

## **SEMICONDUCTOR MATERIALS AND DEVICES**

*by H.L. Kwok*

**Objective:** The purpose of these notes is to familiarize students with semiconductors and devices including the P-N junction, and the transistors.

## **Part 1: Semiconductor Materials**

### **What we need to learn.**

- a. Semiconductor structures
- b. Electronic properties
- c. Intrinsic and extrinsic semiconductors

## **Part 1: Semiconductor Materials**

## A. Properties

- Semiconductors occupy a small fraction of the Periodic Table.

The word semiconductor means a material having conduction properties half-way between a conductor and an insulator.

Solids are often classified according to their conductivity and semiconductors have conductivities in the range of  $10^4 - 10^{-6}$  S/m.

Fig.1.1 shows the conductivity range of semiconductors.

Semiconductors that are of interest to us include: silicon (**Si**), germanium (**Ge**) and gallium arsenide (**GaAs**).

#### a) Crystal structures

- Si and Ge have the *diamond crystal structure*.

The diamond crystal structure is formed by 2 interlaced fcc unit cells. The cells are displaced by 1/4 the distance along the body diagonal. Fig.1.2 shows the diamond crystal structure.

- GaAs has the *zinc-blende* crystal structure (similar to the diamond structure except that the Ga atoms and the As atoms occupy different lattice sites). This is illustrated in Fig.1.3.

In these crystals, the bonding arrangement between the nearest neighbors is tetrahedral (i.e., each atom has 4 nearest neighbors). The bond angle is shown in Fig.1.4.

Semiconductors have covalent bonds. For covalent bonds, the electrons are shared between the nearest neighboring atoms as shown in Fig.1.5.

Example 1.1: At 300K, the lattice constant for Si is 0.543 nm, calculate the number of Si atoms/m<sup>3</sup>.

Solution:

Each unit cell of the fcc structure has 4 atoms. Therefore, the diamond structure has 8 atoms per unit cell.

The volume of the unit cell  $V = (5.43 \times 10^{-10})^3 \text{ m}^3 = 1.6 \times 10^{-28} \text{ m}^3$ .

The number of atoms per unit volume  $N = 8 / (1.6 \times 10^{-28}) / \text{m}^3 = 5 \times 10^{28} / \text{m}^3$ .  
#

- Lattice direction is a vector represented by the symbol:  $[x \ y \ z]$ , where  $x$ ,  $y$ , and  $z$  are the values of the lattice coordinates (of the vector) measured from the origin (a chosen lattice point).

Lattice directions parallel to one another belong to the same family and are represented by the same symbol (in angular brackets).

- A lattice plane is represented by a set of numbers  $(h \ k \ l)$  known as the *Miller indices*.

$h$ ,  $k$ , and  $l$  are the normalized values of the inverse of the interceptions of the plane on each of the coordinates.

Example 1.2: A lattice plane intercepts the principal axes at positions:  $a$ ,  $2a$ ,  $2a$ . Determine the Miller indices.

Solution:

The Miller indices are:  $(211)$ .

This is illustrated in Fig.1.6. #

## B. Electronic properties of semiconductors

- A semiconductor at low temperature is similar to an insulator and the energy bonds are intact.

There are very few **free** electrons at low temperature.

As temperature increases, some energy bonds are broken and the density of electrons increases. The density of electrons is quite small ( $\sim 10^{12-19} /m^3$ ) at room temperature.

- Creation of free electrons is accompanied by the creation of holes.

A hole is a broken energy bond with a missing electron. Holes are positively charged and their motion also leads to current flow.

A hole moves when the hole is filled with an electron. This is illustrated in Fig.1.7.

- For an intrinsic semiconductor, the electron density  $n$  equals to the hole density  $p$ , i.e.,  $n = p$ .

Example 1.3: Intrinsic Si has  $1.45 \times 10^{16} /m^3$  of electrons and holes, respectively. If  $\mu_n = 0.15 \text{ m}^2/V.s$  and  $\mu_p = 0.045 \text{ m}^2/V.s$ , determine the conductivity.

Solution:

The conductivity  $\sigma = nq\mu_n + pq\mu_p = 1.45 \times 10^{16} /m^3 \times 1.6 \times 10^{-19} C \times (0.15 + 0.045) m^2/V.s = 0.45 \times 10^{-3} S/m.$  #

(Remember we mentioned earlier on that  $\sigma$  for a semiconductor lies between  $10^4$  to  $10^{-6} S/m.$ )

### a) Energy bands

- Electrons and holes in a semiconductor are allowed to possess specific energy ranges.

These energy ranges are called energy bands.

Traditionally, engineers are interested in electrons and energy bands are constructed with electrons in mind, i.e., they are plotted in terms of **negative potential energy**. Thus, in a typical **energy band diagram**, electron energy (negative potential energy) is plotted “upward”, while hole energy (positive potential energy) is plotted “downward”.

Fig.1.8 shows the energy band diagram for Si. Crystal momenta (along two different crystal directions) are plotted along the horizontal axes.

The symbol E stands for electron energy.

Since electrons normally reside in the lowest energy states in a solid, electrons will first fill up the lower energy bands. When these are filled, they start to fill the higher energy bands. As a

result, some higher energy bands will be partly filled. These are called conduction bands. They are illustrated in Fig.9.

Partially emptied energy bands have holes and they are called valence bands.

The energy difference between the conduction band edge and the valence band edge is the energy gap. Electrons and holes are not allowed to have energies falling within the energy gap.

Effectively, energy gap is the amount of energy needed to form an electron and a hole.

The shape of an energy band changes with temperature, pressure, etc.

For instance: the energy gap  $E_g(T)$  of Si has the following temperature dependence:

$$E_g(T) = 1.17\text{eV} - 4.73 \times 10^{-4} T^2 / (T + 636 \text{ K}) \text{ eV} \quad (1-1)$$

where T is the absolute temperature.

This dependence is shown in Fig.1-10.

Example 1.4: What is  $E_g$  for Si at 300 K?

Solution:



From Eqn.(1-1):  $E_g(T = 300 \text{ K}) = 1.17 - 4.73 \times 10^{-4} \times 300^2 / (300 \text{ K} + 636 \text{ K})$   
 $eV = 1.12 \text{ eV.} \quad \#$

## b) Electron motion

- Motion of electrons and holes is affected by the lattice potential.

This may be viewed as an *interaction* between the electron/hole and the lattice vibrations.

The result is that electrons and holes appear to possess masses different from the **electron rest mass**,  $m_0$ .

These are called effective masses, ( $m_e^*$  for electrons, and  $m_h^*$  for the holes) and they are less than  $m_0$ .

The small masses imply that the lattice vibrations “appear” to assist in the motion of the electrons and the holes.

For Si at 300 K:  $m_e^*$  (electrons) = 0.33  $m_0$ , and  $m_h^*$  (holes) = 0.55  $m_0$ .

## c) Density of states for the electrons and holes

- An energy band contains many energy levels/states.

Electrons occupying energy levels/states obey the Pauli exclusion principle - which states that there can be at most 2 electrons per energy state.

The density of states in an energy band is given by the **density of states function**  $S(E)$  given by:

$$S(E) = 4\pi [2m_b^*/h^2]^{3/2} E^{1/2} \quad (1-2)$$

where  $m_b^*$  is a density of states mass parameter and  $h$  ( $= 0.662 \times 10^{-33}$  J.s) is the Planck's constant.

$S(E)$  has the dimension of per unit volume per unit energy.

To compute the electron density  $n$ ,  $S(E)$  is multiplied to a probability function that characterizes the occupation of the different energy states. This probability function is the **Fermi-Dirac function**  $f(E)$ .

$$f(E) = 1/[1 + \exp((E - E_F)/kT)] \quad (1-3)$$

where  $E_F$  is the **Fermi-level**,  $E$  is the energy of the electron,  $k$  ( $= 1.38 \times 10^{-23}$  J/K) is the Boltzmann constant, and  $T$  is the absolute temperature.

Based on the above, the electron density  $n$  in the conduction band is given by:

$$n = \int_{E_C}^{\infty} S_e(E)/[1 + \exp((E - E_F)/kT)].dE \quad (1-4)$$

where  $S_e$  is the density of states in the conduction band,  $E_F$  is the Fermi-level, and  $E_C$  is the energy at the conduction band edge.

A graphical plot of  $S_e$  and  $f(E)$  versus  $E$  is shown in Fig.1-11.

A similar expression for the hole density  $p$  in the valence band is given by:

$$p = \int_{-\infty}^{E_V} S_h(E) / [1 - \exp((E - E_F)/kT)]. dE \quad (1-5)$$

where  $S_h(E)$  is the density of states in the valence band, and  $E_V$  is the energy at the valence band edge.

If we assume that both  $|E_C - E_F|$  and  $|E_V - E_F| \gg kT$ ,

$$n \approx N_C \exp[-(E_C - E_F)/kT]$$

$$p \approx N_V \exp[-(E_F - E_V)/kT] \quad (1-6)$$

where  $N_C (= 2 \times (2\pi m_b^* kT/h^2)^{3/2})$  is the effective density of states in the conduction band, and  $N_V (= 2 \times (2\pi m_b^* kT/h^2)^{3/2})$  is the effective density of states in the valence band.

At 300 K,  $N_C = 2.8 \times 10^{25} /m^3$  and  $N_V = 1.04 \times 10^{25} /m^3$  for Si.

According to Eqn.(1-6), electron and hole densities depend on the position of the Fermi-level. This is illustrated in Fig.1-12.

Example 1.5: Compute the electron density in Si if the Fermi level is at 0.55 eV below the conduction band edge. Assume  $T = 300$  K ( $kT$  at 300 K = 0.0259 eV).

Solution:

From Eqn.(1-5): The electron density  $n = N_C \exp(- (E_C - E_F)/kT) = 2.8 \times 10^{25} /m^3 \times \exp(- 0.55 \text{ eV} /0.259 \text{ eV}) = 1.67 \times 10^{16} /m^3$ . #

d) Intrinsic semiconductor

- An *intrinsic* semiconductor is one with **no impurities**.

For an intrinsic semiconductor,  $n = p = n_i$ , the intrinsic carrier density.

Using Eqn.(1-6), the intrinsic carrier density is given by:

$$n_i = [N_C N_V \exp(- E_g/kT)]^{1/2} \quad (1-7)$$

where we have substituted  $E_g = E_C - E_V$ .

At 300 K,  $n_i = 1.45 \times 10^{16} /m^3$  for Si.

- Note that:  $pn = n_i^2$ .

This is a very important relationship and is a consequence of the ***law of mass action*** - which states that the product of the electron density and the hole density is a **constant** at a given temperature.

For an intrinsic semiconductor, we have  $E = E_i$ , the *Intrinsic Fermi- level*.

It can be shown that:

$$E_i = E_g/2 + kT/2 \ln(N_V/N_C) \quad (1-8)$$

When  $N_V = N_C$ ,  $E_i$  is at the middle of the energy gap.

e) Extrinsic semiconductors

- Semiconductors become **extrinsic** when impurities are added.

This is achieved with the addition of donors or acceptors. Impurities are called donors/acceptors depending on whether electrons/holes are produced.

Structurally, donors are similar to the semiconductor (host) atoms except that there are 5 electrons in the outermost orbits (versus 4 electrons for the (host) semiconductor). Ionization of the donor atom generates an electron.

Acceptor atoms have 3 electrons in the outermost orbits and the capture of an electron from a neighboring atom will generate a hole in the neighboring atom. This is illustrated in Fig.1-13.

- Because of their low ionization energies, donors and acceptors are “fully” ionized at room temperature.

This results in:  $n = N_D$  or,  $p = N_A$ , where  $N_D$  is the donor density and  $N_A$  is the acceptor density.

A semiconductor is **N-type** when  $n \gg p$ , and **P-type** when  $p \gg n$ .

- In the presence of donors/acceptors, electron and hole densities will change.

Since the law of mass action dictates that  $pn = n_i^2$ ,  $p = n_i^2/N_D$  for a N-type semiconductor and  $n = n_i^2/N_A$  for a P-type semiconductor.

- For an extrinsic semiconductor, the position of the Fermi-level moves away from the intrinsic position.

For an N-type semiconductor ( $n = N_D$ ),

$$E_C - E_F = kT \ln(N_C/N_D)$$

Similarly, for a P-type semiconductor ( $p = N_A$ ),

$$E_V - E_F = -kT \ln(N_V/N_A) \quad (1-9)$$

Fig.1-14 shows the positions of the Fermi-levels in an N-type semiconductor and in a P-type semiconductor, respectively.

If both donors and acceptors are present in a semiconductor, the dopant in greater concentration dominates, and the one in smaller concentration becomes negligible.

- Semiconductors are named after the dominant carrier type.

Thus, an N-type semiconductor has more electrons than holes, while a P-type semiconductor has more holes than electrons.

- The dominant carrier is called the **majority carrier** versus the term **minority carrier** for the carrier in lesser quantity.

It is sometimes more convenient to express  $n$  and  $p$  in terms of the intrinsic Fermi-level  $E_i$ , i.e.,

$$\begin{aligned} n &= N_C \exp(- (E_C - E_F)/kT) \\ &= N_C \exp(- (E_C - E_i)/kT) \exp(- (E_F - E_i)/kT) \\ &= n_i \exp((E_F - E_i)/kT) \end{aligned} \quad (1-10)$$

Similarly,

$$p = n_i \exp((E_i - E_F)/kT) \quad (1-11)$$

Using this representation, one reduces the number of unknown parameters for the above equations of  $n$  and  $p$  from 3 ( $N_C$ ,  $N_V$ ,  $E_C/E_V$ ) to 2 ( $E_i$  and  $n_i$ ).

