LINKÖPING UNIVERSITY
Per Larsson-Edefors
Electronic Devices, Dept. of Physics
E-mail: perla@ifm.liu.se

Lecture notes in

TFFY34 Semiconductor Technology

-

an undergraduate course at Linköping University

by

Per Larsson-Edefors
Electronic Devices, Department of Physics
Linköping University, SWEDEN

Fall 2000

LINKÖPING UNIVERSITY
Per Larsson-Edefors
Electronic Devices, Dept. of Physics
E-mail: perla@ifm.liu.se

Content

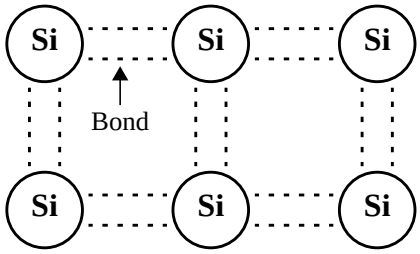
LECTURE 1

1. Some Basic Facts and Concepts

Semiconductor materials of importance are silicon (Si), germanium (Ge), gallium-arsenide (GaAs), zinc-selenide (ZnSe) and alloys such as $Al_xGa_{1-x}As$. Silicon-based semiconductors however totally dominate the present commercial market due to their advanced fabrication technology and this course, thus, mainly will focus on these semiconductors. Silicon is the second most abundant element in the earth's crust but nowhere is it to be found in a pure single-crystal form. It has to be man made from e.g. silica (impure SiO_2).

1.1. Energy Bands (*Secs 3.1.1 - 3.1.3*[1])

Sometimes valence electrons are shared, becoming a bond between two atoms - covalent bonding. This is the bonding type in diamond-crystal lattice semiconductors such as silicon semiconductors. However, it is more interesting to analyze energy-related aspects rather than spatial aspects such as bonds. Therefore the concept of energy bands is coming in handy.



An almost continuous band of allowed energies[2] of electrons comes about when atoms are brought in close proximity to each other, this is because of the interatomic forces and is foreseen in the Pauli exclusion principle. "Almost", well, one energy level is split into *N* levels when *N* atoms are brought together, and these *N* levels can accommodate at most 2*N* electrons due to spin de-

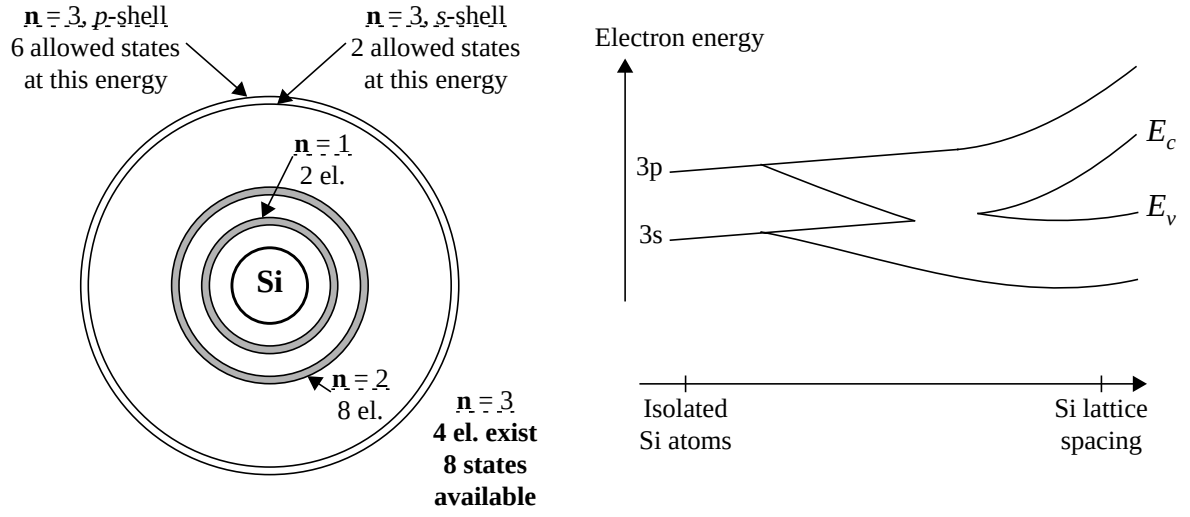


Fig. 1: (Left) Isolated Si atom, having 14 electrons. (Right) Energy bands are forming when a huge number of atoms are brought together.

generacy. Remember, *N* is huge! Now, since the separation between the energy levels within the band is much smaller than the thermal energy possessed by an electron at room temperature

1. Section, figure and equation references in *italic style* refer to the main textbook: Streetman and Banerjee, "Solid State Electronic Devices", 5th ed., Prentice-Hall Intl Editions, ISBN 0-13-025538-6, 2000.
2. See *Fig. 3-3* on page 60 in the main textbook.

the band can be viewed as continuous.

E_c is the lowest possible conduction band energy, while E_v is the highest possible valence band energy. The band gap energy, E_g, is furthermore defined as (E_c - E_v). E_g is the energy it takes to break a bond in the spatial view of the crystal. The band gap energies for some semiconductors at $T = 300$ K are: E_g = 1.42 eV in GaAs and 1.12 eV in Si. You do remember that 1 eV = 1.602×10^{-19} J, don't you?

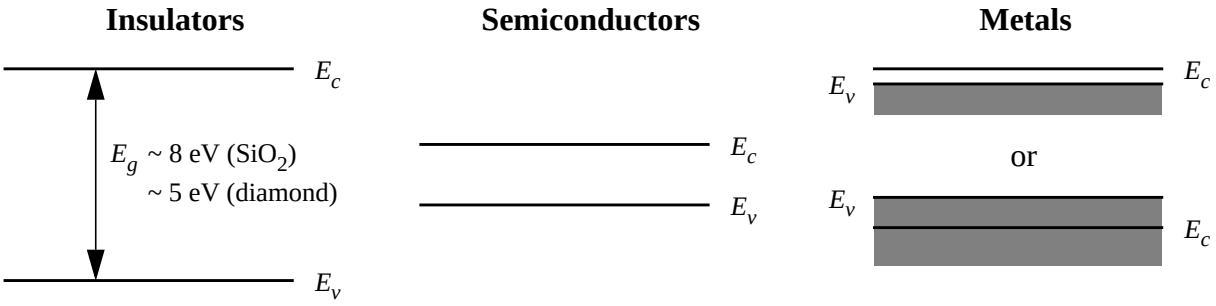


Fig. 2: Energy band gaps in insulators, semiconductors and metals.

1.2. Current Flow and the Concept of Holes (*Sec. 3.2.1*)

Using the concept of energy bands, the pure semiconductor (ideally, i.e. $T \rightarrow 0$ K) contains a completely filled (with electrons) valence band and a completely empty conduction band. Thus no current can flow - i.e. there are no electrons at all in the conduction band and no empty states (i.e. states containing no electrons) in the valence band to which electrons inside this band can move.

A hole is now defined as an empty state in the valence band.

1.3. Free Carriers - Excitation and Doping (*Secs 3.2.3 - 3.2.4*)

If there exist free electrons or holes, so-called charge carriers, charge transport can occur (current can flow).

If a semiconductor is excited by energy in the form of light, temperature or electric fields, electrons in the valence band can jump to the conduction band and take part in a current flow both as electrons in the conduction band and as holes in the valence band. This process is known as electron-hole pair generation (sometimes: intrinsic generation).

Another way to create (almost) free charge carriers is to contaminate a material with impurities that occupy lattice sites in place of the atoms of the pure semiconductor - so-called doping. The amount of doping, the doping density or concentration, is usually given as impurities/cm^3.

1.4. Doping (*Sec. 3.2.4*)

If the pure Si is doped with atoms from group V, i.e. they have one more valence electron, the N_d impurity atoms are called donors. Since four of the valence electrons from the impurity atom are enough to create the covalent bond, the fifth electron is almost free to move around. However, the fifth electron is weakly bound to the impurity atom by the excess positive charge of the nucleus and thus it needs a small amount of energy to become fully free - but when it becomes free only this carrier has been created, the positively charged dopant ion cannot move.

A donor-doped material where there are more electrons than holes is called an n-type material.

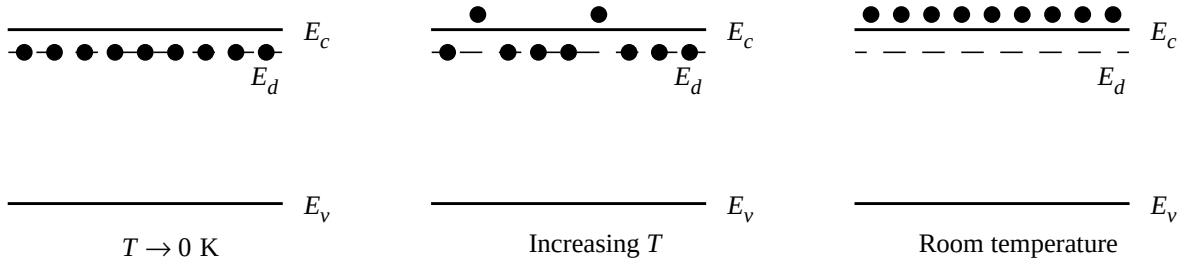


Fig. 3: Thermal excitation of a semiconductor doped with donors.

Instead, if the pure Si is doped with atoms from group III, i.e. they have one less valence electron, the N_a impurity atoms are called acceptors. Since there are only three valence electrons in the impurity atom instead of the four needed to create the covalent bond, a fourth electron has to be borrowed from a nearby bond and in this way a hole is created, a hole that is almost free to move around. Similar to the case of the donor impurity, only a small amount of energy is needed to lift the electron from the valence band into the energy level of the vacant bond - but when it becomes free only this carrier has been created, the negatively charged acceptor ion cannot move.

An acceptor-doped material where there are more holes than electrons is called a p-type material.

The majority carrier is the most abundant carrier in a given semiconductor sample; electrons in n-type and holes in p-type materials. Similarly the minority carrier is the least abundant carrier in a given semiconductor sample; holes in n-type materials and electrons in p-type materials.

1.5. Effective Mass (*Sec. 3.2.2*, part of *Sec. 3.3.2*)

At the atomic level quantum mechanics rule the world. However, it is possible to use the classical, and simple, second law of Newton for crystals that are large compared to atomic dimensions. In vacuum the force on an electron in the electric field $\bar{E}$ [1] is

$$\bar{F} = -q\bar{E} = m_0 \frac{d}{dt}\bar{v}. \tag{1}$$

By taking into account the periodic potential which is present in a perfect semiconductor crystal (with no scattering!) this equation can be written as

$$\bar{F} = m_n^* \frac{d}{dt}\bar{v}, \tag{2}$$

which is a great simplification to reality and allows us to simplify device analyses. Here m_n^* denotes the effective mass of an electron and $\bar{v}$ is the group velocity of the wave packet that describes the electron motion. In a similar fashion, the empty states in the valence band, the holes, have an effective mass of m_p^*.

