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 in this chapter?

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^{-3}.

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 $V = (5.43 \times 10^{-10})^3\ m^3 = 1.6 \times 10^{-28}\ m^3$.

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

- *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 (~ $10^{12\text{-}19}$ 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 x 10^{16} m^{-3} of electrons and holes, respectively. If μ_n = 0.15 m^2/V.s and μ_p = 0.045 m^2/V.s, determine the conductivity.

Solution:

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

(Remember we mentioned earlier on that σ 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)

increases “upward”, while hole energy (positive potential energy) increases “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 energy bands 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:

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

where T is the absolute temperature.

This dependence is shown in Fig.1-10.

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

Solution:

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

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 assist in the motion of the electrons and the holes.

For Si at 300K: 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:

$$S(E) = 4\pi [2m_{ds}^*/h^2]^{3/2}E^{1/2} \qquad (1\text{-}2)$$

where m_{ds}^* is a "density of states mass" parameter and h (= 0.662 x 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).

$$f(E) = 1/[1 + \exp((E - E_F)/kT)] \qquad (1\text{-}3)$$

where E_F is the Fermi-level, E is the energy of the electron, k (= 1.38×10^{-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:

$$n = \int_{Ec}^{\infty} S_e(E)/[1 + \exp((E - E_F)/kT)]\, dE \qquad (1\text{-}4)$$

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 $S_e(E)$ and f(E) is shown in Fig.1-11.

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

$$p = \int_{Ev}^{-\infty} S_h(E)/[1 - \exp((E - E_F)/kT)]\, dE \qquad (1\text{-}5)$$

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 $|E_v - E_F| >> kT$,

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

$$p = N_V \exp[-(E_F - E_v)/kT] \qquad (1\text{-}6)$$

where N_C (= 2 x $(2\pi\, m_{CB}^* kT/h^2)^{3/2}$) is the effective density of states in the conduction band, and N_V (= 2 x $(2\pi\, m_{VB}^* kT/h^2)^{3/2}$) is the effective density of states in the valence band. m_{CB}^* is the density of states mass parameter for the conduction band, and m_{VB}^* is the density of states mass parameter for the valence band.

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

According to Eqn.(1-6), electron and hole densities are functions of 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 = 300K (kT at 300K = 0.0259 eV).

Solution:

From Eqn.(1-6): The electron density $n = N_C \exp(-(E_C - E_F)/kT) = 2.8 \times 10^{25}$ m^{-3} x exp(- 0.55 eV /0.259 eV) = 1.67×10^{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:

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

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

At 300K, $n_i = 1.45 \times 10^{16}$ 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:

$$E_i = E_g/2 + (kT/2) \ln(N_V/N_C) \qquad (1\text{-}8)$$

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

e) Extrinsic semiconductor

- Semiconductor becomes **extrinsic** when impurities are added

12

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 the 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 semiconductors is **N-type** when $n >> p$, and **P-type** when $p >> n$.

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

Since the law of mass action dictates that: $pn = n_i^2$, we have $p = n_i^2/N_D$ for an 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 ($n = N_D$),

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

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

$$E_V - E_F = - kT \ln(N_V/N_A) \qquad (1\text{-}9)$$

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.

14

$$n = N_C \exp(-(E_C - E_F)/kT)$$

$$= N_C \exp(-(E_C - E_i)/kT) \exp((E_F - E_i)/kT)$$

$$= n_i \exp((E_F - E_i)/kT) \qquad (1\text{-}10)$$

Similarly,

$$p = n_i \exp((E_i - E_F)/kT) \qquad (1\text{-}11)$$

Using this representation, one uses the intrinsic Fermi-level as the energy reference. This also reduces the number of unknown materials parameters for the equations of n and p from three (N_C, N_V, E_C/E_V) to two (E_i and n_i).

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

Solution:

From Eqn.(1-9), $E_C - E_F = kT \ln(N_C/N_D) = 0.0259$ eV x $\ln(2.8 \times 10^{25}\ m^{-3}/(1 \times 10^{22}\ m^{-3})) = 0.2$ eV. #

Example 1.7: If Si is doped with 1×10^{20} donor m^{-3}, compute the carrier densities at 300K.

Solution:

$n = N_D = 1 \times 10^{20}\ m^{-3}$ and $p = n_i^2/N_D = (1.45 \times 10^{16}\ m^{-6})^2/(1 \times 10^{20}\ m^{-3}) = 2.1 \times 10^{12}\ m^{-3}$. #

f) Dopant energy levels

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

16

	Phosphorus	Arsenic	Boron
Si	0.045	0.049	0.045
Ge	0.012	0.0127	0.0104
GaAs	0.026	-	-

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

The dopant energies may be compared to the thermal energy kT (≈ 0.0259 eV) at 300K.

Fig.1-15 shows the position of dopant energy levels in a typical energy band diagram.

g) Temperature dependence of the carrier densities

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

At low temperature, not all of the dopant is ionized and the majority carrier density is small. Nevertheless, its value is still larger than the value of n_i.

At high temperature, the intrinsic carrier density increases rapidly, while the dopant density (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 in this chapter?

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 or a concentration gradient

The simplest types of carrier transport are **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 3kT/2 (3-dimensions)

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

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

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

It can be shown that at T = 300K, 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 inside a semiconductor. This is superimposed on top of the random motion (v_{th}).

Because of the charge, electrons move in opposite direction to the applied electric field and holes move in the same direction.