Example 1.6: Find  $E_C - E_F$  if  $N_D = 1 \times 10^{22} /\text{m}^3$  for Si at 300 K.

Solution:

From Eqn.(1-10),  $E_C - E_F = kT \ln(N_C/N_D) = 0.0259 \text{ eV} \times \ln(2.8 \times 10^{25} /\text{m}^3 / 1 \times 10^{22} /\text{m}^3) = 0.2 \text{ eV}$ . #

Example 1.7: If Si is doped with  $1 \times 10^{20}$  donors/m<sup>3</sup>, compute the carrier densities at 300 K.

Solution:

From Eqn.(1-8):  $n = N_D = 1 \times 10^{20}$  /m<sup>3</sup> and  $p = n_i^2/N_D = (1.45 \times 10^{16})^2/(1 \times 10^{20} /m^3) = 2.1 \times 10^{12}$  /m<sup>3</sup>. #

#### f) Dopant energy levels

Table 1.1 shows the dopant ionization energies in common semiconductors at 300 K:

	<u>Phosphorus</u>	<u>Arsenic</u>	<u>Boron</u>
Si	0.045	0.049	0.045
Ge	0.012	0.0127	0.0104
GaAs	0.026	-	-

Table 1.1: Dopant ionization energies in the common semiconductors (measured in eV from the band edges: i.e.,  $N_C$  in the case of the donors and  $N_V$  in the case of the acceptors).

The ionization energies may be compared to the thermal energy  $kT$  (= 0.0259 eV) at 300 K.

Fig.1-15 shows the positions of dopant energy levels in an energy band diagram.

#### g) Temperature dependence of the carrier densities



- Electron and hole densities in an extrinsic semiconductor can vary with temperature.

At a low temperature, not all of the dopant is ionized and the majority carrier density will be small. Nevertheless, its density is still much larger than the intrinsic carrier density.

At high temperature, the intrinsic carrier density increases rapidly, while  $N_D$  or  $N_A$  remains unchanged. At a sufficiently high temperature, the semiconductor becomes intrinsic.

The variation of the electron density  $n$  versus the inverse temperature in an N-type semiconductor is shown in Fig.1-16.

## **Part 2: Carrier transport**

### **What we need to learn.**

- a. Electronic motion
- b. Conduction properties of electrons and holes
- c. Drift and diffusion
- d. Generation and recombination
- e. Continuity equations

## Part 2: Carrier transport

- Carrier transport refers to the movement of the electrons and holes in a semiconductor under an applied voltage.

The simplest types of carrier transport are by **drift** and **diffusion**.

Drift is charge flow as a consequence of the presence of an applied electric field, while diffusion is charge flow due to the presence of a carrier density gradient.

### A. Random motion of the electrons

- At room temperature, electrons possess thermal energy and its value is  $\approx 3kT/2$ .

Equating this thermal energy to the kinetic energy of an electron will give:

$$m_e^* v_{th}^2/2 = 3kT/2 \quad (2-1)$$

where  $m_e^*$  is the effective mass of the electron, and  $v_{th}$  is the thermal velocity.

It can be shown that at  $T = 300$  K, the thermal velocity  $v_{th} \approx 1 \times 10^5$  m/s.

This suggests that electrons in a semiconductor are moving at a very high speed.

## B. Electrons in an electric field

An electric field  $E'$  imposes a preferred direction of motion for the electrons. This is superimposed on top of the random motion.

Because of their charge, electrons will move in opposite direction to the applied electric field (holes actually move in the same direction)

- The motion of the electrons and the holes will stop when they encounter collisions.

This is illustrated in Fig.2.1.

Assuming a mean free time  $\tau_c$  between collisions for the electrons, Newton's law gives:

$$-qE' = m_e^* v_d / \tau_c \quad (2-2)$$

where  $v_d$  is the drift velocity of the electrons.

This leads to:

$$v_d = -q\tau_c E' / m_e^* \quad (2-3)$$

By definition: the electron mobility  $\mu_n = -v_d / E'$ . We have:

$$\mu_n = q\tau_c / m_n^* \quad (2-4)$$

Note that:  $\mu_n$  is positive.

Collision of a hole can be viewed as an interruption in the filling of the hole by an electron. The drift velocity is then given by:

$$v_d' = q\tau_c^* E'/m_h^* \quad (2-5)$$

where  $\tau_c'$  is the mean free time of the holes, and  $m_h^*$  is the hole effective mass.

This leads to:

$$\mu_p = q\tau_c^* E'/m_h^* \quad (2-6)$$

According to Eqns.(2-5) and (2-6), the mean free times of the electrons and the holes determine their mobilities.

Collision is also referred to as scattering. Theoretically, it can be shown that:  $\tau_c \sim T^{-3/2}$  for lattice scattering, and  $\tau_c \sim T^{3/2}$  for ionized impurity scattering.

When more than one type of scattering is present, the shortest mean free time dominates.

This also applies to the carrier mobility since mobility is directly proportional to the mean free time.

In general, the combined mobility  $\mu_T$  in the presence of more than one type of scattering is given by:

$$\mu_T^{-1} = \sum_i \mu_i^{-1} \quad (2-7)$$

where  $\mu_i$  is the mobility due to the  $i$ th scattering.

The process is shown in Fig.2.2.

From the carrier mobility, one can compute the drift velocity.

Example 2.1: At 300 K,  $\mu_n = 0.15 \text{ m}^2/\text{V.s}$ . Determine its value at 200 K if lattice scattering dominates.

Solution:

$$\mu_n(T = 200 \text{ K}) = \mu_n(T = 300 \text{ K}) (300 \text{ K}/200 \text{ K})^{3/2} = 0.15 \text{ m}^2/\text{V.s} \times 1.84 = 0.275 \text{ m}^2/\text{V.s.} \quad \#$$

- Current density is the charge flow through a unit area cross-section in unit time.

Consider a unit volume in a semiconductor, the amount of electrons  $Q$  inside this unit volume is given by:  $Q = -nq$ , where  $n$  is the electron density, and  $q$  is the electron charge.

If this quantity of charge  $Q$  moves through a distance  $\Delta x$  in time  $\Delta t$ , the electron current density  $J_n (= Q \Delta x / \Delta t = Qv_d)$  is given by:

$$J_n = -qnv_d \quad (2-8)$$

Similarly, hole current density  $J_p = qp v_d'$ .

The total current density  $J_T$  is given by:

$$J_T = J_n + J_p = -qnv_d + pqv_d' \quad (2-9)$$

Since  $\mu_n = -v_d/E'$  and  $\mu_p = v_d'/E'$ ,

$$J_T = (nq\mu_n + pq\mu_p) E' = \sigma E' \quad (2-10)$$

where  $\sigma$  is the conductivity of the semiconductor.

Note that the inverse to conductivity is the resistivity.

For a measurement of resistivity, a 4-point probe method is frequently used. The measurement setup is shown in Fig.2.3.

In the measurement, a current is supplied to the semiconductor wafer through the outer probes and a voltage is measured across the inner probes.

For a thin sample, resistivity is given by:

$$\rho = V/I \times 4.54 t' \Omega.m \quad (2-11)$$

where  $V$  is the measured voltage,  $I$  is the current through the outer probes, and  $t'$  is the thickness of the wafer.

Note that the probe separation has been assumed to be small compared to the other dimensions.

Example 2.2: If  $\mu_p = 0.045 \text{ m}^2/\text{V.s}$  at 300 K, compute the drift velocity when  $E' = 100 \text{ V/m}$ .

Solution:

From Eqn.(2-5),  $v_d' = \mu_p E' = 0.045 \text{ m}^2/\text{V.s} \times 100 \text{ V/m} = 4.5 \text{ m/s}$ . #

Example 2.3: In a 4-point probe measurement, if  $V = 1 \text{ V}$  and  $I = 1 \text{ mA}$ , determine  $\rho$  if  $t' = 200 \text{ }\mu\text{m}$ .

Solution:

From Eqn.(2-11),  $\rho = V / I \times 4.54 t' \text{ }\Omega\cdot\text{m} = 1 \text{ V} / 1 \times 10^{-3} \text{ A} \times 4.54 \times 200 \times 10^{-6} \text{ m} = 0.91 \text{ }\Omega\cdot\text{m}$ . #

### C. Diffusion current

- Diffusion of carriers occurs in the presence of a carrier density gradient.

For electrons, the rate of electron flow per unit area  $\phi_n$  in 1-dimension is given by:

$$\phi_n = - D_n \text{ dn/dx} \quad (2-12)$$

where  $D_n$  is the electron diffusivity.

Electron diffusion current density is then given by:

$$J_n = - q\phi_n = qD_n \text{ dn/dx} \quad (2-13)$$

Similarly, hole diffusion current density is given by:

$$J_p = - qD_p \text{ dp/dx} \quad (2-14)$$

where  $D_p$  is the hole diffusivity.



Fig.2.4 shows the relationship between the diffusion currents and the carrier density gradients.

- Sometimes, diffusivity is expressed in terms of a mean free path  $\lambda$  for the carrier.

For electrons, we can write:  $D_n = v_{th}\lambda$ , where  $v_{th}$  is the thermal velocity. The mean free path may be expressed in terms of the collision time, i.e.,  $\lambda = v_{th}\tau_c$ .

a) Einstein relation

In 1-dimension, for a collection of electrons we have:  $m_e^* v_{th}^2 / 2 = kT/2$ , and  $D_n = v_{th}\lambda = v_{th}^2\tau_c = kT \tau_c/m_e^*$ .

Since  $\mu_n = q\tau_c/m_e^*$ ,

$$D_n = \mu_n kT/q \quad (2-15)$$

This equation is known as **Einstein relation**. Einstein relation allows one to relate  $D_n$  to  $\mu_n$ .

A similar expression exists for the holes.

Example 2.4: If  $\mu_n = 0.15 \text{ m}^2/\text{V.s}$  at 300 K for a semiconductor, compute  $D_n$ .

Solution:

From Eqn.(2-15):  $D_n = \mu_n kT/q = 0.15 \text{ m}^2/\text{V.s} \times 0.0259 \text{ V} = 3.89 \times 10^{-3} \text{ m}^2/\text{s}$ .  
#

b) Drift and diffusion current densities

- Currents for electrons and holes are due to drift and diffusion.

In 1-dimension, the electron and hole current densities are given by:

$$J_n = qn\mu_n E' + qD_n dn/dx$$

$$J_p = qp\mu_p E' - qD_p dp/dx \quad (2-16)$$

The total current density  $J_T$  is given by:

$$J_T = J_n + J_p \quad (2-17)$$

D. Carrier injection

In thermal equilibrium, we have  $pn = n_i^2$ .

- Carrier injection in a semiconductor results in:  $pn > n_i^2$ .

Assuming that the *equilibrium* carrier densities are given by:  $n_0$  and  $p_0$ , we can define the excess carrier densities as:  $\Delta n = n - n_0$ , and  $\Delta p = p - p_0$ .

- The extent of charge injection depends on the amount of excess carriers.

High-injection implies that:  $\Delta n \gg n_0$  and  $\Delta p \gg p_0$ , while low-injection implies either  $\Delta n \gg n_0$  and  $\Delta p \ll p_0$  (in a P-type semiconductor), or  $\Delta n \ll n_0$  and  $\Delta p \gg p_0$  (in an N-type semiconductor).

Example 2.5: If  $n_0 = N_D = 1 \times 10^{22} / \text{m}^3$  for a N-type Si sample, and  $\Delta n = \Delta p = 1 \times 10^{15} / \text{m}^3$ , what type of injection occurs.

Solution:

This will be low-injection as:  $p_0 = n_i^2 / N_D = (1.45 \times 10^{16} / \text{m}^3)^2 / 1 \times 10^{22} / \text{m}^3$  ( $p_0 \ll \Delta p$ ). #

## E. Generation and recombination in semiconductors

In addition to diffusion, other processes affecting the carrier densities are also present.

These include: **generation** and **recombination**.

- Generation may be due to: i) thermal generation; ii) optical generation; or iii) other types of carrier generation such as ionization, multiplication, etc.

- Recombination includes: i) direct recombination from band-to-band; and ii) indirect recombination via impurities.

In steady state, generation rate  $G$  equals to recombination rate  $R$ . This is illustrated in Fig.2-5.

## a) Band-to-band recombination

- Recombination depends on the densities of the recombinants.

In equilibrium, band-to-band recombination rate  $R$  is given by:

$$R = \beta^* np \quad (2-18)$$

where  $\beta^*$  is a constant.

Since at thermal equilibrium  $n = n_0$  and  $p = p_0$ , this leads to:  $R_{th} = \beta^* n_0 p_0$ .

- In equilibrium, thermal generation rate equals the recombination rate, i.e.,  $G_{th} = R_{th}$

Deviation from equilibrium results in net recombination or generation.

For low-injection where  $n = n_0$  and  $p = \Delta p + p_0$ , the net recombination rate  $U$  is:

$$U = R - R_{th} = \beta^* n_0 (\Delta p + p_0) - \beta^* n_0 p_0 = \beta^* n_0 \Delta p \quad (2-19)$$

If one defines  $U$  as  $\Delta p / \tau_p$ , where  $\tau_p$  is the hole lifetime, then we have:  $\tau_p = 1 / \beta^* n_0$ .

Similarly, for a p-type semiconductor, we have:  $\tau_n = 1 / \beta^* p_0$ , where  $\tau_n$  is the electron lifetime.

Example 2.6: If the excess hole density is  $1 \times 10^{15} /m^3$  and its lifetime is 100  $\mu s$ , determine the net recombination rate.