1. The electric field is always written either as a vector $\bar{E}$ or as a scalar in one-dimension E_x in these notes.

Using the formula for kinetic energy and the relationship for the electron momentum, $\bar{p} = m\bar{v} = \hbar\bar{k}$, we get

$$E = \frac{1}{2} m\bar{v}^2 = \frac{1}{2m}\hbar^2 \bar{k}^2 . \tag{3}$$

From the (E, $\bar{k}$) relationship of Eq. (3) we can write

$$m^* = \hbar^2 \left(\frac{d^2 E}{d\bar{k}^2} \right)^{-1} .$$

It is very important to note that the quantity of effective mass varies with temperature as well as direction in the crystal. Also it should be noted that there exist different types of effective masses such as conductivity effective mass and density of states effective mass - in Si $m_n^* = 0.26\, m_0$ and $1.18\, m_0$, respectively. As a tensor quantity effective mass exists as conductivity effective mass and density of states effective mass, which are calculated as

$$\frac{1}{m_c} = \frac{1}{3}\left(\frac{1}{m_x} + \frac{1}{m_y} + \frac{1}{m_z} \right)$$

and

$$m_d = (m_x\, m_y\, m_z)^{1/3},$$

respectively.

What makes the concept of effective mass so important in this course is that the effective masses for electrons and holes differ, which greatly affects the behavior of semiconductor devices. For example, in Si holes are three times heavier than electrons when we discuss conductivity, which means that fast Si transistors are always based primarily on n-type semiconductors.

"Weakly bound" in Sec. 1.4, how much is that in for example Si? It is possible to use the energy estimation for a hydrogen atom (according to Bohr) - by replacing m_0 with m_n^* for Si conductivity and using the relative dielectric constant of Si ($\varepsilon_r \approx 11.8$) - to find a value of the <u>binding energy</u> (E_c - E_d in Fig. 3):

$$E_B \approx -\frac{m_n^* q^4}{2(4\pi\, \varepsilon_r \varepsilon_0\, \hbar)^2} = \frac{m_n^*}{m_0}\frac{1}{\varepsilon_r^2} E_H \approx -0.025 \text{ eV}$$

1.6. Intrinsic and Extrinsic Materials (*Secs 3.2.3 - 3.2.4*)

A semiconductor material with no impurities added is called an <u>intrinsic</u> semiconductor. In this material obviously the number of electrons in the conduction band must equal the number of holes in the valence band - this is due to the fact that any electron in the conduction band has been excited there and left a vacant state, a hole, in the valence band (electron-hole pair generation). We usually refer to the density of electrons and holes in the intrinsic semiconductor as n_i and p_i, respectively, and give these values as carriers/cm^3. **NOTE** that n_i and p_i depend strong-

ly on temperature[1] since these densities are due to electron-hole pair generation.

Typical values of n_i are: 2×10^6/cm^3 in GaAs, 1×10^{10}/cm^3 in Si and 2×10^{13}/cm^3 in Ge at $T = 300$ K. Compare with, in Si, 5×10^{22} atoms/cm^{3}and with four bonds (valence electrons) per atom this yields a total of 2×10^{23} valence electrons/cm^3. With an n_i of 10^{10}/cm^3, less than one bond in 10^{13} is broken in Si at room temperature.

When impurities are added, by doping, to a semiconductor it is said to be an extrinsic semiconductor. At equilibrium[2] the extrinsic semiconductor is said to have carrier concentrations n_0 and p_0, both different from n_i. **NOTE** this exception: at $T \to 0$ K, $n_0 = p_0 = n_i = 0$. This state is sometimes called freeze-out.

1. In *Sec. 3.3.3* and Sec. 4.2 in Lecture 2.
2. Equilibrium means a case where there is no external excitation except temperature and no net motion of charge.

LECTURE 2

2. Density of States (*Appendix IV*)

The density of states function, $N(E)$, describes the distribution of energy states, i.e. the number of states per unit energy and unit volume:

$$N_c(E) = \frac{4\pi\,(2m_{dn}^*)^{3/2}\,(E-E_c)^{1/2}}{h^3}, \tag{4}$$

and

$$N_v(E) = \frac{4\pi\,(2m_{dp}^*)^{3/2}\,(E_v-E)^{1/2}}{h^3}.$$

How come these functions look like this?

We consider a cubic region of a crystal with dimensions L along the three perpendicular directions and impose the condition that the electron wave functions become zero at the boundaries of a cube defined by values of x, y and z equal to 0 and L. The boundary conditions are satisfied by a wave function of the form

$$\Psi_k(r) = U_k(r)\sin k_x x \sin k_y y \sin k_z z$$

where $U_k(r)$ is a periodic function. The boundary conditions lead to

$$k_x L = 2\pi n_1$$

$$k_y L = 2\pi n_2$$

$$k_z L = 2\pi n_3$$

where n_1, n_2 and n_3 are integers. Each allowed value of $\bar{k}$ with coordinates k_x, k_y and k_z occupies a volume $((2\pi)/L)^3$ in $\bar{k}$-space. In other words, the density of allowed points in $\bar{k}$-space is $V/(2\pi)^3$ where $V = L^3$ is the crystal volume.

The spherical volume in $\bar{k}$-space defined by vectors $\bar{k}$ and $\bar{k} + d\bar{k}$ is, when $d\bar{k} \to 0$, $4\pi k^2\,dk$. Hence, the total number of states with k values between k and $k + dk$ is

$$dN = 4\pi k^2\,dk \cdot \frac{V}{(2\pi)^3}, \text{ and with spin taken into account}$$

$$dN = \frac{8\pi V k^2\,dk}{(2\pi)^3}$$

Based on the relationship $E = \frac{1}{8\pi^2 m} h^2 \bar{k}^2$ we can, for electrons in the conduction band of a semiconductor, write

$$k^2 = \frac{8\pi^2 m_{dn}^*(E-E_c)}{h^2}, \text{ and}$$

$$2kdk = \frac{8\pi^2 m_{dn}^* dE}{h^2}$$

Inserting the last two equations in the expression for *dN*, we obtain

$$dN = \frac{8\pi V}{(2\pi)^3}\left(\frac{8\pi^2 m_{dn}^*(E-E_c)}{h^2}\right)^{1/2}\left(\frac{4\pi^2 m_{dn}^* dE}{h^2}\right) = \frac{4\pi(2m_{dn}^*)^{3/2}(E-E_c)^{1/2}}{h^3}\cdot VdE$$

3. The Fermi-Dirac Function (*Sec. 3.3.1*)

Whereas the density of states function tells one how many states exist at a given energy *E*, the Fermi-Dirac function *f*(*E*) specifies how many of the existing states at energy *E* will be filled with an electron. More formally, *f*(*E*) specifies, **under equilibrium conditions**, the probability that an available state at an energy *E* will be occupied by an electron.

$$f(E) = \frac{1}{1+e^{\frac{(E-E_F)}{kT}}}, \tag{5}$$

where E_F is the Fermi level or Fermi energy, *k* is the Boltzmann constant and *T* is the temperature in Kelvin (kT = 0.0259 eV at room temperature).

Some observations: At $T \to 0$ K $f(E<E_F) = 1$ and $f(E>E_F) = 0$, i.e. no electrons occupy states above the Fermi level, but they are confined to all states below E_F. Also, for $T > 0$ K, $f(E = E_F) = 0.5$.

4. Carrier Concentrations at Equilibrium (*Sec. 3.3.2*)

The carrier concentration between the energy levels E_1 and E_2, *n*, can now be expressed as the integral

$$n = \int_{E_1}^{E_2} N(E-E_1)\ f(E-E_1)\ dE\,.$$

Consequently, the electron concentration in the conduction band at equilibrium, n_0, can be written as a combination of Eq. (4) and Eq. (5):

$$n_0 = \frac{4\pi}{h^3}(2m_{dn}^*)^{3/2}\int_{E_c}^{E_T}\frac{(E-E_c)^{1/2}}{1+e^{\frac{(E-E_F)}{kT}}}\,dE\,,$$

where E_T is the upper band edge of the conduction band.

The expression for n_0 can be simplified in two aspects:

1. The semiconductor materials that we discuss here are so-called non-degenerate semiconductors. These materials are fairly lightly doped and thus the exponential term in $1+e^{(E-E_F)/kT}$ is large compared to unity, i.e. this expression can be approximated by $e^{(E-E_F)/kT}$.

2. Since there are not many electrons in the upper part of the conduction band, E_T can be

replaced by infinity.

Now, in the integral, variables are replaced such that

$$x = \frac{E - E_c}{kT} \text{ and } \eta = \frac{-(E_c - E_F)}{kT},$$

which implies that $dE = kT\,dx$ and that the lower limit of the integration becomes 0. Hence

$$n_0 = \frac{4\pi}{h^3}(2m_{dn}^*)^{3/2}\left(\int_0^\infty \frac{(x \cdot kT)^{1/2}}{e^{(x-\eta)}}\,kT\,dx\right) =$$

$$= \frac{4\pi}{h^3}(2m_{dn}^*)^{3/2}(kT)^{3/2}\,e^{\eta}\left(\int_0^\infty \frac{x^{1/2}}{e^x}\,dx\right), \tag{6}$$

where the integral is a gamma function with the solution $\frac{\pi^{1/2}}{2}$. Thus, we have

$$n_0 = 2\left(\frac{2\pi\,kT\,m_{dn}^*}{h^2}\right)^{3/2} e^{\frac{-(E_c - E_F)}{kT}},$$

which can be formulated as

$$n_0 = N_c\,e^{\frac{-(E_c - E_F)}{kT}} \tag{7}$$

where N_c is the effective density of states in the conduction band.

In a similar way the carrier concentration in the valence band can be obtained as

$$p_0 = N_v\,e^{\frac{-(E_F - E_v)}{kT}}, \tag{8}$$

where

$$N_v = 2\left(\frac{2\pi\,kT\,m_{dp}^*}{h^2}\right)^{3/2}.$$

4.1. Non-Degenerate and Degenerate Semiconductors (*Sec. 10.1.1*)

As mentioned earlier, non-degenerate semiconductors are fairly lightly doped. The formal condition for being such a semiconductor is that, at equilibrium, the location of the Fermi level is inside the band gap and at least $3kT$ from the conduction and valence band edges. In the previous calculations the implication of this was that the exponential term in the Fermi-Dirac function would dominate over unity, leading to an algebraic simplification ending up in the Maxwell-Boltzmann function.

When the Fermi level is within $3kT$ of either the conduction or valence band edge then the semiconductor is said to be a degenerate semiconductor.

The most important implication of heavier doping is that the bandgap starts to become

more narrow due to bandtail states in the perturbed density of states function. When the average spacing of the impurity atoms approaches the Bohr radius of 100 Å, due to heavy doping, the potential seen by each impurity electron (or hole) is affected by the neighboring impurity atoms and impurity energy bands are formed. Here, heavy is defined as a doping density of more than ~10^{18} atoms/cm^3. Already at a doping density of 10^{19} atoms/cm^3 the bandgap narrowing ΔE_g can be more than 10% of E_g.