- The motion of the electrons and the holes will stop when they encounter a ***collision***

This is illustrated in Fig.2.1.

Assuming a ***mean free time*** τ_c between collisions for the electrons, Newton's law gives an expression of the form:

$$- qE' = m_e^* v_d/\tau_c \qquad (2\text{-}2)$$

where v_d is the drift velocity of the electrons.

This leads to:

$$v_d = - q\tau_c E'/m_e^* \qquad (2\text{-}3)$$

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

We have:

$$\mu_n = q\tau_c/m_e^* \qquad (2\text{-}4)$$

Note that: μ_n is positive.

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

$$v_d' = q\tau_c^* E'/m_h^* \qquad (2\text{-}5)$$

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

This leads to:

$$\mu_p = q\tau_c^*/m_h^* \qquad (2\text{-}6)$$

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 smallest 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* μ_T in the presence of more than one type of scattering is given by:

$$\mu_T^{-1} = \sum_i \mu_I^{-1} \qquad (2\text{-}7)$$

where μ_i is the mobility of the ith scattering.

The summation process is illustrated in Fig.2.2. From the carrier mobility, one can determine the drift velocity.

Example 2.1: At 300K, μ_n = 0.15 m^2/V.s. Determine its value at 200K if lattice scattering dominates.

Solution:

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

- *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 Δx in time Δt, the electron current density J_n (= Q $\Delta x/\Delta t$ = Q v_d) is given by:

$$J_n = - qnv_d \qquad (2\text{-}8)$$

Similarly, hole current density $J_p = qpv_d$'.

The *total current density* J_T is given by:

$$J_T = J_n + J_p = - qnv_d + pqv_d' \qquad (2\text{-}9)$$

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

$$J_T = (nq\mu_n + pq\mu_p) E' = \sigma E' \qquad (2\text{-}10)$$

where σ is the ***conductivity*** of the semiconductor.

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:

$$\rho = V/I \times 4.54\ t'\ \Omega.m \qquad (2\text{-}11)$$

where V is the measured voltage, I is the current through the outer probes, and t' is the thickness of the wafer (see Fig.2.3).

Note that in obtaining Eqn.(2-11), the probe separation has been assumed to be small compared to the other dimensions.

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

24

Solution:

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

Example 2.3: In a 4-point probe measurement, if V = 1 V and I = 1 mA, determine ρ if t' = 200 μm.

Solution:

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

C. Diffusion current

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

For electrons, the flux (rate of electron flow per unit area) φ_n in 1-dimension is given by:

$$\varphi_n = - D_n \frac{dn}{dx} \qquad (2\text{-}12)$$

where D_n is the <u>electron diffusivity</u>.

Electron diffusion current density is then given by:

$$J_n = - q\varphi_n = qD_n \frac{dn}{dx} \qquad (2\text{-}13)$$

Similarly, hole diffusion current density is given by:

$$J_p = - qD_p \frac{dp}{dx} \qquad (2\text{-}14)$$

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*** λ 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 $\mu_n = q\tau_c/m_e^*$,

$$D_n = \mu_n kT/q \qquad (2\text{-}15)$$

This equation is known as **Einstein relation**. Einstein relation allows one to relate D_n to μ_n. A similar expression exists for the holes.

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

Solution:

26

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

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:

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

$$J_p = qp\mu_p E' - qD_p\, dp/dx \qquad (2\text{-}16)$$

The ***total current density*** J_T is given by:

$$J_T = J_n + J_p \qquad (2\text{-}17)$$

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: $\Delta n >> n_0$ and $\Delta p >> p_0$, while *low-injection* implies either $\Delta n >> n_0$ and $\Delta p << p_0$ (in a P-type semiconductor), or $\Delta n << n_0$ and $\Delta p >> p_0$ (in an N-type semiconductor).

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

Solution:

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

E. Generation and recombination in semiconductors

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

These are: **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; or ii) indirect recombination via impurities

In *steady state*, generation rate G equals to recombination rate R. This is illustrated in Fig.2.5.

28

a) Band-to-band recombination

- Recombination depends on the densities of the recombinants

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

$$R = \beta^* np \qquad (2\text{-}18)$$

where β^* 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:

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

If one defines U as $\Delta p/\tau_p$, where τ_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 τ_n is the ***electron lifetime***.

Example 2.6: In a semiconductor if the excess hole density is 1×10^{15} m^{-3} and its lifetime is 100 μs, 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 $= 1 \times 10^{19}$ $m^{-3}.s^{-1}$. #

- In the steady state, the net recombination rate U is balanced by the generation rate G (assume $G >> G_{th}$)

Since $G = \Delta p/\tau_p$, this leads to (for the case of low-injection):

$$p = p_0 + G\tau_p \qquad (2\text{-}20)$$

b) Dynamic response (recombination)

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

$$dp/dt = G - R = - (R - G) \qquad (2\text{-}21)$$

In the absence of ext

where we have 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 with time towards it equilibrium value p_0. This is illustrated in Fig.2.6.

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

Solution:

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

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 as 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:

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

$$dp/dt = -(1/q)\, dJ_p/dx + (G_p - R_p) \qquad (2\text{-}23)$$

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

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

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

$$dp_n/dt = -p_n\mu_p dE'/dx - \mu_p E'\, dp_n/dx + D_p d^2p_n/dx^2 + G_p - (p_n - p_{n0})/\tau_p \qquad (2\text{-}24)$$

The product $(D_p\tau_p)^{1/2} = L_p$ is called the hole diffusion length.