Solution:

From Eqn.(2-19):  $U = \Delta p/\tau_p = 1 \times 10^{15} /m^3/1 \times 10^{-4} s = 1 \times 10^{19} /m^3.s. \#$

- In steady state, the net recombination rate  $U$  is balanced by the generation rate  $G$  (assume  $G \gg G_{th}$ ) and  $G = \Delta p/\tau_p$ .

This leads to (for low-injection):

$$p = p_0 + G\tau_p \quad (2-20)$$

b) Dynamic response

The dynamic change of hole density in a semiconductor is given by:

$$dp/dt = G - R \quad (2-21)$$

In the absence of external generation,  $G = G_{th}$ , we have:  $dp/dt = -U = (p - p_0)/\tau_p$ .

This leads to:

$$p = p_0 + G\tau_p \exp(- t/\tau_p ) \quad (2-22)$$

where we assumed that the initial hole density is:  $p_0 + G\tau_p$ .

The hole density ( $= (p_0 + G\tau_p)$  at  $t = 0$ ) relaxes exponentially towards its equilibrium value  $p_0$ . This is illustrated in Fig.2-6.

Example 2.7: At  $t = 0^-$ ,  $G = 1 \times 10^{24} / \text{m}^3 \cdot \text{s}$ . Determine the hole density in a semiconductor at  $t = 1 \text{ ms}$  if  $G = 0$  when  $t > 0$ . Neglect the value of  $p_0$  and assume  $\tau_p = 0.1 \text{ ms}$ .

Solution:

Based on Eqn.(2-22):  $p = p_0 + G\tau_p \exp(-t/\tau_p) = 1 \times 10^{-4} \text{ s} \times 1 \times 10^{24} / \text{m}^3 \cdot \text{s} \exp(-1 \times 10^{-3} \text{ s} / 1 \times 10^{-4} \text{ s}) = 4.54 \times 10^{15} / \text{m}^3$ . #

### c) Indirect recombination

- Indirect recombination occurs at impurity centers with energy level(s) in the energy gap.

When compared with band-to-band recombination, such a process is favored because it involves smaller energy transitions.

Fig.2.7 shows a schematic of the recombination of an electron and a hole via an impurity center. The lifetime of electrons/holes varies with the density of the impurity center.

- To first order, the lifetime of carriers is inversely proportional to the impurity density  $N_I$ .

## d) Continuity equation

The 1-dimensional **continuity equations** for electrons and holes include drift, diffusion, generation and recombination. They are given by:

$$dn/dt = (1/q) dJ_n/dx + (G_n - R_n)$$

$$dp/dt = - (1/q) dJ_p/dx + (G_p - R_p) \quad (2-23)$$

The subscripts p and n denote the electron and hole contributions.

In full, these equations can be written as (for the minority carriers only):

$$dn_p/dt = n_p \mu_n dE'/dx + \mu_n E' dn_p/dx + D_n d^2 n_p/dx^2 + G_n - (n_p - n_{p0})/\tau_n$$

$$dp_n/dt = - p_n \mu_p dE'/dx - \mu_p E' dp_n/dx + D_p d^2 p_n/dx^2 + G_p - (p_n - p_{n0})/\tau_p$$

(2-24)

The product  $\sqrt{(D_p \tau_p)}$  ( $= L_p$ ) is the hole diffusion length.

Similarly, the electron diffusion length is given by:  $L_n = \sqrt{(D_n \tau_n)}$ .

Example 2.8: Write down the steady-state continuity equation for hole minority carriers when  $E' = 0$  and  $G_p = 0$ . Assume  $p_n = p_{n0}$  when  $x = \infty$ .

Solution:

From Eqn.(24): in the steady state:  $dp_n/dt = -p_n\mu_p dE'/dx - \mu_p E' dp_n/dx + D_p d^2 p_n/dx^2 + G_p - (p_n - p_{n0})/\tau_p = 0$ .

Since  $E' = 0$  and  $G_p = 0$ , only the last term remains, i.e.,  $D_p d^2 p_n/dx^2 = (p_n - p_{n0})/\tau_p$ . The solution to this equation will be:  $p_n = p_{n0} + (p_n(0) - p_{n0}) \exp(-x/\sqrt{(D_p\tau_p)})$ , where  $p_n(0)$  is the hole density when  $x = 0$ .

Example 2.9: Light is allowed to fall on 1 side of a semiconductor sample and  $p_n(0) = 1 \times 10^{22} /m^3$ . If  $\tau_p = 100 \mu s$  and  $D_p = 1 \times 10^{-3} m^2/s$ , determine  $p_n$  when  $x = 1 mm$ . Assume  $p_{n0} = 1 \times 10^{11} /m^3$ .

Solution:

From the previous example:  $p_n = p_{n0} + (p_n(0) - p_{n0})\exp(-x/\sqrt{(D_p\tau_p)}) = 1 \times 10^{11} /m^3 + (1 \times 10^{22} /m^3 - 1 \times 10^{11} /m^3) \times \exp(-1 \times 10^{-3} m/\sqrt{(1 \times 10^{-4} m^2/s \times 1 \times 10^{-3} s)}) = 4.2 \times 10^{20} /m^3$ . #

Example 2.10: Assume  $p_n = p_{n0}$  at  $x = w$ , determine the hole current density based on the conditions specified in Example 2.9. Assume  $w \ll L_p$ . Given:  $\sinh(u) = (\exp(u) - \exp(-u))/2$ .

Solution:

In this case:  $p_n = p_{n0} + (p_n(0) - p_{n0}) \sinh((w - x)/L_p)/\sinh(w/L_p)$ .

Since  $J_p = -qD_p dp_n/dx$  and  $w \gg L_p$ , it can be shown that:  $J_p = q(p_n(0) - p_{n0})D_p/w$ .

Note in this example that the approximation  $w \ll L_p$  will lead to a *linearly-graded* carrier density profile.



## **Part 3: P-N junctions**

### **What we need to learn.**

- a. Physical structure of a PN junction
- b. I-V characteristics
- c. Ideal and non-ideal currents
- d. Capacitances
- e. Breakdown

### Part 3: P-N junctions

- P-N junctions are formed when a P-type semiconductor and an N-type semiconductor are fused together.

A typical P-N junction is depicted in Fig.3.1.

Because of the differences in the electrical properties on the two sides of the P-N junction, physical changes will take place. One of the results is **rectification** in the current-voltage characteristics.

- Rectification is represented by *asymmetrical* current flow when the polarity of bias voltage is altered.

Fig.3.2 shows the rectification properties of a P-N junction.

#### A. Physical properties of the P-N junction

When a P-type semiconductor is connected to an N-type semiconductor, the following results:

- inter-diffusion of the majority carriers (see Fig.3.3)
- creation of a region depleted of carriers (with the exposure of donor and acceptor charges) (see Fig.3.4)
- creation of a *high-field* junction region (consequence of the exposed charges)

- creation of minority carrier density gradients beyond the edges of the depletion region

P-N junctions are best studied using an energy band diagram. Fig.3.5 shows the energy band diagram of a typical P-N junction.

At equilibrium, hole current density in 1-dimension (in the absence of generation and recombination) is given by:

$$J_p = q\mu_p E' - qD_p dp/dx \quad (3-1)$$

If one replaces  $E'$  by  $1/q dE_i/dx$ , and  $D_p$  by  $\mu_p kT/q$  (Einstein relation), this leads to:

$$J_p = q\mu_p E' - \mu_p kT dp/dx \quad (3-2)$$

Furthermore, since  $p = n_i \exp((E_i - E_F)/kT)$ ,  $dp/dx = p/kT (dE_i/dx - dE_F/dx)$ , one gets:

$$J_p = q\mu_p p dE_F/dx \quad (3-3)$$

There is no current flowing at equilibrium, we have  $dE_F/dx = 0$ .

This implies that the Fermi-level should be **flat** across the entire P-N junction. This is illustrated in Fig.3.6.

#### a) The built-in potential

From the energy band diagram (see Fig.3.6), one can see that there exists a “barrier” across the P-N junction.

This potential barrier is given by:

$$qV_{bi} = (E_i - E_F)|_{p\text{-side}} + (E_F - E_i)|_{n\text{-side}} \quad (3.4)$$

where  $V_{bi}$  the **built-in voltage**.

Since  $(E_i - E_F)|_{p\text{-side}} = kT \ln(N_A/n_i)$ , and  $(E_F - E_i)|_{n\text{-side}} = kT \ln(N_D/n_i)$ ,

$$qV_{bi} = kT \ln(N_D N_A/n_i^2) \quad (3.5)$$

This equation shows that the built-in potential is dependent on the dopant densities.

- The origin of the built-in potential is linked to the charges in the depletion region.

The exposed charges generate an electric field, which opposes electrons moving from the N-side into the P-side and holes from the P-side into the N-side.

Example 3.1: Compute  $V_{bi}$  (the built-in voltage) for a Si P-N junction if  $N_A = 1 \times 10^{24} /m^3$  and  $N_D = 1 \times 10^{21} /m^3$  at 300 K.

Solution:

From Eqn.(3-5), the built-in voltage  $V_{bi} = (kT/q) \ln(N_D N_A/n_i^2) = 0.0259 \text{ V} \times \ln((1 \times 10^{24} /m^3 \times 1 \times 10^{21} /m^3)/(1.45 \times 10^{16})^2) = 0.755 \text{ V}$ . #

b) The depletion region

- The region at/near the interface of the P-N junction where carriers are depleted is the depletion region.

In this section, we examine the depletion layer width, the built-in electric field, and the potential profile.

The simplest dopant profile for a P-N junction occurs when the dopant densities are uniform. This is called an abrupt junction.

Within the depletion region of an abrupt junction, there exists a dipolar charge layer due to the exposed donors and acceptors.

Using Poisson equation, one gets:

$$\begin{aligned} d^2\psi/dx^2 &= qN_A/\epsilon_s && \text{for the P-side} \\ d^2\psi/dx^2 &= -qN_D/\epsilon_s && \text{for the N-side} \end{aligned} \quad (3-6)$$

where  $\psi$  is the electrostatic potential, and  $\epsilon_s$  is the semiconductor permittivity.

Fig. 3.7 shows the potential profile in the depletion region of an abrupt junction.

Charge conservation requires that:

$$N_a x_p = N_d x_n \quad (3-7)$$

where  $x_p$  and  $x_n$  are the widths of the depletion region on either side of the P-N junction.

The depletion layer width  $W$  is given by:

$$W = x_p + x_n. \quad (3-8)$$

Integrating Poisson equation (Eqn.(3-6)) gives:

$$E' = - d\psi/dx = - qN_A(x + x_p)/\epsilon_s \quad \text{P-side}$$

$$E' = - d\psi/dx = qN_D(x + x_n)/\epsilon_s \quad \text{N-side} \quad (3-9)$$

where  $E'$  is the electric field.

It is obvious that the *maximum* electric field  $E_m'$  exists at  $x = 0$  (i.e., at the junction interface).

Further integration of the electric field (Eqn.(3-9)) gives the electrostatic potential  $\psi$ .

Assuming  $\psi = 0$  at  $x = x_p$ ,  $\psi = V_{bi}$  at  $x = -x_n$ , we have:

$$\psi_p = qN_A/\epsilon_s(x^2/2 - xx_p) + K_1$$

$$\psi_n = - qN_D/\epsilon_s(x^2/2 + xx_n) + K_2 \quad (3-10)$$

where  $\psi_p$  and  $\psi_n$  are the electrostatic potentials on the P-side and the N-side, respectively, and  $K_1$  and  $K_2$  are constants.

It can be shown that:  $K_1$  equals to  $K_2$  as  $\psi_p(x) = \psi_n(x)$  at  $x = 0$ .

The built-in potential is the integral of the electric field  $E'$  over the entire depletion region  $-x_n < x < x_p$ . It is given by:

$$V_{bi} = qN_A x_p^2 / (2\epsilon_s) + qN_D x_n^2 / (2\epsilon_s) = E_m' W / 2 \quad (3-11)$$

Rearranging this equation gives:

$$W = [(2\epsilon_s(N_A + N_D)V_{bi} / qN_A N_D)]^{1/2} \quad (3-12)$$

Fig.3.8 shows the electric field distribution and the potential profile across the entire P-N junction.

- In the presence of an applied voltage  $V_a$ ,  $V_{bi}$  (in the previous equations) is replaced by:  $V_{bi} - V_a$ .

Our notation is such that  $V_a$  is positive when the positive potential is connected to the P-side and is negative when the positive potential is connected to the N-side.

The former is called forward bias, while the latter reverse bias.

- Depletion layer widens with increasing reverse bias and narrows with increasing forward bias.

In some instances, the dopant densities across the P-N junction are not uniform and depletion is primarily onto 1 side. This is a 1-sided step junction and is illustrated in Fig.3.9.

In the case when  $N_A \gg N_D$ , we have:

$$\begin{aligned} W &\approx x_n = \sqrt{[2\epsilon_s V_{bi}/(qN_D)]} \\ E' &= -qN_D W(1 + x/W)/\epsilon_s \\ \psi &= V_{bi}(-2x - x^2/W)/W \end{aligned} \quad (3-13)$$

Example 3.2: For a 1-sided step junction with  $N_A = 1 \times 10^{25} /m^3$  and  $N_D = 1 \times 10^{21} /m^3$ , compute the value of  $W$  and  $E_m'$  at  $T = 300$  K.

Solution:

From Eqn.(3-5),  $V_{bi} = (kT/q) \ln(N_A N_D / n_i^2) = 0.0259 \text{ V} \times \ln(10^{25} \times 10^{21} / (2.1 \times 10^{32})) = 0.874 \text{ V}$ .