A practical consequence of heavy doping is that integral in Eq. (6) is not a gamma function but rather becomes an intricate Fermi-Dirac integral of order 1/2 which is not a closed-form expression.

4.2. Carrier Concentrations - Observation 1

If we multiply Eq. (7) and Eq. (8) we arrive at

$$n_0\, p_0 = N_c N_v\, e^{\frac{(E_F - E_c - E_F + E_v)}{kT}} = N_c N_v\, e^{\frac{(E_v - E_c)}{kT}} = N_c N_v\, e^{\frac{-E_g}{kT}} \tag{9}$$

which is constant for one kind of material and does not depend on doping.

In an intrinsic semiconductor $n_0 = p_0$, which means that $(E_c - E_F) \approx (E_F - E_v)$. In most cases an exact equality is not true. The intrinsic Fermi level, E_i, must therefore be close to the middle of the bandgap ($(E_c + E_v)/2$) and consequently Eq. (7) and Eq. (8) can be used to describe the intrinsic carrier concentration as

$$n_i = p_i = N_c\, e^{\frac{-(E_c - E_i)}{kT}} = N_v\, e^{\frac{-(E_i - E_v)}{kT}}. \tag{10}$$

The product n_i^2 can according to Eq. (10) be expressed as

$$n_i^2 = N_c N_v\, e^{\frac{-E_g}{kT}},$$

which combined with Eq. (9) yield a relationship sometimes referred to as the law of mass action:

$$n_i^2 = n_0\, p_0. \tag{11}$$

NOTE that this does only hold for non-degenerate semiconductors, obviously at equilibrium.

4.3. Carrier Concentrations - Observation 2

Solving for N_c in Eq. (10) and substituting this into Eq. (7) yields

$$n_0 = n_i\, e^{\frac{-(E_c - E_F)}{kT} - \frac{-(E_c - E_i)}{kT}} = n_i\, e^{\frac{(E_F - E_i)}{kT}}. \tag{12}$$

Similarly the insertion of N_v (from Eq. (10)) into Eq. (8) yields

$$p_0 = n_i\, e^{\frac{(E_i - E_F)}{kT}}. \tag{13}$$

Quite pedagogically Eq. (12), for example, shows how the electron concentration increases when the Fermi level comes closer the conduction band and vice versa.

5. Charge Neutrality (*Sec. 3.3.4*)

In a piece of a uniformly doped semiconductor each and every section has to be charge-neutral, otherwise a current would flow at equilibrium. In a general case where a material is doped with N_d donors and N_a acceptors, at room temperature all impurities are ionized, so-called complete ionization, so that there exist N_d positive ions and N_a negative ions (Fig. 3). Thus, for charge neutrality to hold we must have

$$p_0 - n_0 + N_d - N_a = 0. \tag{14}$$

Using n_0 from Eq. (11) in Eq. (14) yields an expression that is the foundation for determining the carrier concentration in a p-type material

$$p_0 - \frac{n_i^2}{p_0} + N_d - N_a = 0 \Rightarrow p_0^2 - p_0(N_a - N_d) - n_i^2 = 0,$$

which has a solution based on material parameters which are mostly known

$$p_0 = \frac{N_a - N_d}{2} + \sqrt{\left(\frac{N_a - N_d}{2}\right)^2 + n_i^2}. \tag{15}$$

NOTE that this is under the assumption that all impurities are ionized. A corresponding equation can be written for the electron concentration.

In most practical cases the net dopant concentration is much larger than the intrinsic concentration and Eq. (15) simplifies to

$$p_0 = N_a - N_d,$$

which in turn, with Eq. (11), suggests that

$$n_0 = \frac{n_i^2}{p_0} = \frac{n_i^2}{N_a - N_d}$$

under the usual assumptions that we are dealing with non-degenerate semiconductors at equilibrium.

Also, in most cases one dopant type is in vast majority compared to the other. Then

$$N_a \gg N_d \Rightarrow p_0 \approx N_a \text{ and } N_d \gg N_a \Rightarrow n_0 \approx N_d.$$

5.1. Compensation (*Sec. 3.3.4*)

When one dopant type does not dominate over the other type in number, it is often a characteristic of a compensation situation. Here, both types of impurities are used in the doping. A very important example can be found in the manufacturing of integrated circuits. Starting from a lightly doped p-type piece of Si, tubs of n-type are created on and inside the material by further local doping of donors. In this fashion it is possible to create transistors with different polarity on the same chip.

6. Overview of Carrier Action (*Secs 3.4, 4.3, 4.4*)

When there is a net flow of charge carriers in a semiconductor we are talking about carrier ac-

tion. The equilibrium condition has been the foundation for earlier discussions because it is serves as an important frame of reference for, for example, carrier action. But in practical situations a semiconductor is not used at equilibrium conditions and thus, the next topic will be an overview of the different mechanisms that provide the basis of charge transport in a semiconductor.

There are three kinds of carrier action; drift, diffusion and recombination and generation.

6.1. Drift

Drift is the mechanism in operation when an external electric field $\bar{E}$ is applied to the semiconductor; charged particles respond to the electric field by moving, depending on the charge, along the field or opposite to it. For electrons in one dimension the following expression for current density holds:

$$J_{n,x} = nq\mu_n E_x,$$

where μ_n is the mobility of electrons. The drifting motion is actually superimposed upon the always-present thermal motion of the carriers, which can be approximated by statistical mechanics as

$$\frac{m_n^* v_{th}^2}{2} = \frac{3kT}{2}.$$

For Si at room temperature we can estimate v_{th} in three dimensions as

$$v_{th} = \sqrt{\frac{3kT}{0.26 \cdot m_0}} \approx 2\times10^5 \text{ m/s},$$

i.e. the thermal velocity ~ 1/1000 the velocity of light.

6.2. Diffusion

Diffusion is the process whereby particles tend to spread out or redistribute as a result of their random thermal motion, migrating on a macroscopic scale from regions of high particle concentration into regions of low particle concentration. **NOTE** that not only charged but also neutral particles diffuse.

The drift mechanism is easily understood since it is identical to the usual current transport mechanism in metals. Diffusion however can only be found in semiconductor in contrast to metals, but is it unique, no. Consider a sealed perfume bottle put in the corner of a room. Open the bottle and guess what happens in the opposite corner of the room. In time, our experience tells us, the scent from the perfume will be all over the room, also in the opposite corner. This is an example of diffusion.

Mathematically the diffusion of electrons in a non-uniformly doped semiconductor can be written as

$$J_{n,x} = qD_n \frac{dn}{dx},$$

where D_n is the electron diffusion coefficient.

6.3. Einstein's Relationship

Do drift and diffusion at all relate to each other, the mechanisms seem very different?

By adding the electron current density formulae of drift and diffusion for a non-uniformly doped semiconductor at equilibrium we get

$$J_{n,x} = nq\mu_n E_x + qD_n \frac{dn}{dx} = 0 \tag{16}$$

since no current can flow.

NOTE that there is an electric field present inside the semiconductor - a so-called built-in field! This is compensating for the current due to diffusion, leading to a net current of exactly zero.

We can replace the doping gradient in Eq. (16) by using the derivative of Eq. (12)

$$\frac{dn}{dx} = \left(-\frac{1}{kT}\frac{dE_i}{dx}\right)\cdot n_i\, e^{\frac{(E_F-E_i)}{kT}} = \left(-\frac{1}{kT}\frac{dE_i}{dx}\right)\cdot n\,. \tag{17}$$

As will be discussed in Sec. 7.6, the Fermi level is invariant at equilibrium and thus the derivative of E_F is zero and has already been removed from this equation. However, the derivative of E_i, what is that? The electric field in one dimension is proportional to the gradient of energy inside the semiconductor - see Sec. 7.5 - for convenience sake we can use E_i as reference for energy. We have

$$E_x = \frac{1}{q}\frac{dE_i}{dx} \Rightarrow \frac{dE_i}{dx} = qE_x\,. \tag{18}$$

Inserting Eq. (18) in Eq. (17) yields

$$\frac{dn}{dx} = -\frac{qE_x}{kT}\,n\,. \tag{19}$$

Now we use Eq. (19) to replace the doping gradient in Eq. (16). Hence

$$nq\mu_n E_x - qD_n\left(\frac{qE_x}{kT}\,n\right) = (nqE_x)\,\mu_n - (nqE_x)\,D_n\,\frac{q}{kT} = 0\,.$$

The solution to this equation is the Einstein relationship that links diffusion to drift

$$\frac{D_n}{\mu_n} = \frac{kT}{q}\,. \tag{20}$$

6.4. Recombination and Generation

Recombination-Generation (R-G) is not manifested through carrier transport, but rather affects current densities by changing the carrier concentration. Unlike drift and diffusion the terms recombination and generation do not refer to a single process; there are several similar processes based on R-G as shown in Fig. 4. Either the R-G process goes directly from band to band or it passes through some localized allowed energy state, an R-G center, sometimes referred to as a defect state or trap.

For the sake of completeness, there exists a third kind of recombination, of which we

shall speak little, the Auger recombination. Here collision of carriers essentially leads to a drop in energy level for one carrier, thus transferring energy to the second carrier, which subsequently loses energy in the form of phonons(1). This process is likely to occur in materials which have fairly small bandgaps and high concentrations of carriers.

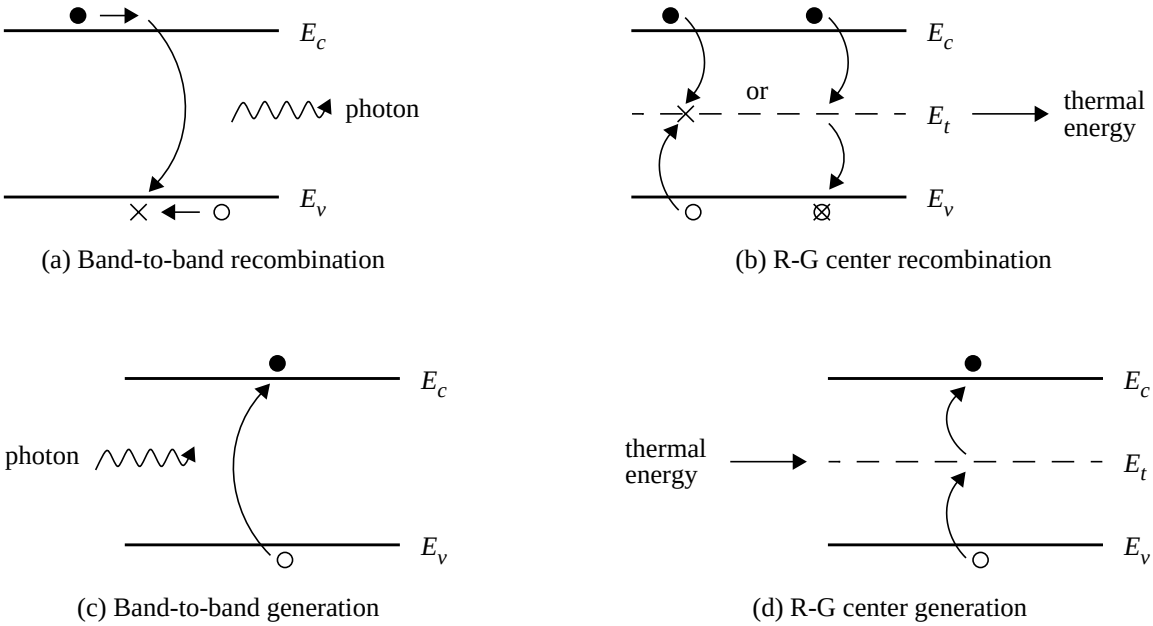


Fig. 4: Energy band visualization of recombination and generation processes.