Similarly, the electron diffusion length is given by: $L_n = (D_n\tau_n)^{1/2}$.

Example 2.8: Write down the steady-state continuity equation for hole minority carriers when $E' = 0$ and $G_p = 0$. Assume $p_n = p_{no}$ when $x = \infty$.

Solution:

32

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 is non-zero, i.e., $D_p\, d^2p_n/dx^2 = (p_n - p_{n0})/\tau_p$. The solution to this equation is: $p_n = p_{n0} + (p_n(0) - p_{n0}) \exp(-x/(D_p\tau_p)^{1/2})$, 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: $p_n = p_{n0} + (p_n\,(0) - p_{n0}) \exp(- x/(D_p\tau_p)^{1/2}) = 1 \times 10^{11}\ /m^3 + (1 \times 10^{22}\ /m^3 - 1 \times 10^{11}\ /m^3) \times \exp(- 1 \times 10^{-3}\ /m/(1 \times 10^{-4}\ m^2/s \times 1 \times 10^{-3}\ s)^{1/2}) = 4.2 \times 10^{20}\ /m^3$. #

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 $w << L_p$. 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 $w >> L_p$, it can be shown that: $J_p = qD_p\,(p_n\,(0) - p_{n0})/w$.

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

Part 3: P-N junctions

What we need to learn in this chapter?

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

34

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

$$J_p = qp\mu_p E' - qD_p dp/dx \qquad (3\text{-}1)$$

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

$$J_p = qp\mu_p (1/q)dE_i/dx - \mu_p kT dp/dx \qquad (3\text{-}2)$$

Furthermore, since $p = n_i \exp((E_i - E_F)/kT)$, $dp/dx = (p/kT).(dE_i/dx - dE_F/dx)$, Eqn.(3-2) becomes:

$$J_p = p\mu_p dE_F/dx \qquad (3\text{-}3)$$

There is no current flowing at equilibrium and 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 physically there exists a "barrier" across the P-N junction.

This potential barrier is given by:

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

where V_{bi} the built-in voltage.

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

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

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

- 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 300K.

Solution:

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

$$d^2\psi/dx^2 = qN_A/\varepsilon_s \qquad \text{for the P-side}$$

$$d^2\psi/dx^2 = -qN_D/\varepsilon_s \qquad \text{for the N-side} \qquad (3\text{-}6)$$

where ψ is the electrostatic potential (in units of volts), and ε_s is the semiconductor permittivity.

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

In the depletion region, charge conservation requires that:

$$N_A x_p = N_D x_n \qquad (3\text{-}7)$$

where x_p and x_n are the widths of the depletion region on the 2 sides of the P-N junction.

The depletion layer width W is given by:

38

$$W = x_p + x_n. \qquad (3\text{-}8)$$

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

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

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

where E' is the electric field.

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

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

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

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

$$\psi_n = - qN_D/\varepsilon_s(x^2/2 + xx_n) + K_2 \qquad (3\text{-}10)$$

where ψ_p and ψ_n are the electrostatic potentials on the P-side and the N-side, respectively.

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

This gives: $K_1 = qN_Ax_p^2/2\varepsilon_s$ and $K_2 = V_{bi} - qN_Dx_n^2/2\varepsilon_s$.

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:

$$V_{bi} = qN_Ax_p^2/(2\varepsilon_s) + qN_Dx_n^2/(2\varepsilon_s) = E_m' W/2 \qquad (3\text{-}11)$$

Rearranging this equation gives:

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

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} (Eqn.(3-12)) 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 $N_A >> N_D$, we have:

$$W \approx x_n = [2\varepsilon_s V_{bi}/(qN_D)]^{1/2}$$

$$E' = - qN_D W(1 + x/W)/\varepsilon_s$$

$$\psi = V_{bi}(- 2x – x^2/W)/W \qquad (3\text{-}13)$$

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 = 300K.

Solution:

From Eqn.(3-5), $V_{bi} = (kT/q) \ln(N_A N_D/n_i^2) = 0.0259$ V x $\ln(10^{25}$ m^{-3} x 10^{21} $m^{-3}/(2.1 \times 10^{32}$ $m^{-6}) = 0.874$ V.

From Eqn.(3-13), $W \approx x_n = (2\varepsilon_s V_{bi}/(qN_D))^{1/2} = 3.37 \times 10^{-7}$ m. #

$E_m' = - qN_D W/\varepsilon_s = - 5.4$ 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:

$$C_j = dQ/dV_a = \varepsilon_s/W \qquad (3\text{-}14)$$

where Q is the charge density per unit area.

For a 1-sided step junction:

$$C_j = [q\varepsilon_s N_B/(2(V_{bi} - V_a))]^{1/2} \qquad (3\text{-}15)$$

where N_B is the substrate dopant density.

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^{-3} on the lightly doped N-side and of 10^{25} m^{-3} 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^{-3}. The governing equation (Eqn.(3-15)) is $1/C^2 = 2(V_{bi} - V_a)/(q\varepsilon_s N_B)$. 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

42

- Low-injection
- No generation and recombination current in the depletion region

Without bias, the electron density in the N-side at equilibrium is given by (see Eqn.(1-10)):

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

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

On the P-side, one has:

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

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:

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

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

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

Under bias, we can replace V_{bi} by $V_{bi} - V_a$; n_{n0} by n_n; and n_{p0} by n_p to Eqn.(3-18).

