electrons. Quasi-continuous distribution of energy as a function of the magnetitude of the momentum. Notice that the energy values extend from 0 to ∞

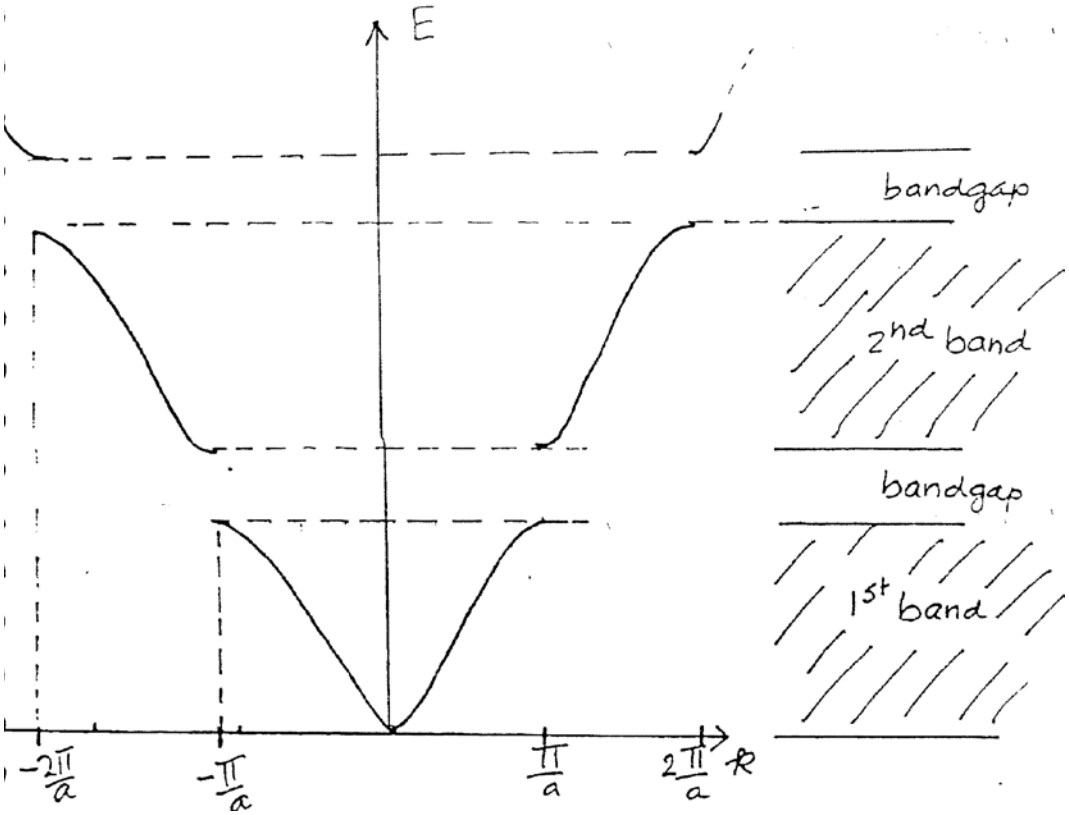


Figure (1.15): A plot of energy as a function of k ($E\ vs\ k$) in the band theory in the one-dimensional approximation. Notice that the energy values are broken into bands of allowed and forbidden values.

As the magnitude of k is increased from 0, the energy increases symmetrically with k for both positive and negative values of k until k becomes equal to $\pm\frac{\pi}{a}$. At this value of k, the energy has a discontinuity and if k is increased beyond $\frac{\pi}{a}$ then the energy again increases. The discontinuity in energy at $\pm\frac{\pi}{a}$ is called the energy gap or the bandgap. The continuous range of energy for values of k between $-\frac{\pi}{a}$ and $+\frac{\pi}{a}$ is called the first energy band. It can be noticed that another energy gap occurs at $-2\frac{\pi}{a}$ and $+2\frac{\pi}{a}$ is called the second band of energy. In general, the bandgap occurs at $k = \pm\frac{n\pi}{a}$ where n is an integer. Thus, according to the band theory of solids, the energy values are split into allowed and forbidden bands of energy. The forbidden band of energy is also referred to as bandgap and the allowed band is referred to as energy band. At the bottom and at the top of the energy band, the energy plot has zero slope. In the middle of the band, the energy has a parabolic dependence on k, i.e., is proportional to k^2 as in the free electron model. Notice that, in the region of the band, the dotted line and the solid line overlap ***showing that the behavior of electrons in this region of the band is similar to the free electron behavior***.

It is shown in text books in solid state physics that in a band, two states whose k value differs by $\frac{2\,n\,\pi}{a}$ where n is a positive or negative integer are equivalent states if they also have the same energy. Referring to the Figure (1.16), A and B are equivalent states because the k values differ by $\frac{2\ \pi}{a}$ and they have the same energy. Similarly, C and D are equivalent states.

Since the velocity of the electron is the group velocity of the wave-packet as given by Equation (1.42), we notice that the electron velocity varies in the bandgap due to the variation of the slope in the E versus k plot i.e., $\frac{dE}{dk}$ in the band. The electron has zero velocity in the bottom and in the top of the band since the slope is zero. It has maximum velocity in the middle of the band where the inflection point occurs.

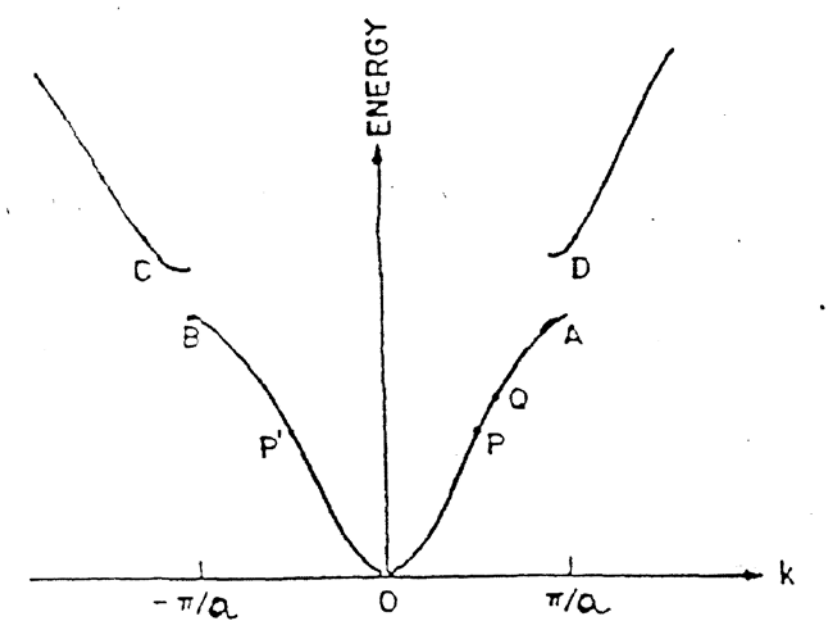


Figure (1.16): One dimensional energy band diagram illustrating equivalent states

Effect of an Applied Electric Field

We will now examine the behavior of the electrons in a band under the application of an external field. Let us assume that the electric field $\boldsymbol{\mathcal{E}}$ is switched on at time $t = 0$. There is a force on the electron equal to – $\boldsymbol{q\mathcal{E}}$. In a short interval of time δt, the electron moves a distance $v\ \delta$t and gains an energy equal to

$$\delta E = -q\ \mathcal{E}\ v\ \delta \mathrm{t} \tag{1.45}$$

The energy of the electron is changed by δE, and it should be now in a new state $k + \delta k$.

The differential in energy δE is given by

$$\delta E = \left(\frac{dE}{dk}\right)\delta k \tag{1.46}$$

The velocity was seen earlier to be equal to the group velocity of the wave packet i.e.

$$v = v_l = \frac{1}{\hbar}\frac{dE}{dk} \tag{1.47}$$

Hence δE can be expressed in terms of the velocity of the electron as

$$\delta E = \hbar\ v\ \delta k \tag{1.48}$$

Equating this expression for δE with that given in equation (1.45), we obtain after rearrangement

$$-q\ \mathcal{E} = \hbar\ \frac{dk}{dt} \tag{1.49}$$

Thus we see that $\hbar \frac{dk}{dt}$ is equal to the external force on the electron due to the applied electric field. We can infer, therefore, that the effect of an external force is to make the k value change with time at a rate given by

$$\frac{dk}{dt} = \frac{external\ force}{\hbar} \tag{1.50}$$

What do we mean by a change in k value? The electron changes or makes a transition from a state of some k value to another of different k value. Under the application of the external field (force) the electron goes from one k state to the next higher k state and from that state to the next higher k state and thus keeps on making these transitions as long as the external force is applied. In other words, *the electron changes its k state under the application of an external force*.

Referring to Figure (1.16), an electron in the state P will jump to the next higher state Q due to the external field. Then it will jump to a next higher k state and keep on moving to higher k states until it reaches the state A. According to Equation (1.45), the velocity of the electron in any given state is proportional to the slope of the energy versus k plot known usually as $E - k$ plot. In the upper left half of the bandgap, the $E - k$ plot is concave downwards which means that the slope (velocity) decreases as k is increased. As the electron changes its state progressively from P to A, the electron velocity is decreasing i.e., the electron is slowing down. Since the slope is zero at A, the electron velocity becomes zero. The k value cannot increase to a value higher than $\frac{\pi}{a}$ since there is an energy discontinuity. However, the state B is equivalent to A and therefore when the electron is in the state $k = \frac{\pi}{a}$, it is equivalent to being in the state $= -\frac{\pi}{a}$. Now, k can increase at the rate $\frac{dk}{dt}$ from the state B, where $k = -\frac{\pi}{a}$. Since $\frac{dE}{dk}$ is negative, the velocity is negative. The electron starts to move in the opposite direction and the magnitude of the velocity in the negative direction increases from zero when it is in the state B to a maximum value when it reaches the state P$'$. Then it starts to slow down and reaches zero velocity in the state denoted by O. As the force continues to act on the electron, the electron starts to move in the positive direction and its velocity increases until it reaches the state P. When the electron goes from state P to a higher k state, it starts to slow down and comes to rest when it reaches the state A. This cycle keeps on repeating as long as the force continues to act on the electron. The electron changes states in the band in a cyclic fashion. When a band is completely full, all the electrons are continually changing states in a cyclic fashion. At any given instant, there are as many electrons traveling in the positive direction as those in the negative direction. This is because electrons occupying states between $k = 0$ (state O) and $k = \frac{\pi}{a}$ (state A) correspond to electrons traveling in the positive direction while electrons occupying states between $k = 0$ (state O) and $k = -\frac{\pi}{a}$ (state B) travel in the negative direction. The average velocity is therefore zero and hence a completely filled band does not contribute to electrical conductivity.

When the electron is in a state between state P and state A, the effect of the external field is to decelerate the electron (i.e., as time progresses its velocity decreases). Normally, when we consider free electrons, the electron will accelerate under the action of the external field. It is as though an electron in states between P and A has a negative mass. Similarly, the electron behavior in states between P' and B is as though the electron has a negative mass. We define therefore, an **effective**

mass m^*, which, when multiplied by the acceleration, gives the force on the electron. Since states in the upper half of the band correspond to deceleration of electrons, they have a negative effective mass while those in the bottom half of the band have a positive effective mass.

The acceleration of the electron is given by

$$a = \frac{dv}{dt} = \frac{d}{dt}\left(\frac{1}{\hbar}\frac{dE}{dk}\right) = \frac{1}{\hbar}\frac{d^2E}{dtdk} = \frac{1}{\hbar}\frac{d^2E}{dk^2}\frac{dk}{dt} \tag{1.51}$$

But the force $-q\mathcal{E}$ is equal to $\hbar\frac{dk}{dt}$, Hence

$$a = \frac{1}{\hbar^2}\frac{d^2E}{dk^2}\left(\hbar\frac{dk}{dt}\right) = \frac{1}{\hbar^2}\frac{d^2E}{dk^2}(-q\,\mathcal{E}) = \frac{1}{\hbar^2}\frac{d^2E}{dk^2} \times Force \tag{1.52}$$

The term $\frac{1}{\hbar^2}\frac{d^2E}{dk^2}$ relates acceleration to the applied force just as mass relates total force and acceleration. We thus define the ***effective mass m^**** as

$$m^* = \frac{\hbar^2}{\frac{d^2E}{dk^2}} \tag{1.53}$$

The curvature $\frac{d^2E}{dk^2}$ of the $E\ vs\ k$ plot is positive in the lower half of the band and hence m^* is positive. The curvature is negative in the upper half of the band, and hence the effective mass is negative.

What is the philosophical interpretation of the effective mass? The effective mass relates the acceleration of the electron to the external force. If one were to include all the forces internal and external, then indeed the relation between the acceleration and the total force will be through the real mass of the electron. We are relating only the external force to the acceleration and hence we need to define an effective mass.

The effective mass is a useful concept since it enables one to determine the behavior of electrons by treating them as though they are free from any internal force (i.e., constant potential energy) and replacing the true mass by the effective mass m^* in the expressions describing its dynamic behavior. In the 3-dimensional solid, the effective mass m^* is a second-order tensor which relates the external force along one direction to the acceleration along another direction. For example, in a general case, unless the external force is applied along one of the three principal axes of the crystal, the resulting acceleration will not be in the same direction as the force. However, for the purposes of our discussion, we will assume that the effectiveness m^* is a scalar quantity i.e., the external force and the acceleration are in the same direction.

Conductors, Insulators and Semiconductors

Let us now consider a material in which the upper most occupied band is partially filled while all the lower bands are completely filled. There is no contribution to the electrical conductivity from the lower bands since they are full as stated earlier. On the other hand, the electrons in the partially filled (highest) band move continuously to higher k states under the action of the eternal force. In a time interval Δt, the electrons would have changed to new states that differ in the k value by

$$\Delta k = \frac{external\ force}{\hbar}\Delta t \tag{1.54}$$

One would expect that the electrons will be continually changing states in this fashion but this does not occur. The electrons *collide with scattering centers* such as impurities and other atoms of the crystal that have been displaced from their normal atomic sites. The scattering causes them to give up their excess energy and return to lower energy states. This process is called a **relaxation process**. The electrons are *never able to go to very high energy states due to the relaxation process*. According to solid state theory, it is possible to assume that, under steady state conditions, the distribution of electrons in the presence of the electric field $\mathcal{E}$, is the same as what one would obtain if the field $\mathcal{E}$ acted on the electrons for a time period τ_c. $\boldsymbol{\tau_c}$ is called the ***relaxation time***. Referring to Figure (1.17), we find that under steady state conditions, all the electrons have shifted to new k states separated from the original states by an amount

$$\Delta k = \frac{1}{\hbar}\,(-q\mathcal{E})\,\tau_c \tag{1.55}$$

All the states below the dotted line in the energy axis have equal number of electrons going in the positive direction as in the negative direction. Only those electrons lying above the dotted line travel in the positive direction without an equal number traveling in the opposite direction. There is thus a net flow of electrons in the positive direction. These electrons contribute to electrical conductivity. We can conclude therefore that only *partially filled bands contribute to electrical conductivity*.

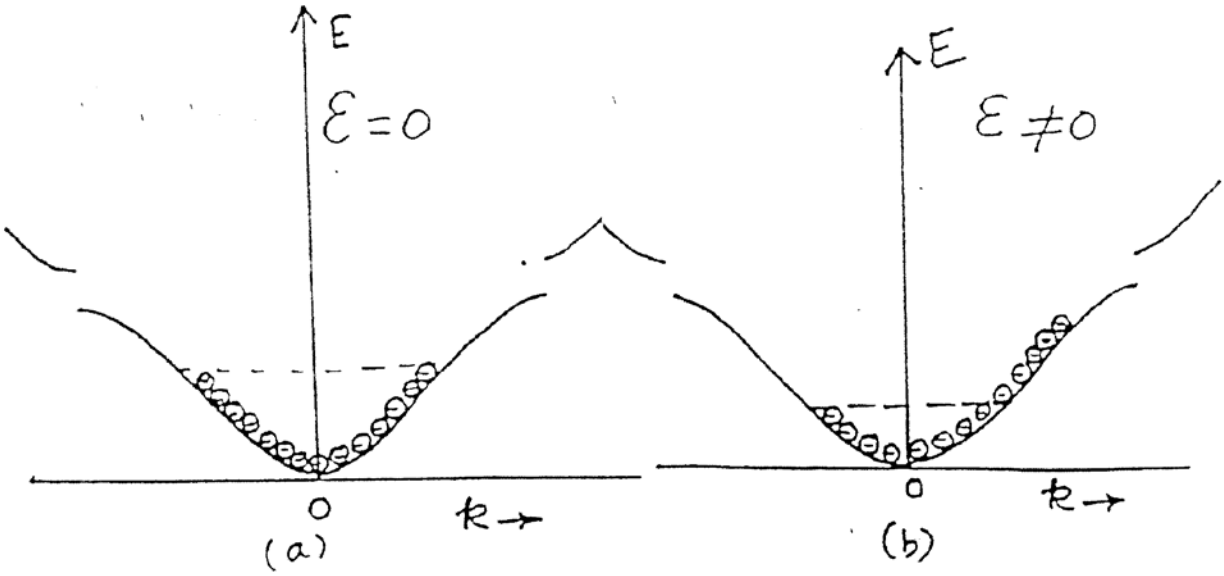


Figure (1.17) Electron distribution in the presence of an electric field

We can distinguish three classes of materials: **conductors**, **insulators** and **semiconductors**. An insulator is one in which the top-most occupied band is completely filled as shown in Figure (1.18). The top most filled band is called the ***valence band*** since the electrons in this band are the ones which give rise to the chemical bonding. The next higher band is called the ***conduction band***, since any electron excited to this band will give rise to electrical conductivity due to the fact that the conduction band

becomes partially filled. The difference in energy between the top of the valence and the bottom of the conduction band is the ***energy gap***. Insulators' energy gap which is also called the bandgap, is very large and hence even at very high temperatures only a negligible number of electrons is excited into the conduction band and the material does not carry electrical current. On the other hand, if the bandgap is small, then, even at moderately low temperatures, electrons will be excited from the valence band into the conduction band and we will get electrical conductivity from both the partially filled valence and conduction bands. Such materials are called **semiconductors**. At absolute zero temperature, the semiconductor is an insulator since there are no electrons in the conduction band. However, as the temperature is raised, electrons are thermally excited from the valence band to the conduction band and the material starts to conduct electricity. Conductors are materials in which the highest occupied band is only partially filled. The band structure for conductors is shown in Figure (1.19). The metallic solids have this feature, i.e., the highest occupied band is partially filled, and hence is a good conductor.

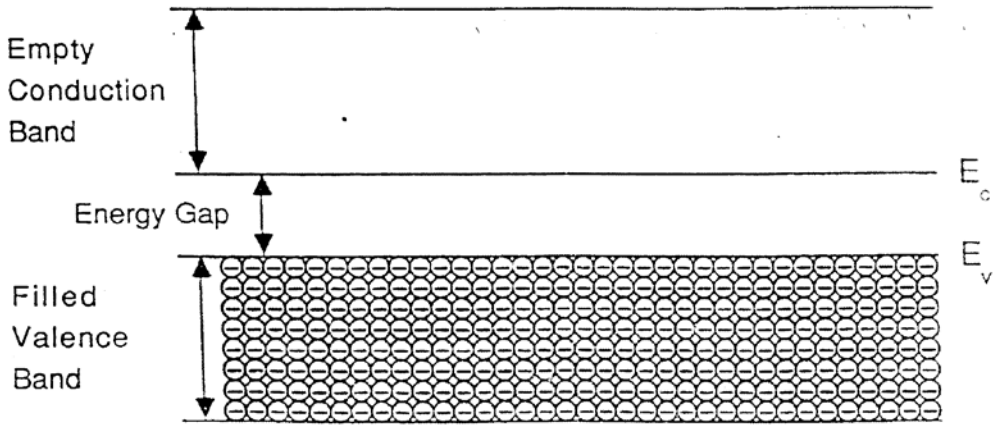


Fig. (1.18): Band Structure of an insulator

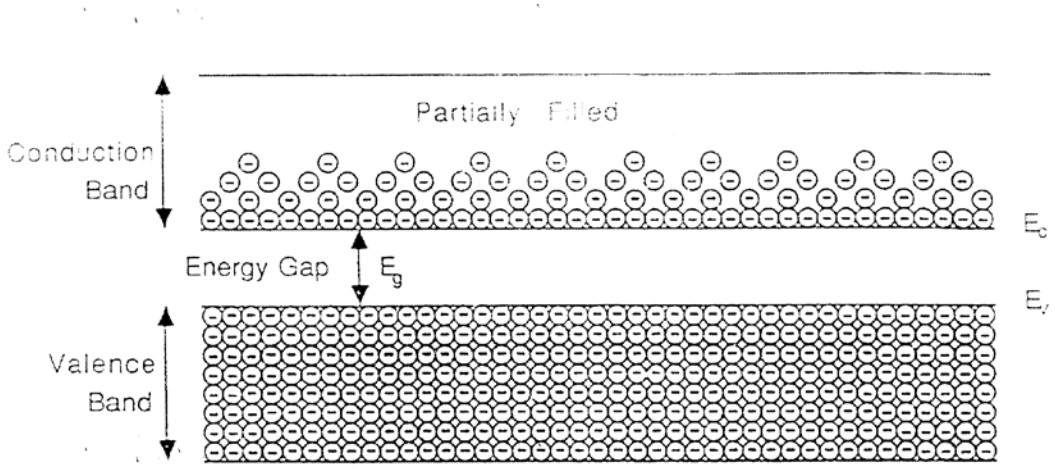


Figure (1.19): Band structure of a conductor

Concept of Holes

We saw in the last section that a semiconductor has a small bandgap with its valence band completely filled while its conduction band is empty at very low temperatures. As the temperature is raised by heating the semiconductor, electrons are thermally excited from the valence band to the conduction band. This gives rise to a partially filled conduction band and the electrons in the conduction band can give rise to electrical conduction. Similarly, the remaining electrons in the valence band can give rise to electrical conduction since the valence band is now only partially filled. However, it is a very difficult task to determine the contribution of the large number of electrons in the valence band. A simple technique can be employed to calculate the contribution of the valence band electrons to electrical conduction.

For the sake of simplicity, let us assume that only one electron is excited from the valence band to the conduction band, leaving a vacant state in the valence band. The electron in the conduction band gives rise to electrical conduction as though it is a free electron with an effective mass, m_c, appropriate to the state it occupies in the conduction band. The electron will occupy only a low energy state in the bottom of the band and therefore its effective mass is positive. On the other hand the vacant state in the valence band is in the top of the band and the effective mass is negative. An electron with an effective mass $-m_v$, can occupy the vacant state in the valence band, and make the valence band fully occupied. Assume that we introduce two fictitious particles one with a charge $-q$ coulomb and a mass $-m_v$ and the other with a charge q coulomb and a mass m_v. The particle with the negative charge and negative effective mass fills the vacant state making the valence band completely filled. However, the particle with the positive charge and positive effective mass will now be free to move around the crystal and contribute to the electrical conductivity. This fictitious particle with a positive charge and a positive effective mass is called a ***hole***. It behaves like a free particle and carries a charge. It therefore contributes to electrical conductivity. For every electron excited from the valence band, there is hole in the valence band. We obtain electrical conductivity due to both electrons and holes.

The number of electrons in the conduction band is small in comparison with the total number of states in the conduction band. Hence, the electrons occupy a narrow range of states in the bottom of the conduction band. We can therefore assume that all these electrons have the same effective mass and calculate the electrical conductivity due to them by using the free electron model (electrons in a box) except that we will use the effective mass m_c in the place of the true mass. Similarly, all the holes are nearly at the top of the valence band and occupy a narrow range of states in the valence band. Therefore they all have the same effective mass, m_v, corresponding to the curvature of the top of the valence band. These holes can be treated as holes in a box. Thus, we can calculate the electrical conductivity of the semiconductor using free electron and free hole models.

Chapter 2

Semiconductor Material Electronic Properties

Intrinsic Semiconductor

We will briefly review, in this chapter, the essential aspects of semiconductor physics that will be relevant to the discussion of device physics. A semiconductor is a material that has an electrical conductivity much less than a metallic conductor but is much more conducting than an insulator. Furthermore, its electrical conductivity is zero at absolute zero degree temperature and increases with temperature. The atoms in a semiconducting crystal are held together by covalent bonds. In order to understand what a covalent bond is consider two hydrogen atoms which are initially infinitely apart. On bring these two atoms closer, the electron in each atom is simultaneously subjected to the attractive forces of both the nuclei and the potential energy of each electron is lowered. The lowering of the potential energy with decrease in the distance between the two nuclei corresponds to an attractive force. The reason the two nuclei do not collapse on each other is because of a repulsive force that arises between the two nuclei (Coulomb repulsion) at very close inter-nuclear distance. The two nuclei are in stable equilibrium at a separation where the attractive force is exactly balanced by the repulsive force. The two electrons have a very high probability of being found midway between the two nuclei and therefore the electrons spend most of the time midway between the two nuclei. This will not violate Pauli's exclusion principle because each atom has only one $1s$ electron. The electron in one atom can have opposite spin to the electron in the other atom and therefore these two electrons can simultaneously exist in the same place. It is as though each atom has two electrons in the 1s orbit. However if we bring two helium atoms together, a covalent bond will not result because of the fact that Pauli's exclusion principle is violated since each atom already has a filled $1s$ state (usually called $1s$ orbital). Therefore we can conclude that for covalent bonds to exist, each atom should have a partially-filled orbital.

The same covalent bond gives rise to the bonding force in many of the solids like carbon, nitrogen, and oxygen. The covalent bond is directional and therefore these atoms crystallize in structures which satisfy the directionality of the covalent bond. An atom with N electrons outside the close-shell structure requires $(8\text{-}N)$ electrons to complete the s and p sub-shell. Since two neighboring atoms each contribute one electron to the bond between them, an atom with N valence electrons requires $(8\text{-}N)$ nearest neighbors and the crystal structure that results has to provide $(8\text{-}N)$ nearest neighbors which are symmetrical situated with respect to the first atom. For example, carbon (diamond), silicon, and germanium, each has two s electrons and two p electrons. Therefore they require 4 nearest neighbors and this is obtained in the diamond structure. Figure (2.1) shows the crystal structure of diamond in which each atom is surrounded by four covalently bonded neighboring atoms. Silicon and Germanium also have similar crystalline structure.

It might be asked how Pauli's exclusion principle is not violated when we have the s-orbitals filled in each of the carbon atoms. This is explained by assuming a rearrangement of the states of the four electrons. One of the $2s$ electrons goes into a $2p$ state and thereby, we have half-filled s-orbitals and half filled p-orbitals. Such a rearrangement is called hybridization. The four new orbitals thus

obtained are (sp^3) hybrids. Figure (2.2) shows how the energy bands are formed by hybridization of s and p orbitals.

The covalent crystals exhibit great hardness, low electrical conductivity at low temperature and in pure state. These usually have strong binding but some of the crystals (semiconductors) have weaker binding. These semiconductors exhibit electronic conductivity at high temperatures.

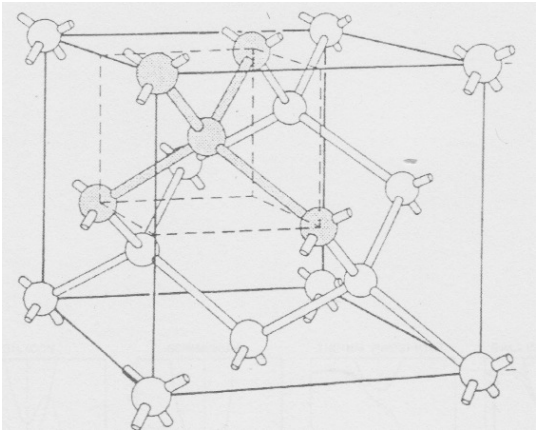


Figure (2.1): The diamond crystal structure

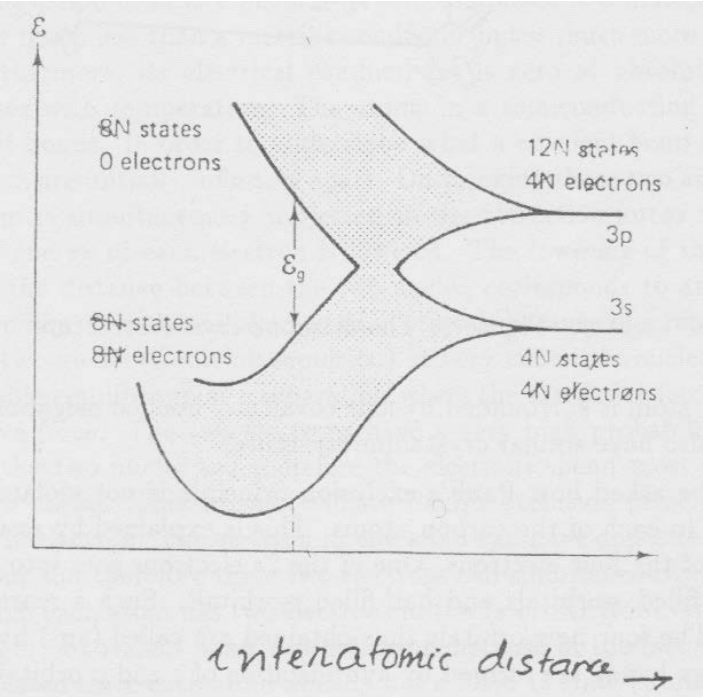


Figure (2.2): Energy band formation in diamond by hybridization of s and p orbitals

Band Structure of Semiconductors

Some of the common semiconductor materials are made up of atoms of Group IV elements like germanium and silicon; these materials are referred to as *elemental semiconductors*. Semiconductor materials such as gallium arsenide or indium antimonide are made up of compounds of atoms of Group

III and Group V elements and therefore these materials are known as compound semiconductors. The energy band structure of some common semiconductors is given in Figure (2.3). In this figure, the energy is plotted as a function of the wave vector (proportional to the momentum) (in two crystallographic directions viz., [111] and [100]. In all the semiconductors, the states in the lower bands are completely filled with electrons and the states in the higher bands are empty at very low temperatures. The electrons in the lower bands are the ones that participate in covalent bonding and as such, they are not free to move around in the crystal. Since the electrons in the lower bands are the ones in the covalent bond, the lower bands are called the valence band. The states in the higher band are empty; therefore, there are no electrons to move around in the crystal. At very low temperatures there is no broken bond and the valence band is completely filled.

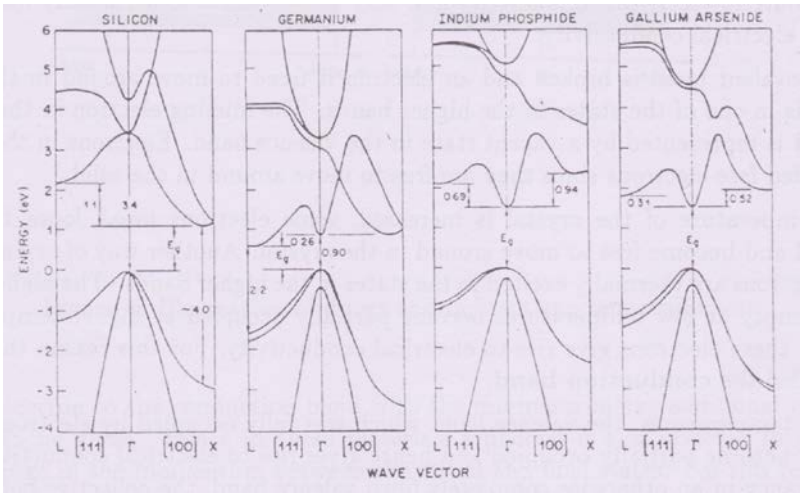


Figure (2.3): Energy band structure of some typical semiconductors

According to the principles of solid state physics, electrons in a completely filled band do not contribute to electrical conductivity and only the electrons in a partially filled band contribute to electrical conductivity.

When a covalent bond is broken and an electron is freed to move around in the solid, this electron is in one of the states in the higher bands. The missing electron in the broken covalent bond is represented by a vacant state in the valence band. Electrons in the higher bands are called free electrons since they are free to move around in the solid.

As the temperature of the crystal is increased, some electrons break loose from the covalent band and become free to move around in the crystal. That is, some electrons are thermally excited to the states in the higher bands. The higher bands (which were empty at low temperatures) become partially occupied at higher temperatures and thus, these electrons give rise to electrical conductivity. For this reason, the higher bands are called the conduction band.

At higher temperatures, the valence band (which was fully occupied by electrons at low temperatures) become partially occupied and hence, gives rise to electrical conductivity. When there is a vacancy in an otherwise completely filled valence band, the collective behavior of the rest of the electrons in the valence band is equivalent to that of a fictitious particle called the hole, which has a positive effective mass and a positive charge. The holes are free to move around in the solid and thus

called “free holes.” An electron in the conduction band is free to move around in the crystal and a hole in the valence band is free to move around in the crystal as shown in Figure (2.4).

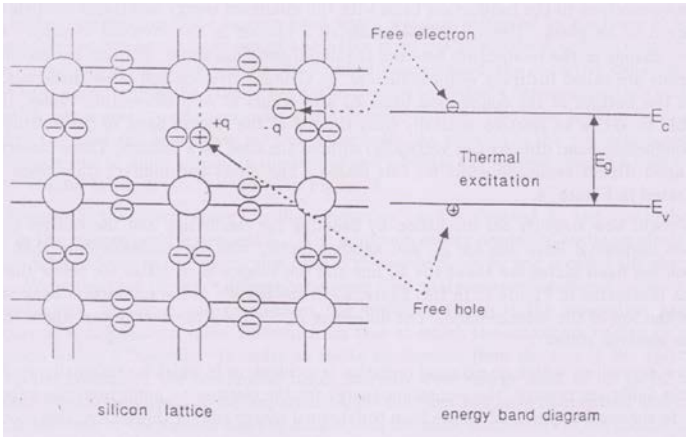


Figure (2.4): Generation of electron-hole pairs

In Figure (2.5), the lowest energy in the conduction band is denoted E_c and the highest energy in the valence band is denoted E_v. The minimum energy needed to excite an electron from the valence band to the conduction band is obviously E_c- E_v and is called the energy gap, E_g. We notice that there is a difference between elemental semiconductors and compound semiconductors. In silicon and germanium, the minimum energy in the conduction band, E_c, occurs at a momentum value different from that at which the maximum energy in the conduction band, E_v, occurs. In order to excite an electron from the top of the valence band to the bottom of the conduction band, not only does energy need to be given, but also some momentum needs to be imparted. A photon does not have momentum and can excite an electron in the valence band to a state with the same momentum in the conduction band. This is referred to as vertical transition. The minimum energy needed to excite an electron to the conduction band is $E_g = E_c$- E_v. However in silicon and germanium, to excite an electron to the conduction band with the minimum energy, additional momentum also needs to be given. Hence, in these materials, a phonon is also involved to supply the needed change in the momentum between the initial and final states. For this reason, these materials are called indirect semiconductors. In gallium arsenide and other materials, in which the bottom of the conduction band, E_c, also occurs at zero momentum values, it is possible to excite an electron optically from the top of the valence band to the bottom of the conduction band directly (or vertically) without the need of a phonon. These materials are called direct semiconductors for this reason. The direct and indirect transitions are illustrated in Figure (2.5).

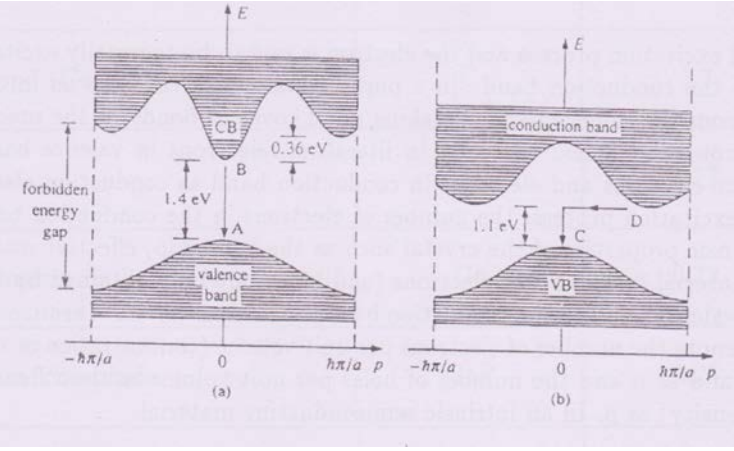


Figure (2.5): Illustration of direct and indirect gap semiconductors

We will now simplify our discussion by denoting the conduction and the valence band by two horizontal lines—one for E_c and another for E_v with the understanding that the conduction band states are above the E_c line and the valence band states are below the E_v line as illustrated in Figure (2.5). In this figure, E_c is the bottom of the conduction band, and E_v is the top of the valence band. The difference E_c- E_v is the energy gap where there are no electron states.

In a crystal, on which no external radiation is incident, or in which no externally applied electric fields are present, the excitation energy for the electron to jump from the valence band to the conduction band comes from the thermal energy. For this reason, this process is called the thermal excitation process and the electron is said to be thermally excited from the valence band to the conduction band. In a pure crystal, which contains no impurities, thermal excitation from the valence band (breaking up of covalent bonds) is the mechanism for generating electrons and holes. (Usually in literature, electrons in valence bands are referred to as valence electrons and electrons in conduction bands as conduction electrons.) Since in a thermal excitation process the number of electrons in the conduction band is a function of the intrinsic properties of the crystal such as the band-gap, effective mass, etc., a semiconducting material in which free electrons (and holes) are only obtained by thermal excitation from the valence band to the conduction band is called an intrinsic semiconductor. It is customary to denote the number of electrons per unit volume (concentration or density) in the conduction band as *n* and the number of holes per unit volume in the valence band (concentration or density as *p*. In an intrinsic semiconducting material

$$n_i = p_i \tag{2.1}$$

where the subscript *i* is purposely used to denote that the material is intrinsic.

The main requirement for a crystal to be a semiconductor is that the bonds between the atoms of the crystal should be covalent and also comparatively a small amount of energy should be needed to break up the bond. Table (2.I) gives the energy gap of some of the typical Group IV elemental semiconductors and Group III-V intermetallic compound semiconductors.

Material	Chemical Symbol	Energy Gap (eV)
Diamond (Carbon)	C	5.3
Silicon	Si	1.17
Germanium	Ge	0.72
Gallium Phosphide	GaP	2.25
Gallium Arsenide	GaAs	1.34
Indium Phosphide	InP	1.27
Indium Antimonide	InSb	0.18

Table (2.1): Energy Gap in typical semiconductdors

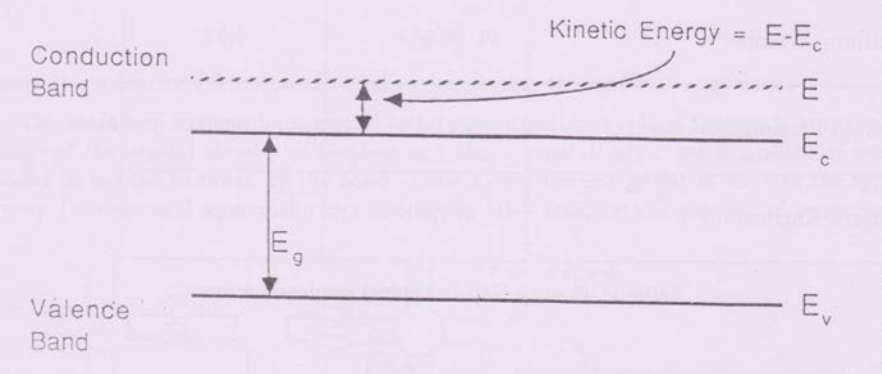


Figure (2.6): An electron in a state with energy E in the conduction band has a kinetic energy equal to $E - E_c$. The kinetic energy is small compared with the band gap energy $E_c - E_v$. This figure is not drawn to scale

Let us now calculate the density (concentration) of electrons in the conduction band. When an electron is in a state which is at the bottom of the conduction band i.e., a state with energy E_c, it has no velocity and therefore no kinetic energy. Hence, we can say that the electron in this state has only potential energy. Therefore the potential energy of the electron is taken as equal to E_c. The kinetic energy of the electron in a state in the conduction band with energy E is therefore equal to $E - E_c$ as shown in Figure (2.6). We have already seen that the conduction electrons can be considered nearly free, and therefore we can apply the techniques of free electron theory i.e., treat them as electrons in a box. The density of states in the conduction band close to the bottom of the band i.e., close to E_c, can be approximated by the density of states of free electrons or electrons in a box with one assumption. The mass of the electron is to be replaced by the effective mass determined by the curvature of the conduction band near E_c. Since only states that are nearly at the bottom of the conduction band are occupied, treating these electrons as free electrons subject to the assumption of the same effective mass for all electrons is valid. The momentum p of the conduction electrons can be readily written from the box model as

$$p = \sqrt{2m_c\,(E - E_c)} \tag{2.2}$$

The density of states for electrons in the conduction band is then given by

$$Z_c(E) = \frac{4\,\pi\,V\,(2m_c)^{\frac{3}{2}}}{h^3}(E - E_c)^{\frac{1}{2}} \tag{2.3}$$

where V is the volume of the semiconductor solid and the subscript c in $Z_c(E)$ has been used to denote the density of states in the conduction band. The density of states for electrons is shown in Figure (2.7).

The number of electrons dn_E, with energy between E and $E + dE$ per unit volume of the crystal, is then given by

$$dn_E = \frac{Z_c(E)f(E)dE}{V} = \frac{4\pi}{h^3}\frac{(2m_c)^{\frac{3}{2}}(E-E_c)^{\frac{1}{2}}dE}{(e^{\frac{E-E_F}{kT}}+1)} \tag{2.4}$$

Where $f(E)$ is the Fermi function.

The concentration (density) of free electrons can now be obtained by integrating Equation (2.4) between the limits E_c and E_{top} where E_{top} is the top of the conduction band i.e., the maximum energy of the conduction band.

$$n = \int_{E_c}^{E_{top}} dn_E$$

Typically, the concentration of free electrons is very low; hence the states in the lower portion of the conduction band will be occupied. Two further assumptions can then be made: one, E_{top} can be assumed to be infinite i.e., the upper limit of the integral can be assumed to be ∞ instead of E_{top} without any appreciable error, and two, the effective mass m_c is assumed to be the same for all the electrons since the curvature of the band is the same for all the states in the bottom of the conduction band. Hence m_c can be considered constant and taken outside the integral. Under these conditions, the density of electrons is given by

$$n = \frac{4\pi}{h^3}\,(2m_c)^{\frac{3}{2}}\int_{E_c}^{\infty}\frac{(E-E_c)^{\frac{1}{2}}dE}{(e^{\frac{E-E_F}{kT}}+1)} \tag{2.5}$$

This integral has to be evaluated numerically as it stands. However, some simplifications can be made. Let us assume that $E_c - E_F >> kT$ i.e., the Fermi energy lies in the band gap at least a few kT below the conduction band minimum E_c. Since the lower value that E can have is E_c, $E >> E_c$, and hence E-E_F >> kT. The equation for the electron density n can now be simplified by neglecting the unity term in the denominator to yield

$$n = \frac{4\pi}{h^3}\,(2m_c)^{\frac{3}{2}}\int_{E_c}^{\infty}(E - E_c)^{\frac{1}{2}}\,e^{\left(\frac{E_F-E}{kT}\right)}dE \tag{2.6}$$

Let us define a new variable $\eta = \frac{E-E_c}{kT}$. Then the above integral equation can be transformed as

$$\int_{E_c}^{\infty}(E - E_c)^{\frac{1}{2}}\,e^{\left(\frac{E_F-E}{kT}\right)}dE = (kT)^{\frac{3}{2}}e^{\left(\frac{E_F-E_c}{kT}\right)}\int\left(\frac{E-E_C}{kT}\right)^{\frac{1}{2}}e^{\left(\frac{E_C-E}{kT}\right)}\,d\,\left(\frac{E}{kT}\right)$$

$$= kT^{\frac{3}{2}} e^{\left(\frac{E_F - E_C}{kT}\right)} \int \eta^{\frac{1}{2}} e^{-\eta} \; d\eta \tag{2.7}$$

Hence

$$n = \frac{4\pi}{h^3} (2m_c)^{\frac{3}{2}} e^{\left(\frac{E_F - E_C}{kT}\right)} (kT)^{\frac{3}{2}} \int_0^{\infty} \eta^{\frac{1}{2}} e^{-\eta} d\eta \tag{2.8}$$

The integral in the right hand side of the above equation is just equal to $\frac{\sqrt{\pi}}{2}$. The equation for n can then be simplified to

$$n = 2 \left(\frac{2\pi m_c kT}{h^2}\right)^{\frac{3}{2}} e^{-\left(\frac{E_C - E_F}{kT}\right)} \tag{2.9}$$

The above equation is usually written as

$$n = N_c e^{-\left(\frac{E_C - E_F}{kT}\right)} \tag{2.10}$$

where

$$N_c = 2 \left(\frac{2\pi m_c kT}{h^2}\right)^{\frac{3}{2}} \tag{2.11}$$

This quantity N_c is called the effective density of states in the conduction band. The reason why N_c is called so is as follows: The factor $e^{-\left(\frac{E_C - E_F}{kT}\right)}$ in the expression for n, gives the probability that a state located at level E_c will be occupied, since it can be shown that when $E_c - E_F \gg kT$,

$$f(E_c) = \frac{1}{e^{\frac{E_C - E_F}{kT}} + 1} \approx e^{-\left(\frac{E_C - E_F}{kT}\right)} \tag{2.12}$$

If there are N_c states located at $E = E_{c,}$ then we will have for n the same expression as derived above.

We can similarly calculate the number of holes in the valence band per unit volume of the crystal. While we can calculate the density of states in the valence band by methods similar to what we did for the conduction band, we must use $\left(1 - f(E)\right)$ as the probability that a hole will exist in a state of energy E, which is the probability that an electron will not exist in that state.

We will consider the holes as free particles in a box. A hole at the top of the valence band has no velocity and therefore has no kinetic energy. As we go to lower energy states in the valence band, the velocity of the hole increases and therefore the kinetic energy increases. It is therefore possible to write the kinetic energy of the hole as $E_v - E$, where E_v is the energy of the top of the valence band. The density of states expression is the same as for electrons except for kinetic energy we write $E_v - E$ and is given by

$$Z_v(E) = \frac{4\pi V (2m_v)^{\frac{3}{2}}}{h^3} (E_v - E)^{\frac{1}{2}} \tag{2.13}$$

The density of states in the valence band is plotted in Figure (2.8).

We can calculate as we did for the electrons, the number of holes per unit volume of the crystal. We again assume that the Fermi Energy is at least a few kT above E_v , the valence band maximum. Then proceeding exactly similar to the case of electrons, we can show

$$p = 2\left(\frac{2\pi m_v kT}{h^2}\right)^{\frac{3}{2}} e^{-\left(\frac{E_F - E_v}{kT}\right)} = N_v e^{-\left(\frac{E_F - E_v}{kT}\right)} \tag{2.14}$$

where N_v is the effective density of states in the valence band, and is defined as

$$N_v = 2\left(\frac{2\pi m_v kT}{h^2}\right)^{\frac{3}{2}} \tag{2.15}$$

When the Fermi energy E_F is a few kT below E_c and also a few kT above E_v, the material is called a non-degenerate semiconductor. When this condition is not satisfied, the material is called a degenerate semiconductor.

We denote the Fermi energy in an intrinsic material E_i. Since $n_i = p_i$ according to Equation (2.1), we can write

$$N_c e^{-\left(\frac{E_c - E_i}{kT}\right)} = N_v e^{-\left(\frac{E_i - E_v}{kT}\right)} \tag{2.16}$$

And also

$$n_i^2 = n_i \times p_i = N_c N_v e^{-\left(\frac{E_c - E_i}{kT}\right)} e^{-\left(\frac{E_i - E_v}{kT}\right)} = N_c N_v e^{-\left(\frac{E_c - E_v}{kT}\right)} = N_c N_v e^{-\frac{E_g}{kT}} \tag{2.17}$$

This is an important result. This shows that the electron density (as well as the hole density) in an intrinsic material depends on the width of the energy gap (band gap). This result helps us to determine the location of E_i, the Fermi energy in an intrinsic material. If we multiply both sides of Equation (2.16) by $\frac{E_i}{kT}$ and rearrange terms we get

$$e^{\frac{2E_i}{kT}} = \frac{N_v}{N_c} e^{\left(\frac{E_c + E_v}{kT}\right)} \tag{2.18}$$

Taking logarithms of both sides and multiplying by $\frac{kT}{2}$, we get

$$E_i = \frac{kT}{2}\ln\left(\frac{N_v}{N_c}\right) + \frac{E_c + E_v}{2} \tag{2.19}$$

We notice first that $\frac{E_c + E_v}{2}$ represents an energy level exactly in the middle of the band gap. Secondly, if the effective mass of the electrons and that of the holes are nearly equal, the first term in the right hand side of the above equation is very nearly equal to 0. Hence we conclude that E_i, the Fermi energy in the intrinsic material, is nearly at the middle of the band gap.

Example

Let us now calculate the effective density of states in a semiconductor at room temperature (T= 300 ^{0}K). Let us assume that the effective mass m_c of the electron is the same as m_0, the mass of an electron in vacuum.

$$m_c \approx m_0 = 9.11 \times 10^{-31}\ kgm$$

$$k = 1.38 \times 10^{-23}\ \frac{Joule}{degree} K$$

$$h = 6.63 \times 10^{-34}\ \frac{Joule}{degree} K$$

Substituting these values in the expression for N_c we get

$$N_c = 2\left(\frac{2 \times \pi \times 9.11\,x\,10^{-31} \times 1.38\,x\,10^{-23} \times 300}{(6.63 \times 10^{-34})^2}\right)^{\frac{3}{2}}$$

$$= 2.51 \times 10^{25}\ meter^{-3} = 2.51 \times 10^{19}\ cm^{-3}$$

If we assume similarly that the effective mass of the hole m_v is the same as that of the electron in vacuum, we will get the same value for N_v.

We can now determine n_i, the density of electrons or holes, in the intrinsic semiconductor. By taking the square root of both sides of Equation (2.17) we can write

$$n_i = \sqrt{N_c N_v}\ e^{-\frac{E_g}{2kT}} \tag{2.20}$$

In order to express the carrier density at any temperature in terms of the number for the density of states that were calculated for T =300 ^0K, we rearrange the terms to yield

$$n_i = 2.51 \times 10^{19} \left(\frac{m_c m_v}{m_0^2}\right)^{\frac{3}{4}} \left(\frac{T}{300}\right)^{\frac{3}{2}} e^{-\frac{E_g}{2kT}}\ cm^{-3} \tag{2.21}$$

Where m_0 is the mass of an electron in vacuum. This expression can more readily be used to calculate n_i, the intrinsic carrier density at any temperature.

Example

Let us now calculate n_i, the intrinsic electron concentration in silicon at room temperature. Substituting the following values

$$m_c = 0.29\ m_0$$

$$m_v = 0.57\, m_0$$

$$E_g(T = 300\ K) = \ 1.12\ eV$$

We get for n_i

$$n_i = \sqrt{4x\left(\frac{2\pi \times 300 \times 1.38 \times 10^{-23} \times 9.1x10^{-31}}{(6.63 \times 10^{-34})^2}\right)^3 \times (0.29 \times 0.57)^{\frac{3}{2}} \times e^{-\left(\frac{1.12\times1.6\times10^{-19}}{2\times300\times1.38\times10^{-23}}\right)}}$$

$$= 2.54\ \times 10^9 cm^{-3}$$

We must point out that the above calculation is in error since it does not take into account the fact that there are six conduction bands in silicon and hence the effective density of states N_c that we calculated must be multiplied by a factor 6. Hence the value for n_i has to be multiplied by a factor $\sqrt{6}$. For our purpose, we approximate n_i as

$$n_i = 1.00\ \times\ 10^{10}\ cm^{-3}$$

We will use this value for n_i for silicon at room temperature throughout this book.

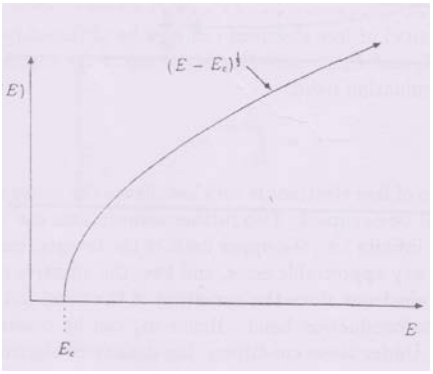


Figure (2.7): Density of states for electrons in the conduction band

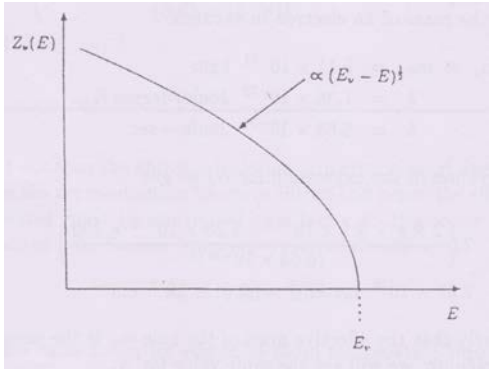


Figure (2.8): Density of states for holes in the valence band

Extrinsic Semiconductor

The properties of the intrinsic (i.e., pure) semiconductor can be altered by adding a small amount of impurity atoms to occupy the lattice sites originally occupied by the atoms of the intrinsic material. The semiconductor is then said to be extrinsic, since its properties depend on the externally added impurities: the impurities are said to be substitutionally added since they substitute the original atom in the lattice site.

Consider, for example, an element belonging to Group III in the Periodic Table which is trivalent, such as boron, aluminum, gallium or indium. Such an atom has three valence electrons and so, when it substitutes for a tetravalent atom of the semiconductor, the site into which it goes has only three of the four bonds completed and has the fourth bond incomplete. The boron atom is in the neutral state. This is illustrated in Figure (2.9 A), with boron as the impurity atom. As electron from one of the covalent bonds of a nearby silicon atom can break free from its bond and jump into the incomplete bond of the boron atom. When that happens, all the four bonds surrounding the boron atom are completed, and an electron (a negative charge) is bound to the boron atom as shown in Figure (2.9 A). The boron atom is now negatively ionized as shown in Figure (2.9 B). In this process, we have a bound electron (at the site of the boron atom) and a broken bond in a nearby semiconductor atom. The broken bond represents a free hole. Since the boron atom (and all other group III elements) accepts an electron, releasing a hole in the valence band, boron and other group III elements are called acceptor impurities. When the acceptor atom accepts an electron, it becomes ionized. Each ionized acceptor atom gives rise to a free hole.

If instead of a group III element, we substitute an atom of the elements of group V which is penta-valent, such as phosphorous, we will then have one electron more than the number needed to complete the four covalent bonds (tetra-valent bond) as shown in Figure (2.10 A). This fifth electron requires very little energy to get freed from the parent phosphorous atom; when it is freed, we have a free electron in the crystal and an ionized phosphorous atom as shown in Figure (2.10 B). We can think of the removal of the fifth electron as a positive charge bound to the site of the phosphorous atom. The bound positive charge in the phosphorous atom donates a free electron. It is called a donor impurity. A donor atom donates an electron when it is ionized. Each ionized donor atom gives rise to a free electron.

In Table (2.2), we give the ionization energy (energy required to free an electron from the donor atom, or to free a hole from the acceptor atom) for some of the common impurity atoms in germanium and silicon.

Table (2.2 A): Ionization Energy in eV of Donor Atoms

	Impurity	Si	Ge
	P	0.044	0.012
4	As	0.049	0.013
	Sb	0.039	0.096
	Bi	0.067	

Table (2.2 B): Ionization Energy in eV of Acceptor Atoms

	Impurity	Si	Ge
	B	0.045	0.01
4	Al	0.057	0.01
	Ga	0.067	0.11
	In		0.11

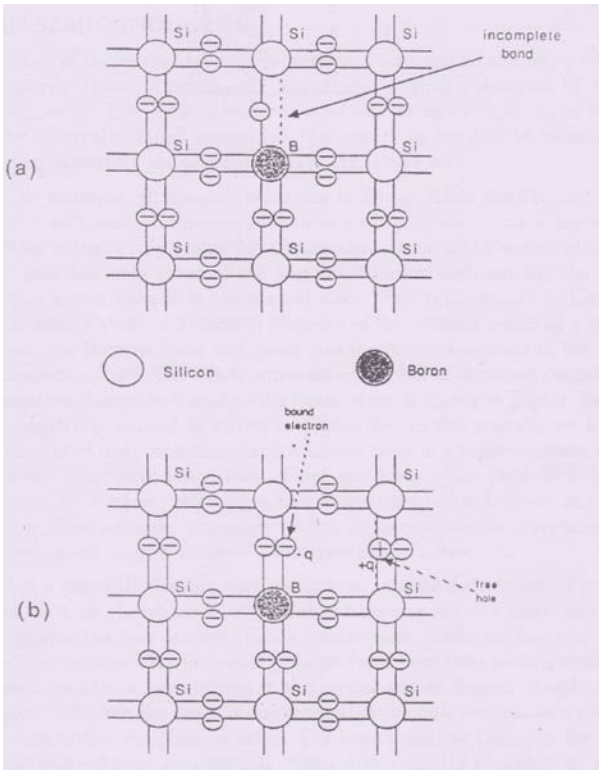


Figure (2.9): Substitution of trivalent boron for one of the tetravalent silicon: A)Boron atom is neutral and an electron is missing from the covalent bond. B) An electron breaks loose from the covalent bond of a neighboring silicon atom and jumps into the covalent bond of the boron atom in which an electron was missing. The boron atom is now negatively charged. The broken covalent bond in the neighboring silicon atom can be thought of as a free hole.

From Table (2.2), it can be seen that the ionization energy is approximately of the same order for all the different impurities in a given semiconductor. In the case of a neutral donor atom, a single (the fifth) electron is orbiting around a positive charge in a space characterized by the dielectric constant of the semiconductor. It is like a hydrogen atom, which can be described by the Bohr model. We can therefore substitute the value of permittivity of the semiconductor for the permittivity of free space in the expression for the ionization energy of the hydrogen atom to obtain the ionization energy of the donor atom. We can treat the acceptor atom similarly, where we assume a positive charge to be revolving around a fixed negative charge. It is therefore usual to assume that the energy level of the acceptor atoms is typically about 0.05 eV above the top of the valence band, and that the energy level

of the donor atoms is typically about 0.05 eV below the bottom of the conduction band in silicon. This is illustrated in Figure (2.11).

In a neutral material, the charge neutrality condition should hold. The magnitude of negative charge per unit volume should be equal to the amount of positive charge per unit volume. If in a material there are N_D donor atoms in unit volume of the crystal, the amount of positive charge per unit volume is equal to

$$q(p + N_D^+)$$

and the magnitude of negative charge per unit volume is equal to

$$qn$$

Equating the two we get

$$n = p + N_D^+ \tag{2.23}$$

Where N_D^+ is the number of positively ionized donor impurities per unit volume. We have free electrons due to two reasons: (1) thermal excitation of the valence electrons, and (2) donor ionization. In the above equation, the first term on the right hand side represents the density of electrons due to thermal excitation from the valence band and the second term represents the density of electrons due to ionization of donor atoms.

Since both electrons and holes carry a charge, they are called charge carriers or, more simply, carriers. In a material in which we have donor type of impurities we have more electrons than holes per unit volume since, from the above equation, it can be inferred that *n* is larger than *p*. For this reason, the electrons are called majority carriers, and the holes are called minority carriers. For the same reason, this material is called a *n*-type semiconductor.

If we have a material with only acceptor type of impurities of density N_A, then charge neutrality requires

$$p = n + N_A^- \tag{2.24}$$

where N_A^- is the number of negatively ionized acceptor atoms per unit volume. Free holes are created due to two reasons: (1) thermal excitation of valence electrons, and (2) acceptor ionization. As before, the first term on the right hand side of the above equation represents the density of holes due to thermal excitation and the second term the density of holes due to ionization of acceptor atoms. Holes are majority carriers in this material, and electrons are minority carriers. This type of material is called *p*-type material. To denote the carrier density and other parameters in an extrinsic material, we use a subscript *p* in *p*-type material, and a subscript *n* in *n*-type material.

In practice, the purest semiconductors prepared in the laboratory have an impurity ration of 1 in 10^{10} or 10^{11}. This corresponds to approximately 10^{12} or 10^{11} impurity atoms per cubic centimeter. Therefore, the acceptor and donor impurities that are added should be greater than 10^{12} per cubic centimeter to sensibly alter the properties of this material.

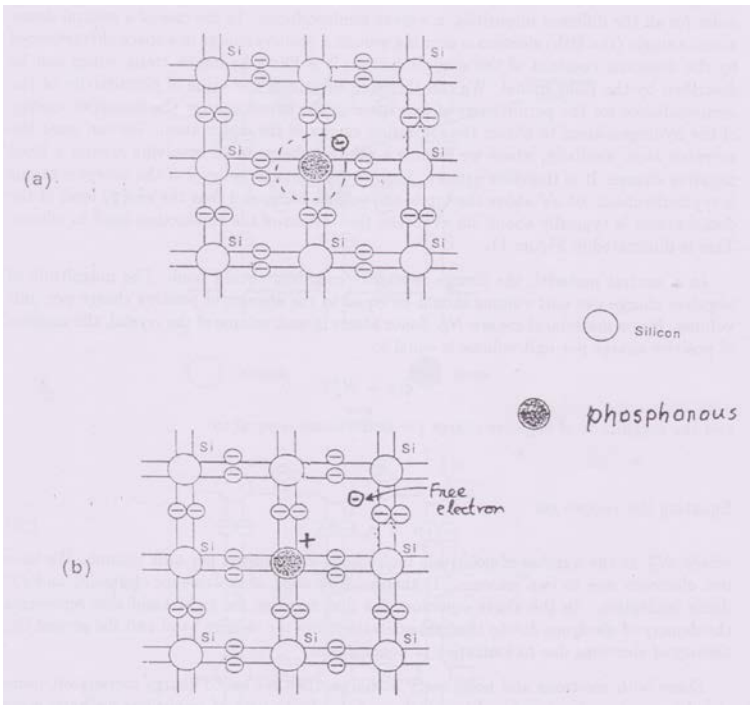


Figure (2.10): Substitution of a pentavalent phosphorus atom for one of the tetravalent silicon: a) Phosphorus atom is neutral and there is an extra electron orbiting around the phosphorous atom similar to the electron in the hydrogen atom b) The orbiting electron breaks loose due to the imparting of a small amount of energy. The phosphorous atom is positively ionized. The electron is now free to move around in the crystal and is called the free electron.

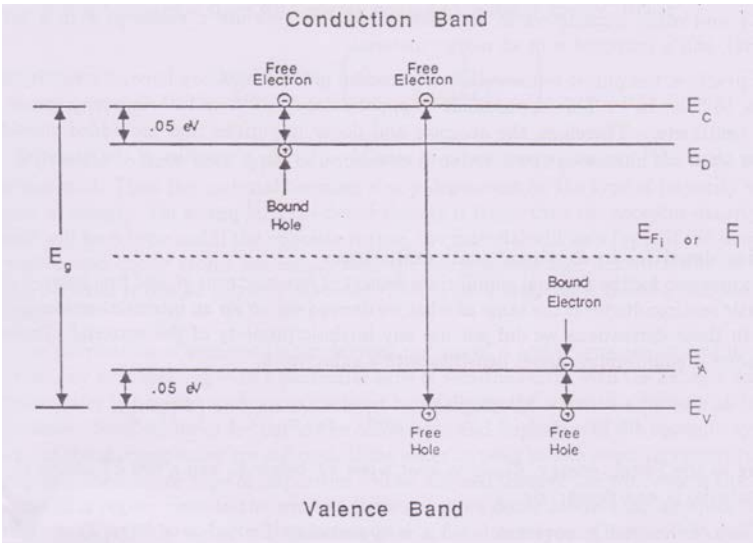


Figure (2.11): Impurity energy levels in silicon and generation of free carriers.

Carrier densities in extrinsic semiconductors

The expression for the thermal equilibrium density of free electrons n and free holes p in an extrinsic semiconductor is the same as what we derived earlier for an intrinsic semiconductor, since in these derivations we did not use any intrinsic property of the material. Denoting the thermal equilibrium carrier densities with a subscript 0,

$$n_0 = N_c e^{-\left(\frac{E_c - E_i}{kT}\right)} \quad and \quad p_0 = N_v e^{-\left(\frac{E_i - E_v}{kT}\right)} \tag{2.25}$$

as long as the Fermi energy, E_F , is at least a few kT below E_c and a few kT above E_v i.e., the material is non-degenerate.

When the impurity concentration N_D, is approximately equal to or larger than N_c in an n-type material, or when the impurity concentration N_A is equal to or greater than N_V in a p-type material, E_F is closer to E_c or E_v by less than a few kT, and the material is said to be degenerate material. In these cases the majority carrier density is given by Equation (2.5) for electrons or a similar one for holes and not by the simplified expressions in Equation (2.25).

The expression for the center densities in Equation (2.25) is the same for both the intrinsic and the non-degenerate extrinsic material. The difference between the values of the carrier concentrations in intrinsic and extrinsic semiconductors arises due to the difference in the positions of the Fermi level in the two cases. Also we note, as before, that in the non-degenerate semiconductor,

$$n_0\, p_0 = N_c N_v e^{-\left(\frac{E_g}{kT}\right)} = n_i^2 \tag{2.26}$$

This relationship is called the law of mass action which is valid even in the extrinsic material as long as it is non-degenerate.

While N_D and N_A are the impurity concentrations in n and p type materials respectively, N_D^+ and N_A^- are respectively the ionized impurity densities. The probability that an acceptor atom will be ionized is the probability that an electron will occupy a state with energy equal to E_A and this is given by the Fermi Dirac statistics. Therefore[1],

$$N_A^- = \frac{N_A}{1 + e^{\frac{E_A - E_F}{kT}}} \tag{2.27}$$

Similarly the probability that donor atom will be ionized is equal to the probability that an electron will not occupy a state with energy equal to the donor energy. Hence[2],

$$N_D^+ = N_D \left(1 - \frac{1}{e^{\frac{E_D - E_F}{kT} + 1}}\right) \tag{2.28}$$

[1] In reality the expression for N_A^- should include a factor in the denominator called the degeneracy factor but we will ignor that here for the purpose of simplicity.

[2] Again we have neglected the degeneracy factor but it will not introduce any serious error.

It is possible to have both types of impurities simultaneously present in the same region of the material. Then the material becomes n or p dependent on the type of impurity which is higher in density. For example, if the donor density is larger than the acceptor density, the material will be n type and if the opposite is true, the material will be p type. If the densities of acceptor and donor atoms are equal, then the electron and hole densities will be equal and the material is like an intrinsic material. This type of material is called a compensated material.

In the fabrication of semiconductor devices, regions of n and p type are formed in the same wafer by first starting with a particular type of semiconductor with the lightest doping. Then the wafer is covered with an oxide layer by oxidizing the wafer in a furnace at high temperature. Next windows are cut in the oxide layer and impurities of the opposite type to the type of the starting wafer are diffused. If the wafer initially had an impurity concentration of acceptors, then donor type of impurities will be diffused through the windows in the oxide such that in a region close to the window there are more donor atoms than acceptor atoms i.e., $N_D > N_A$. This region will therefore have a net donor type of impurities of density $N_D - N_A$, and hence will be n-type. The majority carrier density will be equal to the net donor density. We have to use the net donor density in the calculation of the carrier densities instead of the total donor density. If on the other hand, N_A is greater than N_D, the material will be p-type and we will use the net acceptor density $N_A - N_D$ in the calculation of the carrier density instead of the total acceptor density.

n-type material:

Let us consider a non-degenerate n-type semiconductor. Using the subscript n for the n-type material and the subscript 0 for the thermal equilibrium, we have

$$n_{n0} = N_D^+ + p_{n0}$$

Hence

$$N_c e^{-\left(\frac{E_c - E_i}{kT}\right)} = N_D \left(1 - \frac{1}{e^{\frac{E_D - E_F}{kT}} + 1}\right) + N_v e^{-\left(\frac{E_F - E_v}{kT}\right)} \qquad (2.29)$$

The above expression is difficult to solve, and has to be solved numerically. At absolute zero temperature, no donor atom is ionized and no free carriers exist. Therefore, $n_{n0} = 0$. The Fermi energy lies halfway between E_c and E_D, since a slight increase in the temperature will give rise to an equal number of electrons in the conduction band and ionized donor atoms. As the temperature is gradually raised, more and more donor atoms are ionized, and so the Fermi level moves downward. When the Fermi level is between E_c and E_D, the material is said to be in the freeze-out region. In this range of temperatures, only some of the donor atoms are ionized. This is referred to as partial ionization of the donor atoms. When the Fermi energy is close to E_D, such that the electrons in the conduction band are essentially due to ionization of donor atoms, the material is said to be in the extrinsic region. When the Fermi level goes way below E_D as the temperature is increased further, the material is said to be in the exhaustion region, since in this case the donor energy lies above the Fermi energy and all the donor atoms are ionized. No increase in electron density can arise due to further ionization of donor atoms, since all the donor atoms are already ionized. Any further increase of electron density can arise only due to thermal excitation of the valence electrons and this happens when the temperature is increased to a

much higher value. The Fermi level decreases with temperature and asymptotically approaches the intrinsic Fermi level E_i at high temperature (n_{n0} becomes equal to p_{n0}). At high temperatures the material becomes intrinsic. It can be seen, then that if we want to take advantage of the properties brought about by external impurities in the semiconductor, we must limit the operating temperature so that the material is used in the extrinsic, or exhaustion range. The variation of Fermi level with temperature is shown in the upper half of Figure (2.12) for an n-type material for various concentrations of impurity atoms. The effect of increasing the concentration of impurities is to extend the extrinsic and exhaustion range of temperatures to higher values.

The minority carrier densities in a semiconductor can be obtained from the law of mass action (Equation (2.26)). In an n-type material the thermal equilibrium minority carrier density is therefore obtained as

$$p_{n0} = \frac{n_i^2}{n_{n0}} \tag{2.30}$$

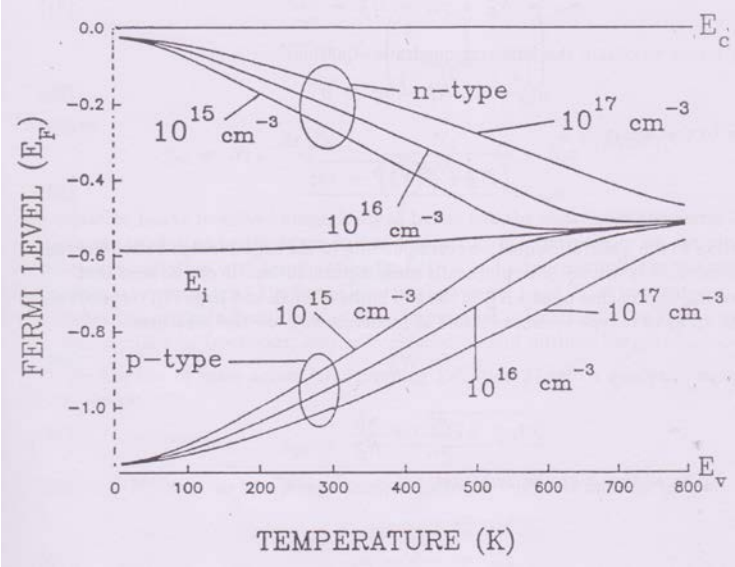


Fig. (2.12): Variation of Fermi Level with Temperature for n- and p-type semiconductor

From Equation (2.23) (charge neutrality expression in an n-type material), we get

$$n_{n0} = N_D^+ + p_{n0} = n_{n0} = N_D^+ + \frac{n_i^2}{n_{n0}} \tag{2.31}$$

Rearranging terms we obtain the following quadratic equation:

$$n_{n0}^2 - n_{n0}N_D^+ - n_i^2 = 0 \tag{2.32}$$

The solution to this equation is

$$n_{n0} = \frac{N_D^+ + \sqrt{\left(N_D^+\right)^2 + 4\, n_i^2}}{2} \tag{2.33}$$

This solution to the quadratic equation corresponding to the negative sign before the radical is neglected since it will not give physically meaningful solution. It can be seen that this expression for n_{n0} approaches n_i and N_D^+ in the two limits of small and large N_D^+ respectively. When $N_D^+ >> n_i$, as in n-type semiconductors at moderate and low temperatures,

$$n_{n0} \approx N_D^+ \tag{2.34}$$

and

$$p_{n0} = \frac{n_i^2}{n_{n0}} = \frac{n_i^2}{N_D^+} \tag{2.35}$$

When $n_i >> N_D^+$ as at very high temperatures,

$$n_{n0} \approx n_i$$

$$p_{n0} = \frac{n_i^2}{n_{n0}} = n_i \tag{2.36}$$

Example

Let us calculate the carrier densities in a n-type material in which the donor density is 10^{14} cm^{-3}. Assume all the donor atoms are ionized at room temperature.

$$n_{n0} = \frac{N_D^+ + \sqrt{\left(N_D^+\right)^2 + 4\, n_i^2}}{2} = \frac{10^{14} + \sqrt{10^{28} + 4 \times 10^{20}}}{2} \approx 10^{14} \text{ cm}^{-3}$$

and

$$p_{n0} = \frac{n_i^2}{n_{n0}} = \frac{10^{20}}{10^{14}} = 10^6 \text{ cm}^{-3}$$

p-type material:

Let us now consider a p-type material. Using the subscripts p for the p type material and 0 for thermal equilibrium in the charge neutrality equation (Equation (2.24)) we get

$$p_{p0} = N_A^- + n_{p0} \tag{2.37}$$

where N_A^- is the density of ionized acceptors. But

$$N_A^- = N_A \left(\frac{1}{e^{\frac{E_A - E_F}{kT}} + 1} \right) \tag{2.38}$$

Therefore

$$p_{p0} = N_v e^{-\left(\frac{E_F - E_v}{kT}\right)} = \left(\frac{N_A}{e^{\frac{E_A - E_F}{kT}} + 1}\right) + N_c e^{-\left(\frac{E_C - E_F}{kT}\right)} \tag{2.39}$$

This equation has to be solved numerically as before and the variation of the Fermi level with temperature is plotted in the lower half of Figure (2.12). For convenience we are plotting the Fermi energy variation with temperature for both n and p type materials in the same figure. In the pure p-type material, the Fermi level lies between E_A and E_v at low temperatures, and approaches the intrinsic Fermi level as the temperature is increased. The p-type material is also characterized by freeze-out, extrinsic, exhaustion and intrinsic ranges of temperatures.

Using the law of mass action, transforming Equation (2.37) into a quadratic equation and solving, we get

$$p_{p0} = \frac{N_A^- + \sqrt{\left(N_A^-\right)^2 + 4\, n_i^2}}{2} \tag{2.40}$$

As before, if $N_A^- >> n_i$ as in a p-type semiconductor at moderate and low temperatures,

$$p_{p0} \approx N_A^- \tag{2.41}$$

and

$$n_{p0} = \frac{n_i^2}{p_{p0}} \approx \frac{n_i^2}{N_A^-} \tag{2.42}$$

When $n_i \gg N_A^-$ as at very high temperatures,

$$p_{p0} \approx n_i \tag{2.43}$$

and

$$n_{p0} = \frac{n_i^2}{p_{p0}} = n_i \tag{2.44}$$

Determination of the Fermi Energy

We will now derive an expression for the Fermi energy location in the band gap for an extrinsic semiconductor. As stated before the material is assumed to be a non-degenerated semiconductor.

n-type material.

In the non-degenerate n-type semiconductor, the majority carrier density can also be expressed from Equation (2.25) as

$$n_{n0} = N_c e^{-\left(\frac{E_C - E_F}{kT}\right)}$$

Dividing both sides by N_c and taking logarithms of each side and rearranging terms, we get

$$E_F = kT\ ln\left(\frac{n_{n0}}{N_c}\right) + E_c \tag{2.45}$$

Since $n_{n0} \approx N_D^+$

$$E_F = kT\ ln\left(\frac{N_D^+}{N_c}\right) + E_c \tag{2.46}$$

Since N_D^+ is less than N_c in the non-degenerate semiconductor, the term $\ln\left(\frac{N_D^+}{N_c}\right)$ is negative. Hence E_F is below E_C by an amount equal to $kT\ ln\left(\frac{N_D^+}{N_c}\right)$. This is illustrated in Figure (2.13 A). At room temperature all the donor atoms can be assumed to be ionized and hence $N_D^+ = N_D$.

Therefore,

$$E_F = kT\ ln\left(\frac{N_D}{N_c}\right) + E_C \tag{2.47}$$

p-type material

The majority carrier density in the non-degenerate p-type semiconductor can be expressed using Equation (2.25) as

$$p_{p0} = N_v e^{\left(\frac{E_v - E_F}{kT}\right)}$$

Proceeding as we did in the case of the n-type material, we get

$$E_F = kT\ ln\left(\frac{N_V}{p_{p0}}\right) + E_v \tag{2.48}$$

But $p_{p0} = N_A^-$. Hence,

$$E_F = kT\ ln\left(\frac{N_v}{N_A^-}\right) + E_v \tag{2.49}$$

The Fermi Energy is above the valence band in a p-type semiconductor by an amount equal to $kT\ ln\left(\frac{N_V}{N_A^-}\right)$ as illustrated in Figure (2.13 B). At room temperature all the acceptor atoms are assumed to be ionized. Hence,

$$N_A^- = N_A$$

Therefore,

$$E_F = kT\ ln\left(\frac{N_v}{N_A}\right) + E_v \tag{2.50}$$

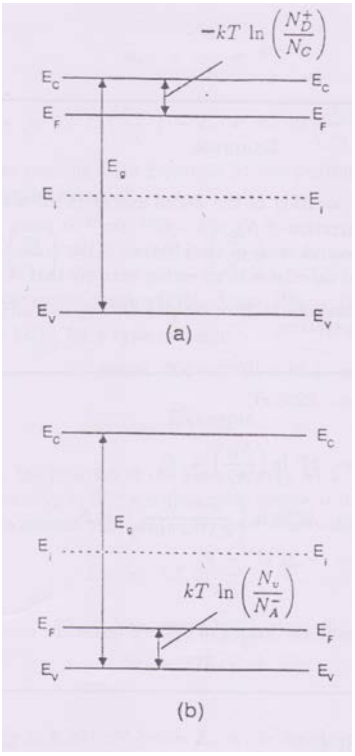


Figure (2.13): Location of the Fermi energy in an extrinsic semiconductor: (A) n-type material (B) p-type material.

Example

Let us now determine the location of the Fermi energy in an n-type material with a net impurity concentration of $N_D = 5 \times 10^{15}$ cm^{-3} at room temperature. Let us assume that the effective mass of the electron is the same as that of the electron in vacuum. We had calculated in an earlier example that N_C under these assumptions is equal to 2.51 x 10^{19} cm^{-3}. All the donor atoms can be assumed to be ionized at room temperature.

$$kT = 1.38 \text{ x } 10^{-23} \text{ x } 300 \quad \text{Joules}$$

$$= 0.02589 \text{ eV}$$

$$E_F = kT\ ln\left(\frac{N_D}{N_c}\right) + E_C$$

$$= 0.0258 \ln\left(\frac{5 \times 10^{15}}{2.51 \times 10^{19}}\right) + E_C$$

$$= -0.22\ eV + E_C$$

The Fermi energy is located in the band gap 0.22 eV below the conduction band.