In the context of R-G, it is important to notice the differentiation between direct and indirect bandgaps(2). As shown in Fig. 5, in direct bandgap materials the energy minima of both the conduction and valence band occur at $k = 0$. In the case of an indirect bandgap material the minimum of the conduction band is displaced to a non-zero momentum in the $\bar{k}$-space. (It so happens that the valence band maxima of Si, Ge and GaAs are at $k = 0$.)

In a direct transition, does the carrier only change energy in the vertical dimension, along the E-axis(3)? **NO**, direct here means that the transition occurs without any R-G center visited in between conduction and valence band, but not necessarily without a change in momentum. Direct transitions not only take place in direct bandgap materials, although they are quite common in these, they can also occur in indirect bandgap materials but it takes a simultaneous three-particle interaction to achieve that; one electron, one hole and one phonon. Such three-particle interactions seldom occur and thus we don't see many direct transitions in indirect bandgap materials.

The unique feature of indirect transitions is that they require an R-G center via which the R-G makes the transition. There is a fairly large probability of occurrence for the two-particle interaction(4), such as that between a free carrier and a phonon that can take place if there are R-G centers into which electrons and holes can make transitions. Consequently, indirect

1. A phonon is a lattice vibration quanta, i.e. a kind of particle defined in a similar way to the photon. The result we can observe from these vibrations is heat.
2. In *Sec. 3.1.4.*
3. Compare to the somewhat misleading *Fig. 3-5* on page 63 in the main textbook.
4. In comparison to the probability of occurrence for simultaneous three-particle interactions, the probability of occurrence for simultaneous two-particle interactions is huge.

transitions are much more likely to occur than direct transitions in indirect bandgap materials.

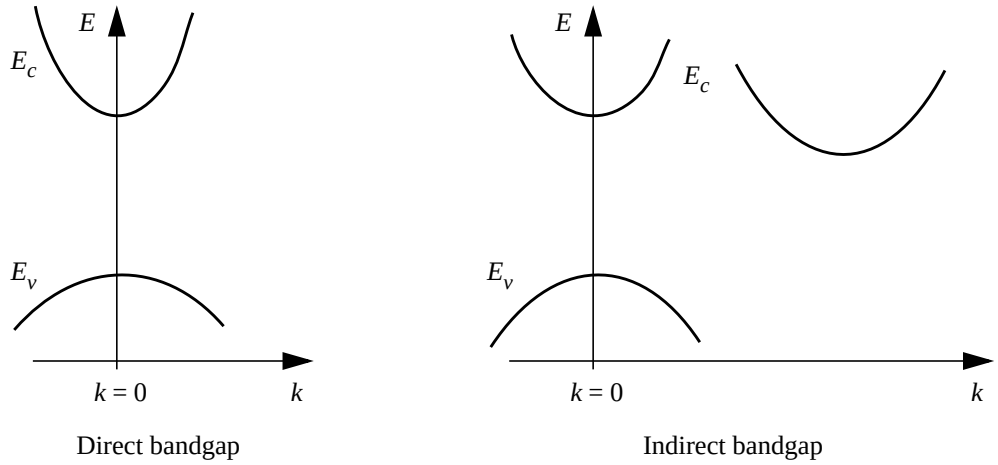


Fig. 5: Energy band examples of direct and indirect bandgap materials.

NOTE that even in many direct bandgap materials R-G processes are mostly based on R-G center transitions rather than on direct band-to-band transitions. This is due to unintentional R-G centers, which always, to some extent, are present even in pure semiconductor materials. Crystal imperfections and contaminants (unwanted impurities) are the most common mechanisms causing these R-G centers to be created.

A final comment on the relation between R-G and the (E, $\bar{k}$) relationship: since phonons carry quite a large momentum but little thermal energy, whereas photons practically have no momentum but a considerable amount of thermal energy, phonon-based transitions are almost horizontal and photon-based transitions are almost vertical in the $\bar{k}$-space. **NOTE** that the arrows are not perfectly vertical nor perfectly horizontal, even though they might look like that in the drawings.

To quantify the transfer of momentum from photon to an electron, we can write an equation for the ratio of the wave vector of the photon involved in the band-to-band transition, k_p, to the maximum electron wave vector, k_{emax}, as

$$\frac{k_p}{k_{emax}} = \frac{2E_g\, a}{hc},$$

where a is the lattice constant. For Si we have a ratio around 10^{-3}.

This course is primarily concentrating on silicon devices, as these represent the vast majority of manufactured devices. However, in some circumstances materials belonging to the III-V category, such as GaAs, have to be used because of their larger bandgap. For GaAs there exists one more recombination mechanism of importance, the surface recombination, which is due to dangling bonds in the crystal structure at the surface. This subject matter is not included in the course, but if you are interested you can read about it in the supplementing textbook by Casey, Sec. 3.7.3 and Sec. 4.7.6.

LECTURE 3

7. Drift (*Sec. 3.4.1*)

From the discussion on effective mass, Eq. (1) and Eq. (2), we remember that an electron inside a perfect semiconductor crystal experiences a force from an electric field such that

$$-qE_x = m_n^* \frac{dv_x}{dt}.$$

This equation suggests a continuous acceleration of the electron, which is a description that is not true to reality. In a real situation the electron velocity is limited by collisions inside the crystal. The collision rate is dependent on the mean time between scattering events, the mean free time τ_{cn}[1], such that the probability for a collision in the time interval dt is dt/τ_{cn}. Thus the differential change in electron velocity due to scattering can be expressed as

$$dv_x = -v_x \frac{dt}{\tau_{cn}},$$

which can be written as

$$\left.\frac{dv_x}{dt}\right|_{\text{collisions}} = -\frac{v_x}{\tau_{cn}}.$$

The general description of an electron in an electric field now becomes

$$-qE_x - \frac{m_n^* v_x}{\tau_{cn}} = m_n^* \frac{dv_x}{dt}, \tag{21}$$

which is a differential equation with the solution

$$v_x = -\frac{q\,\tau_{cn}}{m_n^*}\left(1 - e^{-t/\tau_{cn}}\right) E_x.$$

Eq. (21) can be compared to *Eq. (3-34)*[2], where steady state is assumed. In Eq. (21) we, however, allow for a difference in acceleration by the field and deceleration by the collisions, an impulse difference that is transferred to the electron and increases its velocity. The exponential behavior of the solution to Eq. (21) indicates that the velocity increases exponentially with time to the steady state condition when a "long" time has passed since we turned the electric field on at $t = 0$.

Simplifying the expression of the velocity to steady state, we hence arrive at the current density for n electrons with charge $-q$ as

$$J_{n,x} = \frac{nq^2\tau_{cn}}{m_n^*} E_x = nq\mu_n E_x,$$

1. $\tau_{cn} = \bar{t}$ in the main textbook. Sometimes the mean free time is referred to as momentum relaxation time.
2. On page 94 in the main textbook.

where μ_n is the electron mobility which consequently is defined as

$$\mu_n = \frac{q\tau_{cn}}{m_n^*}. \tag{22}$$

In an analogous way the current density and mobility can be defined for holes:

$$J_{p,x} = pq\mu_p E_x,$$

where

$$\mu_p = \frac{q\tau_{cp}}{m_p^*}.$$

In relation to the previous formulations there are two **NOTES**: **1.** The effective mass made use of here is the conductivity effective mass. **2.** A constant mean free time is assumed. However, as will be shown later in Sec. 7.4 when the electric field is very large and the drift velocity v_x approaches the electron saturation velocity v_{sn}, this assumption is not valid.

7.1. Conductivity and Resistivity (*Secs 3.4.1 - 3.4.2*)

The definition of the conductivity of the semiconductor, σ, as

$$\sigma = nq\mu_n + pq\mu_p$$

allows us to define a relationship between electric field and current density,

$$J_x = \sigma E_x.$$

This can also be formulated with the use of resistivity, $\rho = \frac{1}{\sigma}$, instead, leading to

$$E_x = \rho J_x$$

a relationship that essentially is Ohm's law.

Ex. 1: Drift - An Example

Assignment:

Consider two GaAs samples at room temperature where all impurities are completely ionized. Both samples have a length and a cross-sectional area of 1 cm and 5 mm^2, respectively. Sample 1 has $N_d = 10^7$ cm^{-3} and $N_a = 0$, i.e. it is of n-type. Sample 2 has $N_a = 10^7$ cm^{-3} and $N_d = 0$, i.e. it is of p-type. Find the current through each sample when a voltage of 10 V is applied across the entire length.

Solution:

This is an application of the drift mechanism and we obviously can use

$$I = qA\,(n\mu_n + p\mu_p)\,E_x$$

to find the total current.

Now, q (1.6×10^{-19} C), A (5 mm^2) and E_x (10 V/cm) are known. Also, the mobilities can be found in Appendix III of the main textbook, as μ_n = 8500 cm^2/Vs and μ_p = 400 cm^2/Vs

for GaAs.

What remain to be calculated are the carrier concentrations:

From *Fig. 3-17* on page 89 in the main textbook we can extract n_i as 2×10^6 cm^{-3}. Thus, in both samples the doping levels are modest in comparison to the intrinsic carrier concentration and we cannot assume either $n = N_d$ or $p = N_a$.

Sample 1:
Here we have

$$n_0 = \frac{N_d}{2} + \sqrt{\left(\frac{N_d}{2}\right)^2 + n_i^2} = \frac{10^7}{2} + \sqrt{\left(\frac{10^7}{2}\right)^2 + (2\times10^6)^2}$$

which becomes

$$n_0 = 5\times10^6 + 5.4\times10^6 = 1.04\times10^7 \text{ cm}^{-3}.$$

The intrinsic carrier concentration thus slightly affects the total concentration. But n_i also has an effect on the p_0 value in Sample 1 since

$$p_0 = \frac{n_i^2}{n_0} = \frac{(2\times10^6)^2}{1.04\times10^7} = 3.85\times10^5 \text{ cm}^{-3}.$$

Thus, since Sample 1 is lightly doped the majority and the minority carrier concentrations do not differ very much.

Based on **cm**, for Sample 1 we can now easily calculate the current as

$$I = 1.6\times10^{-19} \cdot 5\times10^{-2} \cdot (1.04\times10^7 \cdot 8500 + 3.85\times10^5 \cdot 400) \cdot 10$$

which yields I = 7.1 nA.