This leads to:

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

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

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

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:

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

To determine the diffusion currents, we combine Eqns.(3-20), (3-21) and the continuity equation (Eqn.(2-24)). This gives (after simplification):

$$D_n \, d^2n_p/dx^2 - (n_p - n_{p0})/\tau_n = 0$$

$$D_p \, d^2p_n/dy^2 - (p_n - p_{n0})/\tau_p = 0 \qquad (3.22)$$

Note that we have used the coordinate transformation: $y = – x$.

44

The boundary conditions are:

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

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

The solutions are:

$$n_p - n_{p0} = n_{p0} [\exp(qV_a/kT) - 1] \exp(- x/L_n)$$

$$p_n - p_{n0} = p_{n0} [\exp(qV_a/kT) - 1] \exp(- y/L_p) \qquad (3\text{-}23)$$

Note that $L_n = (D_n\tau_n)^{1/2}$ and $L_p = (D_p\tau_p)^{1/2}$. The excess carrier densities are plotted in Fig.3.14.

The above equations also lead to:

$$J_n = qD_n \, dn_p/dx = - (qD_n n_{po}/L_n) [\exp(qV_a/kT) - 1]$$

$$J_p = - qD_p \, dp_n/dy = - (qD_p p_{no}/L_p) [\exp(qV_a/kT) - 1]$$

(3-24)

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

The total current density J_T is:

$$J_T = |J_n + J_p| = J_s [\exp(qV_a/kT) - 1] \qquad (3\text{-}25)$$

where $J_s = qD_p p_{no}/L_p + qD_n n_{po}/L_n$.

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}, $\tau_p = \tau_n = 1$ μs and a device area of 1.2×10^{-9} m^2, calculate the ideal saturation current at 300K. (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.(3-25), we can write: $I_s = J_s \times A_{cs} = (qD_p n_i^2/(L_p N_D) + qD_n n_i^2/(L_n N_A)) \times A_{cs} = (q \times n_i^2 \times A_{cs}) \times ((D_p/\tau_p)^{1/2} N_D + (D_n/\tau_n)^{1/2} N_A) = (1.6 \times 10^{-19}$ C $\times 2.1 \times 10^{32}$ $m^{-6} \times 1.2 \times 10^{-9}$ $m^2) \times ((0.00116$ $m^2/s/10^{-6}$ $s)^{1/2}/10^{24}$ $m^{-3}) + ((0.00388$ $m^2/s/10^{-6}$ $s)^{0.5}/10^{22}$ $m^{-3}) = 2.49 \times 10^{-16}$ A. #

Example 3.6: For the device in Example 3.5, compute the forward current at $V_a = + 0.7$ V.

Solution:

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

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

46

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 τ_g, one can express $G = n_i/\tau_g$.

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

$$J_{gen} = \int^{W}_{0} qG\, dW = qn_iW/\tau_g \qquad (3\text{-}26)$$

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

$$J_R = J_s + J_{gen} = qD_pp_{n0}/L_p + qD_nn_{p0}/L_n + qn_iW/\tau_g \qquad (3\text{-}27)$$

- 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 $(\Delta(pn) - n_i^2)$ is the excess of the product of p and n, one can write:

$$U = n_i [\exp(qV_a/kT) - 1]/\tau_r \qquad (3\text{-}28)$$

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 σ_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 σ_o.

- 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:

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

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

The recombination current density J_r is:

$$J_r = q \int^W_0 U_{max}\, dW = q\, U_{max}\, W \qquad (3\text{-}30)$$

The total forward current density now becomes:

$$J_F = J_s \exp(qV_a/kT) + qW\sigma_0 v_{th} N_t n_i [\exp(qV_a/2kT)]/2$$

48

(3-31)

In general, J_F is proportional to $\exp(qV_a/\eta kT)$, where η 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:

$$J_T = J_0 [\exp(qV_a/2kT) - 1] \qquad (3\text{-}32)$$

where J_0 is a constant.

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

Solution:

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

From Eqn.(3-13), $W = (2\varepsilon_s(V_{bi} + V_R)/(qN_D))^{1/2} = (2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} \times (0.695 \text{ V} + 0.5 \text{ V})/(1.6 \times 10^{-19} \text{ C} \times 10^{21} \text{ m}^{-3}))^{1/2} = 1.25 \times 10^{-6}$ m.

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

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

Solution:

From Eqn.(3-13), $W = (2\varepsilon_s(V_{bi} - V_a)/(qN_D))^{1/2}$ = (2 x 11.9 x 8.86 x 10^{-12} F/m x (0.695 V - 0.5 V)/(1.6 x 10^{-19} C x 10^{21} m^{-3}))$^{0.5}$ = 0.5 x 10^{-6} m.

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

b) Charge storage

- At 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:

$$Q_p = q \int_{xn}^{\infty} (p_n - p_{n0})\, dx$$

$$= q \int_{xn}^{\infty} p_{n0} (\exp(qV_a/kT) - 1) \exp(-(x - x_n)/L_p)\, dx$$

$$= qL_p\, p_{n0} (\exp(qV_a/kT) - 1) \qquad (3\text{-}33)$$

50

Since $J_p = (qD_p\, p_{no}/L_p)\,[\exp(qV_a/kT) - 1]$, one can write:

$$Q_p = L_p^2 J_p/D_p = \tau_p J_p \qquad (3\text{-}34)$$

This is illustrated in Fig.3.20.