Carrier Densities in terms of E_i

Starting from the expression for the thermal equilibrium carrier densities n and p in terms of the Fermi energy as given in Equation (2.25), it is easy to show (left as a homework exercise) that

$$n_{x0} = n_i e^{\left(\frac{E_F - E_i}{kT}\right)} \tag{2.51}$$

and

$$p_{x0} = n_i e^{-\left(\frac{E_F - E_i}{kT}\right)} \tag{2.52}$$

where E_i stands for the intrinsic Fermi energy and the subscript x stands for n in a n-type material and for p in a p-type material. Therefore

$$E_F = kT\ ln\left(\frac{n_{x0}}{n_i}\right) + E_i \tag{2.53}$$

and

$$E_F = kT\ ln\left(\frac{n_i}{p_{x0}}\right) + E_i \tag{2.54}$$

In an n-type material,

$$n_{x0} = n_{n0} = N_D^+ \tag{2.55}$$

Hence

$$E_F = kT\ ln\left(\frac{n_{n0}}{n_i}\right) + E_i = kT\ ln\left(\frac{N_D^+}{n_i}\right) + E_i \tag{2.56}$$

It is trivial to show that starting from Equation (2.54) and putting $x = n_i$, we will get the same result. Similarly in a p-type material,

$$E_F = kT\ ln\left(\frac{n_i}{p_{p0}}\right) + E_i = kT\ ln\left(\frac{n_i}{N_A^-}\right) + E_i = E_i - kT\ ln\left(\frac{N_A^-}{n_i}\right) \tag{2.57}$$

The location of the Fermi energy with respect to E_i is illustrated in Figure (2.14 A) for n type material and in Figure (2.14 B) for p type material.

Example

Let us determine the location of the Fermi energy with respect to E_i in a p-type semiconductor with $N_A = 10^{15}$ cm^{-3} at room temperature. At room temperature it is reasonable to assume that all the acceptor atoms are ionized, i.e., $N_A^- \approx N_A$.

$$E_F = kT\ ln\left(\frac{n_i}{N_A}\right) + E_i$$

$$= 0.0259\ ln\left(\frac{10^{10}}{10^{15}}\right) + E_i$$

$$= -0.297\ eV + E_i$$

The Fermi energy is 0.2976 eV below E_i in the bandgap. This is illustrated in Figure (2.14).

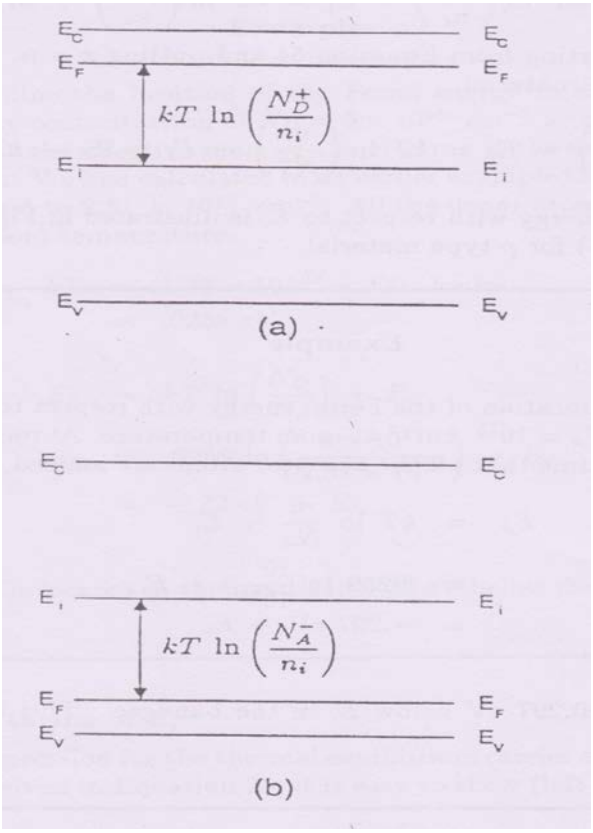


Figure (2.14): Fermi energy in terms of the intrinsic Fermi energy: (A) n-type material (B) p-type material

Electric Current in Semiconductors

We will now study the flow of electric current in semiconductors. According to thermodynamic treatments, the Fermi energy is equal to the sum of chemical potential (energy) and the internal electrostatic potential energy. The electrochemical potential ς for electrons is given by

$$\varsigma = \mu - q\psi$$

where μ is the chemical potential and ψ is the electrostatic potential. Thus we see that the Fermi energy is the electrochemical p[otential in a solid in which equilitbrim conditions exist i.e., no externally applied electric field or no radiation is present. When we apply a voltage across a piece of semiconductor, the

thermal equilibrium conditions are disturbed. Even then, as long as there are no excess carriers, the Fermi energy is equal to the electrochemical potential. The electrochemical potential is different at the two ends of the semiconductor across which the voltage is applied. The difference is equal to the amount of the applied voltage multiplied by the electron charge. Hence, the Fermi level at one end of the semiconductor is shifted in energy with respect to the Fermi level at the other end by q times the applied voltage as shown in Figure (2.15). Since the material is uniform i.e., the carrier density is the same everywhere, the conduction band and the valence band vary in direct accordance with the variation in the Fermi energy. In other words, the conduction band and the valence band are shifted from one end of the crystal to the other by the amount $-qV$ where V is the applied voltage. The electric field is the negative gradient of the electrostatic potential ψ.

$$\vec{\mathcal{E}} = -\vec{\nabla}\psi \tag{2.58}$$

In one dimension

$$\mathcal{E} = -\frac{d\psi}{dx} \tag{2.59}$$

The electrostatic potential ψ is the potential energy divided by charge. We saw earlier that $E_{c,}$ the bottom of the conduction band, is the potential energy of the electron. The electrostatic potential is obtained by dividing by $-q$, and is therefore equal to

$$\psi = -\frac{E_c}{q} = -\frac{E_i}{q} - \frac{E_g}{2q} \tag{2.60}$$

and

$$\mathcal{E} = -\frac{d\psi}{dx} = \frac{1}{q}\frac{dE_c}{dx} = \frac{1}{q}\frac{dE_i}{dx} \tag{2.61}$$

The energy band diagram in the presence of an electric field is represented in Figure (2.15). The electric field is shown as being in the $-x$ direction in this figure. Hence the gradient, according to Equation (2.61) is also negative and hence the bands are tilted downward to the right. The slope is constant since the electric field is constant.

We saw earlier that in the presence of an electric field, the k values and the momentum p values change by an amount proportional to the external force under steady state.

$$\Delta k = \frac{Force}{\hbar}\,\tau_c = \frac{-q\mathcal{E}}{\hbar}\tau_c \tag{2.62}$$

Where τ_c is the scattering relaxation time and Force is the externally applied force due to the electric field. The change in momentum correspondingly is

$$\Delta p = \hbar\Delta k = -q\mathcal{E}\,\tau_c \tag{2.63}$$

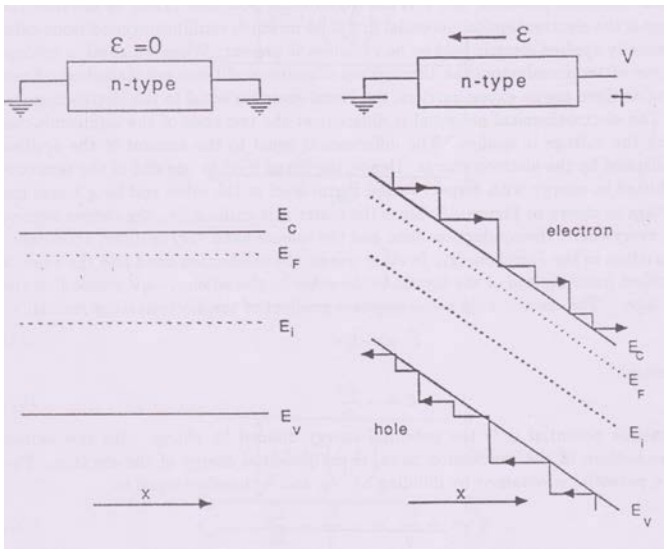


Figure (2.15): Energy band diagram in the presence of an electric field

Before the application of the electric field, the average momentum was zero since for every electron having a particular value of momentum, there was another electron having the opposite momentum. In the presence of the electric field, all the electrons have the same net change in momentum given by Equation (2.63). Since all the electrons suffer the same net change in momentum, the average momentum of the electrons is equal to the net change, and is given by

$$p_{av} = -q\mathcal{E}\,\tau_c \tag{2.64}$$

Since the momentum is given by $p = m^* v$ where m^* is the effective mass of the electron and v is its velocity, the average velocity, which is also called the drift velocity v_d is given by

$$v_d = \frac{p_{av}}{m^*} = -\frac{q\mathcal{E}\,\tau_c}{m^*} \tag{2.65}$$

Equation (2.65) shows that the drift (average) velocity is proportional to the electric field $\mathcal{E}$. This is usually expressed as

$$v_d = -\mu\,\mathcal{E} \tag{2.66}$$

The proportionality constant μ is called the mobility of the electron and is given in units of *velocity per unit electric field*. Typically we express mobility in units of $cm^2V^{-1}sec^{-1}$. Since we will be discussing the drift velocity of electrons as well as that of holes, we denote the parameters with a subscript n for electrons and p for holes. Therefore, for electrons

$$v_{dn} = -\mu_n\,\mathcal{E} \tag{2.67}$$

and

$$\mu_n = \frac{q\ \tau_{cn}}{m_n} \tag{2.68}$$

For holes we must bear in mind that the charge is positive and therefore the externally applied force is $q\mathcal{E}$. Hence,

$$v_{dp} = \frac{q\mathcal{E}}{m_p}\,\tau_{cn} = \mu_n\,\mathcal{E} \tag{2.69}$$

and

$$\mu_p = \frac{q\,\tau_{cp}}{m_p} \tag{2.70}$$

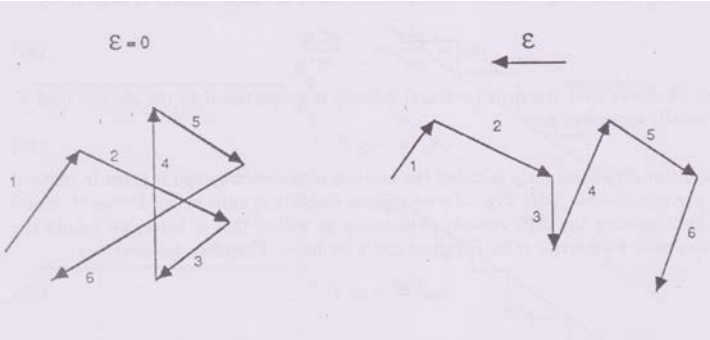


Figure (2.16): Motion of an Electron with and without an Electric Field

In the above equations, m_n and m_p are the effective masses for electrons and holes. The relaxation time due to scattering τ_c is also different for electrons and holes; hence different subscripts are also used to denote the relaxation time. Figure (2.16) shows on the left the random motion of the electron in the absence of an electric field and the average velocity is zero. On the right, the motion of the electrons has a drift velocity v_{dn}superimposed on the random motion in the presence of the electric field. If the charge carriers viz., electrons and holes, have a non-zero drift velocity, then an electric current flows. The particle flux density can be determined as follows: Since all the particles are moving with the same velocity v_{dn}, the particles passing through a unit cross sectional area in the next one second will be only those that are within a distance numerically equal to $v_{dn}$and they are equal to

$$Particle\ Flux\ density = n\,v_{dn} \tag{2.71}$$

The electric current density is obtained by multiplying the particle flux density by the charge –q and hence the electric current density due to electron flow is

$$J_n = -qnv_{dn} = q\,n\,\mu_n\mathcal{E} \tag{2.72}$$

The current flow due to the electric field is called drift current. Similarly, the electric current density due to hole flow is given by

$$J_p = -qnv_{dp} = q\,n\,\mu_p\mathcal{E} \tag{2.73}$$

The total drift current density is $J = J_n + J_p$, Hence

$$J = q\, n\, \mu_n \mathcal{E} + q\, n\, \mu_p \mathcal{E}$$

The total electric current density is proportional to the applied electric field. This is called Ohm's Law. The electrical conductivity is the electric current density per unit electric field and is given by

$$\sigma = \frac{J}{\mathcal{E}}$$

Therefore

$$\sigma = q\,(n\,\mu_n + p\,\mu_p) \tag{2.74}$$

The resistivity ρ is the reciprocal of σ.

$$\rho = \frac{1}{\sigma} = \frac{1}{q\,(n\,\mu_n + p\,\mu_p)}$$

It must be pointed out that the drift current is essentially determined by the majority carriers and the contribution of the minority carriers is negligible since the majority carrier density is orders of magnitude larger than the minority carrier density.

Interpretation of τ_{cn} and τ_{cp}

The mobility defines the ease with which the carriers move in the semiconductor under the application of an external field. It depends primarily on the effective mass of the carrier and the relaxation time $\tau_{c.}$ τ_{cn} can be interpreted as the average or mean time between scattering or collisions suffered by the electron with a similar interpretation for τ_{cp}. Since the random motion of the electron is due to thermal velocity, the average or mean distance traveled by the electron between collisions, which is denoted l_n, is given by

$$l_n = v_{th}\, \tau_{cn} \tag{2.75}$$

Where v_{th} is the average of the magnitude of thermal velocity. The parameter l_n is called the mean free path. $1/\,\tau_{cn}$ can be interpreted as the probability per unit time that an electron will be scattered. Collisions or scattering of the electrons can be due to several mechanisms such as lattice (phonon) scattering and impurity scattering. Therefore when several scattering mechanisms are simultaneously present, the probabilities of scattering due to each of the mechanisms are added.

$$\frac{1}{\tau_{cn}} = \frac{1}{\tau_{n\,ph}} + \frac{1}{\tau_{n\,imp}} \tag{2.76}$$

where $\frac{1}{\tau_{n\,ph}}$ is the probability per unit time of scattering due to lattice vibrations or phonons and $\frac{1}{\tau_{n\,imp}}$ is the probability per unit time of scattering due to impurities. Therefore,

$$\frac{1}{\mu_n} = \frac{1}{\mu_{n\,ph}} + \frac{1}{\mu_{n\,imp}} \tag{2.77}$$

Here $\mu_{n\,ph}$ and $\mu_{n\,imp}$ are respectively the mobilities that the electron would have had if only scattering due to lattice vibrations alone was present or if only scattering due to impurities alone was present. The two scattering mechanisms have different temperature dependence. For example,

$$\mu_{n\,ph}\ \alpha\ T^{-\frac{3}{2}}$$

and

$$\mu_{n\,imp}\ \alpha\ \left(T^{\frac{3}{2}}\right)\frac{1}{N_{imp}}$$

Similar arguments apply for hole mobility.

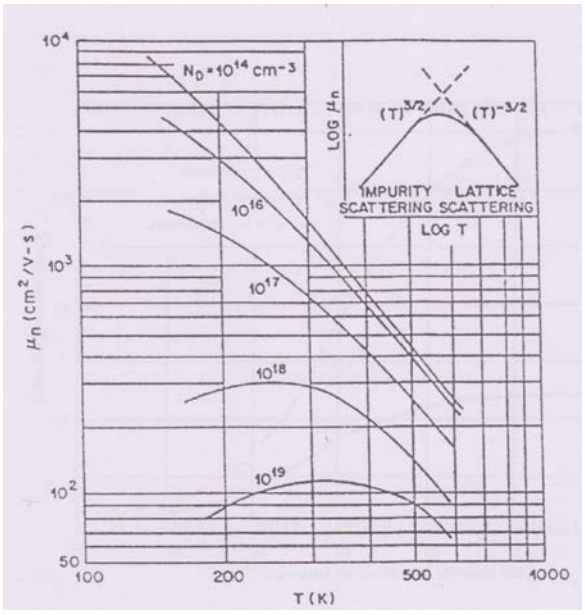


Figure (2.17): Electron mobility versus temperature in silicon (From Sze)

Figure (2.17) gives the variation of electron mobility with temperature for different donor concentrations. The mobility at low temperatures is limited by impurity scattering. The impurity scattering probability decreases with an increase in temperature and the mobility therefore increases with temperature. At higher temperatures the lattice or phonon scattering probability becomes significant and the mobility starts to fall off with temperature since the scattering probability due to lattice vibrations increases with temperature. The mobility thus exhibits a peak at some temperature where the influence of the two scattering mechanisms is comparable. The peak shifts to the right and falls off also in height as the impurity concentrations increase since the probability of scattering due to impurity scattering increases with donor concentration.

Figure (2.18) gives the variation of mobility with impurity concentration. The mobility is high and fairly constant with impurity density until about 10^{16} cm^{-3} after which it falls to a lower value. The figure

gives both the electron and the hole mobility variation with the total impurity concentration. In compensated materials the mobility should be determined from this figure using the total density of impurities (the sum of donor and acceptor impurity densities). The resistivity at room temperature for silicon for different donor or acceptor concentrations is given in Figure (2.19). Once we know the resistivity, we know the impurity concentration. Similarly when we know the impurity concentration, we know the resistivity.

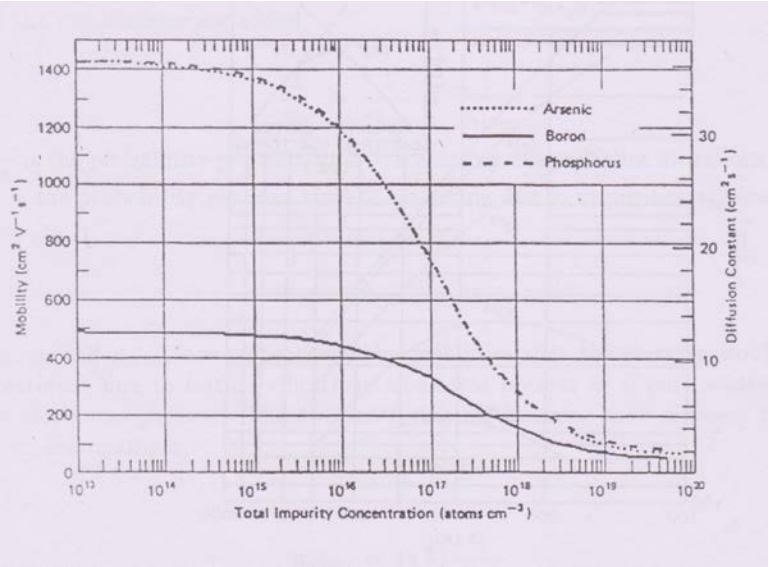


Figure: (2.18) Electron and Hole mobility variation with impurity concentration in silicon at T = 300 ^{0}K (From Mueller and Kamins)

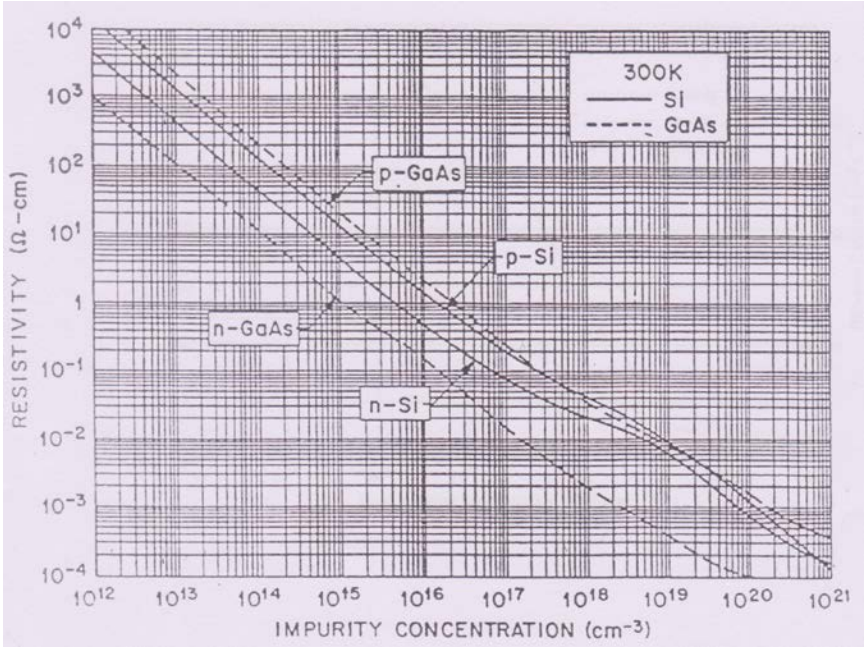


Figure (2.19): Resistivity variation with dopant concentration in silicon. (From Sze)

Diffusion

An electric current can also flow in a semiconductor due to a variation of the density of electrons and holes. The process by which the concentration variation produces carrier flow is called diffusion. In a diffusion process the particles flow from regions of high concentration to regions of low concentration. In Figure (2.20), we have assumed an electron concentration that increases with x; hence electrons tend to flow from right to left. The particle flux density is proportional to the concentration gradient. The direction of the flow is opposite to the gradient.

$$Particle\ flux\ density = -D_n \nabla n \tag{2.78}$$

Where D_n is the proportionality constant. It is called the electron diffusion constant. The negative sign in the equation for the particle flux density indicates that the flux of particles is in the opposite direction to the concentration gradient. ∇n is the electron concentration gradient. Since the dimension of the left – hand side of the above equation is $cm^{-2}sec^{-1}$and the dimension of the concentration gradient is cm^{-4}, we infer that the dimension of the diffusion constant, D_n, is cm^2sec^{-1}. Since each electron carries a charge equal to $-q$ coulombs, the electric current density due to diffusion is given by

$$J_n = \mathrm{q}\, D_n \nabla n \tag{2.79}$$

In one dimension this becomes equal to

$$J_n = \mathrm{q}\, D_n \frac{dn}{dx} \tag{2.80}$$

When one calculates the current flow due to hole diffusion noting that each carries a charge $+q$, we get for the hole current density

$$J_p = -\mathrm{q}\, D_p \frac{dp}{dx} \tag{2.81}$$

The electric current due to the density gradient is called the diffusion current. We saw earlier that the electric current cue to the presence of the electric field is called the drift current. The electric current density caused by electron flow is due to both diffusion and drift and is given by

$$J_n = \mathrm{q\ n}\, \mu_n \mathcal{E} + \mathrm{q}\, D_n \frac{dn}{dx} \tag{2.82}$$

and similarly the hole electric current density is given by

$$J_p = \mathrm{q\ p}\, \mu_p \mathcal{E} - \mathrm{q}\, D_p \frac{dp}{dx} \tag{2.83}$$

The total electric current density is the sum of electron and hole electric current densities and is given by

$$J = J_n + J_p \tag{2.84}$$

In our discussions, we will be concerned with the diffusion current due to the minority carriers only. The diffusion current due to majority carrier density gradient will be seen to be balanced by a drift current arising from a self-induced electric field.

Einstein Relation

The mobility and diffusion constants are related as follows:

$$\mu_n = \frac{q}{kT} D_n \tag{2.85}$$

and

$$\mu_n = \frac{q}{kT} D_n \tag{2.86}$$

This relationship is called Einstein relationship. Although this relationship can be derived, it is sufficient for our purposes to assume the relationship. It is easy to remember this relationship if one bears in mind the dimensions of mobility has the dimension $cm^2V^{-1}sec^{-1}$. Since $\frac{kT}{q}$ has the dimension of volt it is easy to remember that the mobility times $\frac{kT}{q}$ is the diffusion constant.

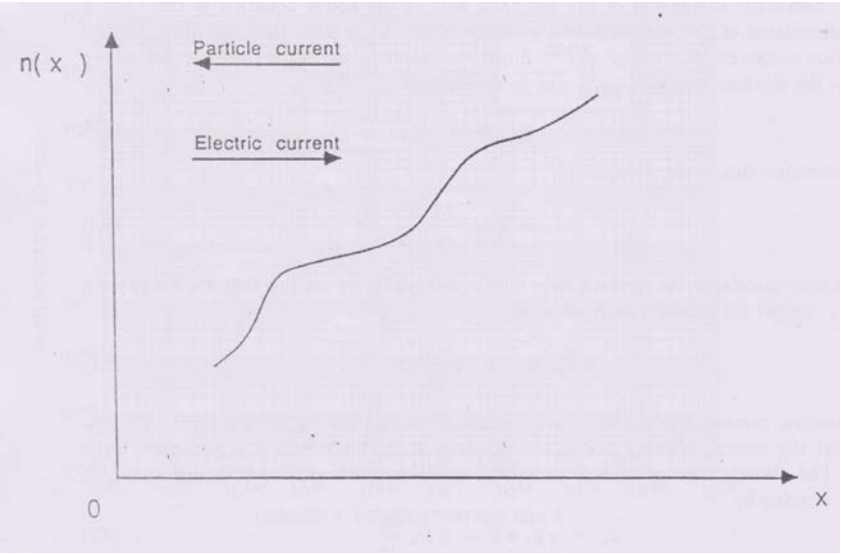


Figure (2.20): Variation of electron concentration with distance

Generation Recombination Process

We saw that free electrons and holes were created by thermal excitation of electrons from the valence band into the conduction band. The process of creating electrons and holes is called generation. Generation can be either due to thermal excitation i.e., by imparting to the electron in the valence band an amount of thermal energy sufficient to be excited to the conduction band, or by other means such as

optically exciting the valence band or by bombardment with high energy particles. Electrons in the valence band are those that are localized in the covalent bonds. As shown in figure (2.21), when a covalent bond is broken the free electron is the electron in the conduction band and the vacancy represented in the broken bond is the hole in the valence band.

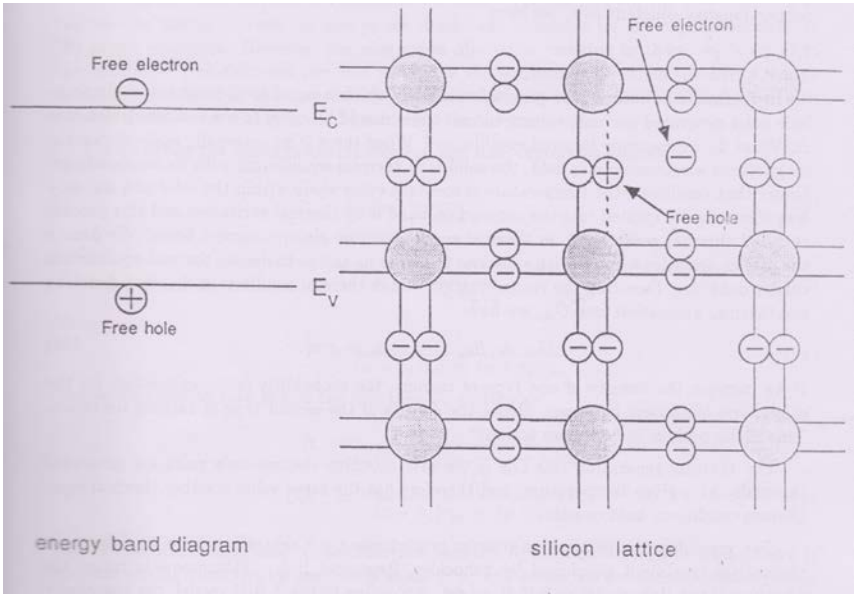


Figure (2.21): Illustration of connection between free electrons and free holes in the semiconductor and electrons in the conduction band and hole in the valence band

A free electron moving in the crystal can recombine with a free hole resulting in the loss of both the electron and hole. This process is called the recombination of an electron and a hole.

There are two mechanisms by which generation-recombination process takes place: 1) band to band generation or recombination; 2) generation or recombination through an impurity or a defect. In the first process electrons and holes recombine directly by an electron in the conduction band jumping back to a vacant state in the valence band. Physically, the recombination process is one in which the free electron wanders to the site of a broken bond and the covalent bond gets completed when the electron jumps into the broken bond. This results in the loss of a free electron and a free hole.

In the second process electrons and holes recombine through the intermediary of an impurity or trap which has an energy level in the band gap. Such impurities or traps are called generation-recombination centers or g-r centers. Recombination takes place by the recombination center first capturing an electron or a hole and then subsequently capturing a carrier of the opposite charge. This gives rise to the annihilation or recombination of an electron and a hole. The recombination centers also generate electrons and holes. The generation takes place by the center emitting an electron or a hole first and subsequently emitting a carrier of the opposite charge. Generation requires imparting of energy to the solid and recombination results in the release of energy. In band to band recombination energy is released in the form of light or photons. In recombination through g-r centers energy is released as thermal (phonon) energy.

These processes of generation and recombination always require the involvement of both an electron and a hole each time the process takes place and it is customary to say that the recombination

of an electron-hole pair or the generation of an electron-hole pair occurs. In silicon the generation-recombination process occurs mostly through g-r centers and the band to band process is highly improbable.

Independent of the actual process, direct or indirect, the rate of recombination R, equal to the number of electron-hole pairs lost due to recombination per unit volume in unit time, is proportional to the density of electrons and the density of holes. During the proportionality constant as r, we have

$$R = r\, n\, p \tag{2.87}$$

In thermal equilibrium, the generation rate G, which is equal to the number of electron-hole pairs generated per unit volume in unit time, should be equal to the recombination rate R. What do we mean by thermal equilibrium? When there is no externally applied electric, magnetic or electromagnetic field, the solid is in thermal equilibrium with its surroundings. Under that condition, the temperature is constant everywhere within the solid and the only way electrons are excited into the conduction band is by thermal excitation and this process is called thermal generation. In thermal equilibrium, no electric current flows. We denote the carrier densities n and p with subscript 0 such as n_0 and p_0 to denote thermal equilibrium carrier densities. Denoting the recombination rate in thermal equilibrium R_{th} , and denoting the thermal generation rate G_{th} , we have

$$G_{th} = R_{th} = r\, n_0\, p_0 = r\, n_i^2 \tag{2.88}$$

If we increase the density of one type of carriers, the probability of recombination for the other type of carriers increases. Hence the density of the second type of carriers decreases. This is the basis of law of mass action.

The thermal generation rate G_{th} is the rate at which electron-hole pairs are generated thermally at a given temperature, and therefore has the same value whether thermal equilibrium conditions hold no or not.

The recombination-generation process involving a g-r center has been modeled by a theoretical treatment developed by Schockley, Read and Hall. This model is called the Schockley-Read Hall model or S-R-H model. According to the S-R-H model, the constant r is given as,

$$r = \frac{1}{\frac{n+n_1}{N_t \sigma_p v_p} + \frac{P+p_1}{N_t \sigma_n v_n}} \tag{2.89}$$

where

N_t = density of traps or g-r centers,

σ_n, σ_p = capture cross-sections for electrons and holes respectively,

v_n, v_p =thermal velocity of electrons and holes respectively and is equal to $\sqrt{\frac{3\,k\,T}{m^*}}$, where m^* is the effective mass of the carrier,

$$n_1 = N_c\, e^{\left(-\frac{E_C - E_t}{kT}\right)}, \tag{2.90}$$

$$p_1 = N_v\, e^{\left(-\frac{E_t - E_v}{kT}\right)}, \tag{2.91}$$

E_t = the trap energy level.

The capture cross-section σ's relate to the ability (probability) of the g-r centers to capture the charge carriers. n_1 and p_1 are short-hand notations for the expression given in the above equations. However, one can assign a physical meaning to them. n_1 is equal to the number of free electrons per unit volume in the semiconductor when the Fermi energy is at the same level as the trap energy level. Similarly, p_1 is the number of holes per unit volume if the Fermi energy is at the same level as the trap level.

Let us define the terms involving electron and hole capture as follows:

$$\tau_{n0} = \frac{1}{N_t \sigma_n v_n} \tag{2.92}$$

and

$$\tau_{p0} = \frac{1}{N_t \sigma_p v_p} \tag{2.93}$$

Then

$$r = \frac{1}{(n + n_1)\,\tau_{p0} + (p + p_1)\,\tau_{n0}} \tag{2.94}$$

The recombination rate R and G_{th} can then be written as

$$R = \frac{n\,p}{(n + n_1)\,\tau_{p0} + (p + p_1)\,\tau_{n0}}$$

$$G_{th} = \frac{n_i^2}{(n + n_1)\,\tau_{p0} + (p + p_1)\,\tau_{n0}} \tag{2.95}$$

The net recombination rate is the difference between R and G_{th}. In regions where there is thermal equilibrium, the np product is equal to n_i^2 and hence the net recombination rate is zero. In regions where there is no thermal equilibrium, if the np product is more than n_i^2, a net recombination will occur. If the np product is less than n_i^2, then a net generation of electron-hole pairs will occur. Let us now illustrate these points by considering an extrinsic semiconductor.

Extrinsic semiconductor

We will consider an n-type semiconductor although the discussion will be similar for a p-type semiconductor. Let us consider the thermal equilibrium recombination rate first.

$$R_{th} = \frac{n_{n0}\, p_{n0}}{(n_{n0} + n_1)\,\tau_{p0} + (\,p_{n0} + p_1)\,\tau_{n0}}$$

$$= \frac{n_i^2}{n_{n0}\tau_{p0} + n_1\tau_{p0} + p_1\tau_{n0}} \tag{2.96}$$

We neglected the term involving p_{n0} in the denominator of the above equation since $p_{n0} \ll n_{n0}$. Expressing n_1 and p_1 in terms of n_i,

$$R_{th} = \frac{n_i^2}{n_{n0}\tau_{p0} + n_i\tau_{p0}e^{\frac{E_t - E_1}{kT}} + n_i\tau_{n0}e^{-\frac{(E_t - E_i)}{kT}}}$$

$$= \frac{p_{n0}}{\tau_{p0} + \frac{n_i}{n_{n0}}\left[\tau_{p0}e^{\frac{E_t - E_1}{kT}} + \tau_{n0}e^{-\frac{(E_t - E_i)}{kT}}\right]} \tag{2.97}$$

Equation (2.97) can be rewritten as

$$R = \frac{p_{n0}}{\tau_p} \tag{2.98}$$

where

$$\frac{1}{\tau_p} = \frac{1}{\tau_{p0} + \frac{n_1}{n_{n0}}\tau_{p0} + \frac{p_1}{n_{n0}}\tau_{n0}} \tag{2.99}$$

We can infer from Equation (2.98) that $\frac{1}{\tau_p}$ can be interpreted as the probability per unit time that a hole, (i.e., a minority carrier) will combine. Using this interpretation one can show that τ_p is the average time that a minority carrier spends before it is lost due to recombination. Hence it is called the minority carrier lifetime. Equation (2.99) tells us that the minority carrier lifetime depends on the energy of the g-r center is near the middle of the band gap.

When the trap levels are close to the middle of the gap, then the minority carrier lifetime τ_p becomes

$$\tau_p = \tau_{po} \tag{2.100}$$

The thermal generation rate, G_{th} is equal to R_{th} and hence,

$$G_{th} = \frac{n_i^2}{n_{n0}\,\tau_{p0} + n_1\tau_{p0} + p_1\,\tau_{n0}}$$

$$= \frac{p_{n0}}{\tau_p} \tag{2.101}$$

We can rewrite the above equation as

$$p_{n0} = G_{th}\,\tau_p \tag{2.102}$$

Which gives the interpretation that the thermal equilibrium minority carrier density is the thermal generation rate times the lifetime.

Till now we considered a g-r center with a discrete energy level. However, generally the g-r centers are distributed over a range of energies. Then the minority carrier lifetime is obtained by adding the contributions of all the g-r centers.

Excess carriers

Suppose that excess carriers over and above the thermal equilibrium densities are created by illuminating the semiconductor with light or by electrical injection as we will see later on in our study of p-n junctions. This corresponds to a non-thermal equilibrium situation. The carrier densities n and p can be written as

$$n = n_{n0} + \Delta n \tag{2.103}$$

$$p = p_{n0} + \Delta p \tag{2.104}$$

Where n_{n0} and p_{n0} are thermal equilibrium carrier densities and Δn and Δp are called excess carrier densities. When excess carrier densities are created optically, equal number of excess holes and electrons are created and hence $\Delta n = \Delta p$. When excess minority carriers are inject ted as in a p-n junction, excess majority carriers will flow into this region to preserve charge neutrality and hence again $\Delta n = \Delta p$. The recombination rate in the presence of excess carriers is therefore given by

$$R = \frac{n\,p}{(n+ n_1)\,\tau_{p0}+(p+ p_1)\,\tau_{n0}}$$

$$R = \frac{(n_{n0}+\Delta p)(p_{n0}+\Delta p)\,p}{(n_{n0}+\Delta p+ n_1)\,\tau_{p0}+(p_{n0}+\Delta p+ p_1)\,\tau_{n0}} \tag{2.105}$$

The net recombination rate, U, is the difference between the recombination rate and the generation rate. The latter is the thermal generation rate

$$U = R - G_{th}$$

$$= \frac{(n_{n0}+\Delta p)(p_{n0}+\Delta p)\,p}{(n_{n0}+\Delta p+ n_1)\,\tau_{p0}+(p_{n0}+\Delta p+ p_1)\,\tau_{n0}} - \frac{n_i^2}{n_{n0}\,\tau_{p0}+ n_1\tau_{p0}+p_1\,\tau_{n0}}$$

$$= \frac{n_{n0}p_{n0}+(n_{n0}+p_{n0})\Delta p+\Delta p^2}{(n_{n0}+\Delta p+ n_1)\tau_{p0}+(p_{n0}+\Delta p+ p_1)\tau_{n0}} - \frac{n_i^2}{n_{n0}\,\tau_{p0}+ n_1\tau_{p0}+p_1\,\tau_{n0}} \tag{2.106}$$

We will now distinguish two limiting cases, one called low injection and the other called high injection. The low injection conditions are obtained when $\Delta p \geq n_{n0}$.

Under low injection conditions, Δp^2 term is negligible and p_{n0} and Δp are also negligible in comparison with n_{n0}. Hence,

$$U = \frac{(n_{n0}+p_{n0})\Delta p+ n_{n0}p_{n0}}{(n_{n0}+\Delta p+ n_1)\tau_{p0}+(p_{n0}+\Delta p+ p_1)\tau_{n0}} - \frac{n_i^2}{(n_{n0}+n_1)\tau_{p0}+p_1\,\tau_{n0}}$$

$$= \frac{(n_{n0}\Delta p)+n_i^2}{(n_{n0}+n_1)\tau_{p0}+p_1\,\tau_{n0}} - \frac{n_i^2}{(n_{n0}+n_1)\tau_{p0}+p_1\,\tau_{n0}}$$

$$= \frac{\Delta p}{\tau_{p0}+\frac{n_1}{n_{n0}}+\frac{p_1}{n_{n0}}\tau_{n0}} - \frac{n_i^2}{(n_{n0}+n_1)\tau_{p0}+p_1\,\tau_{n0}}$$

$$= \frac{\Delta p}{\tau_{p0}+\frac{n_1}{n_{n0}}\tau_{p0}+\frac{p_1}{n_{n0}}\tau_{n0}} = \frac{\Delta p}{\tau_p} \qquad (2.107)$$

The net recombination is due to excess carriers. Therefore U is the rate of excess carrier recombination and is proportional to Δp. In other words, $\frac{1}{\tau_p}$ is the probability per unit time that an excess carrier will be lost due to recombination. Therefore the excess carrier lifetime is also $\tau_p$which is minority carrier lifetime.

Under high injection $\Delta p > n_{n0}$ and hence the term $(n_{n0} + n_1)\Delta p$ is negligible in comparison with Δp^2. Then

$$n = n_{n0} + \Delta p \;\approx\; \Delta p \qquad (2.108)$$

Also G_{th} is negligible in comparison with R. Under this condition

$$U = \frac{\Delta p^2}{(n_{n0} + \Delta p + n_1)\tau_{p0} + (p_{n0} + \Delta p + p_1)\tau_{n0}}$$

$$= \frac{\Delta p^2}{\Delta p\left(\tau_{p0} + \tau_{n0}\right)}$$

$$= \frac{\Delta p}{\tau_{p0}+\tau_{n0}} = \frac{\Delta p}{\tau_{hi}} \qquad (2.109)$$

Where τ_{hi} is the lifetime of the excess carrier at high injection. Under high injection conditions, the excess carrier lifetime becomes the sum of τ_{p0} and τ_{n0}. At intermediate injection, the lifetime is between the high injection value and the low injection value.

Depletion Region

We will see in our study of p-n junctions with an applied voltage that it is possible to have regions within the semiconductor under non-thermal equilibrium condition, in which the carrier densities n and p are much less than the thermal equilibrium carrier densities. Under this condition ($n \approx p \approx 0$) the recombination rate R is negligible for

$$R = \frac{n\,p\,N_t v_0 \sigma_0}{n+n_1+p+p_1} \approx 0 \qquad (2.110)$$

The thermal generation rate G_{th} is equal to

$$G_{th} = \frac{n_i^2}{(n + n_1)\,\tau_{p0} + (p + p_1)\,\tau_{n0}}$$

$$\approx \frac{n_i^2}{n_1\tau_{p0} + p_{1\tau_{n0}}} = \frac{n_i}{\tau_{p0}e^{\frac{E_t - E_i}{kT}} + \tau_{n0}e^{\frac{-(E_t - E_i)}{kT}}}$$

$$= \frac{n_i}{2\tau_g} \tag{2.111}$$

Where τ_g is defined as the generation lifetime and equal to

$$\tau_g = \frac{1}{2}\left[\tau_{p0}e^{\frac{E_t - E_i}{kT}} + \tau_{n0}e^{\frac{-(E_t - E_i)}{kT}}\right] \tag{2.112}$$

The generation rate is maximum only when the trap level E_t is in the middle of the band gap i.e., when $n_1 = p_1 = n_i$. Another way of stating this is that only those traps which are within a few kT above or below E_i will be effective as generation centers. When $E_t = E_i$, $\tau_g$is a minimum and equal to

$$\tau_g = \frac{\tau_{p0} + \tau_{n0}}{2} \tag{2.113}$$

Thus we conclude that the thermal generation rate is $\frac{n_i}{2\tau_g}$ in the non-thermal equilibrium depletion region, while in the neutral region it is equal to $\frac{p_{n0}}{\tau_p}$.

Continuity Equation

Let us consider an elementary volume ΔV in an n-type semiconductor. In this elementary volume minority carriers are being thermally generated at a rate G_{th} and let us further assume that minority carriers are also generated at a rate G_{opt} due to external light sources. Minority carriers are also lost due to recombination at rate R. In addition, if there is an outward flow of current out of this elementary volume, then the minority carriers will be lost if the minority carriers are holes, and minority carries will increase in number if the minority carriers are electrons. Since we are considering an n-type semiconductor, the minority carriers are holes in our example.

The total outward flow of current out of a volume surrounded by a surface is equal to

$$\int_{surface} \overrightarrow{J_p} \cdot \overrightarrow{dS} = \int_{vol} \nabla \cdot \overrightarrow{J_p} \cdot dV \tag{2.114}$$

In an elementary volume ΔV, the outward flow of current reduces to $\nabla \cdot \overrightarrow{J_p}\,\Delta V$. The outward flow represents the charge flow, and hence the number of holes lost in unit time is equal to $\frac{1}{q}\nabla \cdot \overrightarrow{J_p}\Delta V$. Hence we can write the rate of increase of the number of holes in the elementary volume as

$$\frac{\delta p}{\delta t}\Delta V = \Delta V(-R + G_{th} + G_{opt} - \frac{1}{q}\nabla \cdot \overrightarrow{J_p}) \quad (2.115)$$

Dividing the above equation by ∇V we get

$$\frac{\delta p}{\delta t} = -R + G_{th} + G_{opt} - \frac{1}{q}\nabla \cdot \overrightarrow{J_p} \quad (2.116)$$

This equation is called the continuity equation. From our earlier discussion, we know that

$$R - G_{th} = \frac{\Delta p}{\tau_p} \quad (2.117)$$

Hence the continuity equation reduces to

$$\frac{\delta p}{\delta t} = \frac{\delta(\Delta p)}{\delta t} = -\frac{\Delta p}{\tau_p} + G_{opt} - \frac{1}{q}\nabla \cdot \overrightarrow{J_p} \quad (2.118)$$

But

$$J_p = q\mu_p p\mathcal{E} - qD_p\nabla p \quad (2.119)$$

In one dimension, the continuity equation reduces to

$$\frac{\delta(\Delta p)}{\delta t} = -\frac{\Delta p}{\tau_p} + G_{opt} - \frac{1}{q}\frac{\mathrm{dJ_p}}{\mathrm{dx}} \quad (2.120)$$

The expression for the hole current density is given by

$$J_p = q\mu_p p\mathcal{E} - qD_p\frac{\delta p}{\delta x} \quad (2.121)$$

Let us assume that the electric field $\mathcal{E}$ is zero.

$$J_p = -qD_p\frac{\delta p}{\delta x}$$

Then

$$\frac{\delta(\Delta p)}{\delta t} = -\frac{\Delta p}{\tau_p} + G_{opt} + D_p\frac{\delta^2 \mathrm{p}}{\delta^2 \mathrm{x}} \quad (2.122)$$

A simple example:

Let us assume that light is incident on the semiconductor to generate excess carrier uniformly within the semiconductor.

Growth of excess carriers

Let the light be switched on a time t = 0. Since the excess carriers are uniformly generated everywhere in the semiconductor, there is no concentration gradient, and hence the term due to the current flow can be dropped in the continuity equation. The continuity equation governing the growth of the excess carriers is then given by

$$\frac{d\Delta p}{dt} = -\frac{\Delta p}{\tau_p} + G_{opt} \tag{2.123}$$

The solution to this equation can be readily seen to be

$$\Delta p(t) = G_{opt}\tau_p\left[1 - e^{-\frac{t}{\tau_p}}\right] = \Delta p_0\left(1 - e^{-\frac{t}{\tau_p}}\right) \tag{2.124}$$

Where $\Delta p_0 = G_{opt}\tau_p$ is the excess carrier density at t $= \infty$, i.e., when steady state conditions have been reached. Equation (2.124) describes that the excess carrier density grows exponentially with time, from zero to its steady state value.

Decay of excess carriers

Let us assume that the light has been on for a long time, and that steady state conditions have been reached. Now, let the light be turned off at time t $= \infty$. The decay of excess carrier density will be governed by putting $G_{opt} = 0$ in the continuity equation, which will then be

$$\frac{d\Delta p}{dt} = -\frac{\Delta p}{\tau_p} \tag{2.125}$$

The solution to this equation is

$$\Delta p(t) = \Delta p_0 e^{-\frac{t}{\tau_p}} \tag{2.126}$$

Where Δp_0 is the excess carrier density just when the light was turned off, i.e., at $t = 0$.

From the above treatment, we can conclude that the time taken to reach steady state during the growth and the decay of excess carriers is determined by the minority carrier lifetime.

Summary

A semiconductor behaves like an insulator at very low temperatures, but it exhibits increasing electrical conductivity as the temperature is increased. The electrical conductivity of a semiconductor at room temperature is much lower than that of a conductor but very much higher than that of an insulator. As the temperature is raised, in a semiconductor, some of the electrons get enough thermal energy to break loose from the covalent bond and become free electrons. Similarly, the broken bonds represent free holes. In terms of a band picture, at very low temperatures the valence band is completely full and the conduction band is empty. Additionally, the band gap (the separation in energy between the conduction and valence bands) is small. As the temperature is increased, the electrons get excited from the valence band to the conduction band. The electrons in the conduction band are the free electrons while the vacant states in the valence band are the fee holes. Since an electron has a charge of $-q$ coulombs and a hole has a charge of q coulombs, the charge moves with them when they move. Hence electrons and holes are called *charge carriers*, or simply carriers.

An intrinsic semiconductor is one in which electrons and holes are generated only by thermal excitation from the valence band to the conduction band. The free electron density is equal to the free hole density ($n_i = p_i$). The electron has no velocity when it occupies a state at the bottom of the conduction band, and hence has no kinetic energy. Hence E_C, the bottom of the conduction band, is the potential energy of the electron. An electron that occupies a state with energy E has a kinetic energy is $E - E_C$. Similarly, the potential energy of the hole is E_V and its kinetic energy is $E_V - E$. In other words, the kinetic energy of the hole increases as we go deeper into the valence band.

Free holes and free electrons can be treated as particles in a box by assuming the appropriate effective masses for the electrons and holes. The probability of occupation of a state by an electron is given by the Fermi function, $f(E)$. The probability of occupation of a state by a hole is the probability that an electron will not occupy that state, $(1 - f(E))$.

In deriving the expression for the density of electrons and holes, a numerical integration is necessary. This procedure can be simplified by assuming that the Fermi energy is at least a few kT below the conduction band and at least a few kT above the valence band. When the assumption is not valid in a material, the semiconductor is said to be degenerate. The Fermi function approximates the Boltzmann distribution function. Hence this approximation is also called the Boltzmann approximation.

The electron (and hole) density in an intrinsic semiconductor is determined by the band gap energy E_g and the effective masses. It is also a strong function of temperature. The Fermi energy is approximately at the middle of the band gap in an intrinsic material. If the effective masses of the holes and the electrons are assumed to be equal to the mass of the electron in vacuum, then the Fermi energy lies exactly at the middle of the band gap. The Fermi energy in an intrinsic material is denoted by E_i, and hence it is customary to denote the middle of the bandgap as E_i.

An extrinsic semiconductor is one in which controlled amount of trace impurities are substitutionally added to increase one type or the other type of carriers. When elements of group V are added as impurities in silicon, the electron density is increased without a corresponding increase in the free hole density. These impurities are called donors, and the material is called an n-type semiconductor. In an n-type semiconductor, electrons are called majority carriers and holes are called minority carriers since electron density is much larger than hole density. On the other hand, when

impurity atoms belonging to an element in group III of the periodic table are added, the hole density is larger than electron density. This type of material is called a p-type semiconductor. In this type of material, holes are called majority carriers and electrons are called minority carriers.

In an n-type semiconductor, free electrons arise due to two reasons: thermal excitation from the valence band, and thermal excitation from the donor atoms (ionization of the donor atoms). The holes in an n-type material arise only due to thermal excitation of electrons from the valence band. The positive charge in an n-type semiconductor arises due to free holes and positively ionized donor atoms while the negative charge arises due to free electrons.

In a p-type semiconductor, free holes arise due to two reasons: thermal excitation of electrons from the valence band, and ionization of acceptor atoms. The electrons arise only due to thermal excitation from the valence band. The positive charge in a p-type material arises due to free holes, while the negative charge in an n-type material arises due to electrons and negatively ionized acceptor atoms. In a neutral semiconductor, (n-type or p-type) the net charge density is zero. When both acceptors and donors are present in the same region, the net impurity density (the donor density minus the acceptor density) determines the type of semiconductor. Such a material is called a compensated material.

The expression for the carrier density in an extrinsic semiconductor can be derived similarly to the expression for an intrinsic material. For a high density of impurities, the carrier density can only be obtained by numerical; integration. However, when the impurity density is not that high, the material can be considered to be non-degenerate and the expression for the carrier density is of the same form as what is obtained in the intrinsic case under non-degenerate assumption. In a non-degenerate semiconductor, $np = n_i^2$. This is called the law of mass action. When the electron density is increased by adding donor impurities, the hole density decreases. Similarly, when the hole density is increased by adding acceptor impurities, the electron density decreases. The number of ionized impurities increases with temperature. At moderate and low temperatures, the majority carrier density is equal to the ionized net impurity atoms.

In an n-type material the Fermi energy lies in the upper half of the band gap. As the donor density is increased, the Fermi energy moves closer to the conduction band. In a p-type material the Fermi energy lies in the lower half of the band gap. As the acceptor density is increased, the Fermi energy moves closer to the valence band.

At high temperatures, all the impurity atoms are ionized and an increase in carrier density arises only by thermal excitation from the valence band. At still higher temperatures, the electron and hole densities become equal and the material becomes intrinsic.

An electric current in a semiconductor can arise due to either a drift process or a diffusion process. In the drift process, an externally applied electric field drives the charge carriers. The resulting current is called the drift current. The drift current can be expressed in terms of a parameter called the mobility. The drift current is generally due to majority carriers. In the diffusion process, carriers diffuse from regions of high concentration to regions of low concentration. The resulting current is called the diffusion current. The diffusion current is expressed in terms of a constant called the diffusion constant. The diffusion current is due to minority carriers only.

The diffusion constant is related to the mobility by a factor $\frac{kT}{q}$. This relationship is called the Einstein relation. The mobility is limited by scattering from impurity atoms or scattering from phonons. The two processes have different temperature dependence.

In a semiconductor, electrons and holes are constantly being created (generated) and lost due to recombination. Recombination is the process in which an electron in the conduction band jumps back to a vacant state in the valence band. In this manner, both an electron and a hole have been annihilated. In steady state, the generation rate is equal to the recombination rate.

There are two mechanisms for generation-recombination to occur. One is what is called the band to band generation-recombination. The other is called generation-recombination through a trap or a generation-recombination (g-r) center. In silicon, the second is the dominant mechanism. The average time that a minority carrier spends in its band before it is lost due to recombination is called the minority carrier lifetime. The thermal equilibrium minority carrier density is given by the product of the thermal generation rate and the minority carrier lifetime. The minority carrier lifetime is inversely proportional to the density of g-r centers. When excess carriers are generated, the process is called injection of carriers. The average time an excess carrier spends in its band before it is lost due to recombination is called excess carrier lifetime. The steady state excess carrier density is given by the product of the excess carrier generation rate and the excess carrier lifetime.

When the excess carrier density is much less than the thermal equilibrium majority carrier density, it is called a low injection process. When the excess carrier density is comparable to or higher than the thermal equilibrium majority carrier density, then it is called a high injection process. Under low injection, the excess carrier lifetime is the same as the minority carrier lifetime. The growth and decay of excess carriers is governed by the excess carrier lifetime when the generation process is switched on or off respectively.

The flow and growth (or decay) of excess carriers is determined by solving the continuity equation.

Glossary

dn_E = number of electrons with energy between $E \; and \; E + dE$ per unit volume of the crystal

dV = elementary volume

D_n = electron diffusion constant

D_p = hole diffusion constant

E = energy

E_A = acceptor energy level

E_C = bottom of the conduction band, also the potential energy of an electron in the bottom of the conduction band

E_D = donor energy level

E_F = Fermi energy

E_g = bandgap equal to $E_C - E_V$

E_i = Fermi energy in an intrinsic material, usually referred to as the intrinsic Fermi energy

E_t = trap level or energy level of the g-r center

E_{top} = top of the conduction band or the maximum energy of the conduction band

ε = electric field

$f(E)$ = Fermi function; the probability of occupation by an electro

G = generation rate

G_{opt} = rate at which minority carriers are generated by an external light source

G_{th} = thermal generation rate

h = Planck's constant

J = total current density

J_n = electric current density due to electron flow

J_p = electric current density due to hole flow

k = Boltzmann constant

l_n = average distance travelled by the electron between collisions

m_0 = the mass of an electron in vacuum

m_C = effective mass of an electron in the conduction band

m_n, m_p = effective masses for electrons and holes respectively

m_v = effective mass of hole in the valence band

m^* = effective mass of the carrier

$n_0 \& p_0$ = thermal equilibrium carrier densities

n_1 = number of free electrons per unit volume in the semiconductor when the Fermi energy is at the same level as the trap energy level

n = electron density

n_i = density of electrons or holes in the intrinsic semiconductor

n_0 = thermal equilibrium electron density in the n-region

n_{x0} = thermal equilibrium electron density in the x-type of semiconductor where x is n or p

N_C = effective density of states in the conduction band

N_t = density of traps or g-r centers

N_V = effective density of states in the valence band

N_A = density of acceptor atoms

N_A^- = density of ionized acceptors, number of negatively ionized atoms per unit volume

N_D = density of donor atoms

N_D^+ = density of ionized donors, number of positively ionized atoms per unit volume

p = density of free holes

P = momentum of carriers

p_1 = number of holes per unit volume if the Fermi energy is at the same level as the trap level

p_{n0} = thermal equilibrium hole (minority carrier) density in an n-type semiconductor

p_{p0} = thermal equilibrium hole (majority carrier) density in a p-type semiconductor

p_{x0} = thermal equilibrium hole density in x-type material where x is n or p

q = charge

r = proportionality constant for recombination rate

R = recombination rate

R_{th} = recombination rate at thermal equilibrium

t = time

T = temperature in degrees Kelvin

U = net recombination rate, the recombination rate minus the generation rate

v = velocity

v_d = average velocity, also called the drift velocity

v_{dn} = drift velocity of electrons

v_{dp} = drift velocity of holes

v_n, v_p = thermal velocity of electrons and holes respectively

V = volume of a semiconductor solid

$Z_C(E)$ = density of states for electrons in the conduction band

Δn = excess electron density

Δp = excess hole density

Δp = change in momentum

Δp_0 = excess carrier density at $t = \infty$

ΔV = elementary volume

η = ideality factor

μ = mobility of the carriers

μ_n = electron mobility

$\mu_{n\,imp}$ = mobility that an electron would have if scattering were due only to impurities

$\mu_{n\,ph}$ = mobility that an electron would have if scattering were due only to lattice vibration

μ_p = hole mobility

ρ = resistivity, the reciprocal of σ

σ = electrical conductivity

σ = capture cross-section

σ_n, σ_p = capture cross-sections for electrons and holes respectively

τ_C = scattering relaxation time or average time between collisions

τ_{CN} = average time between scattering or collisions suffered by the electron

τ_{Cp} = average time between scattering or collisions suffered by the hole

τ_p = hole lifetime, i.e. average time that a minority carrier spends before it is lost due to recombination

$\emptyset$ = electrostatic potential

ψ = electrostatic potential

Problems

1. Derive an expression for the density of holes in a non-degenerate semiconductor using a method similar to that used in the text for the derivation of an expression for the density of electrons.

2. Given that $n_i = 10^{10}\ cm^{-3}$ at $T = 300^0 K$, calculate n_i at $T = 200^0 K$ and at $T = 150^0 K$ assuming $E_g = 1.12\ eV$ at all temperatures (i.e., neglecting variation of E_g with temperature).

3. A piece of silicon is doped with $2x10^{15}$ phosphorus atoms per cm^3. What are the majority and minority carrier concentrations at room temperature?
4. Suppose that the hole concentration in a piece of silicon at room temperature is $10^5\ cm^{-3}$, find

   (a) the electron concentration
   (b) the location of the Fermi energy

5. A sample of silicon is first doped with $10^{15}\ cm^{-3}$ boron atoms, and then doped with $4x10^{15}$ cm^{-3} arsenic atoms.
   (a) What is the type of the semiconductor?
   (b) Find the location of the Fermi energy.

6. In an n-type semiconductor, the temperature is lowered such that only half the donor atoms are ionized. Neglecting the degeneracy factor, show that

7. In an n-type semiconductor say n-type, at higher temperatures the electrons come from ionization of the donor atoms as well as from excitation of electrons from the valence band to the conduction band. If N_D^+ is the donor density, show that the intrinsic carrier density, at the temperature at which the electron density is twice that of the hole density, is $\sqrt{2}N_D^+$.

8. Find the resistivity of the same in Problem 4. Take the mobility values from the figure in the text.

9. Using the mobility curves in the text, find the resistivity of the sample in the Problem 5.

10. A piece of n-type silicon has a resistivity of 5 $\Omega - cm$ at room temperature $(27^0\ C)$. Find the thermal equilibrium concentration at $17^0 C$.

11. Two scattering mechanisms are present in a semiconductor. The mobility, if the first mechanism alone is present is 500 $cm^2\ V^{-1}\ s^{-1}$. The mobility, if the second mechanism alone is present, is 900 $cm^2\ V^{-1}\ s^{-1}$. What is the mobility of the carrier in the sample?

12. If an intrinsic piece of silicon is illuminated such that carriers are generated at the rate of 10^{18} $cm^{-3}\ s^{-1}$ uniformly everywhere in the sample, and if the lifetime of the carriers is 10^{-6} s, determine the change in the electrical conductivity.

13. Write the expression for the total electron current in a semiconductor. From this write the relation between the concentration gradient $\frac{dn}{dx}$ and $n(x)$ under conditions of no current flow.

14. Show that the probability that a hole will survive without recombining for t seconds is given by $e^{-\frac{t}{\tau_p}}$ assuming that it was created at time $t = 0$ and that $\frac{1}{\tau_p}$ is the probability that a hole will recombine in unit time.

15. Show that the average time that a hole lasts without recombining is τ_p seconds. Assume that the hole was created at time $t = 0$. (This is why τ_p is called the hole lifetime.)

16. Consider an n-type silicon sample with $N_d = 10^{16} cm^{-3}$. Calculate the location of the Fermi level at a) 300^0K, and b) 200^0K.

Chapter 3

P- N Junction

A p-n junction is a device structure in which a single crystal of a semiconductor say silicon, has one region doped with donor impurities to make it n-type while the rest of the semiconductor is doped with acceptor impurities to make it p-type. For example, Figure (3.1) shows the impurity variation from one end of the crystal to the other. This figure is called the impurity profile. The ordinate is the net impurity concentration N_D- N_A which is positive in the n-type region and negative in the p-type region. The abscissa is x, position.

Let us now consider an isolated p-type semiconductor. The net impurity concentration N_D- N_A is negative. The density of states in the two bands is plotted in Figure (3.2 A). The Fermi distribution function is plotted for a p-type material in Figure (3.2 B). The Fermi energy is close to the valence band as shown in Figure (3.2 B). The probability of occupation by electrons i.e., the Fermi function, is nearly zero in value for states in the conduction band whereas the probability of occupation by holes i.e., 1 minus the Fermi function, is much larger for states in the valence band. The distribution of electrons in the conduction band and the holes in the valence band is obtained by multiplying the density of states and the probability of occupation, and is given in Figure (3.2 C). The area under these distribution curves gives the number of carriers in the conduction and the valence band. As is to be expected in a p-type semiconductor, a small density of electrons and a large density of holes are obtained.

Let us now similarly consider an isolated n-type semiconductor. The net impurity concentration N_D- N_A is positive. Again the density states in two bands is plotted in Figure (3.3 A). The probability functions for the occupation of states by electrons and holes are plotted in Figure (3.3 B). The distribution of carriers in the two bands is given in Figure (3.3 C). We see now in the n-type material a small density of holes and a large density of electrons are obtained.

Before going further we will now show that the gradient in Fermi energy is zero when no electric current flows. Let us consider the electric current due to electron flow. The electric current is given by

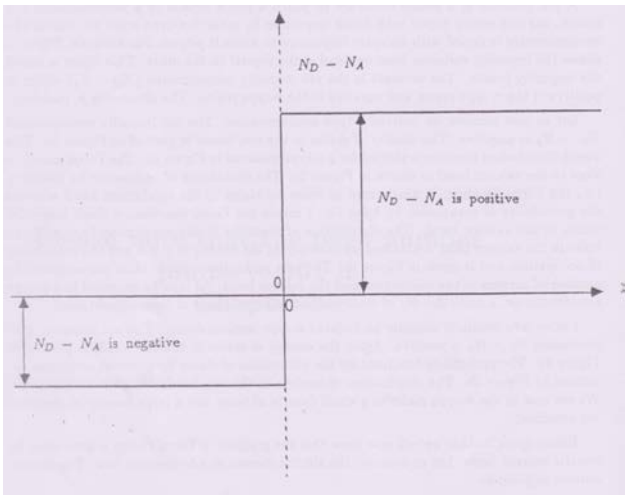
$$J_n = q\, n\, \mu_n\, \mathcal{E} + q\, D_n \frac{dn}{dx} \tag{3.1}$$

But we know that under thermal equilibrium,

$$n_0 = n_i e^{\frac{E_F - E_i}{kT}} \tag{3.2}$$

Where we have used the subscript $_0$ to denote thermal equilibrium densities, and

$$\mathcal{E} = -\frac{d\psi}{dx} = \frac{1}{q}\,\frac{dE_i}{dx} \tag{3.3}$$



Figure(3.1): Impurity profile in a p-n junction. This junction is called an abrupt junction since the net impurities change from one type to the other abruptly.

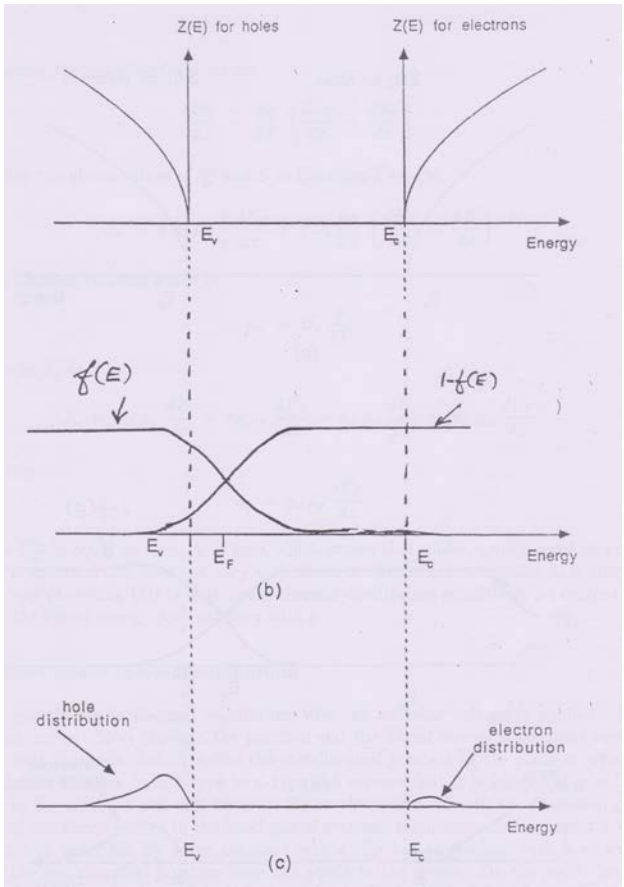


Figure (3.2): Carrier distribution in an isolated p-type semiconductor. A) Plot of the distribution of density of states in the valence and in the conduction band. B) Plot of $f(E)$ (Fermi function) and $1 - f(E)$. c) Distribution of electrons and holes in the respective bands.

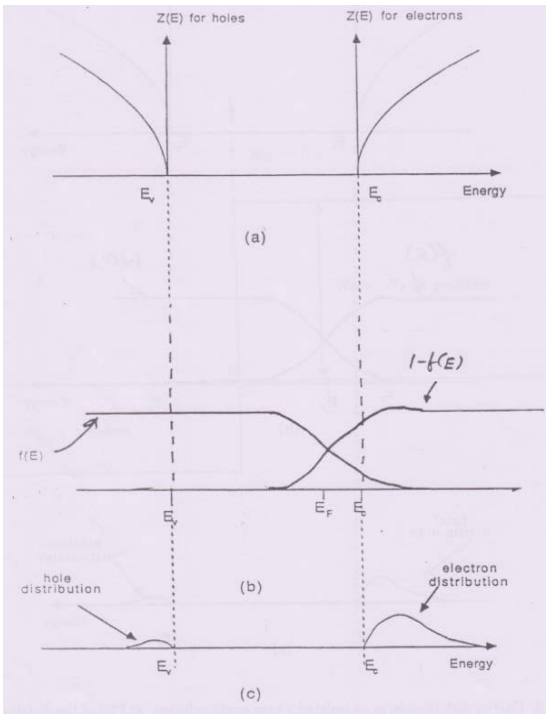


Figure (3.3): Carrier distribution in an isolated n-type semiconductor. A) Plot of the distribution of density of states in the valence and in the conduction band. B) Plot of $f(E)$ (Fermi function) and $1 - f(E)$. C) Distribution of electrons and holes in the respective bands.

differentiating the equation for n we get

$$\frac{dn_0}{dx} = \frac{n_0}{kT}\left(\frac{dE_F}{dx} - \frac{dE_i}{dx}\right) \tag{3.4}$$

Substituting the above values of $\frac{dn}{dx}$ and $\mathcal{E}$ in Equation (3.1) we get

$$J_n = q\, n\, \mu_n \frac{1}{q}\, \frac{dE_i}{dx} + q\, D_n \frac{n_0}{kT}\left(\frac{dE_F}{dx} - \frac{dE_i}{dx}\right) \tag{3.5}$$

Recalling Einstein relation which is

$$\mu_n = D_n \frac{q}{kT} \tag{3.6}$$

we can write J_n as

$$J_n = n_0\, \mu_n\, \frac{dE_i}{dx} + n_0\, \mu_n\, \frac{dE_F}{dx} - n_0\, \mu_n\, \frac{dE_i}{dx} = n_0\, \mu_n\, \frac{dE_F}{dx} \tag{3.7}$$

Similarly

$$J_p = p_0\, \mu_p\, \frac{dE_F}{dx} \tag{3.8}$$

When $\frac{dE_F}{dx}$ is equal to zero, J_n is zero, which means that under conditions of no current flow, E_F is constant and does not vary with distance. Under that condition J_p is also zero. Another way of stating this is that under thermal equilibrium conditions, no current flows and that the Fermi energy does not vary with x.

P-N junction Under Thermal Equilibrium

The p-n junction is in thermal equilibrium when no external voltage is applied. Hence no electric current flows through the junction and the Fermi energy is constant and does not vary with distance. Let us define the metallurgical junction as the plane at which the semiconductor changes from p-type to n-type and assume that it is located at $x = 0$. E_F is closer to E_V on the p-side and closer to E_C on the n-side. (Recall the discussion on the location of the Fermi energy in the band gap of extrinsic semiconductor in chapter 2.) This means that in order for E_F to be constant with x, E_C has to decrease with x as we pass through the metallurgical junction from the p-side to the n-side. On the p-side far away from the junction, there is no electric field and hence E_C is constant i.e., does not vary with distance and is separated from E_F by an amount determined by the net ionized acceptor concentration N_A^-. E_F is below the intrinsic Fermi energy level E_i, by an amount given by

$$E_i - E_F = kT\ ln\left(\frac{N_A^-}{n_i}\right) \tag{3.9}$$

Similarly on the n-side far away from the junction, E_C is constant and is separated from E_F by an amount determined by the net ionized donor concentration N_D^+. E_F is above the intrinsic Fermi energy level E_i by an amount given by

$$E_F - E_i = kT\ ln\left(\frac{N_D^+}{n_i}\right) \tag{3.10}$$

However, close to the junction, E_C (potential energy) and therefore E_i and E_V vary gradually from their respective values on the p-side to their values on the n-side. This region where the potential energy varies is called the transition region. In this region there is an electric field since E_C and the electrostatic potential varies. Outside the transition region the electric field is zero. Since the electric field lines emanate (flow out) from the positive charges and terminate on negative charges, the transition region comprises positive charges on one side and negative charges on the other side. The charge density in the transition region is non-zero whereas the charge density is zero outside the transition region. For this reason the transition region is called the **space-charge region** since it has a net charge density. The region outside the transition region is called the **neutral region** since it has no net charge density. How does a space charge arise? In the transition region on the n-side, electron density is much less than the thermal equilibrium majority carrier density and therefore a charge density equal to the magnitude of the electrical charge times the ionized donor atom density arises in this region. Similarly hole density is less than the thermal equilibrium majority carrier density value in the transition region on the p-side. Thus there is a region on the p-side close to the metallurgical junction with a net negative charge density equal to the electron charge times the ionized acceptor atom density. Electric lines of force emanate from the positive charge on the n-side and terminate on the negative charge on the p-side.

Another way of looking at the origin of the transition region is as follows: Again let us assume that the p-n junction is formed by joining a p-type semiconductor and a n-type semiconductor and fusing them. As soon as the two pieces of semiconductor have been joined together, electrons, which are large in number in the n-region diffuse to the p-region where there are only a few of them. Thus there is a diffusion current. The diffusion of electrons charges the p-region negative with respect to the n-region and therefore the potential energy of the electrons on the p-side raises relative to the n-side i.e., a potential energy barrier arises between the n and p regions. This give rise to an electric field in the transition region due to the gradient of the potential energy since the potential energy changes with x from the p-side to n-side. This electric field causes a drift flow of electron current in the direction from the p-side to the n-side i.e., in the opposite direction to the diffusion current. The potential energy barrier in the transition region reduces the diffusion current and hence fewer electrons diffuse and therefore the potential energy rises less. Ultimately a steady state condition is reached in which the potential energy barrier is of such a magnitude that the drift current is exactly equal and opposite to the diffusion current. Hence no net current flows through the junction. In the transition region on the n-side there is a net positive charge density due to the ionized donor atoms since electrons left this region to go initially to the p-side. Similarly in the transition region on the p-side, there is a net negative charge density due to the majority carriers viz., holes having been neutralized by the electrons that *came from the n-side*. We discussed this model as though electrons initially went from the n-side to the p-side. We would have arrived at the same conclusion if we had assumed that holes initially *diffused form the p-side to the n-side.*

Let us now define the electrostatic potential $\emptyset$ as

$$\emptyset = -\frac{E_C}{q} + \ constant = \ -\frac{E_i}{q} - \frac{E_g}{2q} + constant \tag{3.11}$$

The constant is for the purpose of referencing our electrostatic potential to a particular reference or zero value. Since E_F is constant with distance, choose the constant as $\frac{E_g}{2q} + \frac{E_F}{q}$. Then $\emptyset$ becomes

$$\emptyset = \ -\frac{E_i}{q} + \frac{E_F}{q} \ = \frac{(E_F - E_i)}{q} \tag{3.12}$$

$\emptyset$, the electrostatic potential, varies with distance x in the transition region. Since n_0, the thermal equilibrium density of electrons, is given by

$$n_0 = n_i \, e^{\frac{(E_F - E_i)}{kT}} \tag{3.13}$$

We can express n_0 as

$$n_0 = \ n_i e^{\frac{q\emptyset}{kT}} \tag{3.14}$$

Similarly, since p_0, the thermal equilibrium density of holes, is given by

$$p_0 = n_i \, e^{\frac{(E_i - E_F)}{kT}} \tag{3.15}$$

we can express p_0 as

$$p_0 = n_i e^{-\frac{q\emptyset}{kT}} \tag{3.16}$$

Thus both the electron and the hole densities vary with distance x in the transition region. By taking the logarithm of both sides of Equations (3.14) and (3.16), we can express the electrostatic potential in terms of the thermal equilibrium carrier densities.

$$\emptyset = \frac{kT}{q} \ln\left(\frac{n_0}{n_i}\right) \tag{3.17}$$

and also

$$\emptyset = -\frac{kT}{q} \ln\left(\frac{p_0}{n_i}\right) \tag{3.18}$$

These two equations relate the potential variation with thermal equilibrium carrier density variation.

$\emptyset$, the electrostatic potential, defined as above, can also be used in the neutral semiconductor. In a neutral n-type semiconductor, n_0, the thermal equilibrium electron density, is equal to n_{n0}, the thermal equilibrium majority carrier density which is equal to N_D^+, the ionized donor density, Substituting N_D^+ for n_0 in Equation (3.14) we find that $\emptyset$, the electrostatic potential in the neutral semiconductor, is given by

$$\emptyset_n = \frac{kT}{q} \ln\left(\frac{N_D^+}{n_i}\right) \tag{3.19}$$

where we have used a subscript n in $\emptyset$ to denote that we talking about the electrostatic potential in a neutral n-type semiconductor. Similarly we can define an electrostatic potential for a p-type material as

$$\emptyset_p = -\frac{kT}{q} \ln\left(\frac{N_A^-}{n_i}\right) \tag{3.20}$$

where we have used a subscript p in $\emptyset$ to denote the electrostatic potential in a neutral p-type semiconductor.

Example

Let us now calculate the electrostatic potential $\emptyset$ for a p-type silicon which has a concentration of 10^{16}cm^{-3} acceptor atoms at room temperature. We assume that at room temperature all the impurity atoms are ionized, i.e., $N_A^- = N_A$. Let us take n_i as equal to $1 \times 10^{10} cm^{-3}$.

$$\emptyset_p = -\frac{kT}{q} \ln\left(\frac{N_A^-}{n_i}\right)$$

$$= -.02585 \times \ln\left(\frac{10^{16}}{1 \times 10^{10}}\right)$$

$$= -0.02585 \times 13.81 = -0.357\ V$$

Let us now go back to the space-charge region. Differentiating n_0 given in Equation (3.14) we obtain

$$\frac{dn_0}{dx} = n_0 \frac{q}{kT} \frac{d\emptyset}{dx} \tag{3.21}$$

This yields

$$dn_0 = n_0 \frac{q}{kT} d\emptyset \tag{3.22}$$

which can be written as

$$d\emptyset = \frac{kT}{q} \frac{dn_0}{n_0} \tag{3.23}$$

We can integrate Equation (3.23) to obtain the potential difference between two points say A and B in a semiconductor. The potential difference $\emptyset_{A-B}$ is equal to

$$\emptyset_{A-B} = \int_B^A d\emptyset = \frac{kT}{q} \int_B^A \frac{dn_0}{n_0} = \frac{kT}{q} ln\left(\frac{n_A}{n_B}\right) \tag{3.24}$$

We can use this result to obtain an expression for the potential barrier that exists between the n-side and the p-side in a p-n junction. *This potential barrier is called the built-in voltage* since it is already in existence in a p-n junction under thermal equilibrium. This ***built-in voltage*** which is denoted by $\boldsymbol{V_{bi}}$ is the potential difference between the neutral n-side and the neutral p-side and according to the above equation, is equal to

$$V_{bi} = \frac{kT}{q} ln\left(\frac{n_{n0}}{n_{p0}}\right) = \frac{kT}{q} ln\left(\frac{n_{n0} p_{p0}}{n_i^2}\right) = \frac{kT}{q} ln\left(\frac{N_A^- N_D^+}{n_i^2}\right) \tag{3.25}$$

Equation (3.25) can be rewritten as

$$V_{bi} = \frac{kT}{q}\left(ln\left(\frac{N_D^+}{n_i}\right) + ln\left(\frac{N_A^-}{n_i}\right)\right) \tag{3.26}$$

We could have arrived at the same result by a different route also. We know $\emptyset_p$ and $\emptyset_n$, the electrostatic potentials of the neutral p and the n region respectively. The built-in voltage V_{bi} is the electrostatic potential difference between the neutral n and the p regions.

Hence

$$V_{bi} = \emptyset_n - \emptyset_p \tag{3.27}$$

Since $\emptyset_p$ is negative from Equation (3.20) we can write

$$V_{bi} = \emptyset_n + |\emptyset_p| = \frac{kT}{q}\left(ln\left(\frac{N_D^+}{n_i}\right) + ln\left(\frac{N_A^-}{n_i}\right)\right) \quad (3.28)$$

which is the same result that we obtained before.

Example

Let us now determine the built-in potential for a p-n junction in which the n-region is uniformly doped with a net donor concentration of 2 x 10^{16} cm^{-3} and the p region is uniformly doped with a net acceptor concentration of 1 x 10^{15} cm^{-3}. Assume room temperature and that all the impurity atoms are ionized. From Equation 25 we find

$$V_{bi} = 0.02585 \times ln\left(\frac{2x10^{16}x1x10^{15}}{(1 \times 10^{10})^2}\right)$$

$$= 0.02585 \times 26.02$$

$$= 0.673\ V$$

An abrupt p-n junction is one in which the net acceptor density is uniform in the p-region and the net donor density is uniform in the n-region. The impurities change from acceptor to donor type abruptly as we go across the metallurgical junction. In Figure (3.4 A) the abrupt junction is shown with the metallurgical junction at $x = 0$. The p and the n regions are connected externally by a copper wire. The junction is therefore in thermal equilibrium. In Figure (3.4 B) is shown the band-bending i.e., the bending of E_C, E_i and E_V, that occurs in the transition region as we go from the p-side to the n-side. The built-in voltage is the difference between the potential energy in the p-side and that in the n-side divided by q. In other words, it is equal to the amount of band-bending divided by q. In Figure (3.4 C) the space charge distribution is shown in the transition region. The charge density in the transition region is due to the ionized impurity atoms that are not compensated by equal number of majority carriers. Hence the charge density in the space charge region is $\rho = -q\ N_A$ on the p-side and $\rho = +q\ N_D$ on the n-side. Since the p-n junction is in thermal equilibrium, E_F is constant and does not vary with x. On the other hand E_i varies with x in the depletion region. The electron density and the hole density in the specie-charge region are given by

$$n = n_i\, e^{\frac{E_F - E_i}{kT}}$$

and

$$p = n_i\, e^{-\frac{E_F - E_i}{kT}}$$

Hence n and p vary with position, x, in the depletion region. The electron density decreases from n_{n0}, the thermal equilibrium majority carrier concentration on the n-side, to n_{p0}, the thermal equilibrium minority carrier concentration on the p-side. The electron density in the portion of the transition (space charge) region lying on the n-side is less than n_{n0}, the thermal equilibrium majority carrier density in the

neutral n-region. Hence this region has a net charge density equal to a N_D^+ and is depleted of electrons. By a similar argument, the portion of the transition region on the p-side is depleted of holes and the charge density in this region Is $-qN_A^+$. The carrier densities are smaller than the thermal equilibrium majority carrier densities in respective portions of the entire space-charge region. Hence the space-charge region is said to be depleted of charge carriers and the space-charge region is usually referred to as the depletion region. The product of the electron and the hole densities anywhere in the transition region is equal to the square of the intrinsic carrier concentration since the junction is in thermal equilibrium, i.e., the law of mass action is obeyed. At the two edges of the space-charge region the charge density gradually approaches zero as shown in Figure (3.4 C) since the carrier densities change gradually. However for ease of calculation and analysis, we approximate the space charge density distribution as a rectangular distribution in which the charge density becomes zero abruptly at the two edges as shown in Figure (3.4 D). We assume here that the negative charge density due to ionized acceptors exists between $x = -x_p$ and $x = 0$ and the positive charge density due to ionized donors exists between $x = 0$ and $x = x_n$. The electric field variation in the space region is shown in Figure (3.4 E). Outside the space charge region, the electric field is zero. Inside the space charge region, the electric field varies linearly and the magnitude reaches a maximum value at the metallurgical junction. The electric field is the negative gradient of the electrostatic potential. Therefore the line integral of the electric field yields the electrostatic potential difference between two points. The area under the plot of the electric field in Figure (3.4 E) is equal to the built-in voltage V_{bi}. The linear variation of the electric field in the space-charge region is due to the assumption that the impurity density is constant in each region and that it changes from one type to the other abruptly at the metallurgical junction.