Sample 2:
Since N_d in Sample 1 is equal to N_a of Sample 2 and only one dopant type was used in either case, n_0 and p_0 of Sample 2 are equal to p_0 and n_0, respectively, of Sample 1. Thus

$$n_0 = 3.85\times10^5 \text{ cm}^{-3}$$

and

$$p_0 = 1.04\times10^7 \text{ cm}^{-3}.$$

The current through Sample 2 consequently is

$$I = 1.6\times10^{-19} \cdot 5\times10^{-2} \cdot (3.85\times10^5 \cdot 8500 + 1.04\times10^7 \cdot 400) \cdot 10$$

which yields I = 0.59 nA.

These samples are not very good at conducting current and that's due to the extremely light doping they have experienced.

NOTE that even a very pure sample will probably have an impurity density of at least 10^{12} cm^{-3}, but these are impurities of all sorts!

7.2. The Dependence of Mobility on Temperature (*Sec. 3.4.3*)

When the temperature increases, the periodic lattice of the crystal vibrates increasingly. Thus,

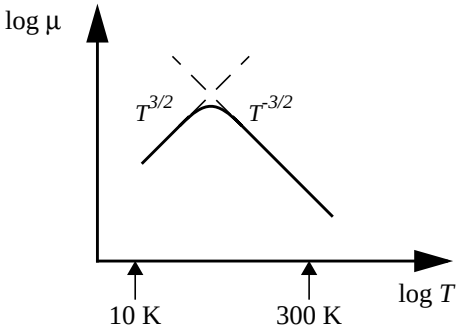


Fig. 6: Temperature dependence of mobility.

the mobility is reduced at high temperatures due to lattice scattering (also called phonon scattering. **NOTE** that the lattice scattering only is due to the displacement of lattice atoms from their lattice positions, the internal field associated with the stationary array of atoms is already taken into account in the effective mass formulation.

For lower temperatures there is a corresponding scattering mechanism although this depends on the low momentum of the carrier. In the impurity scattering scenario the carriers travel so slowly that the inevitable collisions with impurity atoms cause the carriers to be scattered more severely than at high temperatures.

NOTE the order of magnitude for “low” and “high” temperatures in Fig. 6 and compare to *Fig. 3-22*[1].

7.3. The Dependence of Mobility on Doping (*Sec. 3.4.3*)

It is obvious that if we increase the amount of ions in the semiconductor, we also increase the probability of impurity scattering. Thus, with increasing levels of doping the mobility goes down. This is especially obvious in compensation situations (Sec. 5.1) - mobility degradation imposes a practical limit to the number of compensations that can be done in one piece of semiconductor.

7.4. The Hot-Carrier Effect - Velocity Saturation (*Sec. 3.4.4*)

When the drift velocity approaches the thermal velocity the kinetic energy of the carriers is becoming very high. At some critical velocity, the electron saturation velocity v_{sn}, a new high-energy (or “optical”) phonon scattering process comes into action:

$$v_{sn} \approx \sqrt{\frac{h\ \omega_l}{4\pi\ m_n}},$$

where the scattering has a frequency of ω_l. This new scattering process is very effective in transferring energy from the hot electrons to the lattice and is the major reason why there is a saturation velocity. The name “hot” stems from the effective temperature T_e that is commonly associated with the carriers that have attained energies above the ambient thermal energy.

1. On page 98 in the main textbook.

So is this phenomenon of practical interest? Oh yes, velocity saturation is becoming noticeable at electric fields above a few thousand volts per centimeter, which is equivalent to a few hundred millivolts across a micrometer. In today's integrated circuits (ICs) the electric fields can be several volts across say 0.1-0.5 micrometer long semiconductors. In fact, low voltage IC techniques have been proposed so as to avoid, for example, velocity saturation problems.

7.5. Energy Levels under the Influence of an Electric Field (*Sec. 4.4.2*)

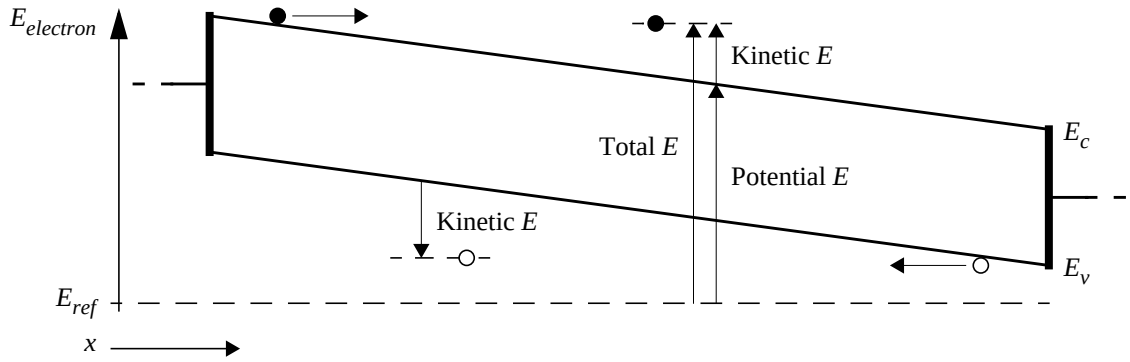


Fig. 7: Energy band bending under an external electric field.

Fig. 7 shows a semiconductor across which there is a voltage of V_A. Here it is assumed that 0 V is connected to the leftmost end and $+V_A$ V to the rightmost end. Now there is a gradient in the energy levels. Consequently electrons will move to the right, thereby minimizing their energy and conversely holes will travel to the left to minimize their energy.

Elementary physics tells us that the potential energy for an electron in an electrostatic potential V is $-qV$. Also, Fig. 7 shows that this energy also can be expressed as (E_c - E_{ref}). Thus,

$$V = -\frac{1}{q}(E_c - E_{ref}). \tag{23}$$

The electric field is, in one dimension,

$$E_x = -\frac{dV_x}{dx}. \tag{24}$$

Consequently, using Eq. (23) in Eq. (24), we have

$$E_x = \frac{1}{q}\frac{dE_c}{dx}. \tag{25}$$

Since E_c and E_v both have the same gradient[1] when experiencing the same electric field, these can be used interchangeably. Also, E_i can be used since this is a level **defined** and used as the intrinsic equilibrium Fermi level, even in non-equilibrium situations.

7.6. The Fermi Level - at Equilibrium and at Non-Equilibrium (*Secs 3.5, 4.3.3*)

At equilibrium no discontinuity or gradient in the Fermi level can arise in a semiconductor be it an intrinsic one or a non-uniformly doped one. In the previous section however, we went a

1. In heterogeneous semiconductors this may not be true.

step further in assuming a non-equilibrium case, namely a case where an external electrical field was imposed on the semiconductor. When an electric field is applied to a semiconductor, it is no longer at equilibrium.

We should be very careful with the concept of the Fermi level as this is only valid as long as the material is at equilibrium. Outside this specific domain so-called quasi-Fermi levels, F_n and F_p, have to be employed. The quasi-Fermi levels are defined in a framework similar to Eq. (12) and Eq. (13), such that the carrier concentrations can be written as

$$n = n_i \, e^{\frac{(F_n - E_i)}{kT}} \tag{26}$$

and

$$p = n_i \, e^{\frac{(E_i - F_p)}{kT}} . \tag{27}$$

Eq. (11), the law of mass action, can now be reformulated for non-equilibrium conditions as

$$np = n_i^2 \, e^{\frac{(F_n - F_p)}{kT}} . \tag{28}$$

Ex. 2: Thermal Noise in a Resistor

With a background in the drift discussion we can easily broaden our scope to noise issues. As an example, the fluctuations in electron transport due to the thermal velocity will now be considered:

One electron carries a current of

$$i_1 = \frac{q \, v_x}{L}$$

where L is the distance traversed.

Consequently the total current, carried by N electrons, is

$$i = \sum_{n=1}^{N} \frac{q \, v_{xn}}{L} .$$

The average current is zero, unless there is an applied bias across the resistor. Assume now that we apply a voltage across the resistor, then we will have a net velocity v_d that drives the current through the circuit, however $v_d \ll \langle |v_{xn}| \rangle$. The fluctuation in the current can be written as

$$\langle i^2 \rangle = i_t^2 = \frac{q^2 \langle v_{xn}^2 \rangle}{L^2} \cdot N .$$

How can this expression be simplified into known quantities?

First, from thermodynamics we know that a Maxwell-Boltzmann distributed gas can be described as

$$\frac{m <v_x^2>}{2} = \frac{kT}{2}$$

in one dimension. Thus,

$$<v_x^2> = \frac{kT}{m}. \tag{29}$$

Secondly, the total number of electrons simply is

$$N = n A L,$$

where n is the electron density and A is the cross-sectional area.

Now the current fluctuation, i.e. the thermal noise, can be written as

$$i_t^2 = \frac{q^2}{L^2} \cdot \left(\frac{kT}{m}\right) \cdot (n A L) = \frac{nq^2 kT}{m} \cdot \frac{A}{L}$$

and with the aid of the resistance expression derived earlier (e.g. in the main textbook, *Eq. (3-44)*)

$$R = \frac{L}{A} \cdot \frac{m}{nq^2 \tau_{cn}},$$

the geometrical ratio in the thermal noise can be replaced by a function of resistance such that

$$i_t^2 = \frac{nq^2 kT}{m} \cdot \frac{m}{nq^2 \tau_{cn}} \cdot \frac{1}{R} = \frac{kT}{R} \cdot \frac{1}{\tau_{cn}}.$$

The next derivation is a bit too simplified, but basically the noise can be written as a function of the bandwidth:

$$i_{t,\Delta f}^2 = \frac{i_t^2}{1/\tau_{cn}} \cdot \Delta f = \frac{kT}{R} \cdot \Delta f.$$

Both the noise current and voltage can, based on this, be formulated as

$$i_{t,\Delta f} = \sqrt{\frac{kT}{R} \cdot \Delta f}$$

and

$$v_{t,\Delta f} = \sqrt{kTR \cdot \Delta f}.$$

To be exact, these expressions in a formally correct derivation have to be compensated by a constant such that

$$i_{t,\Delta f} = \sqrt{\frac{4kT}{R} \cdot \Delta f}$$

and

$$v_{t,\Delta f} = \sqrt{4kTR \cdot \Delta f}\,.$$

Ex. 2(a): Thermal Noise in a Resistor - An Example

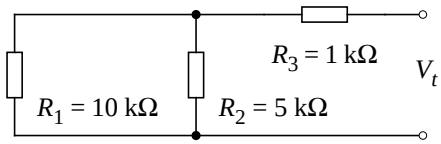


Fig. E.1: Resistor circuit with thermal noise voltage V_t.

Connecting three resistors in a topology as the one in Fig. E.1 leads to a thermal noise voltage over the terminals according to previous calculations. Let us find the thermal noise voltage per square root of bandwidth.