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

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

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

The diffusion capacitance per unit area C_d is given by:

$$C_d = \Delta Q_p/\Delta V_a$$

$$= (q^2 L_p\, p_{n0}/kT) \exp(qV_a/kT) \qquad (3\text{-}36)$$

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 excessive 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 a Si 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.

52

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

The *breakdown voltage* V_B is:

$$V_B = E_C'W/2 \qquad (3\text{-}37)$$

Assuming $E_C' = E_m' = qN_BW/\varepsilon_s$ as in the case of a 1-sided step junction (Eqn.(3-13)), Eqn.(3.37) becomes:

$$V_B = E_C^2\, \varepsilon_s/(2qN_B) \qquad (3\text{-}38)$$

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^{-3}. This leads to (Eqn.(3.28)): $V_B = = E_C'^2\, \varepsilon_s/(2qN_B) = (8 \times 10^7$ V/m$)^2 \times 11.9 \times 8.86 \times 10^{-12}$ F/m /$(2 \times 1.6 \times 10^{-19}$ C $\times 5 \times 10^{23}$ m^{-3}) V = 4.21 V. #

Part 4: Bipolar transistor

What we need to learn in this chapter?

Current components in a bipolar transistor

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

54

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

The 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}, respectively and they are shown in Fig.4.3.

The total emitter current I_E is given by:

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

The total collector current I_C is given by:

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

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

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

These are illustrated in Fig.4.4.

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

56

This is expressed in terms of the ***common-base current gain*** α_0 which is given by:

$$\alpha_0 = I_{Cp}/I_E = I_{Cp}/(I_{Ep} + I_{En})$$

$$= I_{Ep}/(I_{Ep} + I_{En}) \times I_{Cp}/I_{Ep} \qquad (4\text{-}4)$$

The first term appearing in Eqn.(4-4) is the ***emitter efficiency*** γ and the second term is the ***transport factor*** α_T.

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

Furthermore, we have:

$$I_C \approx \alpha_0 I_E + I_{Cn} \qquad (4\text{-}5)$$

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 α_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-dimension hole (minority carrier) density in the base is given by:

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

This equation is subject to the following boundary conditions:

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

$$p_n(W_B) = 0 \qquad (4\text{-}7)$$

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:

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

$$- \sinh(x/L_p)/\sinh(W_B/L_p)) \qquad (4\text{-}8)$$

where we have used the relationship: $L_p^2 = D_p\tau_p$.

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

58

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

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

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

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

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

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.

The values are:

$$n_E = n_{E0} + n_{E0} (\exp(qV_{EB}/kT) - 1) \exp((x + x_E)/L_E) \quad x > - x_E$$

$$n_C = n_{C0} - n_{C0} \exp(- (x - x_C)/L_C) \quad x' > x_C \qquad (4\text{-}11)$$

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:

$$I_{Ep} = A_{cs} [- qD_p \, dp_n/dx]_{x=0}$$

$$= (qA_{cs}D_p p_{n0}/L_p) \coth(W_B/L_p) [(\exp(qV_{EB}/kT) - 1) +$$

$$1/\cosh(W_B/L_p)] \qquad (4\text{-}12)$$

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

For $W_B < L_p$,

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

Similarly, at $x = W_B$,

$$I_{Cp} = qA_{cs} D_p p_{n0}/\{L_p \sinh(W_B/L_p)\} [(\exp(qV_{EB}/kT) - 1) +$$

$$\cosh(W_B/L_p)]$$

$$\approx (qA_{cs} D_p p_{n0}/W_B) [\exp(qV_{EB}/kT) – 1] \qquad (4\text{-}14)$$

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

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

$$I_{En} = A_{cs} [- qD_E \, dn_E/dx]_{x = - xE}$$

$$= qA_{cs}\, D_E n_{E0}\, (\exp(qV_{EB}/kT) - 1)/L_E$$

$$I_{Cn} = A_{cs}\, [-\, qD_C\, dn_C/dx]_{x\,=\,-\,xc}$$

$$= qA_{cs} D_C n_{C0}/L_C \qquad (4\text{-}15)$$

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

We can write:

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

When $W_B < L_p$, we have:

$$a_{11} = qA_{cs}\, [D_p n_i^2/(N_B W_B) + D_E n_{E0}/L_E]$$

$$a_{21} = qA_{cs}\, D_p n_i^2/(N_B W_B) \qquad (4\text{-}17)$$

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 $w_B > L_p$,

$$a_{21} = qA_{cs}\, D_p n_i^2/(N_B W_B)$$

$$a_{22} = qA_{cs}\, [D_p n_i^2/(N_B W_B) + D_C n_{C0}/L_C] \qquad (4\text{-}18)$$

In addition,

$$I_B = I_E - I_C \qquad (4\text{-}19)$$

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 emphasized out that the base current actually arises from carrier recombination as the carrier transit through the base. Based on Eqn.(4-6): $D_p\, d^2p_n/dx^2 - (p_n - p_{n0})/\tau_p = 0$, $J_p = -qD_p dp_n/dx$ = constant if recombination is absent. The base current is therefore suppressed when τ_p is large.