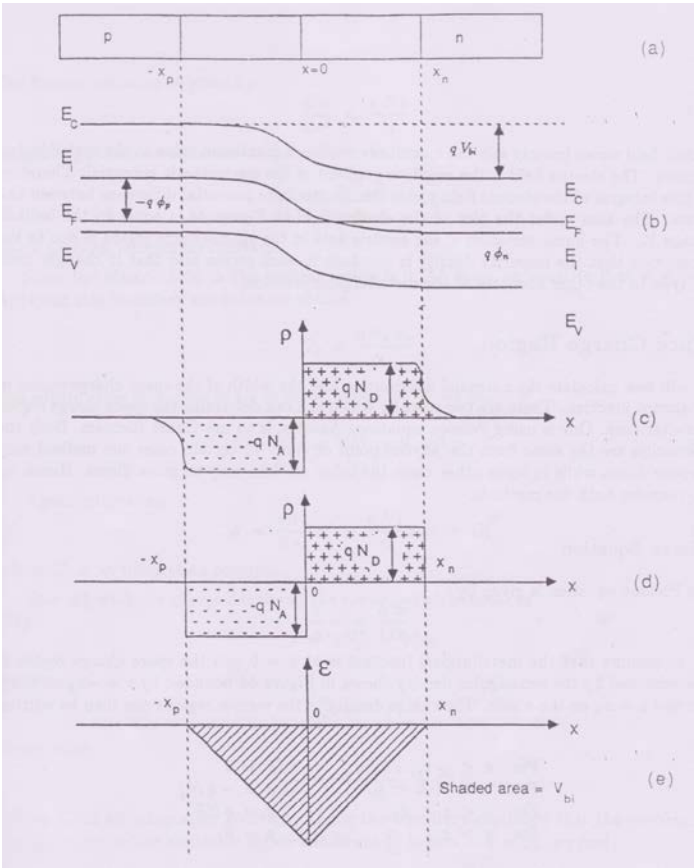


Figure (3.4): (A) An abrupt p-n junction with the metallurgical junction at $x = 0$. (B) Bending of the band at the junction. The built-in voltage is the difference in the potential energy between the two sides divided by the electron charge q. (C) The actual space charge region with the charge density gradually approaching zero at the two edges. (D) Approximation by a rectangular charge density in which the density abruptly goes to zero at the edges. (E) Electric field variation in the space charge region.

Space Charge Region

We will now calculate the potential distribution and the width of the space charge region in the abrupt junction. There are two ways by which one can determine the space charge region characteristics. One is using ***Poisson equation***. Another is to use ***Gauss theorem***. Both the approaches are the same from the physics point of view. In certain cases one method may be more direct while in some other cases the other method may be more direct. Hence we will consider both the methods.

Poisson Equation

The Poisson equation is given by

$$\frac{d^2\emptyset}{dx^2} = -\frac{\rho}{\epsilon_s} \tag{3.29}$$

Let us assume that the metallurgical junction is at $x = 0$ and the space charge region is approximated by the rectangular density shown in Figure (34 D) bounded by $x = -x_p$ on the p side and $x = x_n$ on the n side. The charge density in the various regions can then be written as

For $x < -x_p,$ $\rho = 0$

For $-x_p < x < 0$ $\rho = -qN_A^-$

For $0 < x < x_{n,}$ $\rho = -qN_D^+$

For $x > x_{n,}$ $\rho = 0$

In the space-charge region, under steady state conditions, all the impurity atoms are ionized even at low temperatures[3] i.e., $N_D^+ = N_D$ and $N_A^- = N_A$. This is in contrast to the neutral region where the impurity atoms are partially ionized at low temperatures.

Let us consider the space charge region (the depletion region) on the p side.

$$\rho = -qN_A^- = -qN_A$$

The Poisson equation is given by

$$\frac{d^2\emptyset}{dx^2} = \frac{qN_A}{\epsilon_s} \tag{3.30}$$

[3] The theoretical reason for this is that the electric field in the depletion region ionizes all the impurity atoms even at low temperatures

Integrating the above equation we get

$$\frac{d\emptyset}{dx} = \frac{qN_A x}{\epsilon_s} + C_1 \tag{3.31}$$

where C_1 is an integration constant.

Since the electric field in the neutral region is 0, $\frac{d\emptyset}{dx}$ must be equal to 0 at $x = -x_p$. Applying this boundary condition we obtain

$$C_1 = \frac{qN_A x_p}{\epsilon_s} \tag{3.32}$$

And substituting this value of C_1 in the equation for $\frac{d\emptyset}{dx}$ we get

$$\frac{d\emptyset}{dx} = \frac{qN_A}{\epsilon_s}\left(x + x_p\right) \tag{3.33}$$

Again integrating

$$\emptyset = \frac{qN_A}{2\epsilon_s} x^2 + \frac{qN_A}{\epsilon_s} x_p x + C_2 \tag{3.34}$$

Where C_2 is an integration constant.

Starting with the charge density in the n-side, which is equal to

$$\rho = q\, N_D^+ = q\, N_D$$

We get

$$\frac{d^2\emptyset}{dx^2} = -\frac{qN_D}{\epsilon_s} \tag{3.35}$$

Integrating,

$$\frac{d\emptyset}{dx} = -\frac{qN_D x}{\epsilon_s} + C_1' \tag{3.36}$$

Where C_1' is an integration constant. Using the boundary condition that the electric field $\left(-\frac{d\emptyset}{dx}\right)$ is zero in the neutral n region and hence $\frac{d\emptyset}{dx}$ is zero at $x = x_n$, we find

$$C_1' = \frac{qN_D x_n}{\epsilon_s}$$

Hence,

$$\frac{d\emptyset}{dx} = -\frac{qN_D x}{\epsilon_s} + \frac{qN_D x_n}{\epsilon_s} = \frac{qN_D(x_n - x)}{\epsilon_s} \tag{3.37}$$

Integrating

$$\emptyset = -\frac{qN_D}{2\epsilon_s}x^2 + \frac{qN_D}{\epsilon_s}x_n x + C_3 \tag{3.38}$$

where C_3 is an integration constant.

Realizing that $\emptyset$ given by Equations (3.34) and (3.38), should have the same value at $x = 0$, we put $C_2 = C_3$. Therefore, the potential $\emptyset$ at $-x = -x_p$, is given, from Equation (3.34), by

$$\emptyset(x = -x_p) = \frac{qN_A x_p^2}{2\epsilon_s} - \frac{qN_A x_p^2}{\epsilon_s} + C_2 = -\frac{qN_A x_p^2}{2\epsilon_s} + C_2 \tag{3.39}$$

Similarly, the potential $\emptyset$ at $x = x_n$ is obtained from Equation (3.38) as

$$\emptyset(x = x_n) = -\frac{qN_D x_n^2}{2\epsilon_s} - \frac{qN_D x_n^2}{\epsilon_s} + C_2 = \frac{qN_D x_n^2}{2\epsilon_s} + C_2 \tag{3.40}$$

The total potential difference between the neutral n side and the neutral p side is given by

$$V_{bi} = \emptyset(x = x_n) - \emptyset(x = -x_p)$$

$$= \frac{qN_D x_n^2}{2\epsilon_s} + C_2 - \left(-\frac{qN_A x_p^2}{2\epsilon_s} + C_2\right)$$

$$= \frac{qN_D x_n^2}{2\epsilon_s} + \frac{qN_A x_p^2}{2\epsilon_s} \tag{3.41}$$

The amount of positive charge in the n-side of the space-charge region, per unit area of the junction, is equal to $qN_D x_n$. Similarly the amount of negative charge in the p-side of the space-charge region, per unit area of the junction, is equal to $-qN_A x_p$. Since the positive charge in the n side of the depletion region should be equal to the magnitude of the negative charge in the p side, we get

$$N_A x_p = N_D x_n \tag{3.42}$$

Using this equality we can express V_{bi} as

$$V_{bi} = \frac{qN_D}{2\epsilon_s}x_n^2 + \frac{qN_A}{2\epsilon_s}x_p x_p = \frac{qN_D}{2\epsilon_s}x_n^2 + \frac{qN_D x_n x_p}{2\epsilon_s}$$

$$= \frac{qN_D x_n}{2\epsilon_s}(x_n + x_p) = \frac{qN_D x_n}{2\epsilon_s}\left(x_n + \frac{N_D x_n}{N_A}\right)$$

$$= \frac{qN_D x_n^2}{2\epsilon_s}\left(1 + \frac{N_D}{N_A}\right) = \frac{qN_D x_n^2}{2\epsilon_s}\frac{(N_A + N_D)}{N_A}$$

Therefore

$$x_n = \sqrt{\frac{2\epsilon_s V_{bi}}{q}\frac{N_A}{N_D(N_A + N_D)}} \tag{3.43}$$

Similarly it can be shown that

$$x_p = \sqrt{\frac{2\epsilon_s V_{bi}}{q}\frac{N_D}{N_A(N_A+N_D)}} \tag{3.44}$$

The total depletion region width x_d is equal to

$$\begin{aligned} x_d &= x_n + x_p \\ &= \sqrt{\frac{2\epsilon_s V_{bi}}{q(N_A+N_D)}}\left(\sqrt{\frac{N_A}{N_D}} + \sqrt{\frac{N_D}{N_A}}\right) \\ &= \sqrt{\frac{2\epsilon_s V_{bi}}{q(N_A+N_D)}}\left(\frac{N_A+N_D}{\sqrt{N_D N_A}}\right) \\ &= \sqrt{\frac{2\epsilon_s V_{bi}(N_A+N_D)}{q(N_D N_A)}} \end{aligned} \tag{3.45}$$

It must be pointed out that the temperature dependence of x_d arises due to the temperature dependence of V_{bi} as expressed in Equation (3.28). In that equation we expressed the built-in voltage as sum of two components $\emptyset_n$ and $\emptyset_p$. We have to remember that these two components are not the potential drop across the depletion region on the n side and the potential drop across the depletion region on the p side. In Equation (3.41) we again expressed V_{bi} as sum of two components. However in this equation, the first component expresses the potential drop across the depletion region in the n side while the second component is the potential drop across the depletion region in the p side.

Let us now express the potential as a function of distance in the depletion region. The electrostatic potential $\emptyset$ given in Equation (3.34) is measured with respect to some arbitrary zero reference by an appropriate choice of C_2. Let us measure the potential with respect to the neutral p-side i.e., choose the electrostatic potential on the neutral p-side as zero, and denote this potential by ψ. ψ can be obtained from $\emptyset$ proper choice of C_2. Using Equation (3.34) we obtain ψ for the region $-x_p < x < 0$ as

$$\psi = \frac{qN_A}{2\epsilon_s}x^2 + \frac{qN_A}{\epsilon_s}x_p x + C_2' \tag{3.46}$$

where C_2' is the new constant which should make equal to zero at $x = -x_p$. By putting this condition in the above equation we obtain

$$C_2' = \frac{qN_A x_p^2}{2\epsilon_s} \tag{3.47}$$

By substituting this in the above equation for ψ we get

$$\psi = \frac{qN_A}{2\epsilon_s}x^2 + \frac{qN_A}{\epsilon_s}x_p x + \frac{qN_A x_p^2}{2\epsilon_s} = \frac{qN_A}{2\epsilon_s}\left(x_p + x\right)^2 \tag{3.48}$$

ψ at $x = 0$ is obtained as

$$\psi(x = 0\,) = \frac{qN_A x_p^2}{2\epsilon_s} \tag{3.49}$$

Equation (3.48) describes the potential variation in the depletion region on the p side, the potential being measured with respect to the neutral p region. Since we are measuring the potential, ψ, with respect to the neutral p region, the value of ψ at $x = -x_p$ given by Equation (3.49) is equal to the potential drop across the segment of the transition region lying on the p-side.

Using Equation (3.38) we can express the potential variation in the depletion region on the n side with respect to the same zero reference also by using the same constant C_2' , and show that the potential $\psi(x\,)$ in the depletion region on the n side i.e., for $0 < x < x_n$ as

$$\psi(x\,) = -\frac{qN_D}{2\epsilon_s}x^2 + \frac{qN_D}{\epsilon_s}x_n x + \frac{qN_A x_p^2}{2\epsilon_s} \tag{3.50}$$

This is equivalent to putting the constant C_3 in Equation (3.38) as equal to

$$\frac{qN_A x_p^2}{2\epsilon_s}$$

We could have started also with the zero reference for the potential as that on the neutral n region and derived a similar set of expressions for the potential in the depletion regions on the two sides of the metallurgical junction.

We summarize in Table (3.1), the step by step solution of Poisson equation in the space charge region of an abrupt *p-n* junction. The solutions to the Poisson equation with two different zero references are also given in this table.

Table (3.1). The summary of the step by step solution of Poisson equation in the space charge region of an abrupt *p-n* junction

	Region	Region
	$-x_p < x < 0$	$0 < x < x_n$

$\frac{d^2\emptyset}{dx^2}$	$\frac{qN_A}{\epsilon_s}$	$-\frac{qN_D}{\epsilon_s}$
$\frac{d\emptyset}{dx}$	$\frac{qN_Ax}{\epsilon_s}+C_1$	$-\frac{qN_Dx}{\epsilon_s}+C_1^{'}$
Boundary condition	$\frac{d\emptyset}{dx}=0 \quad at \quad x=x_p$	$\frac{d\emptyset}{dx}=0 \quad at \quad x=x_n$
Integration constant	$C_1=\frac{qN_Ax_p}{\epsilon_s}$	$C_1^{'}=\frac{qN_Dx_n}{\epsilon_s}$
$\frac{d\emptyset}{dx}$	$\frac{qN_A}{\epsilon_s}(x+x_p)$	$\frac{qN_D(x_n-x)}{\epsilon_s}$
$\emptyset$	$\frac{qN_A}{2\epsilon_s}x^2+\frac{qN_A}{\epsilon_s}x_px+C_2$	$-\frac{qN_D}{2\epsilon_s}x^2+\frac{qN_D}{\epsilon_s}x_nx+C_3$
Boundary condition at zero: $\emptyset(x=0)$	C_2	$C_3=C_2$
Zero reference	$\emptyset(x=-x_p)=0$	
Integration Constant	$C_2=\frac{qN_Ax_p^2}{2\epsilon_s}$	$\frac{qN_Ax_p^2}{2\epsilon_s}$
$\emptyset$	$\frac{qN_A(x_p+x)^2}{2\epsilon_s}=\frac{qN_Ax_p^2}{2\epsilon_s}\left(1+\frac{x}{x_p}\right)^2$	$-\frac{qN_D}{2\epsilon_s}x^2+\frac{qN_D}{\epsilon_s}x_nx+\frac{qN_Ax_p^2}{2\epsilon_s}$
Zero reference		$\emptyset(x=x_n)=0$
Integration constant	$C_2=-\frac{qN_Dx_n^2}{2\epsilon_s}$	$C_3=-\frac{qN_Dx_n^2}{2\epsilon_s}$
$\emptyset$	$\frac{qN_A}{\epsilon_s}\left(x_px+\frac{x^2}{2}\right)-\frac{qN_Dx_n^2}{2\epsilon_s}$	$-\frac{qN_D}{2\epsilon_s}(x_n-x)^2=-\frac{qN_Dx_n^2}{2\epsilon_s}\left(1-\frac{x}{x_n}\right)^2$

Gauss Theorem Approach

The origin of the electric field in the depletion region is more intuitively understood if the depletion region is examined using Gauss Theorem. We will now derive the same result using Gauss theorem. As before let us assume that the depletion region charge is rectangular as shown in Figure (3.5 A).

If we were imagine a cylindrical section of the depletion region, with unit area of cross-section, as shown in Figure (3.5 B), then the total negative charge contained between $-x_p$ and $x = 0$ in the cylindrical section is equal in magnitude to the total positive charge in the cylinder between $x = 0$ an x_n. This equality of charge magnitude is required according to Gauss theorem due to the fact that the electric field is zero outside the depletion region i.e., in the neutral n and p regions. The electric field at any point x in the depletion region can be obtained by partitioning the cylinder with a plane at x and considering the charge contained in either part of the cylinder. For example, in Figure (3.5 C), the left part of the cylinder contained between $x = x_p$ and x is shown. The electric field at x is the charge contained in the left part of the cylinder divided by ϵ_s, the permittivity of the semiconductor.

$$\mathcal{E}(x) = -\frac{qN_A\left(x-(-x_p)\right)}{\epsilon_s} = -\frac{qN_A(x+x_p)}{\epsilon_s} \tag{3.51}$$

In the above discussion we assume that x is located between $-x_p$ and 0 i.e., in the p-side of the depletion region. The field lines come from the right to the left and hence the field is negative. x in Equation (3.51) is also negative.

The electric field, $\mathcal{E}(x)$, at x is also equal to the charge contained in the right part of the cylinder between x and x_n as shown in Figure (3.5 D), divided by ϵ_s , the permititivity. Hence,

$$\mathcal{E}(x) = -\left[\frac{qN_A(x-0)}{\epsilon_s} + \frac{qN_D(x_n-0)}{\epsilon_s}\right] = -\left[\frac{qN_Ax}{\epsilon_s} + \frac{qN_Dx}{\epsilon_s}\right] \tag{3.52}$$

Since $N_Dx_n = N_Ax_p$, , this expression for $\mathcal{E}$, the electric field, can be seen to be the same as the one in Equation (3.51). It is always convenient to consider the charge contained in that part of the cylindrical section in which the impurity atoms are of the same type. Hence Equation (3.51) is the preferred choice to express $\mathcal{E}(x)$ for $-x_p < x < 0$.

If x was chosen to be between 0 and x_n i.e., in the n-side of the depletion region, then we can write $\mathcal{E}(x)$ as the charge contained in the right part of the cylinder divided by ϵ_s.

$$\mathcal{E}(x) = -\frac{qN_D(x_n - x)}{\epsilon_s} \tag{3.53}$$

We can obtain the potential variation as a function of x by taking the line integral of $\mathcal{E}(x)$. If we take the zero reference for the potential as that at the neutral p-region, then

$$\psi(x) = -\int_{-x_p}^{x} \mathcal{E}(x)dx$$

$$= \int_{-x_p}^{x} \frac{q\,N_A(x + x_p)}{\epsilon_s}dx$$

$$= \frac{q\,N_A}{\epsilon_s}\left[\frac{x^2}{2} + x_px\right]_{-x_p}^{x}$$

$$= \frac{q\,N_A}{\epsilon_s}\left[\frac{x^2}{2} + x_px - \left(\frac{x_p^2}{2} - x_p^2\right)\right]$$

$$= \frac{q\,N_A}{\epsilon_s}\left[\frac{x^2}{2} + x_p x + \frac{x_p^2}{2}\right]$$

$$= \frac{q\,N_A}{2\epsilon_s}\left[x^2 + 2x_p x + x_p^2\right]$$

$$= \frac{q\,N_A}{2\epsilon_s}\left[x + x_p\right]^2 \qquad (3.54)$$

This is the same result that we obtained in Equation (3.48) by solving Poisson equation.

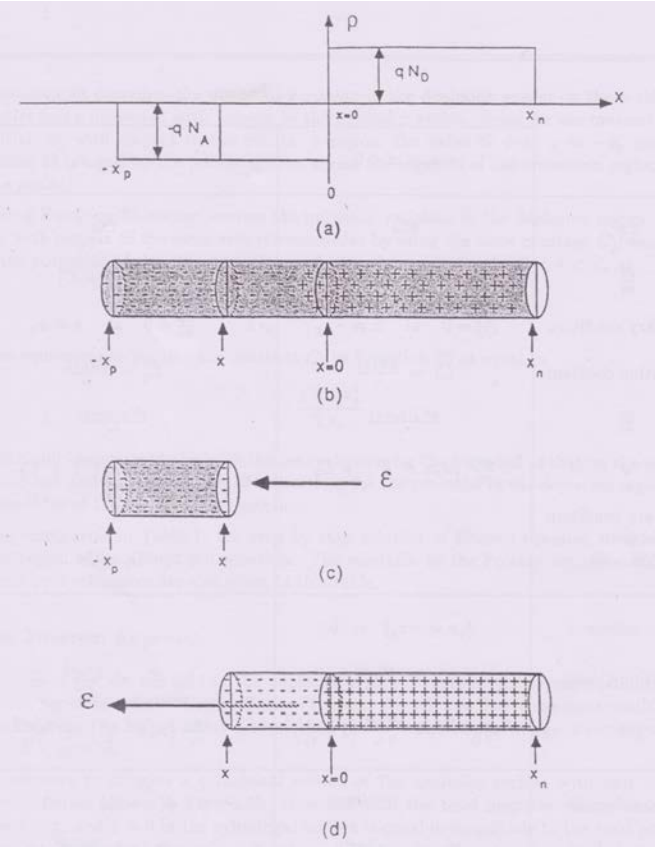


Figure (3.5): The charge distribution in the depletion region of an abrupt *p-n* junction: A) rectangular profile of the charge distribution. B) a cylindrical section of unit cross-sectional area of the entire depletion region. C) the left part of the cylindrical section obtained by partitioning the cylindrical section with a plane at x. We have chosen x to be in the p-side of the depletion region. D) the right part of the cylindrical section.

A review of the p-n junction under thermal equilibrium

Let us now examine the p-n junction under thermal equilibrium i.e., when no external voltage is applied across the junction. Free carriers are present in the two neutral regions with their densities equal to the thermal equilibrium values. In the depletion region, the carrier densities decrease from the thermal equilibrium majority carrier density values at one end of the depletion region to the thermal equilibrium minority carrier density value at the other end. Therefore there is a diffusion of electrons

across the depletion region from the n-side to the p-side, and a diffusion of holes from the p-side to the n-side. The electric current due to the diffusion action is called the ***diffusion current***.

The electric field in the depletion region drives the minority carriers entering the depletion region from one neutral side to the opposite neutral side. For example, electrons entering the depletion region (due to the random thermal motion) from the neutral p-region where they are minority carriers will be driven by the depletion region electric field to the neutral n-region. The holes will similarly be driven from the neutral n-region to the neutral p-region. The resulting electric current is called the ***drift current***. The direction of the drift current balances out exactly the diffusion current and hence the net current through the junction is zero.

This process of diffusion and drift can be understood by reference to Figure (3.6 A). This figure shows in addition to the band diagram of the p-n junction under thermal equilibrium, the electron and hole distributions in the two neutral regions. The carrier distribution functions are rotated by 90 degrees from the way they are drawn in Figure (3.2 C) and Figure (3.3 C) since the variation of energy is vertical in this figure. The electrons in the n-region lying above the dotted line represent those that have the kinetic energy larger than the barrier height and therefore will be able to diffuse to the p-region.

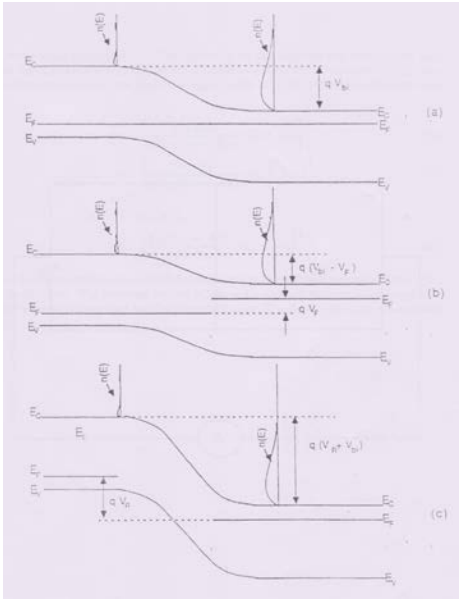


Figure (3.6): Energy band diagram of a p-n junction and the carrier distribution under (A) thermal equilibrium (B) a forward bias and (C) a reverse bias.

P- N Junction under Forward Bias

When an external voltage source such as a battery is connected across the p-n junction such that the positive terminal of the battery is connected to the p-side and the negative terminal is connected to the n-side as shown in Figure (3.7), the p-n junction is said to be forward biased. The potential energy of the electrons on the n-side is increased relative to the potential energy of the electrons on the p-side by an amount equal to qV_F where V_F is the externally applied voltage. The

junction is said to be forward biased under this condition. The band diagram under forward bias is illustrated in Figure (3.6 B). The amount of band-bending is equal to the difference between E_C (the potential energy of the electrons) in the neutral p region and that in the neutral n region: The potential energy barrier height, is now $q(V_{bi} - V_F)$. The barrier height is hence reduced by an amount qV_F under forward bias. In the neutral regions far away from the depletion region, the carrier densities are at their thermal equilibrium values and the Fermi energy is located in the band gap at a level determined by the net impurity (acceptor and donor) concentrations. The Fermi energy in the neutral n region (far away from the depletion region) is shifted upward by qV_F from the Fermi energy in the neutral p region (far away from the metallurgical junction).

The expression for the depletion region width under forward bias can be derived exactly the same way as we did for the thermal equilibrium case either by solving the Poisson equation or by using Gauss theorem. The expression for the depletion region width will have essentially the same form excerpt the potential barrier height V_{bi} will be replaced by (V_{bi}-V_F). Using the same notation for the depletion region width as in the thermal equilibrium case, we have

$$x_p = \left[\frac{2\epsilon_s N_D}{qN_A(N_A+N_D)}(V_{bi} - V_F)\right]^{\frac{1}{2}} \tag{3.55}$$

$$x_n = \left[\frac{2\epsilon_s N_A}{qN_D(N_A+N_D)}(V_{bi} - V_F)\right]^{\frac{1}{2}} \tag{3.56}$$

and

$$x_d = x_n + x_p$$

$$= \left[\frac{2\epsilon_s(N_A+N_D)}{qN_D N_A}(V_{bi} - V_F)\right]^{\frac{1}{2}} \tag{3.57}$$

We denote that the depletion region width is smaller under forward bias conditions than under thermal equilibrium. The potential barrier is less under forward bias conditions, and hence less charge is needed to sustain in a smaller potential barrier. The depletion region has a smaller width.

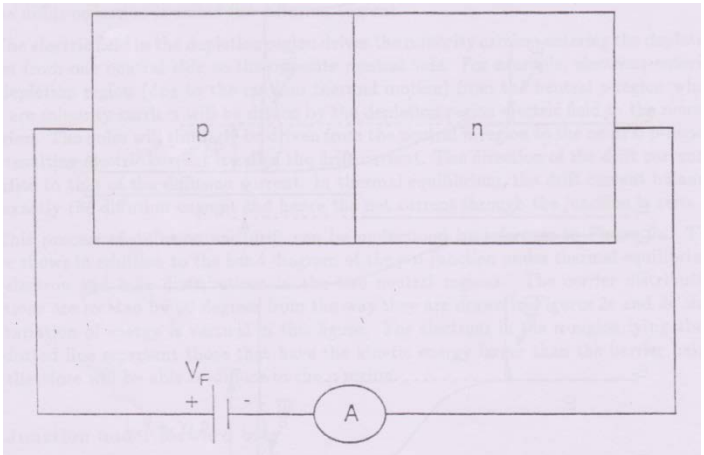


Figure (3.7): A p-n junction under forward bias. V_F is the voltage applied across the p-n

---

Example

Calculate the depletion region width in an abrupt p-n junction with a net impurity density of $N_A = 10^{16}\ cm^{-3}$ on the n-side at room temperature (a) in thermal equilibrium and (b) at a forward applied voltage bias of 0.4 V. Take n_i as equal to $10^{10}\ cm^{-3}$. Using Equation (3.25) we get

$$V_{bi} = 0.0259\ \times\ ln\left(\frac{10^{16}\times10^{15}}{10^{20}}\right) = 0.656\ V$$

(a) Thermal equilibrium:

$$x_p = \left[\frac{2x11.9\times8.854\times10^{-14}\times10^{15}}{1.6x10^{-19}\times10^{16}\times(10^{16}+10^{15})}\right]^{\frac{1}{2}}$$

$$= 3.627\times10^{3}\times\sqrt{\frac{10^{15}}{10^{16}x1.1x10^{16}}}\times\sqrt{0.656}$$

$$= 0.0886\ \times\ 10^{-4}\ \text{cm}$$

Similarly

$$x_n = 3.627\times10^{3}\times\sqrt{\frac{10^{16}}{10^{15}\times1.1\times10^{16}}}\times\sqrt{0.656}$$

$$= 0.886\times10^{-4}\ \text{cm}$$

$$x_d = (0.886+0.0886)\times10^{-4}\ \text{cm}$$

$$= 0.975\ \times\ 10^{-4}\ \text{cm}$$

(b) Forward Bias: The barrier height is now $V_{bi} - V_F$ equal to 0.656 − 0.4= 0.256 V

$$x_p = 3.627\times10^{3}\times\sqrt{\frac{10^{15}}{10^{16}\times1.1\times10^{16}}}\times\sqrt{0.256}$$

$$= 0.553\times10^{-4}\ \text{cm}$$

Similarly

$$x_n = 3.627\times10^{3}\times\sqrt{\frac{10^{16}}{10^{15}\times1.1\times10^{16}}}\times\sqrt{0.256}$$

$$= 0.553\ \times10^{-4}\ \text{cm}$$

$$x_d = (0.553+0.553)\text{x}10^{-4}\ \text{cm}$$

Referring to Figure (3.6 B), the electrons that are distributed above the dotted line have adequate energy to clear the barrier, and diffuse to the other side. This will result in diffusion current. This diffusion current is larger than in thermal equilibrium since there are more electrons with kinetic energy adequate to go over the smaller barrier height. On the other hand, the drift current of electrons will be the same since it does not depend on the barrier height. Similarly more holes will be able to diffuse than in the thermal equilibrium case causing a larger hole diffusion current while the hole drift current will be the same as in thermal equilibrium.

Due to the diffusion of electrons from the n-side to the p-side, the (electron) minority carrier density in the p-side increases from its thermal equilibrium value; hence, this process is called minority carrier injection. Similarly, holes are also injected into the neutral n region from the p region under this process.

We saw earlier that under conditions of no current flow, the electron densities at two different locations x_A and x_B are related by the following expression:

$$n(x_A) = n(x_B)\, e^{\frac{q\psi_{AB}}{kT}} \tag{3.58}$$

and

$$p(x_A) = p(x_B)\, e^{-\frac{q\psi_{AB}}{kT}} \tag{3.59}$$

Where ψ_{AB} is the electrostatic potential difference between x_A and x_B. This is called the **Boltzmann relation**.

In the case of a p - n junction under thermal equilibrium, the condition of zero net current is realized, due to two large, equal and opposite currents. One of these currents is due to diffusion, and the other to drift. However, when a forward bias is applied, a current flows across the junction due to the minority carrier injection process. We will assume that the current through the device under forward bias is very small in comparison with the drift and drift components of current flowing under thermal equilibrium. This is tantamount to assuming that the drift and diffusion components are nearly equal under forward bias. We can therefore assume that the carrier densities are still given by the Boltzmann relation (Equation (3.58)) even under forward bias.

As before, we will choose the origin $x = 0$ at the metallurgical junction. With the boundary of the depletion region on the p - side at $x = -x_p$, and that of the n - side at $x = x_n$, as shown in Figure (3.8). The width of the neutral n region is the distance W_n between the boundary of the depletion region $(x = x_n)$ and the ohmic contact to the neutral n region. The ohmic contact is what is used to apply a voltage to the device and send a current. In our example we have taken the location of the ohmic contact to the neutral n region as $x = x_n + W_n$. Similarly, the ohmic contact to the neutral p region is located at $x = -x_p - W_p$, where W_p is the width of the neutral p region.

The electron densities at $x = x_n$ and $x = -x_p$ are related by the Boltzmann distribution, and therefore,

$$n_p(x = -x_p) = n_n(x = x_n)\, e^{\frac{-q(V_{bi}-V_F)}{kT}}$$

Similarly, the hole density at $x = x_n$ is given by

$$p_n(x = x_n) = p_p(x = -x_p)\, e^{\frac{-q(V_{bi}-V_F)}{kT}}$$

The minority carrier density under thermal equilibrium is n_{po} on the p side, and p_{no} on the n side.

The extra (also called excess) minority carrier densities at $x = x_n$ and at $x = -x_p$ under forward bias are given by

$$\Delta p(x = x_n) = p_n(x = x_n) - p_{no}$$

and

$$\Delta n(x = -x_p) = n_p(x = -x_p) - n_{po}$$

Due to the fact that we have excess minority carriers in the neutral regions, equal amounts of excess majority carriers flow into the neutral regions from their respective ohmic contacts to maintain charge neutrality. Therefore, the majority carrier densities are also increased from their thermal equilibrium values, i.e.,

$$p_p(x = -x_p) = p_{po} + \Delta p(x = -x_p) = p_{po} + \Delta n(x = -x_p) \qquad (3.60)$$

and

$$n_{n0}(x = x_n) = n_{no} + \Delta n(x = x_n) = n_{no} + \Delta p(x = x_n) \qquad (3.61)$$

We distinguish two cases: one called high injection, and the other called low injection.

When

$$\Delta n(x = -x_p) \text{ is } > p_{po}\,,$$

or when

$$\Delta p(x = x_n) \text{ is } > n_{no}\,,$$

we call it a high injection condition. When

$$\Delta n(x = -x_p) < p_{po}, \text{ i.e., } p_p(x = -x_p) \approx p_{po} \quad (\text{see Eq. (3.60)})$$

and when

$$\Delta p(x = x_n) < n_{no}, \text{ i.e., } n_n(x = x_n) \approx n_{no} \quad (\text{see Eq. (3.61)})$$

we have a low and moderate injection condition. We will restrict our discussion to low and moderate injection cases only. Therefore,

$$n_n(x = x_n) \approx n_{no} \quad \text{and} \quad p_p(x = -x_p) \approx p_{po} \tag{3.62}$$

The expression for $n_p(x = -x_p)$ can be simplified as

$$\begin{aligned} n_p(x = -x_p) &= n_n(x = x_n)\, e^{\frac{-qV_{bi}}{kT}} e^{\frac{qV_F}{kT}} \\ &= n_{no}\, e^{\frac{-qV_{bi}}{kT}}\, e^{\frac{qV_F}{kT}} \\ &= n_{po}\, e^{\frac{qV_F}{kT}} \end{aligned} \tag{3.63}$$

In the above equation, we made use of the Boltzmann relation for the thermal equilibrium carrier densities, i.e.,

$$n_{po} = n_{no}\, e^{\frac{-qV_{bi}}{kT}}$$

Similarly, the expression for $p_n(x = x_n)$ can be obtained as

$$p_n(x = x_n) = p_{no} e^{\frac{qV_F}{kT}} \tag{3.64}$$

The excess minority carrier densities at $x = x_n$ and $x = -x_p$ are obtained as

$$\Delta p(x = x_n) = p_{no} e^{\frac{qV_F}{kT}} - p_{no} = p_{no}\left(e^{\frac{qV_F}{kT}} - 1\right) \tag{3.65}$$

and

$$\Delta n(x = -x_p) = n_{po} e^{\frac{qV_F}{kT}} - n_{po} = n_{po}\left(e^{\frac{qV_F}{kT}} - 1\right) \tag{3.66}$$

These excess minority carrier charges will now diffuse into the two respective neutral regions. Before we study the diffusion process, we will discuss another concept called the ***Quasi-Fermi Level***.

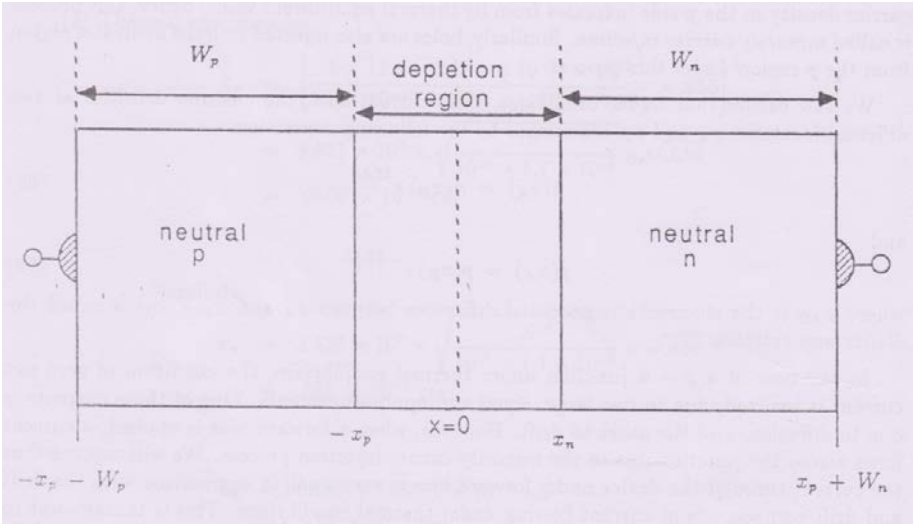


Figure (3.8): The depletion and neutral regions in a forward biased p-n junction

Example

Let us calculate the excess carrier density in the neutral p and regions of a p-n junction, at the boundary of the depletion region. Assume $N_A = 10^{16} cm^{-3}$ on the p-side, and $N_D = 5 \text{ x} 10^{16}\ cm^{-3}$ on the n-side. Assume a forward bias voltage of 0.4 *V*.

$$p_{po} \approx 10^{16} cm^{-3} \text{ and } \quad n_{po} = \frac{n_i^2}{p_{p0}} \times \frac{10^{20}}{10^{16}} = 10^4 cm^{-3}$$

$$n_{n0} = 5 \times 10^{15} cm^{-3} \text{ and } p_{n0} = \frac{n_i^2}{p_{p0}} = \frac{10^{20}}{5 \times 10^{15}} = 2 \times 10^4 cm^{-3}$$

$$\Delta p(x = x_n) = n_{po}\left(e^{\frac{qV_F}{kT}} - 1\right)$$

$$\frac{kT}{q} = 0.0259\ V$$

$$\Delta p(x = x_n) = 10^4\left(e^{\frac{0.4}{0.0259}} - 1\right) 10^4 \times [5.1 \times 10^6 - 1] \approx 5.1 \times 10^{10} cm^{-3}$$

$$\Delta n(x = -x_p) = 2x10^4 \times \left(e^{\frac{0.4}{0.0259}} - 1\right) \times 10^4 x 5.1 \times 10^6 = 1.02 \times 10^{11} cm^{-3}$$

Quasi-Fermi Level

In thermal equilibrium, we expressed the minority carrier density in terms of Fermi energy. However, under forward bias it will not be possible to do so. We therefore define a parameter called the ***quasi-Fermi energy***, to enable us to write the expression for the minority carrier density in the same form as in the thermal equilibrium case.

We define a quasi-Fermi energy level, E_{F_n}, for electrons such that the electron density can be written as

$$n_p = n_i e^{\frac{(E_{F_n} - E_i)}{kT}}$$

Similarly, the hole density can be written as

$$p_n = n_i e^{-\frac{(E_{F_p} - E_i)}{kT}}$$

where E_{F_p} is the quasi-Fermi energy level for holes. The quasi-Fermi levels for electrons and holes are plotted in Figure (3.9). Since we are considering low and moderate injection conditions, the majority carrier densities are nearly the same as the thermal equilibrium values, and the majority carrier quasi-Fermi levels are nearly the same as Fermi energy under thermal equilibrium conditions. We make a further assumption that the quasi-Fermi levels in the depletion region of the forward biased *p-n* junction is at the same level as the majority carrier quasi-Fermi level. This is illustrated in Figure (3.9). At distances far from the junction the excess carrier densities would become zero, as discussed in the next section. Hence, the quasi-Fermi level for the minority carriers would be at the same level at the thermal equilibrium Fermi energy. For example, at $x = -x_p$, E_{F_n} is at a higher level than the Fermi energy under thermal equilibrium, and gets lower and lower and approaches the thermal, and gets lower and lower and approaches the thermal equilibrium value as one proceeds further into the neutral *p* region.

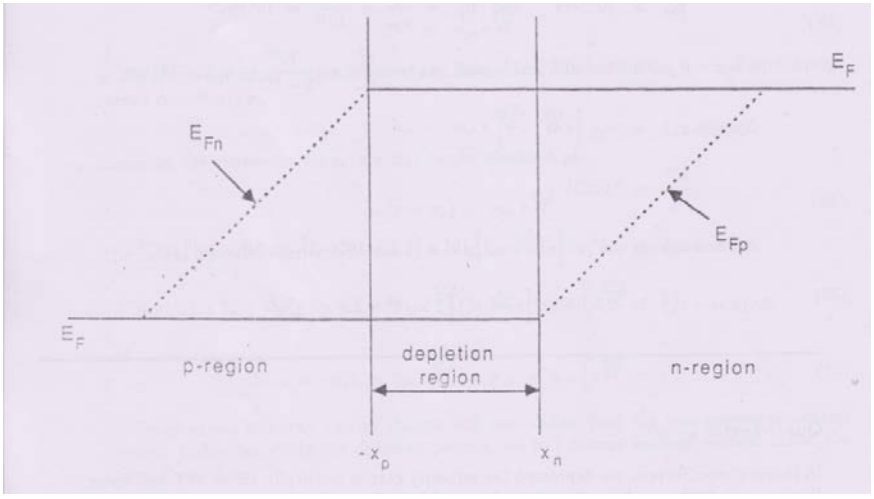


Fig. (3.9): Quasi-Fermi Levels in a forward biased p-n junction

Minority Carrier Diffusion

The diffusion of the excess minority carriers in the neutral regions can be analyzed by setting up the continuity equation. Let us first consider the neutral *n* region. The continuity equation is given by

$$\frac{\partial \Delta p}{\partial t} = -\frac{\Delta p}{\tau_p} - \frac{1}{q}\frac{dJ_p}{dx} \tag{3.67}$$

We assume that there is no electric field in the neutral *n* region, and hence, the hole current is only due to diffusion.

$$J_p = -qD_p \frac{\partial p}{\partial x} = -qD_p \frac{\partial \Delta p}{\partial x}$$

Substituting this into the continuity equation, we get

$$\frac{\partial(\Delta p)}{\partial t} = -\frac{\Delta p}{\tau_p} + D_p \frac{\partial^2(\Delta p)}{\partial x^2} \tag{3.68}$$

Under D.C. conditions (i.e. when Δp does not vary with time, this equation is simplified by setting the left hand side as equal to zero, and rewriting it as:

$$D_p \frac{\partial^2(\Delta p)}{\partial x^2} = \frac{\Delta p}{\tau_p}$$

or

$$\frac{\partial^2(\Delta p)}{\partial x^2} = \frac{\Delta p}{D_p \tau_p} \tag{3.69}$$

By considering the dimensions of D_p and τ_p, it can be seen that the term $D_p \tau_p$ has units of square length. We define a parameter L_p as given by

$$L_p{}^2 = D_p \tau_p \tag{3.70}$$

The equation of continuity is now written as

$$\frac{\partial^2(\Delta p)}{\partial x^2} = \frac{\Delta p}{L_p{}^2}$$

Let us define a new variable $x' = x - x_n$, (i.e. we choose the origin of the horizontal axis at $x = x_n$.) The continuity equation, written in terms of x', is given by

$$\frac{\partial^2(\Delta p)}{\partial x^2} = \frac{\Delta p}{L_p{}^2} \tag{3.71}$$

The general solution to this equation is

$$\Delta p(x') = Ae^{\frac{-x'}{L_p}} + Be^{\frac{-x'}{L_p}} \tag{3.72}$$

A and B are integration constants that will be determined by the boundary conditions at $x' = 0$ and $x' = W_n$. The boundary condition at $x' = 0$ is given by

$$\Delta p(x' = 0) = p_{no}(e^{\frac{qV_F}{kT}} - 1)$$

according to Equation (3.65).

The excess carrier density at an ohmic contact is by definition zero. Hence, the boundary condition at $x' = W_n$ is

$$\Delta p(x' = W_n) = 0$$

Applying these boundary conditions to determine A and B, we can arrive at the solution to the continuity equation, as done in the next chapter, and obtain

$$\Delta p(x') = \Delta p(o)\frac{\sinh(\frac{W_n - x'}{L_p})}{\sinh(\frac{W_n}{L_p})} \tag{3.73}$$

Where $\Delta p(o)$ is defined as $\Delta p(x' = 0)$ However, for the present we limit our discussion to two limiting approximations

$$W_n >> L_p \quad \text{and} \quad W_n << L_p$$

Wide Base Case ($W_n >> L_p$)

When $W_n >> L_p$, we call it a wide base case. Let us consider Equation (3.72), which gives the solution to the continuity equation. If we evaluate Δp at ($x' = W_n$), we get

$$\Delta p(W_n) = Ae^{\frac{-W_n}{L_p}} + Be^{\frac{W_n}{L_p}}$$

Where $\Delta p(W_n)$ is Δp evaluated at $x' = W_n$. Since $W_n >> L_p$, the first term on the right hand side is negligible and hence

$$\Delta p(W_n) = Be^{\frac{W_n}{L_p}}$$

But our boundary condition at the ohmic contact requires $\Delta p(W_n)$ to be zero and $e^{\frac{W_n}{L_p}}$ is a large quantity. In order to satisfy the boundary condition, we set $B = 0$. The solution to the continuity equation then becomes

$$\Delta p(x') = Ae^{\frac{-x'}{L_p}}$$

If we apply the boundary condition at $x' = 0$, we can evaluate A and obtain

$$\Delta p(x') = \Delta p(o)e^{\frac{-x'}{L_p}} \tag{3.74}$$

where $\Delta p(o)$ is $\Delta p(x' = o) = p_{no}(e^{\frac{qV_F}{kT}} - 1)$, as stated earlier.

The plot of Δp as a function of distance as given by Equation (3.74), is given in Figure (3.10 A). We find that the excess minority carrier density decays exponentially with distance. What is the physical interpretation? A hole that is injected from the left at $x' = 0$ diffuses to the right, and as it travels in the neutral n region, it recombines with an electron. Hence, the excess minority carrier density decreases with x'. We can now explain why it decays exponentially.

Let $P(x')dx'$ be the probability that an excess hole injected at $x' = 0$, will recombine in an elementary distance dx' between x' and $x' + dx'$. Then $P(x')$ is the probability per unit distance that an excess minority carrier (hole) will recombine and be lost. The number of excess holes that recombine (and therefore be lost) in the interval dx' is equal to the product $\Delta p(x)$, i.e., the number of excess hole

density at x', multiplied by $P(x')\ dx$'. The difference in excess hole density at x' and at $x' + dx$' is what is lost due to the recombination in the interval dx'. Hence

$$d\big(\Delta p(x')\big) = \Delta p(x' + dx') - \Delta p(x') = -\Delta p(x')P(x')dx'$$

Assuming $P(x')$ does not vary with distance and is equal to P_o, this equation can be rewritten as

$$\frac{d(\Delta p(x'))}{\Delta p(x')} = -P_o dx' ,$$

and directly integrated to obtain

$$\Delta p(x') = A' e^{-P_o x'}$$

The constant A' is obtained as $\Delta P(0)$ by applying the boundary condition at $x' = 0$:

$$\Delta p(x') = \Delta p(o)\ e^{-P_o x'}$$

By comparing the above to Equation (3.74), we see that

$$P_o = \frac{1}{L_p}$$

This means that $\frac{1}{L_p}$ is the probability per unit distance that a hole will recombine. That is, $\frac{dx'}{L_p}$ is the probability that an excess hole will recombine in an interval dx'

Using this argument, it can be shown that L_p is the average distance an excess hole (minority carrier) will travel before it recombines. (This is left as a homework problem.) For this reason, L_p is called the ***minority carrier diffusion length***.

Figure (3.10 A) gives the profile of the excess minority carrier density. As mentioned before in order to maintain charge neutrality, excess majority carrier density equal to the excess minority carrier density will be present in the neutral n region, with a profile exactly equal to that of the minority carrier density. This is illustrated in Figure (3.10 B).

Due to the concentration variation of the excess minority carrier density, a diffusion current will flow. The resulting electric current density is given by

$$\begin{aligned} J_p(x') &= -\mathrm{q}D_p \frac{d(\Delta p)}{dx'} \\ &= \mathrm{q}D_p \frac{\Delta p(o)}{L_p} e^{\frac{-x'}{Lp}} \end{aligned} \qquad (3.75)$$

The diffusion current also decays exponentially with distance.

While there is a diffusion current due to the exponential decay of the minority carrier density, there will be no similar majority carrier diffusion current arising from the exponential variation of the excess majority carrier density. The reason for this is that the excess majority carrier density is

maintained by the presence of a small electric field, whose value is proportional to the ratio of $\Delta n(x')$ and (x') . $n(x')$ is equal to $N_D + \Delta n(x')$ where N_D is the donor density in the n region. The drift current of electrons due to this electric field, given by $J_{drift} = qn(x')\mu_n \mathcal{E}(x')$, is exactly balanced by the excess electron diffusion current, given by $J_{diff} = qD_n \frac{d\Delta n}{dx'}$. In the case of minority carriers, there is a drift current due to this electric field, but it is negligibly small compared to the diffusion current. This is the reason why we take into account the minority carrier diffusion current, and not the majority carrier diffusion current.

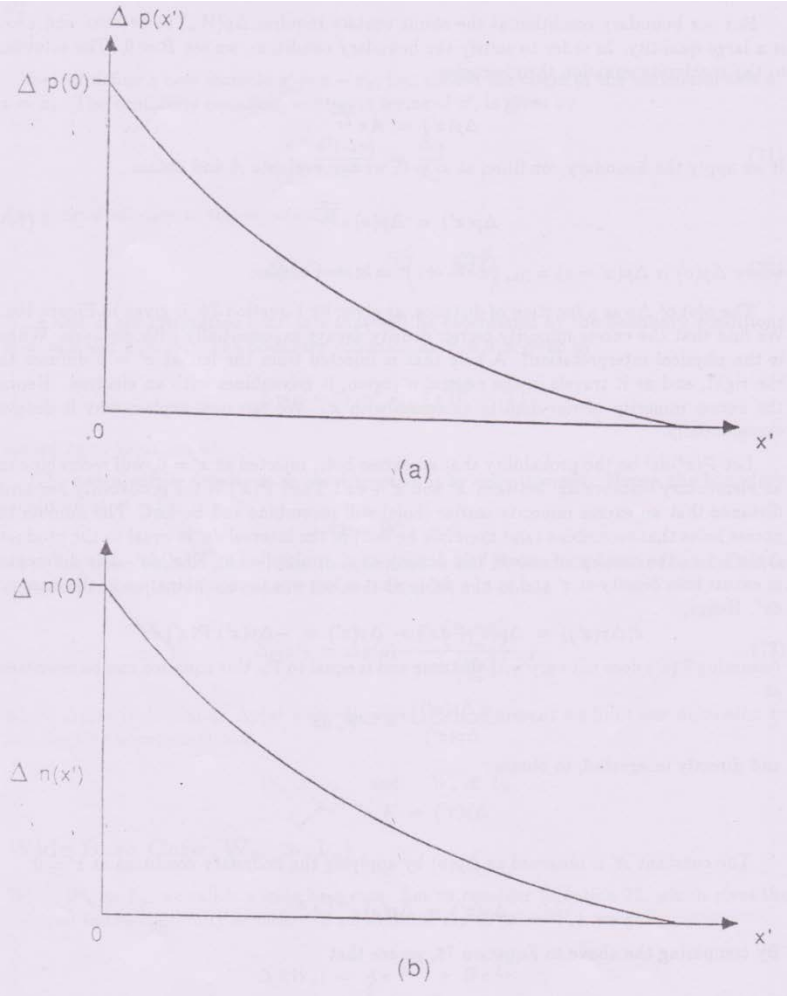


Figure (3.10): Plot of the excess carrier density in the neutral n region for a wide base case: (A) Plot of the excess minority carrier (hole) density, and (B) Plot of the excess majority carrier (electrons) density

In Figure (3.11), the minority carrier diffusion current is plotted as a function of distance in the neutral n region. For each hole lost due to recombination, there must also be the loss of an electron. Therefore, electrons must flow into the n region in the opposite direction from the ohmic contact to replace the electrons lost due to recombination. Since $W_n >> L_p$, all the injected holes are lost, due to recombination, and the hole diffusion current becomes zero at large distances. The (recombinating) electron current density flowing from the ohmic contact into the n region should be equal to the injected hole current density at $x' = 0$. The recombinating electron current density falls off as

$(1 - e^{\frac{-x'}{L_p}})$ with a decrease in x'. The total electric current density (due to the injection of minority carriers in the n region) is the sum of injected hole diffusion current density and the recombining electron current density and is constant in the n region as shown in Figure (3.11). The total electric density is equal to the injected hole current density at $x' = 0$, where the recombining electron current density is zero, and is therefore obtained by putting $x' = 0$ in Equation (3.75), as

$$J_p = \mathrm{q}D_p \frac{\Delta p(0)}{L_p} = \frac{\mathrm{q}D_p p_{no}}{L_p}(e^{\frac{qV_F}{kT}} - 1) \tag{3.76}$$

The above equation gives us the current density flowing in the *p-n* junction due to the injection of minority carriers (holes) in the n region.

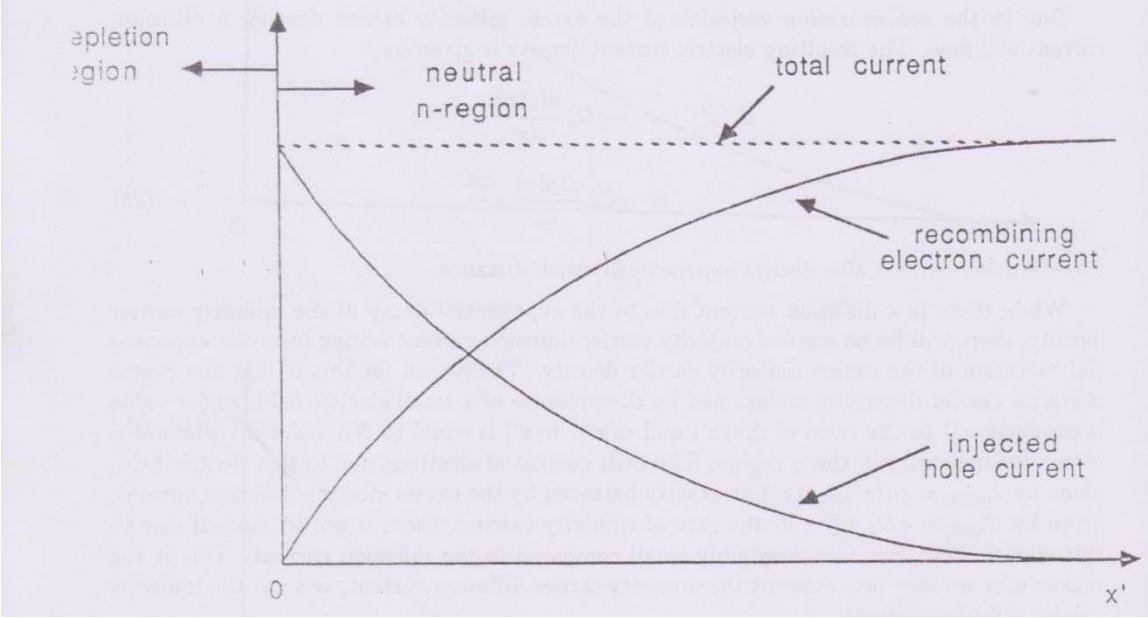


Figure (3.11): Injected hole diffusion current density and the recombining electron current density in the neutral n region

We can similarly treat the injection of minority carriers (electrons) in the p region. We will find that the injected minority carrier (electron) density decays exponentially with distance in the neutral p region, and the minority carrier diffusion length, L_n, will be given by $L_n = \sqrt{D_n \tau_n}$ in the neutral p region. The current density through the junction, due to injection of electrons (minority carriers) in the p region, will be given by

$$J_n = \frac{\mathrm{q}D_n n_{po}}{L_p}(e^{\frac{qV_F}{kT}} - 1)$$

The total electric current density through the junction will be the sum of J_s in the equation above and J_p in Equation (3.76) and is given by

$$J = J_n + J_p = [\frac{qD_n n_{po}}{L_n} + \frac{qD_p p_{no}}{L_p}](e^{\frac{qV_F}{kT}} - 1)$$

(3.77)

This is usually written as

$$J = J_s\,(e^{\frac{qV_F}{kT}} - 1)$$

where

$$J_s = \frac{qD_n n_{po}}{L_n} + \frac{qD_p p_{no}}{L_p} \tag{3.78}$$

J_s is called the ***saturation current density***.

Figure (3.12) shows the various components of current that flow in the forward biased *p-n* junction. J_{p2} is the injected hole diffusion current density. J_{p3} is the hole current density that is flowing in the neutral p region, in order to inject holes in the n region. J_{n1} is the electron current density in the neutral n region to replace the electrons lost due to recombination with holes in the neutral n region. Similarly, J_{n2} is the injected electron diffusion current density in the neutral p region. J_{n3} is the electron current density flowing in the neutral n region to inject electrons in the p region. J_{p1} is the recombining hole current density in the neutral p region. The total electric current density, J_t , is the sum of all these various components.

The current flowing through a forward biased junction is obtained by multiplying the current density by the area A of the junction

$$I = I_s\,(e^{\frac{qV_F}{kT}} - 1) \tag{3.79}$$

where

$$I_s = AJ_s = A[\frac{qD_n n_{po}}{L_n} + \frac{qD_p p_{no}}{L_p}] \tag{3.80}$$

I_s is called the saturation current. The current in the forward biased junction is plotted in Figure (3.13). At values of V_F larger than a few kT, it can be seen from Equation (3.79) that the term -1 is negligible in comparison with the exponential term, and hence, *I* is given by

$$I \approx I_s\,(e^{\frac{qV_F}{kT}}) \tag{3.81}$$

The current increases exponentially with the forward bias voltage V_F.

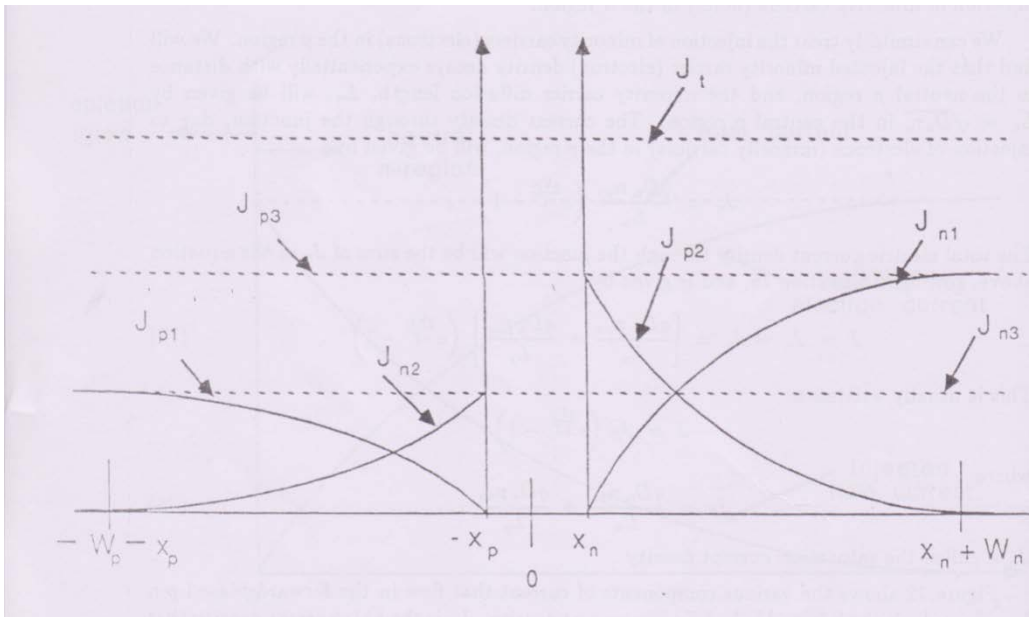


Figure (3.12): Various components of electric current density in the forward biased p-n junction. The arrows indicate the direction in which the electrons or hole flow.

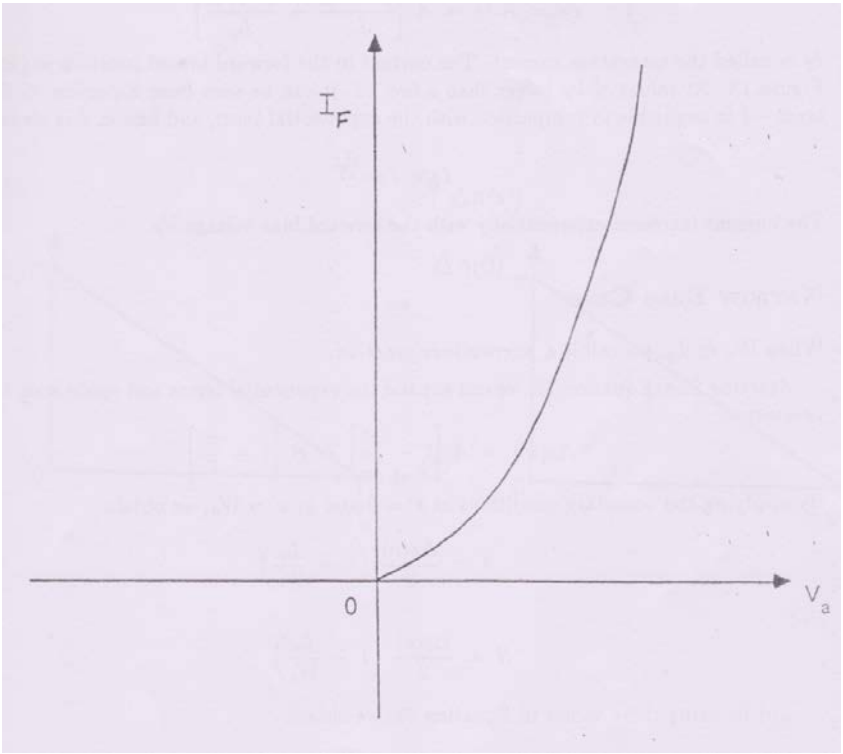


Figure (3.13): Forward Bias current through the junction

Narrow Base Case

When $W_n << L_p$, we call it a narrow base junction.

Starting from Equation (3.72), we can expand the exponential terms and retain only the first two terms:

$$\Delta p(x') = A\left[1 - \frac{x'}{L_p}\right] + B\left[1 + \frac{x'}{L_n}\right] \tag{3.82}$$

By applying the boundary conditions at $x' = 0$ and at $x' = W_n$, we obtain

$$A = \frac{\Delta p(0)}{2}\left(1 + \frac{L_p}{W_n}\right)$$

and

$$B = \frac{\Delta p(0)}{2}\left(1 - \frac{L_p}{W_n}\right)$$

Substituting these values in Equation (3.82), we obtain

$$\Delta p(x') = \Delta p(0)\left(1 - \frac{x'}{W_n}\right) \tag{3.83}$$

The excess minority carrier (hole) density in the neutral n region is plotted in Figure (3.14 A). As stated before, a neutralizing excess majority carrier density $\Delta n(x')(= \Delta p(x'))$ also exists and is plotted in Figure (3.14 B). The injected hole diffusion current density is obtained from Equation (3.83) as

$$J_p(x') = -qD_p \frac{d(\Delta p)}{dx'}$$

$$= \frac{qD_p \Delta p(0)}{W_n}$$

$$= \frac{qD_p p_{n0}}{W_n}\left(e^{\frac{qV_F}{kT}} - 1\right)$$

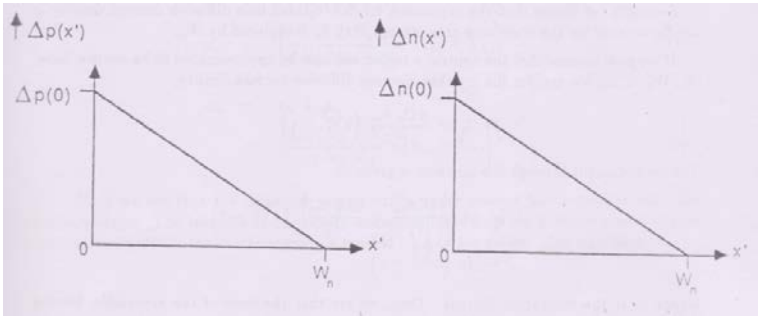


Figure (3.14): Excess carrier density in the neutral n region of a forward biased p-n junction for a narrow base approximation. (A) excess minority carrier (hole) density, and (B) excess majority carrier (electron) density

First, we notice that the injected hole diffusion current density is constant and independent of x'. This means that any hole current injected at $x' = 0$ from the p region reaches the ohmic contact. There is no recombination of minority carriers in the n region. Since W_n is very small in comparison with the diffusion length L_p, the probability of an injected hole recombining in the n region is negligible. Hence the injected hole current density reaches the ohmic contact without any attenuation.

Secondly, we notice that the expression for the injected hole diffusion current density is similar to that for the wide base case, except that L_p is replaced by W_n.

If we now assume that the neutral p region can also be approximated to be narrow base, i.e., $W_p \ll L_n$, we get, for the injected electron diffusion current density,

$$J_n = \frac{qD_n n_{p0}}{W_p}\left(e^{\frac{qV_F}{kT}} - 1\right)$$

The total current through the junction is given by

$$I = A\left[\frac{qD_n n_{p0}}{W_p} + \frac{qD_p p_{n0}}{W_n}\right]\left(e^{\frac{qV_F}{kT}} - 1\right)$$

$$= I_S\left(e^{\frac{qV_F}{kT}} - 1\right)$$

where I_S is the saturation current. Thus, we see that the form of the expression for the current under forward bias of a narrow base p-n junction is similar to that for a wide base junction.

It should be noted that it is possible that in a p-n junction, one side can be approximated as narrow base, and the other as wide base. In such a case, the current through the junction will be expressed as an appropriate combination of narrow base and wide base expressions.

P-N Junction under Reverse Bias

When an external voltage is connected such that the positive terminal is connected to the n-side and the negative terminal to the p-side of the junction, then the junction is said to be reverse-biased. The potential energy of the electrons on the p-side is increased relative to that of the electrons on the n- side by an amount equal to qV_R where V_R is the externally applied reverse bias voltage. The band diagram in a p-n junction under reverse bias is illustrated in Figure (3.6 C). The height of the potential energy barrier as before is the difference between E_C on the neutral p-side and E_C on the n side and therefore equal to $q(V_{bi} + V_R)$. The band-bending, i.e., the barrier height is thus increased from the thermal equilibrium value of qV_{bi} to $q(V_{bi} + V_R)$.

In the neutral regions far away from the junction, the carrier densities are at their thermal equilibrium values and the Fermi energy is located in the band gap at a level determined by the net impurity (acceptor and donor) concentrations. Therefore, the Fermi energy in the neutral region (far away from the junction) is shifted downward by qV_R from the Fermi energy in the neutral p region (far away from the junction).

The expression for the depletion region width under reverse bias can be derived exactly the same way as we did earlier either by solving the Poisson equation or by using Gauss theorem. The expression for the depletion region width will have essentially the same form except the potential barrier height V_{bi} will be replaced by $(V_{bi} + V_R)$. Using the same notation for the depletion region width as earlier, we have

$$x_p = \left[\frac{2\epsilon_s N_D}{qN_A(N_A+N_D)}(V_{bi} + V_R)\right]^{\frac{1}{2}} \tag{3.84}$$

$$x_n = \left[\frac{2\epsilon_s N_A}{qN_D(N_A+N_D)}(V_{bi} + V_R)\right]^{\frac{1}{2}} \tag{3.85}$$

and

$$x_d = x_n + x_p$$

$$= \left[\frac{2\epsilon_s(N_A+N_D)}{qN_D N_A}(V_{bi} + V_R)\right]^{\frac{1}{2}} \tag{3.86}$$

Thus we see that the depletion region width under reverse bias increases with the reverse voltage. The increase is physically understandable because a larger space charge (and hence a wider space charge region) is needed for a larger potential energy barrier.

Example

Calculate the depletion region width in an abrupt p-n junction with a net impurity density of $N_A = 10^{16} cm^{-3}$ on the p-side and $N_D = 10^{15} cm^{-3}$ on the n-side. Assume room temperature and a reverse bias voltage of 6 V. We determined the depletion region width for this diode under thermal equilibrium case and under a forward bias of 0.4 V in the example worked earlier in the section on forward bias. We can therefore use the value, 0.656 V for V_{bi} from the previous example.

$$V_{bi} + V_R = 0.656 + 6 = 6.656\, V$$

$$x_p = 3.627 \times 10^3 \times \sqrt{\frac{10^{15}}{10^{16} \times 1.1 \times 10^{16}}} \times \sqrt{6.656}$$

$$= 0.282 \times 10^{-4}\ \text{cm}$$

Similarly

$$x_n = 3.627 \times 10^3 \times \sqrt{\frac{10^{16}}{10^{15} \times 1.1 \times 10^{16}}} \times \sqrt{6.656}$$

$$= 2.81 \times 10^{-4} \text{ cm}$$

$$x_d = (2.821 + 0.282) \times 10^{-4} \text{ cm}$$
$$= 3.103 \times 10^{-4} \text{ cm}$$

Due to the increase in the barrier height, the hole (minority carrier) density in the neutral n region at the boundary of the depletion region, i.e., $x = x_n$, is reduced from its thermal equilibrium value since every hole that enters the depletion region will be pulled towards the p side by the electric field in the depletion region. Another way of stating this is that while there was a balance between the drift and the diffusion currents in thermal equilibrium, under reverse bias condition the diffusion current is reduced due to the increase in the barrier height. Thus only the drift current remains. Yet another way of looking at this is by considering the carrier density at the edges of the depletion region using the Boltzmann relation in the following manner:

$$p(x = x_n) = p\big(x = -x_p\big)e^{\frac{-q(V_{bi}+V_R)}{kT}} = p_{p0}e^{\frac{-q(V_{bi}-V_F)}{kT}} = p_{n0}e^{\frac{-qV_R}{kT}} \qquad (3.87)$$

We see that when qV_R is much larger than kT, p at $x = x_n$ is nearly zero. Using the above relation we see that the excess hole density at $x = x_n$ is negative and is given by

$$\Delta p(x = x_n) = p_{n0}e^{\frac{-qV_R}{kT}} - p_{n0} = p_{n0}\left(e^{\frac{-qV_R}{kT}} - 1\right) \qquad (3.88)$$

We can now set up the continuity equation which will be the same as Equation (3.68). Again setting up a new variable x', as we did in the forward bias case we obtain the same equation as Equation (3.71) which is given below

$$\frac{d^2(\Delta p)}{dx'^2} = \frac{\Delta p}{L_p^2}$$

Solving this equation under the boundary condition that the excess carrier density is equal to 0 at x' $= W_n$ and is equal to $p_{n0}\left(e^{\frac{-qV_R}{kT}} - 1\right)$ at $x' = 0$, we get for the wide base diode,

$$\Delta p(x') = p_{n0}\left(e^{\frac{-qV_R}{kT}} - 1\right) e^{-\frac{x'}{L_p}} \qquad (3.89)$$

and for the narrow base diode,

$$\Delta p(x') = p_{n0}\left(e^{\frac{-qV_R}{kT}} - 1\right) e^{-\frac{x'}{L_p}}\left(1 - \frac{x'}{W_n}\right) \qquad (3.90)$$

When we consider the electron density in the neutral p region, we get similar results. The excess minority carrier densities in the neutral n and p regions are negative and are plotted in Figure (3.15).

We will find that the diffusion current is of the same form as we got for the forward-biased junction except that V_F will be replaced by – V_R. The current through the diode will be similar in form as the forward bias current but opposite in direction. The current is called the reverse current, I_{rev}.

$$I_{rev} = I_s\left(e^{\frac{-qV_R}{kT}} - 1\right) \tag{3.91}$$

In the case of wide base diode,

$$I_s = AJ_s = A\left[\frac{qD_n n_{p0}}{L_n} + \frac{qD_p p_{n0}}{L_p}\right] \tag{3.92}$$

I_s is the same as what we obtained under forward bias for a wide base diode. In the case of the narrow base diode,

$$I_s = AJ_s = A\left[\frac{qD_n n_{p0}}{W_p} + \frac{qD_p p_{n0}}{W_n}\right] \tag{3.93}$$

I_s is again the same as what we obtained under forward bias for a narrow base diode.

The current in the reverse biased junction is plotted in Figure (3.16). At values of V_R larger than a few kT, it can be seen from Equation (3.91) that the exponential term is negligible in comparison with the term -1 and hence, I approaches or saturates to a value

$$I = -I_s \tag{3.94}$$

and is independent of the reverse voltage. This is the reason why I_s is called the saturation current.

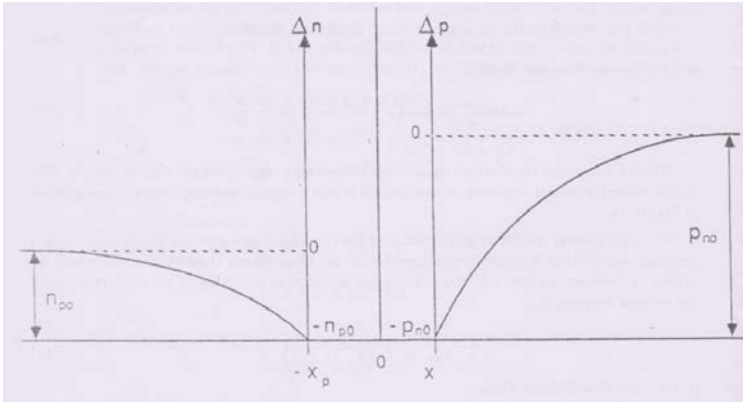


Figure (3.15): The Plot of the excess (depleted) carrier density in the reverse biased junction.

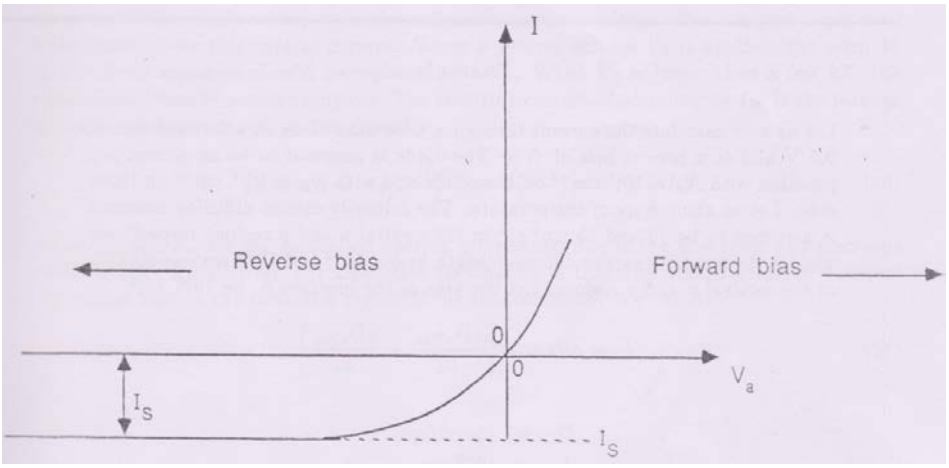


Figure (3.16): Reverse current in the p-n junction.

Example

Let us now calculate the current through a wide-base diode at a forward bias of $0.5\ V$ and at a reverse bias of $0.5\ V$: The diode is assumed to be an abrupt p-n junction with $N_A = 10^{16} cm^{-3}$ on the p-side and with $N_D = 10^{15} cm^{-3}$ on the n side. Let us assume room temperature. The minority carrier diffusion constant is assumed to be 10 and 25 $cm^2\ s^{-1}$ in the neutral n and p regions respectively and let the minority carrier diffusion length be 5×10^{-3} and 10^{-2} cm respectively in the neutral n and p regions. Let the area of the junction A, be 10^{-3} cm^2.

$$I_s = AJ_s = A\left[\frac{qD_n n_{p0}}{L_n} + \frac{qD_p p_{n0}}{L_p}\right]$$

$$D_n = 25\ cm^2 s^{-1}$$

$$L_n = 10^{-2}\ cm$$

$$n_{p0} = \frac{10^{20}}{10^{16}} = 10^4 cm^{-3}$$

$$D_p = 10\ cm^2 s^{-1}$$

$$L_p = 5x10^{-3}\ cm$$

$$p_{n0} = \frac{10^{20}}{10^{15}} = 10^5 cm^{-3}$$

Substituting these values in the equation for I_s we get

$$I_s = 3.60 \times 10^{-14}\ A$$

$$I_F = 3.60\ x10^{-14} \times \left(e^{\frac{0.5}{0.02589}} - 1\right) = 9.41 \times 10^{-6} A$$

$$I_R = 3.60\,x10^{-14} \times \left(e^{-\frac{5}{0.02589}} - 1\right) = -3.60\ \times\ 10^{-14} A$$

Interpretation of Reverse Current

We saw in the last section that the current through the diode is given by

$$I = qA\left[\frac{D_p p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n}\right]\left(e^{\frac{qV_F}{kT}} - 1\right) \tag{3.95}$$

for a wide base diode where V_F is the externally applied voltage. We can give a physical interpretation for this reverse current. When a reverse voltage V_R is applied, the term V_F in the above equation should be replaced by $-V_R$. When V_R is large than a few kT, the exponential term becomes negligible. The resulting current, I, denoted by I_R, is the leakage or saturation current -I_S and is given by

$$I_R = -I_S = -qA\left[\frac{D_p p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n}\right] = -I_{sp} - I_{sn} \tag{3.96}$$

where I_{sp} and I_{sn} are the saturation current components due to the flow holes and electrons respectively. Recall from our discussion in Chapter 2 that the minority carrier thermal generation rate in the neutral n region of the semiconductor is given by

$$g_{th} = \frac{p_{n0}}{\tau_p} \tag{3.97}$$

Also

$$L_p = \sqrt{D_p \tau_p} \tag{3.98}$$

We can therefore write

$$\frac{D_p p_{n0}}{L_p} = \frac{D_p L_p p_{n0}}{L_p^2} = \frac{L_p p_{n0}}{\tau_p} = L_p \times\ g_{th} \tag{3.99}$$

The reverse current can therefore be expressed as

$$I_R = -I_S = -I_{sp} - I_{sn} = -q\ A\left(g_p L_p - g_n L_n\right) \tag{3.100}$$

Where g_n and g_p are the thermal generation rates of electrons and holes in the neutral p and n regions respectively.

What is the physical interpretation of this equation? A minority carrier that enters the depletion region will be swept away to the other side due to the electric field in the depletion region. This gives rise to the reverse current. For example, an electron that enters the depletion region from the p side where it is a minority carrier will be pushed to the n side by the electric field in the depletion region. Similarly a hole will be pushed from the n side to the p side. Since a minority carrier will travel on an average a distance equal to the diffusion length before it recombines, carriers generated in the neutral region farther than a diffusion length away from the edge of the depletion region will not reach the depletion region. It is only those minority carriers that are generated within a diffusion length that will

produce the reverse current. The first term in the above equation represents the current flowing through the junction due to minority carriers (holes) generated in the neutral n region within a diffusion length L_P from the depletion region and similarly the second term represents the current due to the minority carriers (electrons) generated in the neutral p region within a diffusion length L_n from the depletion region. Such a physical interpretation is not readily possible for the reverse current in the narrow base diode.

Non-Ideal Current Characteristics

The current characteristics that we discussed so far are called the ideal current characteristics. We will now discuss additional components of current that flow in a p - n junction and these additional components of current are called non-ideal currents. In our treatment we neglected the component of current due to generation of electron-hole pairs in the depletion region in the reverse current and the component of current due to recombination of electron-hole pairs in the depletion region in the forward current.