The current source due to noise from R_1 and R_2 can be defined as I_p. Then

$$I_p^{\,2} = \left(\sqrt{\frac{4kT}{R_1}}\right)^2 + \left(\sqrt{\frac{4kT}{R_2}}\right)^2$$

as we must sum the square values to find the noise. Together with the parallel resistance originating from R_1 and R_2, R_p, we have an equivalent circuit describing the resistor noise circuit, see Fig. E.2. In this figure all resistors are ideal and the noise contributions come from explicit

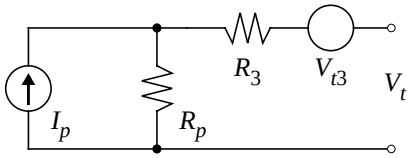


Fig. E.2: Equivalent circuit with ideal resistors and noise current sources and noise voltage generators.

sources.

It is now easy to write an expression for the total noise voltage, in square of course!

$$V_t^{\,2} = I_p^{\,2} R_p^{\,2} + V_{t3}^{\,2}\,.$$

Evaluating the variables leads to

$$V_t^{\,2} = \left(\frac{4kT}{R_1} + \frac{4kT}{R_2}\right)\left(\frac{R_1R_2}{R_1+R_2}\right)^2 + 4kTR_3 = \left(\frac{4kT(R_1+R_2)}{R_1R_2}\right)\left(\frac{R_1R_2}{R_1+R_2}\right)^2 + 4kTR_3$$

which in turn can be simplified into

$$V_t^{\,2} = 4kT \cdot \left(\frac{R_1R_2}{R_1+R_2} + R_3\right).$$

Thus

$$V_t = \sqrt{4kT \cdot R_{tot}}$$

where we notice that R_{tot} is the total resistance of the circuit in Fig. E.1. Using numbers now, R_{tot} becomes 13/3 kΩ and we consequently have

$$V_t = \sqrt{1.6\times10^{-20}\ \text{W/Hz} \cdot \frac{13}{3}\ \text{k}\Omega} = 8.33\times10^{-9}\ \text{V}/\sqrt{\text{Hz}}\,.$$

In this example we have assumed that we have the same T for all devices.

As a conclusion to this example, it is important to remember that the example in itself is not a central part of the course - i.e. it is important to remember not to remember. The example is given so as to show that by using the drift concept we can fairly easily find a model for thermal noise, which is an important phenomenon. Resistances are everywhere, think about all wires that are used, does anyone have zero resistance?

LECTURE 4

8. Diffusion (*Sec. 4.4.1*)

Through a cut of a non-uniformly doped semiconductor there will be a diffusion of carriers. In Fig. 8 there is a sketch showing a situation where the semiconductor has a non-zero doping gradient and where there has to exist diffusion of carriers that balances the difference in concentration. No electric field is present here. Assuming, as usual, electrons as carriers, the mean free path l_{cn} [1] has been marked in the sketch. l_{cn} is the distance covered by an electron in the mean free time, τ_{cn}. Thus, l_{cn} can be written as $l_{cn} = v_{th}\,\tau_{cn}$.

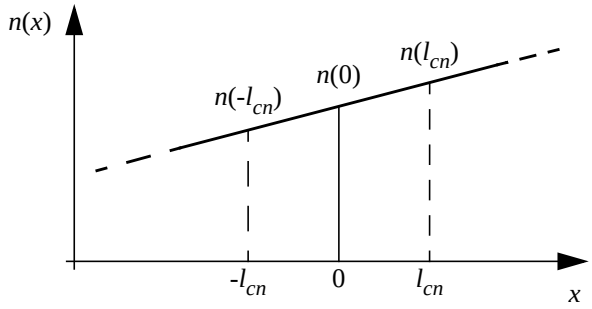


Fig. 8: Diffusion due to a doping gradient.

Based on the random motion of the electrons, we can calculate the flow of electrons through $x = 0$ from left to right, $F_{n,x}$. First there is the contribution from left directed to the right

$$\frac{1}{2}\,v_{th}\,n(-l_{cn})$$

and secondly we have the contribution from right directed to the left

$$-\frac{1}{2}\,v_{th}\,n(l_{cn})\,.$$

The flow then can be expressed as the sum of the two contributions

$$F_{n,x} = \frac{1}{2}\,v_{th}\,[n(-l_{cn}) - n(l_{cn})]\,.$$

Approximating the carrier densities at $x = \pm l_{cn}$ by the first two terms of the Taylor expansion yields

$$F_{n,x} = \frac{1}{2}\,v_{th}\left[\left(n(0) - \frac{dn}{dx}\,l_{cn}\right) - \left(n(0) + \frac{dn}{dx}\,l_{cn}\right)\right] = -v_{th}\,l_{cn}\,\frac{dn}{dx}.$$

As each electron carries a charge $-q$, the particle flow corresponds to a electron current density due to diffusion

$$J_{n,x} = -qF_{n,x} = qv_{th}\,l_{cn}\,\frac{dn}{dx}.$$

1. l_{cn} is on the order of some nanometers.

Now, since $l_{cn} = v_{th}\,\tau_{cn}$ and since the thermal velocity in one dimension can be found, using statistical mechanics, as

$$\frac{m_n^* v_{th}^{\,2}}{2} = \frac{kT}{2} \Rightarrow v_{th} = \sqrt{\frac{kT}{m_n^*}}.$$

we arrive at

$$J_{n,x} = q v_{th}^{\,2} \tau_{cn} \frac{dn}{dx} = q\left(\frac{kT}{m_n^*}\right)\tau_{cn}\frac{dn}{dx} = qD_n\frac{dn}{dx}, \tag{30}$$

where

$$D_n = kT\,\frac{\tau_{cn}}{m_n^*}$$

is the diffusion coefficient for electrons. In a similar way the current density and diffusion coefficient can be defined for holes:

$$J_{p,x} = -qD_p\frac{dp}{dx}, \tag{31}$$

where

$$D_p = kT\,\frac{\tau_{cp}}{m_p^*}$$

Finally, Einstein's relationship in Eq. (20) asserted that

$$\frac{D_n}{\mu_n} = \frac{kT}{q},$$

which then would lead to

$$\mu_n = \frac{q}{kT}\cdot D_n = \frac{q}{kT}\cdot\left(kT\,\frac{\tau_{cn}}{m_n^*}\right) = \frac{q\tau_{cn}}{m_n^*}.$$

This equation for electron mobility is identical to Eq. (22), which is an adequate conclusion to this section on diffusion.

Ex. 3: Drift and Diffusion - An Example

Assignment:

The electron concentration in a 1 μm long device varies linearly with distance from 10^{16} cm^{-3} to 1.1×10^{17} cm^{-3}. Plot the built-in electric field as a function of distance at T = 300 K.

Solution:

There is no current in the device as it is in equilibrium. Apparently the diffusion mechanism is compensated for by the drift mechanism, which is manifested by an electric field. We thus have

$$J_x = nq\mu_n E_x + qD_n \frac{dn}{dx} = 0,$$

from which we obtain

$$E_x = -\frac{D_n}{n\mu_n}\frac{dn}{dx} = -\frac{kT}{q}\frac{1}{n}\frac{dn}{dx}.$$

The electron density can be described by $n = 10^{16} + x\,(1.1\times10^{17} - 10^{16}) = 10^{16} + x\,10^{17}$, where x is given in μm. Now E_x can be evaluated as

$$E_x = -\frac{kT}{q}\frac{1}{10^{16} + x\,10^{17}}\,10^{17} = -0.0259\,\frac{1}{0.1 + x}\ \text{V/μm}$$

and the plot for this function is given in Fig. E.3.

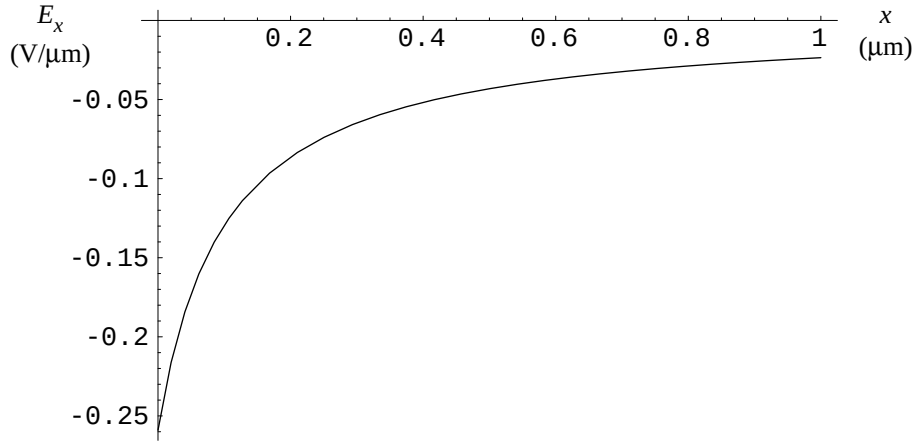


Fig. E.3: Plot of the built-in electric field as function of x.

9. Thermal Recombination-Generation (*Secs 4.3.1 - 4.3.2*)

Although direct transitions between valence and conduction band occur in all semiconductor materials, a complication of the crystal structures makes them unlikely in silicon and germanium except at very high carrier densities. The three-particle interaction it takes in indirect band-gap materials to make a direct transition is very improbable in comparison to two-particle transitions, such as those between a free carrier and a phonon. In this section we shall therefore investigate R-G via R-G center.

We will analyze the electron capture process closely, illustrated as process r_1 in Fig. 9: The rate at which electron capture occurs is proportional to the density of electrons n in the conduction band, the density of **empty** localized states, and the probability that an electron passes near a state and is captured by it.

The density of empty localized states is given by their total density N_t times one minus the probability $f(E_t)$ that they are occupied. Moreover, the probability per unit time that an electron is captured by an R-G state is given by the product of the electron thermal velocity, v_{th}, and a parameter, σ_n, called the capture cross section. This parameter describes the effectiveness of

the localized state in capturing an electron. The product $v_{th}\,\sigma_n$ may be visualized as the volume swept out per unit time by a particle with cross section σ_n. If the localized state lies within this volume, the electron is captured by it. As an example, for gold or iron σ_n is about 10^{-15} cm^2.

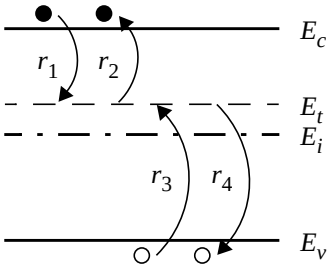


Fig. 9: Free carriers can interact with R-G centers by four processes: r_1 electron capture, r_2 electron emission, r_3 hole capture and r_4 hole emission. The R-G center is of acceptor type, i.e. neutral when empty and negative when full.

Now we can formulate r_1 as

$$r_1 = n[N_t(1 - f(E_t))]v_{th}\sigma_n .$$

The emission of an electron from the R-G center into the conduction band occurs at a rate given by the product of the density of states **occupied** by electrons times a probability e_n that the electron makes the jump:

$$r_2 = [N_t\ f(E_t)]e_n .$$

In the limiting case of equilibrium $r_1 = r_2$, which is a foundation for the law of mass action in Eq. (11)[1], we thus have

$$n_0[N_t(1 - f(E_t))]v_{th}\sigma_n = [N_t\ f(E_t)]e_n$$

and consequently, by using Eq. (5) for $f(E_t)$ and Eq. (12) for n_0, we get after some algebraic exercises

$$e_n = v_{th}\sigma_n\, n_i\, e^{\frac{(E_t - E_i)}{kT}} .$$

Obviously e_n increases as E_t approaches E_c which is in accordance with our intuition. With the knowledge of e_n, we can now determine r_2.