From Eqn.(3-13),  $W \approx x_n = \sqrt{(2\epsilon_s V_{bi} / (qN_D))} = 3.37 \times 10^{-7} \text{ m}$ .  $E_m' = -qN_D W / \epsilon_s = -5.4 \text{ MV/m}$ . #

### c) Depletion capacitance

- The donor and acceptor charge layers in the P-N junction give rise to the *depletion/junction capacitance*.

This is illustrated in Fig.3.11

The depletion capacitance per unit area  $C_j$  is given by:

$$C_j = dQ/dV_a = \epsilon_s/W \quad (3-14)$$

where  $Q$  is the charge density per unit area.

For a 1-sided step junction:

$$C_j = \sqrt{[q\epsilon_s N_B / (2(V_{bi} - V_a))]}^{1/2} \quad (3-15)$$



Fig.3.12 shows a typical plot of  $1/C^2$  versus  $V_a$ .

Example 3.4: An abrupt P-N junction has a doping concentration of  $10^{21}$  /m<sup>3</sup> on the lightly doped N-side and of  $10^{25}$  /m<sup>3</sup> on the heavily doped P-side. From a plot of  $1/C^2$  versus  $V_a$ , comment on values of the slope and the intercept on the voltage axis.

Solution:

$N_B = N_D = 10^{21}$  /m<sup>3</sup>. The governing equation (Eqn.(3-15)) is  $1/C^2 = 2(V_{bi} - V_a)/(q\epsilon_s N_B)$ . Such a plot will give a straight line with a negative slope and the interception on the voltage axis is  $V_{bi}$ . #

## B. I-V characteristics

- I-V characteristics of a P-N junction behave differently under forward bias versus reverse bias.

*Ideal* I-V characteristics are obtained under the following assumptions:

- Abrupt junction
- Carrier densities at the junction boundaries are directly related to the electrostatic potential
- Low-injection
- No generation and recombination current in the depletion region

With bias, the electron density in the N-side at equilibrium is given by (see Eqn.(1-10)):

$$n_{n0} = n_i \exp((E_F - E_{in})/kT) \quad (3-16)$$

where  $E_{in}$  is the intrinsic Fermi-level on the N-side.

On the P-side, one has:

$$n_{p0} = n_i \exp((E_F - E_{ip})/kT) \quad (3-17)$$

where  $E_{ip}$  is the intrinsic Fermi level on the P-side.

The subscript “0” stands for the case of equilibrium.

Since  $E_{ip} - E_{in} = qV_{bi}$ , we have:

$$n_{n0} = n_{p0} \exp(qV_{bi}/kT) \quad (3-18)$$

For an applied voltage equal to  $V_a$ , the Fermi-level will split up as shown in Fig.3.13.

- Electron density at the edge of the depletion region depends exponentially on the built-in voltage.

We can now replace  $V_{bi}$  by  $V_{bi} - V_a$ ;  $n_{n0}$  by  $n_n$ ; and  $n_{p0}$  by  $n_p$  in our equations.

This leads to:

$$n_n = n_p \exp(q(V_{bi} - V_a)/kT) \quad (3-19)$$

At low-injection,  $n_n = n_{n0}$  and we can equate Eqns.(3-18) and (3-19). This leads to:

$$n_p = n_{p0} \exp(qV_a/kT) \quad (3-20)$$

From Eqn.(3-20), we see that across the P-N junction, the minority carrier density is increased by  $\exp(qV_a/kT)$ .

A similar situation exists in the N-side and we have:

$$p_n = p_{n0} \exp(qV_a/kT) \quad (3-21)$$

To determine the diffusion currents, we combine Eqns.(3-20), (3-21) and the continuity equation. This gives:

$$\begin{aligned} D_n \frac{d^2 n_p}{dx^2} - (n_p - n_{p0})/\tau_n &= 0 \\ D_p \frac{d^2 p_n}{dx^2} - (p_n - p_{n0})/\tau_p &= 0 \end{aligned} \quad (3.22)$$

The boundary conditions are:

$$x = y = 0: \quad p_n = p_{n0} \exp(qV_a/kT); \quad \text{and} \quad n_p = n_{p0} \exp(qV_a/kT)$$

$$x = y = \infty: \quad p_n = p_{n0}; \quad \text{and} \quad n_p = n_{p0}$$

They result in:

$$p_n - p_{n0} = p_{n0} [\exp(qV_a/kT) - 1] \exp(-x/L_p)$$

$$n_p - n_{p0} = n_{p0} [\exp(qV_a/kT) - 1] \exp(-y/L_n) \quad (3-23)$$

These excess carrier densities are plotted in Fig.3.14.

The above equations also lead to:

$$\begin{aligned} J_p &= -qD_p \frac{dp_n}{dx} = qD_p p_{n0}/L_p [\exp(qV_a/kT) - 1] \\ J_n &= qD_n \frac{dn_p}{dx} = qD_n n_{p0}/L_n [\exp(qV_a/kT) - 1] \end{aligned} \quad (3-24)$$

where  $J_n$  and  $J_p$  are the electron and hole current densities, respectively.

The total current density  $J_T$  is:

$$J_T = J_n + J_p = J_s [\exp(qV_a/kT) - 1] \quad (3-25)$$

where  $J_s = qD_p p_{n0}/L_p + qD_n n_{p0}/L_n$ .

This is called the **ideal diode equation** and the I-V characteristics are shown in Fig.3.15.

Example 3.5: An ideal P-N junction has  $N_D = 10^{24} /m^3$  and  $N_A = 10^{22} /m^3$ ,  $\tau_p = \tau_n = 1 \mu s$  and a device area of  $1.2 \times 10^{-9} m^2$ , calculate the ideal saturation current at 300 K. (It can be shown that:  $D_p = 0.00116 m^2/s$  and  $D_n = 0.00388 m^2/s$ .)

Solution:

For an ideal device, the recombination current is assumed to be zero.

From Eqn.(3-24), we can out it into the form:  $I_s = J_s \times A_{cs} = (qD_p n_i^2 / (L_p N_D) + qD_n n_i^2 / (L_n N_A)) \times A_{cs} = (q \times n_i^2 \times A_{cs}) \times ((D_p / \tau_p)^{0.5} N_D + (D_n / \tau_n)^{0.5} N_A) = (1.6 \times 10^{-19} \text{ C} \times 2.1 \times 10^{32} / \text{m}^6 \times 1.2 \times 10^{-9} \text{ m}^2) \times ((0.00116 \text{ m}^2/\text{s}/10^{-6} \text{ s})^{0.5}/10^{24} / \text{m}^3) + ((0.00388 \text{ m}^2/\text{s}/10^{-6} \text{ s})^{0.5}/10^{22} / \text{m}^3) = 2.49 \times 10^{-16} \text{ A.} \quad \#$

Example 3.6: For the device in Example 3.5, compute the forward current at  $V_a = +0.7 \text{ V}$

Solution:

From Eqn.(3-25),  $I_F = I_s \times \exp(qV_a/kT) = 2.49 \times 10^{-16} \times \exp(0.7 \text{ V}/0.0259 \text{ V})$   
 $A = 1.36 \times 10^{-4} \text{ A.} \quad \#$

#### a) Other current contributions

Other than the diffusion currents (which occur outside the depletion region), there are also generation current and recombination current within the depletion region.

- Generation current is due to thermal generation of electrons and holes in the depletion region.

The process is illustrated in Fig.3.16. It is important during reverse bias when the depletion layer width is substantially widened.

Assuming a thermal generation rate given by  $G$  and a generation lifetime  $\tau_g$ , one can express  $G = n_i \tau_g$ .

This leads to a generation current density  $J_{gen}$  given by:

$$J_{\text{gen}} = \int_0^W qG \, dW = qn_i W/\tau_g \quad (3-26)$$

The total reverse current density in the presence of generation is:

$$J_R = J_s + J_{\text{gen}} = qD_p p_{n0}/L_p + qD_n n_{p0}/L_n + qn_i W/\tau_g \quad (3-27)$$

- Recombination current exists primarily in the depletion region.

A typical recombination process via an impurity center is shown in Fig.3.17. It differs from the generation process in the sense that it is important under forward bias.

Assuming a carrier recombination rate given by  $U = (\Delta(pn) - n_i^2)/n_i\tau_r$ , where  $\Delta(pn)$  is the excess of the product of  $p$  and  $n$ , one can write:

$$U = n_i [\exp(qV_a/kT) - 1]/\tau_r \quad (3-28)$$

where  $\tau_r = [n_n + p_n + 2n_i \cosh(E_i - E_t)]/(n_i \sigma_0 v_{th} N_t)$ ,  $E_t$  is the energy of the recombination center,  $N_t$  is the density of the recombination center,  $v_{th}$  is thermal velocity, and  $\sigma_0$  is the capture cross-section of the recombination center.

Note that we have assumed the carrier capture cross-sections for the electron and the hole are the same and are equal to  $\sigma_0$ .

- Theoretically, it can also be shown that the most effective recombination center is located when  $E_t = E_i$ .

This statement together with the assumption  $p_n = n_n = n_i \exp(qV_a/2kT)$  (a situation chosen for maximum recombination probability) gives:

$$U_{\max} \approx \sigma_0 v_{th} N_t n_i \exp(qV_a/2kT)/2 \quad (3-29)$$

where we have assumed  $V_a > 3kT/q$ .

The recombination current density  $J_r$  is:

$$J_r = q \int_0^W U_{\max} dW = q U_{\max} W \quad (3-30)$$

The total forward current density now becomes:

$$J_F = J_s \exp(qV_a/kT) + qW\sigma_0 v_{th} N_t n_i \exp(qV_a/2kT)/2 \quad (3-31)$$

In general,  $J_F$  is proportional to  $\exp(qV_a/\eta kT)$ , where  $\eta$  the ideality factor varies between 1 to 2. This is illustrated in Fig.3.18.

- Within the P-N junction, high-injection results in current flow different from the normal currents.

During high-injection, we have  $p_n = n_n = n_i \exp(qV_a/2kT)$  and the total current density  $J_T$  is given by:

$$J_T = J_0 [\exp(qV_a/2kT) - 1] \quad (3-32)$$

Thus, the current density increases at a slower rate with increasing bias voltage at high-injection. A typical plot showing high-injection in a P-N junction is given in Fig.3.19.

Example 3.7: For an ideal P-N junction with  $N_A = 10^{23} /\text{m}^3$  and  $N_D = 10^{21} /\text{m}^3$ . Assume that it contains  $10^{21} /\text{m}^3$  generation and recombination centers located 0.02 eV above the intrinsic Fermi level with  $\sigma_n = \sigma_p = 10^{-19} \text{ m}^2$ . For  $v_{th} = 1 \times 10^5 \text{ m/s}$ , calculate the generation current at  $V_a = -0.5 \text{ V}$ .  $A_{cs} = 10^{-8} \text{ m}^2$ ,  $\tau_g = 0.27 \mu\text{s}$ , and  $\tau_p = 0.1 \mu\text{s}$ .

Solution:

From Eqn.(3-5),  $V_{bi} = kT/q \times \ln(N_A N_D / n_i^2) = 0.0259 \text{ V} \times \ln(10^{23} /\text{m}^3 \times 10^{21} /\text{m}^3 / 2.1 \times 10^{32} /\text{m}^6) = 0.695 \text{ V}$ .

From Eqn.(3-13),  $W = (2\varepsilon_s(V_{bi} + V_R)/(qN_D))^{0.5} = (2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} \times (0.695 + 0.5) \text{ V} / (1.6 \times 10^{-19} \text{ C} \times 10^{21} /\text{m}^3))^{0.5} = 1.25 \times 10^{-6} \text{ m}$ .

From Eqn.(3-26),  $I_{gen} = J_{gen} \times A_{cs} = q n_i W A_{cs} / \tau_g = 1.6 \times 10^{-19} \text{ C} \times 1.45 \times 10^{16} /\text{m}^3 \times 1.25 \times 10^{-6} \text{ m} \times 10^{-8} \text{ m}^2 / 0.27 \times 10^{-6} \text{ s} = 1.1 \times 10^{-10} \text{ A}$ . #

Example 3.8: With Example 3.7, repeat the calculations for the recombination current when  $V_a = 0.5 \text{ V}$ . Assume  $\tau_r (= \sigma_0 v_{th} N_t / 2) = 0.1 \mu\text{s}$ .

Solution:

From Eqn.(3-13),  $W = (2\varepsilon_s(V_{bi} - V_a)/(qN_D))^{0.5} = (2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} \times (0.695 \text{ V} - 0.5 \text{ V}) / (1.6 \times 10^{-19} \text{ C} \times 10^{21} /\text{m}^3))^{0.5} = 0.5 \times 10^{-6} \text{ m}$ .

From Eqns.(29) and (30),  $I_{rec} = qW n_i A_{cs} \exp(qV_a/2kT) / \tau_g = 1.6 \times 10^{-19} \text{ C} \times 0.5 \times 10^{-6} \text{ m} \times 1.45 \times 10^{16} /\text{m}^3 \times 10^{-8} \text{ m}^2 \times \exp(0.5 \text{ V} / (2 \times 0.0259 \text{ V})) / 0.1 \times 10^{-6} \text{ s} = 1.6 \times 10^{-6} \text{ A}$ . #



## b) Charge storage

- Under forward bias, minority carriers are stored in the P-N junction.

Such charges will give rise to diffusion capacitance.