Reverse Bias

In the depletion region of the reverse-bias junction, the $n\ p$ product is less than n_i^2. Recall from our discussion in Chapter 2, that under this condition, there is a net generation in the depletion region equal to

$$g_{th} = \frac{n_i}{2\tau_g} \tag{3.101}$$

Each electron-hole pair that is generated in the depletion region will be acted upon by the electric field in the depletion region such that the electron will be propelled to the neutral n region and the hole will be propelled to the neutral p region. This results in an electric current through the junction equal to the flow of the charge of an electron for each electron-hole pair that is generated. Since each electron is carrying a charge of q coulombs, and the electron-hole pairs are generated throughout the depletion region of volume equal to $A\ x_d$ where A is the area of the junction and x_d is the width of the depletion region, the reverse current is given by

$$I_R = q\ A\ \left(\frac{D_p p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n}\right)\left(e^{\frac{-qV_R}{kT}} - 1\right) - \ qA\ x_d\ \frac{n_i}{2\tau_g} = I_s\left(e^{\frac{-qV_R}{kT}} - 1\right) - qA\ x_d\ \frac{n_i}{2\tau_g} \tag{3.102}$$

Since x_d, the depletion region width, is dependent on the applied reverse voltage, the term due to generation in the depletion region has a voltage dependence as shown in Figure (3.17).

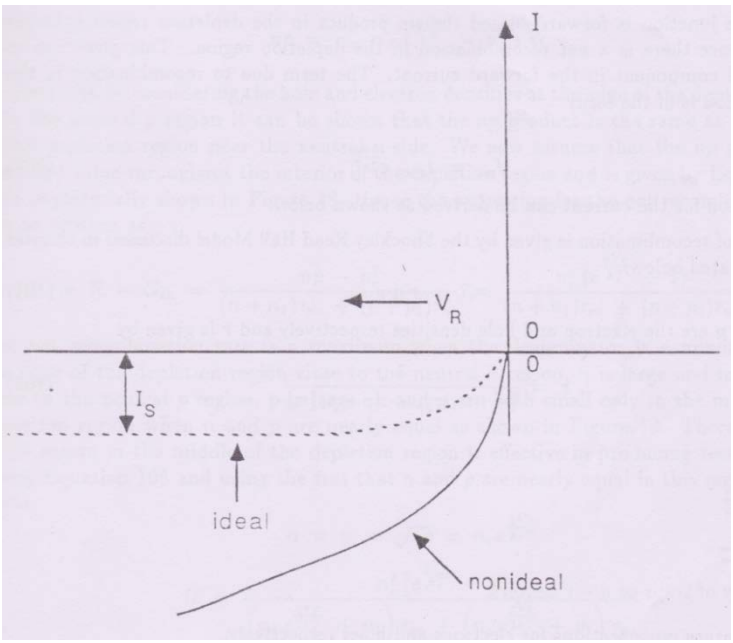


Figure (3.17): Non-ideal current characteristics under a reverse bias

When the junction is forward biased the $n\ p$ product in the depletion region is larger than n_i^2. Hence there is a net recombination in the depletion region. This gives rise to an additional component in the forward current. The term due to recombination in the depletion region is of the form

$$I_{rec} = I_{s\,rec}\, e^{\frac{qV_F}{2kT}} \tag{3.103}$$

This expression for the current can be derived as shown below:

The rate of recombination is given by the Schockley Read Hall Model discussed in chapter 2 and is repeated below:

$$R = r\ n\ p \tag{3.104}$$

Where n and p are the electron and hole densities respectively and r is given by

$$r = \frac{1}{(n+n_1)\tau_{p0}+(p+p_1)\tau_{n0}} \tag{3.105}$$

Where

$$\tau_{p0} = \frac{1}{N_t \sigma_p v_p}$$

$$\tau_{n0} = \frac{1}{N_t \sigma_n v_n}$$

N_t = density of traps or g-r centers.

σ_n, σ_p = capture cross-sections for electrons and holes respectively,

v_n, v_p =thermal; velocity of electrons and holes respectively and is equal to $\sqrt{\frac{3\,kT}{m^*}}$, and m^* is the effective mass of the carrier,

$$n_1 = N_C \exp\left(-\frac{E_C - E_t}{kT}\right)$$

$$= n_i exp\left(\frac{E_t - E_i}{kT}\right) \tag{3.106}$$

$$p_1 = N_V \exp\left(-\frac{E_t - E_V}{kT}\right)$$

$$= n_i exp\left(-\frac{E_t - E_i}{kT}\right) \tag{3.107}$$

where E_t = the trap energy level

The value of n at the edge of the depletion region near the neutral n-side is equal to the majority carrier density n_{n0} and p has a value equal to $p_{n0}e^{\frac{qV_F}{kT}}$. Hence the product of n and p at the edge of the depeletion region near the neutral n region is given by

$$n\,p = n_{n0}p_{n0}e^{\frac{qV_F}{kT}} = n_i^2 e^{\frac{qV_F}{kT}} \tag{3.108}$$

Similarly by considering the hole and electron densities at the edge of the depletion region near the neutral p region it can be shown that the $n\,p$ product is the same as at the edge of the depletion region near the neutral n-side. We now assume that the $n\,p$ product has the same value throughout the interior of the depletion region and is given by Equation (3.108). This is pictorially shown in Figure (3.18). Hence the expression for the net recombination rate can be written as

$$U = R - G_{th} = \frac{np - n_i^2}{(n+n_1)\tau_{p0} + (p+p_1)\tau_{n0}} = \frac{n_i^2\left(e^{\frac{qV_F}{kT}} - 1\right)}{(n+n_1)\tau_{p0} + (p+p_1)\tau_{n0}} \tag{3.109}$$

The net recombination rate is a maximum when the denominator is a minimum. Near the edge of the depletion region close to the neutral n region, n is large and in the region close to the neutral p region, p is larger. n and p are both small only in the middle of the depletion region when n and p are nearly equal as shown in Figure (3.18). Therefore only a small region in the middle of the depletion region is effective in producing recombination. Using Equation (3.108) and using the fact that n and p are nearly equal in this region, we can write

$$n = p = \sqrt{np} = n_i e^{\frac{qV_F}{2kT}} \tag{3.110}$$

$$U = \frac{n_i^2\left(e^{\frac{qV_F}{kT}}-1\right)}{\left(n_i e^{\frac{qV_F}{2kT}}+n_1\right)\tau_{p0}+\left(n_i e^{\frac{qV_F}{2kT}}+p_1\right)\tau_{n0}} \tag{3.111}$$

Substituting the expression for n_1 and p_1 in terms of $(E_t - E_i)$, we obtain

$$U = \frac{n_i^2\left(e^{\frac{qV_F}{kT}}-1\right)}{n_i e^{\frac{qV_F}{2kT}}(\tau_{p0}+\tau_{n0})+n_i e^{\frac{(E_t-E_i)}{kT}}\tau_{p0}+n_i e^{-\frac{(E_t-E_i)}{kT}}\tau_{n0}}$$

$$= \frac{n_i\left(e^{\frac{qV_F}{kT}}-1\right)}{e^{\frac{qV_F}{2kT}}(\tau_{p0}+\tau_{n0})+ \tau_{p0}e^{\frac{(E_t-E_i)}{kT}}+ \tau_{n0}e^{-\frac{(E_t-E_i)}{kT}}} \tag{3.112}$$

When $|E_t - E_i| > kT$, the denominator is large and hence U is nearly zero. Only when the trap is in the middle of the gap i.e., $E_t \approx E_i$, U becomes not negligible.

$$U \approx \frac{n_i\left(e^{\frac{qV_F}{kT}}-1\right)}{(\tau_{p0}+\tau_{n0})\left(e^{\frac{qV_F}{2kT}}+1\right)} \tag{3.113}$$

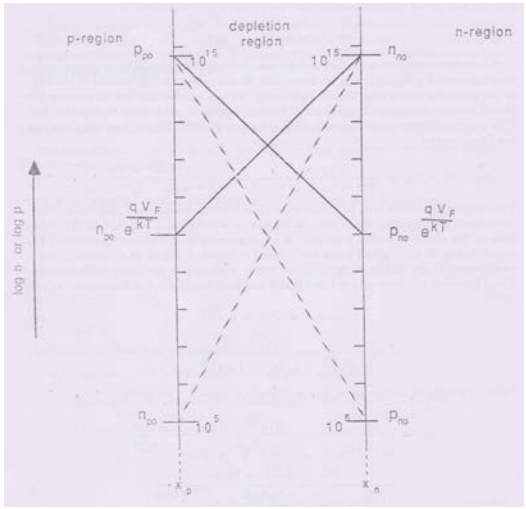


Figure (3.18). Electron and hole density variation in the depletion region of a forward biased p-n junction. The solid lines represent the carrier density variation under forward bias. The dotted lines represent the carrier density variation in the depletion region of a p-n junction under thermal equilibrium. In thermal equilibrium the np product is equal to n_i^2. Under forward bias the np product is assumed to be equal to $n_i^2 e^{\frac{qV_F}{kT}}$ throughout the depletion region since the np product has this value at the boundary of the depletion region on either side.

The term 1 in both the numerator and the denominator is negligible in comparison with the exponential terms and can therefore be omitted.

$$U \approx \frac{n_i e^{\frac{qV_F}{kT}}}{e^{\frac{qV_F}{2kT}}(\tau_{p0}+\tau_{n0})} = \frac{n_i e^{\frac{qV_F}{2kT}}}{(\tau_{p0}+\tau_{n0})} \tag{3.114}$$

When an electron and a hole recombine, an electron flows from the neutral n region into the depletion region and similarly a hole flows into the depletion region from the neutral p region. This constitutes a current flow through the junction in the same direction as the ideal forward current that we considered in the previous section. Hence the additional component of current due to recombination in the depletion region is given by

$$I_{rec} = A\, x_{\text{deff}}\, q\, U = \frac{A\, x_{\text{deff}} q\, n_i e^{\frac{qV_F}{2kT}}}{(\tau_{p0}+\tau_{n0})} = I_{S\,rec}\; e^{\frac{qV_F}{2kT}} \tag{3.115}$$

where x_{deff} is the small region in the middle of the depletion region where the electron and hole densities are both minimum and equal.

The total forward current including recombination in the depletion region is given by

$$I_F = I = qA\left[\frac{D_p p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n}\right] e^{\frac{qV_a}{kT}} + I_{S\,rec}\; e^{\frac{qV_F}{2kT}} = I_{S\,ideal}\; e^{\frac{qV_F}{kT}} + I_{S\,rec}\; e^{\frac{qV_F}{2kT}} \tag{3.116}$$

If one were to calculate $I_{S\,ideal}$ and $I_{S\,rec}$ for modern devices as illustrated in the worked out example below, the former will be much smaller than the latter. Hence, when the forward bias is not that large, the forward current is dominated by the recombination current and depends on the voltage as $e^{\frac{qV_F}{2kT}}$. At much larger forward bias voltages, the forward current is due mainly to the ideal junction current and depends on the bias voltage as $e^{\frac{qV_F}{kT}}$. At intermediate voltages, the forward current is due to both the mechanisms and the bias voltage dependence is approximated by $e^{\frac{qV_F}{\eta kT}}$. The parameter η is called the ideality factor and is usually used as a parameter to define how good a junction is.

Example

Consider a p-n junction with the following characteristics where N_A and N_D are the net impurity concentration on the n and p sides respectively. A is the area of the junction.

$A = 10^{-4} cm^2$, $N_D = 10^{16} cm^{-3}$, $N_A = 10^{15} cm^{-3}$, $D_n = 30\ cm^2 s^{-1}$, $D_p = 4\ cm^2 s^{-1}$, $\tau_n = 10^{-4} s$, $\tau_p = 10^{-5} s$

The ideal current prefactor $I_{S\,ideal}$ and the recombination current prefactor $I_{S\,rec}$ can now be calculated using the values for the universal constants q and k and assuming room temperature i.e., $T = 300\ K$. Taking the intrinsic carrier density n_i as equal to $10^{10}\ cm^{-3}$, the thermal equilibrium minority densities can be calculated as

$$p_{n0} = \frac{n_i^2}{N_D} = \frac{10^{20}}{10^{16}} = 10^4\ cm^{-3} \tag{3.117}$$

and

$$n_{p0} = \frac{n_i^2}{N_A} = \frac{10^{20}}{10^{15}} = 10^5\ cm^{-3} \tag{3.118}$$

The minority carrier diffusion lengths are calculated as

$$L_p = \sqrt{\tau_p D_p} = \sqrt{10^{-5} x 4}\ = 6.32\ x\ 10^{-3}\ \ \text{cm} \tag{3.119}$$

$$L_n = \sqrt{\tau_n D_n} = \sqrt{10^{-4} x 30}\ = 5.48\ \text{x}10^{-2}\ \text{cm} \tag{3.120}$$

Substituting these values in the equation for $I_{S\ ideal}$ we get

$$I_{S\ ideal}\ = 1.6x10^{-19} \times\ 10^{-4} \times \left(\frac{4\times10^4}{6.32 \times 10^{-3}} + \frac{30\times 10^5}{5.48 \times 10^{-2}}\right)$$

$$=\ 1.6x10^{-23} \times\ (5.7 \times 10^6 + 5.47 \times 10^7)$$

$$=\ 9.66\ \times 10^{-16}\ \text{Amperes}$$

For calculating $I_{S\ rec}$ we use the following typical parameters. Let the density of the traps, N_t be $10^{15}\ cm^{-3}$ and $\sigma_n = \sigma_n = 10^{-16} cm$. Assume $v_n = v_p = v_{th}$ where

$$v_{th} = \sqrt{\frac{3kT}{m}} = \sqrt{\frac{3\times1.38\times10^{-23}\times300}{9.11x10^{-31}}}\ = 1{,}17\ \times\ 10^5 \tag{3.121}$$

And we have assumed the effective mass of the carriers to be the same as for electrons in vacuum. Since we will be using mostly cm as the unit of length we take v_{th} as $1.17x10^7\ cm\ s^{-1}$. Let us assume tht the width of the region wwhere the recombination is effective, x_{diff}, is 3 x10^{-5} cm. τ_{n0} is equal to τ_{p0} due to our assumption of equality of σ_n and σ_p as well as v_n and v_p. Therefore

$$\tau_{n0} + \tau_{p0} = \frac{1.6x10^{-19}x\ 10^{-4}x\ 3\ x\ 10^{-5}x\ 10^{10}}{1{,}71\ x\ 10^{-6}}$$

$$= 2.81\ x\ 10^{-12}\ \textit{Amperes}$$

Thus we see that $I_{S\ rec}$ is larger than $I_{S\ ideal}$ and therefore dominates at low bias voltages. At higher bias voltages, the ideal current dominates because of the absence of the factor 2 in the denominator of the exponent of the voltage dependent term.

In the discussion so far we considered recombination generation in the bulk of the semiconductor. Normally the depletion region intersects a surface and the g-r centers (called surface-states or interface states) on the surface of the depletion region can also provide a recombination

current under forward bias and a generation current under reverse bias in addition to the previously discussed components of current. The net recombination rate per unit area of the surface, U_s is given by

$$U_s = \frac{np - n_i^2}{\frac{n+n_1}{N_{st}\sigma_p v_p} + \frac{p+p_1}{N_{st}\sigma_n v_n}} \tag{3.122}$$

Where N_{st} is the number of g-r centers per unit area of the surface (surface density), and all other parameters have the same meaning as before for the bulk recombination. Under forward bias the $n\ p$ product in the surface region is much larger than n_i^2 and hence a surface recombination current flows in addition to the bulk recombination current. The surface recombination current also has the same voltage dependence as the bulk recombination current i.e., the current is proportional to $e^{\frac{qV_F}{2kT}}$.

When the junction is under reverse bias, the $n\ p$ product is negligible in comparison with n_i^2 and hence the net generation rate under the assumption that only the surface states whose energy is in the middle of the gap contribute to the generation, is given by

$$U = \frac{n_i^2}{\frac{n_1}{N_{st}\sigma_p v_p} + \frac{p_1}{N_{st}\sigma_n v_n}} = -\frac{n_i}{\frac{1}{N_{st}\sigma_p v_p} + \frac{1}{N_{st}\sigma_n v_n}} = -\frac{n_i s}{2} \tag{3.123}$$

where s equal to

$$= \frac{2}{\frac{1}{N_{st}\sigma_p v_p} + \frac{1}{N_{st}\sigma_n v_n}}$$

is called the ***surface recombination velocity***. Recall that we defined a minority carrier lifetime in the bulk to describe the generation rate. We now define a surface recombination velocity for the surface to describe the generation at the surface. The electron-hole pairs created at the surface depletion region will give rise to another component of reverse current which we will call the surface generation current. This component of surface generation current will add to the bulk generation current under reverse bias.

A typical current voltage plot is given in Figure (3.19) in which the current is plotted on a logarithmic scale while the voltage is plotted on a linear scale. The slope at small values of forward bias voltage is half that at larger values of bias voltage. The recombination component $I_{S\,rec}$ is larger than the ideal $I_{S\,ideal}$ and hence at small voltages the recombination current dominates and the slope is $\frac{q}{2kT}$. At higher voltage the ideal component dominates and the slope is $\frac{q}{kT}$. It is customary to express the current as

$$I_F = I_S\, e^{\frac{qV_F}{\eta kT}} \tag{124}$$

Where η is called the ideality factor. The more ideal a diode is, closer is η to unity.

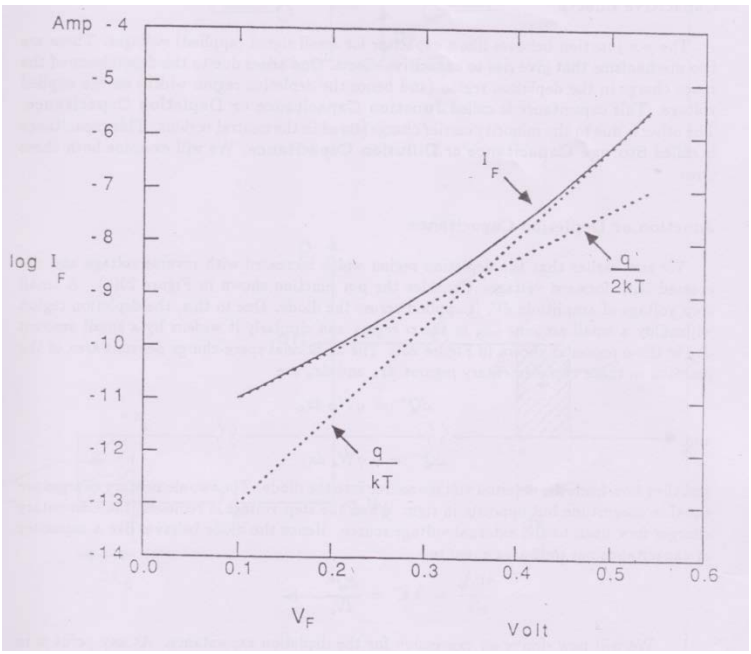


Figure (3.19): Plot of the forward current in a typical p-n junction as a function of the bias voltage. The current is on a logarithmic scale. Notice how the slope increases by a factor 2 between the lower and higher voltages.

Capacitive Effects

The p-n junction behaves like a capacitor for small signal (applied) voltages. There are two mechanisms that give rise to capacitive effects. One arises due to the dependence of the space charge in the depletion region (and hence the depletion region width) on the applied voltage. The capacitance is called **Junction Capacitance** or **Depletion Capacitance**. The other is due to the minority carrier charge stored in the neutral regions. This capacitance is called **Storage Capacitance** or **Diffusion Capacitance**. We will examine both these now:

$$dQ^+ = qN_D dx_n$$

and

$$dQ^- = -qN_A dx_p$$

and they flow from the external voltage source into the diode. The two elementary charges are equal in magnitude but opposite in sign. When the step voltage is removed, the elementary charges flow back to the external voltage source. Hence the diode behaves like a capacitor of capacitance per unit area equal to

$$C = \frac{dQ^+}{dV_a}$$

We will now derive an expression for the depletion capacitance. At any point x in the depletion region, the increase in the electric field, d$\mathcal{E}$ due to the small step voltage, is obtained from Gauss theorem as

$$\mathrm{d}\mathcal{E} = \frac{dQ^+}{\epsilon_s}$$

It is independent of x. The incremental change in the applied voltage across the diode dV_a can be obtained from the line integral of $\mathrm{d}\mathcal{E}$ across the depletion region. Since $\mathrm{d}\mathcal{E}$ is constant with x, the line integral is

$$dV_a = x_\mathrm{d} \times \mathrm{d}\mathcal{E}$$

Substituting the expression for $\mathrm{d}\mathcal{E}$ in terms of dQ^+ we get

$$dV_a = x_\mathrm{d} \times \frac{dQ^+}{\epsilon_s}$$

Dividing both sides of the equation by dQ^+, we obtain for the depletion capacitance per unit area of the junction,

$$C_d = \frac{dQ^+}{dV_a} = \frac{\epsilon_s}{x_\mathrm{d}} \tag{3.125}$$

We find that this is exactly the same expression that we would have had if we have a parallel plate capacitor with the semiconductor as the dielectric between the parallel plates of unit area and a spacing of W between the plates. Thus we see that the junction behaves like a parallel plate capacitor. What we calculated was the expression for the capacitance for unit area of the junction. The capacitance of a junction of area A is

$$C_J = A\, C_d = A\, \frac{\epsilon_s}{x_\mathrm{d}} \tag{3.126}$$

What we derived for the junction capacitance is valid for a junction even if it did not have uniform impurity concentrations i.e., if the diode was not an abrupt junction. We can arrive at the expression for the capacitance of an abrupt junction by an alternate method also. The total positive charge in the depletion region, Q^+ per unit area of the junction is given by

$$Q^+ = q\, N_D x_d = q\, \frac{N_D N_A}{N_D + N_A} x_\mathrm{d}$$

Where we have used the relation

$$N_D x_\mathrm{n} = N_A x_\mathrm{p}$$

Differentiating this expression for Q^+ with respect to V_a we can get

$$C_d = \frac{\epsilon_s}{x_\mathrm{d}}$$

which is the same as what we obtained earlier. It is left as an exercise for the student to show this.

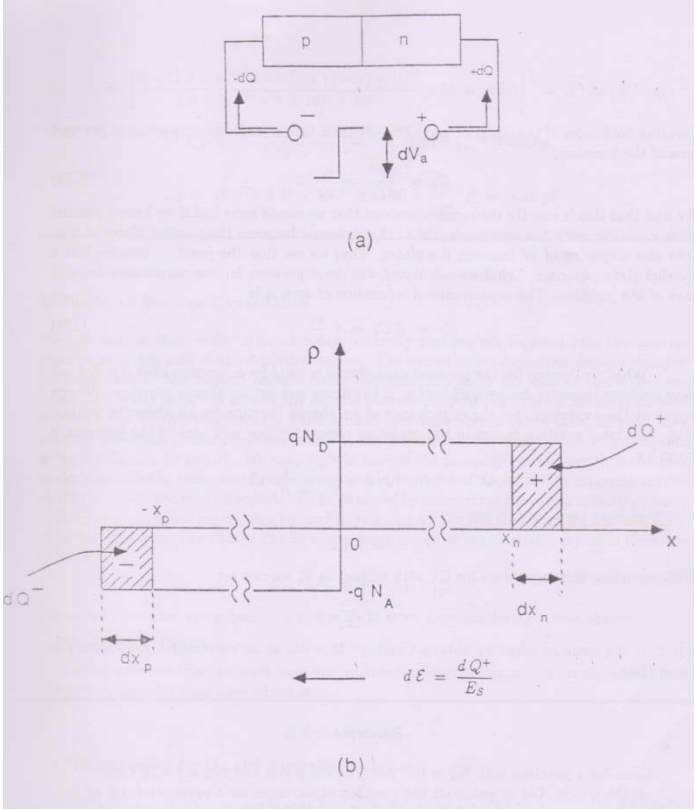


Figure (3.20): a) Flow of elementary charges dQ^+ and dQ^- into the junction diode due to the application of a step voltage: b) The widening of the space charge region due to the step voltage: dx_{n} and dx_{p} are the elementary increase in the space charge on the n and p sides respectively.

Example

Consider a junction with $N_A = 10^{15}cm^{-3}$ on the p side and $N_D = 3x10^{15}cm^{-3}$ on the n side. Let us calculate the junction capacitance for a reverse voltage of $5V$. Assume the area of the junction to be $10^{-3}cm^2$. We saw that the expression for x_d in terms of $V_R$is given as

$$x_d = \left[\frac{2\epsilon_s(N_A+N_D)}{qN_DN_A}(V_{bi}+V_R)\right]^{\frac{1}{2}}$$

$$V_{bi} = 0.0259 \times \ln\left(\frac{10^{15}\times 3\times 10^{15}}{10^{20}}\right) = 0.625\ V$$

Substituting the values for V_{bi} and other parameters, we get

$$x_d = \left[\frac{211.9\times 8.84\times 10^{-14}(3+1)\times 10^{15}}{1.6x10^{-19}\times 3\times 10^{15}\times 10^{15}}(5+0.625)\right]^{\frac{1}{2}} = 3.13 \text{ x } 10^{-4} \text{ cm}$$

$$C_d = \frac{11.9\times8.84\times10^{-14}}{3.13 \text{ x } 10^{-4}} = 3.36 \text{ x } 10^{-9} \quad \text{F } cm^{-2}$$

and

$$C_J = 10^{-3} \times 3.36 \times 10^{-9} = 3.36 \times 10^{-12} F = 3.36\, pF$$

Diffusion or Storage Capacitance

We saw earlier that under a forward bias minority carriers are injected into the neutral region on either side of the depletion region. The excess minority carrier density decayed with distance in the neutral regions either exponentially as in the case of the wide base diode or linearly as in the case of the narrow base diode. Although the excess carriers were diffusing and therefore constantly moving until they recombined (wide base) or until they reached the ohmic contact (narrow base), at any given time they had a density distribution given by Figure (3.10) and (3.14). We can think as though the minority carriers equal to the area under the curve in these two figures are stored in the neutral regions. The minority carrier charge stored in the neural regions can be obtained by integrating the excess minority carrier density in the neutral region and by multiplying this by the electron charge q and the area, A, of the junction. The charge due to excess holes stored in the neutral n region is therefore given by

$$Q_{ps} = q\, A \int_0^{W_n} \Delta p(x')\, dx'$$

where we have used the subscript s to denote that we are considering stored charge.

Wide Base Diode

Let us consider the wide base diode. Substituting the expression for the excess hole density in the wide base case which is

$$\Delta p(x') = \Delta p(o) e^{\frac{-x'}{L_p}}$$

into the expression for Q_{ps} and integrating we get,

$$Q_{ps} = q\, A\, \Delta p(o) \int_0^{W_n} e^{\frac{-x'}{L_p}}\, dx' = Aq\Delta p(o) L_p \qquad (3.127)$$

We can relate this to the current through the diode. The current due to hole injection is

$$I_p = -q\, A\, D_p \left.\frac{d\Delta p(x')}{dx'}\right|_{x'=0} = \frac{q\, A\, D_p \Delta p(o)}{L_p}$$

Dividing the expression for Q_{ps}, by I_p and rearranging, we get

$$Q_{ps} = I_p \frac{{L_p}^2}{D_p} = I_p \, \tau_p \tag{3.128}$$

Thus we see that the stored charge in the neutral region is the injected minority carrier current times the life time. The stored charge is dependent on the applied forward bias, V_F, since

$$\Delta p(o) = p_{n0}\left(e^{\frac{qV_F}{kT}} - 1\right)$$

Let us assume that a step voltage dV_F is applied in series with V_F as shown in Figure (3.21 A). Due to this step voltage the stored charge will increase by an amount dQ_{ps} as shown in Figure (3.21 B). This extra charge dQ_{ps}, flowed into the neutral n region from the external source dV_a through the neutral p region. In order to maintain charge neutrality the electron density in the neutral n region also has to increase and the increase in the negative charge dQ_{ns} will be exactly equal and oppoissite to $dQ_{p,}$ and has to flow from the external source through the neutral n region as shown in Figure (3.21 C). Therefore the diode appears to the external source as a capacitor of capacitance equal to

$$C_{\text{diff}\,p} = \frac{dQ_{ps}}{dV_F} \tag{3.129}$$

Hence we can obtain the diffusion capacitance as

$$C_{\text{diff}\,p} = \frac{d(I_p \, \tau_p)}{dV_F} = \tau_p \frac{dI_p}{dV_F} \tag{3.130}$$

But

$$I_p = I_{sp}\left(e^{\frac{qV_F}{kT}} - 1\right)$$

which becomes equal to

$$I_p \approx I_{sp} e^{\frac{qV_F}{kT}}$$

Under the approximation $dV_F \gg kT$. Differentiating this with respect to V_F we get

$$\frac{dI_p}{dV_a} = \frac{dI_p}{dV_F} = \frac{qI_p}{kT} \tag{3.131}$$

Therefore the diffusion capacitance becomes equal to

$$C_{\text{diff}\,p} = \tau_p \frac{qI_p}{kT} \tag{3.132}$$

The DC current through the diode also increases by a small increment dI_p due to the small step voltage and therefore the diode appears to the step voltage as a resistor of resistance equal to

$$r_{fp} = \frac{dV_a}{dI_p}$$

or a conductor of conductance equal to

$$g_{fp} = \frac{dI_p}{dV_a} \tag{3.133}$$

Using Equation (3.132), we can write the diffusion capacitance as

$$C_{\text{diff}\,p} = \tau_p g_{fp} \tag{3.134}$$

So far we considered only the small signal circuit elements due to the hole injection in the neutral n region. We have similar diffusion capacitance due to electron injection in the neutral p region so that we can write for the total diffusion capacitance as

$$C_{\text{diff}\,p} = \tau_p \frac{qI_p}{kT} + \tau_n \frac{qI_n}{kT} = \tau_p g_{fp} + \tau_n g_{fn} \tag{3.135}$$

where I_n is the component of forward bias current due to electron injection in the neutral p region and g_{fp} and g_{fn} are the forward conductance due to hole and electron injection respectively. Similarly the forward conductance will be equal to

$$g_f = \frac{q}{kT}\left(I_p + I_n\right) = \frac{qI_F}{kT} = g_{fp} + g_{fn}$$

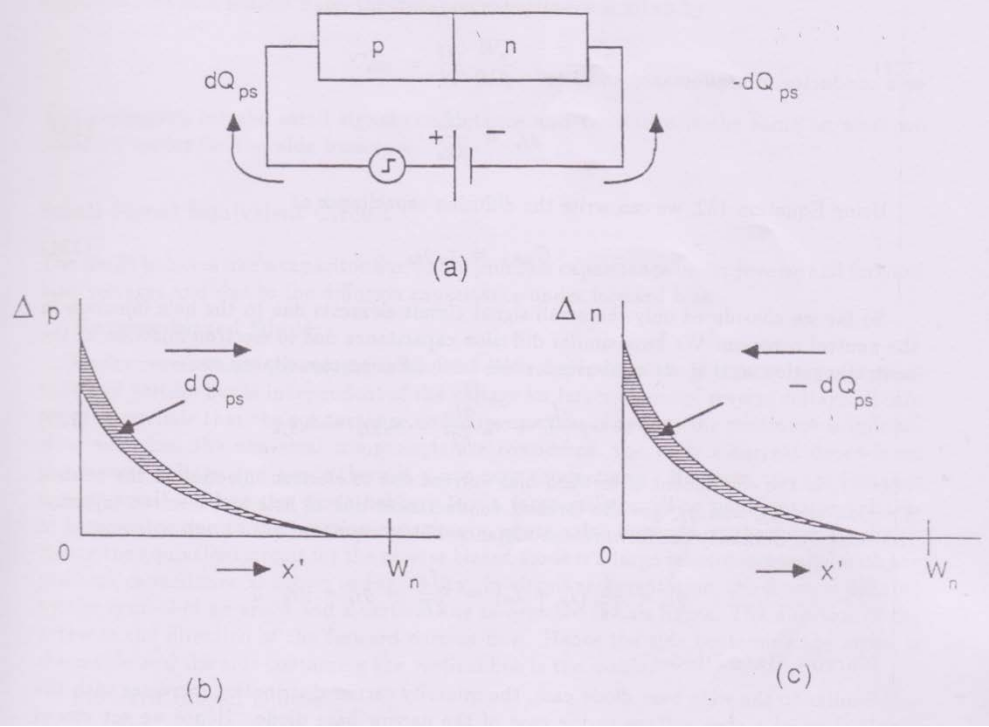


Figure (3.21): The effect of the application of a step voltage on a forward biased junction diode a) The step voltage, dV_a is applied in series with the forward bias. B) The increase in the injected minority carrier distribution in the neutral n-region is shown by the shaded area. C) The increase in the majority carrier distribution in the neutral n-region to maintain the charge neutrality is shown by the shaded area.

Narrow Base Diode

Similar to the wide base diode case, the minority carrier distribution increases with the application of a step voltage in the case of the narrow base diode. Hence we get stored charge effects and the diode gives rise to diffusion capacitance. It is left as an exercise for the student to show that the stored charge due to the injection of minority carriers in the neutral n region is given by

$$Q_{ps} = \frac{\mathrm{q\ A}\ \Delta p(o) W_n}{2}$$

Similarly the diffusion capacitance can be shown to be

$$C_{\mathrm{diff}\,p} = \frac{{W_n}^2}{2D_p}\frac{qI_p}{kT}$$

In the next chapter, it will be shown that $\frac{{W_n}^2}{2D_p}$ is the transit time for the holes in the neutral n-region i.e., the time taken for the hole to traverse the n-region. A diffusion capacitance arises also due to the injection of electrons in the neutral p-region. The total diffusion capacitance is the sum of both the diffusion capacitances and is given by

$$C_{\mathrm{diff}} = \frac{qI_p}{kT}\frac{{W_n}^2}{2D_p} + \frac{qI_n}{kT}\frac{{W_p}^2}{2D_n} \tag{3.136}$$

If one side, say the p-side, is approximated as a wide base and the other side (n-side) is approximated as a narrow base, the diffusion capacitance is given by

$$C_{\mathrm{diff}} = \frac{qI_p}{kT}\frac{{W_n}^2}{2D_p} + \frac{\tau_n qI_n}{kT} \tag{3.137}$$

The expression for the small signal conductance and resistance is the same as what we obtained earlier for the wide base case.

Small Signal Equivalent Circuit

The diode behaves like a capacitor due to the junction capacitance under reverse and forward bias voltages and due to the diffusion capacitance under forward bias.

Reverse-biased Diode

Under reverse bias, the current in an ideal diode depends on the reverse voltage for small values of voltage but is independent of the voltage for larger values of reverse voltage. Hence we can conclude that the conductance under reverse bias is zero or the resistance is infinite. However when the non-ideal components are considered, the reverse current depends on the reverse voltage. Hence there is a non-zero conductance. However, the conductance is very small. Hence the diode behaves like a large resistor. The junction also behaves as a capacitor due to the junction capacitance whose value depends

on the reverse voltage. Hence the equivalent circuit for the reverse biased diode is a large resistor in parallel with the junction capacitance as shown in Figure (3.22 A). In circuit representation, the diode is denoted by the symbol of an arrow and a vertical line as depicted in this figure. The direction of the arrow is the direction of the forward current flow. Hence the side containing the arrow is the p-side and the side containing the vertical line is the n-side.

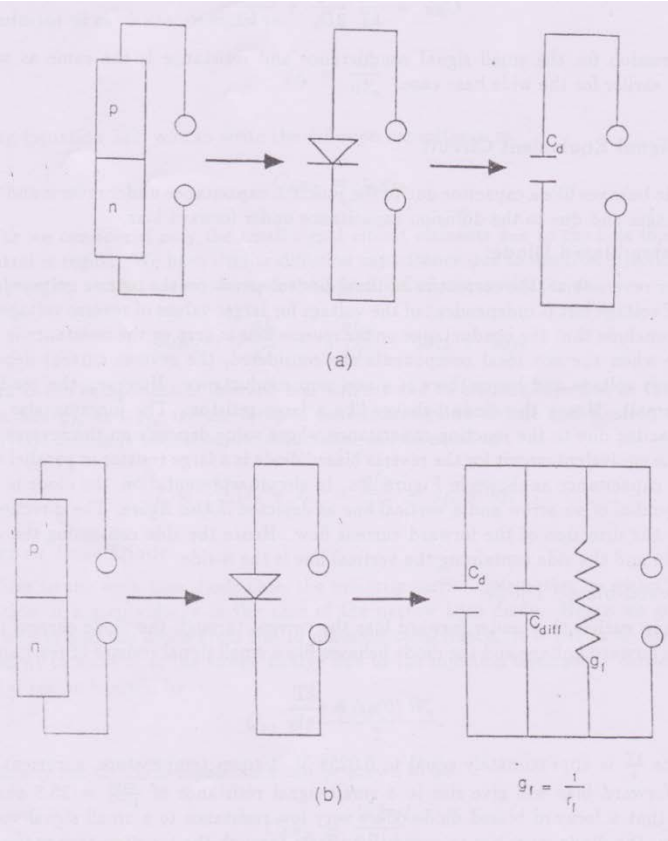


Figure (3.22). Equivalent circuit of a diode under A) reverse bias, and B) forward bias

Forward-biased Diode

We saw earlier that under forward bias the current through the diode current increases with the forward voltage and diode behaves like a small signal resistor of resistance equal to

$$r_f = \frac{kT}{qI_F}$$

Since $\frac{kT}{q}$ is approximately equal to 0.0258 V at room temperature, a current of *1 mA* under forward bias will give rise to a small resistance of $\frac{0.0258}{10^{-3}} = 25.8\ ohm$. Thus we see that a forward biased diode offers very low resistance to a small signal voltage. In addition, the diode gives rise to capacitive effects through the junction capacitance and the diffusion capacitance. Therefore the response of the diode to a small signal voltage can be represented by the equivalent circuit shown in Figure (3.22 B). The

capacitors and the resistor are all in parallel because the incremental charge or the incremental current all flow independent of each other and only depend on the small signal voltage. The resistance has a small value.

Junction Breakdown

The reverse current in an ideal diode remains constant with increasing reverse voltage until at a certain reverse voltage called the ***breakdown voltage*** the current suddenly increases to an extremely large value as shown in Figure (3.23). If current limiting resistors were not included in the circuit the current will become so large as to destroy the junction. Under this condition, the junction is said to have suffered a breakdown and the reverse voltage at which the breakdown occurs is called the junction breakdown voltage. There are three mechanisms that cause the junction to break down and they are: a) Thermal mechanism, b) Avalanche breakdown and c) Tunneling breakdown.

Although the reverse current, which is called the diode leakage current, is extremely small at room temperature, it is very sensitive to temperature and increases very rapidly with temperature. The power dissipated in the junction, is the product of the leakage current and the reverse voltage. At large enough reverse voltage, the power dissipation can become excessive and hence can cause a heating of the junction. This increases the junction temperature and hence the leakage current increases further and this in turn increases the junction power dissipation. This cycle of increased power dissipation and increase in temperature keeps on recurring until the current becomes very large and the diode is destroyed. This is called the *thermal breakdown*. The thermal breakdown used to be a big problem in the days when germanium instead of silicon was used as the semiconducting material for fabricating junction devices. However with modern silicon devices, in which adequate heat sinks are used as part of the device design, thermal breakdown is not usually a problem.

In modern devices the breakdown due to the other two mechanisms limits the maximum reverse voltage that can be applied across the junction. The avalanche mechanism can be explained by reference to Figure (3.24). In this figure, the band diagram of a junction across which a larger reverse voltage has been applied is shown. The depletion region is wide and also has a large electric field. An electron entering the depletion region is accelerated by the large electric field, and within a short distance its kinetic energy increases to a sufficiently large value such that it is able to create an electron-hole pair by impact ionization. The electron loses most of its kinetic energy in the impact ionization process. It starts to accelerate again and within a short distance again causes the generation of another electron-hole pair. Thus by the time the electron reaches the other end of the depletion region, it has generated many more electrons. This process is called *avalanche multiplication*. The ratio of the number of electrons coming out of the depletion region to the electrons entering the depletion region is called the *multiplication factor*. The holes can also cause avalanche multiplication. At sufficiently large reverse voltage the multiplication factor becomes infinite and the current increases infinitely to cause a junction breakdown.

The *tunneling mechanism* can be explained with reference to Figure (3.25 A). In this figure a junction in which the impurity concentrations on the two sides are very high is shown. The depletion region width is very narrow because of this. In order to go to the conduction band the electrons in the valence band need not thermally get excited but can tunnel through a triangle barrier to the conduction band as shown in Figure (3.25 B).

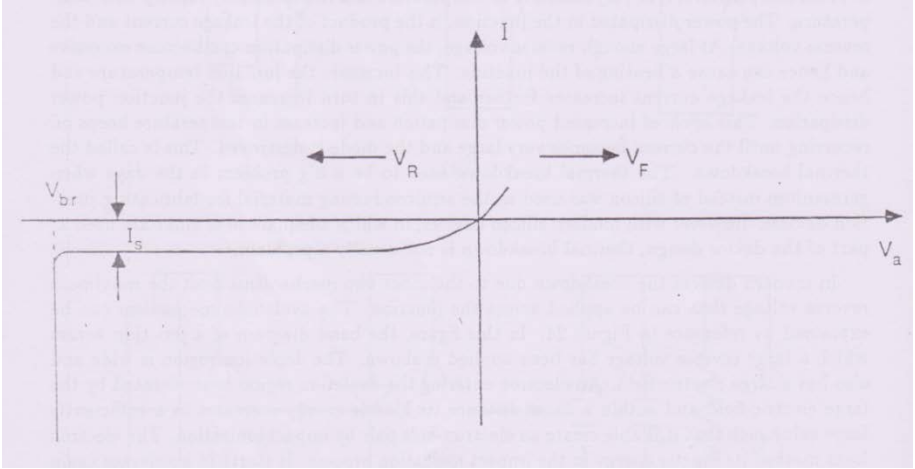


Figure (3.23): Breakdown in a junction diode. The voltage V_{br} at which the current rapidly increases is defined as the breakdown voltage

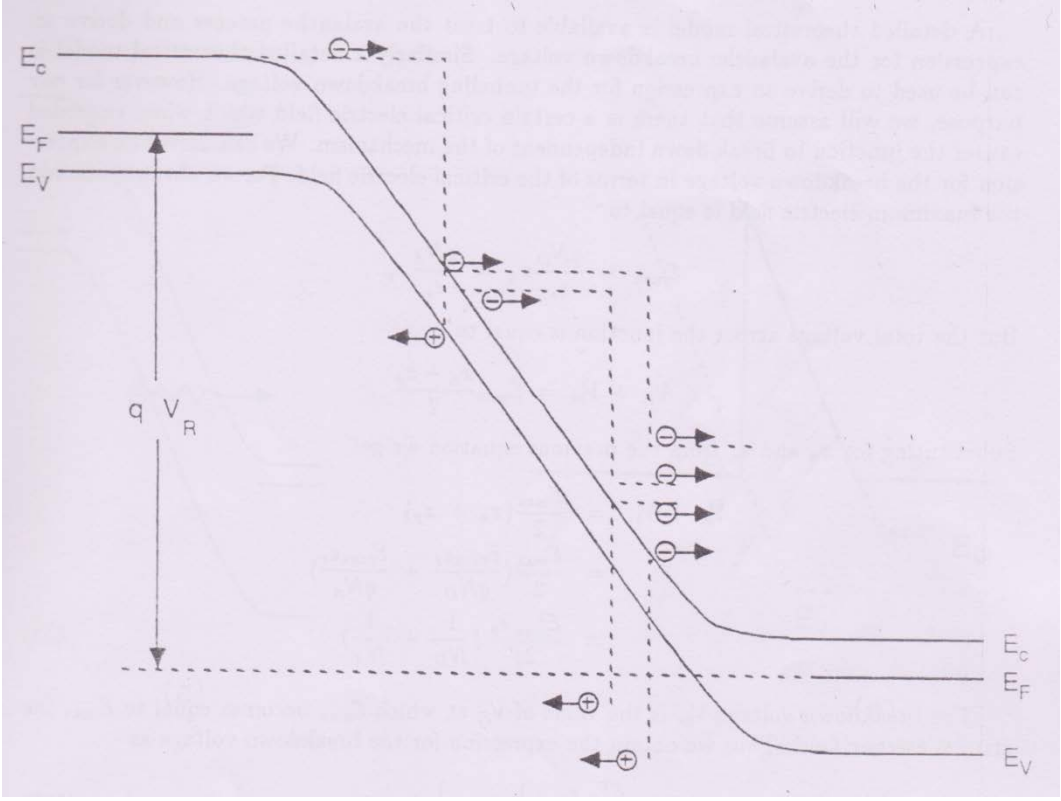


Figure (3.24): The avalanche breakdown mechanism

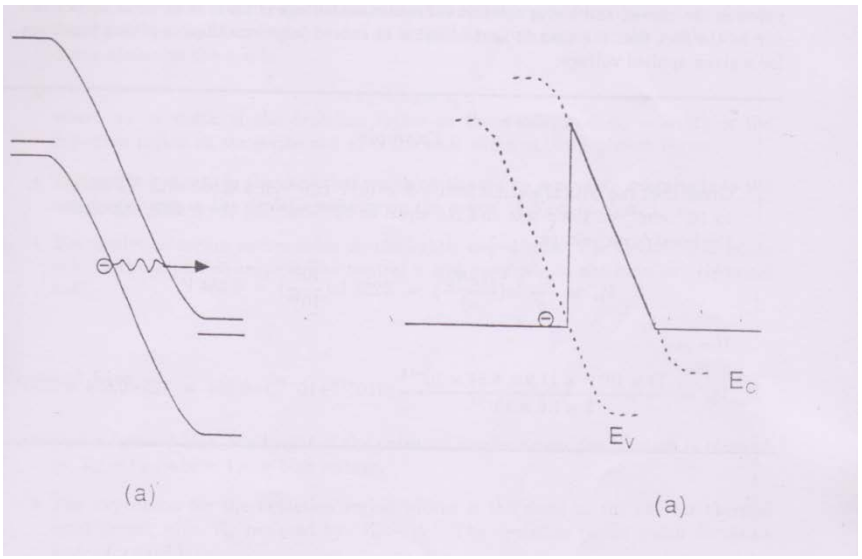


Figure (3.25): The tunneling breakdown mechanism: A) the band diagram under tunneling conditions and B) the triangular barrier through which the electron tunnels through

A detailed theoretical model is available to treat the avalanche process and derive an expression for the avalanche breakdown voltage. Similarly a detailed theoretical model is can be used to derive an expression for the tunneling breakdown voltage. However for our purpose we will assume that there is a certain critical electric field which when exceeded causes the junction to breakdown independent of the mechanism. We can derive an expression for the breakdown voltage in terms of the critical electric field. For an abrupt junction the maximum electric field is equal to

$$\mathcal{E}_{max} = \frac{qN_D}{\epsilon_s} x_n = \frac{qN_A}{\epsilon_s} x_p$$

But the total voltage across the junction is equal to

$$V_R + V_{bi} = \mathcal{E}_{max} \frac{x_n + x_p}{2}$$

Substituting for x_n and x_p from the previous equation we get

$$\begin{aligned} V_R + V_{bi} &= \mathcal{E}_{max} \frac{x_n + x_p}{2} \\ &= \frac{\mathcal{E}_{max}}{2} \left(\frac{\mathcal{E}_{max} \epsilon_s}{qN_D} + \frac{\mathcal{E}_{max} \epsilon_s}{qN_A} \right) \\ &= \frac{{\mathcal{E}_{max}}^2 \epsilon_s}{2q} \left(\frac{1}{N_D} + \frac{1}{N_A} \right) \end{aligned} \quad (3.138)$$

The breakdown voltage V_{br} is the value of V_R at which $\mathcal{E}_{max}$ becomes equal to $\mathcal{E}_{crit}$, the critical electric field. Thus we obtain the expression for the breakdown voltage as

$$V_{br} = \frac{{\mathcal{E}_{crit}}^2 \in_S}{2q}\left(\frac{1}{N_D} + \frac{1}{N_A}\right) - V_{bi} \tag{3.139}$$

There are two points that must be mentioned with respect to the junction breakdown. One is that the avalanche mechanism is the dominant mechanism in diodes in which the breakdown voltage is larger than $\frac{6E_g}{q}$ and the tunneling mechanism is the dominant one in diodes in which the breakdown voltage is less than $\frac{4E_g}{q}$. For silicon diodes avalanche process is dominant in diodes in which the breakdown voltage is above 6.6 V and tunneling is the dominate process for diodes with a breakdown voltage less than 4.4 V. Secondly the avalanche breakdown process has a positive temperature coefficient i.e., the breakdown voltage increases with temperature whereas the tunneling breakdown process has a negative temperature coefficient with the breakdown voltage decreasing with increasing temperature. Silicon diodes which have a breakdown voltage between 4.4 and 6.6 volts have both mechanisms present and hence have very little temperature dependence of the breakdown voltage. Usually diodes which are in the breakdown mode are used a voltage reference at a voltage of V_{br} and these are called Zener diodes.

The expression we derived for the breakdown voltage is for planar diodes i.e., ones in which the junction is planar. However in practical diodes the sides and corners of the junction are curved and hence the actual breakdown voltage is lower than what we derived due to the fact that the electric field is higher in curved junctions than in planar junctions for a given applied voltage.

Example

Given that the critical electric field is $3x10^5 V cm^{-1}$ for a diode with N_A equal to $10^{15} cm^{-3}$ on the p side and N_D equal to $10^{16} cm^{-3}$ on the n side, calculate the breakdown voltage.

$$V_{bi} = \frac{kT}{q} ln\left(\frac{N_A N_D}{n_i^2}\right) = 0.0258 \ln\left(\frac{10^{31}}{10^{20}}\right) = 0.654\ V$$

Hence

$$V_{bi} = \frac{(3x10^5)^2 \times 11.9 \times 8.84x10^{-14}}{2 \times 1.6 \times 10^{-19}} \times\ (10^{-15} + 10^{-16}) - 0.654$$
$$= 325 - 0.654 \approx 324\ V$$

Summary

- In thermal equilibrium, a potential barrier exists across the p-n junction such that the diffusion current is balanced by a drift current in the opposite direction. The height of the potential barrier is denoted by V_{bi}.
- A depletion region is formed on either side of the metallurgical junction, and the charge density in the depletion region is due to ionized acceptor atoms on the p-side, and to donor atoms on the n-side.

$$x_d = x_n + x_p$$

  Where x_n = width of the depletion region on the n-side, and x_p = width of the depletion region on the p-side and x_d is the total width of the depletion region.
- The positive charge in the depletion region on the n-side is equal in magnitude to the negative charge in the depletion charge on the n-side ($N_A x_p = N_D x_n$)
- The depletion region occurs more on the lightly doped side. The electric field exists only in the depletion region. The neutral n and p regions do not have any electrical field.

Forward Bias

- Under forward bias, the height of the potential barrier across the junction is reduced to $V_{bi} - V_{F,}$ where $V_F =$ bias voltage.
- The expression for the depletion region widths is the same as the one for thermal equilibrium, with V_{bi} replaced by $V_{bi} - V_F$. The depletion region width decreases under forward bias.
- Under forward bias, excess minority carriers are injected into the two neutral regions. The value of the excess minority density at the boundary of the deletion region and the neutral region is

$$\Delta p(o) = p_{no}(e^{\frac{qV_F}{kT}} - 1) \quad \ldots\ldots\ldots\ldots\ldots n\text{-side}$$

$$\Delta n(o) = n_{po}(e^{\frac{qV_F}{kT}} - 1) \quad \ldots\ldots\ldots\ldots\ldots P \text{ side}$$

- A net current flows across the junction due to the diffusion of the injected excess minority carrier's density.
- Under ideal conditions.

$$I_F = I_S(e^{\frac{qV_F}{kT}} - 1),$$

  Where

$$I_S = qA\left[\frac{D_p p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n}\right]$$

When $W_n \gg L_p \ and \ W_p \gg L_n$

and

$$I_S = qA\left[\frac{D_p p_{n0}}{W_n} + \frac{D_n n_{p0}}{W_p}\right]$$

When $W_n \ll L_p \ \ and\ W_p \ll L_n$

The first term in the expression for I_F represents the diffusion current, and the second one represents the drift current.

Reverse Bias

- Under reverse bias, the height of the potential barrier increases to $V_{bi} + V_R$, where V_R= reverse voltage.
- The expression for the depletion region under reverse bias is the same as the one for thermal equilibrium, with V_{bi} replaced by $V_{bi} + V_R$. The depletion region width increase under forward bias.
- The net current under reverse bias has the same form as the forward current, with V_F replaced by $-V_{R.}$

$$I_R = I_S\left(e^{\frac{-qV_R}{kT}} - 1\right)$$

Again, as in the forward bias case, the first term represents the diffusion current, and the second term the drift current. When $qV_R >> kT$, the reverse current is only due to the drift current.

I_S is called the saturation current.

I_R is called the leakage (or reverse) current.

- It can be shown that I_S is due to the generation of minority carriers within a diffusion length from the boundary of the depletion region, in the two neutral regions.

Non-ideal Current

- Current due to recombination of electron-hole pairs also contributes to the current in the p-n junction under forward bias. This component of current has a voltage dependence given by $e^{\frac{qV_F}{kT}}$, and is dependent on the number of traps (g-r centers) per unit volume.
- Similarly, an additional component due to the generation of electron-hole pairs in the depletion region flows in the junction under reverse bias.

Capacitance

- The depletion region behaves like a parallel plate capacitor with spacing equal to the depletion region width, with respect to applied small signal voltage. The capacitance is called the junction capacitance.

- Injected minority carriers give rise to a stored charge in the two neutral regions. In the one-sided abrupt $p^+ - n$ junction,

$$Q_s = I\tau_p$$

Where

$Q_s =$ stored charge

$\tau_p =$ minority carrier lifetime

$I =$ current through the diode

A capacitance effect arises due to the stored charge, and this is called the diffusion capacitance.

- When the diode is switched from 'on' to 'off', there is a delay in the reverse current attaining the steady-state value. This delay time is related to the minority carrier lifetime.

Breakdown Voltage

- As the reverse voltage is increased at sufficiently large voltages, when the maximum electric field in the depletion region becomes equal to the critical electrical field, the reverse current increases enormously. This voltage is called the Breakdown Voltage.
- There are three mechanisms:
    1) Thermal instability
    2) Tunneling process
    3) Avalanche process
- The breakdown voltage due to tunneling has a negative temperature coefficient.
- The cylindrical and spherical junctions have lower breakdown voltage than planar junctions.

Glossary

A = area of the junction

A' = an integration constant

A and B = two points in a semiconductor

A and B = integration constants that will be determined by the boundary conditions at $x = 0$ and $x' = W_n$

C = capacitance

C_1 = an integration constant

C_1' = an integration constant

C_1' = an integration constant

C_2 = an integration constant

C_1' = an integration constant

C_3 = an integration constant

C_d = depletion capacitance per unit area

C_{diff} = diffusion or storage capacitance

$C_{\text{diff}\,p}$ = diffusion capacitance due to stored holes

C_J = junction capacitance

dV_a = step voltage

D_n = electron diffusion constant

D_p = hole diffusion constant

E_C = energy at the bottom of the conduction band, al;so the potential energy of electrons

E_F = Fermi energy

E_{Fn} = quasi-Fermi energy level for electrons

E_{Fp} = quasi-Fermi energy level for holes

E_g = energy gap or bandgap energy

E_i = intrinsic Fermi energy level

E_t = trap energy level

E_V = energy at the top of the valence band

$\mathcal{E}$ = electric field

$\mathcal{E}_{crit}$ = critical electric field

$\mathcal{E}_{max}$ = maximum electric field in the space charge region

$\mathcal{E}(x)$ = the electric field at x

$f(x)$ = Fermi function

g_f = small signal forward conductance

g_{fn} =forward conductance due to electron injection

g_{fp} = forward conductance due to hole injection

g_n = thermal generation rates of electrons in the p region

g_p = thermal generation rates of holes in the n region

g_{th} = thermal generation rate

I = current flowing through the junction

I_F = current through a forward biased junction

I_n = forward current due to electron flow

I_p = forward current due to hole flow

I_R = current through a reverse biased junction

I_{rec} = component of forward bias current due to recombination in the space-charge region

I_{rev} = reverse current in the p-n junction

I_S = saturation current

$I_{S\,ideal}$ = ideal current prefactor

I_{Sn} = saturation current component due to the flow of electrons

I_{Sp} = saturation current component due to the flow of holes

$I_{S\,rec}$ = prefactor in the expression for current due to recombination in the space-charge region

J = total electric current density through the junction

J_n = electric current density due to electron flow

J_{n1} = electron current density in the neutral n region to replace the electrons lost due to recombination with holes in the neutral n region

J_{n2} = (injected) electron diffusion current density in the neutral p region

J_{n3} = electron current density flowing in the neutral n region to inject electrons in the p region

$J_{n\,\mathrm{diff}}$ = excess electron diffusion current

$J_{n\,\mathrm{drift}}$ = drift current density due to electron flow

J_p = electric current density due to hole flow

J_{p1} = recombining hole current density in the neutral p region

J_{p2} = (injected) hole diffusion current density

J_{p3} = hole current density that is flowing in the neutral p region to inject holes in the n region

J_S = saturation current density

J_t = the total electric current density

k = Boltzmann constant

L_p = minority carrier diffusion length (of holes) in the n region

L_n = minority carrier diffusion length (of electrons) in the p region

m^* = effective mass of the carrier

n = electron density

n_1 = electron density in the semiconductor if the Fermi energy at trap level

n_0 = thermal equilibrium density of electrons

n_{n0} = thermal equilibrium majority carrier (electron) density

n_n = electron density in the n region

n_i = electron density in an intrinsic semiconductor

n_p = electron density in the neutral p region

n_{p0} = thermal equilibrium minority carrier (electron) density

N_A = acceptor density

N_A^- = ionized acceptor density

N_D = donor density

N_D^+ = ionized donor density

$N_D - N_A$ = net impurity concentration

N_t = density of traps or g-r centers

p = hole density

p_1 = hole density in the semiconductor if the Fermi energy is at trap level

p_0 = thermal equilibrium density of holes

p_n = hole density in the neutral n region

p_{n0} = thermal equilibrium minority carrier (hole) density

p_p = hole density in the p region

p_{p0} = thermal equilibrium majority carrier (hole) density

$P(x')dx'$ = probability that an excess hole, injected at $x' = 0$, will recombine in an elementary distance dx' between $\boldsymbol{x}'$and $x' + dx'$

$P(x')$ = probability per unit distance that an excess minority carrier (hole) will recombine and be lost

P_0 = probability that an excess minority carrier will recombine and that does not vary with x

q = total electric charge, expressed in coulombs

Q = charge

Q_{ns} = stored charge due to excess electrons

Q_{ps} = stored charge due to excess holes

Q_S = stored charge

r = proportionality constant in the recombination rate

r_f = small signal forward resistance

R = recombination rate

s = surface recombination velocity

t = time

T = Temperature

U_s = net recombination rate per unit area of the surface

v_n, v_p = thermal velocity of electrons and holes respectively and is equal to $\sqrt{\frac{3kT}{m^*}}$

v_{th} = thermal velocity

V_a = applied voltage

V_{bi} = built-in voltage, potential difference between the neutral n-side and the neutral p-side

V_{br} = breakdown voltage

V_F = externally applied forward bias voltage

V_R = externally applied reverse bias voltage

W_n = width of the neutral n region

W_p = width of the neutral p region

x = distance measured with the origin at the metallurgical junction

x' = distance measured with the origin at $x = x_n$

x_d = total depletion region width = $x_n + x_p$

$x_{d\,\text{eff}}$ = part of the depletion region where the electron and hole are both minimum and equal

x_n = boundary of the depletion region on the n-side, or the width of the depletion region on the n-side

x_p = width of the depletion region on the p-side

$-x_p$ = boundary of the depletion region on the p-side

Δn = excess electron density

$\Delta n(x = -x_p)$ = excess minority carrier densities at $x = -x_p$ under forward bias

Δp = excess hole density

$\Delta p(0)$ = excess carrier (hole) density at $x' = 0$, i.e., at $x = x_n$

$\Delta p(W_n)$ = Δp evaluated at $x' = W_n$

$\Delta p(x = x_n)$ = excess minority carrier density at $x = x_n$ under forward bias, same as $\Delta p(0)$

$\in_s$ = permittivity of a semiconductor

$\boldsymbol{\eta}$ = ideality factor

μ_n = electron mobility

μ_p = hole mobility

ρ = charge density

σ_n,σ_p = capture cross-sections for electrons and holes respectively

τ_g = generation lifetime

τ_n = minority carrier (or excess carrier) lifetime in the p-region

$\tau_{n0} = \frac{1}{N_t \sigma_n v_n}$

τ_p = minority carrier (or excess carrier) lifetime in the n-region

$\tau_{p0} = \frac{1}{N_t \sigma_p v_p}$

$\emptyset$ = electrostatic potential

$\emptyset_{A-B}$ = potential difference between two points A and B

$\emptyset_n$ = electrostatic potential in an n-type material

$\emptyset_p$ = electrostatic potential in a p-type material

$\boldsymbol{\psi}$ = electrostatic potential difference

$\boldsymbol{\psi}$(x) = electrostatic potential difference at some point x

ψ_{AB} = electrostatic potential difference between x_A and x_B

Problems

1. Calculate the electrostatic potential $\emptyset_n$ or $\emptyset_p$ as the case may be, at room temperature for each of the following impurity concentrations: a) $10^{15}cm^{-3}$ boron atoms; b) $10^{16}cm^{-3}$ gallium atoms and $10^{17}cm^{-3}$ phosphorous atoms.
2. Redo problem above at T=100 ^{0}K. (Assume that all the impurity atoms are ionized at this temperature.)
3. Consider a n-type silicon sample with $N_D = 10^{16}cm^{-3}$. Calculate $\emptyset_n$ at a) 300 ^{0}K and b) 200 ^{0}K
4. An abrupt p-n junction has an impurity concentration of$10^{15}cm^{-3}$ boron atoms on the p-side and $10^{17}cm^{-3}$ antimony atoms on the n-side. Calculate the built-in potential of the junction at room temperature.
5. Calculate the depletion region width on the n-side and that on the p side as well as the total depletion region width for the case described in the previous problem. Do you expect your answer to depend on the temperature? (The impurity atoms are completely ionized in the depletion region even though in the neutral region the impurity atoms will only be partially ionized at low temperatures.) If so, why?
6. Consider an abrupt p-n junction with the net impurity concentration of $4 \times 10^{15}cm^{-3}$ on the n-side and $6 \times 10^{16}cm^{-3}$ on the p-side. Determine the built-in potential.
7. For the sample in the previous problem, calculate the depletion region width under thermal equilibrium at T=300 ^{0}K
8. For the sample in the previous problem,
   (a) What is the maximum electric field $\mathcal{E}_{max}$, in the depletion region?
   (b) At what values of x in the depletion region will the electric field be $\frac{1}{2}$ of $\mathcal{E}_{max}$?
9. Consider a one sided abrupt p^+- n junction in which the acceptor density on the p-side is $10^{18}cm^{-3}$ and the donor density on the n-side is $10^{15}cm^{-3}$. Calculate the depletion region width in thermal equilibrium.
10. An abrupt p-n junction has $10^{15}cm^{-3}$ donor atoms on the n-side and $10^{16}cm^{-3}$ acceptor atoms on the p-side. Determine $V_{bi,}x_p, x_n$ and x_d

11. Consider an abrupt junction with $N_D = 10^{16} cm^{-3}$ on the n-side and $N_A = 10^{15} cm^{-3}$ on the p-side. Assume that the edge of the depletion region occurs at $-x_p$ and x_n. Assume n_i to be $10^{10} cm^{-3}$ at room temperature.
    (a) What is the electron density at $x = -\frac{x_p}{2}$ in thermal equilibrium?
    (b) What is the hole density at the same point under the same condition?
12. The equations in the text give the potential variation in the p-side and in the n-side respectively for an abrupt junction, taking the zero reference for the potential as that at the neutral p region. Write down similar equations for the potential taking the zero reference for the potential as that at the neutral n region.
13. Consider a linear junction where the impurity concentration varies as $N = -ax$ and a is a constant. Derive an expression for the electric field and the potential variation in the depletion region assuming the boundary of the depletion region to be at $x = -\frac{W}{2}$ and $x = \frac{W}{2}$.
14. A diffused p-n junction has a uniform doping of $2 \times 10^{14} cm^{-3}$ on the n-side. The impurity concentration varies linearly from the n-side to the p-side with $a = 10^{19} cm^{-4}$. At zero bias the depletion region on the p-side is 0.8 μm. Find the depletion region width, built-in potential and maximum electric field in the depletion region.
15. Show that, for a narrow base diode, the excess carrier density variation in the neutral n region is given by

$$\Delta p(x') = \Delta p(0)\left(1 - \frac{x'}{W_n}\right)$$

16. Write down the expression for the ideal diode current when the n region is approximated as a narrow base and the p region is approximated as a wide base.
17. Consider an abrupt p-n junction in which the acceptor density on the p-side is $10^{16} cm^{-3}$ and the donor density on the n-side is $6x10^{15} cm^{-3}$. Calculate the depletion region widths on the p and on the n side for a forward bias voltage of 0.6 *V*.
18. Assume that the area of the junction for the device in the last problem is $10^{-4} cm^2$. Calculate the current through the device under a forward bias voltage of 0.5 *V* assuming that the diode is a wide base diode. Take the minority carrier lifetime in the p and in the n regions to be 10^{-6} and 10^{-5} seconds respectively. Use the mobility curve in chapter 2 in your text.
19. Consider as abrupt junction with $N_D = 10^{16} cm^{-3}$ on the n side and $N_A = 2x10^{15} cm^{-3}$ on the p side. Let the area of the junction be $10^{-3} cm^2$. Given the diffusion constant to be 20 and 10 $cm^2 s^{-1}$ and the lifetime to be 80 and 10 μs in the n and p regions respectively, determine the ideal reverse current for a reverse voltage of a) 0.06 and b) 6*V*.
20. Given that $N_t = 10^{15} cm^{-3}$, $\sigma_n = \sigma_p = 10^{-16} cm^2$ and $v_{th} = 10^7\ cm\ s^{-1}$ for both the electrons and the holes in the diode in the previous problem, calculate the forward current due to recombination in the depletion region for a forward bias of 0.4 *V*. Assume that the effective depletion region width in which the recombination is effective is $x_{\text{deff}} = \frac{x_d}{3}$ where x_d is the depletion region width.

21. Given that non-ideal reverse current is only due to generation in the depletion region occurring in the bulk of the device (i.e., neglecting surface generation) calculate the reverse current at a reverse bias of 6V.
22. Consider an abrupt p-n junction with the metallurgical junction located at $x = 0$ and the boundaries of the depletion region located at $x = -x_n$ on the n-side and at $x = x_p$ on the p-side.
    (a) What is the ratio of the electric field at $x = -r_1 x_n$ and that at $x = r_2 x_p$ where r_1 and r_2 are constants of value less than 1.
    (b) What is the built-in voltage at room temperature if $N_A = 10^{16} cm^{-3}$ in the p-side and $N_D = 10^{15} cm^{-3}$ on the n-side? Assume $n_i = 1x10^{10} cm^{-3}$.
23. An abrupt p-n junction is fabricated with $N_D = 2x10^{16} cm^{-3}$ on the n-side and $N_A = 5x\ 10^{15} cm^{-3}$ on the p-side. Let A the area of the junction be $10^{-4} cm^2$.
    (a) What fraction of the forward current is due to the minority carrier injection in the neutral p-side? Assume that D_n and L_n on the p-side to be 40 $cm^2 s^{-1}$ and $0.02\ cm$ respectively and that D_p and L_p on the n-side to be 10 $cm^2 s^{-1}$ and $0.005\ cm$ respectively.
24. An abrupt p-n junction has $10^{15} cm^{-3}$ donor atoms on the n-side and $10^{16} cm^{-3}$ acceptor atoms on the p-side. If the area of the junction is $10^{-4} cm^2$ and the minority carrier lifetime is 0.4 and 1 μs on the n and p side respectively, find the diode current at a) 0.6 and b) -1 volt bias applied across the device,
25. An abrupt p-n junction has $10^{15} cm^{-3}$ donor atoms on the n-side and $10^{19} cm^{-3}$ acceptor atoms on the p-side. Find the current through the diode for a forward bias of 0.5 volt given that the diffusion lengths in the n and p regions are 10^{-2} and $3 \times 10^{-4}\ cm$ respectively. Assume the area of the junction to be $2 \times 10^{-4} cm^2$. Take the diffusion constants from the curve in chapter 2 of the text.
26. An abrupt p-n junction has $10^{15} cm^{-3}$ donor atoms on the n-side and $10^{19} cm^{-3}$ acceptor atoms on the p-side. Calculate the junction capacitance a) at a forward bias of 0.4 volt and b) at a reverse bias of 10 volt.
27. In an abrupt p-n junction $N_D = 5x10^{16} cm^{-3}$, $N_A = 10^{17} cm^{-3}$, and the area of the junction is $10^{-3} cm^2$. Given that the minority carrier diffusion constant and lifetime are $20\ cm^2 s^{-1}$ and $2x10^{-5}\ s$ on the p side and $8\ cm^2 s^{-1}$ and $2x10^{-6}\ s$ on the n-side, calculate the current through the diode at room temperature under
    (a) A forward bias voltage of 0.5 V.

Chapter 4

Bipolar Transistor

In this chapter, we will be discussing a device called bipolar junction transistor or a bipolar transistor. The principle of the bipolar transistor action is based on the minority carrier injection into a neutral region using a forward-biased p-n junction, and the collection of the injected minority carriers by placing a second p-n junction (which is reverse-biased) very close to the forward-biased p-n junction. In order to illustrate this principle, let us first consider an ideal p-n junction. Under forward bias, the current through the junction comprises two components: one due to electron injection into the p side, the other due to hole injection into the n side. The forward biased p-n junction represents a low-impedance junction: for a small voltage applied across the junction, a large amount of current flows through that junction. On the other hand, if a p-n junction is under reverse bias, the current through the junction is due to minority carriers that enter the depletion region, and does not depend on the reverse voltage (barrier height). The same current will essentially flow independent of the reverse voltage. In this sense, it is a constant current source or a high impedance current source: even for a large change in the reverse bias, the increase in the current is negligibly small.

Two types of transistor structure are possible in the bipolar transistors, and are shown in Figure (4.1). They are called ***n*- *p*-*n*** and ***p*-*n* -*p*** transistors. Let us consider the n-p-n transistor. Assume that the first n-p junction is forward-biased, and the second junction is reverse-biased as shown in Figure (4.1A). *The first n region is called the **emitter***, the p *region the **base***, and *the second n region the **collector***. The emitter-base junction is forward biased, and therefore minority carriers (electrons) are injected into the base region. These injected minority carriers diffuse in the base region towards the collector-base junction. If the collector-base junction is close to the emitter-base, then a substantial fraction of the injected minority carriers will reach the collector-base depletion region without being lost in recombination. The carriers reaching the collector-base depletion region are driven by the electric field in the collector-base depletion region into the neutral collector region. Thus, the injected minority carriers in the base region give rise to a current through the collector lead, and this current is called the **collector current**. *The collector current is like a current from a constant current source since it is due to minority carriers that enter a reverse-biased junction*. If a large load resistance is connected between the collector and the base, then a large amount of output power will be delivered to the load resistance. Let us look at the band diagram of the device as given in Figure (4.2). The energy barrier for minority carrier flow in the emitter-base junction is reduced because it is forward-biased, giving rise to electron injection from the emitter into the base. The energy barrier in the collector-base junction is increased because it is reverse-biased. Most of the electrons injected into the base will reach the collector-base junction if the base region is narrow. Let us consider the circuit shown in Figure (4.3). A small signal voltage, $\widetilde{v_s}$ is connected between the emitter and the base in series with the forward-bias voltage source V_{BE}. A load resistor (R_L) is connected between the collector and the base, in series with the collector voltage supply (V_{CC}). Let an alternating current (i_e) flow in the emitter lead due to the small signal voltatge. Let i_c and i_b be the corresponding collector and base currents. The collector current (i_c) flows through the load resistance. The power that the small-signal source, has to deliver to the transistor is given by

$$Power\ Input = i_c^2\ r_E \tag{4.1}$$

where $\boldsymbol{r_E}$ **is the input resistance** of the forward-biased p-n junction, and is given by

$$r_E = \frac{kT}{qI_E} \tag{4.2}$$

with $\boldsymbol{I_E}$ **as the DC emitter current**. The power output is given by

$$Power\ Input = i_c^2\ R_L \tag{4.3}$$

We can now calculate the power gain, which is the ratio of the power output to the power input.

$$Power\ Gain = \frac{i_c^2\ R_L}{i_e^2\ r_E} \tag{4.4}$$

If we design the transistor so that the collector current (i_c) is very nearly equal to i_e, then the power gain is equal to

$$Power\ Gain\ \approx \frac{R_L}{r_E}$$

However, we already noticed that r_E is the resistance of a forward-biased p-n junction and hence is a very small quantity. By choosing the load resistance R_L to be large, we can obtain a large power gain. This is the principle of a bipolar transistor.

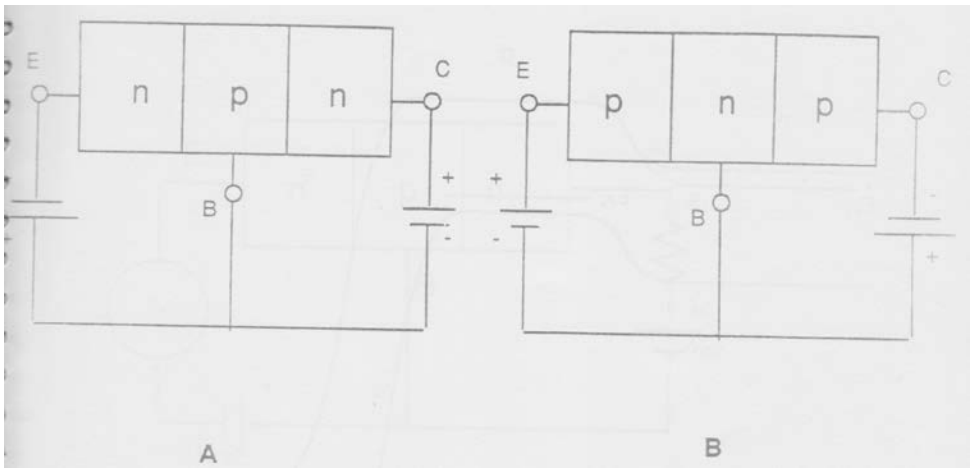


Figure (4.1): Two alternate structures for a bipolar transistor: A) n-p-n transistor B) p-n-p transistor.

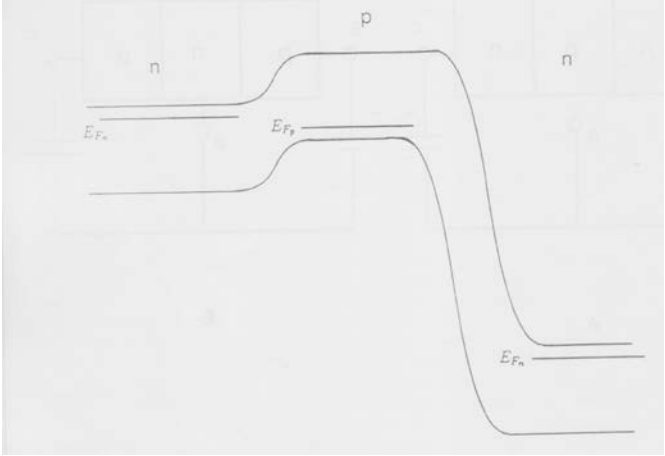


Figure (4.2): The energy band diagram for the two n-p junctions connected back-to-back.

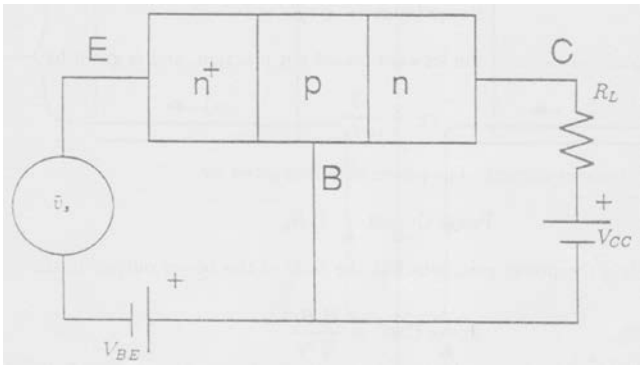


Figure (4.3): Application of a small signal voltage source between the emitter and the base of a n-p-n transistor.

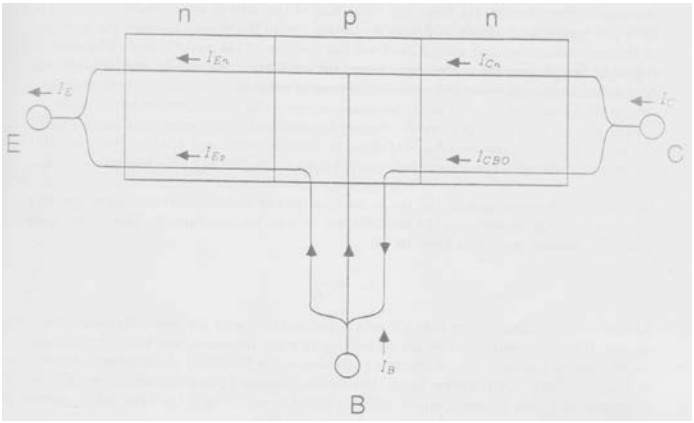


Figure (4.4): The various components of current, flowing in an n-p-n transistor

Let us now consider the various components of current flowing through the device. The DC current flowing through the emitter (which is denoted I_E) comprises two components: one due to electron injection from the emitter into the base (I_{En}), and the other due to hole injection from the base into the emitter (I_{Ep}). The electrons injected into the base diffuse toward the collector region and result in a collector current. However, not all the electrons injected into the base are able to reach the

collector-base depletion region, since some of the injected electrons recombine in the neutral-base region. Hence the electron current (I_{Cn}) that reaches the collector-depletion region and flows through the collector lead is smaller than I_{En}, the electron current injected from the emitter into the base.

In the collector, in addition to I_{Cn}, we also have a leakage current that normally arises in a reverse-biased p-n junction. Let us denote the leakage current I_{CBO}. The collector current (denoted I_C) is therefore equal to the sum of I_{CBO} (the leakage current), and I_{Cn} (the current due to the injection of electrons from the emitter into the base).

Holes have to flow into the base region to participate in the recombination process. This hole current flows through the base lead. Similarly, the holes injected into the emitter form the base also flow through the base lead. Additionally, the collector leakage current, I_{CBO}, also has to flow through the base lead, The flow of the various components of current in an n-p-n transistor is shown in Figure (4.4). In this figure, the arrows indicate the direction of the conventional electric current, and not the direction of the particle flow. The electron (particle) flow is opposite to the conventional electric current flow for I_{En} and I_{Cn}. We can now write the emitter, base and collector currents as equal to

$$I_E = I_{En} + I_{Ep} \tag{4.6}$$

$$I_C = I_{Cn} + I_{CBO} \tag{4.7}$$

$$I_B = I_E - I_C = I_{En} - I_{Cn} + I_{Ep} - I_{CBO} \tag{4.8}$$

We define the current gain of the device as equal to the collector current due to electron injection from the emitter into the base, divided by total emitter current. The **current gain** is usually denoted $\boldsymbol{\alpha_0}$, and is given by

$$\alpha_0 = \frac{I_{Cn}}{I_E} \tag{4.9}$$

The subscript $\mathbf{0}$ in α_0 implies that the gain is evaluated for *very low (zero) frequency input signals*. It must be noted that we are considering an n-p-n transistor, and hence the collector current due to minority carrier (electron) injection in the base from the emitter is denoted with a subscript n. On the other hand, if we were to consider a p-n-p transistor, the collector current arising from minority carrier injection from the emitter into the base will be due to holes and will be denoted with a subscript p in place of n.