Also, in a similar manner r_3 and r_4 can be found:

$$r_3 = p[N_t\ f(E_t)]v_{th}\sigma_p$$

and

$$r_4 = [N_t(1 - f(E_t))]e_p ,$$

where the equilibrium condition leads to

1. This is valid only for non-degenerate semiconductor at equilibrium, as you remember.

$$e_p = v_{th}\sigma_p n_i e^{\frac{(E_i - E_t)}{kT}}.$$

NOTE that all four r_1 - r_4 can be used for non-equilibrium conditions and that gives us the capability to find the net rate of recombination minus generation.

Let us dwell here for a minute: At equilibrium we have $r_1 = r_2$ and $r_3 = r_4$, while a non-equilibrium material presents us with $r_1 \neq r_2$ and $r_3 \neq r_4$.

Now assume we have a case with an n-type semiconductor where suddenly holes are increased in number above their equilibrium value. This would cause r_3 to increase as there are more holes available. The effect of this increase would be to increase r_4 or r_1, both of which eliminate holes at E_t. If most of the holes disappear from E_t via r_1, they will remove electrons, and the R-G center will be an effective recombination center. If the holes are removed from the level at E_t predominantly by an increase in r_4, they will return to the valence band, and the site will be effective as a hole trap. A given R-G center will generally be effective in only one way; either as a trap or as a recombination center.

Around 1950 Shockley[1] and Read[2] on one hand and Hall[3] one the other derived the previous equations on recombination and generation through R-G centers. Therefore this process is frequently referred to as Shockley-Hall-Read (or SHR) recombination. According to this model, when non-equilibrium occurs in a semiconductor, the overall population of the R-G centers is **not** greatly affected. The reason for this is that recombination centers quickly capture majority carriers, as there are so many of these, but have to wait for minority carriers. Thus, these states are nearly always full of majority carriers whether under equilibrium or not.

Taking U as the recombination rate we have

$U = r_1 - r_2 = r_3 - r_4$.

Solving $U = r_1 - r_2$ for $f(E_t)$ we can insert the achieved solution for $f(E_t)$ in $U = r_3 - r_4$ and arrive at

$$U = \frac{np - n_i^2}{\tau_p\left(n + n_i e^{\frac{(E_t - E_i)}{kT}}\right) + \tau_n\left(p + n_i e^{\frac{(E_i - E_t)}{kT}}\right)}, \qquad (32)$$

where $\tau_p = (N_t v_{th}\sigma_p)^{-1}$ and $\tau_n = (N_t v_{th}\sigma_n)^{-1}$ are the excess carrier lifetimes. These lifetimes have come to be interpreted as the average time an excess minority carrier will live in a sea of majority carriers before recombining.

NOTE that U is the rate of **thermal** recombination minus the rate of **thermal** generation. If light would be shone on the semiconductor we would have to add a generation component due to the light.

1. William Shockley is a famous researcher in the semiconductor field. He will show up later again.
2. W. T. Read also invented the IMPATT diode.
3. This Hall, R. N. Hall, is not the same Hall that discovered the Hall effect in 1879, E. H. Hall.

9.1. R-G at Low Injection Levels (*Sec. 4.3.3*)

A condition where $\delta p \ll n_0$, $n \approx n_0$ [1] in an n-type material and $\delta n \ll p_0$, $p \approx p_0$ in a p-type material is called low-level injection. The perturbation in carrier numbers does not significantly affect the majority carrier number, whereas the number of minority carriers may, and routinely does, increase by many orders of magnitude.

Let us make some assertions: First, we know that $n_i^2 = n_0\, p_0$. Secondly, the maximum of U in Eq. (32) is attained when E_i is close to E_t - effective R-G centers is a property of importance in real devices. Based on these assertions we can write U as

$$U = \frac{np - n_0\, p_0}{\tau_p(n + n_i) + \tau_n(p + n_i)}. \tag{33}$$

Now, in the case of n-type material with low-level injection, i.e. $n \gg p$ and $n_0 \approx n$, the first term in the denominator of Eq. (33) dominates over the second. Moreover, we have $n \gg n_i$. Thus, we can simplify Eq. (33) into

$$U = \frac{np - np_0}{\tau_p(n + n_i)} = \frac{n \cdot \delta p}{n \cdot \tau_p} = \frac{\delta p}{\tau_p}.$$

With the same line of reasoning, for a p-type material we have

$$U = \frac{np - n_0\, p}{\tau_n(p + n_i)} = \frac{p \cdot \delta n}{p \cdot \tau_n} = \frac{\delta n}{\tau_n}. \tag{34}$$

Here it is obvious that the recombination rate depends upon the minority carriers. This is in accordance with the SHR model that states that recombination centers have to wait for minority carriers, which therefore become the rate-limiting step in the recombination process.

10. The Continuity Equation (*Sec. 4.4.3*)

Having reviewed the three cornerstones of carrier action, we can now sum them up in a continuity equation. A continuity equation can be written for both majority and minority carriers, but solutions of the minority-carrier continuity equation in semiconductors have special importance in many device applications since excess majority carriers are so few in comparison to the equilibrium majority carriers.

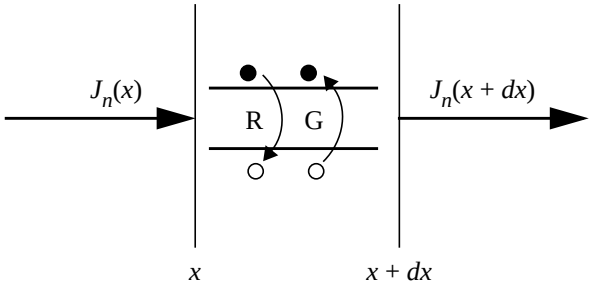


Fig. 10: The increase in the electron density in a p-type slice of thickness *dx* is related to the net flow of electrons into the slice and the excess of generation over recombination.

1. As in the main textbook, $\delta p = p - p_0$ and $\delta n = n - n_0$.

To derive a one-dimensional minority-carrier continuity equation for electrons, we consider the p-type slice of thickness *dx* located at *x* as shown in Fig. 10. The number of electrons inside the slice may increase because of a net flow into the volume and from net carrier generation inside the slice. The overall rate of electron increase equals the sum of the number of electrons flowing into the slice minus the number of the electrons flowing out, plus the rate at which electrons are generated minus the rate at which they recombine. Formally we thus have

$$\frac{\partial n}{\partial t}\ A\ dx\ =\ \left(\frac{J_n(x)-J_n(x+dx)}{-q}\right)A-U_n\ A\ dx\,. \tag{35}$$

Taking the Taylor expansion of the current density yields

$$J_n(x)-J_n(x+dx)\ =\ J_n(x)-\left(J_n(x)+\frac{\partial J_n}{\partial x}\cdot dx\right)\ =\ -\frac{\partial J_n}{\partial x}\cdot dx\,. \tag{36}$$

Eq. (34) and Eq. (36) are now substituted into Eq. (35)

$$\frac{\partial n}{\partial t}\ =\ \frac{1}{q}\frac{\partial J_n}{\partial x}-\frac{\delta n}{\tau_n}$$

and we are staring at the continuity equation for electrons. For holes the continuity equation would look like

$$\frac{\partial p}{\partial t}\ =\ -\frac{1}{q}\frac{\partial J_p}{\partial x}-\frac{\delta p}{\tau_p},$$

in which only the sign for the charge differs from the electron case.

In the continuity equation for holes, let us now substitute J_p for the total current density due to drift and diffusion, as the general case. Then we have

$$\frac{\partial p}{\partial t}\ =\ -\frac{1}{q}\frac{\partial}{\partial x}\left[p(x)\ q\mu_p\ E_x(x)-qD_p\frac{\partial p}{\partial x}\right]-\frac{\delta p}{\tau_p},$$

which in turn, fully evaluated, becomes

$$\frac{\partial p}{\partial t}\ =\ -\mu_p E_x\frac{\partial p}{\partial x}-p\mu_p\frac{\partial E_x}{\partial x}+D_p\frac{\partial^2 p}{\partial x^2}-\frac{\delta p}{\tau_p}.$$

Although μ_p and D_p are not fully independent of *x*, more elaborate versions of this equation are seldom considered.

There are a few directions we can take from this point; some important observations we can make from the continuity equation. To review two of these directions briefly we first take a look at the diffusion equation and then get to know the concept of diffusion length.

10.1. The Diffusion Equation (*Sec. 4.4.3*)

The diffusion equation is based on the continuity equation but assumes the special case of $E_x\ =\ 0$. Hence, for electrons it is written as

$$\frac{\partial n}{\partial t} = D_n \frac{\partial^2 n}{\partial x^2} - \frac{\delta n}{\tau_n}.$$

Further simplification can be achieved if we assume that the equilibrium carrier concentration is independent of time and position:

$$\frac{\partial}{\partial t}(\delta n) = D_n \frac{\partial^2}{\partial x^2}(\delta n) - \frac{\delta n}{\tau_n}.$$

This is a much more useful form of the equation, which will become apparent as we discuss the diode.

10.2. The Diffusion Length (*Sec. 4.4.4*)

Assuming a number of excess carriers are created in one end of a long uniformly doped bar, at $x = 0$. Many of these carriers will not only diffuse into the bar, but also recombine as they travel through the bar.

Assuming steady-state conditions, i.e. the number of excess carriers is constant over time, the diffusion equation for, say, electrons can be written as

$$0 = D_n \frac{d^2}{dx^2}(\delta n) - \frac{\delta n}{\tau_n}. \tag{37}$$

The boundaries of this differential equation are determined by the physical characteristic of the experiment with an "infinitely" long bar; $\delta n(0) = \delta n(0)|_{t=0}$ [1] and $\delta n(\infty) = 0$. The solution to the equation thus becomes

$$\delta n(x) = \delta n(0)|_{t=0} \cdot e^{-\frac{x}{\sqrt{D_n \tau_n}}},$$

where the diffusion length L_n is defined as

$$L_n = \sqrt{D_n \tau_n}.$$

The diffusion length is the average length an excess carrier diffuses before recombining and this quantity has great significance in the future sections on diodes and transistors.