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

$$\gamma = I_{Ep}/(I_{Ep} + I_{En})$$

$$= (1 + I_{En}/I_{Ep})^{-1}$$

$$\tilde{} \; [1 + (qA_{cs}D_E n_{E0}/L_E)\, (\exp(qV_{EB}/kT) - 1)/\{(qA_{cs}\, D_p p_{n0}/W_B)$$

$$(\exp(qV_{EB}/kT) - 1)\}]^{-1}$$

$$\approx (1 + D_E n_{E0} W_B/(D_p p_{n0} L_E))^{-1} \qquad (4\text{-}20)$$

62

Thus,

$$\alpha_T = I_{Cp}/I_{Ep}$$

$$\approx \{qA_{cs}D_p p_{n0}(\exp(qV_{EB}/kT) - 1)/(L_p \sinh(W_B/L_p)\}/[(qA_{cs} D_p p_{n0}/W_B) (\exp(qV_{EB}/kT) - 1)]$$

$$\approx 1 - W_B^2/(2L_p^2) \qquad (4\text{-}21)$$

- For an efficient device, it is desirable to have both γ and α_T close to 1

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

Solution:

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

The depletion layer width in the base-collector junction $= (2\varepsilon_s(N_{base} + N_{collector})(V_{bi} + V_{CB})/(qN_{base}N_{collector}))^{1/2} = (2 \times 11.9 \times 8.86 \times 10^{-12}\ F/m \times (10^{22}\ m^{-3} + 10^{21}\ m^{-3} \times (0.636\ V + 5\ V)/(1.6 \times 10^{-19}\ C \times 10^{22}\ m^{-3} \times 10^{21}\ m^{-3}))^{1/2} = 2.85\ \mu m$.

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

A similar calculation for the emitter-base junction gives a depletion region in the base of 0.217 μm.

The neutral base width = (1 - 0.26 - 0.217) μm = 0.523 μm. #

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 μs, respectively, determine the current components: I_{Ep}, I_{Cp} and I_B.

Solution:

It can be shown that: L_{En} = 1.41 μm, L_{Bp} = 10 μm and L_{Cn} = 59.2 μm.

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

Similarly, it can be shown that (see Eqn.(4-13)): I_{Cp} = 4.6650 mA. #

$I_B = I_{Ep} - I_{Cp}$ = 6.38 μA. #

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 x 10^{21} m^{-3}, compute α_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 x 10^{-6} m.

Solution:

From Eqn.(4-20): $\gamma \approx 1/(1 + D_E n_{E0} W_B/(D_p p_{n0} L_E)) = 1/(1 + 10^{-4}\ m_2/s \times 10^{23}\ /m^3 \times 0.5 \times 10^{-6}\ m\ /(10^{-3}\ m^2/s \times 10^{25}\ /m^3 \times 10^{-6}\ m)) = 0.9995$. $\alpha_T \approx 1 - W_B^2/(2L_p^2) = 0.9987$. 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:

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

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

(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, α_F is the *forward common-base current gain*, and α_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)):

66

$$i_E = I_E + i_E'$$

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

$$i_C = I_C + i_C'$$

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

$$i_B = I_B + i_B' \qquad (4\text{-}25)$$

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:

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

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

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

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

$$i_C = I_C + d(i_C)/dV_{EB}|_{VEC}\, v_{EB}' = I_C + i_C'$$

$$i_B = I_B + d(i_B)/dV_{EB}|_{VEC}\, v_{EB}' = I_B + i_B' \qquad (4\text{-}27)$$

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.

For convenience, we define $\beta_F = i_C/i_B = \alpha_F/(1 - \alpha_F)$ as the common-emitter current gain.

In addition,

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

$$g_{EB} = d(i_B)/dV_{EB}|_{VEC} = q\ (1 - \alpha_F)\ I_E/kT = qI_B/kT$$

(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* α is given by:

$$\alpha = \alpha_0/(1 + jf/f_\alpha) \qquad (4\text{-}29)$$

where α_0 is the low frequency common-base current gain, and f_α is the *common-base cutoff frequency*.

Similarly, the *frequency-dependent common-emitter current gain* β is given by:

$$\beta = \beta_0/(1 + jf/f_\beta) \qquad (4\text{-}30)$$

where β_0 is the low frequency common-emitter current gain, and f_β is the *common-emitter cutoff frequency*.

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

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

Both f_β and f_α are called the *3-dB frequencies* meaning that at these frequencies, β and α 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,

$$\beta = \beta_0/(1 + jf_T/f_\beta)) = 1$$

$$f_T = f_\beta \approx (\beta_0^2 - 1) \approx \beta_0 f_\beta \approx \alpha_0 f_\alpha \qquad (4\text{-}32)$$

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*** τ_B.

Transit time is given by:

$$\tau_B = \int^{W_B}_0 (1/v)\, dx = \int^{W_B}_0 (qp_nA_{cs}/I_p)\, dx \qquad (4\text{-}33)$$

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

$$\tau_B = \int^{W_B}_0 [W_B(1 - x/W_B)/D_p]\, dx$$

$$= W_B^2/(2D_p) \qquad (4\text{-}34)$$

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_{off} and an ON-resistance R_{on}.

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

These resistances are given by:

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

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

where V_C is the bias voltage, V_{CE} is the collector-emitter voltage and I_C is the collector current.