The current gain, α_0 *, can be related* to two other parameters, which are $\boldsymbol{\gamma}$ **(the emitter injection efficiency)**, and $\boldsymbol{\alpha_T}$ **(the base transport factor)**. The emitter injection efficiency γ is defined as the ratio of the emitter current due to minority injection from the emitter into the base and the total emitter current. Hence it is equal to

$$\gamma = \frac{I_{En}}{I_E} = \frac{I_{En}}{I_{En}+I_{Ep}} \tag{4.10}$$

The base transport factor, α_T , is defined as the ratio of the collector current and the minority carrier current injected from the emitter into the base, i.e., the fraction of the injected minority carrier current that reaches the collector. This relationship is expressed by

$$\alpha_T = \frac{I_{Cn}}{I_{En}} \tag{4.11}$$

It is readily seen that α_0 is given by the product of γ and α_T , that is

$$\alpha_0 = \frac{I_{Cn}}{I_E} = \frac{I_{En}}{I_E}\frac{I_{Cn}}{I_{En}} = \gamma\ \alpha_T \tag{4.12}$$

We can now write the collector current as equal to

$$\begin{aligned} I_C &= I_{Cn} + I_{CBO} = \alpha_T I_{En} + I_{CBO} \\ &= \alpha_T \gamma\ I_E + I_{CBO} = \alpha_0 I_E + I_{CBO} \end{aligned} \tag{4.13}$$

and similarly, the base current can be expressed in terms of the emitter current as

$$I_B = I_E - I_C = (1 - \alpha_0) I_E - I_{CBO} \tag{4.14}$$

In principle, the emitter injection efficiency is made very close to unity by making the emitter impurity concentration very large in comparison to the base impurity concentration. Similarly, the base transport factor can be made very close to unity by making the width of the neutral base region (i.e., the distance which the minority carriers that are injected from the emitter into the base have to travel to reach the collector-base depletion region) very small compared with the diffusion length of the minority carriers in the base. Under these conditions, the current gain will be very close to unity.

Example

Let us now calculate the power gain of an n-p-n transistor, based on our discussions so far. Let us assume that the transistor is carrying a DC emitter current I_E, equal to 1 mA. Let the load resistance R_L be equal to 259 $K\Omega$. Let the current gain α_0 be taken as 0.99. Assume a small signal voltage $\tilde{v}_s$ is applied between the emitter and the base giving rise to an emitter current and i_e and a collector current i_c.

$$\begin{aligned} Power\ Output &= i_c^2\ \mathrm{R_L} \\ &= (0.99\ i_e)^2\ \mathrm{R_L} \\ &= (0.99 i_e)^2 \times 259\ x10^3\ W \end{aligned}$$

$$Power\ Input = i_e^2\ \mathrm{r_e}$$

But

$$\begin{aligned} r_e &= \frac{kT}{qI_E} \\ &= \frac{0.0259}{10^{-3}}\ \Omega \\ &= 25.9\ \Omega \end{aligned}$$

$$Power\ Input = i_e^2\ x\ 25.9$$

$$Power\ Gain = \frac{(0.99 i_e)^2 x\ 259\ x10^3}{i_e^2\ x\ 25.9}$$

$$= 0.99^2\ x\ 10^4$$

$$= 0.98\ x\ 10^4 = 9800$$

From the above example, we see that a large power gain is obtained by choosing a large value of R_L. From this rudimentary discussion it appears that infinite power gain can be obtained by choosing R_L to be infinitely large. This is not so. Later on, we will see that our initial assumption that the collector-base junction acts as a current source with infinite impedance is in error, and that the output impedance of the collector-base junction (the ratio of the incremental change in the collector voltage to the incremental change in the collector current) is large but not infinite. Hence the maximum power that can be drawn from the device is what is obtained with R_L equal to the output impedance of the collector-base junction.

The bipolar device is denoted by the symbol shown in Figure (4.5A) for n-p-n transistors, and that in Figure (4.5B) for p-n-p transistors. The lead with an arrow is the emitter and the direction of the arrow indicates (as in a p-n junction) the direction of the forward-bias current.

D.C. Characteristics

Let us now investigate the characteristics of the bipolar transistor in a quantitative manner. We make the following assumptions: 1) The impurity concentration in the emitter, base and collector, regions are uniform. This means that there is no spatial variation of the impurity concentration within any one of the three regions. 2) The injection of minority carriers in the base corresponds to a low-level injection. 3) There is no recombination or generation in the depletion region of the emitter-base junction, or in the collector-base junction. In other words, we treat the two junctions as ideal junctions. 4) There is no series resistance in the bulk regions of the device.

Ideal Transistor Characteristics

The characteristics of a device under the assumption of ideal junctions are called ideal transistor characteristics. Let us examine the *base region*. Referring to Figure (4.6), the metallurgical junctions on the emitter-base side and on the collector-base side are located at $x = 0$ and $x = W_{BJ}$, respectively. However, the active neutral region of the base is obtained by subtracting x_{PE} (the width of the depletion-region occurring in the base region due to the emitter-base junction), and x_{PC} (the width of the depletion-region occurring in the base region due to collector-base junction) from the width W_{BJ}. We choose a new coordinate system, x', where the origin is chosen at the edge of the neutral base

region on the emitter side. The edge of the depletion region on the collector side occurs at $x' = W'$. W' is then equal to

$$W' = W_{BJ} - x_{PE} - x_{PC} \tag{4.15}$$

From now on, we will use W' to denote the width of the neutral base region. The minority carriers injected from the emitter at $x' = 0$ diffuse in the neutral base region and travel toward the collector-based junction. This gives rise to a diffusion current. In order to derive an expression for the diffusion current due to injection of minority carriers in the base, we first set up the continuity equation in the base region. Since we are considering an n-p-n transistor, electrons are the minority carriers in the base. The continuity equation for electrons in the base region is given by

$$\frac{\partial \Delta n}{\partial t} = \frac{1}{q}\nabla \cdot J_n - \frac{\Delta n}{\tau_n} \tag{4.16}$$

Here, Δn stands for the excess minority carrier in the base region due to minority carrier injection from the emitter, J_n is the minority carrier current density, and τ_n is the minority carrier lifetime. If we assume a one-dimensional case, this continuity equation reduces to

$$\frac{\partial \Delta n}{\partial t} = \frac{1}{q}\frac{\partial}{\partial x'}(J_n) - \frac{\Delta n}{\tau_n} \tag{4.17}$$

where

$$J_n = nq\mu_n\mathcal{E} + qD_n\frac{\partial \Delta n}{\partial x'} \tag{4.18}$$

Let us assume that there is no electric field in the neutral base region. Then the expression for the current density has only the diffusion term, and the gradient of J_n becomes equal to

$$\frac{\partial J_n}{\partial x'} = qD_n\frac{\partial^2 \Delta n}{\partial x'^2} \tag{4.19}$$

Substituting this expression for the gradient of J_n in the continuity equation, we get

$$\frac{\partial \Delta n}{\partial t} = D_n\frac{\partial^2 \Delta n}{\partial x'^2} - \frac{\Delta n}{\tau_n} \tag{4.20}$$

We can solve the continuity equation first under DC conditions. We assume that *sufficient time has been allowed to elapse after the application of the DC bias so that steady-state conditions can be assumed.*

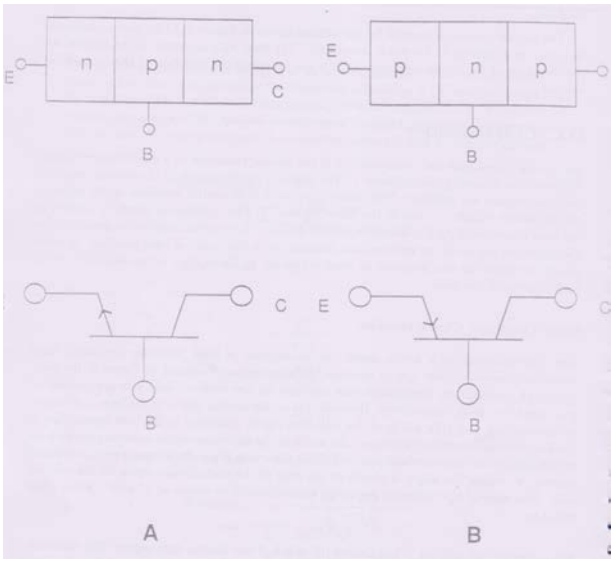


Figure (4.5): The circuit symbol for the bipolar transistor. A) n-p-n transistor and B) p-n-p transistor.

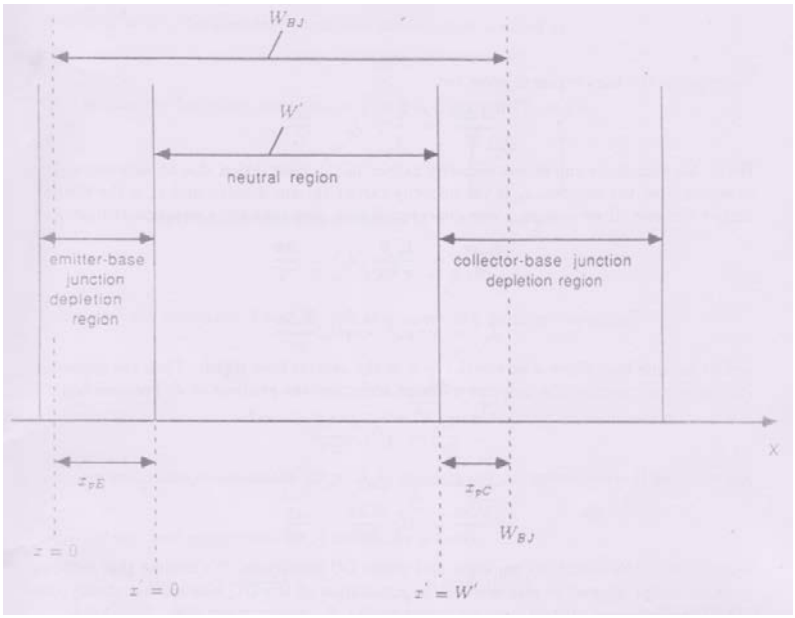


Figure (4.6): The definition of the neutral base region. The region between the metallurgical junction on the emitter side and that on the collector side represents the total base region. However, the neutral base region width is obtained by subtracting appropriate segments of the depletion region widths on the two sides.

Steady State (DC Conditions)

Under steady-state conditions, partial of Δn with respect to time is equal to zero. Hence, the continuity equation can be rearranged to obtain

$$\frac{\partial^2 \Delta n}{\partial x'^2} = \frac{\Delta n}{D_n \tau_n} = \frac{\Delta n}{L_n^2} \tag{4.21}$$

where $L_n = \sqrt{D_n \tau_n}$ is the diffusion length, which we came across in our study of the p-n junction. The solution to the continuity equation is seen to be

$$\Delta n = A\, e^{\frac{-x'}{L_n}} + B e^{\frac{x'}{L_n}} \tag{4.22}$$

The boundary conditions for the excess carrier density in the base region are:

At $x' = 0$

$$\Delta n(x' = 0) = \Delta n_0 = n_{p0}\left[e^{\frac{q\,V_{BE}}{kT}} - 1\right] \tag{4.23}$$

And at $x' = W'$, the boundary condition[4] can be approximated as

$$\Delta n = 0 \tag{4.24}$$

By applying the boundary condition at $x' = 0$ in Equation (4.22), we get

$$\Delta n_0 = \mathrm{A} + \mathrm{B} = n_{pB0}\left[e^{\frac{q\,V_{BE}}{kT}} - 1\right] \tag{4.25}$$

Similarly, by applying the boundary condition at $x' = W'$, we get

$$0 = A\, e^{\frac{-W'}{L_n}} + B e^{\frac{W'}{L_n}} \tag{4.26}$$

Solving for the constants A and B, (left as a homework problem) we obtain

$$B = \Delta n_0 \frac{e^{\frac{-W'}{L_n}}}{e^{\frac{-W'}{L_n}} - e^{\frac{W'}{L_n}}} \tag{4.27}$$

$$A = \Delta n_0 \frac{e^{\frac{W'}{L_n}}}{e^{\frac{W'}{L_n}} - e^{\frac{-W'}{L_n}}} \tag{4.28}$$

Substituting these expressions for A and B, we obtain

$$\Delta n(x') = \Delta n_0 \frac{e^{\frac{(W'-x')}{L_n}}}{e^{\frac{W'}{L_n}} - e^{\frac{-W'}{L_n}}} + \Delta n_0 \frac{e^{-\frac{(W'-x')}{L_n}}}{e^{-\frac{W'}{L_n}} - e^{+\frac{W'}{L_n}}} = \Delta n_0 \frac{\sinh(\frac{(W'-x')}{L_n})}{\sinh(\frac{W'}{L_n})} \tag{4.29}$$

[4] A more exact boundary condition than the one we have used is

$$\Delta n(x' = W') = n_{pB}\left[e^{\frac{-q\,V_{CB}}{kT}} - 1\right]$$

However, the resulting expression for $\Delta n(x')$ will look more cumbersome. Very little error results due to our simpler assumption that $\Delta n(x' = W') = 0$

This expression for Δn *is the most general expression for excess carrier density in the base region*. It may be recalled from our discussion of the p-n junction that the excess carrier density in the neutral region decreased linearly with distance when the width of the neutral region was small compared to the minority carrier diffusion length (narrow base diode) and decayed exponentially with distance when the width of the neutral region was large compared to the diffusion length (wide base diode). It is easy to show that Equation (4.29) reduces to an exponential or linear relation under appropriate assumptions.

As we will see later on, in order to make α_T (the base transport factor) as close to unity as possible, W' (the neutral base width) has to be kept very small in comparison with L_n (the minority carrier diffusion length). When $\boldsymbol{W' \ll L_n}$, the expression for $\Delta n(x)$ can be simplified as

$$\Delta n(x') \approx \frac{\Delta n_0 \frac{(W'-x')}{L_n}}{\frac{W'}{L_n}} = \Delta n_0 \left(1 - \frac{x'}{W'}\right) \tag{4.30}$$

$n(x')$ given by Equation (4.29) is plotted in Figure (4.7) as a function of x' for various values of $\frac{W'}{L_n}$. It can be noticed that when $W' \ll L_n$, a linear plot is obtained; when $W' \gg L_n$, an exponential plot is obtained. We can now calculate the diffusion current in the base due to the injected minority carriers by substituting the expression for $\Delta n(x')$ given by Equation (4.29) in the expression for electron diffusion current. The electron diffusion current density is given

$$J_n(x') = qD_n \frac{d\Delta n}{\partial x'} = -q \frac{D_n \Delta n_0}{L_n} \frac{\cosh\left(\frac{(W'-x')}{L_n}\right)}{\sinh\left(\frac{W'}{L_n}\right)} \tag{4.31}$$

Base Transport Factor:

The electron current density at $x' = 0$, is obtained by putting $x' = 0$ in the above equation.

$$J_n(x' = 0) = -q\, D_n \frac{\Delta n_0}{L_n} \frac{\cosh\left(\frac{W'}{L_n}\right)}{\sinh\left(\frac{W'}{L_n}\right)} = -q\, D_n \frac{\Delta n_0 \coth\left(\frac{W'}{L_n}\right)}{L_n} \tag{4.32}$$

The emitter current due to electrons injected from the emitter into the base (I_{En}) is equal to

$$I_{En} = A\, J_n(x' = 0) = -q\, A\, D_n \frac{\Delta n_0 \coth\left(\frac{W'}{L_n}\right)}{L_n} \tag{4.33}$$

The electron current density at $x' = W'$ is obtained by putting $x' = W'$ in Equation (4.31)

$$J_n(x' = W') = -q\, D_n \frac{\Delta n_0}{L_n} \frac{1}{\sinh\left(\frac{W'}{L_n}\right)} \tag{4.34}$$

This is the collector current density arising due to minority carrier injection in the base. **The collector current** $(\boldsymbol{I_{Cn}})$ is equal to

$$I_{Cn} = A\,J_n(x' = W') = -q\,A\,D_n\frac{\Delta n_0}{L_n}\frac{1}{\sinh(\frac{W'}{L_n})} \tag{4.35}$$

The difference between I_{En} and I_{Cn} represents the decrease in current due to recombination in the neutral base region. The base transport factor is the ratio of I_{Cn} to I_{En}, and is given by

$$\alpha_T = \frac{I_{Cn}}{I_{En}} = \frac{J_n(x'=W')}{J_n(x'=0)} = \frac{1}{\cosh(\frac{W'}{L_n})} \tag{4.36}$$

When W' becomes negligibly small in comparison with L_n, $cosh(\frac{W'}{L_n}) \approx 1$ and α_T becomes equal to 1.
This can be seen from a comparison of the electron current density at $x' = 0$ and that at $x' = W'$.
When $W' \ll L_n$,

$$J_n(x' = 0) = -\frac{q\,D_n\Delta n_0}{W'} \tag{4.37}$$

$J_n(x' = W')$ is also equal to $-\frac{q\,D_n\Delta n_0}{W'}$. Therefore, $\alpha_T = 1$. This reflects the fact that there is no recombination in the base region.

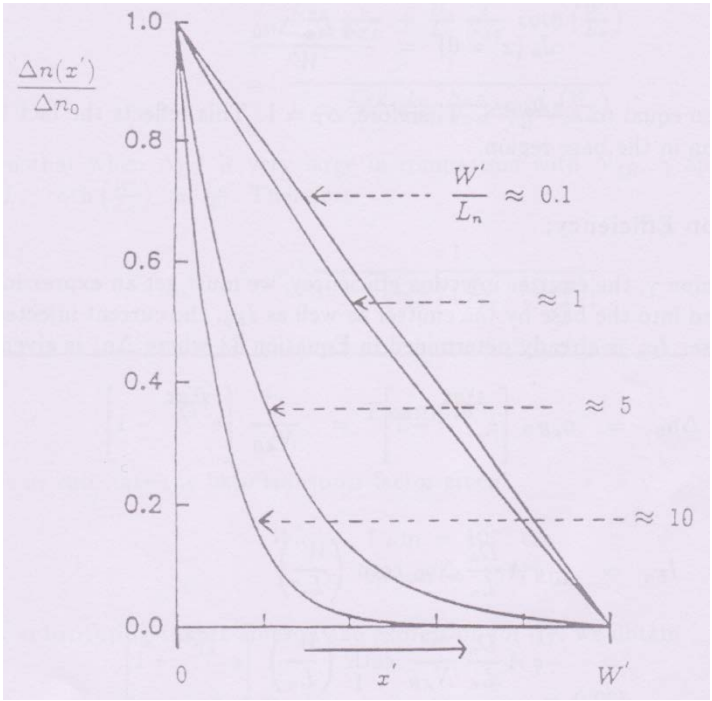


Figure (4.7): A plot of the excess minority carrier density in the base region for various values of $\frac{W'}{L_n}$.

Emitter Injection Efficiency:

In order to determine γ, the emitter injection efficiency, we must get an expression for I_{En}, the current injected into the base by the emitter as well as I_{Ep}, the current injected into the emitter by the base, I_{En} is already determined in Equation (4.33) where Δn_0 is given by

$$\Delta n_0 = n_{pB0}\left[e^{\frac{qV_{BE}}{kT}} - 1\right] = \frac{n_i^2}{N_{AB}}\left[e^{\frac{qV_{BE}}{kT}} - 1\right] \tag{4.38}$$

Therefore

$$I_{En} = -q\,A\,\frac{D_n}{L_n}\,\Delta n_0 \coth\,(\frac{W'}{L_n})$$

$$= -q\,A\,\frac{D_n}{L_n}\frac{n_i^2}{N_{AB}}\,\coth\,(\frac{W'}{L_n})\left[e^{\frac{qV_{BE}}{kT}} - 1\right] \tag{4.39}$$

The component I_{En} of the emitter current due to hole injection from the base into the emitter can similarly be obtained by solving the continuity equation for holes in the neutral emitter region. Let us assume that the width of the neutral emitter region, W_E, is large compared with the minority carrier (hole) diffusion length (L_{PE}) in the emitter. Then we can use the wide base approximation in the emitter region. We will get an expression for the injected hole current density which decays exponentially (because of our assumption that W_{nE} is much large than L_{pE}). We determine the expression for the hole current in the emitter just at the boundary of the emitter-base depletion region, and take that as the current injected into the emitter by the base. Hence I_{Ep} is given by

$$I_{Ep} = -q\,A\,\frac{D_{pE}}{L_{pE}}p_{nE0}\left[e^{\frac{qV_{BE}}{kT}} - 1\right]$$

$$= -q\,A\,\frac{D_{PE}}{L_{pE}}\,\frac{n_i^2}{N_{DE}}\,\left[e^{\frac{qV_{BE}}{kT}} - 1\right] \tag{4.40}$$

The emitter injection efficiency, γ, can be obtained as

$$\gamma = \frac{I_{En}}{I_E} = \frac{I_{En}}{I_{En} + I_{Ep}}$$

$$= \frac{\frac{D_n}{L_n}\frac{1}{N_{AB}}\coth\,(\frac{W'}{L_n})}{\frac{D_{pE}}{L_{pE}}\frac{1}{N_{DE}} + \frac{D_n}{L_n}\frac{1}{N_{AB}}\coth\,(\frac{W'}{L_n})}$$

$$= \frac{1}{1 + \frac{D_{pE}}{D_n}\frac{L_n}{L_{pE}}\frac{N_{AB}}{N_{DE}}tanh(\frac{W'}{L_n})} \tag{4.41}$$

It is seen that when N_{DE} is very large in comparison with N_{AB}, γ approach 1. When $W' \ll L_n$, $\coth\,(\frac{W'}{L_n}) \approx \frac{L_n}{W'}$. Therefore

$$\gamma = \frac{1}{1 + \left[\left(\frac{D_{pE}}{D_n}\right)\left(\frac{W'}{L_{pE}}\right)\left(\frac{N_{AB}}{N_{DE}}\right)\right]} \tag{4.42}$$

Example

Let us calculate the base transport factor given:

$$W' = 1\ \mu m = 10^{-4} cm$$

$$L_n = 10\ \mu m = 10^{-3} cm$$

By substituting these values in the expression for α_T, we obtain

$$\alpha_T = \frac{1}{\cosh(\frac{W'}{L_n})} = \frac{1}{1.005} = 0.995 \tag{4.43}$$

Example

Let us now calculate the emitter injection efficiency, given

$$N_{DE} \approx 10^{19} cm^{-3}$$

$$N_{AB} = 10^{17} cm^{-3}$$

$$W' = 1.0\ \mu m$$

$$L_{pE} = 10\ \mu m$$

$$D_{pE} = 1\ cm^2/sec$$

$$D_n = 10\ cm^2/sec$$

By substituting these values in the expression for γ, we obtain

$$\gamma = \frac{1}{1 + \left[\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{100}\right)\right]} = \frac{1}{1 + \frac{1}{10^4}} \approx 0.9999$$

DC Current-Voltage Characteristics

The DC current-voltage characteristics of the bipolar transistor can be measured in either of two modes: the common base mode and the common emitter mode.

Common Base Mode

In the common base mode, the collector current is measured as a function of the collector-base voltage for a specified value of the emitter current using the experimental configuration shown in Figure

(4.8). The emitter-base voltage (V_{BE}) is adjusted to a suitable value, such that a specified emitter current is obtained. The collector current is measured for different values of the collector-base voltage, V_{CB}, while adjusting V_{BE} to keep I_E constant at the specified value. Thus a family of curves can be obtained, each curve corresponding to a specific value of emitter current, as shown in Figure (4.9). The x-axis in this figure is the collector-base voltage, V_{CB}, and the y –axis is the collector current, I_C. When V_{CB} is positive[5], the collector-base junction is reversed-biased. Since α_0 is very close to unity, the collector current is very nearly equal to the emitter current. When V_{CB} is negative, the collector-base junction is forward-biased, and the collector also injects minority carriers into the base just as the emitter does. Hence as the forward bias on the collector is increased, the collector current decreases and ultimately becomes zero. When the collector-base junction is forward-biased, the transistor is said to be saturated, or said to be operating in the region of saturation. When the collector-base junction is reverse-biased, the excess carrier density at $x' = W'$ (the edge of the depletion region of the collector-base junction) is negligibly small. The excess carrier density profile in the neutral base region under this condition is shown in Figure (4.10A). On the other hand, when the collector-base junction is forward-biased, the excess carrier density at $x' = W'$ is increased due to injection of minority carriers from the collector into the base. The excess carrier density profile in the neutral base region under this condition is shown in Figure (4.10B). The slope of the excess carrier density profile is less now and hence the collector current is reduced.

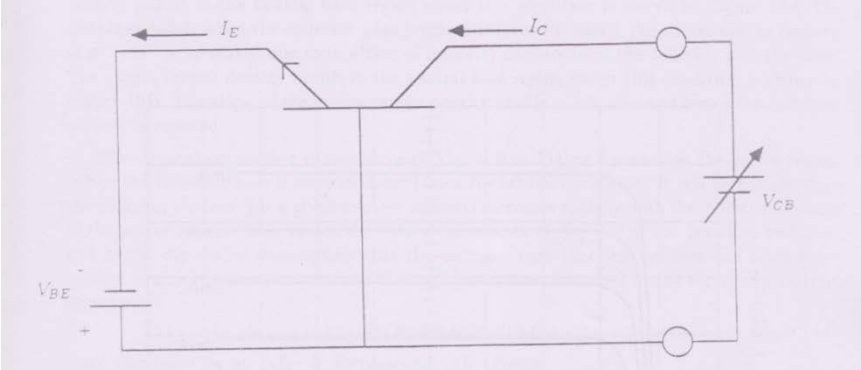


Figure (4.8): Experimental measurement of $I_C - V_{CB}$ characteristics.

[5] Remember we are dealing with an n-p-n transistor

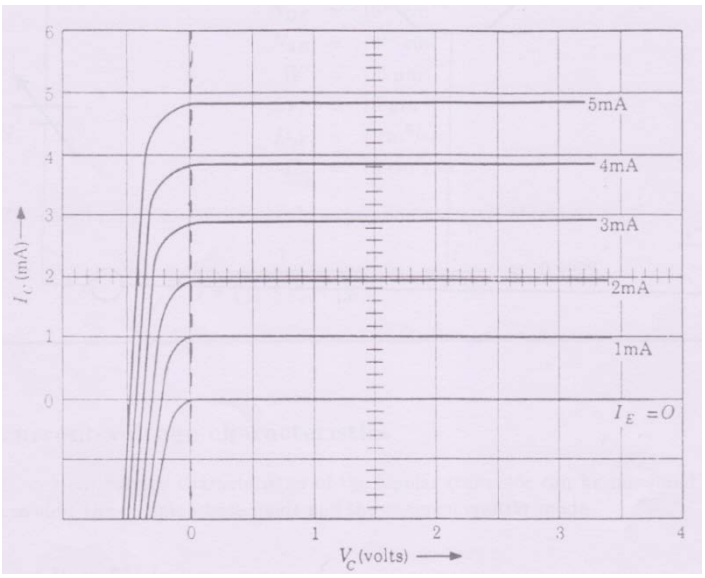


Figure (4.9): The collector current-voltage characteristics of an n-p-n transistor in the common base mode.

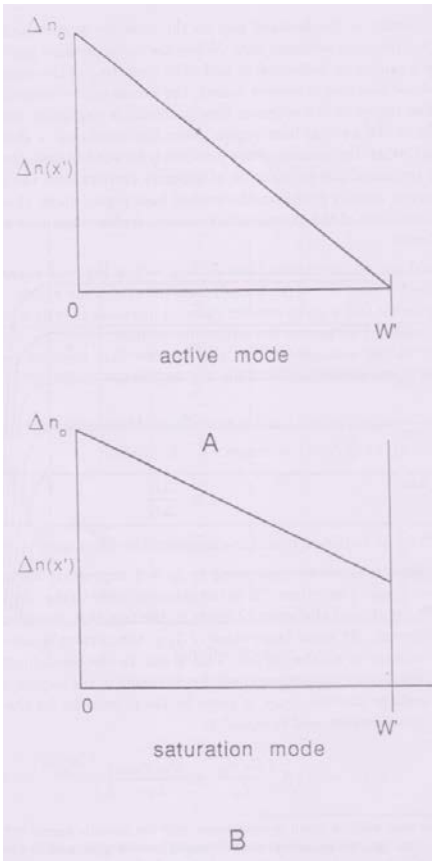


Figure (4.10): The excess carrier density in the base region A) when the collector-base junction is reverse-biased, the device is said to operate in the active mode. B) When the collector-base junction is forward-biased, the device is said to operate in the saturation mode.

The vertical dotted line drawn through V_{CB}= 0 in Figure (4.9) separates the active region (when the collector-base is reverse-biased) from the saturation region. It will be noticed that the collector current (for a given emitter current) increases slightly with the collector voltage in the active region. This causes the output impedance $\left(\left(\frac{dI_c}{dV_{CB}}\right)^{-1}\right)$ of the device to be finite, and hence our earlier assumption that the collector-base junction behaves like a constant current source is not strictly valid. This will be further discussed in the section on output impedance.

$$I_C = \alpha_0 I_E + I_{CBO} \tag{4.44}$$

If we increment I_E by ΔI_E, I_C increases by ΔI_C. Hence

$$\alpha_0 = \frac{\Delta I_C}{\Delta I_E} \tag{4.45}$$

Therefore, α_0 can be obtained from the characteristics illustrated in Figure (4.9).

The characteristic curve corresponding to $I_E = 0$, represents the collector-base junction leakage current, I_{CBO}. The letters CB in the subscript refer to the fact that it is the collector-base junction current, and the letter o refers to the fact that the emitter is open, i.e. there is no emitter current. At some large value of V_{CB}, the current is seen to increase suddenly with collector voltage in all the curves. This is due to the breakdown of the collector-base junction. The breakdown phenomenon will be discussed in the section on voltage limitations. The collector leakage current, I_{CBO}, is given by the expression for the leakage current for a reverse-biased p-n junction and is equal[6] to

$$I_{CBO} = -qA\left[\frac{D_n n_{p0}}{L_n} + \frac{D_{pC} p_{nC0}}{L_{pC}}\right]\left[e^{\frac{-qV_{CB}}{kT}} - 1\right] \tag{4.46}$$

Common Emitter Mode

In the common emitter mode, the collector current-voltage characteristics are measured as a function of the collector-emitter voltage, keeping the base current constant. The experimental set-up to obtain these characteristics is illustrated in Figure (4.11). As in the previous method, the collector current is measured at different values of the collector-emitter voltage while adjusting the base-emitter voltage to keep the base current constant at a specified value. As before, a family of $I_C - V_{CE}$ curves is obtained, each curve corresponding to a particular value of the base current (I_B) as shown in Figure (4.12). The voltage V_{CE} represents the sum of the voltage drop across the collector-base junction and that across the base-emitter junction.

$$V_{CE} = V_{CB} + V_{BE}$$

Hence, when V_{CE} becomes less than V_{BE}, V_{CB} becomes negative, and the collector-base junction becomes forward-biased. The device becomes saturated under this condition. In Figure (4.12), the

[6] Although the base width is small in comparison with the minority carrier diffusion length, we use the wide base expression for leakage current since minority carriers generated in the base within a diffusion length from the collector base junction will contribute to the leakage current.

portion of the curve lying to the left of the dotted line marked $V_{CB} = 0$ corresponds to the saturation region of operation, while the portion to the right corresponds to the active region of operation. As expressed before, the collector current I_C is equal to

$$I_C = \alpha_0 I_E + I_{CBO}$$

But

$$I_B = (1 - \alpha_0)\, I_E - I_{CBO}$$

Referring terms in this equation to obtain I_E, we get

$$I_E = \frac{I_B}{(1-\alpha_0)} + \frac{I_{CBO}}{(1-\alpha_0)} \tag{4.47}$$

Substituting this expression for I_E in the equation for I_C , we get

$$I_C = \alpha_0 \left(\frac{I_B}{1-\alpha_0} + \frac{I_{CBO}}{1-\alpha_0}\right) + I_{CBO}$$

$$= \left(\frac{\alpha_0}{1-\alpha_0}\right) I_B + \left(\frac{\alpha_0}{1-\alpha_0}\right) I_{CBO} + I_{CBO} \tag{4.48}$$

If we give an increment ΔI_B in the base current, the collector current increases by ΔI_C. We now define a current gain, β_0 in the common emitter mode as equal to

$$\beta_0 \equiv \frac{\Delta I_C}{\Delta I_B} \tag{4.49}$$

From the last equation for I_C,

$$\Delta I_C = \left(\frac{\alpha_0}{1-\alpha_0}\right) \Delta I_B \tag{4.50}$$

since I_{CBO} does not change due to a change in I_B. Hence

$$\beta_0 = \frac{\alpha_0}{1-\alpha_0} \tag{4.51}$$

Hence, I_C can be written in terms of I_B as

$$I_C = \beta_0 I_B + \beta_0 I_{CBO} + I_{CBO} = \beta_0 I_B + (\beta_0+1)\, I_{CBO} \tag{4.52}$$

The collector current with the base open, i.e., with $I_B = 0$, is denoted $I_{CEO,}$ and from this equation we can see that

$$I_{CEO} = (\beta_0+1)\, I_{CBO} \tag{4.53}$$

Again, the letters CE refer to the fact that the current is flowing between the collector and the emitter, and the letter o signifies the fact that the base is open, i.e., no base current is flowing. The leakage current in the common emitter mode, $I_{CEO,}$ is larger than the leakage current in the common base mode, I_{CBO}, by a factor $(\beta_0 + 1)$.

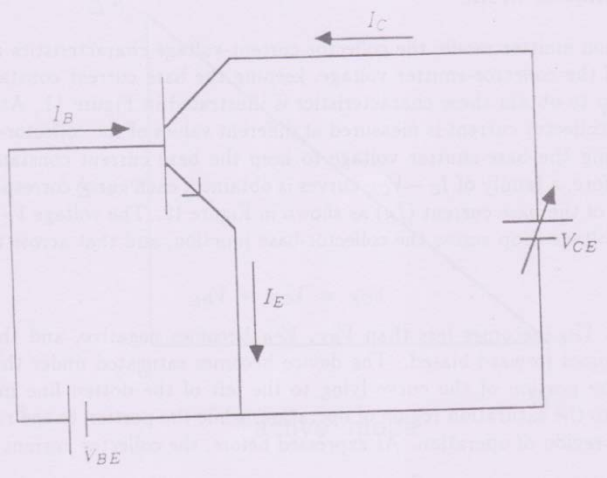


Figure (4.11): Experimental measurement of $I_C - V_{CE}$ characteristics.

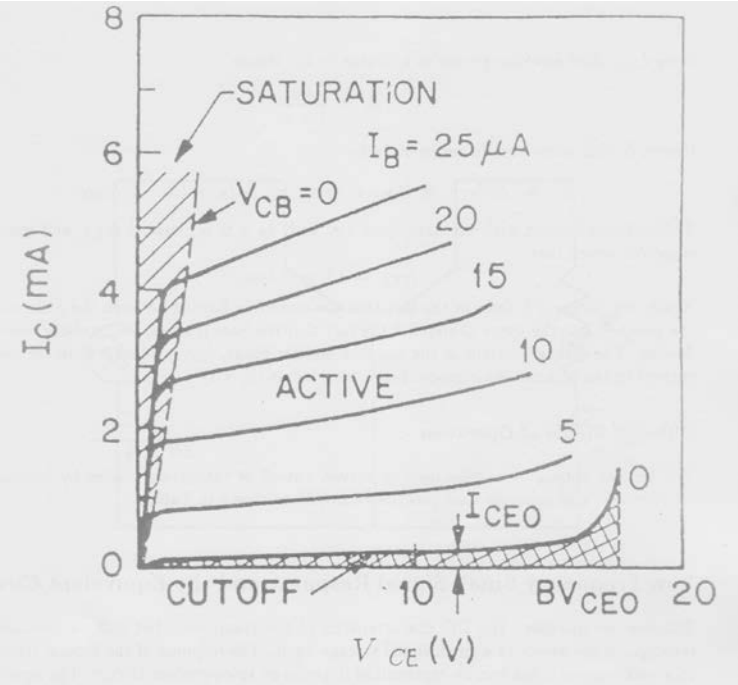


Figure (4.12): The collector current-voltage characteristics of a n-p-n transistor in the common emitter mode.

Different Modes of Operation

The bipolar transistor can be used in active, cut-off or saturation modes by biasing the emitter-base and collector-base junctions suitably as shown in Table 1.

Mode	E-B Junction Bias	C-B Junction Bias
active	forward	reverse
cut-off	reverse	reverse
saturation	forward	forward
inverse	reverse	forward

Table 1: Modes of Operation

Low Frequency Small Signal Response and the Equivalent Circuit

Till now we discussed the DC characteristics of the transistor. Let us now consider the response of the device to a small signal voltage input. The response of the bipolar transistor to a small signal input can be represented in terms of an equivalent circuit. The equivalent circuit will be different for common base and common emitter modes of operation.

Common Base Mode

Consider the circuit shown in Figure (4.13). V_{CC} is the DC collector supply voltage. A small signal voltage source ($\tilde{v}_s$) is connected between the emitter and the base in series with the DC bias voltage (V_{BE}). We assume that the frequency of the small signal voltage source is very low and that the injected minority carrier density everywhere in the base is varying in step with voltage variation. Since $\tilde{v}_s$ is varying sinusoidally with time, the injected carrier density in the base varies sinusoidally as shown in Figure (4.14). The top line in this figure shows the minority carrier density in the base when the sinusoidal function ($\tilde{v}_s$) is at the most positive excursion. Assuming $\tilde{v}_s = v_s \sin \omega t$, the base emitter voltage at its most positive excursion is $V_{BE} + v_s$. The middle line in this figure represents the carrier distribution when the same function is zero. The bottom line represents the carrier density when $\tilde{v}_s$ is at its most negative excursion, with the instantaneous base-emitter voltage equal to $V_{BE} - v_s$. The excess carrier density Δn at $x' = 0$ is given by

$$\Delta n_0 = \Delta n(x' = 0) = n_{pBO}\left[e^{\frac{q(V_{BE}+\tilde{v}_S)}{kT}} - 1\right] \tag{4.54}$$

Denoting the total emitter current as I_{Et}, we obtain

$$I_{Et} = I_{ES}\left[e^{\frac{q(V_{BE}+\tilde{v}_S)}{kT}} - 1\right] \approx I_{ES}\, e^{\frac{q(V_{BE}+\tilde{v}_S)}{kT}} = I_{ES}\, e^{\frac{qV_{BE}}{kT}} e^{\frac{q\tilde{v}_S}{kT}} \tag{4.55}$$

Expanding $e^{\frac{q\tilde{v}_S}{kT}}$ as an exponential series and retaining only the leading terms, we obtain

$$I_{Et} \approx I_{ES}\, e^{\frac{qV_{BE}}{kT}}\left(1 + \frac{q\tilde{v}_S}{kT}\right) \tag{4.56}$$

I_{Et} therefore comprises a DC term and a sinusoidally varying term, and can expressed as

$$I_{Et} = I_E + \tilde{\imath}_e \tag{4.57}$$

where I_E is the DC current and $\tilde{\imath}_e$ is the sinusoidally varying component of current. By comparing Equations (4.56) and (4.57), we obtain

$$I_E = I_{ES}\, e^{\frac{qV_{BE}}{kT}} \tag{4.58}$$

which is the DC emitter current when no sinusoidally varying voltage is applied.

$$\tilde{\imath}_e = I_E\, e^{\frac{q\tilde{v}_S}{kT}} \tag{4.59}$$

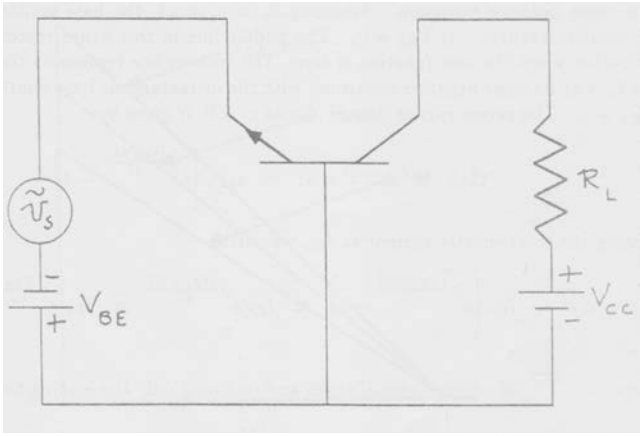


Figure (4.13): Common base mode amplifier connection. $\tilde{v}_S$ is the small signal sinusoidal input voltage. R_L is the load resistor in the collector circuit.

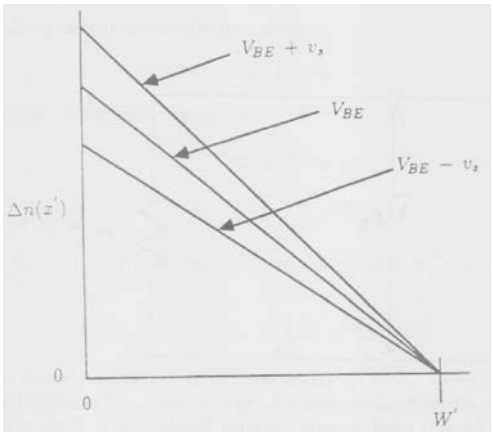


Figure (4.14): Modulation of the injected minority carrier density in the base in step with a small sinusoidal signal voltage.

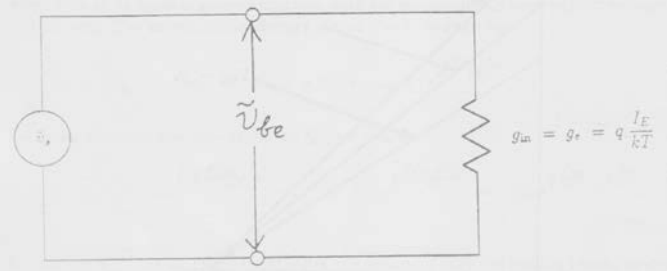


Figure (4.15): Equivalent circuit of the input circuit in the common base mode.

The ratio of the sinusoidal input voltage, $\tilde{v}_s$, to the sinusoidal emitter current (input current) $\tilde{\imath}_e$ defines an equivalent resistor which has a conductance equal to $\left(\frac{q}{kT}\right) I_E$. The input circuit is equivalent to a resistance (or conductance) and is shown in Figure (4.15). The ratio $\frac{\tilde{\imath}_e}{\tilde{v}_s}$ is the small signal *input conductance*, $\boldsymbol{g_{in}}$, and $\frac{1}{g_{in}} = r_{in} = r_E$ is called the input resistance. Due to $\Delta n(x)$ varying sinusoidally in step with $\tilde{v}_s$, a small signal collector current flows in step with $\tilde{v}_s$, the total collector current I_{Ct} is given by

$$I_{Ct} = \alpha_0 I_{Et} + I_{CBO} = \alpha_0 I_E + \alpha_0 \tilde{\imath}_e + I_{CBO} \tag{4.60}$$

The DC current in the collector is

$$I_C = \alpha_0 I_E + I_{CBO}$$

and the small signal sinusoidal collector current is

$$\tilde{\imath}_c = \alpha_0 \tilde{\imath}_e \tag{4.61}$$

The small signal current gain is equal to

$$\frac{\tilde{\imath}_{out}}{\tilde{\imath}_{in}} = \frac{\tilde{\imath}_c}{\tilde{\imath}_e} = \alpha_0 \tag{4.62}$$

The output voltage is

$$\tilde{v}_{out} = \tilde{\imath}_c R_L = \alpha_0 \tilde{\imath}_e R_L = \alpha_0 \frac{q}{kT} I_E \, \tilde{v}_s \, R_L = g_m \tilde{v}_s \, R_L \tag{4.63}$$

where

$$g_m = \alpha_0 \frac{q}{kT} I_E = \frac{\alpha_0}{r_E} \tag{4.64}$$

is called the ***transconductance***. In our treatment, $\tilde{v}_s = \tilde{v}_{be}$ where $\tilde{v}_{be}$ is the sinusoidally varying small signal emitter-base voltage. Transconductance relates the output (collector) current to the input (emitter-base) voltage. Since $\tilde{\imath}_c$ flows through a reverse-biased p-n junction, $\tilde{\imath}_c$ is (nearly) independent of the value of R_L, the load resistor. This is equivalent to a constant current source, $\tilde{\imath}_c$. The output circuit behaves as though a current source $g_m \tilde{v}_s$ $(g_m \tilde{v}_{be})$ exists between the collector and the emitter. The equivalent circuit for the output circuit is shown in Figure (4.16). Combining the equivalent circuits

for the input and the output sides, the equivalent circuit for the transistor is shown in Figure (4.17). The rectangular box enclosed by the dotted line outlines the transistor. The voltage gain equals

$$\frac{\tilde{v}_{out}}{\tilde{v}_{in}} = \frac{g_m \tilde{v}_s R_L}{\tilde{v}_s} = g_m R_L \tag{4.65}$$

The power gain equals

$$\frac{\tilde{\imath}_{out}\tilde{v}_{out}}{\tilde{\imath}_{in}\tilde{v}_{in}} = \alpha_0 \, g_m \, R_L = \alpha_0{}^2 \, g_{in} \, R_L = \alpha_0{}^2 \frac{R_L}{r_E} \tag{4.66}$$

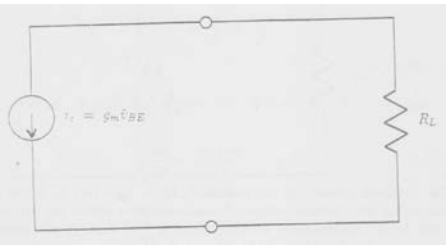


Figure (4.16): Equivalent circuit of the output circuit in the common base mode

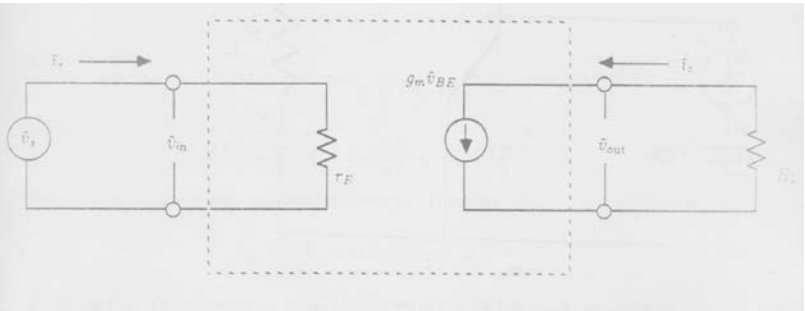


Figure (4.17): Equivalent circuit for the bipolar transistor in the common base mode.

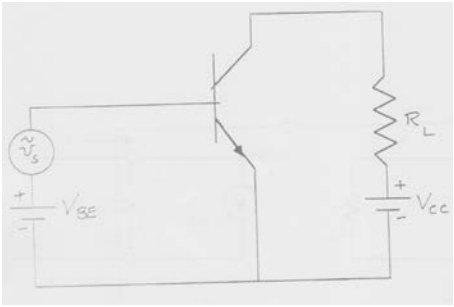


Figure (4.18): Circuit connection for common emitter mode amplifier. $\tilde{v}_s$ is the small signal sinusoidal input voltage, and R_L is the load resistor.

Common Emitter Mode

Consider the circuit shown in Figure (4.18). In the common emitter mode, the small signal sinusoidal input voltage is connected in series with the DC bias base-emitter voltage across the base and the emitter. V_{CC} is the collector supply voltage. The load resistor R_L is connected between the collector

and the emitter in series with the collector supply voltage. Thus the emitter is common to both the input and the output circuit. The input current is the base current. The total base current (I_{Bt}) is given by

$$I_{Bt} = (1-\alpha_0)\, I_{Et} - I_{CBO} = (1-\alpha_0)\, I_E + (1-\alpha_0)\, \tilde{\imath}_e - I_{CBO} = I_B + \tilde{\imath}_b \qquad (4.67)$$

The base current thus contains a DC current (I_B) equal to $(1-\alpha_0)\, I_E - I_{CBO}$, and a small signal current ($\tilde{\imath}_b$) equal to $(1-\alpha_0)\, \tilde{\imath}_e$. The input current is equal to

$$\tilde{\imath}_{in} = \tilde{\imath}_b = (1-\alpha_0)\, \tilde{\imath}_e \qquad (4.68)$$

But

$$\tilde{\imath}_e = \frac{\tilde{v}_s}{r_E}$$

Hence

$$\tilde{\imath}_b = (1-\alpha_0)\frac{qI_E}{kT}\tilde{v}_s = \tilde{v}_s \frac{1-\alpha_0}{r_E} \qquad (4.69)$$

The ratio $\frac{\tilde{\imath}_{in}}{\tilde{v}_s}$ is called the ***input conductance, g_{in}***. Therefore

$$g_{in} = \frac{\tilde{\imath}_{in}}{\tilde{v}_s} = (1-\alpha_0)\frac{qI_E}{kT} = \frac{1-\alpha_0}{r_E} \qquad (4.70)$$

Recalling that we defined r_E as equal to $\frac{kT}{qI_E}$, we can write the input resistance as

$$r_{in} = \frac{1}{g_{in}} = \frac{kT}{qI_E\,(1-\alpha_0)} = \frac{r_E}{(1-\alpha_0)} \qquad (4.71)$$

The input circuit therefore behaves as though there were a resistance, r_{in}, between the base and emitter terminals. The output current is equal to

$$\tilde{\imath}_{out} = \tilde{\imath}_c = \alpha_0 \tilde{\imath}_e = \alpha_0 \frac{qI_E}{kT}\tilde{v}_s = \alpha_0 \frac{\tilde{v}_s}{r_E} = g_m \tilde{v}_s \qquad (4.72)$$

As before, the output circuit behaves as though a current source, $g_m \tilde{v}_s$, exists between the collector and the emitter. The equivalent circuit is shown in Figure (4.19). In this instance, the current gain is equal to

$$\frac{\tilde{\imath}_{out}}{\tilde{\imath}_{in}} = \frac{\tilde{\imath}_c}{\tilde{\imath}_b} = \frac{g_m \tilde{v}_s}{\tilde{\imath}_{in}} = \frac{\alpha_0 \frac{qI_E}{kT}\tilde{v}_s}{(1-\alpha_0)\frac{qI_E}{kT}\tilde{v}_s} = \frac{\alpha_0}{1-\alpha_0} = \beta_0 \qquad (4.73)$$

Thus the current gain in the common emitter mode β_0. The subscript 0 in the α_0 and β_0 refers to the fact we are considering a very low frequency input signal, i.e., at nearly zero frequency. We are neglecting capacitive effects. The voltage gain equals

$$\frac{\tilde{v}_{out}}{\tilde{v}_{in}} = \frac{g_m \tilde{v}_s R_L}{\tilde{v}_s} = g_m R_L \qquad (4.74)$$

and the power gain equals

$$\frac{\tilde{\imath}_{out}}{\tilde{\imath}_{in}}\frac{\tilde{v}_{out}}{\tilde{v}_{in}} = \beta_0 g_m R_L = \frac{\alpha_0\beta_0}{r_E}R_L \tag{4.75}$$

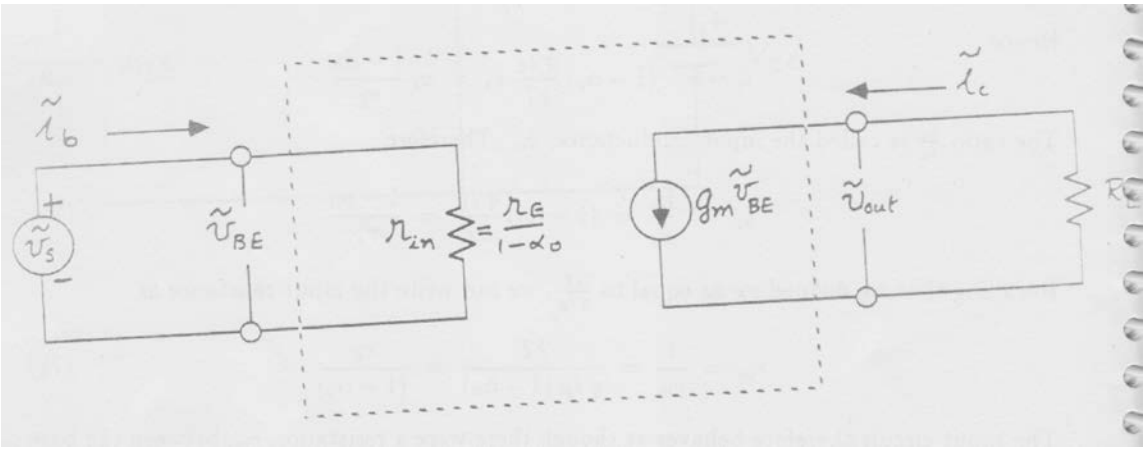


Figure (4.19): Equivalent circuit of the transistor in the common emitter mode.

Voltage Limitation

The maximum collector voltage that can be applied is limited by one of the following physical mechanisms: thermal, punch-through, the breakdown of the collector-base junction in the common base mode, or the collector breakdown in the common emitter mode. In a practical device, the limit on the magnitude of the collector voltage that can be applied is due to that mechanism which yields the lowest breakdown voltage. In some devices the punch-through voltage may be the limiting mechanism while in some other devices the collector breakdown maybe the limiting mechanism. We will now examine the three voltage limiting mechanisms.

Punch-Through Voltage, V_{pt}

When the reverse collector-base junction voltage (V_{CB}) is increased, the depletion region widens, and hence the width of the neutral base region decreases. This is illustrated in Figure (4.20). W_{BJ} is the width of the base region between the metallurgical junctions on the emitter side, and that on the collector side. The width of the segment of the emitter-base depletion region occurring on the base side is denoted x_{pE}. Similarly, the width of the segment of the collector-base depletion region occurring on the base side is denoted x_{pC}. The width of the neutral base region is denoted W', and is equal to

$$W' = W_{BJ} - x_{pE} - x_{pC} \tag{4.76}$$

However, we know from our study of the p-n junctions that

$$x_{pE} = \sqrt{\frac{2\,\epsilon_s(V_{biE}-V_{BE})N_{DE}}{q\,N_{AB}(N_{DE}+N_{AB})}} \approx \sqrt{\frac{2\,\epsilon_s(V_{biE}-V_{BE})}{q\,N_{AB}}} \tag{4.77}$$

Since $N_{DE} \gg N_{AB}$ and

$$x_{pC} = \sqrt{\frac{2\,\epsilon_s(V_{biC}+V_{BC})N_{DC}}{q\,N_{AB}(N_{DC}+N_{AB})}} \tag{4.78}$$

Hence

$$W' = W_{BJ} - \sqrt{\frac{2\,\epsilon_s(V_{biE}-V_{BE})}{q\,N_{AB}}} - \sqrt{\frac{2\,\epsilon_s(V_{biC}+V_{BC})N_{DC}}{q\,N_{AB}(N_{DC}+N_{AB})}} \tag{4.79}$$

where V_{biE} and V_{biC} are the built-in voltages of the emitter-base and collector-base junctions. N_{DE}, N_{AB} and N_{DC} are the impurity concentrations in the emitter, base and collector regions, respectively.

As the collector-base applied voltage V_{CB} is increased, x_{pC} which is the third term on the right-hand side of Equation (4.79), representing the portion of the collector-base depletion region lying on the base side increases. Thus, W' decreases as V_{CB} is increased. At some value of voltage $V_{CB} \equiv V_{pt}$ (called the ***punch-through voltage***) W' becomes zero. V_{pt} can be determined by putting $W' = 0$ in the above equation and solving for V_{pt}. Hence,

$$V_{pt} = \frac{q\,N_{AB}(N_{AB}+N_{DC})}{2\,\epsilon_s N_{DC}}\left(W_{BJ} - \sqrt{\frac{2\,\epsilon_s(V_{biE}-V_{BE})}{q\,N_{AB}}}\right)^2 - V_{biC} \tag{4.80}$$

Under the punch through conditions, the depletion region extends all the way from the collector to the emitter, and the base current (or voltage) has no control on the collector current. The collector current that flows under this condition is a space-charge limited current, and is called the ***punch-through current.*** The punch-through phenomenon is important only in transistors in which the metallurgical base region which is small.

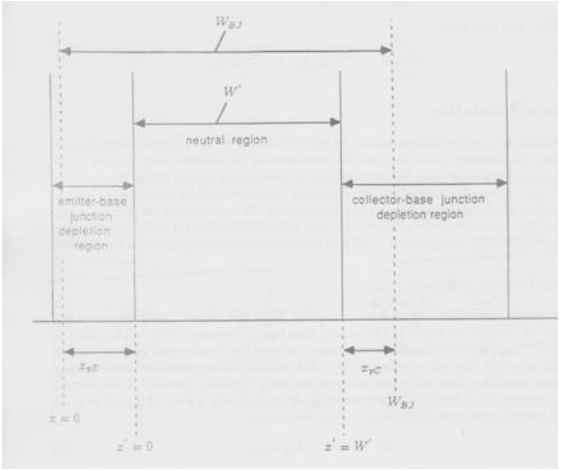


Figure (4.20): The neutral base region.

Example

Given that the impurity concentration in the emitter, base and collector regions are $N_{DE} = 10^{19} cm^{-3}$, $N_{AB} = 10^{16} cm^{-3}$ and $N_{DC} = 10^{15} cm^{-3}$, and the width of the base region is $W_{BJ} = 2\ \mu m$, determine the punch-through voltage, given the emitter-base junction forward voltage ot be 0.5 V. The built-in voltage, V_{biE}, of the emitter base junction is

$$V_{biE} = \frac{kT}{q} \ln\left(\frac{N_{DE} N_{AB}}{n_i^2}\right) = 0.0258 \left(\frac{10^{35}}{10^{20}}\right) = 0.893\ V$$

$$x_{pE} \approx \sqrt{\frac{2\ \epsilon_s (V_{biE} - 0.5)}{q\ N_{AB}}} = \sqrt{\frac{2 \times 11.9 \times 8.84 \times 10^{-14} \times 0.393}{1.6x10^{-19} \times 10^{16}}} = 0.227 \mu m$$

The built-in voltage of the collector-base junction is

$$V_{biC} = \frac{kT}{q} \ln\left(\frac{N_{AB} N_{DC}}{n_i^2}\right) = 0.0258 \left(\frac{10^{31}}{10^{20}}\right) = 0.655\ V \approx 0.66\ V$$

Substituting these values into Equation 80, we obtain

$$V_{pt} = \frac{1.6 \times 10^{-19} \times 10^{16} \times (1.1\ x\ 10^{16})}{2 \times 11.9 \times 8.84 \times 10^{-14} \times 10^{15}} (2 \times 10^{-4} - 0.227 \times 10^{-4})^2 - 0.66$$

$$= 2.63.33 - 0.66 \approx 262.7\ V$$

Collector- Base Breakdown Voltage, BV_{CBO}

In the common-base mode, the maximum voltage that can be applied on the collector is limited by the breakdown characteristics of the collector-base junction. Recall our discussion on avalanche breakdown. The breakdown voltage of the collector-base junction is calculated the same way as the p-n junction breakdown voltage. That is, the collector-base junction voltage is determined by the applied voltage needed to attain critical electrical field in the depletion region, and is therefore a function of the dopant impurity concentrations in the base and collector regions.

BV_{CBO} represents the maximum collector voltage that can be applied in the common-base configuration unless punch-through occurs earlier, i.e., at a lower collector voltage. BV_{CBO} can be calculated by using the expression for the breakdown voltage for the p-n junction in terms of the critical electric field, $\mathcal{E}_{crit,}$ discusse in Chapter 3.

$$BV_{CBO} = \frac{\mathcal{E}_{crit}{}^2\ \epsilon_s}{2q} \left(\frac{1}{N_{DC}} + \frac{1}{N_{AB}}\right) - V_{biC} \tag{4.81}$$

In the avalanche process, the carriers entering the depletion region get multiplied due to impact ionization in the depletion region. The multiplication factor M is given by

$$M = \frac{1}{1-\left(\frac{V}{V_{br}}\right)^{\eta}} \tag{4.82}$$

where V is the voltage applied across the junction, V_{br} is the breakdown voltage of the junction, and η is an empirical parameter. By putting $V = V_{br}$, we can see that the multiplication factor M increases to infinity at breakdown. The symbol that is used to denote the common-base collector breakdown voltage is BV_{CBO}. The 0 is in the subscript implies that the emitter is open. Equation (4.82) can be rewritten in terms of BV_{CBO} as

$$M = \frac{1}{1-\left(\frac{V_{CB}}{BV_{CBO}}\right)^{\eta}} \tag{4.83}$$

where V_{CB} is the voltage across the collector-base junction.

Example

For the transistor discussed in the previous example, let us calculate BV_{CBO} given that $\mathcal{E}_{crit} = 200{,}000\ \mathrm{V/cm}$. Substituting the values for the various quantities, we have

$$BV_{CBO} = \frac{\left(2\times 10^{5}\right)^{2}\times 11.9\times 8.84\times 10^{-14}}{2\times 1.6\times 10^{-19}}\left(\frac{1}{10^{15}}+\frac{1}{10^{16}}\right) - 0.66$$

$$= 144.6 - 0.66 \approx 143.9\ V$$

Collector-Emitter Breakdown Voltage, BV_{CEO}

In the common-emitter mode of operation, recall that the base current is given by

$$I_B = (1-\alpha_0)\, I_E - I_{CBO} \tag{4.84}$$

When the base is open, I_B is 0, and hence

$$I_E = \frac{I_{CBO}}{(1-\alpha_0)} \tag{4.85}$$

This is the leakage current that flows between the collector and the emitter when the base lead is open. However, at larger collector voltages, the collector multiplication factor arises, and hence the current through the collector-base junction increases by a factor M, i.e., it gets multiplied by M.

$$I_C = M(\alpha_0 I_E + I_{CBO}) \tag{4.86}$$

When the base is open, $I_C = I_E$. Hence

$$I_E = \frac{M\, I_{CBO}}{1-\alpha_0 M} \tag{4.87}$$

As the collector-emitter voltage is increased, M becomes larger than 1 (as seen from Equation (4.82)), and hence I_E increases. Ultimately when $\alpha_0 M = 1$, I_E becomes infinite, and breakdown occurs. The

collector voltage at which I_E becomes infinite is denoted BV_{CEO}. The notation in the subscript is similar to what we had for BV_{CBO}; 0 indicates that the base is open. The condition for this breakdown to occur is given by

$$\alpha_0 M = 1 = \frac{\alpha_0}{1-\left(\frac{V_{Cbr}}{BV_{CBO}}\right)^{\eta}} = \frac{\alpha_0}{1-\left(\frac{BV_{CEO}}{BV_{CBO}}\right)^{\eta}} \tag{4.88}$$

In writing the above equation we have assumed that at breakdown, the collector to emitter voltage BV_{CEO}, is the same as the collector to base voltage V_{Cbr}. Rearranging terms, we obtain

$$BV_{CEO} = BV_{CBO}(1-\alpha_0)^{\frac{1}{\eta}} \tag{4.89}$$

Since $\beta_0 \approx \frac{1}{1-\alpha_0}$, we obtain

$$BV_{CEO} \approx \frac{BV_{CBO}}{\beta_0^{\frac{1}{\eta}}} = \frac{BV_{CBO}}{\sqrt[\eta]{\beta_0}} \tag{4.90}$$

For silicon, it is reported that η has a value between 2 and 6.

Example

Let us calculate BV_{CEO} by assuming that $\eta = 3$ and $\beta_0 = 125$ for the transistor used in the previous example.

$$BV_{CEO} = \frac{BV_{CBO}}{\sqrt[\eta]{\beta_0}}$$

From our last example, BV_{CBO} is calculated to be 143.8 V

$$\sqrt[\eta]{\beta_0} = \sqrt[3]{125} = 5$$

Hence

$$BV_{CEO} = \frac{BV_{CBO}}{5} = \frac{143.9}{5} = 28.8\ V$$

Thus we see that BV_{CEO} is much smaller than BV_{CBO}. In the common-emitter mode BV_{CEO} imposes the limit on the collector voltage unless punch-through occurs earlier.

Base Stored Charge

The minority carriers injected in the base are diffusing towards the collector base junction. At any instant of time, the minority carrier density distribution in the base region looks like what is shown in Figure (4.21A). This distribution is stationary and does not change with time, although individual electrons (minority carriers) are constantly moving. It is as though so many minority carriers are stored in the base region. The area under the plot in Figure (4.21A) multiplied by qA, where A is the area of cross-section of the base region is called the stored minority carrier charge. The minority carrier charge stores in the base is given by

$$Q_B = -q\,A\int_0^{W'}\Delta n(x')dx' \tag{4.91}$$

To maintain charge neutrality, an excess majority carrier density ($\Delta p(x')$) equal to $\Delta n(x')$ also arises. The distribution of excess majority carriers is identical to that of excess minority carrier distribution, and is shown in Figure (4.21B). The stored majority carrier charge is equal in magnitude but opposite in sign to Q_B. (Recall our discussion in chapter 3.)

In a well-designed transistor, most of the minority carriers emitted from the emitter reach the collector. Very few minority carriers are lost due to recombination in the base. Hence the distribution is almost linear. We can therefore define a term τ_B, called the ***Base-Transit time***. The base-transit time denotes the average time taken by the injected minority carrier to traverse the neutral base region, and reach the collector. The current in the base region due to the injected electrons is given by

$$I_{Cn}(x') = -q\,A\,v(x')\Delta n(x') \tag{4.92}$$

Where $v(x')$ is the velocity of electrons as a function of x'. The velocity of the minority carrier at some point x' in the neutral base region is

$$v(x') = \frac{-I_{Cn}(x')}{q\,A\,\Delta n(x')} \tag{4.93}$$

The time taken to traverse a distance dx' between x' and $x'+dx'$ is

$$dt = \frac{dx'}{v(x')} = -\frac{q\,A\,\Delta n(x')}{I_{Cn}(x')}dx' \tag{4.94}$$

The transit time is obtained by integrating dt, and is equal to

$$\tau_B = \int dt = \int_0^{W'}\frac{-q\,A\,\Delta n(x')}{I_{Cn}(x')}dx' \tag{4.95}$$

Since we assume that there is negligible recombination in the base region, the current is constant in the base, and equal to I_{Cn}. Hence, we can take $I_{Cn}(x')$ outside the integral, and obtain

$$\tau_B = \frac{1}{I_{Cn}}\int_0^{W'} -q\,A\,\Delta n(x')\,dx' = \frac{Q_B}{I_{Cn}} \tag{4.96}$$

$$Q_B = \tau_B\,I_{Cn} = \tau_B\,\alpha_0\,I_E \tag{4.97}$$

Since we assumed that there was negligible recombination in the base, the excess carrier density can be approximated to vary linearly with x', and is given by

$$\Delta n(x') = \Delta n_0 \left(1 - \frac{x'}{W'}\right) \tag{4.98}$$

$$Q_B = -\int_0^{W'} q\, A\, \Delta n(x')\, dx' = -q\, A\, \frac{\Delta n_0 W'}{2} \tag{4.99}$$

But

$$I_{Cn} = q\, A\, D_n \frac{\delta \Delta n}{\delta x'} = -q\, A\, D_n \frac{\Delta n_0}{W'} \tag{4.100}$$

$$\tau_B = \frac{Q_B}{I_{Cn}} = \frac{W'^2}{2D_n} \tag{4.101}$$

Example

Consider the base region in which the minority carrier diffusion constant is equal to 20 cm^2/sec. Assume $W' = 2$. Then the base transit time, τ_B is equal to

$$\tau_B = \frac{(2x10^{-4})^2}{2\ x\ 20} = 10^{-9}\ sec$$

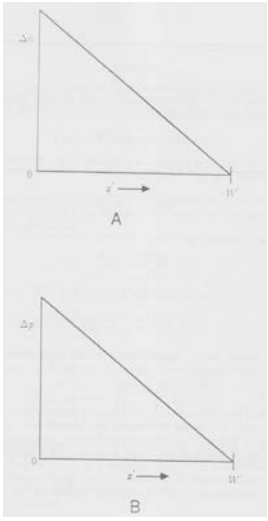


Figure (4.21): Excess carrier distribution in the base region: A) minority carriers B) majority carriers.

Although we assumed no recombination in the base region for the purpose of determining the base transit time, in reality there is some recombination and the base transport factor α_T, is less than unity. Recall that

$$\alpha_T = \frac{1}{\cosh \frac{W'}{L_n}} \tag{4.102}$$

If we now expand the denominator under the assumption $W' \ll L_n$, we obtain

$$\alpha_T \approx \frac{1}{1+\frac{W'^2}{2L_n{}^2}} \approx 1 - \frac{W'^2}{2L_n{}^2} = 1 - \frac{W'^2}{2D_n\tau_n} \tag{4.103}$$

where τ_n is the minority carrier lifetime in the base.

$$\alpha_T \approx 1 - \frac{W'^2}{2D_n\tau_n} \approx 1 - \frac{\tau_B}{\tau_n} \tag{4.104}$$

The above equation can be used to give a physical interpretation to the base transport factor. We know that $\frac{dt}{\tau_n}$ is the probability that the minority carrier in the base will recombine in the base in a time interval dt. Hence if the base transit time τ_B is small compared with τ_n, then $\frac{\tau_B}{\tau_n}$ is the probability that a minority carrier will recombine in the base region in its transit to the collector. Hence $1 - \frac{\tau_B}{\tau_n}$ is the probability that an injected minority carrier will reach the collector without being lost by recombining. By decreasing the base-transit time τ_B, in comparison with the lifetime τ_n, α_T can be made closer to unity. When τ_B is very much smaller than τ_n, $\alpha_T \approx 1$.

Stored Charge Capacitance

The minority carriers stored in the base region give rise to capacitive effects. The Capacitor associated with minority carrier storage is called the diffusion capacitance, or storage capacitance. Consider the carrier density distributions in the base region shown in Figure (4.23A) for excess minority carriers and Figure (4.23B) for excess majority carriers when the emitter base junction is forward biased at some DC bias voltage, V_{BE}. Let us now apply a small incremental step voltage, dV in series with V_{BE} so that the net forward bias voltage is $V_{BE} + dV$ as shown in Figure (4.22). Due to dV, the excess minority carrier density is increased in the base to correspond to the new emitter-base voltage. The stored negative charge[7] in the base is increased by an amount, $dQ_B = -dQ$, where

$$dQ_B = -dQ = -qA \int_0^{W'} \partial(\Delta n)\, dx' \tag{4.105}$$

where $\partial(\Delta n)$ is the incremental change in the excess minority carrier density. The shaded region in Figure (4.23A) represents the increment in the negative charge, $-dQ$. This incremental negative charge entered the base region through the emitter from the step voltage source as shown in Figure (4.22). In order to preserve charge neutrality, the excess majority carrier in the base also increases by the same amount as the excess minority carriers. The incremental change in the positive charge due to majority carriers is equal to dQ, and is shown in Figure (4.22B) by shaded lines. The incremental positive (majority carrier) charge (dQ) flows into the base region through the base ohmic contact from the step voltage source as shown in Figure (4.22). Thus we see that due to the application of the step voltage (dV), a charge $+ dQ$ flows into the base through the base contact, and a charge $- dQ$ flows into the

[7] Since electrons are minority carriers, the increase in minority carriers corresponds to an increase in the negative charge.

base through the emitter, both from the step voltage source. When the step voltage is reduced to zero, the incremental charges flow back to the step voltage source from the base, the same way they entered the base region. Thus the transistor behaves as though there were a capacitor between the emitter and the base lead. This capacitor which is called the ***diffusion capacitor*** (or storage capacitor) has a capacitance, $\boldsymbol{C_{diff}}$, equal to

$$C_{diff} = \frac{dQ}{dV} = \frac{d|Q_B|}{dV_{BE}} \tag{4.106}$$

We can evaluate this capacitance by writing Q_B explicitly. Recall that in a previous section, we expressed Q_B as

$$Q_B = I_{Cn}\,\tau_B = \alpha_0\, I_E\, \tau_B$$

$$C_{diff} = \frac{d|Q_B|}{dV_{BE}} = \alpha_0\,\tau_B \;\; \frac{d|I_E|}{dV_{BE}} = \alpha_0\,\tau_B \;\; \frac{q|I_E|}{kT} = \frac{\alpha_0\,\tau_B}{r_E} \tag{4.107}$$

Example

Let us calculate the diffusion capacitance of a bipolar transistor carrying an emitter current of $1\ mA$ with a transit time equal to 10^{-8} sec, and α_0 equal to 0.99. Let $T = 300\ K$.

$$r_E = \frac{kT}{q\ I_E} = \frac{0.0259\ V}{1\ mA} = 25.9\ \Omega$$

$$C_{diff} = \frac{\alpha_0\,\tau_B}{r_E} = \frac{0.99 \times 10^{-8}}{25.9} = 3.82 \times 10^{-10}\ F$$

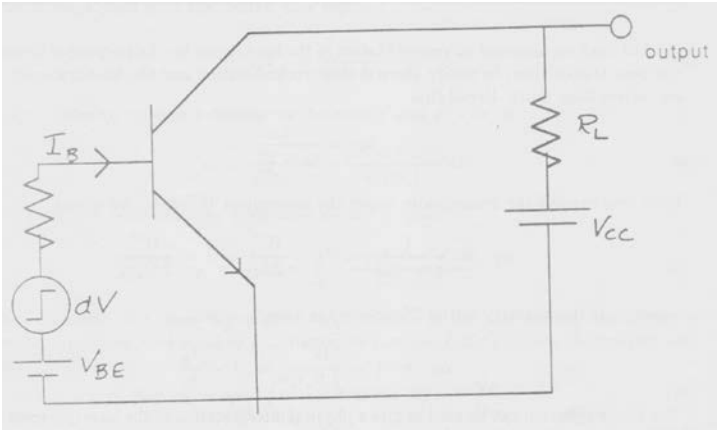


Figure (4.22): Application of a step voltage dV, in series with V_{BE} so as to increase the injected minority carriers in the base.

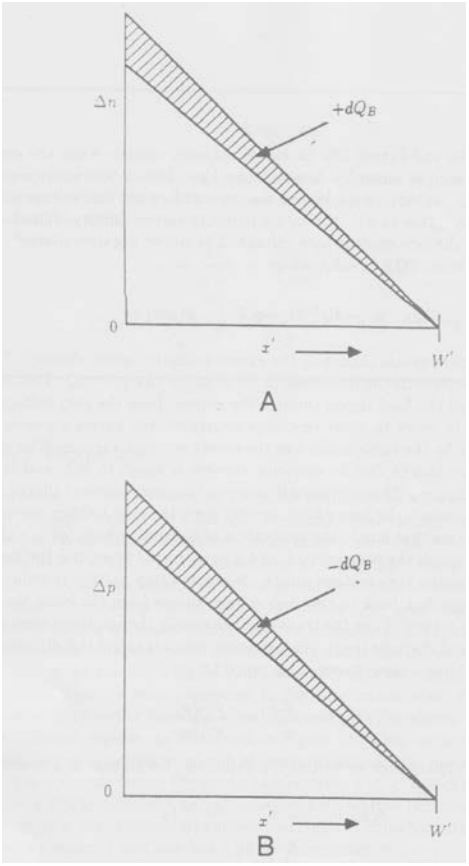


Figure (4.23): Incremental change in excess carrier densities A) minority carriers. The increase in stored minority carrier charge comes about due to $+\ dQ_B$ flowing from the emitter into the base. B) majority carries. The increase in stored majority carrier charge comes about due to $-\ dQ_B$ flowing into the base region from the base ohmic contact.

Frequency Response

We consider earlier the response of a bipolar transistor to a sinusoidally varying small signal input voltage. We had assumed that the small signal voltage varied sinusoidally at a very low frequency, so that the injected carrier density varied in step with the small signal input emitter base voltage. This is illustrated in Figure (4.14). When the sinusoidal input voltage is at its maximum value, the excess carrier density is increased everywhere in the base, as indicated by the top line in Figure (4.14). When $\tilde{v}_s$ is at its most negative value, the excess carrier density is decreased everywhere, as indicated by the bottom line in Figure (4.14). The middle line represents the excess carrier density due to the DC emitter-base voltage. On the other hand, when the input signal varies at a very high frequency such that the period of the sinusoidal signal is small compared to the base-transit time, τ_B (i.e. $\frac{1}{\omega} < \tau_B$) the excess carrier density in the base has a spatial variation. The excess carrier density varies in phase in different regions,

as illustrated in Figure (4.24). This variation arises because the excess carrier density injected at $x' = 0$ at some instant of time has not had enough time to travel to the collector before the excess carrier density at $x' = 0$ changes to a new value. Thus there is a phase difference and attenuation between the (sinusoidal) current injected by the emitter and the (sinusoidal) current reaching the collector. Hence the current gain decreases with frequency and also has a phase dependence.

We can now represent the high frequency response of the transistor in terms of an equivalent circuit. Let us initially consider the frequency response of the device, neglecting all parasitic capacitances and resistances. We consider the junction (depletion) capacitance as a parasitic element, and similarly the bulk resistances of the collector, base, and contact resistances[8] are considered parasitic elements. On the other hand, the diffusion capacitance, C_{diff}, is not neglected and is included as part of the transistor. The diffusion capacitance arises due to the injection of minority carriers in the base, which is vital to the transistor action. A transistor in which the parasitic elements are neglected, is called an ***intrinsic transistor***. In the equivalent circuit for an intrinsic transistor, we included only those elements that are necessary for the transistor action.

Common Base Mode

Let us now consider the common-base mode. The equivalent circuit for the intrinsic transistor in the common-base mode is shown in Figure (4.25). In this circuit, the diffusion capacitance is shown to be connected parallel to r_E (the emitter resistance). The reason for placing the diffusion capacitance in parallel with the input emitter resistance r_E is that the input voltage supplies both the incremental stored charge and the injected current. The collector current is given as before by the current source ($g_m \tilde{v}_{BE}$) and is in phase with $\tilde{v}_{BE}$; the collector current is in phase with the current that flows through r_E. The emitter current flows partially through C_{diff} and partially through r_E. There is therefore a phase difference between the emitter current and the current through r_E. Hence the emitter current and the collector current are not in phase. As the frequency of the input signal is increased, comparatively more and more of the emitter current flows through C_{diff}, and hence the collector current decreases. Thus the current gain decreases.

Let us assume that a sinusoidal input current of $\tilde{\imath}_e = i_e \sin \omega t$ is applied at the input terminals. Then the base-emitter voltage $\tilde{v}_{BE}$ is given by

$$\tilde{v}_{BE} = i_e \frac{r_E \times \left(\frac{1}{j\omega C_{diff}}\right)}{r_E + \frac{1}{j\omega C_{diff}}} = \frac{i_e r_E}{1 + j\omega C_{diff} r_E} \tag{4.108}$$

The collector current us given by

$$i_c = g_m \tilde{v}_{BE} = \frac{g_m r_E}{1 + j\omega C_{diff} r_E} i_e \tag{4.109}$$

[8] The emitter bulk resistance is negligible because of the heavy doping

Hence the current gain α, which is a function of frequency, is given by

$$\alpha(\omega) = \frac{i_c}{i_e} = \frac{g_m r_E}{1+ j\omega\, C_{diff}\, r_E} \tag{4.110}$$

At very low frequency ($\omega \to 0$), $\alpha(\omega)$ approaches α_0, the low frequency gain. By putting $\omega = 0$ in the equation for $\alpha(\omega)$, we obtain

$$\alpha(\omega = 0) = \alpha_0 = g_m r_E \tag{4.111}$$

Hence

$$\alpha(\omega) = \frac{\alpha_0}{1+ j\omega\, C_{diff}\, r_E} \tag{4.112}$$

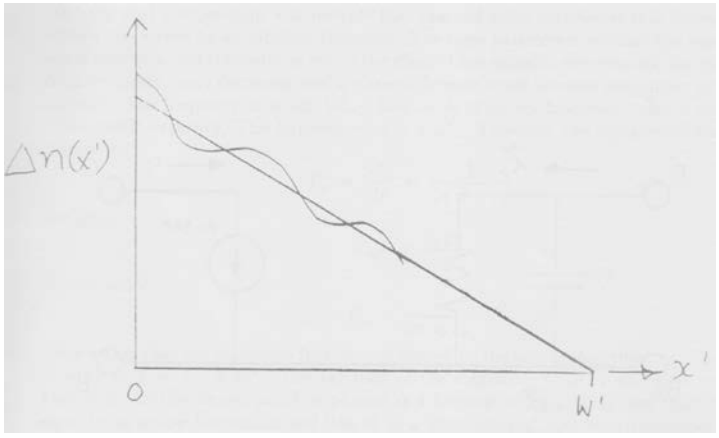


Figure (4.24): Spatial variation in the base of the excess minority carrier charge density due to a high frequency input signal.