In a forward bias P-N junction, the hole charge per unit area stored in the N-region is given by:

$$\begin{aligned}
 Q_p &= q \int_{x_n}^{\infty} (p_n - p_{n0}) dx \\
 &= q \int_{x_n}^{\infty} p_{n0} (\exp(qV_a/kT) - 1) \exp(-(x - x_n)/L_p) dx \\
 &= qL_p p_{n0} (\exp(qV_a/kT) - 1)
 \end{aligned} \tag{3-33}$$

Since  $J_p = qD_p p_{n0}/L_p [\exp(qV_a/kT) - 1]$ , one can write:

$$Q_p = L_p^2 J_p / D_p = \tau_p J_p \tag{3-34}$$

This is illustrated in Fig.3.20.

Similarly,  $Q_n = \tau_n J_n$ . Thus,

$$Q_T (= Q_p + Q_n) = \tau_p J_p + \tau_n J_n. \tag{3-35}$$

If  $\tau_p = \tau_n$ ,  $Q_T \propto J_T$ .

The diffusion capacitance per unit area  $C_d$  is given by:

$$\begin{aligned} C_d &= \Delta Q_p / \Delta V_a \\ &= q^2 L_p p_{n0} / kT \exp(qV_a / kT) \end{aligned} \quad (3-36)$$

Note that diffusion capacitance increases with increasing forward bias. This is illustrated in Fig.3.21.

Fig.3.22 shows the equivalent circuit of a P-N junction including the capacitances.

#### d) Junction breakdown

- Breakdown occurs when there is excess current flowing through the P-N junction.

Normally, breakdown is reversible provided there is a large external resistance that limits the current.

In a P-N junction, breakdown is normally either due to tunneling or avalanche multiplication. These processes are shown in Figs.3.23 and 3.24.

Tunneling occurs in heavily doped P-N junctions. At reverse bias, tunneling occurs when the *filled states* in the P-side is right opposite to the *unfilled states* in the N-side so that electrons can

tunnel through without energy change. This can result in a large current.

Avalanche multiplication is due to impact ionization of the carriers. Under a strong reverse bias, the electrons and holes gain enough energy before suffering a collision and upon impact with an atom, additional electrons and holes will be generated. Such a multiplication process is called *avalanche multiplication*.

Fig. 3.25 shows the (carrier) ionization rates during breakdown under different field intensities.

To estimate the breakdown voltage, one needs to define a critical field  $E_c'$ .

The *breakdown voltage*  $V_B$  is:

$$V_B = E_C'W/2 \quad (3-37)$$

Assuming  $E_C' = E_m' = qN_BW/\epsilon_s$  in the case of a 1-sided step junction, one gets:

$$V_B = E_C'^2 \epsilon_s / (2qN_B) \quad (3-38)$$

where  $N_B$  is the substrate dopant density.

Note that the breakdown voltage is inversely proportional to the substrate dopant density.

Fig.3.26 shows a plot of critical electric field versus the substrate dopant density.

Example 3.9: For a 1-sided step junction, what is the critical breakdown voltage when tunneling becomes important?

Solution:

From Fig.3.26, tunneling becomes important when  $E_C' = 8 \times 10^7$  V/m for  $N_B = 5 \times 10^{23}$  /m<sup>3</sup>. This leads to (from Eqn.(3.28)):  $V_B = E_C'^2 \epsilon_s / (2qN_B) = (8 \times 10^7 \text{ V/m})^2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} / (2 \times 1.6 \times 10^{-19} \text{ Coulomb} \times 5 \times 10^{23} \text{ /m}^3) \text{ V} = 4.21 \text{ V.} \quad \#$

## **Part 4: Bipolar transistor**

### **What we need to learn.**

Current components in a bipolar transistor

- a. Gain parameters
- b. Modes of operation
- c. Frequency characteristics
- d. Switching characteristics

## **Part 4: Bipolar transistor**

- The word transistor refers to a transfer-resistance device.

This implies that the terminal resistances at the input and the output of the transistor are different.

Structurally, a bipolar transistor is made up of 2 back-to-back P-N junctions. It is therefore possible to have either a NPN or a PNP transistor.

For simplicity we shall examine a 1-dimensional PNP transistor.

This will be a 3 terminal device as shown in Fig.4.1. We label the terminals as emitter, base and collector. The dopant densities in each of the 3 regions will be different.

The operation of the transistor depends on the voltage bias (which characterizes the different modes of operation):

- For normal active operation, the emitter-base is forward biased and the collector base reverse biased. Its band diagram is shown in Fig.4.2.

Note that the splitting of the Fermi-levels is indicative of the different biases.

### A. Current components in the active transistor

The following occurs during the formation of the PNP transistor:

- Hole current flowing from the emitter into the base

A portion of this current reaches the base-collector junction and they are labeled as  $I_{Ep}$  and  $I_{Cp}$ .

- Recombination of the holes occurs in the base
- Electrons flow from the collector into the base due to the reverse bias and electrons flow into the emitter from the base due to the forward bias

The currents are labeled as  $I_{Cn}$ ,  $I_{En}$  and they are shown in Fig.4.3.

The total emitter current  $I_E$  is given by:

$$I_E = I_{Ep} + I_{En} \quad (4-1)$$

The total collector current  $I_C$  is given by:

$$I_C = I_{Cp} + I_{Cn} \quad (4-2)$$

Since the transistor is a 3 terminal device, the base current  $I_B$  is:

$$I_B = I_E - I_C = I_{Ep} + I_{En} - I_{Cp} - I_{Cn} \quad (4-3)$$

These are illustrated in Fig.4.4.

- The efficiency of a PNP transistor is determined by the amount of hole current reaching the collector.

This is expressed in terms of the common-base current gain  $\alpha_0$  given by.

$$\begin{aligned}\alpha_0 &= I_{Cp}/I_E = I_{Cp}/(I_{Ep} + I_{En}) \\ &= I_{Ep}/(I_{Ep} + I_{En}) \times I_{Cp}/I_{Ep}\end{aligned}\quad (4-4)$$

The first term appearing in Eqn.(4-4) is the emitter efficiency  $\gamma$ . and the second term is the transport factor  $\alpha_T$ .

We can write:  $\alpha_0 = \gamma \alpha_T$ .

Furthermore, one has:

$$I_C = \alpha_0 I_E + I_{Cn} \quad (4-5)$$

The collector current in the active mode consists of:

- a fraction of the emitter current
- generation and diffusion currents in the reverse bias base collector junction

Since  $\alpha_0$  is always less than 1 and  $I_{Cn}$  is small,  $I_C < I_E$ .

#### a) Small-signal characteristics

The small signal characteristics are obtained under the following assumptions:

- dopant densities are uniform



- low-injection
- absence of generation and recombination currents in the depletion regions
- absence of series resistance

Due to current injection from the emitter, the 1-dimensional hole (minority carrier) density in the base is given by:

$$D_p \frac{d^2 p_n}{dx^2} - (p_n - p_{n0})/\tau_p = 0 \quad (4-6)$$

This equation is subject to the following boundary conditions:

$$p_n(0) = p_{n0} \exp(qV_{EB}/kT)$$

$$p_n(W_B) = 0 \quad (4-7)$$

where  $x = 0$  is at the edge of the emitter-base depletion region in the base, and  $W_B$  is the base width.

The general solution is:

$$p_n = p_{n0} (\exp(qV_{EB}/kT) - 1) \frac{\sinh((W_B - x)/L_p)}{\sinh(W_B/L_p)} + p_{n0} (1 - \sinh(x/L_p)/\sinh(W_B/L_p)) \quad (4-8)$$

where we have made use of the relationship:  $L_p^2 = D_p \tau_p$ .

Eqn.(4-8) gives the excess minority carrier distribution in the base which is shown in Fig.4.5. Note that  $\sinh(z) = (\exp(z) - \exp(-z))/2$ .

For most P-N junctions,  $W_B < L_p$  and we have:

$$p_n = p_{n0} (\exp(qV_{EB}/kT) - 1) (1 - x/W_B) \quad (4-9)$$

This is the excess minority carrier density distribution in the base.

The excess minority carriers stored in the base  $Q_B$  is given by:

$$Q_B = qA_{cs}W_B p_{n0} [\exp(qV_{EB}/kT) - 1]/2 \quad (4-10)$$

where  $A_{cs}$  is the area cross-section of the base.

In a similar manner, we can obtain the minority carrier density distributions in the emitter and the collector.

These are shown in Fig.4.6.

Their values are:

$$\begin{aligned} n_E &= n_{E0} + n_{E0} (\exp(qV_{EB}/kT) - 1) \exp((x + x_E)/L_E) \quad x, -x_E \\ n_C &= n_{C0} - n_{C0} \exp(-(x - x_C)/L_C) \quad x' > x_C \end{aligned} \quad (4-11)$$

where  $n_E$  is the electron density in the emitter,  $n_{E0}$  is the equilibrium value of  $n_E$ ,  $-x_E$  is the edge of the depletion region in the emitter side,  $L_E$  is the diffusion length for the electrons in the emitter,  $n_C$  is the electron density in the collector,  $n_{C0}$  is the equilibrium value of  $n_C$ ,  $W_B + x_C$  is the edge of the depletion region in the collector side, and  $L_C$  is the diffusion length for the electrons in the collector.

Note that  $x$  is negative in the emitter.

We can now compute the diffusion currents in the different region:

$$\begin{aligned}
 I_{Ep} &= A_{cs} [-qD_p dp_n/dx]_{x=0} \\
 &= (qA_{cs}D_p p_{n0}/L_p) \coth(W_B/L_p) [(\exp(qV_{EB}/kT) - 1) + \\
 &\quad 1/\cosh(W_B/L_p)] \quad (4-12)
 \end{aligned}$$

where we have the relationship:  $\cosh(z) = (\exp(z) + \exp(-z))/2$  and  $\coth(z) = \cosh(z)/\sinh(z)$ .

For  $W_B < L_p$ ,

$$I_{Ep} \approx (qA_{cs} D_p p_{n0}/W_B)(\exp(qV_{EB}/kT) - 1) + qA_{cs}D_p n_i^2/(N_B W_B) \quad (4-13)$$

Similarly, at  $x = W_B$ ,

$$\begin{aligned}
 I_{Cp} &= qA_{cs} D_p p_{n0}/\{L_p \sinh(W_B/L_p)\} [(\exp(qV_{EB}/kT) - 1) + \\
 &\quad \cosh(W_B/L_p)] \\
 &\approx (qA_{cs} D_p p_{n0}/W_B) [\exp(qV_{EB}/kT) - 1] \quad (4-14)
 \end{aligned}$$

Thus,  $I_{Ep} \approx I_{Cp}$ .

For the electron current (assuming  $D_E = D_C = D_n$ ),

$$\begin{aligned}
I_{En} &= A_{cs} [-qD_E dn_E/dx]_{x=-x_E} \\
&= qA_{cs} D_E n_{E0} (\exp(qV_{EB}/kT) - 1)/L_E \\
I_{Cn} &= A_{cs} [-qD_C dn_C/dx]_{x=-x_C} \\
&= qA_{cs} D_C n_{C0}/L_C
\end{aligned} \tag{4-15}$$

where  $D_E$  is the electron diffusivity in the emitter, and  $D_C$  is the electron diffusivity in the collector.

We can write:

$$I_E = a_{11} (\exp(qV_{EB}/kT) - 1) + a_{12} \tag{4-16}$$

When  $W_B < L_p$ , we have:

$$\begin{aligned}
a_{11} &= qA_{cs} [D_p n_i^2 / (N_B W_B) + D_E n_{E0} / L_E] \\
a_{21} &= qA_{cs} D_p n_i^2 / (N_B W_B)
\end{aligned} \tag{4-17}$$

where  $N_B$  is the dopant density in the base.

Note that  $I_E = a_{12}$  when  $V_{EB} = 0$ .

This will be the diffusion current in the base when  $p_n(0) = p_{n0}$  and  $p_n(W_B) = 0$ .

Similarly, we can write:

$$I_C = a_{21} (\exp(qV_{EB}/kT) - 1) + a_{22} .$$

For  $w_B > L_p$ ,

$$\begin{aligned} a_{21} &= qA_{cs} D_p n_i^2 / (N_B W_B) \\ a_{22} &= qA_{cs} [D_p n_i^2 / (N_B W_B) + D_C n_{C0} / L_C] \end{aligned} \quad (4-18)$$

In addition,

$$I_B = I_E - I_C \quad (4-19)$$

Since  $Q_B = qA_{cs} W_B p_{n0} (\exp(qV_{EB}/kT) - 1)/2$ , it can be shown that both  $I_E$  and  $I_C$  are proportional to  $Q_B$ , the base charge.

It should be pointed out that the base current actually arises from carrier recombination as the carriers transit through the base.

From Eqn.(4-6):  $D_p d^2 p_n / dx^2 - (p_n - p_{n0}) / \tau_p = 0$ ,  $J_p = -qD_p dp_n / dx =$  constant if recombination is absent. The base current is therefore suppressed when  $\tau_p$  is large.