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:

$$Q_B = qA_{cs} \int^{WB}_0 (p_n - p_{n0})\, dx \qquad (4\text{-}36)$$

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

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

Integrating from x = 0 to x = W_B gives:

$$I_p(0) - I_p(W_B) = dQ_B/dt + Q_B/\tau_p \qquad (4\text{-}38)$$

where we have used the relationship $J_p = I_pA_{cs}$ and $Q_B = q(p_n - p_{n0})A_{cs}$. A_{cs} is the area cross-section of the device.

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

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

$$Q_B = i_B\tau_p (1 - \exp(- t/\tau_p)) \qquad (4\text{-}39)$$

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

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

Or,

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

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:

$$Q_B = Q_B(t_2) \exp(- (t - t_2)/ \tau_p) \qquad (4\text{-}41)$$

72

- 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:

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

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

$$i_C = (Q_B(t_3)/ \tau_p) \exp(- (t - t_3)/ \tau_p) \qquad (4\text{-}43)$$

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

Solution:

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

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

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

74

Part 5: MIS diode and MOS transistor

What we need to learn in this chapter?

Physical properties of the MIS diode

Capacitance characteristics

Physical structure of the MOS transistor

I-V characteristics

Other effects

Current 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.

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

76

where Φ_m is the ***metal work function***, χ is the ***electron affinity***, and ψ_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:

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

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 $Q_{SC} = - qN_AW$, where W is the depletion layer width.

Inside the MIS diode, the charge states in the semiconductor are best described by the potential ψ. Since the energy bands are drawn for electrons, positive potential (and ψ) increases ownward.

The value of ψ right at the oxide-semiconductor interface is the *surface potential*, ψ_s.

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

ψ_B is the energy difference between the Fermi-level and the intrinsic Fermi-level.

Note that ψ is zero in the bulk and this is shown in Fig.5.4.

78

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

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

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

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

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

$$n_p = n_i \exp(- q(\psi_B - \psi)/kT) \qquad (5\text{-}4)$$

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.,

$$d^2\psi/dx^2 = - \rho/\varepsilon_s = qN_A/\varepsilon_s \qquad (5\text{-}5)$$

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

$$\psi = \psi_s (1 - x/W)^2 \qquad (5\text{-}6)$$

where $\psi_s = qN_AW^2/(2\varepsilon_s)$.

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:

$$W_m = (2\varepsilon_s\psi_B/(qN_A))^{1/2} \qquad (5\text{-}7)$$

Similarly,

$$Q_{SC} = - qN_AW_m \qquad (5\text{-}8)$$

Example 5.1: For an ideal Si-SiO_2 MIS diode with d = 30 nm and N_A = 5 x 10^{21} 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 x 8.86 x 10^{-12} F/m/(3 x 10^{-8} m) = 1.15 x 10^{-3} F/m^2.

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

Ideal diode implies V_{FB} = 0 and $V_a = \psi_B + (2\varepsilon_s q N_A \psi_B)^{1/2}/C_o$ = 0.33 V + (2 x 11.9 x 8.86 x 10^{-12} F/m x 1.6 x 10^{-19} C x 5 x 10^{21} m^{-3} x 0.33 V)$^{1/2}$/1.15 x 10^{-3} F/m^3 = 0.53 V. #

ii) For strong inversion: $\psi = 2\psi_B$. This leads to: V_a = 0.95 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:

$$V_a = V_0 + \psi_s \qquad (5\text{-}9)$$

where V_0 is the voltage drop across the oxide and ψ_s is the surface potential.

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

81

One region is the oxide layer and there we have: $V_0 = Q_s d/\varepsilon_{ox}$, where $d/\varepsilon_{ox} = 1/C_0$, 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:

$$C_{SC} = \varepsilon_s / W = (q\varepsilon_s N_A/(2\psi_s))^{1/2} = [q\varepsilon_s N_A/(2(V_a - V_0))]^{1/2} \quad (5\text{-}10)$$

At strong inversion, $Q_s = qN_A W_m$.

This leads to:

$$V_a = qN_A W_m/C_0 + 2\psi_B \quad (5\text{-}11)$$

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.

82

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*:*

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

where Φ_{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_2 MIS diode at 300K 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: $C_0 = 1.15 \times 10^{-3}$ F/m^2.

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

84

The I-V characteristics of the MOS transistor are 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:

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

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 $Q_{SC} = -(2\varepsilon_s q N_A (V_a + 2\psi_B))^{1/2}$, one gets:

$$Q_n = C_0 (V_G - V_a - 2\psi_B) + (2\varepsilon_s q N_A (V_a + 2\psi_B))^{1/2} \qquad (5\text{-}14)$$

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

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

$$\Delta R = \Delta y/(Z\mu_n Q_n) \qquad (5\text{-}15)$$

where Z is the width of the device, μ_n is the electron mobility, and Δy 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:

$$\int_0^L I_D \, dy = \int_0^{VD} Z\mu_n Q_n \, dV \qquad (5\text{-}16)$$

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

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

$$I_D = Z\mu_n(C_0/L)\,[(V_G - 2\psi_B - V_D/2)\,V_D - \{(8\varepsilon_s q N_A)^{1/2}/3C_0\}\,((V_D + 2\psi_B)^{3/2} - (2\psi_B)^{3/2})] \qquad (5\text{-}17)$$

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

86

A graphical plot is shown in Fig.5.8.