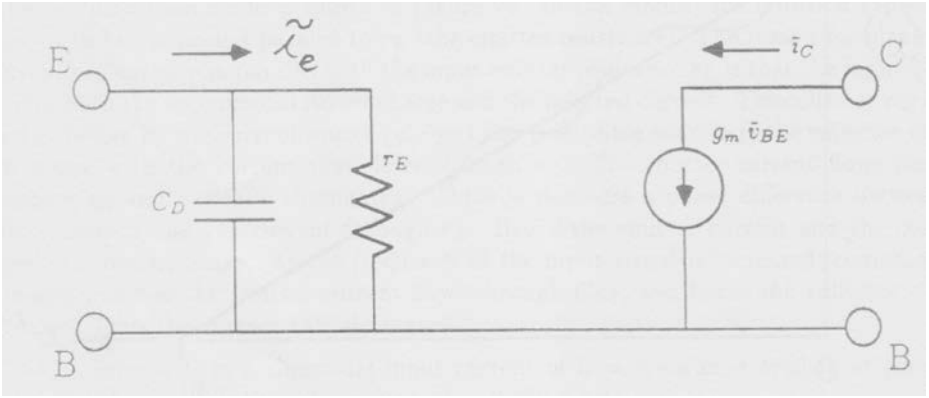


Figure (4.25): High Frequency equivalent circuit of the intrinsic transistor in the common base mode

Let us now define a parameter, $\omega^*_{c\alpha}$, as

$$\omega^*_{c\alpha} \equiv \frac{1}{C_{diff}\, r_E} \tag{4.113}$$

Then

$$\alpha(\omega) = \frac{\alpha_0}{1+ j\frac{\omega}{\omega^*_{c\alpha}}} \tag{4.114}$$

We have used a superscript $*$ to indicate that this and other parameters that follow with a superscript $*$ refer to an intrinsic transistor. The same parameters without the superscript $*$ will refer to a real transistor in which the effect of the parasitic elements are also included. As ω increases, $\alpha(\omega)$ decreases, and a phase difference arises between the output and input currents. The frequency at which $|\alpha(\omega)|$ falls to $\frac{1}{\sqrt{2}}$ of its low frequency value is called the ***alpha-cutoff frequency***. This happens when $\omega = \omega^*_{c\alpha}$. Therefore, the alpha-cutoff frequency is given by

$$f^*_{c\alpha} = \frac{\omega^*_{c\alpha}}{2\pi} = \frac{1}{2\pi\, C_{diff}\, r_E} \tag{4.115}$$

Recalling

$$C_{diff} = \frac{\alpha_0 \tau_B}{r_E}$$

We can write

$$f^*_{c\alpha} = \frac{1}{2\pi\, \alpha_0 \tau_B} \tag{4.116}$$

Thus we see that the maximum frequency is limited by the base-transit time, τ_B. The phase of $\alpha(\omega)$ at $f = f_{c\alpha}$ is 45^0. The variation of the magnitude of $\alpha(\omega)$ with ω is shown in Figure (4.26). In this figure, $|\alpha(\omega)|$ is plotted as a function of $\log \omega$. It is seen that $|\alpha(\omega)|$ is equal to α_0 at low frequencies and falls off as ω approaches $\omega^*_{c\alpha}$. At high frequencies,

$$\alpha(\omega) \approx \frac{\alpha_0}{j\frac{\omega}{\omega^*_{c\alpha}}} = \frac{\alpha_0 \omega^*_{c\alpha}}{j\omega} \tag{4.117}$$

The gain falls off by a factor of 2 when the frequency (or ω) is doubled. In this region the gain is said to fall off at the rate of 6 db/octave.

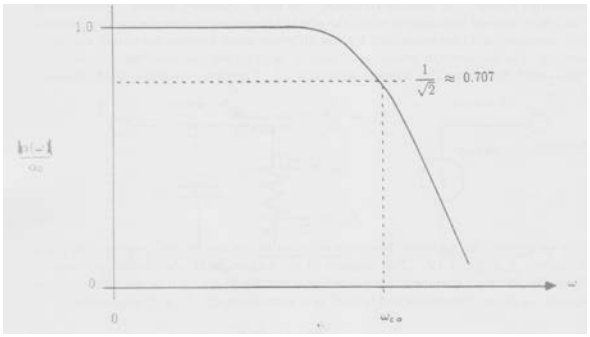


Figure (4.26): Plot of the magnitude of the current gain, $\alpha(\omega)$, as a function of ω. $\omega_{c\alpha}$ is the value of ω at which $|\alpha(\omega)|$ is equal to 0.707 of α_0.

Example

Let us now determine the alpha cut-off frequency of an intrinsic n-p-n bipolar transistor, given:

$$W' = 2\ \mu m$$

$$D_n = 15\ cm^2/sec$$

$$\alpha_0 = 0.995$$

$$\tau_B = \frac{W'^2}{2D_n} = \frac{\left(1\times 10^{-4}\right)^2}{2 \times 15} = 1.33 \text{ x } 10^{-9}\ sec$$

$$f_{c\alpha}^{*} = \frac{1}{2\pi\ \alpha_0 \tau_B}$$

Therefore,

$$f_{c\alpha}^{*} = \frac{1}{2\pi \times 0.995\ \times 1.33\ \times\ 10^{-9}} = \frac{1}{8.34 \times\ 10^{-9}} = 1.2\ \times\ 10^{8}\ Hz$$

Common Emitter Mode

Let us now treat the frequency response of the device in the common emitter mode. The equivalent circuit of the intrinsic transistor is the same as what we derived earlier for the low frequency case, except that the input resistance, r_{in}, which is in parallel with the diffusion capacitor (C_{diff}), is equal to $\frac{r_E}{1-\ \alpha_0}$. The equivalent circuit for the intrinsic transistor in the common emitter mode is shown in Figure (4.27). The input current (which is now the base current $\tilde{\imath}_b$) divides itself between two paths: one through r_{in} , and the other through $C_{diff.}$ The collector current ($\tilde{\imath}_c$) is determined by the current that flows through r_{in}, and therefore there is a phase difference between $\tilde{\imath}_c$ and $\tilde{\imath}_b$. The base-emitter voltage ($\tilde{v}_{BE}$) is given by

$$\tilde{v}_{BE} = \tilde{\imath}_b \frac{\left(\frac{r_E}{1-\ \alpha_0}\right)\frac{1}{j\omega C_{diff}}}{\frac{r_E}{1-\ \alpha_0} + \frac{1}{j\omega C_{diff}}} = \frac{\frac{r_E}{1-\ \alpha_0}}{1+\ j\omega\ C_{diff}\frac{r_E}{1-\ \alpha_0}}\ \tilde{\imath}_b \qquad (4.118)$$

The collector current is obtained as

$$\tilde{\imath}_c = g_m \tilde{v}_{BE} = \frac{g_m \frac{r_E}{1-\ \alpha_0}}{1+\ j\omega\ C_{diff}\frac{r_E}{1-\ \alpha_0}}\ \tilde{\imath}_b \qquad (4.119)$$

The current gain is now a function of frequency, and is given by

$$\beta(\omega) = \frac{\tilde{\imath}_c}{\tilde{\imath}_b} = \frac{g_m \frac{r_E}{1- \alpha_0}}{1+ j\omega\, C_{diff} \frac{r_E}{1- \alpha_0}} \tag{4.120}$$

At very low frequency, (i.e. $\omega \to 0$), $\beta(\omega)$ is obtained by putting $\omega = 0$, and is equal to β_0.

$$\beta(\omega)\,|_{\omega=0} = \beta_0 = g_m \frac{r_E}{1- \alpha_0} \tag{4.121}$$

Therefore $\beta(\omega)$ can be expressed in terms of β_0 as

$$\beta(\omega) = \frac{\beta_0}{1+ j\omega C_{diff} \frac{r_E}{1- \alpha_0}} \tag{4.122}$$

Let us define a parameter $\omega^*_{C\beta}$ given by

$$\omega^*_{C\beta} \equiv \frac{1}{C_{diff} \frac{r_E}{1- \alpha_0}} \tag{4.123}$$

Since

$$C_{diff} r_E = \alpha_0 \tau_B \tag{4.124}$$

$$\omega^*_{C\beta} = \frac{1}{\frac{\alpha_0 \tau_B}{1- \alpha_0}} = \frac{1}{\beta_0 \tau_B} \tag{4.125}$$

Then

$$\beta(\omega) = \frac{\beta_0}{1+ j\frac{\omega}{\omega^*_{C\beta}}} \tag{4.126}$$

The magnitude of the current gain ($|\beta(\omega)|$) is plotted in Figure (4.28) as a function of log ω. We define the beta cutoff frequency as that at which the magnitude of the gain, falls off to $\frac{1}{\sqrt{2}}$ of its low frequency value. This happens when $\omega = \omega^*_{C\beta}$. Hence beta cut-off frequency $\left(f^*_{C\beta}\right)$ is given by

$$f^*_{C\beta} = \frac{\omega^*_{C\beta}}{2\pi} = \frac{1}{2\pi\, C_{diff} \frac{r_E}{1- \alpha_0}} = \frac{1}{2\pi\, \beta_0 \tau_B} \tag{4.127}$$

The phase of the collector current differs from that of the base current by 45^0 when $f = f^*_{C\beta}$. At very high frequencies,

$$\beta(\omega) \approx \frac{\beta_0}{j\frac{\omega}{\omega^*_{C\beta}}} = j\, \frac{\beta_0 \omega^*_{C\beta}}{\omega} \tag{4.128}$$

In the range of frequencies over which the above equation is valid, the magnitude of $|\beta(\omega)|$ multiplied by ω is a constant, given by

$$\omega|\beta(\omega)| = \beta_0 \omega^*_{C\beta} = \frac{\beta_0}{\beta_0 \tau_B} = \frac{1}{\tau_B} \tag{4.129}$$

As the frequency is increased, the gain falls by an amount such that the product of gain and frequency is a constant. This constant is called the ***gain-bandwidth product***. The decrease in the gain with frequency at high frequencies, again demonstrates a 6db/octave characteristic, i.e. $|\beta|$ decreases by a factor of 2 when the frequency is doubled. The value of ω at which $|\beta(\omega)|$ becomes unity is denoted ω_T^*, and is found in Equation (4.129) to be

$$\omega_T^* = \frac{1}{\tau_B} \tag{4.130}$$

According to Equation (4.129), the frequency at which $|\beta(\omega)|$ becomes unity is equal to

$$f_T^* = \frac{\omega_T^*}{2\,\pi} = \frac{1}{2\,\pi\,\tau_B} \tag{4.131}$$

Using Equation (4.130) we can express $\omega_{C\beta}^*$ as

$$\omega_{C\beta}^* = \frac{1}{\beta_0 \tau_B} = \frac{\omega_T^*}{\beta_0} \tag{4.132}$$

To summarize, the alpha cut-off frequency is

$$f_{c\alpha}^* = \frac{1}{2\pi}\ \frac{1}{\alpha_0}\ \frac{1}{\tau_B} \tag{4.133}$$

The beta cut-off frequency is

$$f_{C\beta}^* = \frac{1}{2\pi}\ \frac{1}{\beta_0}\ \frac{1}{\tau_B} \tag{4.134}$$

The gain bandwidth product is

$$f_T^* = \frac{1}{2\pi}\ \frac{1}{\tau_B} \tag{4.135}$$

If any two of these three frequencies are known, the third is readily obtained. Or alternately, if τ_B and either α_0 or β_0 are known, the three frequencies are readily obtained. Thus, of the five quantities, $f_{c\alpha}^*, f_{C\beta}^*, f_T^*, \tau_B$ and either α_0 or β_0, we can determine the other three if any two are known.

Example

Let us now determine the limiting frequencies $f_{C\beta}^*$ and f_T^* for the transistor discussed in the early example. We saw that $\tau_B = 1.33 \times 10^{-9}\ sec.$ We were given α_0 as equal to 0.995. Hence

$$f_T^* = \frac{1}{2\pi}\ \frac{1}{1.33 \times 10^{-9}} = 1.194 \times 10^8\ Hz$$

$$\beta_0 = \frac{\alpha_0}{1-\alpha_0} = \frac{0.995}{0.005} = 199$$

$$f_{c\beta}^* = \frac{1}{2\pi} \times \frac{1}{\beta_0} \times \frac{1}{\tau_B} = 5.998 \times 10^5\ Hz$$

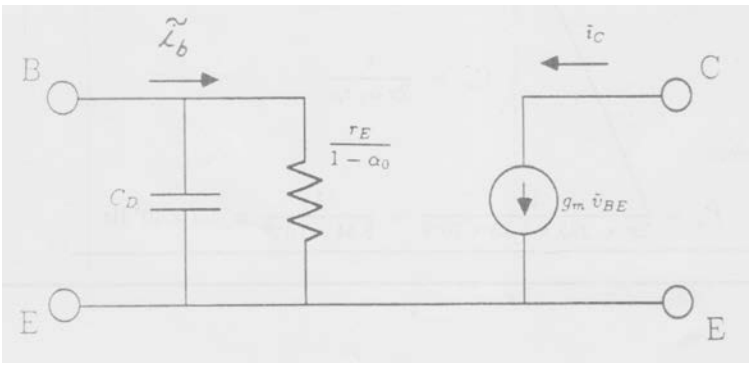


Figure (4.27): The equivalent circuit of the intrinsic transistor in the common emitter mode.

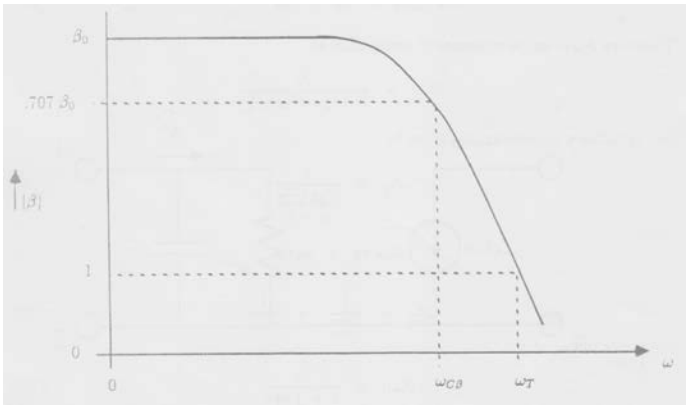


Figure (4.28): Plot of the magnitude of the current gain in the common emitter mode as a function of log $\boldsymbol{\omega}$

Inclusion of Junction Capacitances

Let us next consider the equivalent circuit for the bipolar transistor by including the emitter-base and collector-base junction capacitances. Previously, the α-cutoff frequency and the β- cutoff frequency of an intrinsic transistor were discussed. The actual α - and β- cutoff frequencies of the transistors will be different from what we obtained for the intrinsic transistor, due to the presence of the emitter-base junction capacitances. The equivalent circuits that we previously drew for the common-base mode and the common-emitter mode need to be modified to include the junction capacitances.

Common Base Mode

The equivalent circuits for the common-base mode will have the emitter-base junction capacitance in parallel with the diffusion capacitance between the emitter and the base, and the collector-base capacitance will be in parallel with the current source in the output circuit between the collector and the base, as shown in Figure (4.29). If we go through the analysis exactly the same way as we did for the intrinsic transistor, we will find that the current gain is frequency dependent, with a time constant that is determined not only by the diffusion capacitance but also by the emitter-base junction

capacitance. Without duplicating every step of the analysis, we can readily write the expression for α as a function of frequency:

$$\alpha(\omega) = \frac{\alpha_0}{1+j\omega(\alpha_0\tau_B + r_E C_{JEB})} \tag{4.136}$$

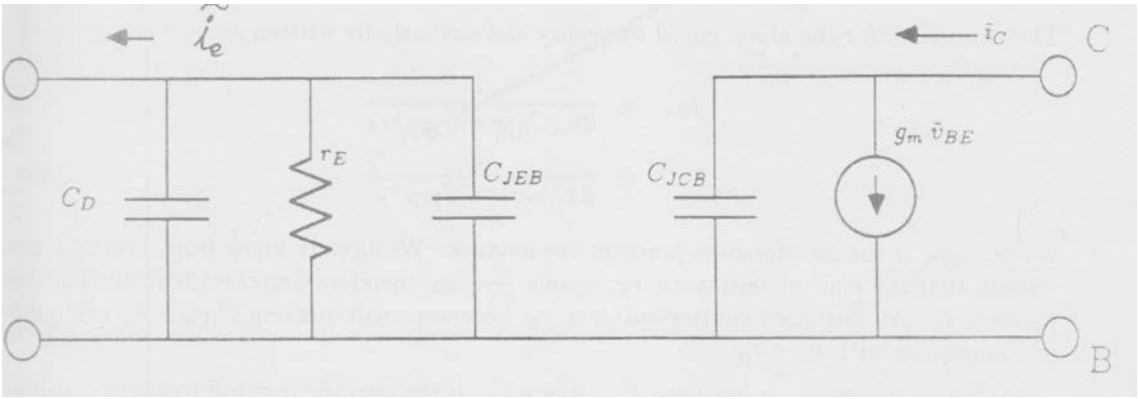


Figure (4.29): The equivalent circuit of the bipolar transistor in the common-base mode, including the emitter-base and the collector-base capacitances.

The expression differs from that for the intrinsic transistor due to the fact that the time constant of the input circuit is increased from $\alpha_0\tau_B$ to $\alpha_0\tau_B + r_E C_{JEB}$. In terms of C_{diff}, $\alpha(\omega)$ can be written as

$$\alpha(\omega) = \frac{\alpha_0}{1+j\omega r_E(C_{diff}+C_{JEB})} \tag{4.137}$$

The expression for the alpha-cutoff frequency can similarly be written as

$$f_{C\alpha} = \frac{1}{2\pi(C_{diff} + C_{JEB})r_E}$$

$$= \frac{1}{2\pi(\alpha_0\tau_B + C_{JEB}r_E)} \tag{4.138}$$

where C_{JEB} is the emitter-base junction capacitance. We already know from previous discussion that the emitter resistance, r_E equals $\frac{kT}{qI_E}$, and therefore decreases with the emitter current, I_E. At very high emitter-currents, r_E becomes small, making $C_{JEB} \times r_E$ negligible in comparison with $\alpha_{0\times}\tau_B$.

At low currents, $f_{C\alpha}$ is less than $f^*_{C\alpha}$, where $f^*_{C\alpha}$ is the intrinsic α-cutoff frequency, and at high currents, $f_{C\alpha}$ approaches $f^*_{C\alpha}$. If we measure the α- cutoff frequency at different values of the emitter current, I_E, and then plot the reciprocal of $f_{C\alpha}$ versus $\frac{1}{I_E}$, we will obtain a straight-line plot as show in Figure (4.30). If we now extrapolate the linear plot, the intercept on the vertical axis at $\frac{1}{I_E}$ = 0 will be equal to $\frac{1}{f^*_{C\alpha}}$. Thus we can obtain the intrinsic α - cutoff frequency by measuring the α - cutoff frequency of a real transistor at different emitter-current.

Common Emitter Mode

The equivalent circuit for the bipolar transistor in the common emitter mode including the junction capacitances is shown in Figure (4.31). The input capacitance is increased by having C_{JEB}, the emitter-base junction capacitance in parallel with the diffusion capacitance, C_{diff}. In the common-emitter mode, the collector-base junction capacitance appears as a feedback capacitance between the output and the input terminals. The feedback capacitance is usually referred to as a ***Miller Feedback Capacitance***.

As we discussed in the common-base mode, the time constant of the input circuit now is increased from its intrinsic value to a larger value given by

$$Input\ time\ constant\ = \frac{r_E}{(1-\alpha_0)}\left(C_{diff} + C_{JEB}\right)$$

The expression for $\beta(\omega)$ can be derived as before. We can write

$$\begin{aligned}\beta(\omega) &= \frac{\beta_0}{1+j\omega\left[\beta_0\tau_B + \frac{r_E}{(1-\alpha_0)}C_{JEB}\right]} \\ &= \frac{\beta_0}{1+\frac{j\omega}{(1-\alpha_0)}\left[\alpha_0\tau_B + r_E C_{JEB}\right]} \\ &= \frac{\beta_0}{1+\frac{j\omega}{(1-\alpha_0)}\left[r_E C_{diff} + r_E C_{JEB}\right]} \\ &= \frac{\beta_0}{1+j\omega\frac{r_E}{(1-\alpha_0)}\left[C_{diff} + C_{JEB}\right]}\end{aligned} \qquad (4.139)$$

Thus we see that the effect of including the emitter-base junction capacitance is as though the input circuit time constant is increased by $\frac{r_E}{(1-\alpha_0)}\ C_{JEB}$. We can readily write the expression for the β- cutoff frequencies without any further analysis as

$$\begin{aligned}f_{C\beta} &= \frac{1}{2\pi}\frac{1-\alpha_0}{r_E(C_{diff} + C_{JEB})} \\ &= \frac{1-\alpha_0}{2\pi(\alpha_0\tau_B + r_E C_{JEB})} \\ &= \frac{1}{2\pi\beta_0(\tau_B + \frac{1}{\alpha_0}r_E C_{JEB})}\end{aligned} \qquad (4.140)$$

The β- cutoff frequency is thus seen to be less than the intrinsic β- cutoff frequency. Again noticing that the emitter resistance, r_E, is inversely proportional to the emitter-current. We can now determine the intrinsic β - cutoff frequency by measuring the actual β - cutoff frequency at several different emitter-current values. This procedure is to measure the β - cutoff frequency at different values of the emitter-

current and plot the reciprocal of the measured β- cutoff frequency, $\frac{1}{f_{C\beta}}$, as a reciprocal function of the emitter current, $\frac{1}{I_E}$ as illustrated in Figure (4.32). A linear plot will be obtained which when extrapolated to intercept the vertical axis at $\frac{1}{I_E} = 0$ yields the reciprocal of the intrinsic β - cutoff frequency, $\frac{1}{f_{C\beta}}$.

Example

Let us calculate the α - cutoff frequency and the β - cutoff frequency of a transistor including junction capacitances. Let us consider the transistor which we used to calculate $f_{C\alpha}^*$ and $f_{C\beta}^*$ in the previous examples. Assume $I_E = 1\ mA, C_{JEB} = 340\ pF$.

r_E is calculated to be

$$r_E = \frac{kT}{qI_E} = \frac{25.9\ mV}{1\ mA} = 25.9\ \Omega$$

Using the value for β_0, $f_{C\alpha}^*$ and $f_{C\beta}^*$ in the previous examples, we calculate the following.

$$\alpha_0 = \frac{\beta_0}{\beta_0+1} = 0.995$$

$$\tau_B = 1.33\ x\ 10^{-9}\ sec$$

$$f_{C\alpha}^* = \frac{1}{2\pi \times (\alpha_0\tau_B + r_E C_{JEB})} = \frac{1}{2\pi \times (0.995 \times 1.33 \times 10^{-9} + 25.9 \times 340 \times 10^{-12})}$$

$$= 1.57\ \times\ 10^7\ Hz$$

$$f_{C\beta}^* = \frac{1}{2\pi \times (\beta_0\tau_B + \frac{r_E}{1-\alpha_0} C_{JEB})}$$

$$= \frac{1}{2\pi \times (199 \times 1.33 \times 10^{-9} + \frac{25.9}{0.005} \times 340 \times 10^{-12})} = 78.56\ KHz$$

Thus we see that both α - cutoff and β- cutoff frequencies are reduced. The junction capacitances decrease the frequency response of the transistor.

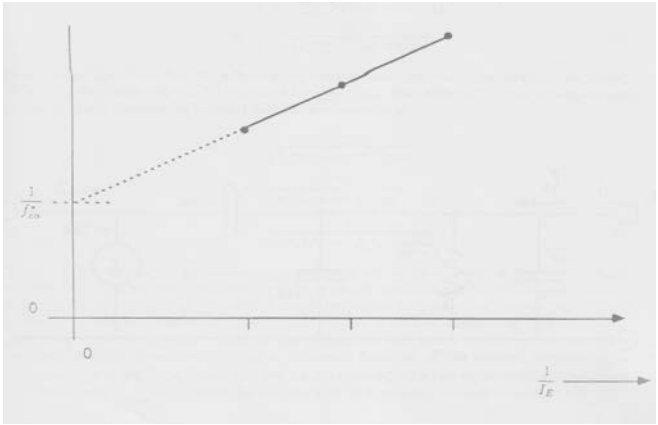


Figure (4.30): Plot of the reciprocal α - cutoff frequency, $\frac{1}{f_{C\alpha}}$, as a function of the reciprocal emitter-current, $\frac{1}{I_E}$.

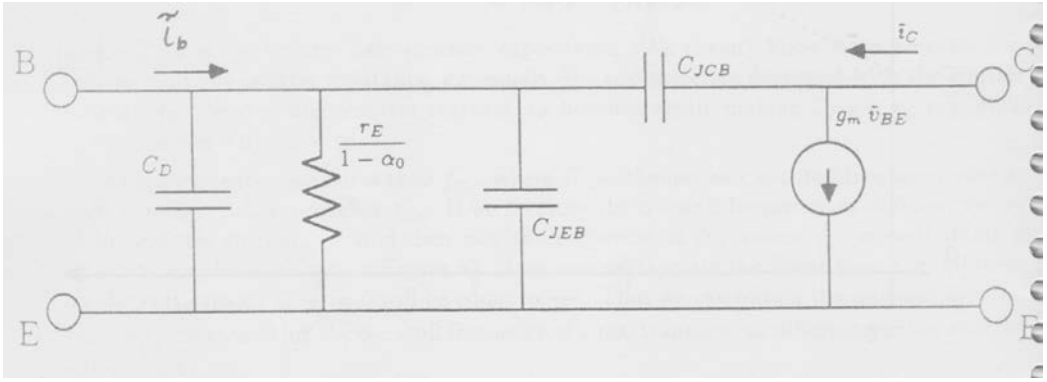


Figure (4.31): Equivalent circuit in the common emitter mode, including junction capacitance.

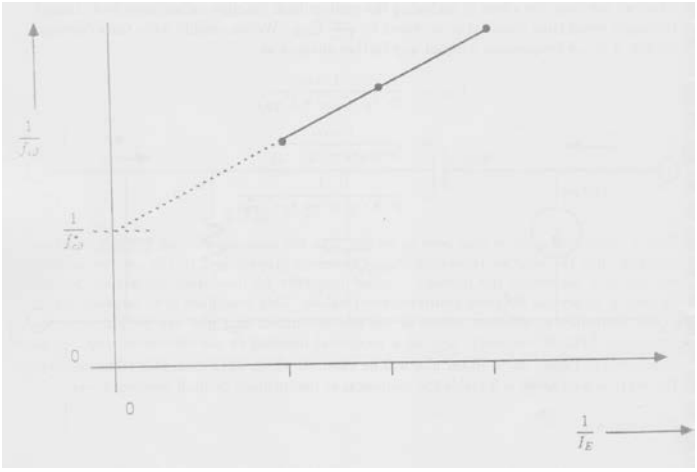


Figure (4.32): Plot of $\frac{1}{f_{C\beta}}$ as a function of $\frac{1}{I_E}$. The intercept on the vertical axis at $\frac{1}{I_E} = 0$ corresponds to $\frac{1}{f_{C\beta}^*}$.

Switching Transistors

A bipolar transistor can also be used as an inverter or as a switch. In digital applications, a bipolar transistor is used in either of these two functions. A bipolar transistor is typically usually used as what is called a ***saturated inverter***, or ***a switch***. The transistor is either off or on: when it is on, the transistor is in the saturation mode. When it is off, it is in the cutoff mode. Let us consider the circuit shown in Figure (4.33A). The transistor has a load resistance (R_L) in series with the collector-supply voltage (V_{CC}) connected between the collector and the emitter. The base current is applied to the base terminal as shown in this figure. Denoting the collector-emitter voltage as the output voltage V_o,

$$V_o = V_{CC} - R_L I_C$$

As I_C increases, V_o decreases. Finally when

$$I_C = \frac{V_{CC}}{R_L}$$

V_o is zero. When the device is off, I_C is nearly zero and $V_o = V_{CC}$. Thus when the input current goes up from 0 to I_B, the output voltage (which is the collector-emitter voltage) goes down from V_{CC} to 0. Hence the output voltage swing is opposite in phase to the input, and this circuit is called an ***inverter circuit***. *We assume that the transistor is initially off (i.e., the emitter-base voltage is reverse-biased). At time $t = 0$, a base current (I_B) is applied. The base current is chosen to be large enough so that the device under steady-state condition will be driven into saturation*.

Due to the base current, minority carriers are injected into the base from the emitter and the stored charge (Q_B) in the base increases. The collector current which is equal to $\frac{Q_B}{\tau_B}$, increases. As the current increases in the collector circuit, the collector to the base voltage which was initially at a voltage nearly equal to the collector supply voltage starts to decrease. Ultimately, when sufficiently large collector current flows, the collector voltage falls below the base emitter voltage. The collector base junction becomes forward-biased, driving the transistor into saturation. When the collector to emitter voltage becomes 0, no further increase in collector current can arise. Thus the collector current reaches a maximum value equal to $\frac{V_{CC}}{R_L}$. However, the base current is still flowing, and the stored charge in the base continues to increase. The stored charge in the base increases even though the collector current has reached its maximum value and stopped increasing.

Since we are discussing an n-p-n transistor, the injected (and stored) minority carriers are electrons. The base current provides the majority carriers (holes) needed to neutralize the stored minority carrier charge, and to provide the hole current for: a) injecting holes into the emitter, b) providing majority carriers to recombine with the minority carriers in the base, and c) the flow of the collector-base junction leakage current, I_{CBO}. We will first assume that the hole injection current into the emitter (I_{Ep}) is negligible. This assumption is tantamount to letting the emitter injection efficiency (γ) to be very nearly unity. Let us further assume that I_{CBO} is negligibly small in comparison with the base current bias I_B. The base current under these assumptions is given by

$$I_B = \frac{d|Q_B|}{dt} + \frac{Q_B}{\tau_n} \tag{4.141}$$

This equation is called the ***charge-control equation***. The first term on the right hand side denotes the base current needed for the buildup of majority carrier charge to neutralize the growing Q_B and the second term on the right hand side denotes the hole current needed for the recombination.

The boundary condition for solving the charge-control equation, when I_B is turned on at time $t = 0$, is

$$Q_B = 0 \quad \text{for } t < t_0 \tag{4.142}$$

The solution is then easily seen to be

$$|Q_B| = I_B\tau_n \left(1 - e^{-\frac{t}{\tau_n}}\right) \tag{4.143}$$

A plot of the growth of Q_B as a function of time is shown in Figure (4.34). The stored charge is seen to grow exponentially with time. When $t > \tau_n$.

$$|Q_B| = I_B \tau_n \tag{4.144}$$

As long as the transistor is in the active mode (i.e., the collector-base junction is reverse biased), the collector current (I_C) and Q_B are related to each other as

$$Q_B = I_C \tau_B \tag{4.145}$$

Initially as Q_B grows, I_C grows at the same rate, and the collector-base voltage, V_{CB}decreases. When V_{CB} becomes zero, the transistor is at the edge of saturation. When V_{CB} becomes negative, the device is in saturation and the collector voltage is nearly equal to the emitter voltage. I_C is at its maximum value since I_C cannot increase beyond $\left(\frac{V_{CC}}{R_L}\right)$. Let us assume that at $t = t_1$, the device gets saturated and that I_C becomes equal to $\frac{V_{CC}}{R_L}$. The plot of the growth of I_C with time is shown in Figure (4.35). I_C grows exponentially with time as long as $t < t_1$. At $t = t_1$, I_C reaches the maximum value $\left(\frac{V_{CC}}{R_L}\right)$, and stays constant at this value. We can think of it as though I_C and Q_B becomes decoupled when $t > t_1$. While Q_B continues to grow exponentially with time even for $t > t_1$, I_C remains constant at $\frac{V_{CC}}{R_L}$ for $t > t_1$.

Let us define $Q_B$at $t = t_1$ as Q_s. Therefore when $Q_B > Q_S$, the device is saturated.

$$Q_s = Q_B(t_1) = I_B \tau_n \left(1 - e^{-\frac{t_1}{\tau_n}}\right) \tag{4.146}$$

$$I_C(t_1) = \frac{V_{CC}}{R_L} \tag{4.147}$$

But

$$Q_S = I_C(t_1)\,\tau_B = \frac{V_{CC}}{R_L}\,\tau_B \tag{4.148}$$

Therefore

$$\frac{V_{CC}}{R_L}\,\tau_B = I_B\,\tau_n \left(1 - e^{-\frac{t_1}{\tau_n}}\right) \tag{4.149}$$

From this equation, we can obtain t_1 as

$$t_1 = \tau_n \ln\left(\frac{1}{1-\frac{V_{CC}\,\tau_B}{R_L\,I_B\,\tau_n}}\right) \tag{4.150}$$

t_1 represents the time taken by the device to reach steady state collector current, and is called the ***turn-on time***. To keep t_1 small so as to obtain a high clock frequency in digital applications, τ_n should be kept

small and $\frac{V_{CC}\tau_B}{R_L I_B}$ should be kept small compared to τ_n. The latter can be achieved by keeping τ_B small and increasing I_B. Beyond $t = t_1$, Q_B continues to grow while I_C remains constant. I_C is determined by the gradient of $\Delta n(x')$ in the base. In Figure (4.36), $\Delta n(x')$ is plotted in the base region at several instants of time. When $t < t_1$, the slope of $\Delta n(x')$ plot increases with time indicates an increase of I_C with time. When $t \geq t_1$, I_C stays constant which means the slope has to remain constant. However, the growth of Q_B occurs now in a manner such that $\Delta n(x')$ increase while the slope remains constant. $\Delta n(x' = W')$ increases from zero when $t > t_1$, since the collector is also injecting minority carriers into the base. Ultimately $\Delta n(x' = W')$ reaches a steady state value at $t = \infty$, determined by the requirement that

$$Q_B(\infty) = I_B \tau_n$$

The area under the plot of $\Delta n(x')$ is described by a right-angled triangle for $t < t_1$ and by a trapezoid for $t > t_1$, Q_S is equal to q times the area of the triangle at $t = t_1$. If $\Delta n(W')$ is the excess carrier density at $x' = W'$ for $t > t_1$, then the area of the trapezoid A_{trap} is equal to

$$\frac{Q_B(\infty)}{qA} = A_{trap} = \Delta n(W')W' + \frac{Q_S}{qA}$$

What we have discussed so far is called the turn-on behavior. Let us now discuss the device behavior, when the transistor which has been on in the saturation mode for a long time with a large base current drive, is switched off by turning off I_B to zero at time $t = 0$. The transistor current will decrease to zero only after some time delay, and this behavior is called the ***turn-off transient***.

Let us make the assumption that a base current I_B has been flowing for a long time and that I_B is sufficiently large to drive the transistor into saturation. The base current as a function of time is shown in Figure (4.37). The charge control equation under this assumption becomes $t = 0$.

$$\frac{d|Q_B|}{dt} = -\frac{|Q_B|}{\tau_n} \tag{4.151}$$

The boundary condition is that at $t = 0$,

$$|Q_B(t = 0)| = I_B \tau_n \tag{4.152}$$

The solution can be readily seen to be

$$|Q_B| = I_B \tau_n e^{-\frac{t}{\tau_n}} \tag{4.153}$$

The stored charge decays exponentially with time, as illustrated in Figure (4.38). The collector current continues to remain constant at the value it had before the base current was turned off. This situation continues until the device gets out of saturation. The device gets out of saturation when Q_B becomes equal to or less than Q_S. We define the time that it takes for the device to get out of saturation as the storage time and denote it by the symbol $t_{s.}$ Therefore by our definition, at $t = t_{s,}$ the stored charge is equal to Q_S. At $t = t_{s,}$

$$Q_B(t_{s.}) = Q_S = I_B \tau_n e^{-\frac{t_S}{\tau_n}} \tag{4.154}$$

Therefore, t_s is obtained as

$$t_s = \tau_n \ln\left(\frac{I_B\tau_n}{Q_S}\right) = \tau_n \ln\left(\frac{I_B\tau_n}{\frac{V_{CC}\tau_B}{R_L}}\right) \quad (4.155)$$

When $t > t_s$, the device is in the active mode and I_C is proportional to Q_B. Hence I_C decays with time in step with Q_B. This is illustrated in Figure (4.39). The decay of I_C is given by

$$I_C = \frac{Q_S}{\tau_B} e^{-\frac{t-t_S}{\tau_n}} \quad (4.156)$$

To decay to $\frac{1}{e}$ of the initial value of the collector current, $t - t_{s,}$ should be equal to τ_n. The total time taken for the collector current to decay to $\frac{1}{e}$ of its initial value from the instant of time when I_B is reduced to zero is

$$t = t_s + \tau_n \quad (4.157)$$

To speed up this process, t_s should be decreased. This is equivalent to saying that the device should not be operated in deep saturation.

Example

Let us calculate the switching characteristics of an n-p-n transistor used as a saturated inverter, given $R_L = 2\ K\Omega, V_{CC} = 10\ V, I_B = 150\ \mu A,\ \beta_0 = 100$, and $\tau_B = 10^{-9}\ sec$. Let us further assume that the emitter injection efficiency is nearly unity, so that we can assume $\beta_0 = \frac{\tau_n}{\tau_B}$.

<u>Turn-on</u>

$$I_{cmax} = \frac{V_{CC}}{R_L} = \frac{10}{2 \times 10^3} = 5\ mA$$

$$Q_S = I_{cmax}\tau_B = 5 \times 10^{-3} \times 10^{-9} = 5 \times 10^{-12} C$$

$$\tau_n = \beta_0\tau_B = 100 \times 10^{-9} = 10^{-7} sec$$

$$Q_B(\infty) = I_B\tau_n = 150 \times 10^{-6} \times 10^{-7} = 1.5 \times 10^{-11} C$$

$$t_1 = -\tau_n \ln\left[1 - \frac{V_{CC}\tau_B}{R_L I_B \tau_n}\right]$$

$$= -10^{-7} \times \ln\left[1 - \frac{5 \times 10^{-12}}{1.5 \times 10^{-11}}\right] = 10^{-7} \times 0.405$$

$$= 4.05 \times 10^{-8}\, sec$$

Turn-off

$$t_s = -\tau_n \ln\left[\frac{V_{CC}\tau_B}{R_L I_B \tau_n}\right] = -10^{-7} \times \ln\left[\frac{5\ x\ 10^{-12}}{1.5\ x\ 10^{-11}}\right]$$

$$= 1.1 \times 10^{-7}\, sec$$

Total time taken for I_C to fall off to $\frac{1}{e}$ of its initial value is equal to

$$t_s + \tau_n = 10^{-7} + 1.1\ x\ 10^{-7} = 2.1 \times 10^{-7}\ \text{sec}$$

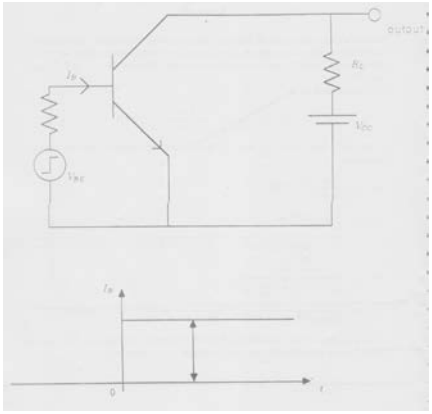


Figure (4.33): A) A circuit configuration for using a bipolar transistor as a saturated inverter. B) The base current as a function of time.

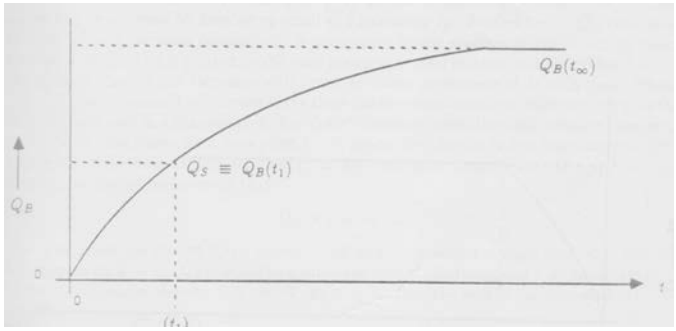


Figure (4.34): The growth of the stored charge as a function of time, due to a constant base current drive, I_B.

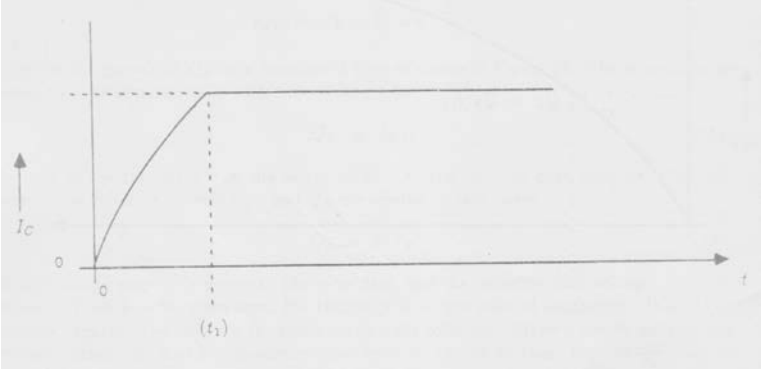


Figure (4.35): The increase in collector current as a function of time due to the base current drive I_B. I_B is assumed to be large enough to drive the device into saturation. The device gets saturated at $t = t_1$,

For $t < t_1$, $I_C = \frac{Q_B(t)}{\tau_B}$.

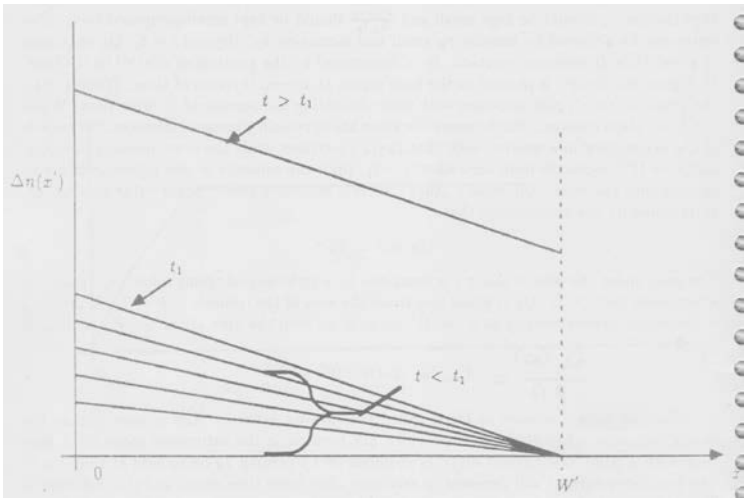


Figure (4.36): The growth of excess minority carrier density in the base in a saturated inverter.

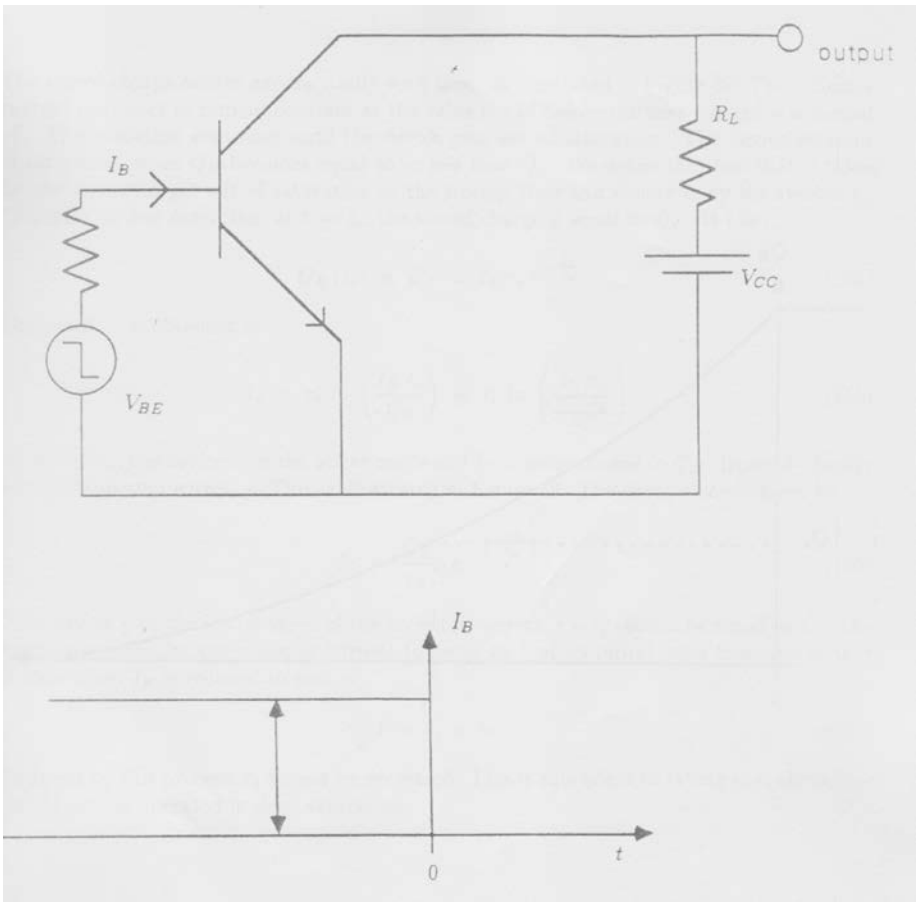


Figure (4.37): The base current during the turn-off phase.

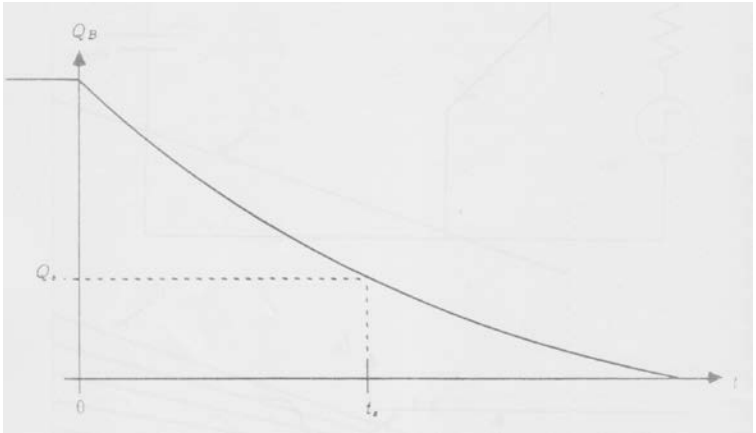


Figure (4.38): The decay of stored charge in the base in a saturated inverter, when the base current drive, I_B, is turned off at $t = 0$.

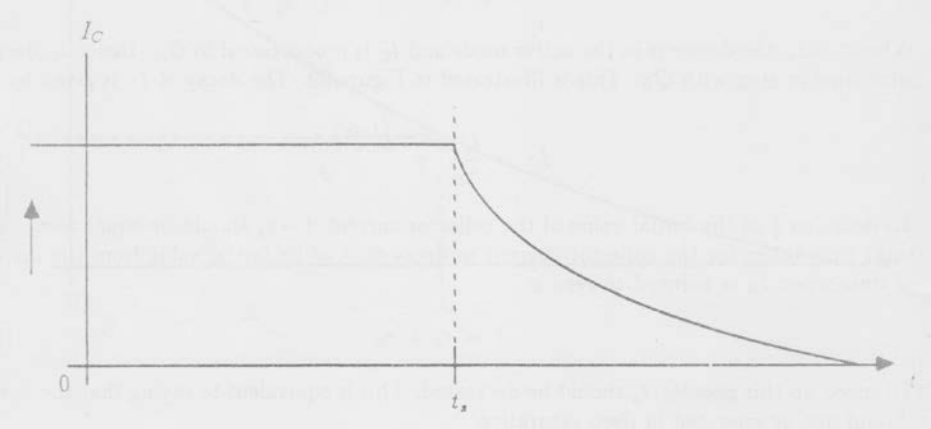


Figure (4.39): The decay of collector current in a saturated inverter. The current starts to decay only after a time delay equal to the storage time.

Ebers-Moll Model

The model is a basic model for the bipolar transistor which relates dc currents to the junction voltages and is useful for analyzing the behavior of the devices under large signal input.

We saw earlier that in the active region of operation of the bipolar transistor, the current in the emitter, I_E is given as

$$I_{Es}\left(e^{\frac{qV_{BE}}{kT}}-1\right),$$

and the current in the collector is given as

$$I_C = \alpha_0 I_{Es}\left(e^{\frac{qV_{BE}}{kT}}-1\right) + I_{Cs}\left(e^{\frac{qV_{BC}}{kT}}-1\right)$$

The current in the collector comprises a collector base junction diode current and a current due to minority carrier injection in the base from the emitter. If we want to describe the behavior of the device under saturation as well as in the inverse mode of operation, we must take into account the emitter current due to injection of minority carriers in the base from the collector also. This is what is done in the Ebers-Moll Model and is illustrated in Figure (1).

The emitter current, the base current and the collector current are all shown as going into the device. The emitter current comprises two components, one due to the junction diode current and the other due to minority carrier injection from the collector into the base.

$$I_E = -I_{Es}\left(e^{\frac{qV_{BE}}{kT}}-1\right) + \alpha_r I_{Cs}\left(e^{\frac{qV_{BC}}{kT}}-1\right) \quad (1)$$

where α_r is the alpha of the device in the inverse or reverse mode i.e., when the collector is used as the emitter and the emitter is used as the collector. The collector current is now written as

$$I_C = \alpha_f I_{Es}\left(e^{\frac{qV_{BE}}{kT}} - 1\right) - I_{Cs}\left(e^{\frac{qV_{BC}}{kT}} - 1\right) \quad (2)$$

where α_f is the alpha of the device in the normal mode. The base current is equal to

$$I_B = -(I_E + I_C) = -(1-\alpha_f)\, I_{Es}\left(e^{\frac{qV_{BE}}{kT}} - 1\right) - (1-\alpha_r)\, I_{Cs}\left(e^{\frac{qV_{BC}}{kT}} - 1\right) \quad (3)$$

The transistor is shown as two diodes connected back to back with a current source connected in parallel to each diode. The current is shown in the emitter and also in the collector as consisting of two components, one due to the diode current and the other from a current source due to minority carrier injection from the opposite region. The currents are controlled by the internal junction voltages. It is customary to write the equations as

$$I_E = a_{11}\left(e^{\frac{qV_{BE}}{kT}} - 1\right) + a_{12}\left(e^{\frac{qV_{BC}}{kT}} - 1\right) \quad (4)$$

and

$$I_C = a_{21}\left(e^{\frac{qV_{BE}}{kT}} - 1\right) + a_{22}\left(e^{\frac{qV_{BC}}{kT}} - 1\right) \quad (5)$$

where

$$a_{11} = -I_{Es}$$

$$a_{12} = \alpha_r I_{Cs}$$

$$a_{22} = -I_{Cs}$$

and

$$a_{21} = \alpha_f I_{Es}$$

Using the reciprocity relationship of the two-port device, we can write $a_{12} = \alpha_{21}$. Hence,

$$\alpha_r I_{Cs} = \alpha_f I_{Es}$$

Therefore only three out of the four parameters viz., $\alpha_f, \alpha_r, I_{Es}$ and $I_{Cs,}$ are needed to model the device.

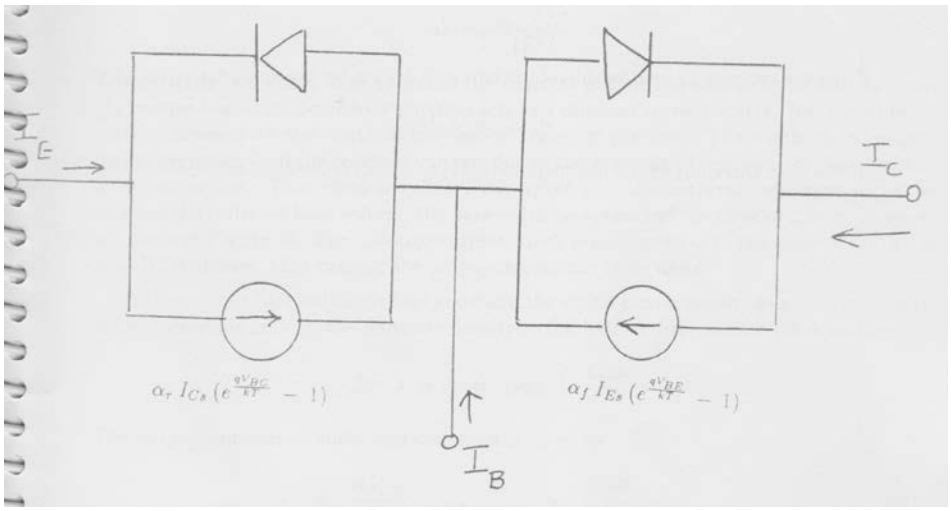


Figure 1: Equivalent circuit diagram for the bipolar transistor according to Ebers-Moll Model

Output Impedance

When we examined the DC characteristics of the transistor, we observed that the collector current increased slightly with collector voltage in the common-base mode, whereas in the common-emitter mode the collector current increased much more with collector voltage. The rate of increase of collector current with collector voltage is called the ***collector output impedance***. The output impedance (r_o) is defined as

$$r_o \equiv \frac{\partial V_C}{\partial I_C} \tag{4.158}$$

Theoretically, we would have expected the collector output impedance to be infinite, since the reverse-biased collector-base junction acts as a constant current source. But the collector current depends on the width of the neutral region of the base. ***The width of the neutral region decreases with the collector voltage, due to the widening of the collector-base junction depletion region*****. The effect is called *Early effect*,** and is illustrated in Figure (4.40). If we increase the collector-base voltage, the base width decreases and the slope of $\Delta n(x')$ increases as shown in Figure (4.40). The collector current which is proportional to the slope of $\Delta n(x')$ at $x' = W'$ increases, thus causing the output impedance to be finite.

If the emitter-base voltage is kept constant, the excess carrier density at $x' = 0$ is constant since it depends only on the impurity density in the base and the emitter-base voltage, i.e.

$$\Delta n(x' = 0) = n_{PBO}\left(e^{\frac{qV_{BE}}{kT}} - 1\right)$$

The output impedance under this condition is given by

$$r_o \equiv \frac{\partial V_{CB}}{\partial I_{Cn}} \Big|_{V_{BE=constant}} = \frac{\frac{\partial V_{CB}}{\partial W'}}{\frac{\partial I_{Cn}}{\partial W'}} \Big|_{V_{BE=constant}} \quad (4.159)$$

$\frac{\partial W'}{\partial V_{CB}}$ can be obtained by differentiating Equation (4.79) with respect to $V_{CB.}$

Common-Base Mode Output Impedance

Let us now look at the output impedance when the transistor is used in the common-base mode. In Figure (4.9), the DC collector current-voltage characteristics were plotted in the common-base mode with the emitter current as a parameter with specific values. The output impedance is obtained from this plot under conditions of constant emitter current as

$$r_o \equiv \frac{\partial V_{CB}}{\partial I_C} \Big|_{I_{E=constant}} \quad (4.160)$$

The emitter current is determined both by the emitter-base voltage and the gradient of the excess minority carrier density in the base. As the base width decreases, the slope increases and the emitter current will also increase. In order to keep the emitter current constant, the emitter-base forward voltage will have to decrease as illustrated in Figure (4.41). Thus the increase in the collector current with collector voltage is not as large as what would result if it were due to the decrease in the base width, with the emitter base voltage remaining constant.

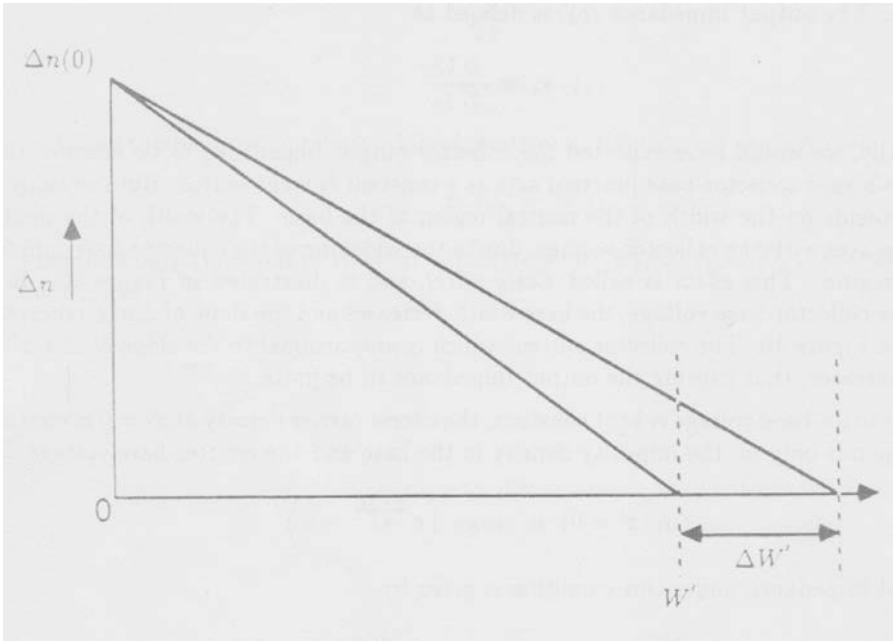


Figure (4.40): Base width modulation due to Early Effect

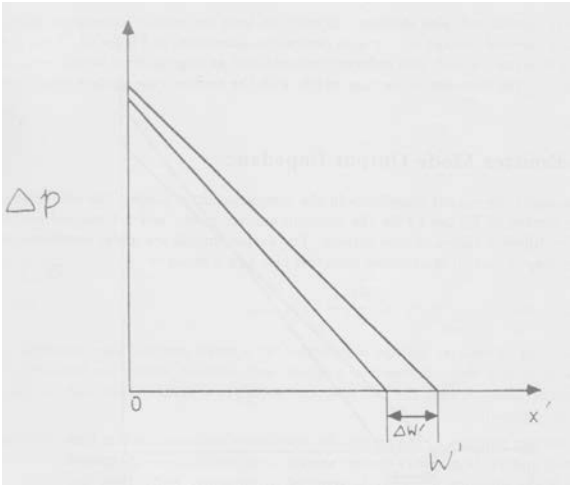


Figure (4.41): The excess carrier density distribution in the base for two different collector voltages in the common-base mode. The base-emitter voltage has to decrease to keep the emitter current constant when the collector voltage is increased.

Common-Emitter Mode Output Impedance

Let us now consider the output impedance in the common-emitter mode. The DC characteristics were plotted in Figure (4.12) for the common emitter mode, and a family of curve was plotted for different values of base current. The output impedance under conditions of constant base output current is obtained from this plot and is equal to

$$r_o = \frac{\partial V_{CE}}{\partial I_C} \Big|_{I_{B=constant}} \tag{4.161}$$

When the collector to emitter voltage is increased by a small amount, this increment is partially applied across the collector-base junction and partially across the base-emitter junction. This division between the two junctions occurs in a manner such that the base current remains constant.

The base current supplies a) the current for injection of minority carriers from the base into the emitter, and b) the majority carriers needed for recombination with minority carriers in the base. If the injection efficiency is assumed to be nearly unity, then the former is negligible, and the latter is the dominant component of the base current. The base current is then equal to

$$I_B \approx \frac{|Q_B|}{\tau_n} \tag{4.162}$$

where Q_B is the stored minority carrier charge, and is proportional to the area under the plot of the excess minority carrier density in the base, as shown in Figure (4.42). I_B is therefore constant when $Q_B$is constant. When V_{CE} is increased, the base width decreases due to the increase in the collector-base voltage. In order to keep the stored charge Q_B constant, the excess carrier density at $x' = 0$ (height of the triangle) has to increase so that the area of the triangle is constant, as illustrated in Figure (4.42).

The collector current increase is therefore not only due to the increase in slope arising from a decrease in base width, but also due to an increase in the injected minority carriers at $x' = 0$. Thus the output impedance is much smaller than what is obtained in the common base mode.

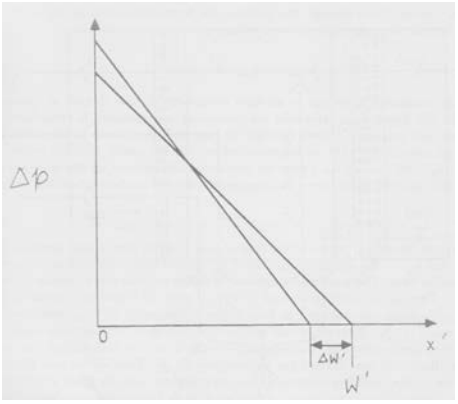


Figure (4.42): The excess carrier density in the base for two different values of the collector to emitter voltage in the common emitter mode under conditions of constant base current. The area under the plot has to remain constant if the base current is kept constant.

Non-ideal base current

In our discussion of the bipolar transistor so far, we considered the emitter-base junction to be an ideal p-n junction. However, the emitter-base junction has non-ideal components, due to recombination in the depletion region. This is similar to our discussion of the forward-biased non-ideal p-n junction. We saw in our treatment of p-n junctions, that an additional component of current flows through the junction due to recombination in the depletion region. This current depends on the forward voltage as

$$I_{rec} = I_{S\,rec}\, e^{\frac{qV_F}{2kT}} \tag{4.163}$$

The non-ideal component gives rise to an extra component of base current as shown in Figure (4.43). In this figure, $I_{B\,rec}$ represents the component corresponding to recombination in the emitter-base junction. It flows through the base lead, in addition to the other two components of base current that we discussed before. We saw earlier that the other two components (I_{Ep} and $(1 - \alpha_T)I_{En}$) have an exponential dependence on the emitter-base forward voltage. In other words, these two components depend on the emitter-base voltage given by

$$I_{Ep} + (1 - \alpha_T)I_{En} = I_{E\,Sp}\, e^{\frac{qV_{BE}}{kT}} + (1 - \alpha_T)I_{E\,Sn}\, e^{\frac{qV_{BE}}{kT}} \tag{4.164}$$

where $I_{E\,Sp}$ is the saturation current of the emitter-base junction due to hole injection into the emitter, and $I_{E\,Sn}$ is the saturation current of the emitter-base junction due to electron injection into the base. The non-ideal component is added to these two to obtain to total base current. The plot of the logarithm of I_B (V_{BE}) is given in Figure (4.44). In this figure, at low values of emitter-base voltage, the base current is dominated by $I_{B\,rec}$, and therefore the slope is equal to $\frac{q}{2kT}$. However, as the emitter-base voltage is increased, the other two components (the ideal components) increase faster with the

emitter-base voltage, and start to dominate. Hence the base-current has a slope given by $\frac{q}{kT}$. Thus we see that the base current can be approximated as comprising two linear regions: one for small emitter-base voltage with a slope $\frac{q}{2kT}$, and another with a slope equal to $\frac{q}{kT}$ for larger emitter-base voltage. The collector-current on the other hand is determined only by the ideal components of base current, and hence the plot of the logarithm of I_C is linear, with a slope equal to $\frac{q}{kT}$, and the ratio of I_C to $I_{B\ ideal}$ remains constant until we come to really low values of V_{BE} where the leakage current (I_{CBO}) dominates. This plot of the logarithm of I_C versus V_{BE} and the logarithm of I_B versus V_{BE} in the same figure is called the ***gummel plot***. At low values of V_{BE}, the recombination in the depletion region of the emitter-base junction dominates, and hence the base current has the exponential $\frac{qV_{BE}}{2kT}$ dependence. At higher values of V_{BE}, I_{Ep} and $(1 - \alpha_T)I_{En}$ dominate, and hence the base-current has a $\frac{qV_{BE}}{kT}$ dependence. The collector-current depends exponentially on V_{BE}, since it is equal to α_T (I_{En}). β_0 is defined as the ratio of I_C to I_B, and therefore varies with V_{BE}. In Figure (4.45), β_0 is plotted as a function of V_{BE}, the emitter-base voltage (or the emitter current I_B). At low values of V_{BE}, β_0 decreases with a decrease in V_{BE} because in this region recombination in the emitter-base junction depletion region dominates. For moderate values of V_{BE}, β_0 is constant, since in this region both I_C and I_B have an exponential voltage dependence with a slope to $\frac{q}{kT}$. At very high values of V_{BE}, the current appears to decrease, and this is due to high-injection effects. We will discuss the high-injection effects in a later section.

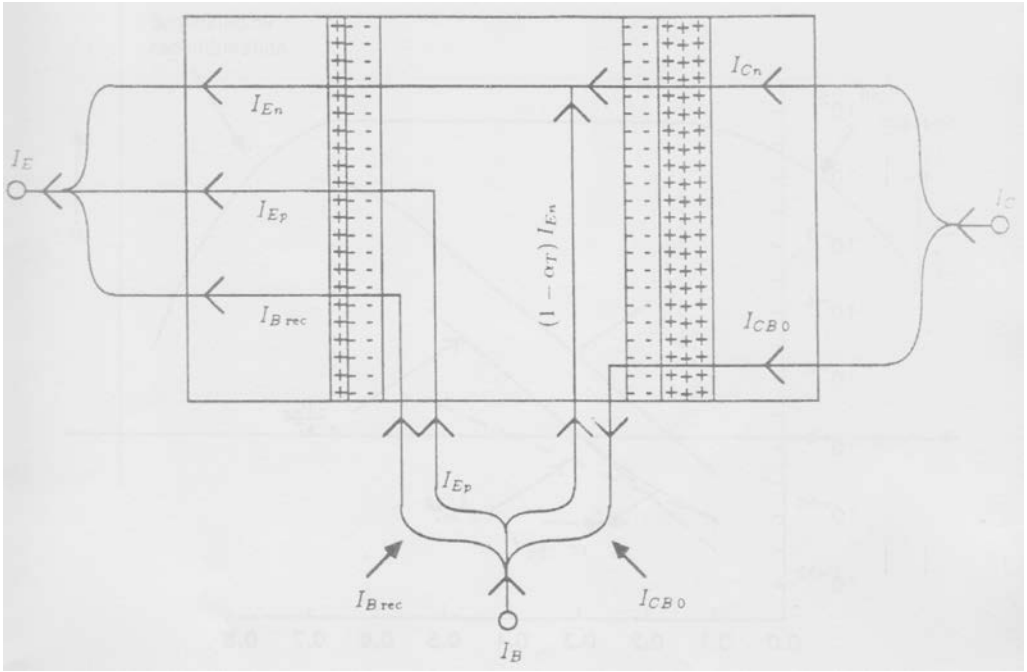


Figure (4.43): Various components of current in the bipolar transistor including the non-ideal current in the emitter-base junction.

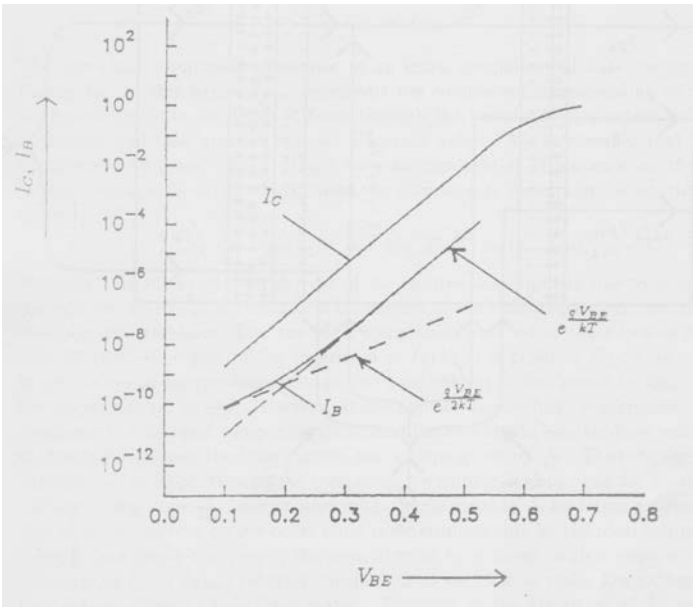


Figure (4.44): Plot of the base and collector currents as a function of the emitter-base voltage.

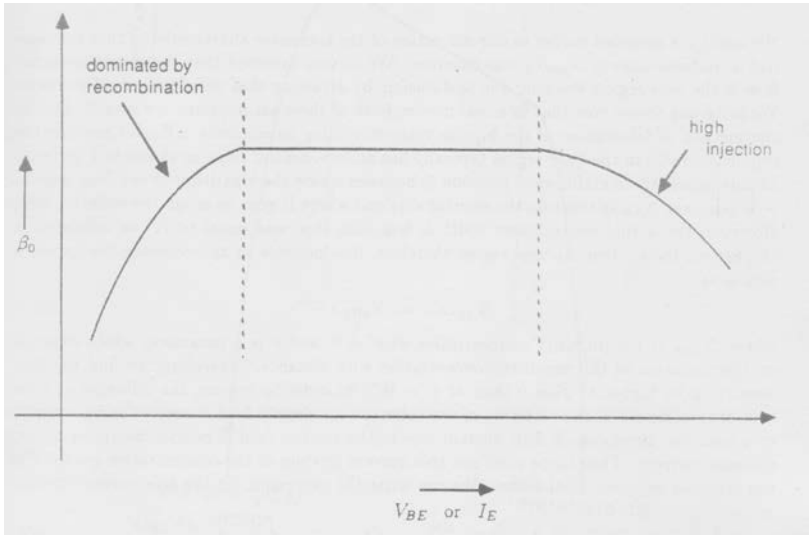


Figure (4.45): Plot of β_0 as a function of V_{BE} or $I_{E.}$

Non-uniform doping in the base

We also have assumed earlier in our discussion of the transistor characteristics that the base had a uniform doping i.e., N_A was constant. We further assumed that the minority-carrier flow in the base region was only due to diffusion, by assuming that the electric field was zero. We are going to see now that in a real device, both of these assumptions are invalid. Due to the method of fabrication of the bipolar transistor using an impurity diffusion process, the impurity profile in the base region typically has an exponential slope as shown in Figure (4.46). In this figure, the metallurgical junction is between where the logarithm of net N_{AB} goes to - ∞ (i.e., net N_{AB} is zero) on the emitter side and where it goes

to ∞ on the collector side. However, the actual neutral-base width is less than this, and equal to W' as indicated in the figure. In this neutral-base region therefore, it is possible to approximate the impurity profile as

$$N_{AB}\,(x') = \; N_{ABE}\, e^{\frac{-\eta x'}{W'}} \tag{4.165}$$

where N_{ABE} is the impurity concentration at $x' = 0, and\ \eta$ is a parameter which depends on the variation of the impurity concentration with distance. Therefore, we find the hole density to be higher at $x' = 0$ than at $x' = W'$. In order to prevent the diffusion of holes (majority carriers) in the direction of the collector, an electric field arises in the base region in a negative direction. A drift current due to the electric filed is exactly balanced by the diffusion current. Thus there is no net hole current in spite of the concentration gradient of the acceptor impurity in the base. We can write the expression for the hole current density as

$$J_p = \; q\; p_{p0}\mu_p\mathcal{E} - q\; D_{p0}\,\frac{dp_0}{dx'} = q\; N_{AB}\,(x')\mu_p\mathcal{E} - qD_p\,\frac{dN_{AB}\,(x')}{dx'} \tag{4.166}$$

where p_{p0} is the majority-carrier density in thermal equilibrium, and therefore equal to N_{AB} , the acceptor density in the base-region. In order for the two currents to balance, J_p has to equal 0. By setting the equation for J_p to 0, the electric field is expressed as

$$\mathcal{E} = \; \frac{D_p}{\mu_p} \times \frac{dN_{AB}\,(x')}{dx'} \; \times \frac{1}{N_{AB}\,(x')} = \frac{kT}{q} \; \times \frac{dN_{AB}\,(x')}{dx'} \; \times \frac{1}{N_{AB}\,(x')} \tag{4.167}$$

We had assumed earlier that N_A varies exponentially with x', and substituting this exponential function for N_{AB} we obtain the electric field as

$$\mathcal{E} = \; -\frac{kT}{q}\,\frac{\eta}{W'} \tag{4.168}$$

The electric field arises internally in the base region to prevent the diffusion of holes from the highly doped side to the lightly doped side. For this reason the electric field is called the *built-in electric field*. The electric field influences the minority-carrier flow through the base region from the emitter to the collector. The direction of the built-in electric field is such as to aid the minority-carrier flow towards the collector. Hence the transit time (τ_B) is reduced. Since the transit time is reduced, the frequency response improves. There are some additional advantages due to the inhomogeneous doping of the base. These are 1) increase in punch-through voltage, 2) increase in output impedance, and 3) decrease in base resistance. Increase in punch-through voltage arises because as the depletion-region widens in the collector-base junction, a larger increase in the collector-voltage is needed to produce a given incremental charge in the depletion region, Hence it takes a larger amount of collector-voltage to punch through the base-region than what it would have taken if the entire base-region had a uniform doping concentration equal in value to what is obtained at $x' = W'$. Because of this, the *Early effect* also gets reduced, and therefore the output impedance increases. Due to the increase in the impurity concentration in the base, the base resistance decreases. However, there are also some disadvantages due to the non-uniform doping of the base region. First, the injection efficiency (γ) decreases, since the doping concentration on the base side of the emitter-base junction is larger than in a uniformly doped

base. Secondly, the emitter-base junction depletion-region width is smaller now, and hence the junction capacitance (C_{JEB}) increases.

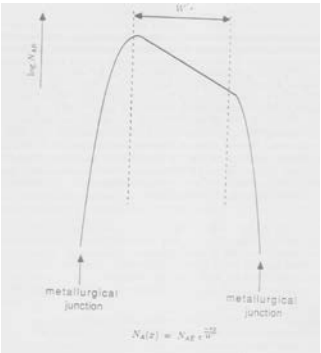


Figure (4.46): Impurity Profile in the base region in a typical device.

Summary

- The bipolar transistor works on the principle of minority carrier injection into the lightly doped side of a one-sided abrupt p-n junction and the flow (or collection) of these minority carriers into a reverse-biased p-n junction which is placed nearby.

- The typical structure of a bipolar transistor comprises a heavily doped n^+ region and a lightly doped n-region separated by a narrow p-region. The n^+ region is the emitter, the p-region the base and the n-region the collector. Such a transistor is called an n-p-n transistor. A p-n -p transistor is one in which the emitter is a p^+ region, and the collector a p-region with the intervening base layer an n-region. In normal operation, the emitter-base junction is forward-biased and the collector-base junction reverse biased.

- The total emitter current is the sum of minority carrier current injected from the emitter into the base and the minority carrier current injected from the base into the emitter. The emitter injection efficiency, γ, is defined as the ratio of minority carrier current injected into the base from the emitter to the total emitter current. In a well designed transistor, γ is made as close to 1 as possible.

- The minority carriers injected from the emitter diffuse through the base region to the collector-base junction. Before they reach the collector-base junction, some of the minority carrier current reaching the collector-base junction is less than the emitter injected minority carrier current. The ratio of the two is called the base transport factor, α_T . In a well designed transistor, α_T (which is less than 1) is made as close to unity as possible.

- The minority carriers reaching the collector-base junction from the base are swiftly swept into the collector region by the electric field in the collector-base depletion region. This results in a collector current. In addition, a reverse leakage current due to the collector-base junction also flows and adds to the collector current.

- The current through the base is the difference between the current due to majority carriers recombining with minority carriers in the base and the reverse leakage current in the collector-base junction.

- The transistor can be normally operated in either of two modes viz., common-base or common-emitter. In the common-base mode the input signal is applied between the emitter and the base and the output signal is obtained between the collector and the base. In the common-emitter mode the input signal is applied between the base and the emitter and the output signal is obtained between the collector and the emitter.

- The short circuit current gain of the transistor is defined as the ratio of the incremental change in the output current and an incremental change in the input current with the output AC short

circuited. In the common-base mode, the current gain is called alpha of the device and is denoted by α_0 where the subscript 0 refers to the fact that the frequency is very low

$$\alpha_0 = \gamma\, \alpha_T$$

In the common-emitter mode, the short circuit current gain at low frequency is called beta of the device and is denoted by β_0.

$$\beta_0 = \frac{\alpha_0}{1 - \alpha_0}$$

- In a well designed transistor, α_0 is made as close to unity as possible. This is done by making the emitter heavily doped in comparison with base doping to get γ as close to unity as possible and making the base region as narrow as possible in comparison with the minority carrier diffusion length in the base to get α_T as close to unity as possible.

- The transistor is said to be saturated or operating in the saturation regime when the collector-base junction also is forward biased. In saturation, the excess minority density in the base at the edge of the collector base space-charge region is increased due to injection from the collector. The transistor is said to be in the active regime of operation when the emitter-base junction is forward-biased and the collector-base junction is reverse biased.

- In the common-base mode of operation, the input resistance to a small signal input voltage is the small signal resistance of a forward biased p-n junction. The resistance is called the emitter resistance and is equal to $r_E = \frac{kT}{qI_E}$, where I_E is the DC emitter current. The output circuit can be represented by having a dependent current source with a current value $g_m \tilde{v}_{BE}$ where $\tilde{v}_{BE}$ is the sinusoidal signal emitter-base voltage.

- In the common emitter mode the small signal input resistance is equal to $\frac{r_E}{1-\alpha_0}$and therefore $(\frac{1}{1-\alpha_0})$ times larger than the input resistance in the common base mode. The output circuit can be represented by the same equivalent circuit as for the common base mode.

- Due to injection of excess minority carriers in the base, at any given instant of time, there is an extra amount of charge in the base. This charge is called stored base charge,. In order to maintain charge neutrality majority carriers flow into the base from the base contact to keep an excess majority carrier density everywhere in the base exactly equal to the excess minority carrier density. A small increment in the emitter base voltage causes an increase in the stored charge by an incremental charge flowing into the base from the emitter while an incremental charge of same magnitude but of opposite polarity flows into the base from the base contact to maintain charge neutrality. Thus a capacitive effect arises due to the stored charge. This is called the diffusion capacitance.

- The time taken for the minority carriers to traverse the neutral base region from the emitter-base depletion region to the collector-base depletion junction when the device is in the active mode is called the base transit time, τ_B. The collector current, I_C times the base transit time, gives the stored base charge, Q_B

$$Q_B = I_C \tau_B$$

The diffusion capacitance is related to τ_B as

$$C_{diff} = \frac{\alpha_0}{r_E} \tau_B$$

- Due to the diffusion capacitance the excess minority carrier density does not change instantaneously with a change in the emitter base voltage. Hence the gain of the transistor falls off. The equivalent circuit for the transistor has the diffusion capacitor in parallel with the input resistance. The input circuit has a time constant due to C_{diff} and r_{in}. When we ignore the effect of junction capacitances but include only the effect of the diffusion capacitance we call the device an intrinsic transistor. In an intrinsic transistor, the alpha falls off at high frequencies and the frequency at which $|\alpha|$ falls to $\frac{1}{\sqrt{2}}$ of its low frequency is called the alpha-cut-off frequency, $f_{c\alpha}^*$. The gain $|\alpha|$ falls of at a rate of 6 db/octave i.e., decreases by a factor of 2 for an increase in frequency by a factor of 2. In this range of frequencies the product of f and $|\alpha|$ is a constant. The asterisk denotes that an intrinsic transistor is being considered. The beta also decreases with frequency and the frequency at which $|\beta|$ falls to $\frac{1}{\sqrt{2}}$ of its low frequency is called the beta cut-off frequency, $f_{c\beta}^*$. Hence also, $|\beta|$ decreases by a factor of 2 when the frequency is increased by a factor of 2 at frequencies larger than $f_{c\beta}^*$. The frequency at which $|\beta|$ becomes equal to 1, is called the gain-band width product and is denoted by f_T^*. The product of f and $|\beta|$ at frequencies larger than $f_{c\beta}^*$ is a constant and equal to f_T.

- When the junction capacitances are included, the time constant of the input circuit is increased. In the output circuit capacitive shunting effects arise. Due to this, the frequency response is poorer than in the intrinsic device. The alpha cut-off frequency, the beta cut-off frequency and the gain band width product in an actual transistor are all reduced from the intrinsic transistor values.

- When the transistor is used as a saturated inverter, there is delay in the turn-on time. The delay in the turn-off time is longer than the turn-on time by the amount of storage time. The storage time is defined as the time needed to get off saturation regime and reach active regime.