Finally, we can evaluate the emitter efficiency and the transport factor. They are given as:

$$\begin{aligned} \gamma &= I_{Ep} / (I_{Ep} + I_{En}) \\ &= 1 / (1 + I_{En} / I_{Ep}) \\ &\approx 1 / [1 + (qA_{cs} D_{En} E_0 / L_E) (\exp(qV_{EB}/kT) - 1) / \{(qA_{cs} D_p p_{n0} / W_B)\}] \end{aligned}$$

$$\begin{aligned}
 & (\exp(qV_{EB}/kT) - 1)\} \\
 & \approx 1/(1 + D_E n_{E0} W_B / (D_p p_{n0} L_E)) \quad (4-20)
 \end{aligned}$$

Thus,

$$\begin{aligned}
 \alpha_T &= I_{Cp} / I_{Ep} \\
 &\approx \{qA_{cs} D_p p_{n0} (\exp(qV_{EB}/kT) - 1) / (L_p \sinh(W_B/L_p))\} / \{[qA_{cs} \\
 & D_p p_{n0} / W_B] (\exp(qV_{EB}/kT) - 1)\} \\
 &\approx 1 - W_B^2 / (2L_p^2) \quad (4-21)
 \end{aligned}$$

- For an efficient device, it is desirable to have both  $\gamma$  and  $\alpha_T$  close to 1

Example 4.1: A silicon P<sup>+</sup>-N-P transistor has impurity densities of  $5 \times 10^{24} / \text{m}^3$ ,  $10^{22} / \text{m}^3$  and  $10^{21} / \text{m}^3$  in the emitter, base and collector, respectively. The base width is  $1 \mu\text{m}$  and the device cross-section is  $3 \text{ mm} \times 3 \text{ mm}$ . When the emitter-base junction is forward-biased to  $+0.5 \text{ V}$  and the base-collector junction is reverse-biased to  $-5 \text{ V}$ , calculate the neutral base width (physical base width minus the depletion layer widths).

Solution:

From Eqn.(3-5):  $V_{bi} = kT/q \times \ln(N_C N_V / n_i^2) = 0.0259 \text{ V} \times \ln(10^{22} / \text{m}^3 \times 10^{21} / \text{m}^3 / 2.1 \times 10^{26} / \text{m}^3) = 0.636 \text{ V}$ .

The depletion layer width in the base-collector junction =  $(2\epsilon_s(N_{\text{base}} + N_{\text{collector}})(V_{bi} + V_{CB}) / (qN_{\text{base}}N_{\text{collector}}))^{0.5} = (2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} \times (10^{22}$

$$/m^3 + 10^{21} /m^3 \times (0.636 V + V)/(1.6 \times 10^{-19} C \times 10^{22} /m^3 \times 10^{21} /m^3)^{0.5} = 2.85 \mu m.$$

The depletion region on the collector side that extends into the base =  $W_{BC}/(1 + N_{base}/N_{collector}) = 2.85 \mu m/(1 + 10^{22} /m^3/10^{21} /m^3) = 0.26 \mu m.$

A similar calculation for the emitter-base junction gives a depletion region in the base of  $0.217 \mu m.$

The neutral base width =  $(1 - 0.26 - 0.217) \mu m = 0.523 \mu m. \quad \#$

Example 4.2: For the transistor in Example 4.1, assume that the diffusivities of the minority carriers are  $0.0002, 0.001$  and  $0.0035 m^2/s$  and the corresponding lifetimes are  $0.01, 0.1$  and  $1 \mu s$ , respectively, determine the current components:  $I_{Ep}, I_{Cp}$  and  $I_B.$

Solution:

It can be shown that:  $L_{En} = 1.41 \mu m, L_{Bp} = 10 \mu m$  and  $L_{Cn} = 59.2 \mu m.$

From Eqn.(4-12):  $I_{Ep} = qA_{cs}D_p p_{n0}/L_p \coth(W_B/L_p) [(exp(qV_{EB}/kT) - 1) + 1/\cosh(W_B/L_p)] = 1.6 \times 10^{-19} C \times (3 \times 10^{-3} m)^2 \times 0.0002 m^2/s \times 2.1 \times 10^{26} /m^6/10^{22} /m^3 \times \coth(0.523 \mu m/10 \mu m) \times (exp(0.5 V/0.0259 V) - 1) + 1/(\cosh(0.523 \mu m/10 \mu m)) = 4.6714 mA. \quad \#$

Similarly, it can be shown that (based on Eqn.(4-13)):  $I_{Cp} = 4.6650 mA. \quad \#$

$$I_B = I_{Ep} - I_{Cp} = 6.38 \mu A. \quad \#$$

Example 4.3: If  $N_A$  in the emitter is  $10^{25} /m^3, N_D$  in the base is  $10^{23} /m^3$  and  $N_A$  in the collector is  $5 \times 10^{21} /m^3$ , compute  $\alpha_0$  for  $D_E = 10^{-4} m^2/V.s, D_p = 10^{-3} m^2/V.s, L_E = 10^{-6} m, L_p = 10^{-5} m,$  and  $W_B = 0.5 \times 10^{-6} m.$

Solution:

From Eqn.(4-20):  $\gamma \approx 1/(1 + D_{E0}n_{E0}W_B/(D_p p_{n0}L_E)) = 1/(1 + 10^{-4} \text{ m}^2/\text{s} \times 10^{23} / \text{m}^3 \times 0.5 \times 10^{-6} \text{ m} / (10^{-3} \text{ m}^2/\text{s} \times 10^{25} / \text{m}^3 \times 10^{-6} \text{ m})) = 0.9995$ .  $\alpha_T \approx 1 - W_B^2/(2L_p^2) = 0.9987$ .

Therefore,  $\alpha_0 = \gamma\alpha_T = 0.9982$ . #

## B. Modes of operation

The following are the different modes of operation for the bipolar transistor:

- Active mode -  $V_{EB}$  is at forward bias while  $V_{BC}$  is at reverse bias
- Saturation mode -  $V_{EB}$  is at forward bias while  $V_{BC}$  is also at forward bias
- Cutoff mode -  $V_{EB}$  is at reverse bias while  $V_{BC}$  is also at reverse bias
- Inverted mode -  $V_{EB}$  is at reverse bias while  $V_{BC}$  is at forward bias

Fig. 4.7 shows the output characteristics of a typical bipolar transistor showing the active region, saturation and cutoff.

Active mode corresponds to the case when the output behaves as a current source dependent on the value of the input voltage.



Saturation occurs when both the input junction and output P-N junction are at forward bias and there is very little resistance across the terminals of the transistor.

Cutoff implies that little or no current is allowed to pass through the terminals.

#### D. Equivalent circuit models for the bipolar transistor

Eber-Moll model - This model is constructed using 2 back-to-back P-N junctions as shown in Fig.4.14.

The governing equations are:

$$I_E = I_{F0} (\exp(qV_{EB}/kT) - 1) - \alpha_R I_{R0} (\exp(qV_{CB}/kT) - 1)$$

$$I_C = \alpha_F I_{F0} (\exp(qV_{EB}/kT) - 1) - I_{R0} (\exp(qV_{CB}/kT) - 1) \quad (4-24)$$

where  $I_{F0}$  is the forward saturation current of the emitter-base junction,  $I_{R0}$  is the reverse saturation current of the base-collector junction,  $\alpha_F$  is the forward common-base current gain, and  $\alpha_R$  is the reverse common-base current gain.

The above equations can apply to the different modes of operation including the use of large signals.

#### E. Frequency response of the bipolar transistor

The response of the bipolar transistor can be separated into the dc and ac components.

The frequency response is usually included in the small-signal analysis. In general, the currents and the terminal voltages can be expressed as (Total = dc + ac (small signal)):

$$i_E = I_E + i_E'$$

$$V_{EB} = V_{EB} + v_{EB}'$$

$$i_C = I_C + i_C'$$

$$V_{CB} = V_{CB} + v_{CB}'$$

$$i_B = I_B + i_B' \quad (4-25)$$

where  $i_E$ ,  $i_B$  and  $i_C$  are the total currents,  $I_E$ ,  $I_B$  and  $I_C$  are the dc components, and  $i_E'$ ,  $i_B'$  and  $i_C'$  are the ac components.

The subscripts E, B and C stand for the emitter, base and collector, respectively. Similar notations apply to the voltages.

For the active mode of operation, we have:

$$i_E = I_{F0} \exp(qV_{EB}/kT)$$

$$i_C = \alpha_F I_{F0} \exp(qV_{EB}/kT)$$

$$i_B = (1 - \alpha_F) I_{F0} \exp(qV_{EB}/kT) \quad (4-26)$$

Thus, for a small change in the emitter-base voltage  $v_{EB}'$ , we have:

$$i_C = I_C + d(i_C)/dV_{EB}|_{V_{EC}} v_{EB}' = I_C + i_C'$$

$$i_B = I_B + d(i_B)/dV_{EB}|_{V_{EC}} v_{EB}' = I_B + i_B' \quad (4-27)$$

This leads to:  $i_C' = g_m v_{EB}'$  and  $i_B' = g_{EB} v_{EB}'$ , where  $g_m$  is the transconductance, and  $g_{EB}$  is the conductance of the emitter-base junction.

In addition,

$$g_m = d(i_C)/dV_{EB}|_{V_{EC}} = (q \alpha_F I_{F0}/kT) \exp(qV_{EB}/kT) = \alpha_F q I_E/kT$$

$$g_{EB} = d(i_B)/dV_{EB}|_{V_{EC}} = q (1 - \alpha_F) I_E/kT = q I_B/kT \quad (4-28)$$

From Eqn.(4-27),  $i_C'$ ,  $i_B'$  and  $v_{EB}'$  form the small-signal ac parameters of the device. Fig.4.15 shows the small-signal equivalent circuit for the bipolar transistor.

The frequency response of the bipolar transistor can be linked to the gain parameters as shown in Fig.4.16.

In general, the frequency-dependent common-base current gain  $\alpha$  is given by:

$$\alpha = \alpha_0 / (1 + jf/f_\alpha) \quad (4-29)$$

where  $\alpha_0$  is the low frequency common-base current gain, and  $f_\alpha$  is the common-base cutoff frequency.

Similarly, the frequency-dependent common-emitter current gain  $\beta$  is given by:

$$\beta = \beta_0 / (1 + jf/f_\beta) \quad (4-30)$$

where  $f_\beta$  is the common-emitter cutoff frequency.

Since  $\beta = \alpha / (1 - \alpha)$ , we have:

$$f_\beta = (1 - \alpha_0)f_\alpha \text{ and } f_\beta < f_\alpha . \quad (4-31)$$

Both  $f_\beta$  and  $f_\alpha$  are called the 3-dB frequencies meaning that at these frequencies,  $\beta$  and  $\alpha$  both reduce to 0.707 of their initial values.

Note that the 3 dB frequency points are the half-power points.

The unity-gain cutoff frequency  $f_T$  is often quoted in the specifications of transistors. It occurs when  $\beta = 1$ .

Thus,

$$\begin{aligned} \beta &= \beta_0 / (1 + jf_T/f_\beta) = 1 \\ f_T &= f_\beta \sqrt{(\beta_0^2 - 1)} \approx \beta_0 f_\beta \approx \alpha_0 f_\alpha \end{aligned} \quad (4-32)$$

This implies:  $f_T \approx f_\alpha$ .

The ultimate frequency limit for the operation of the bipolar transistor is given by the inverse of the time required for an injected carrier to cross the base as illustrated in Fig.4.17. This is called the transit time  $\tau_B$ .

Transit time is given by:

$$\tau_B = \int_0^{W_B} (1/v).dx = \int_0^{W_B} (q p_n A_{cs} / I_p).dx \quad (4-33)$$

By setting  $p_n = p_{n0} (1 - x/W_B)$  and  $I_p = q A_{cs} D_p p_{n0} / W_B$ , we have:

$$\begin{aligned} \tau_B &= \int_0^{W_B} [W_B(1 - x/W_B)/D_p].dx \\ &= W_B^2 / (2D_p) \end{aligned} \quad (4-34)$$

The base width theoretically limits the frequency response of the bipolar transistor and  $f_{\max} \sim 1/\tau_B$ .

## F. Switching transients in the bipolar transistor

In digital circuits, transistors are often required to switch from active to cutoff. The switching response depends on the charge storage in the base region. The simplest way to model the switching bipolar transistor is to assume the existence of an OFF-resistance  $R_{\text{off}}$  and an ON-resistance  $R_{\text{on}}$ .

The circuit schematic is shown in Fig.4-18.

These resistances are given by:

$$R_{\text{off}} = V_C / I_C(\text{off})$$

$$R_{on} = V_{CE(on)}/I_C \quad (4-35)$$

The switching time is given by the time required to switch between these 2 states.

For a common-emitter transistor to switch from off to on, the base charge simply builds up from cutoff to saturation as shown in Fig.4-19.

The base charge  $Q_B$  is given by:

$$Q_B = qA_{cs} \int_0^{W_B} (p_n - p_{n0}) dx \quad (4-36)$$

The time evolution of the base charge can be obtained in the following manner:

$$dp_n/dt = - 1/q dJ_p/dx - (p_n - p_{n0})/\tau_p \quad (4-37)$$

Integrating from  $x = 0$  to  $x = W_B$  gives:

$$I_p(0) - I_p(W_B) = dQ_B/dt + Q_B/\tau_p \quad (4-38)$$

Since  $I_p(0) - I_p(W_B) = i_B$ , we have:

$$dQ_B/dt + Q_B/\tau_p = i_B$$

$$Q_B = i_B \tau_p (1 - \exp(- t/\tau_p)) \quad (4-39)$$

Thus,  $Q_B$  changes from 0 (OFF) to  $i_B \tau_p$  (ON). This is reflected in the lowering of the base voltage as shown in Fig.4.20.

Assuming that the bipolar transistor is saturated (this corresponds to the situation when the base-collector junction also becomes forward bias) at  $t = t_1$  and  $Q_B = Q_{sat}$ , then

$$Q_{sat} = i_B \tau_p (1 - \exp(- t_1 / \tau_p))$$

$$t_1 = \tau_p \ln(1/(1 - Q_{sat}/I_B \tau_p)) \quad (4-40)$$

In general,  $Q_{sat} = V_C \tau_p / R_L$ .

The turn-off transient should be the reverse of the turn-on transient.

If  $t = t_2$  is the initial turn-off time, one gets:

$$Q_B = Q_B(t_2) \exp(- (t - t_2) / \tau_p) \quad (4-41)$$

- Turn-off comes in 2 stages.