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:

$$I_D \approx Z\mu_n (C_0/L) (V_G - V_T) V_D \qquad (5\text{-}18)$$

where $V_T = (4\varepsilon_s q N_A \psi_B)^{1/2}/C_0 + 2\psi_B$.

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

$$g_D = Z\mu_n (C_o/L) (V_G - V_T) \qquad (5\text{-}19)$$

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

Eqn.(5-14) now has the form: $x^2 + (2)^{1/2}K_2 x - V_G = 0$, where $x = (V_a + 2\psi_B)^{1/2}$ and $K_2 = (\varepsilon_s q N_A)^{1/2}/C_0$.

Solving this leads to: $x = [- (2)^{1/2} K_2 +/- (2K_2^2 + 4 V_G)^{1/2}]/2$.

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

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

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

$$I_{Dsat} \approx Z\mu_n\varepsilon_{ox} (V_G - V_T)^2/(2dL) \qquad (5\text{-}21)$$

where d is the oxide thickness.

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

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

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

Example 5.3: Consider a long channel MOSFET with L = 3 μm, Z = 21 μm, N_A = 5 x 10^{21} m^{-3}, C_0 = 1.5 x 10^{-3} F/m^2 and V_T = 1.5 V. Determine V_{Dsat} when V_G = 4 V.

Solution:

The substrate potential ψ_B = kT/q x ln(N_A/n_i) = 0.0259 V x ln(5 x 10^{21} m^{-3}/1.56 x 10^{16} m^{-3}) = 0.33 V.

K = $(\varepsilon_s q N_A)^{1/2}/C_0$ = (11.9 x 8.86 x 10^{-12} F/m x 1.6 x 10^{-19} C x 5 x 10^{21} /m^3)$^{1/2}$/1.5 x 10^{-3} F/m^2 = 1.93 x 10^{-2} $V^{1/2}$.

88

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

b) Frequency response

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

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

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:

$$f_T = \omega/2\pi = g_m/(2\pi\, C_0 ZL) = \mu_n V_D/(2\pi\, L^2) \qquad (5\text{-}24)$$

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:

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

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:

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

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

$$I_D \approx \exp(- q(V_G - V_T')/kT) \quad (5\text{-}27)$$

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,

$$V_T' = V_{FB} + 2\psi_B + (4\varepsilon_s q N_A \psi_B)^{1/2}/C_0 \qquad (5\text{-}28)$$

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} m^{-3}. Compute the threshold voltage if the substrate is: i) P-type; and ii) N-type. Assume an oxide thickness of 65 nm.

91

Solution:

i) For a P-type substrate: $q\Phi_{ms}$ = - 0.96 eV (see Fig.E5.1).
$\psi_B = kT/q \ln(N_A/n_i) = 0.0259 \text{ V} \times \ln(10^{20} \text{ m}^{-3}/(1.45 \times 10^{16} \text{ m}^{-3}) = 0.229 \text{ V}$.
$C_{ox} = \varepsilon_{ox}/d = 3.9 \times 8.86 \times 10^{-12} \text{ F/m}/(650 \times 10^{-10} \text{ m}) = 5.3 \times 10^{-4} \text{ F/m}^2$. #

$V_{TN}' = V_{FB} + 2\psi_B + (4\varepsilon_s q N_A \psi_B)^{1/2}/C_0 = - 0.96 \text{ V} + 2 \times 0.229 \text{ V} + (2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} \times 10^{20} \text{ m}^{-3} \times 0.458 \text{ V})^{1/2}/(5.3 \times 10^{-4} \text{ F}) = - 0.42 \text{ V}$. #

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

$V_{TP}' = V_{FB} + 2\psi_B + (4\varepsilon_s q N_D \psi_B)^{1/2}/C_0 = - 0.52 - 2 \times 0.229 - (2 \times 11.9 \times 8.86 \times 10^{-12} \text{ F/m} \times 10^{20} \text{ m}^{-3} \times 0.458 \text{ V})^{1/2}/(5.3 \times 10^{-4} \text{ F}) = - 1.04 \text{ V}$. #

Note that as ψ_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:

$$W_m' = (2\varepsilon_s (2\psi_B + V_{BS})/qN_A)^{1/2} \qquad (5\text{-}29)$$

Similarly,

$$V_{TN}' = V_{FB} + 2\psi_B + [2\varepsilon_s q N_A (2\psi_B + V_{BS})]^{1/2}/C_0$$

92

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

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 $v_{drift} = v_{sat}$, $I_{Dsat} = Zqv_{sat}$?$_0^{xi}$ n.dx, where x_i is the thickness of the depletion layer.

93

Since $Q_n = \int^{xi}_0 n\, dx \sim C_0 (V_G - V_T)$, one gets:

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

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

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:

$$L' = L/k,$$
$$d' = d/k,$$
$$Z' = Z/k,$$
$$V_a' = V_a/k \qquad (5\text{-}32)$$

This leads to:

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

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

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

$$P_{dc}' = I' V_a' = P_{dc}/k^2 \qquad (5\text{-}33)$$

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

94

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:

$$L_{min} = 0.4\,[r_j d(W_s + W_D)^2]^{1/2} \qquad (5\text{-}34)$$

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 μm. (The gate length is the channel length plus twice the junction depth.) If the junction depth is 0.2 μm, 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 μm - 2 x 0.2 μm = 0.35 μm. From Eqn.(5-34), $L_{min} = 0.4\,[r_j d(W_s + W_D)^2]^{1/3}$.

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

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

END

95