- The punch-through voltage, V_{pt} , is the collector voltage at which the collector-base depletion region widens all the way until it touches the emitter-base junction depletion region, V_{pt} poses a limit on the maximum voltage that can be applied on the collector

- In the common base mode, the collector-base junction breakdown voltage represents the maximum voltage that can be applied on the collector unless punch-through occurs earlier. This voltage is denoted BV_{CBO}.

- In the common-emitter mode, the maximum collector voltage that can be applied is less than BV_{CBO} due to collector multiplication factor. This voltage is denoted BV_{CEO}.

- The output impedance of the transistor, $\left(\frac{\partial I_C}{\partial V_{CB}}\right)^{-1}$, in the common base mode is not infinite due to Early effect. Early effect is the reduction of the neutral base width with increasing collector voltage.

- The output impedance in the common emitter mode is smaller than in the common base mode due to the fact that the injection of minority carriers from the base has to increase in order to keep the base current constant when the collector voltage is increased.

- In an actual transistor, non-ideal components due to recombination in the emitter, base junction depletion region and surface recombination arise. This has the effect of lowering the gain at low emitter current whereas at higher emitter currents the gain remains at a high value.

- In a real device, the impurity profile is non-uniform in the base with the density decreasing from the emitter side to the collector side. Such a non-uniform impurity profile arises as a result of the diffusion process in the fabrication of the device. Due to this an internal electric field arises which helps the minority carriers to reach the collector in a shorter time interval. The base transit time is reduced. Hence the frequency response is improved.

Glossary

A = area of cross-section of the base region

A, B = integration constants

BV_{CBO} = collector-base junction breakdown voltage

BV_{CEO} = collector breakdown voltage in the common emitter mode

C_{diff} = diffusion capacitance of the transistor

C_{JEB} = emitter-base junction capacitance

D_n = electron diffusion constant

D_p = hole diffusion constant

D_{pC} = hole diffusion constant in the collector

dQ = incremental positive (majority carrier) charge

dt = incremental time or time interval

dV = small incremental step voltage

$\mathcal{E}$ = electric field

$\mathcal{E}_{crit}$ = critical electric field

$f_{C\alpha}$ = α- cut off frequency

$f_{C\alpha}^{*}$ = intrinsic α- cut off frequency

$f_{C\beta}$ = β - cut off frequency

$f_{C\beta}^{*}$ = intrinsic β - cut off frequency

f_{T} = gain bandwidth product

f_{T}^{*} = intrinsic gain bandwidth product

F = Farad. Unit of capacitance

g = conductance

g_m = small signal input conductance

$\tilde{\imath}_b$ = small signal sinusoidal base current

$\tilde{\imath}_c$ = small signal sinusoidal collector current

$\tilde{\imath}_e$ = small signal sinusoidal emitter current

$\tilde{\imath}_{in}$ = small signal sinusoidal input current

$\tilde{\imath}_{out}$ = small signal sinusoidal output current

I_B = base current

$I_{B\,rec}$ = component of base current due to recombination in the emitter-base junction depletion region

I_{Bt} = total base current

I_C = collector current

I_{CB0} = collector-base junction leakage current

I_{CEO} = collector current with the base open

I_{Cn} = collector current due to the injection of electrons from the emitter into the base

$I_{Cn}(x)$ = current in the base region due to injected electrons

I_{Ct} = total collector current

I_{DC} = DC current

I_E = DC emitter current

I_{En} = electron current injected from the emitter into the base

I_{Ep} = DC emitter current due to hole injection from base into the emitter

I_{ES} = total saturation current of the emitter-base junction

I_{ESn} = saturation current of the emitter-base junction due to electron injection into the base

I_{ESp} = saturation current of the emitter-base junction due to hole injection into the emitter

$I_{S\,rec}$ = saturation current due to recombination in the emitter-base depletion region

I_{Et} = total emitter current

j = $\sqrt{-1}$

J_n = minority carrier (electron) current density

$J_n(x')$ = electron current density in the base

J_p = hole current density

k = Boltzmann constant

L_n = minority carrier (electron) diffusion length in the base

L_{pC} = minority carrier (hole) diffusion length in the collector

L_{pE} = minority carrier (hole) diffusion length in the emitter

M = avalanche multiplication factor

n = electron density

n_i = intrinsic carrier density

n_{pBO} = thermal equilibrium minority carrier (electron) density in the base

N_A = acceptor density

N_{AB} = acceptor impurity concentration in the base

N_{ABE} = acceptor impurity concentration at $x' = 0$

N_{DC} = donor impurity concentration in the collector

N_{DE} = donor impurity concentration in the emitter

p_0 = majority carrier density in thermal equilibrium

p_{nC0} = thermal equilibrium minority carrier (hole) density in the collector

p_{nE0} = thermal equilibrium minority carrier (hole) density in the emitter

q = electron charge

Q_B = minority carrier charge stored in the base

Q_s = stored charge, Q_B at t =t_1

r_E = emitter resistance

r_0 = output impedance

R_L = load resistance

t = time

t_1 = turn-on transient

t_s = storage time, the time it takes for the device to get out of saturation

T = absolute temperature

$\tilde{v}_{BE}$ = base-emitter voltage

$\tilde{v}_{in}$ = small signal sinusoidal input voltage

$\tilde{v}_{out}$ = small signal sinusoidal output voltage

$\tilde{v}_0$ = small signal sinusoidal voltage

$v(x')$ = velocity of electrons as a function of x'

V = voltage applied across the junction

V_{BC} = voltage across the collector-base junction

V_{biC} = built-in voltage of the collector-base junction

V_{biE} = built-in voltage of the emitter-base junction

V_{BE} = base-emitter voltage

V_{br} = breakdown voltage of the junction

V_{CB} = collector-base voltage

V_{CC} = DC collector supply voltage

V_{CE} = collector-emitter voltage

V_{pt} = punch-through voltage

W' = width of the neutral base region

W_{BJ} = width of the base region between the metallurgical junction on the emitter side, and that on the collector side

W_E = width of the neutral emitter region

x = coordinate

x_{pC} = width of the depletion-region occurring in the base region due to the collector-base junction

x_{pE} = width of the depletion-region occurring in the base region due to the emitter-base junction

x' = coordinate system in the neutral base region where the origin is chosen at the edge of the neutral base region on the emitter side

α = current gain

α_0 = low frequency current gain in the common-base mode

α_T = base transport factor

$|\alpha(\omega)|$ = alpha as a function of $\omega = 2\pi$ times frequency

β = current gain in the common emitter mode

$\beta(\omega)$ = beta as a function of $\omega = 2\pi$ times frequency

β_0 = low frequency current gain in the common emitter mode

∂ = differential

$\partial(\Delta n)$ = incremental change in the excess minority carrier density

∂t = short time interval

$\partial x'$ = elementary distance

Δ = increment

ΔI_B = incremental change in base current I_B

ΔI_C = incremental change in collector current I_C

Δn = excess minority carrier in the base region due to minority carrier injection from the emitter

ΔI_E = incremental change in emitter current I_E

$\Delta n(x')$ = excess carrier (electron) density as a function of x'

ϵ_s = permittivity of semiconductors

γ = emitter injection efficiency

η = an empirical parameter in avalanche multiplication expression

η = an empirical parameter defining the impurity density variation in the base

μ_n = electron mobility

μ_p = hole mobility

τ_B = base transit time

τ_n = minority carrier (electron) lifetime

ω = 2π times frequency

$\omega_{c\alpha}^*$ = 2π times alpha cut-off frequency

$\omega_{c\beta}^*$ = 2π times beta cut-off frequency

ω_T^* = value of ω at which $|\beta(\omega)|$ becomes unity

∇ = divergence

Problems

(Assume room temperature in all problems except when it is specifically states otherwise. Also assume silicon as the semiconductor material. Take n_i, as $1 \times 10^{10}\ cm^{-3}$)

1. The expression for the excess minority carrier density in the neutral base region was derived in the text in terms of hyperbolic functions. Show under appropriate conditions that a) it becomes a linear function and b) it becomes an exponentially decaying function.

2. A n-p-n transistor has impurity concentrations of $5 \times 10^{19}\ cm^{-3}$ in the emitter, $5 \times 10^{16}\ cm^{-3}$ in the base and $2 \times 10^{15}\ cm^{-3}$ in the collector. Assume the distance between the metallurgical junction on the emitter side and that on the collector side called the metallurgical base width is equal to $2\ \mu m$. Assume the diffusion constant and the lifetime of the minority carriers in the base region to be $30\ cm^2/sec$ ane $3 \times 10^{-6}\ sec$ respectively. Assume that the emitter-base voltage is $-0.5\ V$ and the collector-base voltage is $5\ V$. Plot on a graph sheet the excess carrier density in the base region.

3. We derived in the class the expression for the base-transport factor, the emitter injection efficiency and the current gain for a n-p-n transistor. Do the same for a p-n-p transistor.

4. For the device in problem 2, the area of the emitter-base junction and that of the collector-base junction are the same and equal to $10^{-4}\ cm^2$. Assume the minority carrier mobility and lifetime in the emitter to be 500 $cm^2/V - sec$ and $0.25\ x\ 10^{-6}$ sec respectively. What is the injection efficiency at the voltage biases given in problem 2.

5. For the device in problem 2, determine the base transport factor at the specified voltage.

6. A p-n-p transistor has impurity concentrations of $5 \times 10^{19}\ cm^{-3}, 2 \times 10^{16}\ cm^{-3}$ and $5 \times 10^{15}\ cm^{-3}$ in the emitter, base and collector regions respectively. Assume that the distance between the two metallurgical junctions, W_{BJ} is $2.5 \times 10^{-4}\ cm$. Assume the emitter-base junction is forward biased by $0.5\ V$ and that the collector-base junction is reverse biased by 10 V. The minority carrier parameters are given below in each of the three regions.

	Emitter	Base	Collector
Lifetime (sec)	10^{-6}	10^{-4}	2x10^{-4}
Diffusion Constant (cm^2/sec)	10	15	16

Calculate $\gamma, \alpha_{T,} \alpha_0\ and\ \beta_0$.

7. For the device in problem 6, calculate I_{CBO} and I_{CEO} . Take the area of both the collecetor and the emitter junctions to be $10^{-3}\ cm^2$.
8. For the device in problem 6, draw the equivalent circuit for a) common base mode and b) common emitter base mode.

9. For the device in problem 6, calculate the punch-through voltage, BV_{CBO} andBV_{CEO} . Assume that the critical electric filed for breakdown in the collector-base junction is 200,000 V/cm. Take as η = 3

10. An n-p-n transistor has impurity concentrations of $10^{19}\ cm^{-3}$, $10^{16}\ cm^{-3}$ and $10^{15}\ cm^{-3}$ in the emitter, base and collector regions respectively. Assume that the distance between the two metallurgical junctions, W_{BJ} is $2.5\ \times\ 10^{-4}\ cm$. Assume that the emitter-base junction is forward biased to 0.5 V. Find the punch-through voltage.

11. Assume that for the device in problem 10, $\tau_B = 5\ \text{x}10^{-8}$ sec and the minority carrier lifetime in the base is 10^{-5} sec. Assume the transistor to be intrinsic. What is BV_{CEO} , assuming that γ is nearly unity. Take the value of the parameter η as 3 and $BV_{CBO}$to be 250V.

12. For the device in problem 11, determine $f_{C\alpha}$, $f_{C\beta}$ and f_T . Assume the device is intrinsic i.e., you can neglect the parasitic elements and the junction capacitances.

13. Assume an area of $10^{-3}\ cm^2$ for emitter-base junction for the device in problem 10. Assume τ_B is not the value given in problem 11. Assume the width of the neutral base region W' to be $\sqrt{2}\ \times\ 10^{-4}$ cm. Let the minority carrier diffusion constant in the base be 16 $cm^2 sec^{-1}$. Calculate $f_{C\alpha}$ and $f_{C\beta}$ at an emitter-base bias voltage of 0.6 V and a collector-base reverse bias of 5 V. (Note: You are still assuming that γ is nearly unity.)

14. For the operating voltage given in the last problem determine the equivalent circuit. Assuming the area of the collector-base junction to be the same as that for the emitter-base junction.

15. Assume that V_{CC} is $10\ V$ for the device in the last problem. Calculate the storage time when the device is used as an inverter with a load resistor of resistance equal to $10\ K\Omega$. Assume that the base current drive is $2\ mA$ and that the current has been on for a long time for the purpose of calculating the storage time.

16. A switching transistor has the following characteristics $f_{C\beta}^* = \ 10^6 Hz$ and $f_T^* = 10^8\ Hz$. The device is used as a saturated inverter and assume (a) γ, the emitter injection efficiency is nearly

unity, b) the load resistor is $5K\Omega$ and (c) the collector supply voltage is $10\ V$. A base drive current of $50\ \mu A$ is applied at $t\ = 0$. Calculate the time required to turn the device on.

17. In the device in problem 16, assume that the base current of $50\ \mu A$ is flowing for sufficiently long time to establish steady state condition. Assume that at $t\ = 0$, the base current is switched off. Calculate the turn-off time which is defined as the time required for the collector current to decay to a predetermined fraction such as $\frac{1}{e}$ or 10% of the initial value.

18. For the condition given in problem 17, determine the emitter-base voltage at time $t\ = 0$, assuming that the device is a p-n-p transistor and the base doping is $10^{16}\ cm^{-3}$.

19. The impurity concentration in the base of a n-p-n transistor is given approximately by the function

$$N_B(x') = N_{BE}\exp(\frac{-\eta x'}{W'})$$

Where N_{BE} = impurity concentration at $x' = 0$ and η is a parameter. Take W' as equal to 2 μm.

(a) What is the value of η if $N_{BC} = N_B(x' = W') = 10^{-3}N_{BE}$. Do not confuse this η with the one that we used to calculate $BV_{CEO.}$

(b) What is the value of the electric field in the base?

20. Design a n-p-n transistor with the following specifications. It should have a minimum β_0 of 40 and it should be capable of withstanding a collector voltage of $100\ V$. It should have an intrinsic beta cut-off frequency of $150\ KHz$ or larger. Assume you can alter the minority carrier lifetime to any specific design value by adding g-r centers with a capture cross-section of $10^{-16}\ cm^2$. Assume that the starting material has a lifetime of $100\ \mu sec$. (Note: There is no unique design.)

Chapter 5

MOS Devices

MOS devices refer to a class of devices in which a thin insulating layer such as silicon-di-oxide (usually a thermally grown oxide) is sandwiched between a metallic plate and a semiconductor. The term MOS stands for Metal-Oxide-Semiconductor. These types of devices are called field-effect devices, since a voltage applied between the metallic plate and the semiconductor gives rise to an electric field perpendicular to the interface between the semiconductor and the insulating layer. The electric field alters the electrical properties of the semiconductor in a region close to the interface by inducing a space charge. The term MIS is used to describe in which, instead of an oxide, any insulting layer is used.

MOS Capacitor

We will first consider a simple structure called an MOS capacitor, which is also sometimes called an MOS diode. The device is just a parallel-plate capacitor with the metallic layer as one plate, and the semiconductor as the other parallel plate, with the insulating layer as the dielectric region between the parallel plates. The device is illustrated in Figure (5.1). This type of structure is fabricated by oxidizing the surface of silicon in a furnace which is heated to a very high temperature, such that a thin oxide layer of thickness typically less than 1000 Å is grown on the surface. Silicon has a great affinity for oxygen, and therefore the thermal oxidation of silicon is easy to accomplish. The interface between silicon and the oxide has a low interfacial charge. A metallic film is then deposited on top of the oxide layer to form the MOS sandwich.

The metal is called the gate. In modern devices the gate is made of polysilicon instead of a metallic layer. The semiconductor is called the substrate. A voltage applied on the gate with respect to the substrate give rise to a perpendicular electric field in the oxide layer. The electric field induces charge in the semiconductor. The induced charge generally exists over a certain spatial region of the semiconductor, and hence it is called the spacer charge. The space charge is induced in a region very close to the semiconductor-oxide interface. This field-effect is illustrated in Figure (5.2). $\mathcal{E}_{ox}$ is the electric field in the oxide layer. The electric field lines terminate on the charges in the space charge region. Choosing the interface between the oxide and the substrate as the origin of the x-axis in Figure (5.2), the space charge region extends up to a depth of x_d and the neutral region is below the space charge region.

In the neutral region of the semiconductor, the net charge density is zero, and in the space charge region of the semiconductor, the charge density is not zero. The total charge of the free carriers in a neutral semiconductor, which is mostly due to majority carriers, is equal and opposite to that of the ionized impurities. Hence, in the neutral region the charge density ρ is zero, as shown in the following equation.

$$\rho = q[p_0 - n_0 + N_D^+ - N_A^-] = 0 \tag{5.1}$$

As done in earlier chapters, we use a subscript 0 to denote the thermal equilibrium carrier density. On the other hand, in the space-charge region, the free-carrier densities are less than in the neutral region,

and hence a net charge density results. The electrostatic potential, $\emptyset$, varies in the space charge region as given by Poisson equation.

$$\nabla^2 \emptyset = -\frac{\rho}{\epsilon_s} \tag{5.2}$$

The potential energy of the electron, which is equal to E_c, varies in the space-charge region with distance, and therefore E_i and E_v also vary with distance as we saw in our discussion of the depletion region in pn junctions. This is usually called the bending of the bands in the space-charge region. In the neutral region of the semiconductor, where the electric field is zero, the potential and therefore the potential energy are constant with distance, and hence the band is said to be flat. This is illustrated in Figure (5.3). Since there is no electric current in the direction of x, the Fermi energy, E_F, does not vary with x. In the example shown in Figure (5.3) we have assumed the substrate be p-type.

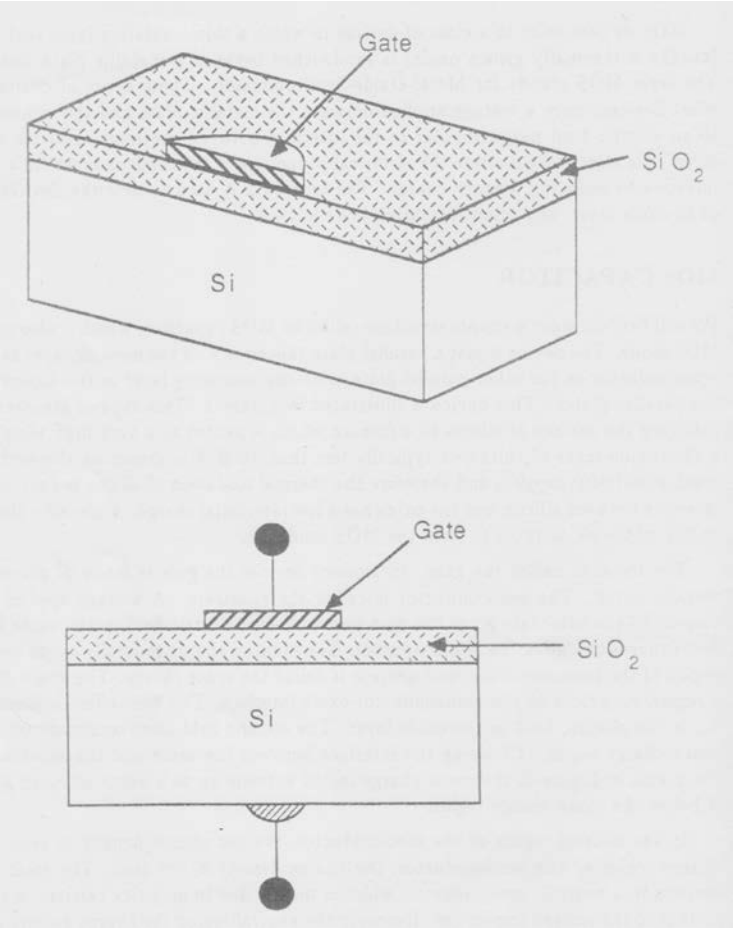


Figure (5.1): Cross-section of an MOS capacitor

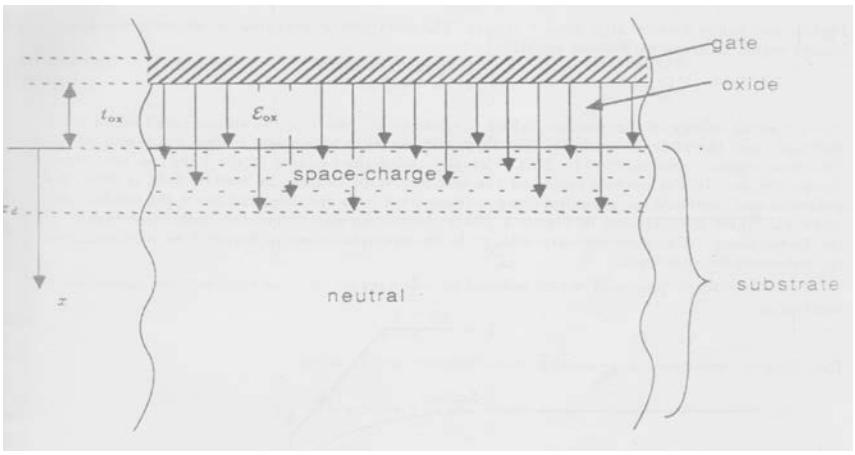


Figure (5.2): Illustration of the field-effect in the MOS structure. A sheet charge is induced in the gate and a space-charge is induced in the semiconductor

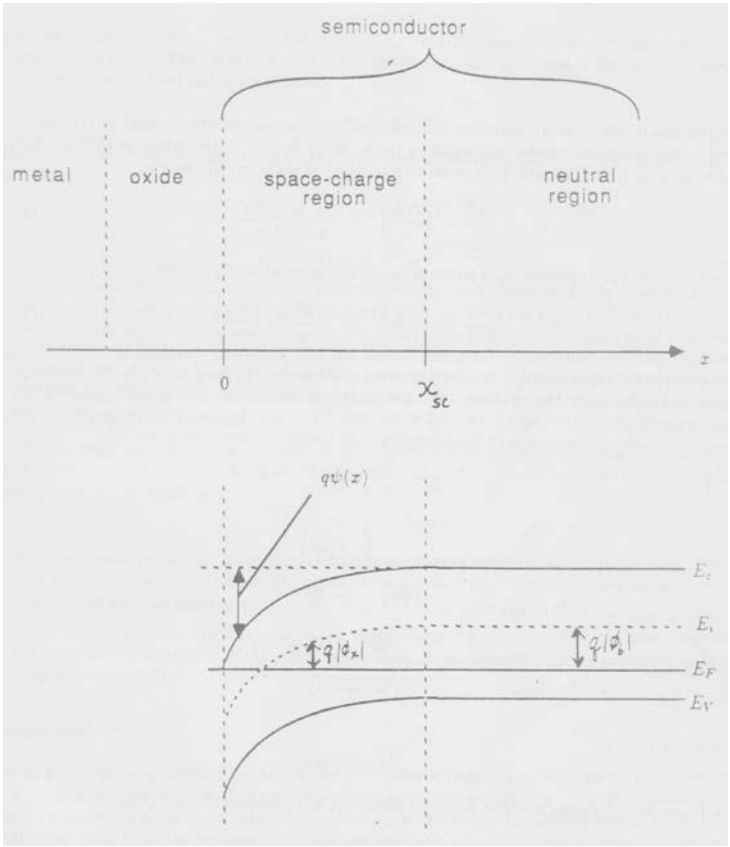


Figure (5.3): The space-charge region in the substrate of an MOS capacitor. (a) Four distinct regions of the MOS capacitor (b) The band-bending in the space-charge region of the semiconductor

The electrostatic potential, which is equal to $-\frac{E_c}{q}$ plus an additive constant, can therefore be written as

$$\emptyset = \frac{E_F - E_i}{q} \tag{5.3}$$

Recalling our previous discussions on carrier density, we can write

$$n = n_i e^{\left(\frac{E_F - E_i}{kT}\right)} = n_i e^{\left(\frac{q\emptyset}{kT}\right)} \tag{5.4}$$

and

$$p = n_i e^{\left(-\frac{E_F - E_i}{kT}\right)} = n_i e^{\left(-\frac{q\emptyset}{kT}\right)} \tag{5.5}$$

The neutral region beneath the space-charge region is referred to as the bulk, where the carrier densities are at their thermal equilibrium values. The carrier densities in the bulk are denoted n_b and p_b, and are given by

$$n_b = n_i e^{\left(\frac{q\emptyset_b}{kT}\right)} \tag{5.6}$$

and

$$p_b = n_i e^{\left(-\frac{q\emptyset_b}{kT}\right)} \tag{5.7}$$

Here we have defined a bulk electrostatic potential as equal to

$$\emptyset_b = \frac{E_F - E_{ib}}{q} \tag{5.8}$$

where E_{ib} is the intrinsic Fermi energy level, and $E_F - E_{ib}$ represents the separation between the Fermi energy and the intrinsic Fermi energy in the bulk. We can now write the expression for the electrostatic potential $\emptyset_b$ in the bulk in terms of the thermal equilibrium carrier densities.

$$\emptyset_b = \frac{kT}{q} ln\left(\frac{n_b}{n_i}\right) = -\frac{kT}{q} ln\left(\frac{p_b}{n_i}\right) \tag{5.9}$$

In an n-type substrate, n_b is larger than n_i, and hence $\emptyset_b$is positive, while in a p-type substrate p_b is larger than n_i, and therefore it is negative. Again we can express this in terms of the impurity concentrations in the bulk by assuming that the majority carrier density is equal to the donor density in the n-type substrate, and equal to the acceptor density in the p-type substrate. For example, in an n-type substrate, n_b is equal to n_{n0} which is equal to N_D^+. Hence

$$\emptyset_b = \frac{kT}{q} ln\left(\frac{N_D^+}{n_i}\right) \approx \frac{kT}{q} ln\left(\frac{N_D}{n_i}\right) \tag{5.10}$$

Similarly, in a p-type substrate, $p_b$is equal to p_{p0}, which is equal to N_A^-. Hence

$$\emptyset_b = -\frac{kT}{q} ln\left(\frac{N_A^-}{n_i}\right) \approx -\frac{kT}{q} ln\left(\frac{N_A}{n_i}\right) \tag{5.11}$$

The reader should easily recall that $\emptyset_b$is the same electrostatic potential $\emptyset_n$ or $\emptyset_p$ we defined in Chapter 2. In the above two equations we have made the assumption that at room temperature all the impurity atoms are ionized, i.e., $N_D^+ = N_D$ and $N_A^- = N_A$. In the space-charge region, the electrostatic potential varies due to band bending, and hence the carrier densities vary.

$$n(x) = n_i \, e^{\left(\frac{q\,\emptyset(x)}{kT}\right)} \tag{5.12}$$

and

$$p(x) = n_i \, e^{\left(-\frac{q\,\emptyset(x)}{kT}\right)} \tag{5.13}$$

Multiplying both sides of the equation by $e^{\left(\frac{q\emptyset_b}{kT}\right)} \times e^{\left(-\frac{q\emptyset_b}{kT}\right)}$, which is equal to 1, we get

$$n(x) = n_i e^{\left(\frac{q\emptyset_b}{kT}\right)} e^{\left(\frac{q(\emptyset(x)-\emptyset_b)}{kT}\right)}$$

$$= n_b \, e^{\left(\frac{q(\emptyset(x)\,-\,\emptyset_b)}{kT}\right)} \tag{5.14}$$

and similarly,

$$p(x) = p_b e^{\left(-\frac{q(\emptyset(x)-\emptyset_b)}{kT}\right)} \tag{5.15}$$

We will now define as **electrostatic potential difference**, $\boldsymbol{\psi(x)}$ between x and x_d ,as

$$\psi(x) = \emptyset(x) - \emptyset_b \tag{5.16}$$

We can think of $\psi(x)$ as representing the electrostatic potential at some point x, measured with respect to the bulk. The carrier densities at x in the space-charge region can be now expressed in terms of this electrostatic potential difference ψ as

$$n(x) = n_b \, e^{\left(\frac{q\,\psi(x)}{kT}\right)} \tag{5.17}$$

and

$$p(x) = p_b \, e^{\left(-\frac{q\,\psi(x)}{kT}\right)} \tag{5.18}$$

We notice that

$$n(x)\, p(x) = n_b \, p_b = n_i^2$$

The law of mass action is valid in the space-charge region, since the space-charge region is in thermal equilibrium. The value of ψ at $x = 0$ is denoted ψ_s, and is defined as the surface potential (or more correctly, the interface potential).

$$\psi_s = \psi\,(x = 0) \tag{5.19}$$

The electron and hole concentrations at the interface are obtained by substituting ψ_s in Equations (5.17) and (5.18)

$$n_s = n(x = 0) = n_b e^{\left(\frac{q\psi_s}{kT}\right)} \tag{5.20}$$

$$p_s = p(x = 0) = p_b e^{\left(-\frac{q\psi_s}{kT}\right)} \tag{5.21}$$

$-\,\mathrm{q}\,\psi(x)$ denotes the value by which the potential energy (and therefore $\mathrm{E_c}$) is different at some value x, from the value in the bulk. In other words, $q\,\psi(x)$ denotes the amount of band bending. If $\psi(x)$ is positive, the potential energy at x is less than in the bulk, and the bands bend downward. The electron concentration increases, and the hole concentration decreases from their respective bulk values (i.e., the values in the neutral region) as the interface is approached. If $\psi(x)$ is negative, the band bends upward as the surface is approached. The hole concentration increases and the electron concentration decreases.

The amount of space charge induced in the semiconductor under a unit area of cross-section at the interface is defined as the space-charge density Q_{sc}. If x_{sc} is the width of the space-charge region, then Q_{sc} is equal to

$$Q_{sc} = \int_0^{x_{sc}} \rho\,(x)\,dx \tag{5.22}$$

Q_{sc} can be visualized as the amount of charge in the space-charge region contained in a cylinder of unit area of cross-section and length equal to the width of the space-charge region, as shown in Figure (5.4).

We will now consider the behavior of the space-charge region under different bias conditions. For the purposes of discussion of the principles involved, let us consider a p-type substrate, although we could have (equally well) chosen an n-type semiconductor.

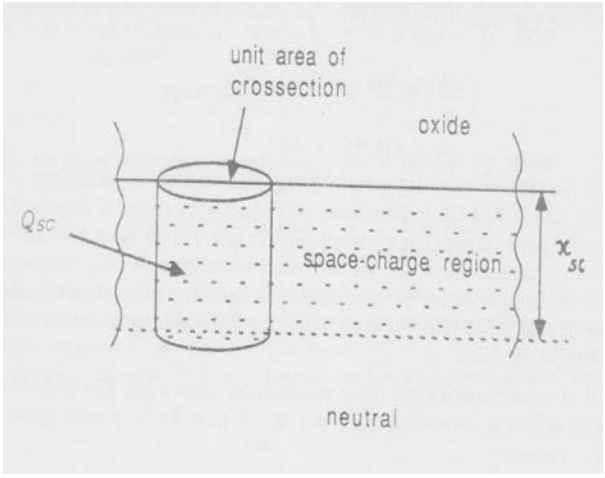


Figure (5.4): The space-charge density Q_{sc}, is defined as the amount of charge in the space-charge region under a unit area of cross-section at the interface. Therefore Q_{sc} is equal to the charge contained in a cylinder in the space-charge region of length equal to the width of the space-charge region and unit area of cross-section.

Accumulation

Let us first assume a negative charge voltage on the gate. The voltage is divided between the oxide layer and the space-charge region. The potential across space-charge region is the surface

potential ψ_s, and ψ_s is therefore negative. Another way of looking at this is that the electric field at the interface is directed away from the semiconductor towards the gate. Potential increases from the surface as the bulk is approached and hence ψ_s is negative. The induced space-charge is therefore positive. The hole concentration increases from its bulk value, so as to give rise to a net positive charge in the space-charge region. The hole concentration at the surface is given by

$$p_s = p_b\, e^{\left(-\frac{q\psi_s}{kT}\right)} \tag{5.23}$$

For a small increase in $|\psi_s|$, p_s increases enormously because of the exponential dependence on ψ_s, and also because p_b (which is the thermal equilibrium majority carrier density) is large. The space-charge region occurs in a very thin region near the surface, so that the space-charge can be considered as a sheet charge. Since the majority carrier charge is increased at the interface by accumulating majority carriers, the space-charge region is called the accumulation region. Under this condition, the semiconductor is said to be accumulated.

The band-bending under conditions of accumulation is illustrated in Figure (5.5). The space charge region extends to a width $x_{sc} = x_{acc}$. Since the space charge due to accumulation extends over a small region of very narrow width, the accumulation charge can be considered to be a sheet charge for all practical purposes.

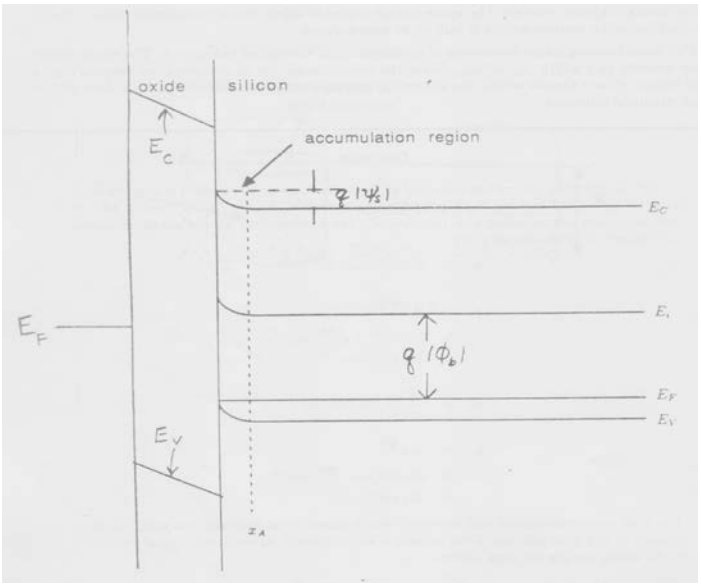


Figure (5.5): Band-bending under conditions of accumulation at the surface of a p-type substrate in an MOS capacitor. In this figure, the conduction band and the valence band in the oxide are also shown.

Example

Let us calculate the carrier densities at the surface, given the surface is accumulated with $\psi_s = -0.1$ V and the bulk impurity concentration $N_A = 5 \times 10^{15} cm^{-3}$. Let us consider room temperature and therefore we can assume that all the acceptor atoms are ionized at room temperature.

$$p_b = 5 \times 10^{15} cm^{-3}$$

$$p_s = p_b \, e^{\left(-\frac{q\psi_s}{kT}\right)}$$

$$= 5 \times 10^{15} e^{\left(\frac{0.1}{0.0259}\right)} cm^{-3}$$

$$= 2.4 \times 10^{17} cm^{-3}$$

$$n_b = \frac{n_i^2}{p_b} = 20{,}000 \; cm^{-3}$$

$$n_s = n_b e^{\left(\frac{q\psi_s}{kT}\right)}$$

$$= 20000 \times e^{\left(-\frac{0.1}{0.0259}\right)} cm^{-3}$$

$$= 421 \; cm^{-3}$$

The hole concentration at the surface is much larger than the hole (majority carrier) density in the bulk and hence the surface is accumulated. Alternately, we can calculate n_s, by using the law of mass action.

$$n_s = \frac{n_i^2}{p_s} = \frac{10^{20}}{2.4 \times 10^{17}} = 421 \; cm^{-3}$$

Depletion

When the gate voltage is positive, ψ_s is positive and the induced space-charge is negative. The hole concentration in the space charge region is decreased from its thermal equilibrium value, and hence the net charge is negative in the space-charge region. The space-charge is assumed to be only due to the ionized impurity atoms, and not due to free carriers. It is therefore like the depletion region that we considered earlier in the p-n junction. To calculate the potential variation in the space-charge region, we make the approximation that, throughout the space-charge region, the free carrier density is small compared with the impurity density. Hence we will refer to the space-charge region under these conditions as the depletion region. The free carrier density varies in the depletion region due to the potential variation as given by Equation (5.17) and (5.18). The hole density decreases and the electron density increases as they move towards the surface. The assumption that the free carrier density in the space-charge region is negligibly small in comparison with the impurity concentration is valid everywhere in the depletion region except near the edge of the depletion region. However, we will assume that even at the edge of the depletion region the free carrier density is negligible, by taking the

majority carrier density to change abruptly from its thermal equilibrium value in the bulk to zero in the depletion region.

We can express n_s in terms of p_b, the thermal equilibrium majority carrier density in the neutral region as

$$n_s = n_b\, e^{\left(\frac{q\psi_s}{kT}\right)} = \frac{n_i^2}{p_b}\, e^{\left(\frac{q\psi_s}{kT}\right)} \tag{5.24}$$

But p_b is given by Equation (5.7). Rearranging Equation (5.7) as

$$n_i = p_b e^{\left(\frac{q\emptyset_b}{kT}\right)} \tag{5.25}$$

and substituting, we obtain

$$\frac{n_i^2}{p_b} = p_b e^{\left(2\frac{q\emptyset_b}{kT}\right)} = p_b e^{\left(-\frac{q\,2\,|\emptyset_b|}{kT}\right)} \tag{5.26}$$

Therefore

$$n_s = p_b\, e^{\left(\frac{q}{kT}(\psi_s - 2\,|\emptyset_b|)\right)} \tag{5.27}$$

From this we see that as long as ψ_s *is* < 2 $|\emptyset_b|$**,** the electron density at the surface is less than p_b, and therefore less than N_A. If the electron density at the surface is less than p_b $(= N_A)$, it is definitely also less than N_A throughout the space-charge region since $n(x)$ is less than n_s when ψ_s is positive. For the entire range of ψ_s values between 0 and 2 $|\emptyset_b|$, the assumption that the space-charge region is just a depletion region is valid, and the semiconductor surface is said to be depleted.

When $\psi_s = |\emptyset_b|$, $n_s = n_i$ as can be readily seen. Therefore, when $\boldsymbol{\psi_s} > |\emptyset_{\boldsymbol{b}}|$, $n_s > n_i$ and $p_s < n_i$, In other words, at the surface, the electron concentration is larger than the hole concentration, and hence the surface region is more like an n-type material as far as the carrier densities are concerned, and hence the surface is said to be weakly inverted. We say that the surface is weakly inverted because at the surface the electron concentration, n_s is still less than the bulk majority carrier density, p_b. It is therefore possible to subdivide the range of ψ_s values of 0 to 2 $|\emptyset_b|$ as depletion when ψ_s lies between 0 and $|\emptyset_b|$ and *weak inversion, when* ψ_s *lies between* $|\emptyset_b|$ *and* $2|\emptyset_b|$.

The band bending and the charge density in the space-charge region are shown in Figure (5.6), when the surface is depleted. The band-bending under the condition ψ that satisfies $0 < \psi_s < |\emptyset_b|$, is shown in Figure (5.6 A), while the band-bending under the condition that ψ_s satisfies $|\emptyset_b| < \emptyset_s < 2|\emptyset_b|$, corresponding to weak inversion is shown in Figure (5.6 B). In both cases the charge density in the space-charge region is still equal to $-qN_A$. For our purposes, the entire range of ψ_s values between 0 and $2|\emptyset_b|$ is said to correspond to depletion condition, for in this range the charge density in the space-charge region is equal to $-qN_A$. When the gate voltage, V_G, is of such a magnitude and polarity as to make ψ_s lie between 0 and $2|\emptyset_b|$, the MOS device is said in the depletion region of operation.

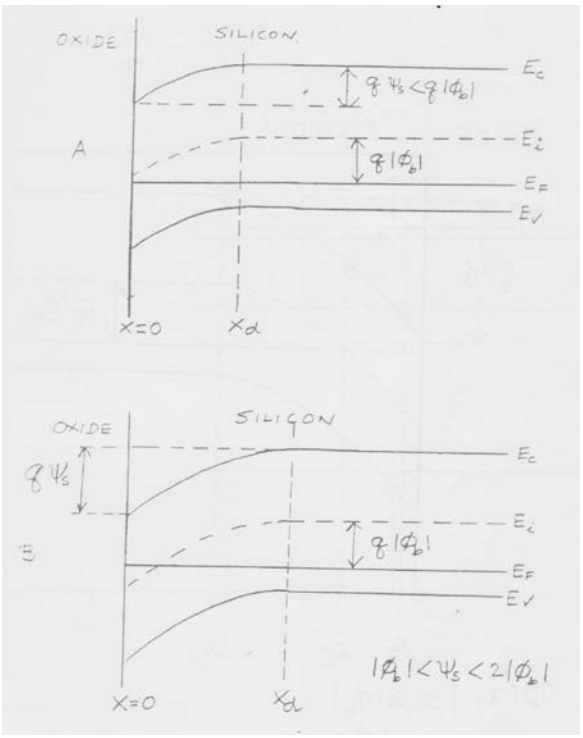


Figure (5.6): Band-bending under conditions of depletion at the surface: A) Depletion B) Weak inversion

Strong Inversion

When ψ_s becomes larger than $2|\emptyset_b|$, according to Equation (5.27) the surface electron density becomes larger than p_b, and hence N_A. The surface is said to be strongly inverted under this conditions. The band-bending and the charge density under strong inversion conditions are shown in Figure (5.7).

The space-charge region under these conditions can be viewed as comprising two regions: one in which the electron density is > N_A, and the other in which the electron density is < N_A. The former, which lies close to the surface is called the inversion region and the latter, that lies between the inversion region and the bulk neutral region, is called the depletion region. The charge density in the inversion region is determined by the electron density (since $n(x) > N_A$) while that in the depletion region is just equal to qN_A since $n(x) < N_A$.

Referring to Figure (5.7), the region lying between the surface $(x = 0)$ and x_I, where $\psi\ (x_I) = 2|\emptyset_b|$, represents the inversion region. The region to the right of x_I is the depletion region. Thus when $\psi_s = 2|\emptyset_b|$, x_I occurs at $x = 0$. The inversion region has zero width. Hence, the condition $\psi_s = 2|\emptyset_b|$ is called the on-set of inversion. As ψ_s increases beyond $2|\emptyset_b|$, the inversion region grows. The surface is said to be strongly inverted. We denote the boundary between the space charge region and the neutral bulk region $x_{d\,max}$. Hence the depletion region occurs between x_I and $x_{d\,max}$.

In reality, the inversion layer is a thin layer, the charge in this region can be approximated as a sheet charge, and x_I can be assumed to be negligibly small.

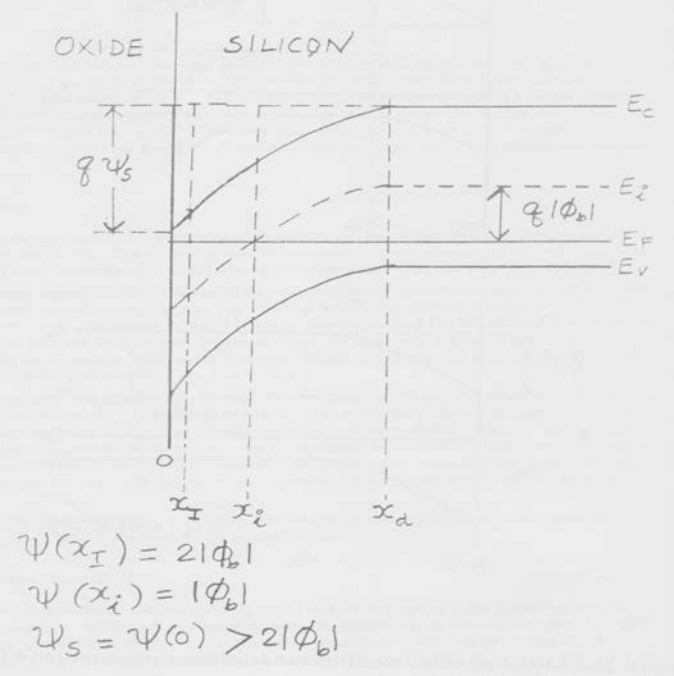


Figure (5.7): Band-bending under inversion conditions

Relation Between V_G and ψ_s

In this section, let us determine the relation between the gate voltage and the surface potential or the space-charge density. Referring to Figure (5.8), when a voltage V_G is applied between the gate and the substrate, the electric field, $\mathcal{E}_{ox}$, in the oxide induces a space-charge of density Q_{SC} in the semiconductor. If $\mathcal{E}_{ox}$ is positive, i.e., (directed into the semiconductor,) Q_{SC} is negative. Using Gauss' law, Q_{SC} is given by

$$Q_{SC} = -\epsilon_{ox}\mathcal{E}_{ox} \tag{5.28}$$

where ϵ_{ox} is the permittivity of the oxide, which is equal to

$$\epsilon_{ox} = K_{ox}\epsilon_o \tag{5.29}$$

where K_{ox} is the dielectric constant of the oxide, and ϵ_o is the permittivity of free space. For silicon-di-oxide, $K_{ox} \approx 3.9$. The voltage across the oxide layer is denoted V_{ox}, and equal to

$$V_{ox} = \mathcal{E}_{ox}t_{ox} \tag{5.30}$$

where t_{ox} is the thickness of the oxide layer. Using Equation (5.28), we can write

$$V_{ox} = -\frac{Q_{SC}}{\epsilon_{ox}}t_{ox} \tag{5.31}$$

A parallel plate capacitor using the oxide layer as the dielectric and with t_{ox} as the thickness of the dielectric layer will have a capacitance per unit area (denoted C_{ox}) equal to

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{5.32}$$

We call C_{ox} the oxide layer capacitance. Equation (5.31) can be rewritten as

$$V_{ox} = -\frac{Q_{SC}}{C_{ox}} \tag{5.33}$$

Since the gate voltage is applied across the oxide layer and the space-charge region, V_G is divided into two components, one, V_{ox}, across the oxide layer, and the other, ψ_s, across the space-charge region.

$$V_G = V_{ox} + \psi_s \tag{5.34}$$

Since Q_{SC}depends on the surface ψ_s, we will denote Q_{SC} as $Q_{SC}(\psi_s)$. Combining the equation for V_{ox} with the above equation, we can write V_G as

$$V_G = -\frac{Q_{SC}(\psi_s)}{C_{ox}} + \psi_s \tag{5.35}$$

We can now determine the relation between ψ_s and V_G for accumulation, depletion and inversion conditions.

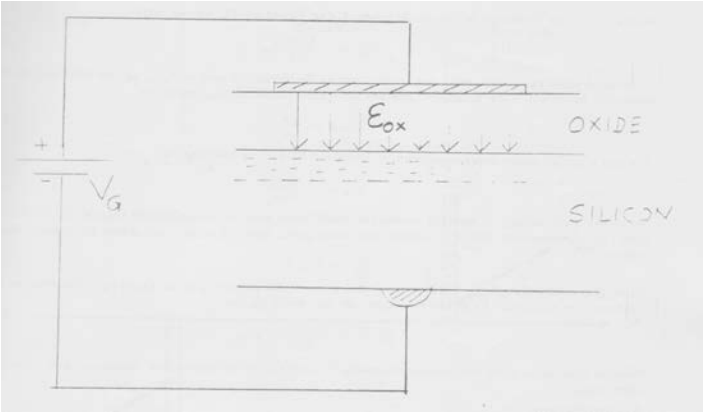


Figure 5.8: The MOS diode under the application of a DC bias

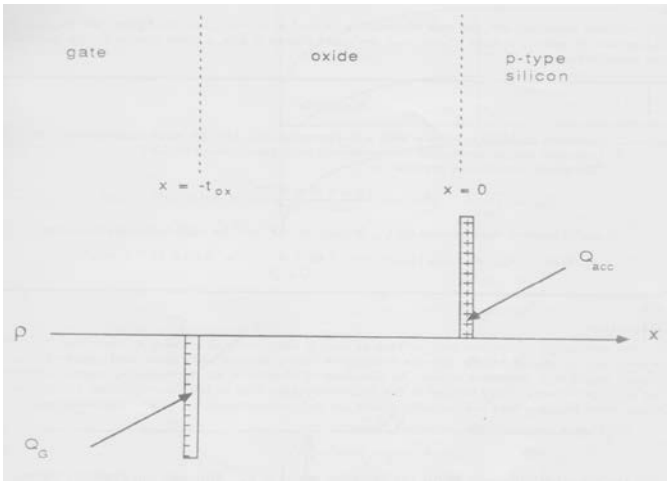


Figure (5.9): Sheet charge of majority carriers under accumulation condition

Accumulation

As stated before, we will assume that the space-charge is a sheet charge under accumulation conditions, and we take $\psi_s \approx 0$. By putting $\psi_s = 0$ in Equation (5.35), we get

$$V_G = -\frac{Q_{SC}}{C_{ox}}$$

Denoting the space-charge density in accumulation as Q_{acc}, we get

$$Q_{acc} = -C_{ox}\, V_G \tag{5.36}$$

In p-type substrates, the accumulation charge is due to holes and it is also customary to write

$$Q_{acc} = Q_p$$

Hence

$$Q_p = Q_{acc} = -C_{ox}\, V_G \tag{5.37}$$

Since V_G is negative in the accumulation condition for a p-type substrate, we see that

$$Q_p = C_{ox} |\, V_G| \tag{5.38}$$

The accumulation sheet charge is illustrated in Figure (5.9). The origin of the x-axis is chosen at the oxide silicon interface, and the gate electrode is located at $x = -t_{ox}$ in this figure. We see that the charge on the gate is a sheet charge and the space-charge is also a sheet charge at the surface of the semiconductor.

Example

Consider an MOS capacitor with a p-type substrate. Let the oxide thickness be 500 Å. Calculate the accumulation charge density at a gate voltage of – 5V.

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{K_{ox}\epsilon_{ox}}{t_{ox}} = \frac{3.9 \times 8.85 \times 10^{-14}}{500 \times 10^{-8}} = 6.9 \times 10^{-8}\ F/cm^2$$

(A useful number to remember for C_{ox} is $3.45 \times 10^{-8}\ F/cm^2$ for 1000 Å of oxide thickness.)

$$Q_{acc} = Q_p = -C_{ox} V_G = -6.9 \times 10^{-8} \times (-5) = 3.45 \times 10^{-7}\ C/cm^2$$

Depletion

We now apply a positive gate voltage of such a magnitude as to keep ψ_s less than $2|\emptyset_b|$. ψ_s is positive. As stated before, the space-charge is due to ionized impurities, and hence the space charge region is a depletion region. As discussed in chapter 3, all the impurity atoms (acceptors for a p-type substrate) are ionized in the depletion even at low temperatures, in contrast to the neutral region where the impurity atoms are only ionized partially at low temperatures.

The charge density in the depletion region is

$$\rho = -q\ N_A \tag{5.39}$$

and the potential difference across the depletion region is ψ_s. Although the depletion region can be treated as similar to that arising in a one-sided abrupt junction, and the expression for the space-charge density written straight away, we derive the expression for $\psi_s(x)$ again using Gauss theorem approach. Consider the band-bending under depletion condition shown in Figure (5.10 A), and the corresponding space-charge region shown in Figure (5.10 B). The electric field $\mathcal{E}$(x) at some point x in the depletion region is due to negatively ionized acceptor atoms lying between x and x_d in the depletion region and the direction of the electric field is positive since the charge is negative.

$$\mathcal{E}(\text{x}) = \frac{qN_A\ (x_d - x)}{\epsilon_s} \tag{5.40}$$

Integrating, we obtain

$$\psi_s(x) = -\int \mathcal{E}\ dx + C$$

$$= -\frac{q\ N_A(x_d\ x - \frac{x^2}{2})}{\epsilon_s} + \text{C} \tag{5.41}$$

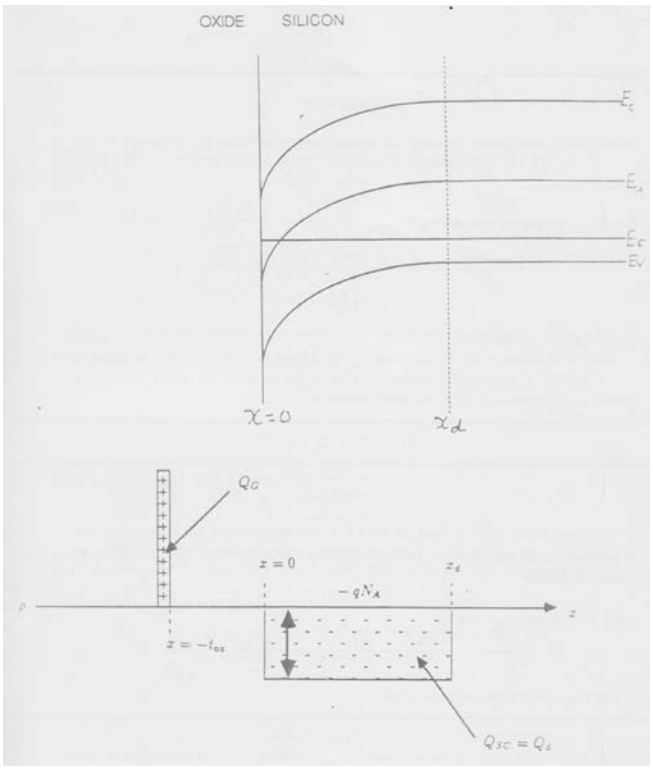


Figure (5.10): A) Band-bending under depletion conditions. B) The space-charge is due to depletion charge.

where C is an integration constant. At $x = x_d,\ \psi_s(x_d) = 0$. Hence

$$C = \frac{q\,N_A\,x_d^2}{2\epsilon_s} \tag{5.42}$$

Substituting this expression for C,

$$\psi_s(x) = -\frac{q\,N_A(x_d x - \frac{x^2}{2} - \frac{x_d^2}{2})}{\epsilon_s}$$

$$= q\,\frac{N_A}{2\epsilon_s}\,(x_d - x)^2 \tag{5.43}$$

at $x = 0$, $\psi = \psi_s$. Hence

$$\psi_s = \frac{q\,N_A\,x_d^2}{2\epsilon_s} \tag{5.44}$$

Hence the expression for $\psi(x)$ can be expressed as

$$\psi(x) = q\,\frac{N_A}{2\epsilon_s}\,(x_d - x)^2$$

$$= q\,\frac{N_A}{2\epsilon_s}x_d^2\,\left(1 - \frac{x}{x_d}\right)^2$$

$$= \psi_s \left(1 - \frac{x}{x_d}\right)^2 \tag{5.45}$$

The space-charge density under depletion conditions is denoted Q_d, since the charge is only due to depletion charge.

$$Q_{SC} = Q_d = -q\, N_A\, x_d \tag{5.46}$$

We can now express V_G as

$$V_G = -\frac{Q_d}{C_{ox}} + \psi_s = \frac{q\, N_A\, x_d}{C_{ox}} + \psi_s \tag{5.47}$$

From Equation (5.44), we can write x_d as

$$x_d = \sqrt{\frac{2\,\epsilon_s\,\psi_s}{q\,N_A}} \tag{5.48}$$

Using this expression for x_d, the depletion charge density Q_d can be expressed as

$$Q_d = -q\, N_A\, x_d = -q\, N_A \sqrt{\frac{2\,\epsilon_s\,\psi_s}{q\,N_A}} = -\sqrt{2\epsilon_s\, q\, N_A\, \psi_s} \tag{5.49}$$

Substituting the expression for Q_d in that for V_G, we get

$$V_G = \frac{\sqrt{2\epsilon_s q\, N_A\, \psi_s}}{C_{ox}} + \psi_s \tag{5.50}$$

Example

Let us calculate the depletion region width in a p-type substrate when the band-bending is $0.5\ V$. Assume the net acceptor density in the substrate is $5 \times 10^{15} cm^{-3}$. $\psi_s = 0.5\ V$

$$x_d = \sqrt{\frac{2\epsilon_s\,\psi_s}{q\,N_A}} = \sqrt{\frac{2\,K_s\epsilon_0\,\psi_s}{q\,N_A}}$$

$$= \sqrt{\frac{2\times 11.9\times 8.85\times 10^{-14}\times 0.5}{1.6\times 10^{-19}\times 5\times 10^{15}}} = \sqrt{1.316\times 10^{-9}}$$

$$= 3.63\times 10^{-5}\, cm = 0.363\ \mu m$$

This result can be scaled for other values of ψ_s by multiplying the result by $\sqrt{\frac{\psi_s}{0.5}}$ and for other values of N_A by multiplying by $\sqrt{\frac{5\times10^{15}}{N_A}}$. A useful number to remember is 1.15 μm at 1 volt drop across the depletion region for an impurity concentration of $10^{15}\ cm^{-3}$.

Example

Let us calculate the gate voltage needed to bend the band by $0.5\ V$ under depletion in an MOS capacitor with an oxide thickness of 500 Å and a net acceptor density, $N_A = 5 \times 10^{15} cm^{-3}$. $C_{ox} = 6.9 \times 10^{-8}\ F/cm^2$ from one of the previous examples.

$$V_G = \frac{\sqrt{2\epsilon_s q\ N_A\ \psi_s}}{C_{ox}} + \psi_s$$

$$= \frac{\sqrt{2\times11.9\times8.85\times10^{-14}\times1.6\times10^{-19}\times5\times10^{15}\times0.5}}{6.9\times10^{-8}} + 0.5$$

$$= 0.92\ V$$

To obtain an expression for ψ_s in terms of V_G, we rewrite equation (5.50) as a quadratic equation in $\psi_s^{\frac{1}{2}}$ as

$$\psi_s + K\ \psi_s^{\frac{1}{2}} - V_G = 0$$

where $K = \frac{\sqrt{2\epsilon_s q\ N_A}}{C_{ox}}$. Solving this quadratic equation

$$\psi_s^{\frac{1}{2}} = \frac{-K \pm \sqrt{K^2+4\ V_G}}{2} \tag{5.51}$$

Substituting this expression for $\psi_s^{\frac{1}{2}}$ in the quadratic equation and rearranging, we get

$$V_G = \psi_s + \frac{-K^2 \pm K\sqrt{K^2+4V_G}}{2}$$

$$= \psi_s + \frac{-K^2 \pm K^2\sqrt{1+\frac{4V_G}{K^2}}}{2} \tag{5.52}$$

Since $\psi_s = 0$ when $V_G = 0$, we take only the positive sign in front of the radical

$$V_G = \psi_s + \frac{K^2}{2}\left(\sqrt{1+\frac{4V_G}{K^2}} - 1\right) \tag{5.53}$$

Or

$$\psi_s = V_G - \frac{K^2}{2}\left(\sqrt{1+\frac{4V_G}{K^2}}-1\right) \tag{5.54}$$

Example

Let us calculate the surface potential of the MOS capacitor discussed in the previous example for a gate voltage of 0.7 V.

$$\psi_s = V_G - \frac{K^2}{2}\left(\sqrt{1+\frac{4V_G}{K^2}}-1\right)$$

$$K = \frac{\sqrt{2\epsilon_s q\, N_A}}{C_{ox}} = \frac{\sqrt{2\times 11.9\times 8.85\times 10^{-14}\times 1.6\times 10^{-19}\times 5\times 10^{15}}}{6.9\times 10^{-8}}$$

$$= 0.59\sqrt{Volt}$$

$$\psi_s = 0.7 - \frac{0.59^2}{2}\left(\sqrt{1+\frac{2.8}{0.59^2}}-1\right)$$

$$= 0.35\, V$$

Onset of Strong Inversion

We can now derive an expression for the gate voltage which corresponds to the onset of strong inversion. Onset of inversion occurs when $\psi_s = 2|\emptyset_b|$. We denote this value $\psi_{s\,inv}$. When the gate voltage exceeds the value corresponding to the onset of inversion, strong inversion occurs. For this reason, the gate voltage is called the turn-on voltage, since the gate voltage has to exceed this voltage to turn the device on. Denoting the **threshold voltage as V_T**, we find

$$V_T = \frac{\sqrt{2\epsilon_s q\, N_A\, \psi_{s\,inv}}}{C_{ox}} + \psi_{s\,inv}$$

$$= \frac{\sqrt{2\epsilon_s q\, N_A 2|\emptyset_b|}}{C_{ox}} + 2|\emptyset_b| \tag{5.55}$$

We are assuming that at the onset of inversion, there is no contribution to the space charge from electrons, and that the space-charge is essentially depletion charge.

Example

Let us determine the threshold voltage, V_T, for the MOS capacitor discussed in the previous examples.

$$N_A = 5 \times 10^{15}\ cm^{-3}$$

Hence

$$\emptyset_b = -\frac{KT}{q}\ln\left(\frac{N_A^-}{n_i}\right) \approx -\frac{KT}{q}\ln\left(\frac{N_A}{n_i}\right) = -0.34\ V$$

$$\psi_{s\,inv} = 2|\emptyset_b| = 0.68\ V$$

$$V_T = = \frac{\sqrt{2\epsilon_s q\ N_A\ \psi_{s\,inv}}}{C_{ox}} + \psi_{s\,inv}$$

$$= \frac{\sqrt{2\times11.9\times8.85\times10^{-14}\times1.6\times10^{-19}\times5\times10^{15}\times0.68}}{6.9\times10^{-8}} + 0.68$$

$$= 1.17\ V$$

Strong Inversion

The space charge, under strong inversion, arise due to both the inversion charge (the minority carrier charge at the surface) and the depletion charge. Denoting the inversion charge per unit area of the surface as $Q_{inv.}$

$$Q_{SC} = Q_{inv} + Q_d \tag{5.56}$$

For a small increase, $\Delta\psi_s$, in the surface potential beyond $\psi_{s\,inv}$ (equal to $2|\emptyset_b|$), the inversion charge density, Q_{inv}, increase enormously since the minority carrier density, n_s, increase exponentially with surface potential. On the other hand, the depletion charge density, Q_d, increase only slightly since it is proportional to $\psi_s^{\frac{1}{2}}$. The exact calculation of $Q_{SC}(\psi_s)$ in strong inversion is treated in advanced text books, and is beyond the scope of this book. Hence we will calculate $Q_{SC}(\psi_s)$ using an approximation called ***Depletion Approximation***. According to this approximation, in strong inversion the depletion region width, and hence the depletion charge density do not increase once inversion sets in i.e., they remain constant at a value they had at the on-set of inversion. i.e., corresponding to $\psi_s = \psi_{s\,inv} = 2|\emptyset_b|$. This means that beyond the on-set of inversion, the space-charge density, Q_{SC}, increases only due to an increase in the inversion charge density. Since the depletion region width does not increase once inversion sets in, the depletion region width is presumed to have attained a maximum value at the on-set of strong inversion and is denoted $x_{d\,max}$ where

$$x_{d\,max} = \sqrt{\frac{2\epsilon_s\ \psi_{s\,inv}}{q\ N_A}} = \sqrt{\frac{2\epsilon_s 2|\emptyset_b|}{q\ N_A}} \tag{5.57}$$

The depletion layer charge density, which also does not increase once inversion sets in, is denoted $Q_{d\,max}$ where

$$Q_{d\,max} = -qN_A x_{d\,max} = -\sqrt{2\epsilon_s q\, N_A 2|\emptyset_b|} \tag{5.58}$$

Therefore

$$Q_{SC} = Q_{inv} + Q_{d\,max} = Q_{inv} - \sqrt{2\epsilon_s q\, N_A 2|\emptyset_b|} \tag{5.59}$$

Under strong inversion, ψ_s can be written as

$$\psi_s = \psi_{s\,inv} + \Delta\psi_s = 2|\emptyset_b| + \Delta\psi_s \tag{5.60}$$

As already stated, for a small $\Delta\psi_s$, Q_{inv} increases enormously. As V_G is increased, the term - $\frac{Q_{inv}}{C_{ox}}$ will increase enormously, while $\Delta\psi_s$ will change only slightly. Hence in the depletion approximation, $\Delta\psi_s$ is neglected and ψ_s at inversion is taken as $2|\emptyset_b|$.

$$V_G = -\frac{Q_{SC}}{C_{ox}} + \psi_s = -\frac{Q_{inv}}{C_{ox}} - \frac{Q_{d\,max}}{C_{ox}} + 2|\emptyset_b| + \Delta\psi_s$$

$$= -\frac{Q_{inv}}{C_{ox}} + \frac{\sqrt{2\epsilon_s q\, N_A 2|\emptyset_b|}}{C_{ox}} + 2|\emptyset_b| + \Delta\psi_s \tag{5.61}$$

Using Equation (5.55), we can write

$$V_G = -\frac{Q_{inv}}{C_{ox}} + V_T + \Delta\psi_s \tag{5.62}$$

Hence we can write

$$V_G = -\frac{Q_{inv}}{C_{ox}} + V_T \tag{5.63}$$

Therefore

$$Q_{inv} = -C_{ox}(V_G - V_T) \tag{5.64}$$

This equation implies that the inversion charge density is zero when $V_G \leq V_T$. However, in reality it is not true, and there is some inversion charge density even when V_G is equal to V_T or slightly less than V_T due to weak inversion. However, in the region of strong inversion, i.e., when $V_G > V_T$, the depletion approximation is reasonable and Equation (5.64) is valid. The inversion charge density increases linearly with the gate voltage. Since the inversion charge density is due to electrons in a p-type substrate, the inversion charge density, Q_{inv}, is also denoted Q_n.

$$Q_{inv} = Q_n = -C_{ox}(V_G - V_T) \tag{5.65}$$

So far we discussed accumulation, depletion and inversion using a p-type substrate as a reference. If we use an n-type substrate, we will get similar expressions for the various parameters, excepting in accumulation we will have a sheet charge of electrons, and in inversion a sheet charge of holes. Also the signs will be opposite for different parameters. Table 1 summarizes the expressions for different parameters in both p-type and n-type substrates when the MOS capacitor is biased into

accumulation, depletion or inversion. Referring to Table 1, it can be seen that the surface potential is of opposite sign in an n-type substrate in comparison to the p-type substrate.

Table 5.1 Difference between substrate types

Bias	Parameter	p-type	n-type
	e.s. potential, $\emptyset_b$	$= -\frac{KT}{q}\ln\left(\frac{N_A^-}{n_i}\right)$	$= \frac{KT}{q}\ln\left(\frac{N_D^-}{n_i}\right)$
Accumulation	Surface Potential, ψ_s	< 0	> 0
	Space charge density, $Q_{SC} = Q_{acc}$	$= Q_p \approx -C_{ox}V_G$	$= = Q_n \approx -C_{ox}V_G$
Depletion	Surface Potential, ψ_s	$-2\emptyset_b > \psi_s > 0$	$-2\emptyset_b < \psi_s < 0$
	Space charge density, Q_{SC}	$= Q_d = -\sqrt{2\epsilon_s q\, N_A\, \psi_s}$	$= Q_d = \sqrt{2\epsilon_s q\, N_D \lvert \psi_s \rvert}$
On-set of Inversion	Surface Potential, ψ_s	$= -2\emptyset_b = 2\lvert\emptyset_b\rvert$	$= -2\emptyset_b$
	Space charge density, Q_{SC}	$= Q_{d\,max} = -\sqrt{2\epsilon_s q\, N_A 2\lvert\emptyset_b\rvert}$	$= Q_{d\,max} = \sqrt{2\epsilon_s q\, N_D 2\emptyset_b}$
Strong Inversion	Surface Potential, ψ_s	$> -2\emptyset_b$ but[1] $\approx -2\emptyset_b$	$< -2\emptyset_b$ but[1] $\approx -2\emptyset_b$
	Space charge density, Q_{SC}	$= Q_n + Q_{d\,max}$	$= Q_p + Q_{d\,max}$
	Threshold Voltage, V_T	$= -\frac{Q_{d\,max}}{C_{ox}} + 2\lvert\emptyset_b\rvert$	$= -\frac{Q_{d\,max}}{C_{ox}} + 2\emptyset_b$
	Inversion charge density, Q_{inv}	$= Q_n = -C_{ox}(V_G - V_T)$	$= Q_p = -C_{ox}(V_G - V_T)$

1- Depletion Approximation

Non-Ideal MOS Device

Till now we discussed as ideal MOS Device. What do we mean by an ideal MOS device? An ideal MOS device is one in which there is no induced space charge in the semiconductor in the absence of the gate voltage. An non-ideal MOS device, on the other hand, is one in which there is a space-charge in the semiconductor even in the absence of a gate voltage. Real MOS device have a space charge in the semiconductor even when the gate voltage V_G is equal to zero. There are three causes for this non-ideal behavior, and they are:

1. Charges in the oxide
2. Difference in work function between the gate and the substrate
3. Surface state charges

We can apply a voltage on the gate that will reduce the induced space charge to zero, and we define that voltage as the **flat-band voltage, V_{FB}**. The reason for using the concept of the flat-band voltage is to describe the behavior of a non-ideal MOS device as though it were an ideal device with a bias voltage applied on the gate equal to the flat-band voltage V_{FB}. For example the threshold voltage of a non-ideal MOS device can be written as

$$V_{T\ non-ideal} = V_{T\ ideal} + V_{FB} \tag{5.66}$$

Actually, this relationship is not exactly valid due to the presence of surface states, but for all practical purposes this can be taken as valid. Let us now determine the flat-band voltage arising from the three different sources of non-ideal behavior.

Charges in the Oxide

Due to the manner in which the silicon atoms are bonded to the oxygen atoms in the growth of the oxide layer in a thermal oxidation process, the oxide layer contains some positive charge. The positive charges in the oxide induced a negative space charge in the semiconductor, and the magnitude of the induced charge will depend on the relative location of the charge in the oxide between the metal and the substrate. Let the positive charge in the oxide have a density ρ(x), and let it vary in some arbitrary fashion as shown in the Figure (5.11). In this figure we have chosen the origin, $(x = 0)$, at the oxide-silicon interface, and the oxide-gate interface at $x = -t_{ox}$ where t_{ox} is the thickness of the oxide layer. Consider the charge in the oxide in an elementary region between x and $x + dx$. This charge can be thought of as a sheet charge of density $\rho(x)dx$. According to the theory of images, the sheet charge $\rho(x)dx$, located at a distance x on the x-axis from the oxide-silicon interface, will induce charges both on the gate and in the semiconductor. The charge induced in the semiconductor is proportional to the distance between the sheet charge and the gate and similarly the charge induced in the gate is proportional to the distance between the sheet charge and the semiconductor. Hence it is possible to write the space charge induced in the semiconductor per unit area of the interface as

$$dQ_{SC} = -\rho(x)dx\frac{x-(-t_{ox})}{t_{ox}} = -\frac{\rho(x)\ (t_{ox}+x)\ dx}{t_{ox}} \tag{5.67}$$

In writing this equation, we must note that x is negative due to our choice of the origin of the x-axis and $\frac{1}{t_{ox}}$ can be shown to be the proportionality constant. The total induced space charge per unit area of the interface due to the entire positive charge distribution in the oxide is then given by

$$Q_{SC} = \int dQ_{SC} = -\int_{-t_{ox}}^{0} \frac{\rho(x)\ (t_{ox}+x)\ dx}{t_{ox}} \tag{5.68}$$

Let us now apply on the gate a voltage equal to $\frac{Q_{SC}}{C_{ox}}$. This voltage will give rise to a charge, Q_{SC} on the gate. According to Gauss' theorem, all the electric field lines emanating from the positive charge

in the oxide now terminate on the charges on the gate. The space-charge in the semiconductor becomes zero. Hence, the flat-band voltage V_{FB} can be expressed as the voltage necessary to apply on the gate such that the charge on the gate is Q_{SC}.

$$V_{FB} = \frac{Q_{SC}}{C_{ox}}$$

$$= -\frac{1}{C_{ox}t_{ox}} \int_{-t_{ox}}^{0} \rho(x)\,(t_{ox} + x)\,dx \tag{5.69}$$

$$= -\frac{1}{C_{ox}} \int_{-t_{ox}}^{0} \rho(x)\,(t_{ox} + x)\,dx \tag{5.70}$$

We define an effective sheet charge density in the oxide as Q_{ox}, which is equal to

$$Q_{ox} = -Q_{SC} = \frac{1}{t_{ox}} \int_{-t_{ox}}^{0} \rho(x)\,(t_{ox} + x)\,dx \tag{5.71}$$

Hence the flat-band voltage is then given by

$$V_{FB} = -\frac{Q_{SC}}{C_{ox}} \tag{5.71}$$

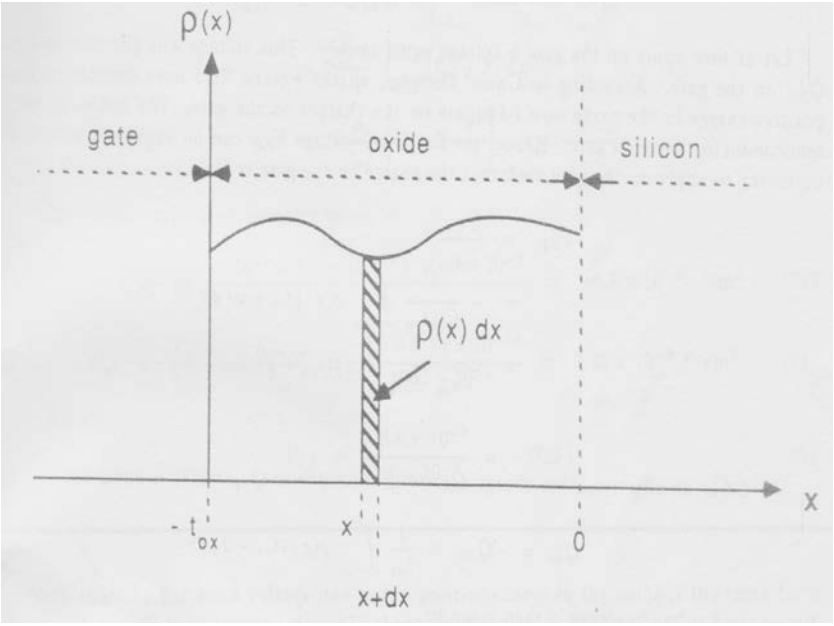


Figure (5.11): Am arbitrary charge density in the oxide.

Example

Let us now consider an example in which we assume that the oxide layer has a thickness of 1000 Å, and has a uniform charge density of 10^{15} ions/ cm^3. Since $\rho(x)$ is a constant, say equal to ρ_0, Q_{SC} (or Q_{ox}) can be evaluated taking ρ outside the integral.

$$Q_{SC} = -Q_{ox} = -\int_{-t_{ox}}^{0} \frac{\rho(x)\,(t_{ox}+x)}{t_{ox}}\,dx$$

$$= \frac{-\rho_0}{t_{ox}} \int_{-t_{ox}}^{0} (t_{ox} + x)\,dx$$

$$= -\frac{\rho_0}{t_{ox}} \left[t_{ox}x + \frac{x^2}{2} \right]_{-t_{ox}}^{0} = -\frac{\rho_0 t_{ox}}{2}$$

In this example, Q_{SC} becomes equal to

$$Q_{SC} = -\frac{10^{15} \times 1.6 \times 10^{-19} \times 1000 \times 10^{-8}}{2} = -0.8 \times 10^{-9}\ c/cm^2 \qquad (5.73)$$

$$C_{ox} = \frac{\epsilon_{ox}}{C_{ox}} = \frac{K_{ox}\epsilon_o}{t_{ox}} = \frac{3.9 \times 8.84 \times 10^{-14}}{1000 \times 10^{-8}} \approx 3.45 \times 10^{-8}\ F/cm^2 \qquad (5.74)$$

Hence

$$V_{FB} = \frac{-0.8 \times 10^{-9}}{3.4 \times 10^{-8}} = -0.023\ V \qquad (5.75)$$

Note that the flat-band voltage due to the positive charge in the oxide is the same for n- and p- type substrates. It is negative since the charge in the oxide is positive.

Work Function Difference

Every material has a characteristic work function. Work function is defined as the difference between the Fermi energy and the surface barrier energy, where the surface barrier energy is the energy that an electron inside the solid must have in order to leave the solid and escape into vacuum. The surface barrier energy therefore, is also called the *vacuum level*, E_{vac}. This is illustrated in Figure (5.12). We denote the work-function of the solid as $q\Phi$ and is given by

$$q\Phi = E_{\text{vac}} - E_{\text{F}} \qquad (5.76)$$

The student should note that Φ, although it is in units of volt, is not the electrostatic potential.

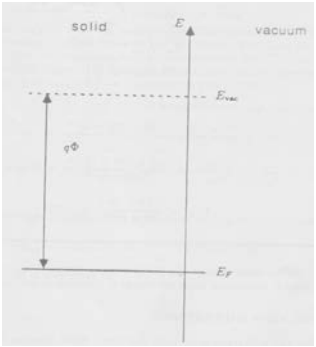


Figure (5.12): Work-function of a solid

Let us now consider two solids A and B with work functions $q\Phi_A$ and $q\Phi_B$ respectively. Let Φ_B be less than Φ_A. Let the two solids be not connected to each other as shown in Figure (5.13 A). Since the two solids are not connected to each other, the vacuum levels for the two solids are at the same level since an electron that is excited with enough energy to overcome the surface barrier energy will have zero kinetic energy in vacuum in both solids. The Fermi energy is at a higher level in B than in A since it has a smaller work-function. If the materials A and B are connected to form a junction as shown in Figure (5.13 B), electrons will move from material B into material A so that the Fermi levels are lined up. Hence B will be positive with respect to A. A potential difference, $\Phi_{\mathrm{AB}} = \Phi_{\mathrm{A}} - \Phi_{\mathrm{B}}$ will arise between B and A. This potential difference is called the contact potential.

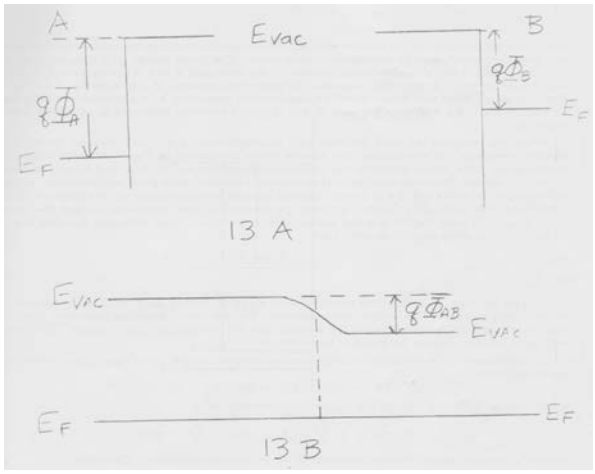


Figure (5.13): Two solids with different work-function: A): The solids are not connected to each other, and B): The solids are connected to each other.

When we consider the work function of a semiconductor, a similar definition follows. The vacuum level and the conduction and valence bands are shown in Figure (5.14). We notice that the location of the Fermi energy in the bandgap is dependent on the type and density of dopant (impurity) atoms. The work-function is therefore not a constant for a given semiconductor material and depends on whether it is n or p type material and also on the resistivity of the material. On the other hand, the difference in energy between the conduction band minimum, E_c , and the vacuum level, E_{vac}, is a constant for a given semiconductor material. This difference is called electron affinity, and denoting affinity as $\boldsymbol{\chi}$ we have

$$\chi = \frac{E_{vac} - E_c}{q} \tag{5.77}$$

In a pn junction, n- and p-type semiconductor materials form a junction. This is equivalent to connecting a p-type and an n-type semiconductor. We saw that a potential difference, called the *built-in voltage*, V_{bi}, exists between the n-region and the p-region in a pn junction. The built-in voltage is just the contact potential as shown below:

Let us denote the work-function of the p-type semiconductor Φ_1, and that of the n-type semiconductor as Φ_2.

$$q\Phi_1 = E_{vac} - E_{F1} = q\chi + E_c - E_{F1} \tag{5.78}$$

Where E_{F1} is the Fermi level in the p-type semiconductor. But

$$E_c - E_{F1} = \frac{E_g}{2} + E_i - E_{F1} = \frac{E_g}{2} + q\left|\emptyset_p\right| \tag{5.79}$$

Where $\emptyset_p$ is the bulk electrostatic potential of the p-type semiconductor. Therefore

$$q\emptyset_1 = q\chi + \frac{E_g}{2} + q\left|\emptyset_p\right| \tag{5.80}$$

Similarly, it can be shown that

$$q\emptyset_2 = q\chi + \frac{E_g}{2} - q\,\emptyset_n \tag{5.81}$$

where $\emptyset_n$ is the bulk electrostatic potential of the n-region.