The first stage involves charge removal while the device remains in saturation and the second stage involves a progressive decrease in the output current. The time  $t_{sat}$  in the first stage is the storage time.

Since  $Q_{sat} = Q_B(t_2) \exp(- (t_3 - t_2) / \tau_p)$ , where  $Q_B = Q_{sat}$  when  $t = t_3$ , we have:

$$t_{sat} = t_3 - t_2 = \tau_p \ln((Q_B(t_3)/Q_{sat})) \quad (4-42)$$

The second stage involves the reduction in the charge gradient in the base and the collector current will decrease as:

$$i_C = (Q_B(t_3) / \tau_p) \exp(- (t - t_3) / \tau_p) \quad (4-43)$$

**Example 4.7:** A switching transistor has a base width of  $0.5 \mu\text{m}$  and a diffusivity of  $0.001 \text{ m}^2/\text{s}$ . The minority carrier lifetime in the base is  $0.1 \mu\text{s}$  and  $V_{CC} = 5 \text{ V}$  and  $R_L = 10 \text{ k}\Omega$ . If the base current is a  $1 \mu\text{s}$  pulse at  $2 \mu\text{A}$ , find the stored charge in the base and the storage time delay.

Solution:

From Eqn.(4-39), the base charge  $Q_B = I_B \tau_p (1 - \exp(- t_1 / \tau_p)) = 2 \times 10^{-6} \text{ A} \times 10^{-7} \text{ s} \times (1 - \exp(- 10^{-6} \text{ s} / 10^{-7} \text{ s})) = 2 \times 10^{-13} \text{ C}$ . #

At saturation,  $Q_{\text{sat}} = V_{CC} \tau_B / R_L = V_{CC} (W^2 / 2D_B) / R_L = 5 \text{ V} \times (5 \times 10^{-7} \text{ m})^2 / (2 \times 10^{-3} \text{ m}^2/\text{s}) / 10^4 \Omega = 6.25 \times 10^{-14} \text{ C}$ .

From Eqn.(4-42), the storage time  $t_{\text{sat}} = \tau_p \ln(Q_B(t_2) / Q_{\text{sat}}) = 10^{-7} \text{ s} \times \ln(2 \times 10^{-13} \text{ C} / 6.25 \times 10^{-14} \text{ C}) = 0.116 \mu\text{s}$ . #



## **Part 5: MIS diode and MOS transistor**

### **What we need to learn.**

Physical properties of the MIS diode

Capacitance characteristics

Physical structure of the MOS transistor

I-V characteristics

Other effects

urrent components in a bipolar transistor

- a. Gain parameters
- b. Modes of operation
- c. Frequency characteristics
- d. Switching characteristics

## **Part 5: MIS diode and MOS transistor**

The electrical properties of the MOS transistor are primarily determined by the properties of the MIS diode. We first consider the ideal MIS diode and its characteristics:

### A. MIS diode

- The MIS diode is a two terminal device similar to a PN junction although structurally it is no different from a capacitor

Since there is no dc current passing through a capacitor, the parameter of interest is its capacitance, which varies with the applied voltage.

#### a) Potential distribution in the MIS diode

The MIS diode is in a way similar to the reverse bias P-N junction. Structurally, it consists of a metal gate, an oxide layer, and a substrate semiconductor, which we assume to be P-type.

This is illustrated in Fig.5.1.

Because of the different possible combination of MIS structures and the presence of non-idealities, we shall begin with the ideal initial condition that all of the energy bands are flat, i.e.,

$$q\Phi_{ms} = q\Phi_m - [q\chi + E_g/2 + q\psi_B] = 0 \quad (5-1)$$

where  $\Phi_m$  is the metal work function,  $\chi$  is the electron affinity, and  $\psi_B$  is the energy difference between the Fermi-level and the intrinsic Fermi-level.

The above is the flat-band condition and it assumes the absence of space charge in the oxide.

Fig.5.2 shows the energy band diagram for the MIS diode under flat-band condition.

When a voltage is applied across the MIS diode, the following may result:

- When  $V_G$  is small and positive, the substrate will be depleted; i.e., holes will be withdrawn from the interface and the acceptors are exposed
- When  $V_G$  is large and positive, the substrate will be inverted near the interface but further into the substrate, there is a depletion region
- When  $V_G$  is negative, holes will be attracted to the oxide-semiconductor interface and a hole accumulation layer will exist

These are known as depletion, inversion, and accumulation, respectively.

The conditions are shown in Fig.5.3.

Similar to a parallel-plate capacitor, charge conservation is observed in the MIS diode and we have:

$$Q_S = Q_n + Q_{SC} = - Q_m \quad (5-2)$$

where  $Q_S$  is the semiconductor charge per unit area,  $Q_n$  is the inverted charge per unit area,  $Q_{SC}$  is the depletion layer charge per unit area, and  $Q_m$  is the electrode charge per unit area.

These charge layers are shown in Fig.5.4.

In the absence of  $Q_n$ , the depletion charge density  $Q_{SC} = - qN_A W$ , where  $W$  is the depletion layer width.

Inside the MIS diode, the charge states in the semiconductor are best described by their potential denoted by  $\psi$ . Since the energy bands are drawn for electrons, positive potential (and  $\psi$ ) will increase downward.

The value of  $\psi$  right at the oxide-semiconductor interface is the surface potential,  $\psi_s$ .

With this notation, *inversion* occurs when  $\psi_s > \psi_B$ ; *depletion* occurs when  $\psi_B > \psi_s > 0$ ; and *accumulation* occurs when  $\psi_s < 0$ .

$\psi_B$  is the energy difference between the Fermi-level and the intrinsic Fermi-level.

Note that  $\psi$  is zero in the bulk and this is shown in Fig.5.4.

The carrier densities in the P-type semiconductor can be written as:

$$p_p = n_i \exp((E_i - E_F)/kT)$$

$$n_p = n_i \exp(- (E_i - E_F)/kT) \quad (5.3)$$

For the P-type substrate, we have  $E_i = - q\psi$  and  $E_F = - q\psi_B$ . This leads to:

$$p_p = n_i \exp(q(\psi_B - \psi)/kT)$$

$$n_p = n_i \exp(- q(\psi_B - \psi)/kT) \quad (5-4)$$

We can now use the above equations to determine the carrier densities.

At the oxide-semiconductor interface, we have  $\psi = \psi_s$ . If  $\psi_s < \psi_B$ ,  $p_p > n_i$ , we have either depletion or accumulation.

On the other hand, if  $\psi_s > \psi_B$ ,  $p_p < n_i$ , and we have inversion.

The electrostatic potential in the semiconductor can be computed using Poisson equation, i.e.,

$$d^2\psi/dx^2 = -\rho/\epsilon_s = qN_A/\epsilon_s \quad (5-5)$$

Assuming a depletion layer width  $W$  and boundary conditions such that  $\psi = 0$  and  $d\psi/dx = 0$  at  $x = W$ , we have:

$$\psi = \psi_s (1 - x/W)^2 \quad (5-6)$$

where  $\psi_s = qN_A W^2 / (2\epsilon_s)$ .

Thus, the surface potential changes as  $W^2$ .

Normally, surface inversion is fully formed when  $\psi_s = 2\psi_B$ . This corresponds to the situation when the surface electron density  $n|_{x=0} \approx N_A$ .

This is known as strong inversion.

Once the inversion layer is fully formed, the depletion layer width reaches its maximum value and it is given by:

$$W_m = \sqrt{(2\epsilon_s \psi_B / (qN_A))} \quad (5-7)$$

Similarly,

$$Q_{SC} = -qN_A W_m \quad (5-8)$$

Example 5.1: For an ideal Si-SiO<sub>2</sub> MIS diode with  $d = 30 \text{ nm}$  and  $N_A = 5 \times 10^{21} / \text{m}^3$ , find the applied voltage required to make the silicon surface: i) intrinsic; and ii) in strong inversion.

Solution:

The oxide capacitance  $C_0 = 3.9 \times 8.86 \times 10^{-12} \text{ F/m} / 3 \times 10^{-8} \text{ m} = 1.15 \times 10^{-3} \text{ F/m}^2$ .

i) For the Si surface to be intrinsic:  $\psi = \psi_B = kT/q \times \ln(N_A/n_i) = 0.0259 \text{ V} \times \ln(5 \times 10^{21} / \text{m}^3 / 1.56 \times 10^{16} / \text{m}^3) = 0.33 \text{ V}$ .

Ideal diode implies  $V_{FB} = 0$  and  $V_a = \psi_B + (2 \epsilon_s q N_A \psi_B)^{0.5} / C_0 = 0.33 \text{ V} + (2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} \times 1.6 \times 10^{-19} \text{ C} \times 5 \times 10^{21} / \text{m}^3 \times 0.33 \text{ V})^{0.5} / 1.15 \times 10^{-3} \text{ F/m}^2 = 0.53 \text{ V}$ . #

ii) For strong inversion:  $\psi = 2\psi_B$ . This leads to:  $V_a = 0.95 \text{ V}$ . #

## b) C-V characteristics of the MIS diode

In the presence of an applied bias, the voltage drop across the MIS diode can be expressed as:

$$V_a = V_0 + \psi_s \quad (5-9)$$

where  $V_0$  is the voltage drop across the oxide and  $\psi_s$  is the surface potential.

In general, we see that the applied voltage is dropped across two different regions.

One region is the oxide layer and there we have:  $V_0 = Q_s d / \epsilon_{ox}$ , where  $d / \epsilon_{ox} = 1 / C_0$ , and  $C_0$  is the oxide capacitance per unit area.

Another component of  $V_a$  is dropped across the space charge region, and the associated capacitance per unit area  $C_{SC}$  is given by:

$$C_{SC} = \epsilon_s / W = \sqrt{(q \epsilon_s N_A / (2 \psi_s))} = \sqrt{[q \epsilon_s N_A / (2 (V_a - V_0))]} \quad (5-10)$$

At strong inversion,  $Q_s = q N_A W_m$ .

This leads to:

$$V_a = q N_A W_m / C_0 + 2 \psi_B \quad (5-11)$$

This is the threshold voltage  $V_T$  for the MIS diode.

Fig.5.5 shows the C-V characteristics of the MIS diode under different bias conditions and at different frequencies.

c) Flat-band voltage



The flat-band condition mentioned earlier is rarely observed in nature and in general  $q\Phi_{ms}$  is not zero.

Flat-band condition can be achieved by introducing a parameter called flat-band voltage  $V_{FB}$ .

In general, the flat-band voltage is given by:

$$V_{FB} = \Phi_{ms} + (Q_f + Q_m + Q_{ot} + Q_{it})/C_0 \quad (5-12)$$

where  $\Phi_{ms}$  is the metal-semiconductor work function,  $Q_f$  is the oxide fixed charge density,  $Q_m$  is the oxide mobile charge density,  $Q_{ot}$  is the oxide trap charge density, and  $Q_{it}$  is the interface trap charge density.

Example 5.2: For a Si-SiO<sub>2</sub> MIS diode at 300 K with  $d = 30$  nm,  $N_A = 5 \times 10^{21} /m^3$ ; the metal work function is 3 eV,  $q\chi = 4.05$  eV,  $Q_f/q = 10^{15} /m^2$ ,  $Q_m/q = 10^{14} /m^2$ ,  $Q_{ot}/q = 5 \times 10^{14} /m^2$  and  $Q_{it} = 0$ , determine the flat band voltage.

Solution:

As shown in Example 5.1:  $C_0 = 1.15 \times 10^{-3} \text{ F/m}^2$ .

From Eqn.(5-12),  $V_{FB} = \phi_{ms} + (Q_f + Q_m + Q_{ot} + Q_{it})/C_0 = 3 \text{ V} - 4.05 \text{ V} + 1.6 \times 10^{-19} \text{ C}(10^{15} \text{ C} + 10^{14} \text{ C} + 5 \times 10^{14} \text{ C})/(1.15 \times 10^{-3} \text{ F/m}^2) = - 1.27 \text{ V}$ . #

## B. MOS transistor

- MOS transistor is also called MOSFET (**metal-oxide-semiconductor field-effect transistor**)

The device consists of a MIS diode with two PN junctions on either side. A schematic of a MOSFET is shown in Fig.5.6.

The input to the device is the **gate** (the top electrode of the MOS diode) and the output is the current passing (laterally) through the PN junctions.

The output terminals are called the **drain** and the **source** (the source is the source of electrons, i.e., the terminal where current leaves).

In the following, we shall examine the I-V characteristics of an N-channel device, i.e., the substrate semiconductor is P-type. A schematic of such a device is shown in Fig.5.7.

For any appreciable current to pass through the device, an *inversion layer* or *channel* underneath the gate must exist. This normally requires a positive gate voltage exceeding the threshold voltage of the MIS.

In addition, we assume that the source and the substrate are grounded and a positive voltage is applied to the drain.

### a) I-V characteristics of the MOS transistor

The I-V characteristics of the MOS transistor is derived with the following ideal conditions (see Fig.5.7):

- flat-band condition
- only drift currents are important
- carrier mobilities are constant
- doping in the channel is uniform
- reverse leakage currents are small
- transverse electric field in the channel is much smaller than the longitudinal electric field

Semiconductor charge present in the MOS transistor can be expressed as:

$$Q_s = Q_n + Q_{SC} \quad (5-13)$$

where  $Q_s$  is the semiconductor charge density;  $Q_n$  is the electron density in the channel, and  $Q_{SC}$  is the space charge density present in the depletion region.