Hence the work-function difference, $\Phi_{12} = \Phi_1 - \Phi_2$ is equal to

$$\Phi_{12} = \Phi_1 - \Phi_2 = \left|\emptyset_p\right| + \emptyset_n = V_{bi} \tag{5.82}$$

In an MOS device when there is a difference between the gate and the semiconductor work function, a potential difference arises. Let Φ_M be the work function of the gate and Φ_S be the work function of the substrate semiconductor. Therefore the substrate is positive with respect to the gate by an amount equal to the contact potential Φ_{MS} where $\Phi_{MS} = \Phi_M - \Phi_S$. This will therefore give rise to a space charge in the semiconductor. If we now apply a voltage Φ_{MS}, it will exactly balance out the contact potential difference, and hence no space charge will be induced in the substrate. We can therefore conclude that the flat-band voltage V_{FB} required to account for the work function difference is given by

$$\mathrm{V}_{FB} = \Phi_{MS} = \Phi_M - \Phi_S \tag{5.83}$$

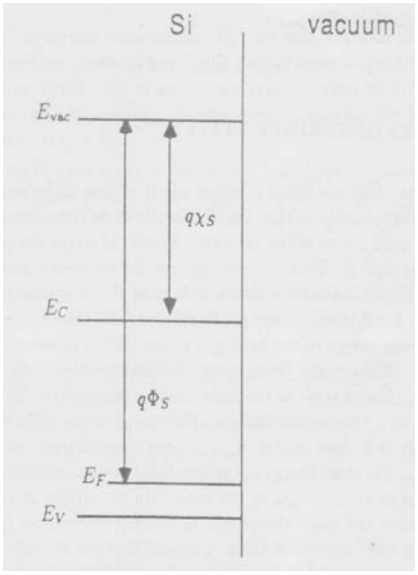


Figure (5.14): Electron affinity and work function in a semiconductor

Surface States (Interface States)

At the interface between the oxide and the substrate, there are a large number of states due to unfilled covalent bonds which are called dangling bonds. These states can be charged or neutral. These are called interface states or surface states. Usually these states are distributed continuously in energy in the band gap. Let us define D_{it} as the number of states per unit area per unit energy interval at some energy level E. Then in an energy level dE between E and $E + dE$, the number of states per unit area is $D_{it}dE$. A surface state will be more positively charged if the Fermi energy is below it. On the other hand, it will be less positively charged if the Fermi energy is above it. Thus, the location of the Fermi energy in the band gap at the surface determines the amount of charge in the surface states. Although the Fermi energy remains constant with distance, its location in the band gap at the surface changes as the band bending changes (i.e., as ψ_s changes). Hence the amount of charge in the surface states changes. The charge in the surface states is located at the interface, and therefore is a sheet charge. If Q_{st} is the sheet charge density at the surface, then Q_{st} is a function of ψ_s. The sheet charge at the interface between the semiconductor and the oxide induces a space charge of density $-Q_{st}$ in the semiconductor. Hence if we apply a charge on the gate equal to $-Q_{st}$, then the space charge will be completely removed from the semiconductor. This is again based on the principle of Gauss' theorem that the electric lines of force emanating from the surface state sheet charge Q_{st} now will terminate on the sheet charge $-Q_{st}$ on the gate. Hence we can write the flat band voltage V_{FB} as equal to

$$V_{FB} = \frac{-Q_{st}(\psi_s=0)}{C_{ox}} \tag{5.84}$$

The total flat band voltage due to all the three sources is

$$V_{FB} = -\frac{Q_{ox}}{C_{ox}} + \Phi_{MS} - \frac{Q_{st}(\psi_s = 0)}{C_{ox}} \tag{5.85}$$

On the other hand, the threshold voltage V_T of a non-ideal MOS structure is

$$\begin{aligned} \mathrm{V}_{T\,non-ideal} &= \mathrm{V}_{T\,ideal} - \frac{Q_{ox}}{C_{ox}} + \Phi_{MS} - \frac{Q_{st}(\psi_s = 2|\emptyset_b|)}{C_{ox}} \\ &= \mathrm{V}_{T\,ideal} + V_{FB} - \frac{\Delta Q_{st}}{C_{ox}} \end{aligned} \tag{5.86}$$

Where

$$\Delta Q_{st} = Q_{st}(|\psi_s| = 2|\emptyset_b|) - Q_{st}(\psi_s = 0) \tag{5.87}$$

In most modern devices, ΔQ_{st} is very small, and hence it is reasonable to express the threshold voltage as

$$\mathrm{V}_T = \mathrm{V}_{T\,ideal} + \mathrm{V}_{FB} \tag{5.88}$$

and neglect the $-\frac{\Delta Q_{st}}{C_{ox}}$ term.

Example

Let us assume that the surface state density, D_{it} , is a constant with energy (i.e., it has the same value throughout the band-gap) with a value equal to $5 \times 10^{10}\ eV^{-1}cm^{-2}$. Let the oxide thickness be 1000 Å. Let us assume a p-type substrate with $N_A = 5 \times 10^{15}/cm^3$. Let us calculate the value for $-\frac{\Delta Q_{st}}{C_{ox}}$. The change in ψ_s between flat-band and on-set of inversion is equal to $\Delta\,\psi_s = 2|\emptyset_b|$.

$$\emptyset_b = -\frac{kT}{q} ln\left(\frac{N_A}{n_i}\right) = -0.0259 \ln\left(\frac{5\times10^{15}}{5\times10^{10}}\right) = -0.298$$

$$\Delta\,\psi_s = 2 \times 0.298 = 0.596\ V$$

ΔQ_{st} is negative, since the bands bend downward at inversion (Fermi energy rises in the band gap at the surface).

$$\Delta Q_{st} = -q \int_{\psi_s=0}^{\psi_s=2|\emptyset_b|} N_{it}\, d\psi = -q\, N_{it}\, \Delta\,\psi_s$$

$$= -1.6 \times 10^{-19} \times 5 \times 10^{10} \times 0.596$$

$$= -4.77 \times 10^{-9}\ c\ cm^{-2}$$

Therefore

$$-\frac{\Delta Q_{st}}{C_{ox}} = \frac{4.77 \times 10^{-9}}{3.45 \times 10^{-8}} = 0.14\ V$$

In modern semiconductor devices, the interface state density is lower by a factor of 10, and the oxide thickness is of the order of 200 Å. Hence $-\frac{\Delta Q_{st}}{C_{ox}}$ is a couple of millivollts, and hence negligible.

Capacitance Effect

In a parallel plate capacitor, which is formed by having a dielectric medium between two parallel metal plates, equal sheet charges but of opposite sign are induced on the two plates when a voltage is applied across this capacitor. The charge on the metal plate which is connected to the positive terminal of the voltage source has a positive sheet charge and that which is connected to the negative terminal has a negative sheet charge. The capacitance is determined only by the geometrical structure of the parallel plate capacitor. The capacitance is inversely proportional to the thickness of the dielectric material between the plates and proportional to the area of the plate. On the other hand, when an MOS capacitor is fabricated with silicon (the substrate) as one of the two parallel plates, the induced charge in the semiconductor is not a sheet charge but a space charge which under certain specific conditions can be approximated as a sheet charge as discussed earlier. The magnitude of the space charge is not directly proportional to the potential drop across the space charge region, (the surface potential). Hence, the capacitance in an MOS capacitor is voltage dependent. In this section we will examine the voltage dependence of the capacitance of an MOS capacitor. Let us first consider an ideal MOS capacitor.

Ideal MOS Capacitor

Let us now assume that a DC voltage V_G is applied on the gate of an MOS capacitor so as to bias the device in one of the three regimes of operation, i.e., accumulation, depletion or inversion. The applied voltage is divided into a voltage, V_{ox}, across the oxide layer and a potential drop, ψ_s, across the space charge region. Let us assume that additionally a step voltage of a very small amplitude dV_G, is applied in series with the DC voltage V_G, as shown in Figure (5.15). The applied voltage dV_G is also distributed partially across the oxide and partially across the space charge region, giving us the relationship

$$dV_G = dV_{ox} + d\psi_s \tag{5.89}$$

where dV_{ox} and $d\psi_s$, are the incremental changes in V_{ox} and ψ_s respectively. dV_G gives rise to a small sheet charge dQ on the gate electrode in addition to the already existing sheet charge due to the bias voltage V_G. Similarly, the space charge density in the semiconductor increases by dQ_{SC}

where

$$dQ_{SC} = -dQ \tag{5.90}$$

We divide the above equation for dV_G by $|dQ_{SC}|$ to obtain

$$\frac{\mathrm{dV}_G}{dQ} = \frac{\mathrm{dV}_{ox}}{|dQ_{SC}|} + \frac{d\psi_s}{|dQ_{SC}|} \tag{5.91}$$

The differential of the gate (sheet) charge Q, with respect to V_G, given by $\frac{dQ}{\mathrm{dV}_G}$, denotes the ratio of the incremental change dQ in the gate charge Q to the incremental change dV_G in the gate voltage, and is defined as the small signal capacitance C of the MOS capacitor of gate area equal to unity.

$$\frac{\mathrm{dV}_G}{dQ} = \frac{1}{\frac{dQ}{\mathrm{dV}_G}} = \frac{1}{C} \tag{5.92}$$

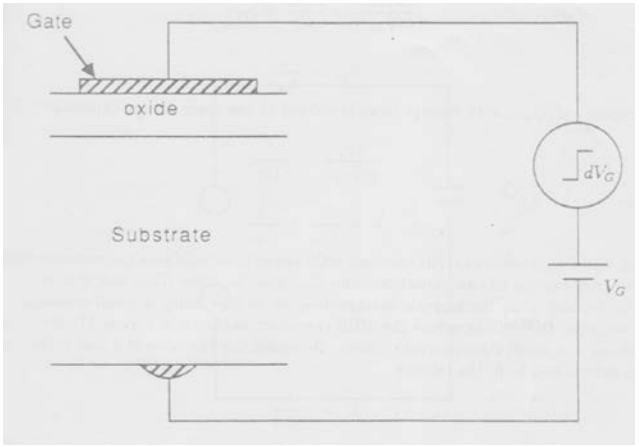


Figure (5.15): A small step voltage in series with a DC bias voltage V_G applied across an MOS capacitor

The differential of Q with respect to V_{ox} similarly is equal to the capacitance of a parallel plate capacitor with the oxide layer between two metallic plates. Denoting this as the oxide capacitance per unit area, $\mathrm{C_{ox}}$, we obtain

$$\frac{\mathrm{dV}_{ox}}{|dQ_{SC}|} = \frac{\mathrm{dV}_{ox}}{dQ} = \frac{1}{C_{ox}} \tag{5.93}$$

where

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{5.94}$$

The differential of $|dQ_{SC}|$ with respect to ψ_s is defined as the space charge capacitance C_{SC} and is given by

$$\frac{d\psi_s}{|dQ_{SC}|} = \frac{1}{C_{SC}} \tag{5.95}$$

Therefore

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{SC}} \tag{5.96}$$

The small signal equivalent capacitance of an MOS capacitor of unit area can therefore be described as a series combination of two capacitors, one C_{ox}, and the other C_{SC}, and this is illustrated in Figure (5.16). Instead of an incremental voltage dV_G, let us now apply a small sinusoidal voltage $\tilde{v}$ in series with the DC voltage across the MOS capacitor, as shown in Figure (5.17). Due to the small signal voltage $\tilde{v}$, a small signal current $\tilde{\imath}$ flows. By measuring the ratio of $\tilde{v}$ and $\tilde{\imath}$, the capacitance C can be determined from the relation

$$C = \frac{1}{j\omega\,\tilde{v}} \tag{5.97}$$

where ω is 2π time the frequency of the small signal voltage $\tilde{v}$. If $\tilde{Q}$ is the component of space charge that varies sinusoidally in step with $\tilde{v}$, then

$$C = \frac{\tilde{Q}}{\tilde{v}} \tag{5.98}$$

The measured capacitance C is

$$C = \frac{C_{ox}\,C_{SC}}{C_{ox} + C_{SC}} \tag{5.99}$$

where the area of the gate is assumed to be unity, C_{ox} is the oxide capacitance per unit area, and C_{SC} is equal to space charge capacitance per unit area. If we were to consider a MOS capacitor of area A, then the measured capacitance, C_m is equal to $A \times C$ where C is the capacitance per unit area.

$$\tilde{Q} = \tilde{Q}_{acc} \tag{5.100}$$

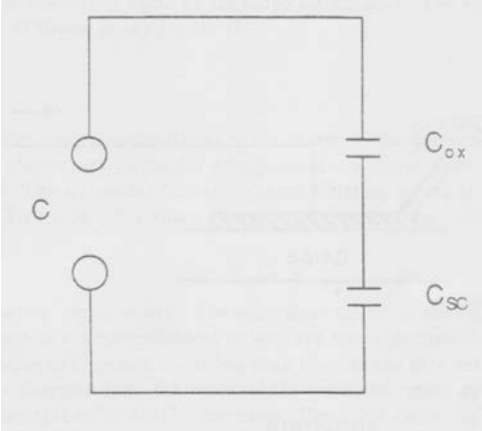


Figure (5.16) Equivalent circuit of an MOS capacitor

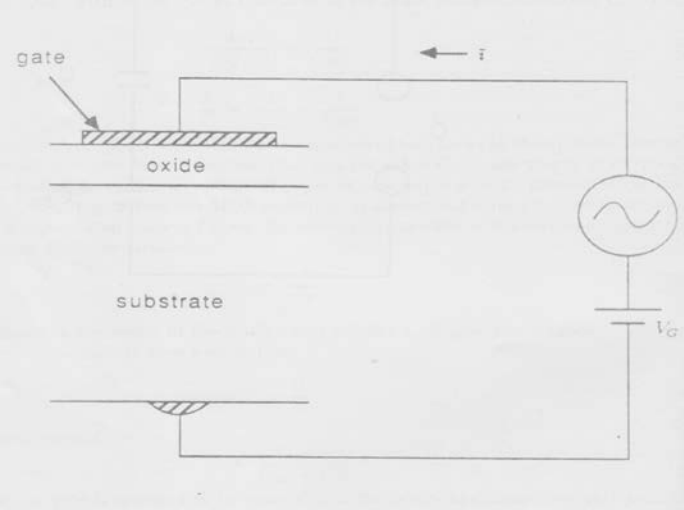


Figure (5.17): A sinusoidal voltage in series with a DC bias voltage applied across an MOS capacitor

Let us now evaluate the small signal capacitance for different regimes.

Accumulation

The charge distribution in the MOS device under accumulation conditions can be visualized as shown in Figure (5.18). Let us assume that $x = 0$ represents the interface between the oxide and silicon. Let the interface between the gate and the oxide be located a $x = -t_{ox}$.

The accumulation charge is located at $x = 0$ as shown in Figure (5.18). $\tilde{Q}_{acc}$ is a small component of the total sheet charge which varies in step with the applied small signal voltage $\tilde{v}$.

$$C_{SC} = \frac{|dQ_{SC}|}{d\psi_s} = \frac{dQ_{acc}}{d\psi_s} \tag{5.101}$$

Since Q_{acc} changes enormously for a small change in ψ_s, C_{SC} is very large. Hence

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{SC}} \approx \frac{1}{C} = \frac{1}{C_{ox}} = \frac{t_{ox}}{\epsilon_{ox}} \tag{5.102}$$

The above equation shows that the measured capacitance of the MOS capacitor in the accumulation regime of operation is essentially equal to the oxide capacitance. The equivalent circuit is just a simple capacitor C_{ox} as shown in Figure (5.19)

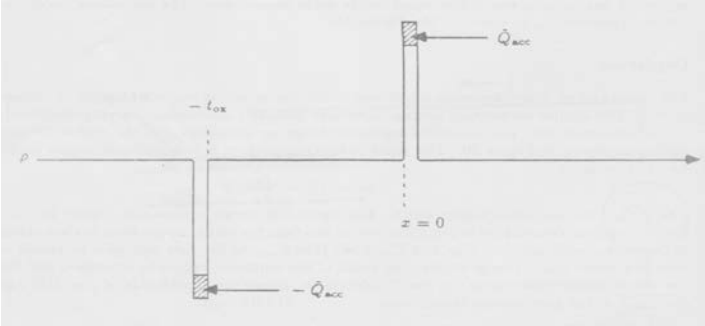


Figure (5.18): Charge distribution in the MOS capacitor under accumulation conditions.

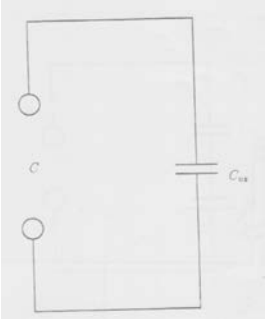


Figure (5.19): Equivalent circuit of an MOS capacitor when the surface is accumulated. The MOS capacitor has a capacitance equal to C_{ox} in accumulation.

Depletion

The space charge under depletion conditions is the charge in the depletion region in the semiconductor. The applied elementary sinusoidal voltage induces a sinusoidally varying depletion charge in the semiconductor. The sinusoidal depletion charge variation occurs at the edge of the depletion region, as shown in Figure (5.20) The space charge capacitance is the depletion region capacitance which is equal to

$$C_{SC} = C_d = \frac{\epsilon_s}{x_d} \tag{5.103}$$

where x_d is the depletion region width. The equivalent circuit is shown in Figure (5.21). Since the depletion region capacitance is comparable to or less than the oxide capacitance, the net capacitance of the series combination of C_{SC} and C_{ox} is less tha C_{ox}. As the gate voltage is increased so as to induce a larger space charge region, the width of the depletion region x_d increases, and therefore the space charge capacitance $C_{SC}\ (= C_d)$ decreases. Hence the capacitance of the MOS capacitor decreases as the gate voltage is increased.

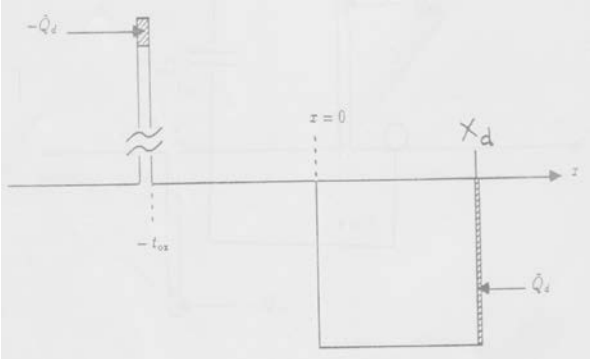


Figure (5.20): Charge distribution in the MOS capacitor under depletion conditions.

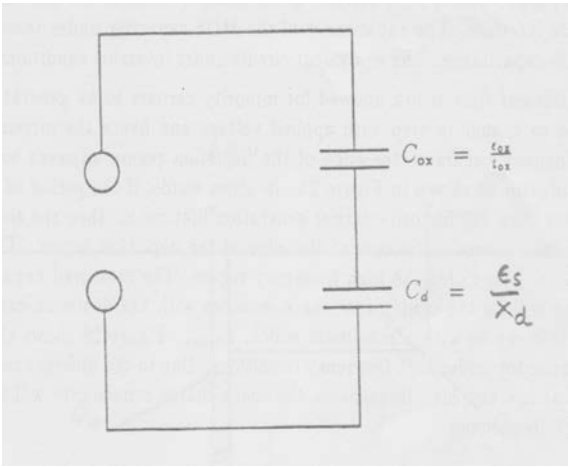


Figure (5.21): Equivalent circuit of an MOS capacitor when the surface is depleted.

Inversion

When the gate voltage is increased to a value larger than the threshold voltage, we saw earlier that the space charge is due to both the inversion charge and the depletion charge. Due to the applied voltage dV_G, an increase in the inversion charge occurs if minority carriers can be generated fast enough in the depletion region. Then the inversion charge has a component that varies sinusoidally in step with the small signal voltage $\tilde{v}$. For the inversion charge to change in step with applied voltage, the sinusoidal voltage should change slowly with time. The period of oscillation of the sinusoidal voltage $\tilde{v}$ has to be much larger than the minority carrier generation lifetime τ_g (i.e., $\omega < \frac{1}{\tau_g}$). The sinusoidal change in the space charge density therefore is equal to

$$\tilde{Q}_{SC} = \tilde{Q}_{inv}$$

and occurs at the silicon surface as shown in Figure (5.22). The range of frequencies for which $\omega < \frac{1}{\tau_g}$ is called the low frequency region.

$$C_{SC} = \frac{dQ_{SC}}{d\psi_s} = \frac{dQ_{inv}}{d\psi_s}$$

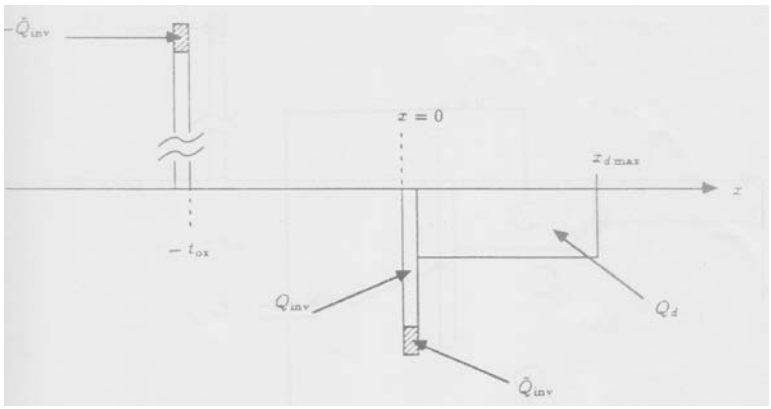


Figure (5.22): Low frequency charger distribution in the MOS capacitor under inversion conditions.

Hence it is large since the inversion charge increases enormously for a small change in ψ_s. At low frequencies the MOS device behaves like a parallel plate capacitor with an oxide dielectric between the two parallel plates since the sinusoidal change occur on the gate and in the semiconductor at the oxide interface. The capacitance of the MOS capacitor under these conditions is the same as the oxide capacitance. The equivalent circuit under inversion conditions is shown in Figure (5.23).

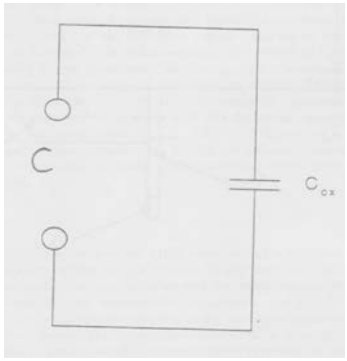


Figure (5.23): Low frequency equivalent circuit of an MOS capacitor under inversion conditions

If sufficient time is not allowed for minority carriers to be generated, the inversion charge is not able to change in step with applied voltage and hence the incremental change in the space charge density occurs at the edge of the depletion region adjacent to the neutral region of the semiconductor as shown in Figure (5.24). In other words, if the period of the sinusoidal voltage $\tilde{v}$ is much less than the minority carrier generation lifetime τ_g, then the elementary sinusoidal change in the space charge dQ occurs at the edge of the depletion region. The range of frequencies for which $\omega > \frac{1}{\tau_g}$ is called the high frequency region. The measured capacitance at high frequencies therefore will be the oxide capacitance in series with the depletion capacitance corresponding to a depletion region with a maximum width, $x_{d\,max}$. Figure (5.25) shows the equivalent circuit of the MOS capacitor under high frequency conditions. Due to the different responses of the space charge density at low and high frequencies, the space charge capacitance will have different values at low and high frequencies.

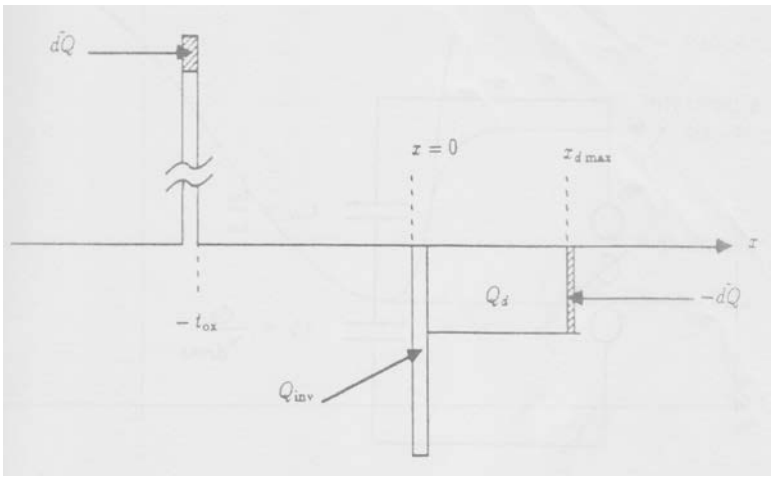


Figure 5.24 High frequency charge distribution in an MOS capacitor under inversion conditions.

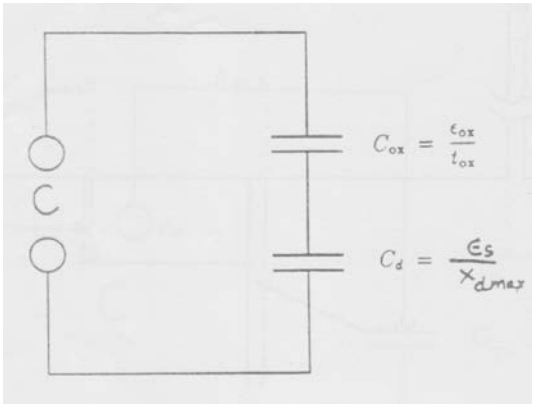


Figure 5.25 High frequency equivalent circuit of an MOS capacitor under inversion conditions.

C-V Curve

The plot of a small signal capacitance of an MOS capacitor as a function of the DC bias (gate) voltage is called a capacitance-voltage (C-V) curve. While the C-V curve will be the same for both high and low frequencies in the accumulation and depletion regions, there will be a difference in the inversion region behavior. At intermediate frequencies, the C-V curve will lie between the high frequency and the low frequency curves, under inversion conditions. This is illustrated in Figure (5.26). C_{min} is the value of the capacitance measured in inversion under high frequency conditions, C_{ox} the oxide capacitance. At intermediate frequencies, the capacitance measured in inversion region has a value intermediate between C_{min} and C_{ox}.

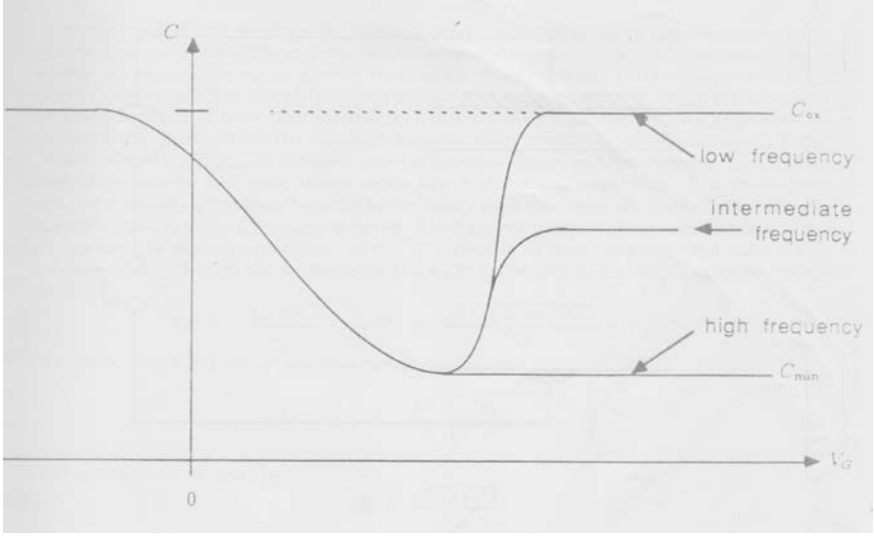


Figure (5.26): The capacitance-voltage (C-V) characteristics of an MOS capacitor at low, intermediate and high measurement frequencies.

Deep Depletion

Let us now assume that a step voltage of amplitude V_G ($> dV_T$) is applied on the gate at time $t = 0$. Let us also apply a small sinusoidal voltage $\tilde{v}$ in series with the step voltage V_G as shown in Figure (5.27). The purpose of $\tilde{v}$ is to measure the small signal capacitance, and it is therefore called the probing signal. Immediately after the application of the step voltage, that is at $t = 0^+$, the induced space charge will essentially be due to the depletion charge. This is because there is no time for the inversion charge to be generated by thermal generation of minority carriers. The depletion charge is increased beyond the maximum depletion charge obtained under steady state conditions. The space charge region under this condition represents the non-equilibrium condition. At $t = 0^+$, the depletion region widens to a large value x_{d0^+} which is much larger than $x_{d\,max}$. Hence, the condition is called "*deep depletion*." The generation and recombination rates in the space charge are not what wll be obtained in thermal equilibrium, and do not balance each other out as required by the law of mass action. Minority carriers are generated in the space charge region in excess of what are lost due to recombination. These minority carriers will be driven towards the interface, where they will form the inversion charge, while the majority carriers proceed in the opposite direction. Part of the majority carriers generated will neutralize the ionized impurities at the edge of the depletion region, thereby reducing the depletion region width while the rest of the majority carriers will be flown out of the ohmic contact. It must be pointed out that the increase in the inversion charge is more than the reduction in the depletion charge. As time progresses, the inversion charge increases and the depletion charge decreases correspondingly. Ultimately at $t = \infty$, a depletion width of $x_{d\,max}$ is obtained, and the inversion charge and the depletion charge attain steady state values. The space charge region now is in thermal equilibrium. The space-charge region thus exhibits a transient behavior before steady state conditions are reached. Therefore the capacitance also exhibits a transient behavior. The transient response of the capacitance is shown in Figure (5.28). In deep depletion, (i.e., at $t = 0^+$), there is no inversion charge and there is

only depletion charge. Hence, the surface potential $\psi_s(0^+)$ is related to the step voltage V_G by the relationship

$$V_G = -\frac{Q_d(0^+)}{C_{ox}} + \psi_s(0^+) = \frac{\sqrt{2\epsilon_s q N_A \psi_s(0^+)}}{C_{ox}} + \psi_s(0^+) \tag{5.104}$$

This can be solved to yield, as was done earlier,

$$\psi_s(0^+) = V_G - \frac{K^2}{2}\left(\sqrt{\frac{4V_G}{K^2} + 1} - 1\right) \tag{5.105}$$

where K is defined as equal to

$$K = \frac{\sqrt{2\epsilon_s q N_A}}{C_{ox}} \tag{5.106}$$

The band-bending at $t = 0^+$ is shown in Figure (5.29). The surface potential ψ_s represents a potential well at $x = 0$ for an electron (remember we are assuming a p-type substrate), of depth equal to $\psi_s(0^+)$. This potential well can be used to store electrons. This storage can be done only momentarily due to the transient nature of deep depletion. By placing a large number of MOS capacitors at close spacing and by pulsing these in a suitable sequence, the charge stored in the potential well under these capacitors can be successfully transferred from one to the next all the way to the end of the string of capacitors. The string of closely spaced MOS capacitors is called the charge couple device (CCD). The charge coupled device has many applications in digital, analog and imaging circuits.

The capacitance at $t = 0^+$ is very small since the depletion region width is large and hence C_d at $t = 0^+$ is small. Hence C is also small. As time increases from $t = 0^+$, ψ_s decreases and the depletion region width decreases and C_d increases. Therefore the capacitance of the MOS capacitor also increases and finally reaches a steady state value corresponding to $x_d = x_{d\,max}$ after a long time. This is called *the transient behavior of the MOS capacitor*.

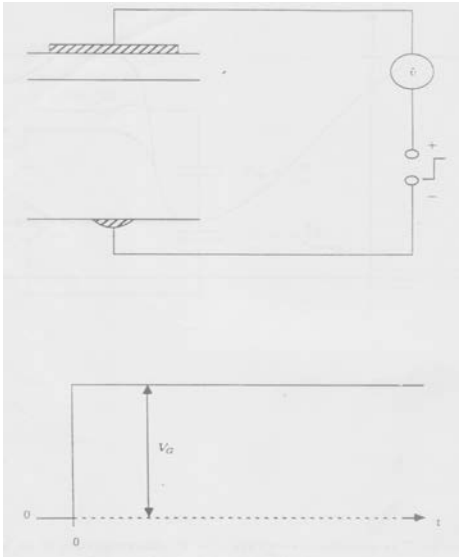


Figure 5.27: Application of a large step voltage on the gate of an MOS capacitor to produce deep depletion.

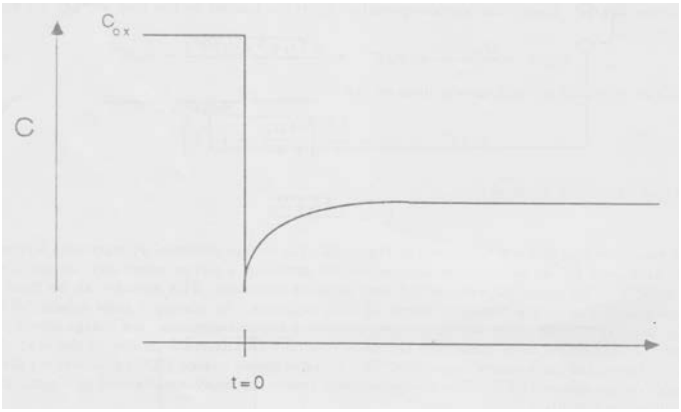


Figure 5.28 Transient behavior of the capacitance of an MOS capacitor.

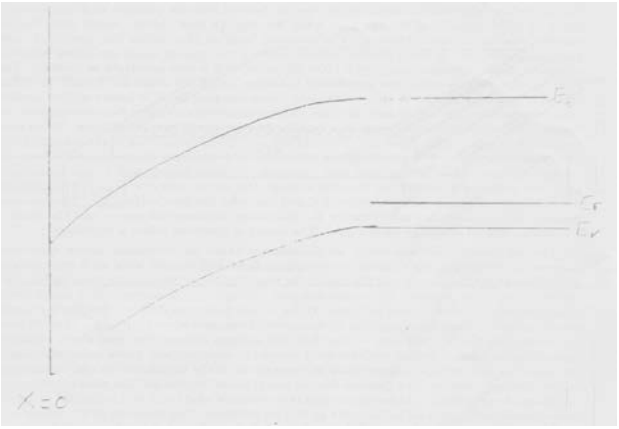


Figure 5.29 Band-bending in deep depletion.

MOS Transistors

The MOS transistor is a device which is based on the principle that a conducting channel is induced in the surface of a semiconductor by an electric field of suitable polarity and magnitude. In our study of MOS capacitors, we saw that when we apply a gate voltage larger than the threshold voltage, the surface gets inverted. The inversion layer at the surface can provide a conducting path, and therefore can carry electric current between two regions which are doped opposite to the substrate. As an example, consider a MOS structure on a p-type substrate as shown in Figure (5.30). Two n^+ regions, one called the source and ther other called the drain are formed by diffusing (or implanting) donor atoms. A thin oxide layer called the gate oxide is grown on the surface of the substrate between the source and the drain. A gate electrode extending from the edge of the source region to the edge of the drain region is deposited on top of the gate oxide layer. When the voltage applied on the gate is less than the threshold voltage V_T, no conducting channel exists between the source and the drain. The path between the source and the drain comprises two n^+- p junctions connected back-to-back. When the

gate voltage V_G is made larger than V_T, an inversion layer of electrons provides a conducting channel between the source and the drain. The conductance is dependent on the inversion charge density, and hence on the gate voltage V_G. Thus we are able to control the conductance (and therefore the resistance) between the source and the drain through the application of a voltage on the gate. The device is therefore called a transistor.

The particular structure which we discussed is called an n-channel device since the charge carriers in the conducting channel are electrons. If we had started with an n-type substrate and p^+ diffusions for the source and the drain, the resulting device would have been called a p-channel device. The drain current, $\mathrm{I_D}$, versus the gate voltage, $\mathrm{V_G}$, characteristic is called the transfer characteristic, and it is shown in Figure (5.31) for n - and p- channel devices. For the n-channel device when the gate voltage V_G is less than V_T no current flows, and when V_G is larger than V_T the current increases with V_G. We will later on see that this increase is linear for small drain voltages. In the case of the p- channel device, the threshold voltage is negative, and hence when the magnitude of the negative gate voltage is larger than the magnitude of the threshold voltage, conduction results in the p-channel device. The symbols for the two types of devices are also shown in Figure (5.31). The direction of the arrow in the substrate lead is to indicate whether it is a p- channel or an n- channel device, and the convention is the same as in a $p\ n$ junction. The direction of the arrow is from the p to the n-region. In the case of the n-channel device, the substrate is p and the channel is n. Therefore the arrow is inward. In the case of the p-channel, the substrate is n-type and the channel is p-type, and therefore the arrow is outward.

Let us now calculate the drain current. Assume that the length of the channel between the source and the drain lies along the y axis with the source at $y = 0$, and the drain at $y = L$. Let W be the width of the channel. Then the current, J, in the channel is given by

$$J = \mu_n n \frac{dE_F}{dy} \tag{5.107}$$

where μ_n is the mobility of the electrons, n is the electron density, and E_F is the Fermi energy. The current density J is in units of amperes per square centimeter. Let us assume that a voltage V_D between the source and the drain. At any point, the voltage drop along the channel measured with respect to the source will be denoted $V(y)$. The gradient of the voltage along the channel is $\frac{dV}{dy}$. The gradient of Fermi energy is related to the gradient of the voltage drop along the channel through the relationship

$$\frac{dE_F}{dy} = -q \frac{dV}{dy} \tag{5.108}$$

We can therefore write the current density as equal to

$$J = -q\mu_n n \frac{dV}{dy} \tag{5.109}$$

Let x_I be the thickness of the inversion layer of electrons. Furthermore, let us assume that the electron density, n, is constant in the inversion layer perpendicular to the channel. Then the drain current I_D is equal to the current density, J, multiplied by W x_I , and is given by

$$I_D = -q\mu_n n W x_I \frac{dV}{dy} \tag{5.110}$$

Consider a small section of the inversion layer a shown in Figure (5.32). Let us now describe a rectangular strip of length dy, contained between y and $y + dy$. The charge contained in this rectangular strip is equal to $Q_n W\ dy$ since Q_n is the electron (inversion) charge per unit area of the surface. But this is also equal to

$$Q_n W\ dy\ = - q\ n\ W x_I\ dy \tag{5.111}$$

Hence

$$Q_n\ = - q\ n\ x_I \tag{5.112}$$

Using this expression for Q_n we can express the drain current as

$$I_D = \ \mu_n\ Q_n\ W \frac{dV}{dy} \tag{5.113}$$

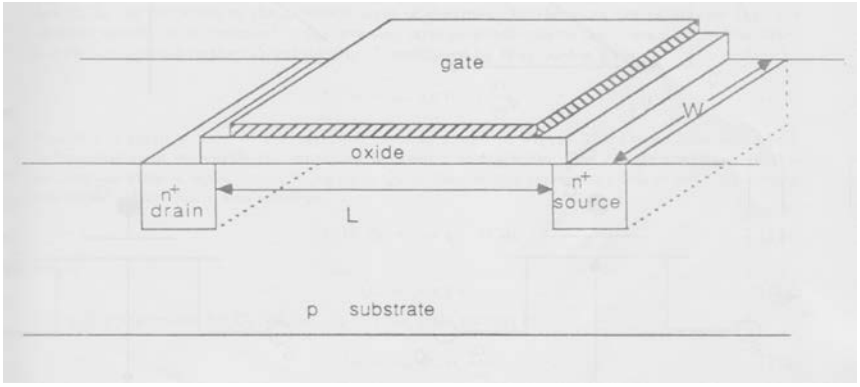


Figure (5.30) A schematic representation of the n-channel MOS transistor.

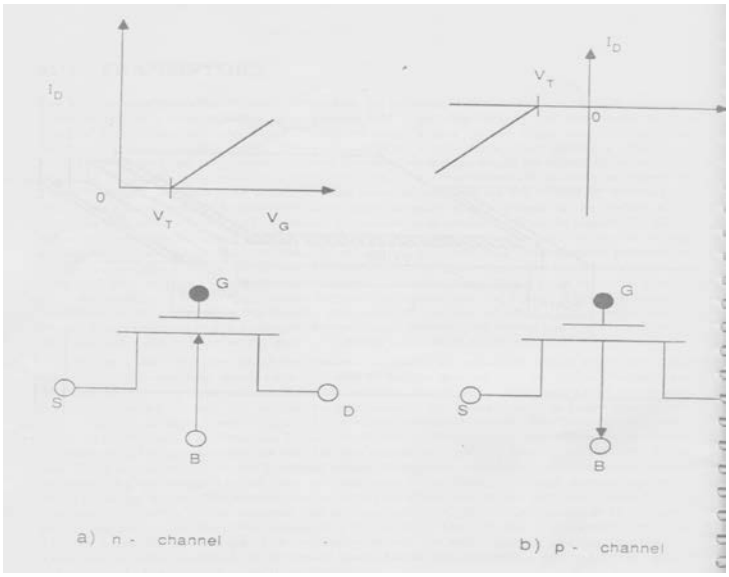


Figure (5-31) The transfer characteristics and the symbol for a) n-channel device and b) p-channel device.

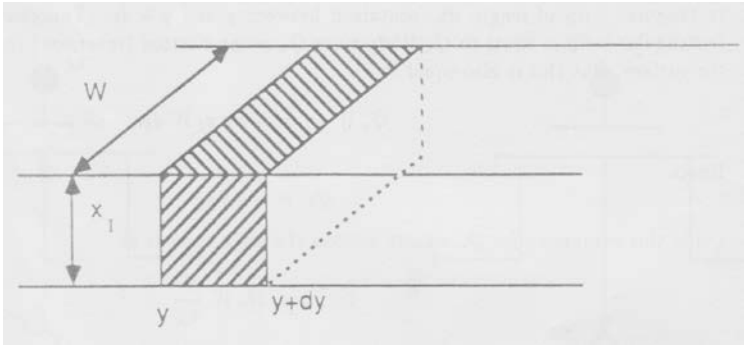


Figure (5-32) A section of the inversion layer between y and $y + dy$

We will now distinguish three cases:

Case 1: Small V_D

Let the source and the substrate be connected together to the ground. We will assume that the drain voltage V_D is small. Therefore it is reasonable to assume that the inversion charge density is the same everywhere in the channel i.e., does not vary with y. The inversion charge density Q_{inv} is given by

$$Q_{inv} = Q_n = -C_{ox}(V_G - V_T) \tag{5.114}$$

where C_{ox} is the oxide capacitance per unit area. The inversion layer now looks like a rectangular sheet of electron charge of dimensions W, L and x_I as shown in Figure (5.33). Therefore the drain current is given by

$$I_D = -\mu_n W C_{ox}(V_G - V_T)\frac{dV}{dy} \tag{5.115}$$

We can now approximate the gradient $\frac{dV}{dy}$ as equal to $\frac{V_D}{L}$ since the drain voltage V_D is very small.

Hence

$$I_D = -\mu_n \frac{W}{L} C_{ox}(V_G - V_T) V_D \tag{5.116}$$

We see therefore that the drain current varies linearly with the drain voltage and also linearly with the gate voltage.

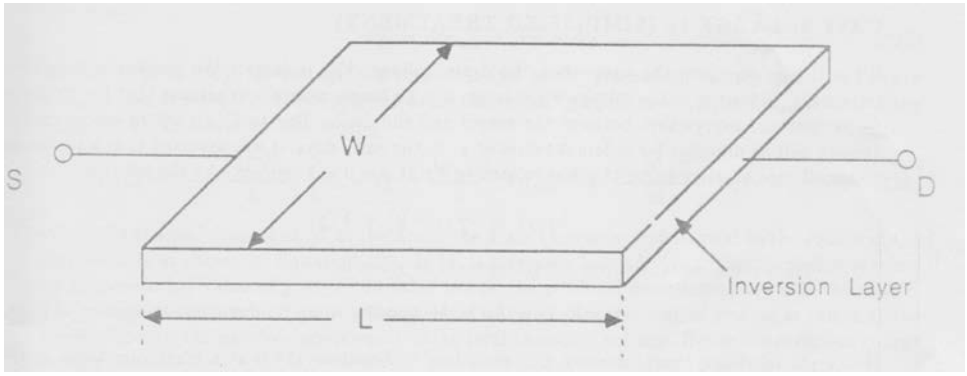


Figure (5-33) Approximation of the inversion layer as a sheet charge.

Case 2: Large V_D (Simplified Treatment)

We will now consider the case when the drain voltage, V_D is large. We present a simplified treatment. Since the drain voltage V_D is large, it is no longer possible to assume that the gradient $\frac{dV}{dy}$ is constant everywhere between the source and the drain. Due to $V(y)$, the inversion charge density will be different for different values of y. In the early days, it was assumed that a threshold voltage $V_T(y)$ at some value of y was related to V_T at $y = 0$ at the source by the relation

$$V_T(y) = V_T(y = 0) + V(y) \tag{5.117}$$

At the drain therefore, the threshold voltage will be given by

$$V_T(y = L) = V_T(y = 0) + V_D \tag{5.118}$$

Hence the inversion charge density, Q_n, according to Equation (5.114) is at a maximum value at the source, decreases along the channel, and is at minimum value at the drain. Another way of looking at this is that the electric field between the gate and the substrate is large at the source and decreases as one proceeds towards the drain. Hence the induced inversion charge gets smaller as you approach the drain. The inversion charge density $Q_n(y)$, at some point y in the channel is therefore given by

$$Q_n(y) = -C_{ox}[V_G - V_T(y)] = -C_{ox}[V_G - V_T(y = 0) - V(y)] \tag{5.119}$$

Substituting this value of Q_n in the expression for drain current given before, we get

$$I_D = -\mu_n W C_{ox}[V_G - V_T(y = 0) - V(y)]\ \frac{dV}{dy} \tag{5.120}$$

Multiplying both sides of the above equation by dy and integrating, we obtain

$$I_D L = -\int_0^{V_D} \mu_n W C_{ox}[V_G - V_T(y = 0) - V(y)]dV$$

$$= -\mu_n W C_{ox}\left[V_G - V_T(y = 0) - \frac{V_D}{2}\right]V_D \tag{5.121}$$

Dividing both sides of the equation by L, the drain current is obtained as

$$I_D = -\frac{\mu_n\ WC_{ox}}{L}\left[V_G - V_T(y = 0) - \frac{V_D}{2}\right]V_D \tag{5.122}$$

Writing the threshold voltage at the source $y = 0$ as V_T , we obtain I_D as equal to

$$I_D = -\frac{\mu_n\ WC_{ox}}{L}\left[V_G - V_T - \frac{V_D}{2}\right]V_D \tag{5.123}$$

At the drain, y is equal to L, and therefore Q_n at $y = L$ is equal to

$$Q_n(y = L) = - C_{ox}[V_G - V_T - V_D] \tag{5.124}$$

The inversion charge density at the drain decreases, as V_D is increased, and at some value of V_D denoted $V_{D\,sat}$, the inversion charge density $Q_n(y = L)$, becomes zero. Q_n at the drain becomes zero when

$$V_D = V_{D\,sat} = V_G - V_T \tag{5.125}$$

At this value of the drain voltage, the channel is said to be "pinched-off" at the drain, and hence $V_{D\,sat}$ is also denoted V_p . We denote the drain current at V_D equal to $V_{D\,sat}$, as $I_{D\,sat}$. Substituting $V_{D\,sat}$ for V_D in Equation (5.122), we get

$$I_{D\,sat} = - \frac{\mu_n \ WC_{ox}}{L}\left[V_G - V_T - \frac{V_{D\,sat}}{2}\right]V_{D\,sat} = \ - \frac{\mu_n \ WC_{ox}}{2\,L}(V_G - V_T)^2 \tag{5.126}$$

When the channel is pinched off at the drain, the drain current is proportional to the square of the gate voltage as shown in Equation (5.126). If V_D is increased beyond $V_{D\,sat}$, the channel is pinched off at some other value of $y \ < \ L$. In other words, the pinch off point moves towards the source. The voltage at the pinch-off point y is now equal to $V_{D\,sat}$. Another way of looking at this is at the pinch off point, the gate voltage is equal to the local threshold voltage. Hence the inversion charge density is zero.

When $V_D > \ V_{D\,sat}$, let L' be the distance of the pinch-off point from the source as shown in Figure (5.34). The region between $y = L'$ and $y = L$, has a voltage drop $V_D \ - V_{D\,sat}$. This region is a depletion region. Although the channel is pinched off at $y = L'$, the current through the device is what would flow in a device with $V_D = \ V_{D\,sat}$ and of length L', since the integration carried out in Equation (5.121) to obtain I_D is now carried out between the source and the pinch off point. In other words, the current is given by

$$I_D(V_D > \ V_{D\,sat}) = \ - \frac{\mu_n \ WC_{ox}}{2\,L'}(V_G - V_T)^2 \tag{5.127}$$

In the olden days, devices were fabricated with very large L, and hence $L - L'$ was negligible in comparison with L. Hence I_D at voltages greater than $V_{D\,sat}$ was essentially the same as what was obtained at $V_D = V_{D\,sat}$. It is though the drain current saturates to a constant value $I_{D\,sat}$.

$$I_D(V_D > \ V_{D\,sat}) = I_{D\,sat} = \ - \frac{\mu_n \ WC_{ox}}{2\,L'}(V_G - V_T)^2 \tag{5.128}$$

However, in modern short channel devices the distance $(L - L')$ between the pinch off point and the drain is not negligible in comparison with L. Hence the drain current does not remain constant with the drain voltage, and increases with the drain voltage in saturation. The drain voltage drain current characteristic is shown in Figure (5.35). The drain conductance g_D is defined as

$$g_D = \ \frac{\partial I_D}{\partial V_D} \ |V_G \ = constant \tag{5.129}$$

and the transconductance g_m is defined as

$$g_m = \ \frac{\partial I_D}{\partial V_G} \ |V_D \ = constant \tag{5.130}$$

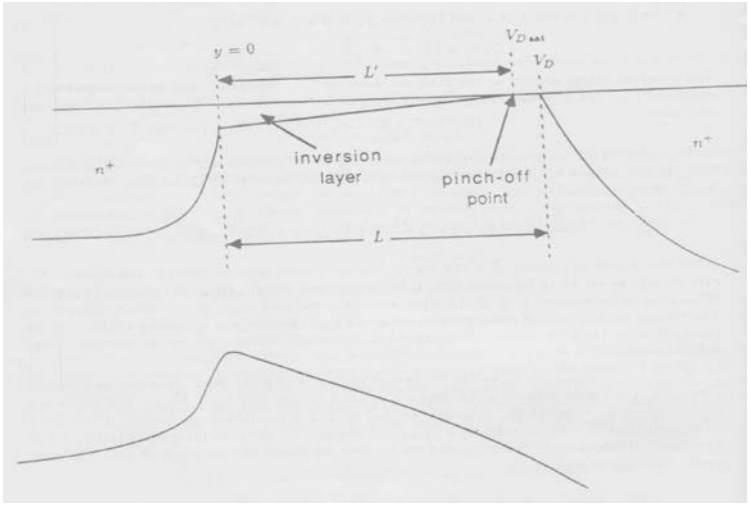


Figure (5.34) Schematics representation of the channel pinch-off.

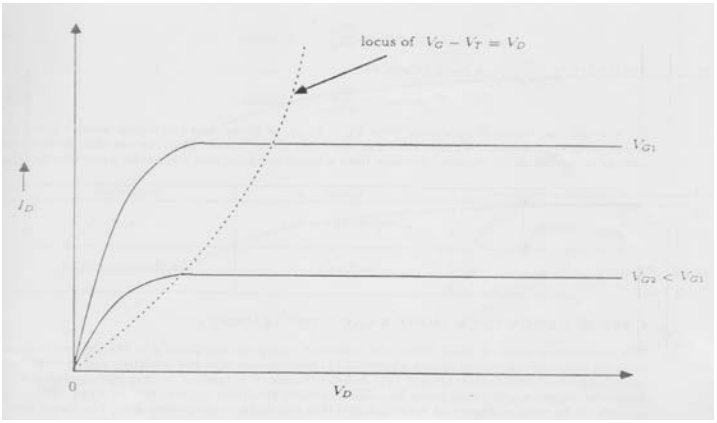


Figure (5.35) The drain current/drain voltage characteristics, called $I_D - V_D$ characteristics, of an n-channel MOS transistor.

We will call the region of operation when $V_D < V_{D\,sat}$ as linear, and the region when $V_D > V_{D\,sat}$ as the saturation region. This is shown in the table given below. Thus we see that as the drain voltage is increased, the current deviates from a linear behavior and ultimately saturates to $I_{D\,sat}$.

	Linear	Saturation
g_D	$\frac{\mu_n \ WC_{ox}}{L}(V_G - V_T)$	≈ 0
g_m	$\frac{\mu_n \ WC_{ox}}{L} V_D$	$\frac{\mu_n \ WC_{ox}}{L}(V_G - V_T)$

Case 3: Large V_D (A More Exact Treatment)

The assumption made in Case 2 that the threshold voltage at any point y in the channel increases from its value at the source by the amount $V(y)$ presupposes that the depletion region width does not change with the voltage drop $V(y)$. In other words, it is implicit in this assumption that the depletion region width is the same from the source to the drain independent of $V(y)$. This is not correct. If we refer to Figure (5.36) we will notice that due to the voltage drop $V(y)$ the Fermi level at point y is decreased from the Fermi level in the bulk by an amount equal to $qV(y)$ and therefore the band has to be bent by $q\ (V(y) + \ 2|\emptyset_b|)$ to produce the onset of inversion. Hence the condition for the inversion to set in at y is dictated by $V(y)$. The band-bending for the onset of inversion at y is given by

$$\psi_{s\,inv}(y) = \ V(y) \ + \ 2|\emptyset_b| \tag{5.131}$$

Under this condition, the depletion region width at y is given by

$$x_d(y) = \ \sqrt{\frac{2\,\epsilon_s}{q\ N_A}[V(y) \ + \ 2|\emptyset_b|]} \tag{5.132}$$

And the depletion charge density at y is given by

$$Q_d(y) = \ -q\ N_A\ x_d(y) = -\sqrt{q\ N_A 2\ \epsilon_s[V(y) \ + \ 2|\emptyset_b|]} \tag{5.133}$$

Now if we were to write the expression for gate voltage in terms of the space charge density and ψ_s, we get

$$V_G = V_{ox} + \ \psi_s\,(y) = \ -\ \frac{Q_{sc}\,(y)}{C_{ox}} + \psi_s\,(y) \tag{5.134}$$

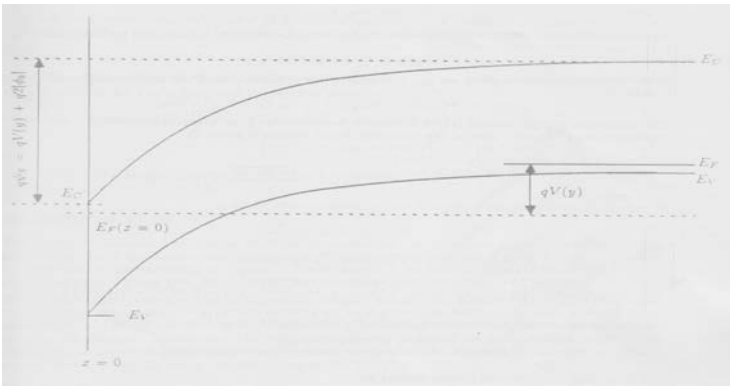


Figure (5.36): Band-bending at y when the voltage drop is $V(y)$.

We will now use the depletion approximation according to which the band bending in strong inversion is the same as the band bending at the onset of inversion. That is, the surface potential continues to remain at

$$\psi_s\,(y) = V(y) \ + \ 2|\emptyset_b| \tag{5.135}$$

even in strong inversion. Using the above two equations, we find that the total space-charge density is given by

$$Q_{sc}\,(y) = \ -\,C_{ox}\ [V_G - \ \psi_s\,(y)] = -\,C_{ox}\ [V_G - \ V(y) - \ 2|\emptyset_b|] \tag{5.136}$$

The inversion charge density is the difference between the total space charge density and the depletion charge density. Hence, the inversion charge density is given by

$$Q_n(y) = Q_{sc}(y) - Q_d(y)$$

$$= -C_{ox}\left[V_G - V(y) - 2|\emptyset_b| - \frac{\sqrt{q\,N_A 2\,\epsilon_s}}{C_{ox}}\,(V(y) + 2|\emptyset_b|)^{\frac{1}{2}}\right]$$

If we substitute this in the expression for the drain current which is

$$I_D = \mu_n\,W\,Q_n(y)\,\frac{dV}{dy}$$

and then integrate after multiplying both sides by $\frac{dy}{L}$, we obtain

$$I_D = -\frac{\mu_n\,C_{ox}\,W}{L}\left[\left(V_G - \frac{V_D}{2} - 2|\emptyset_b|\right)V_D - \frac{2}{3}\frac{\sqrt{q\,N_A 2\,\epsilon_s}}{C_{ox}}\left((V_D + 2|\emptyset_b|)^{\frac{3}{2}} - (2|\emptyset_b|)^{\frac{3}{2}}\right)\right]$$

(5.137)

We could have arrived at the same result by determining the local threshold voltage $V_T(y)$ as

$$V_T(y) = \frac{\sqrt{q\,N_A 2\,\epsilon_s\,[V(y) + 2|\emptyset_b|]}}{C_{ox}} + V(y) + 2|\emptyset_b| \qquad (5.138)$$

and obtaining the inversion charge density as

$$Q_n(y) = -C_{ox}\,[V_G - V_T(y)] \qquad (5.139)$$

This expression for the drain current is valid as long as the channel is not pinched off due to a high drain voltage. The channel is pinched off at $y = L$ if the gate voltage is equal to the threshold voltage at $y = L$. $V_{D\,sat}$ therefore can be obtain from Equation (5.138) by putting $V_T\,(y = L)$ as V_G and $V(y = L)$ as $V_{D\,sat}$. Hence solving for $V_{D\,sat}$ (left as an exercise for the student), we get

$$V_{D\,sat} = V_D - \frac{K^2}{2}\left(\sqrt{\frac{4V_G}{K^2} + 1} - 1\right) - 2|\emptyset_b| \qquad (5.140)$$

where

$$K = \frac{\sqrt{2\epsilon_s q N_A}}{C_{ox}} \qquad (5.141)$$

K is called the **body factor**, and is also sometimes denoted $\boldsymbol{\gamma}$.

The expression for the drain current I_D given in Equation (5.137) is valid in strong inversion regions ($V_G > V_T$). This expression is modified in the presence of a substrate bias voltage. If a reverse

bias voltage V_{SB} is applied between the source and the body, then the amount of band bending needed at the source $(y = 0)$ for the onset of inversion is

$$\psi_s\,(y = 0) = V_{SB}\ +\ 2|\emptyset_b| \tag{5.142}$$

The band-bending required at the onset of inversion at some point y in the channel is therefore equal to

$$\psi_s\,(y) = V_{SB}\ + V(y) +\ 2|\emptyset_b| \tag{5.143}$$

Hence the expression for the drain current in the presence of a substrate bias is modified such that $2|\emptyset_b|$ is replaced by $V_{SB}\ +\ 2|\emptyset_b|$ and is therefore given by

$$I_D = -\frac{\mu_n\, C_{ox}\ W}{L}\left[\left(V_G - \frac{V_{DS}}{2} - V_{SB} -\ 2|\emptyset_b|\right)V_{DS}\ -\ \frac{2}{3}\,\gamma\,\left((V_{DS}\ + V_{SB}\ + 2|\emptyset_b|)^{\frac{3}{2}} - (V_{SB}\ +\ 2|\emptyset_b|)^{\frac{3}{2}}\right)\right] \tag{5.144}$$

where V_{DS} is the drain to source voltage, V_G is measured with respect to the substrate, and γ is equal to K given in Equation (5.141)

Measurement of V_T

The threshold voltage, can be obtained by measuring the drain current as a function of the gate voltage for a small value of drain voltage. When the drain current is plotted as a function of V_G, the characteristics shown in Figure (5.37) are obtained. If we now extrapolate the linear portion to intercept the voltage axis, the intercept on the voltage axis is the threshold voltage.

We note that I_D does not go to zero when the gate voltage, V_G is equal to V_T. This is because the inversion charge density does not actually go to zero at this voltage. In fact, a small drain current flows through the device, even when the gate voltage is less than the threshold voltage under weak inversion conditions. This region of operation is called the subthreshold region. The calculation of the drain current in the subthreshold region is beyond the scope of this course. Hence we will assume that the drain current is zero at $V_G = V_T$.

Equivalent Circuit

We can draw the equivalent circuit for the MOS transistor. This is illustrated in Figure (5.38). The input voltage is applied between the gate and the source, and the output current is taken between the drain and the source. C_{in} and G_{in} are the capacitance and the conductance that the input voltage source looks into. In our case, C_{in} is given by

$$C_{in} = \frac{d\,Q_G}{d\,V_G} \approx W\ L\ C_{ox} \tag{5.145}$$

The value of C_{in} is obtained in strong inversion. The input conductance, G_{in}, is essentially due to leakage current in the oxide. The feedback capacitance between the drain and the gate, C_{fb}, is the sum of the gate overlap capacitance and the fringing field capacitance. The overlap capacitance arises because the gate electrode extends beyond the length of the channel over the source and the drain region. This fringing field capacitance arises due to edge effects. The output conductance g_D is given by

$$g_D = \frac{\delta I_D}{\delta V_D} \tag{5.146}$$

The transconductance g_m is given by

$$g_m = \frac{\delta I_D}{\delta V_G}$$

The output capacitance is the capacitance between the drain and the source region, and that is essentially equal to the capacitances of the source-to-substrate pn-junction and the drain to substrate pn-junction.

We can now calculate the maximum frequency up to which the MOS transistor can be used. The constant current source in the output circuit is equal to $g_m \tilde{v}_g$. This constant current source therefore, is directly related to the voltage across the conductance $G_{in,}$ that is in other words, the current through the conductance $G_{in,}$. Therefore as the frequency is increased, the input current flows more and more through the input capacitance $C_{in,}$, and less through the conductance G_{in}. We can now calculate the gain of the device at any given frequency, just as we did for the bipolar transistor in the last chapter. The maximum frequency is defined as that at which the current through the input capacitance, $C_{in,}$ is equal to $g_m V_G$. The current through C_{in} is given by $\tilde{v}_g j\omega C_{in}$. Equating this to $g_m \tilde{v}_g$ we get

$$\omega_{max} = \frac{g_m}{C_{in}} \tag{5.147}$$

This now leads to the maximum frequency as equal to

$$f_{max} = \frac{g_m}{2\pi C_{in}} \tag{5.148}$$

The ratio $\frac{g_m}{C_{in}}$ is called the figure of merit. In the linear region we saw that the transconductance g_m is given by

$$g_m = \frac{\mu_n C_{ox} W}{L} V_D$$

and

$$C_{in} = W L C_{ox}$$

Substituting these two values to calculate the figure of merit, we find the maximum frequency f_{max} to be

$$f_{max} = \frac{1}{2\pi} \frac{\mu_n V_D}{L^2} \tag{5.149}$$

Assuming V_D is very small because we are considering the linear region, $\frac{V_D}{L}$ is approximately the electric field, and multiplying it by μ_n gives us the velocity, and therefore, the maximum frequency f_{max} is rewritten as

$$f_{max} = \frac{v}{2\,\pi\,L} = \frac{1}{2\,\pi\,\tau_{tr}} \tag{5.150}$$

where v is the velocity of the carriers in the channel, and $\frac{v}{L}$ is the reciprocal of the transit time τ_{tr}. Thus we find that the maximum frequency at which the device can be operated is related to the time of transit of carriers from the source to the drain. The intrinsic limit on the maximum frequency of operation is imposed by the transit time of the carriers. This is true not only in MOS transistors and bipolar transistors but also in all other devices.

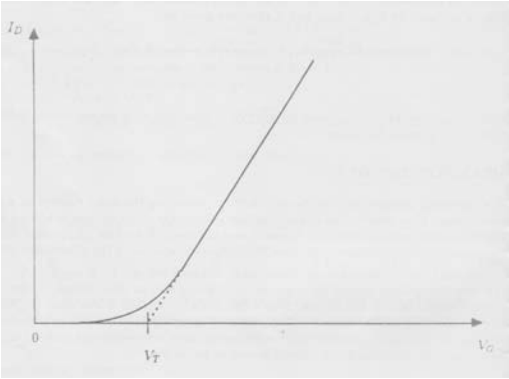


Figure (5.37) Plot of the drain current as a function of the gate voltage to obtain the threshold voltage

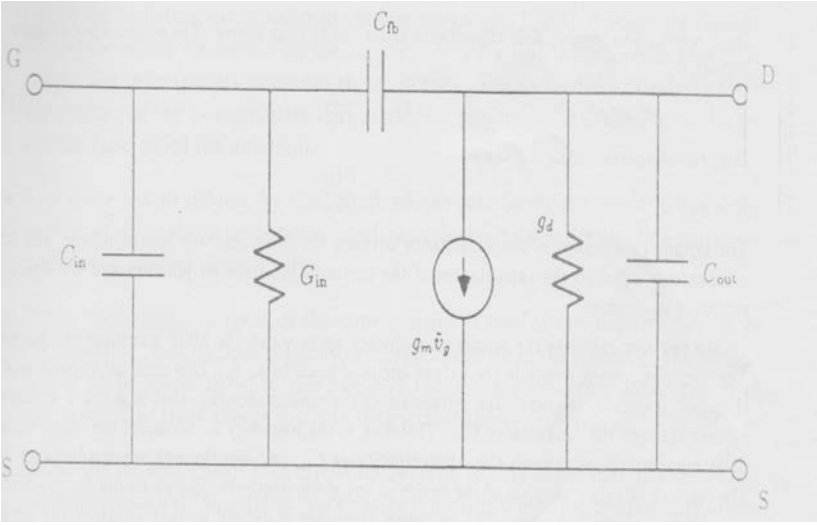


Figure (5.38) Equivalent circuit for the MOS transistor.

Summary

- An MOS device uses the electric field applied perpendicular to the surface of a semiconductor to induce space-charge in the semiconductor. The MOS structure consists of a semiconductor substrate on the surface of which a thin oxide layer is grown. A conducting layer (metal), deposited on top of the oxide, is called the gate and a voltage applied on the gate generates the electric field perpendicular to the semiconductor surface (also called the interface).
- The space charge induced by the voltage on the gate bends the bands in the semiconductor up or down towards the interface according to whether the gate voltage is negative or positive, respectively.
- When the induced charge is of the same polarity as that of the majority carriers in the substrate (bulk), excess majority carriers accumulate at the interface. The surface is said to be accumulated in this case and the device is in accumulation region of operation. The space charge is due to the excess majority carriers.
- When the induced charge is of a polarity opposite to that of the majority carriers in the semiconductor, the induced space charge is initially due to ionized impurity atoms in the substrate and the surface is said to be depleted. The device is said to be in the depletion region of operation. For larger values of gate voltage the minority carrier density at the interface becomes comparable to or large than the majority carrier density in the bulk. Under these conditions, the surface is said to be inverted. The charge due to the excess minority carriers is called the inversion charge and the layer of excess minority carriers is called the inversion layer. The inversion layer is very thin and therefore the inversion charge can be considered as a sheet charge. The space charge is due to both depletion charge and inversion charge.
- The space charge density, which is defined as the charge contained in the space charge under unit area of the interface, determines the electric field in the oxide and hence the potential difference across the oxide.
- The potential difference across the space charge region is the amount of band bending in the space charge region and is called the surface potential. The surface potential determines whether the substrate is accumulated, depleted or inverted.
- The charge in the accumulation layer or in the inversion layer is approximated as a sheet charge and hence the potential difference across these layers is taken as negligible. On the other hand, the depletion region is of non-zero width. The width of the space charge region is therefore approximated as the width of the depletion region.
- Inversion sets in when the bending of the band is such that the difference between the Fermi energy and the intrinsic Fermi energy at the interface is the same in magnitude as the difference between these two energies in the bulk but of opposite polarity. This condition is called “on-set of inversion”. The gate voltage at which the surface starts to be inverted (on-set of inversion) is called the threshold voltage.

- According to the depletion approximation, the depletion region width remains constant in strong inversion at a value it had at the on-set of inversion. The width of the depletion region remains at a maximum value $x_{d\,max}$ under inversion condition.
- An ideal MOS device is one in which there is no space charge in the semiconductor when the gate voltage is zero. A voltage is required to be applied on the gate to induce a space charge.
- A non-ideal MOS device is one in which there is a space charge even when there is no applied voltage on the gate, Charge in the gate oxide, interface states and work function difference between the gate and the substrate give rise to non-ideal behavior.
- Flat band voltage is the voltage needed to be applied on the gate to reduce the space charge to zero so that the device with a voltage on the gate equal to the flat –band voltage looks like an ideal device with zero gate voltage.
- The space charge induced by the oxide charge and the work function difference is independent of band-bending while that due to the interface states depends on band bending.
- Interface states can be acceptor type or donor type. It is conventional to assume that the interface states in the upper half of the band gap are acceptor type while those in the lower half are donor type.
- The space charge region gives rise to small signal capacitance effect. This capacitance is called space charge capacitance and is equal to the rate of increase of space charge with surface potential.
- Since the inversion and accumulation charge increases exponentially with surface potential, the space charge capacitance of accumulation and inversion charges is much larger than the capacitance due to the oxide layer. The capacitance due to the oxide layer called the oxide capacitance is the same as a parallel plate capacitor with the gate oxide as the dielectric. The MOS capacitor behaves equivalent to a series combination of the oxide capacitance and the space charge capacitance.
- Due to the dependence of the space charge capacitance on the surface potential which in turn depends on the gate voltage, the capacitance of a MOS capacitor varies with the gate voltage. The capacitance-voltage characteristic is called the C-V curve.
- In accumulation, the MOS capacitance is equal to the oxide capacitance itself since the space charge capacitance is very much larger.
- As the device is biased into depletion, the space charge capacitance (equal to the depletion region capacitance $\approx \frac{\epsilon_s}{x_d}$) decreases (due to the increase in depletion region width) and hence the capacitance of the MOS device decreases.
- At strong inversion, the capacitance behavior is different at high and low frequency of the measuring small signal. At high frequencies, the inversion charge does not vary in step with the small measuring signal and hence the space charge capacitance remains at a minimum value $\approx \frac{\epsilon_s}{x_{d\,max}}$ and the MOS capacitance remains constant at a minimum value.
- At low frequencies, the inversion charge has time to be generated thermally so that it can vary in step with the measuring signal; hence the space charge capacitance is determined by the

variation with time of the inversion charge and is therefore very large. The MOS capacitance in strong inversion becomes equal to the oxide capacitance as in the accumulation case.

- In the MOS transistor, the conductance between two heavily doped regions, in the substrate, (called the source and the drain) is controlled by the surface inversion of the region between the source and the drain.
- In the simplest model, the drain voltage is assumed to be small and the inversion charge density is assumed to be constant between the source and the drain.
- If the inversion charge is due to electrons providing a conductance between the n^+ source and drain, the device is called an n-channel MOS transistor.
- If the source and drain are heavily doped p regions and the inversion charge is due to holes, the device is called a p-channel MOS transistor.
- When the drain voltage is high, the inversion charge density decreases towards the drain due to the voltage drop along the channel. This is equivalent to the local threshold voltage increasing from its lowest value near the source to the highest value near the drain.
- In one approximate model, the local threshold voltage at some point y in the channel is taken as sum of the voltage drop at y measured with respect to the source and the threshold voltage at the source.
- In a more exact model, the band bending required to attain inversion at some point y in the channel is taken as the sum of voltage drop at y and the band bending needed at the source to produce inversion.
- As the drain voltage is increased, ultimately the inversion charge density at the drain decreases and becomes zero. The channel is said to be pinched off at the drain. When the drain voltage is increased further, the pinch – off point moves toward the source and the drain current increases very slowly with the drain voltage as though the current has saturated. Hence the transistor is said to be operating in the saturation region.
-

Glossary

$\mathring{A}$ = Angstrom (s)
C = an integration constant
C = capacitance per unit area of the MOS capacitor
C_d = depletion layer capacitance
C_{fb} = feedback capacitance between the drain and the gate
C_{in} = input capacitance
C_m = measured capacitance
C_{min} = minimum capacitance in the high frequency $C-V$ curve
C_{ox} = oxide layer capacitance per unit area
C_{SC} = space charge capacitance per unit area
$e.s.$ = electrostatic potential
E_c = energy at the bottom of the conduction band (also potential energy of electrons)
E_F = Fermi energy
E_{F1} = Fermi level in a p-type semiconductor
E_g = energy gap
E_i = intrinsic Fermi energy
E_{ib} = intrinsic Fermi energy level in the bulk
E_{vac} = vacuum level
E_V = energy at the top of the valence band (also potential energy of holes)
$\mathcal{E}$ = electric field
$\mathcal{E}_{ox}$ = electric field in the oxide
$\mathcal{E}(x)$ = electric field at x
f_{max} = maximum frequency
g_D = drain conductance or output conductance
g_m = transconductance
G_{in} = input conductance
$\tilde{\imath}$ = small signal current
I_D = drain current
$I_{D\,sat}$ = drain current at saturation
J = current density
k = Boltzmann constant
K = body factor, also γ
K_{ox} = dielectric constant of the oxide
L = length of the channel
L' = distance of the pinch-off point from the source
n = electron density
n_0 = thermal equilibrium electron density
n_b = electron density in the bulk
n_i = intrinsic carrier density

n_{n0} = thermal equilibrium electron density in n-type material
n_s = electron concentration at the interface or surface
$n(x)$ = electron density at x
N_A = acceptor impurity density
N_A^- = ionized acceptor impurity density
N_D = donor impurity density
N_D^+ = ionized donor impurity density
N_{it} = number of interface states per unit area per unit energy interval at some energy level E
p = hole density
p_0 = thermal equilibrium hole density
p_b = hole density in the bulk
p_s = hole concentration at the interface or surface
$p(x)$ = hole density at x
q = electron charge (magnitude)
$q\Phi$ = work function of the solid
$q\Phi_A$ = work function of solid A
$q\Phi_B$ = work function of solid A
$\tilde{Q}$ = sinusoidally varying small signal charge
Q_{acc} = space charge density in accumulation
Q_d = depletion charge density
$Q_{d\ max}$ = maximum depletion charge density
Q_G = charge on the gate
Q_{inv} = inversion charge density
Q_n = electron density at the surface or interface
Q_{ox} = effective sheet charge density
Q_p = hole density in the surface or interface
Q_{SC} = space charge density
Q_{st} = sheet charge density at the interface or surface
t = time
t_{ox} = thickness of the oxide layer
v = velocity of the carriers in the channel
$\tilde{v}$ = small signal sinusoidal voltage
$\tilde{v}_g$ = small signal sinusoidal gate voltage
V_{bi} = built-in voltage, contact potential
V_D = drain voltage
$V_{D\ sat}$ = saturation drain voltage
V_{DS} = drain to source voltage
V_{FB} = flat-band voltage
p_0 = thermal equilibrium density of holes
V_G = gate voltage
V_{ox} = voltage across the oxide layer
V_p = pinch-off voltage

V_{SB} = reverse substrate bias voltage
V_T = threshold voltage
$V_{T\,ideal}$ = threshold voltage of an ideal MOS device
$V_{T\,non-ideal}$ = threshold voltage of a non-ideal MOS device
$V(y)$ = voltage at y
W = width of the channel
x = position value
x_d = width of the space-charge or depletion region
$x_{d\,max}$ = depletion region width at the on-set of strong inversion
x_I = thickness of the inversion layer of electrons
y = position value
$\Delta\psi_s$ = change in surface potential
ϵ_0 = permittivity of free space
ϵ_{ox} = permittivity of the oxide
ϵ_s = permittivity of the semiconductor
γ = body factor, same as K
$\emptyset$ = electrostatic potential
$\emptyset_b$ = bulk potential
$\emptyset_n$ = bulk electrostatic potential of the n-region
$\emptyset_p$ = bulk electrostatic potential of a p-type semiconductor
$\emptyset\ (x)$ = electrostatic potential at x
Φ = work function
Φ_1 = work function of a p-type semiconductor
Φ_2 = work function of an n-type semiconductor
Φ_{AB} = contact potential
Φ_M = work function of the gate
Φ_{MS} = work function difference between the gate and the substrate
Φ_S = work function of the substrate semiconductor
ρ = charge density
τ_g = minority carrier generation lifetime
τ_{tr} = transit time
$\boldsymbol{\psi}$ = electrostatic potential difference
ψ_s = surface potential or interface potential
ψ_{inv} = band bending at the onset of inversion or surface potential at onset of inversion

$\boldsymbol{\psi}(\mathrm{x})$ = electrostatic potential difference at x i.e., $\emptyset\ (x) - \emptyset_b$
χ = electron affinity
ω = 2 π times the frequency of the small signal voltage $\tilde{v}$
ω_{max} = 2 π times the maximum frequency

Problems

1. Show that $n_s = n_i$ when $\psi_s = -\emptyset_b$ for a MOS device with a p-type substrate.
2. For an ideal MOS diode (capacitor) with $t_{ox} = 500\text{Å}$ and $N_A = 10^{15}cm^{-3}$, Find the value of the surface potential ψ_s and the voltage across the oxide layer required to produce a) intrinsic condition at the interface and b) on-set of strong inversion at the interface.
3. In a MOS capacitor with a p-type substrate determine the space-charge density for a gate voltage of $0.75\ V$ assuming the oxide layer thickness to be 1000 Å and the substrate impurity concentration to be $5 \times 10^{15}cm^{-3}$.
4. Consider a MOS capacitor with a substrate doping of $2 \times 10^{15}cm^{-3}$ of donor atoms and an oxide thickness of 600 Å, Determine the gate voltage at which the surface becomes intrinsic.
5. For the device in the previous problem, find the value of the space-charge density at a gate voltage equal to
    a. The threshold voltage
    b. Twice the threshold voltage
6. For the device in the previous problem, find the inversion charge density and the depletion region charge density when the gate voltage is three times the threshold voltage.
7. For the device in the previous problem, find the high and low frequency capacitance at a gate voltage of $-4V$. Assume an area of $10^{-4}cm^2$ for the capacitor.
8. Show that the potential $\boldsymbol{\psi}$(x) in the depletion region of a MOS capacitor varies as $\boldsymbol{\psi}(\text{x}) = \psi_s\left(1 - \frac{x}{x_d}\right)^2$.
9. Show that in deep depletion, the relation between the gate voltage and the surface potential ψ_s is given by

$$V_G = \frac{K^2}{2}\left(\sqrt{\frac{4V_G}{K^2} + 1} - 1\right) + \psi_s$$

Where K is defined as

$$K = \frac{\sqrt{2\epsilon_s q N_A}}{C_{ox}}$$

10. Consider a MOS device in which the oxide layer contains positive ions with a density of $10^{16}cm^{-3}$. If the oxide thickness is 400 Å calculate the flat-band voltage V_{FB}.
11. Assume the gate is p-type poly-silicon with an impurity concentration of $10^{17}cm^{-3}$ and the substrate to be n-type with a resistivity equal to 5 Ω-cm. Assuming the electron affinity of silicon to be $4.1\ V$, determine the flat-band voltage.
12. Consider a MOS capacitor with a p-type substrate and an oxide thickness of 500 Å . Let the substrate impurity density be $5\ \times 10^{15}cm^{-3}$. Assume the device is ideal.

a. Calculate the high frequency and low frequency capacitances in strong inversion.

b. Assuming that a step voltage of $5\ V$ is applied on the gate at $t = 0$. Calculate the small signal capacitance at $t = 0^+$.

13. Consider a MOS transistor with a p-type substrate containing $10^{15} cm^{-3}$ acceptor impurities and an oxide thickness of 500 Å. The flat-band voltage is $0.5\ V$. Determine the drain conductance for a gate voltage of $5\ V$ in the linear region. Take the mobility μ_n as $900\ cm^2 V^{-1} sec^{-1}$, W as $20\ \mu m$ and L as 10μm.

14. For the device in the previous problem, determine the transconductance in the saturation region.

15. Consider an n-channel MOS transistor with $t_{ox} = 150$ Å. The substrate impurity concentration is $10^{16} cm^{-3}$. Let the channel length, L, be $2\ \mu m$, the channel width, W be $20\ \mu m$ and the mobility of the channel carriers be $750\ cm^2 V^{-1} sec^{-1}$.

    a. What is the value of the transconductance for a drain voltage of $20\ mV$?

    b. Assume that the oxide has positive ions of density equal to $10^{17} cm^{-3}$. Assume that the gate is p-type poly-silicon with an impurity density of $10^{18} cm^{-3}$. Neglect the contribution of surface states to the flat-band voltage. Using the simplified treatment given in case 2, determine the saturation drain voltage, V_{DSAT} for a gate voltage $5\ V$.