Since  $Q_s = - (V_G - \psi_s) C_0$  and  $Q_{SC} = - \sqrt{(2\epsilon_s q N_A (V_a + 2\psi_B))}$ , one gets:

$$Q_n = C_0 (V_G - V_a - 2\psi_B) + \sqrt{2\epsilon_s q N_A (V_a + 2\psi_B)} \quad (5-14)$$

This equation gives the electron density in the inversion layer per unit area.

The resistance associated with this inversion layer can be written as:

$$\Delta R = \Delta y / (Z \mu_n Q_n) \quad (5-15)$$

where  $Z$  is the width of the device,  $\mu_n$  is the electron mobility, and  $\Delta y$  is incremental distance in the direction of current flow.

Since  $\Delta V = I_D \Delta R$ , and  $I_D$  is the current in the inversion layer, one gets:

$$\int_0^L I_D \cdot dy = \int_0^{V_D} Z \mu_n Q_n \cdot dV \quad (5-16)$$

where  $L$  is the channel length, and  $V_D$  is the drain voltage.

Combining Eqn.(5-14) and (5-16) leads to:

$$I_D = Z \mu_n (C_0/L) [(V_G - 2\psi_B - V_D/2) V_D - \{\sqrt{(8\epsilon_s q N_A)/3C_0}\} \cdot ((V_D + 2\psi_B)^{3/2} - (2\psi_B)^{3/2})] \quad (5-17)$$

This equation gives the I-V characteristics of the MOS transistor.

A graphical plot is shown in Fig.5.8.

The I-V characteristics of the MOS transistor can be subdivided into the linear region and the saturation region.

In the linear region,  $V_D$  is small and we can write:

$$I_D \approx Z\mu_n (C_0/L) (V_G - V_T) V_D \quad (5-18)$$

where  $V_T = \sqrt{(4\epsilon_s q N_A \psi_B)/C_0} + 2\psi_B$ .

Thus,  $I_D$  is proportional to  $V_D$  and we can define a *channel conductance*  $g_D$  given by:

$$g_D = Z\mu_n (C_0/L) (V_G - V_T) \quad (5-19)$$

Saturation occurs when the channel pinches off and this happens when  $Q_n = 0$  as  $y = L$ .

Eqn.(5-14) now has the form:  $x^2 + \sqrt{(2)K_2} x - V_G = 0$ , where  $x = \sqrt{(V_a + 2\psi_B)}$  and  $K_2 = \sqrt{(\epsilon_s q N_A)/C_0}$ .

Solving this leads to:  $x = [-\sqrt{(2) K_2} \pm \sqrt{(2K_2^2 + 4 V_G)}]/2$ .

By setting  $V_a = V_{Dsat}$ , one gets:

$$V_{D_{\text{sat}}} = V_G - 2\psi_B + K_2 [1 - (1 + 2V_G/K_2)^{1/2}] \quad (5-20)$$

If we substitute  $V_{\text{dsat}}$  into the I-V characteristics of the MOSFET, this leads to:

$$I_{D_{\text{sat}}} \approx Z\mu_n\epsilon_{\text{ox}} (V_G - V_T)^2 / (2dL) \quad (5-21)$$

where  $d$  is the oxide thickness.

In addition, we define transconductance  $g_m$  given by:

$$g_m = Z\mu_n \epsilon_{\text{ox}}(V_G - V_T)/(dL) \quad (5-22)$$

Fig.5.9 shows the equivalent circuit for the MOSFET in saturation.

Example 5.3: Consider a long channel MOSFET with  $L = 3 \mu\text{m}$ ,  $Z = 21 \mu\text{m}$ ,  $N_A = 5 \times 10^{21} /\text{m}^3$ ,  $C_0 = 1.5 \times 10^{-3} \text{ F}/\text{m}^2$  and  $V_T = 1.5 \text{ V}$ . Determine  $V_{D_{\text{sat}}}$  when  $V_G = 4 \text{ V}$ .

Solution:

The substrate potential  $\psi_B = kT/q \times \ln(N_A/n_i) = 0.0259 \text{ V} \times \ln(5 \times 10^{21} /\text{m}^3 / 1.56 \times 10^{16} /\text{m}^3) = 0.33 \text{ V}$ .

$K = \sqrt{(\epsilon_s q N_A) / C_0} = (11.9 \times 8.86 \times 10^{-12} \text{ F}/\text{m} \times 1.6 \times 10^{-19} \text{ C} \times 5 \times 10^{21} /\text{m}^3)^{0.5} / 1.5 \times 10^{-3} \text{ F}/\text{m}^2 = 1.93 \times 10^{-2} \text{ V}^{0.5}$ .

From Eqn.(5-20),  $V_{Dsat} = V_G - 2\psi_B + K^2[1 - (1 + 2V_G/K^2)^{0.5}] = 4 \text{ V} - 2 \times 0.33 \text{ V} + 3.72 \times 10^{-4} \text{ V} \times (1 - (1 + 2 \times 4 \text{ V}/3.72 \times 10^{-4} \text{ V})^{0.5}) = 3.40 \text{ V}.$  #

## b) Frequency response

When the output of the MOS transistor is short-circuited, the input current is given by:

$$i_{in} = j\omega(C_{GS} + C_{GD}) v_G \approx j\omega C_0 ZL v_G. \quad (5-23)$$

where  $v_G$  is the ac gate voltage.

Since the output current is  $i_{out} = g_m v_G$ , unity gain occurs when  $\omega C_0 ZL = g_m$ .

The cutoff frequency is given by:

$$f_T = \omega/2\pi = g_m/(2\pi C_0 ZL) = \mu_n V_D/(2\pi L^2) \quad (5-24)$$

Note that  $f_T$  is inversely proportional to  $L^2$  and we have used the relationship  $V_D = V_G - V_T$ .

## c) Subthreshold conduction

Below threshold, the MOS channel is not fully formed and a horizontal PNP transistor is present as shown in Fig.5.10.

The lateral current will be primarily due to diffusion and it is given by:

$$I_D = -qA_{cs}D_n \frac{dn}{dy} = qA_{cs}D_n (n(0) - n(L))/L \quad (5-25)$$

Since  $n(0) = n_i \exp(q(\psi_s - \psi_B)/kT)$ , and  $n(L) = n_i \exp(q(\psi_s - \psi_B - V_D)/kT)$ , it can be shown that:

$$I_D = qA_{cs}D_n n_i \exp(-q\psi_B/kT) [1 - \exp(-qV_D/kT)] \exp(q\psi_s/kT)/L \quad (5-26)$$

Since  $\psi_s = V_G - V_T'$ , we have for a small  $V_D$ :

$$I_D \propto \exp(-q(V_G - V_T')/kT) \quad (5-27)$$

where  $V_T'$  is the threshold voltage of the MOS transistor.

The threshold current is exponentially proportional to the gate voltage. This is similar to the case of the bipolar junction transistor where the output collector current is proportional to the emitter-base voltage.



## d) Enhancement mode and depletion mode devices

In addition to N-channel and P-channel MOS transistors, these devices can operate in the enhancement mode or the depletion mode.

Enhancement mode transistors are normally OFF and a gate bias is required to form the channels (ON). The channel of a depletion mode device is formed in the absence of any bias.

## e) Threshold voltage under non-ideal conditions

Most MOS transistors do not exhibit flat-band condition in the absence of bias and a flat-band voltage is required to achieve flat-band condition. In general,

$$V_T' = V_{FB} + 2\psi_B + \sqrt{(4\epsilon_s q N_A \psi_B)/C_0} \quad (5-28)$$

Normally, in a P-channel device, a negative  $V_T'$  implies an enhancement mode device and a gate voltage more negative than  $V_T'$  is required to turn the device ON.

For an N-channel device, a negative  $V_T'$  implies a depletion mode device and a gate voltage more positive than  $-V_T'$  is required to keep the device on.

Example 5.4: If the substrate dopant density of a MOS transistor is  $10^{20}/\text{m}^3$ . Compute the threshold voltage if the substrate is: i) P-type; and ii) N-type. Assume an oxide thickness of 65 nm.

Solution:

i) For a P-type substrate:  $q\Phi_{ms} = -0.96$  eV (see Fig.E5.1).

$$\psi_B = kT/q \ln(N_A/n_i) = 0.0259 \times \ln(10^{20}/(1.45 \times 10^{16})) \text{ V} = 0.229 \text{ V}. \quad C_{ox} = \epsilon_{ox}/d \\ = 3.9 \times 8.86 \times 10^{-12}/(650 \times 10^{-10}) \text{ F/m}^2 = 5.3 \times 10^{-4} \text{ F/m}^2. \quad \#$$

$$V_{TN}' = V_{FB} + 2\psi_B + \sqrt{(4\epsilon_s q N_A \psi_B)/C_0} = -0.96 + 2 \times 0.229 + \sqrt{(2 \times 11.9 \times 8.86 \times 10^{-12} \times 10^{20} \times 0.458)/5.3 \times 10^{-4}} = -0.42 \text{ V}. \quad \#$$

ii) For an N-type substrate,  $q\Phi_{ms} = -0.51$  eV.

$$V_{TP}' = V_{FB} + 2\psi_B + \sqrt{(4\epsilon_s q N_D \psi_B)/C_0} = -0.52 - 2 \times 0.229 - \sqrt{(2 \times 11.9 \times 8.86 \times 10^{-12} \times 10^{20} \times 0.458)/5.3 \times 10^{-4}} = -1.04 \text{ V}. \quad \#$$

Note that as  $\psi_B$  increases,  $V_{TN}'$  increases while  $V_{TP}'$  decreases.

f) Substrate bias

A reverse bias in the substrate affects the space charge in the depletion layer. Substrate bias is normally applied at the source and this increases the depletion layer width such that:

$$W_m' = \sqrt{(2\epsilon_s (2\psi_B + V_{BS})/qN_A)} \quad (5-29)$$

Similarly,

$$V_{TN}' = V_{FB} + 2\psi_B + \sqrt{[2\epsilon_s q N_A (2\psi_B + V_{BS})] / C_0}$$

$$\Delta V_{TN}' = \sqrt{(2\epsilon_s q N_A / C_0) [(2\psi_B + V_{BS})^{1/2} - (2\psi_B)^{1/2}]} \quad (5-30)$$

Substrate bias therefore increases the threshold voltage in a N-channel device.

### g) Device scaling

Device scaling is intended to increase the device density per unit area while not affecting substantially the transistor characteristics

Disadvantages observed in scale-down device are:

- 2-dimensional and high-field effects often limit the device performance.
- High field effects tend to cause breakdown between the drain and the source. Mobility saturation can degrade the frequency performance.

At very high electric field, velocity saturation occurs when  $v_{sat} \approx 1 \times 10^5$  m/s.

When  $v_{drift} = v_{sat}$ ,  $I_{Dsat} = Zq v_{sat} \int_0^{x_i} n \cdot dx$ , where  $x_i$  is the thickness of the depletion layer.

Since  $Q_n = \int_0^{x_i} n \, dx \approx C_0 (V_G - V_T)$ , one gets:

$$I_{Dsat} = ZC_0 v_{sat} (V_G - V_T)$$

$$g_m' = ZC_0 v_{sat} \quad (5-31)$$

Compare this with normal operation (when  $g_m = Z C_0 \mu_n (V_G - V_T)/L$ ),  $g_m'$  is reduced considerably.

For a scaling factor of  $k$  ( $k > 1$ ), we have:

$$L' = L/k,$$

$$d' = d/k,$$

$$Z' = Z/k,$$

$$V_a' = V_a/k \quad (5-31)$$

This leads to:

$$I_{Dsat}' = (Z/k) kC_0 v_{sat} (V_G - V_T)/k = I_{Dsat} /k$$

$$J_{Dsat}' = J_{Dsat} k$$

$$P_{ac}' = C_o' A_{cs}' V_a'^2 / (2\pi R C_o' A_{cs}') = P_{ac} / k^2$$

$$P_{dc}' = I'V_a' = P_{dc}/k^2 \quad (5-32)$$

Note that in this case, only  $J_{Dsat}$  is increased.

To avoid short-channel effect (such as velocity saturation and 2 dimensional effect), there is an empirical relationship for the minimum channel length.

It is given by:

$$L_{min} = 0.4 [r_j d(W_s + W_D)^2]^{1/2} \quad (5-33)$$

where  $r_j$  is the junction curvature, and  $W_s$  and  $W_D$  are the depletion layer widths in the source and the drain, respectively.

Example 5.6: Design a sub-micron MOSFET with a gate length of 0.75  $\mu\text{m}$ . (The gate length is the channel length plus twice the junction depth.) If the junction depth is 0.2  $\mu\text{m}$ , the gate oxide thickness is 20 nm, and the maximum drain voltage is limited to 2.5 V, find the required channel doping so that the MOSFET can maintain its long channel characteristics.

Solution:

The channel length  $L = 0.75 \mu\text{m} - 2 \times 0.2 \mu\text{m} = 0.35 \mu\text{m}$ . From Eqn.(5-33),  
 $L_{min} = 0.4 [r_j d(W_s + W_D)^2]^{1/3}$ .

If  $L_{min} = L$ ,  $W_s + W_D = (0.35 \mu\text{m}/0.4)^3/(0.2 \times 200) = 0.129 \mu\text{m}$ .

Since  $W_D \approx W_s + W_D$  and  $V_R \approx V_R + V_{bi}$ ,  $N_B (\text{min}) = (2 \epsilon_s V_R / (q W_D^2)) = 2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} \times 2.5 \text{ V} / (1.6 \times 10^{-19} \text{ C} \times 11.8 \times (0.129 \times 10^{-6} \text{ m})^2) = 1.97 \times 10^{23} / \text{m}^3$ . #

**END**