1. In the main textbook there is another symbol for the initial excess carrier number: $\delta n(0)|_{t=0} = \Delta n$.

LECTURE 5

11. Gradients in the Quasi-Fermi Levels (*Sec. 4.4.6*)

We are not only prepared to summarize the carrier action into continuity equations, but we are now also mature enough to study the total current in another perspective. The total current density for electrons in the *x*-direction is

$$J_n = J_{n,\,drift} + J_{n,\,diff} = nq\mu_n E_x + qD_n\frac{dn}{dx}.$$

The gradient of the electron concentration *n* can be expressed as a function of the quasi-Fermi level for electrons, F_n, in Eq. (26), i.e.

$$\frac{dn}{dx} = \frac{d}{dx}\left(n_i\, e^{\frac{(F_n - E_i)}{kT}}\right) = \frac{n}{kT}\left(\frac{dF_n}{dx} - \frac{dE_i}{dx}\right).$$

From Eq. (18) we know that we can replace the gradient in the intrinsic Fermi level with qE_x which leads to a total electron current on the form

$$J_n = nq\mu_n E_x + qD_n \cdot \frac{n}{kT}\left(\frac{dF_n}{dx} - qE_x\right).$$

Einstein's relationship, Eq. (20), allows us to substitute the diffusion coefficient for a function of mobility. Then we have

$$J_n = nq\mu_n E_x + q\left(\mu_n \frac{kT}{q}\right) \cdot \frac{n}{kT}\left(\frac{dF_n}{dx} - qE_x\right) = nq\mu_n E_x + n\mu_n \frac{dF_n}{dx} - nq\mu_n E_x,$$

which can be simplified to

$$J_n = n\mu_n \frac{dF_n}{dx}.$$

The same derivation for holes yields

$$J_p = p\mu_p \frac{dF_p}{dx}.$$

In conclusion, if the gradient of any of the two quasi-Fermi levels is non-zero there is a net current flowing in the semiconductor. Oppositely, if there is a net current in a semiconductor at least one of the quasi-Fermi levels has to have a non-zero gradient.

Ex. 4: Quasi-Fermi Levels - An Example

Assignment:

In Eq. (26) and Eq. (27) quasi-Fermi levels were introduced for the description of carrier concentration outside equilibrium:

$$n = n_i\, e^{\frac{(F_n - E_i)}{kT}}$$

and

$$p = n_i \, e^{\frac{(E_i - F_p)}{kT}} .$$

Assume we have an n-type semiconductor at room temperature with $n_0 = 10^{15}$ cm^{-3}, $n_i = 10^{10}$ cm^{-3} and $p_0 = 10^5$ cm^{-3} (could it be Si?). Furthermore assume that at non-equilibrium we have $\delta n = \delta p = 10^{13}$ cm^{-3}. What are the quasi-Fermi levels in this semiconductor in relation to the intrinsic Fermi level?

Solution:

First, using Eq. (12) we find the Fermi level <u>at equilibrium</u> as

$$E_F - E_i \; = \; kT \; \ln\!\left(\frac{n_0}{n_i}\right) \; = \; 0.298 \text{ eV}$$

just to have one more reference energy level that can assist us in grasping the positions of the quasi-Fermi levels.

By having E_i as energy reference we can now take excess carriers into account in the non-equilibrium case, and quickly solve the quasi-Fermi levels by the use of the definitions in Eq. (26) and Eq. (27) so that

$$F_n - E_i \; = \; kT \; \ln\!\left(\frac{n}{n_i}\right) \; = \; kT \; \ln\!\left(\frac{n_0 + \delta n}{n_i}\right) \; = \; 0.298 \text{ eV}$$

and

$$F_p - E_i \; = \; -kT \; \ln\!\left(\frac{p}{n_i}\right) \; = \; -kT \; \ln\!\left(\frac{p_0 + \delta p}{n_i}\right) \; = \; -0.179 \text{ eV} .$$

Obviously, as the number of electrons (majority carriers) is not particularly affected by the addition of excess carriers F_n is not deflected much from E_F that is defined before we add the excess carriers. However, the number of holes is greatly affected by the addition of 10^{13} cm^{-3} excess carriers (there were only 10^5 cm^{-3} holes to start with!). Thus, F_p is situated 0.298 + 0.179 = 0.477 eV below E_F.

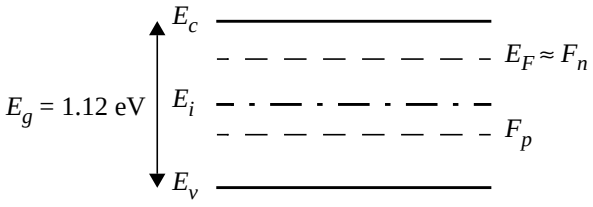


Fig. E.4: Band diagram indicating Fermi level as well as quasi-Fermi levels.

Take a look at Fig. E.4 to get a feeling for the location of all these levels, assuming the material really is Si. As a final comment, E_i in Si is located 0.013 eV below midgap. Thus the approxi-

mation that the intrinsic Fermi level is located at midgap is quite applicable if we consider the size of the bandgap.

12. Overview on Semiconductor Bulk Devices

The elementary semiconductor physics that we need for describing the semiconductor devices of this course has been treated. Let us stop for a while and make some observations:

Up until now the semiconductors have been homogeneous except from variations in doping concentration. Are such so-called bulk semiconductors of interest in practical situations? Well, not that often. But important applications **do** exist;

- Thermistors for measuring temperature -
  μ or n vary with temperature.
- Hall devices for measuring magnetic fields[1].
- Piezo-resistance devices that are sensitive to strain, compression or twist -
  these are probably among the most complicated devices to describe mathematically despite their simple physical structure. μ varies due to a perturbation in the $\bar{k}$-space (a result from the strain) which affects the effective mass through Eq. (3) and thus μ. Because of the involvement of the $\bar{k}$-space, the resistance depends on in which crystal direction the strain is applied.
- Photoconductive devices, so-called photoconductors[2], sensitive to light -
  the resistivity decreases with the increased electron-hole pair generation which comes about by the light absorption. **NOTE** the difference from photodiodes which always comprise a semiconductor junction.
- Gunn diodes[3].

Ex. 5: Gain in a Photoconductor - An Example

Assignment (*Problem 8.6* in the main textbook):

Assume that a photoconductor in the shape of a bar of length L and area A has a constant voltage V applied, and that it is illuminated such that G_{op} EHP/cm^3-s are generated uniformly throughout. If $\mu_n \gg \mu_p$, we can assume that the optically induced change in current ΔI is dominated by the mobility μ_n and the lifetime τ_n for electrons. Show that

$$\Delta I = qAL\, G_{op}\, \frac{\tau_n}{\tau_t},$$

for this photoconductor, where τ_t is the transit time of electrons drifting down the length of the bar.

1. In *Sec. 3.4.5* there is more information on the Hall effect, but this is optional and will not be part of any examination.
2. In *Sec. 4.3.4* photoconductors are discussed. This will be included in the examination.
3. The operation of the Gunn diode is based on peculiar mobility variations (Sec. 7.4) present in some materials. For optional reading, outside the scope of examination, consult *Sec. 10.3* of the main textbook.

Solution:

The average generation rate can be found be using Eq. (34) to describe the **thermal** recombination. This describes the spontaneous **thermal** recombination in a semiconductor **which contains excess electrons** outside the thermally excited carriers. There is also a thermal recombination going on of thermally generated carriers, but that is in fact simplified away in Eq. (34).

Now, assuming an illuminated junction, in which generation, G_{op}, of electron-hole pairs is taking place, we have at steady-state a balance between **thermal** recombination of **optically** generated excess electrons, such that

$$G_{op} = R = \frac{\delta n}{\tau_n},$$

or rather

$$\delta n = G_{op}\,\tau_n.$$

Similarly we have for holes

$$G_{op} = R = \frac{\delta p}{\tau_p},$$

or rather

$$\delta p = G_{op}\,\tau_p.$$

We can write the drift current in a photoconductor of length L and cross-sectional area A, across which there is a voltage V, as

$$I = qA\,(n\mu_n + p\mu_p)\,E_x = qA\,(n\mu_n + p\mu_p)\,\frac{V}{L}.$$

The change in current due to illumination can be described with regard to the change in carrier concentrations

$$\Delta I = qA\,(\delta n\mu_n + \delta p\mu_p)\,\frac{V}{L} = qA\,G_{op}\,(\tau_n\mu_n + \tau_p\mu_p)\,\frac{V}{L},$$

and with the assumption in the problem that $\mu_n \gg \mu_p$ ($\tau_n \approx \tau_p$) we can express this as

$$\Delta I = qA\,G_{op}\,\tau_n\mu_n\,\frac{V}{L}.$$

How can this be related to the transit time? Well, the transit time for an electron can be described as

$$\tau_t = \frac{L}{v_{drift}} = \frac{L}{\mu_n E_x} = \frac{L}{\mu_n}\,\frac{L}{V}$$

which gives

$$\mu_n = \frac{L^2}{\tau_t V}.$$

Thus, we have finally

$$\Delta I = qA\, G_{op}\, \tau_n \left(\frac{L^2}{\tau_t\, V} \right) \frac{V}{L} = qAL\, G_{op}\, \frac{\tau_n}{\tau_t},$$

which is the conclusion of *Problem 8.6*.

However, this is a very interesting expression in that it gives us the possibility to analyze the photoconductive gain Γ, which is defined as the ratio of flow of electrons per second from the device, i.e.

$$\frac{\Delta I}{q}$$

to the rate of generation of electron-hole pairs within the device, i.e.

$$AL\, G_{op}.$$

Thus, the gain of a photoconductor is

$$\Gamma = \frac{\tau_n}{\tau_t} \quad (1)$$

which for fairly high values on *V*, experiences a short enough transit time for the gain to become larger than unity. A surprise perhaps, considering that the photoconductor is nothing but a piece of doped semiconductor, without any p-n junction that usually is associated with amplifying devices.

In some materials such as CdS, trap levels exist. Whilst a carrier is held in a trap, a carrier of the opposite type must be present in the semiconductor to maintain space charge neutrality. This means; since the minority carrier lifetimes increase due to the traps, while the transit time remains constant, an even higher gain is achieved.

Now we will enter a new field of great importance: the theory of junctions between n- and p-type materials, which is the foundation for diodes as well as transistors.

13. The p-n Junction at Equilibrium (*Sec. 5.2*)

Prior to carrying out a thought experiment, where a p-n junction will be created, we have two semiconductors, one of p- and one of n-type, isolated from each other in room temperature, as shown in Fig. 11(a). At this temperature we can view the carriers supplied through doping as free carriers.

When the two pieces are put together, in a so-called step junction[2], diffusion of carriers will immediately start taking place, as depicted in Fig. 11(b). Electrons diffuse from the n-material to the p-material in the conduction band, while holes diffuse from the p-material to the n-material in the valence band. Uncompensated dopant ions are left behind on each side and the

1. Compare with the relation in Sec. 27.3 of the Lecture notes, that states that $\beta = I_C / I_B = \tau_n / \tau_t$.
2. The doping is uniform and constant in both the n- and the p-material, which is an important assumption in the step-junction model.

electric field that is starting to build up, because of these ions, will soon balance the diffusion.

Now balance between drift and diffusion has been established, which is illustrated in Fig. 11(c)[1]. The electric field E_x is forcing electrons to the left and holes to the right, which exactly balances the electron diffusion to the right and the hole diffusion to the left. We now have a p-n junction with a built-in potential which is called the contact potential, V_0. The region around the junction is called the depletion region and according to the depletion approximation it completely lacks free carriers (which is almost true).

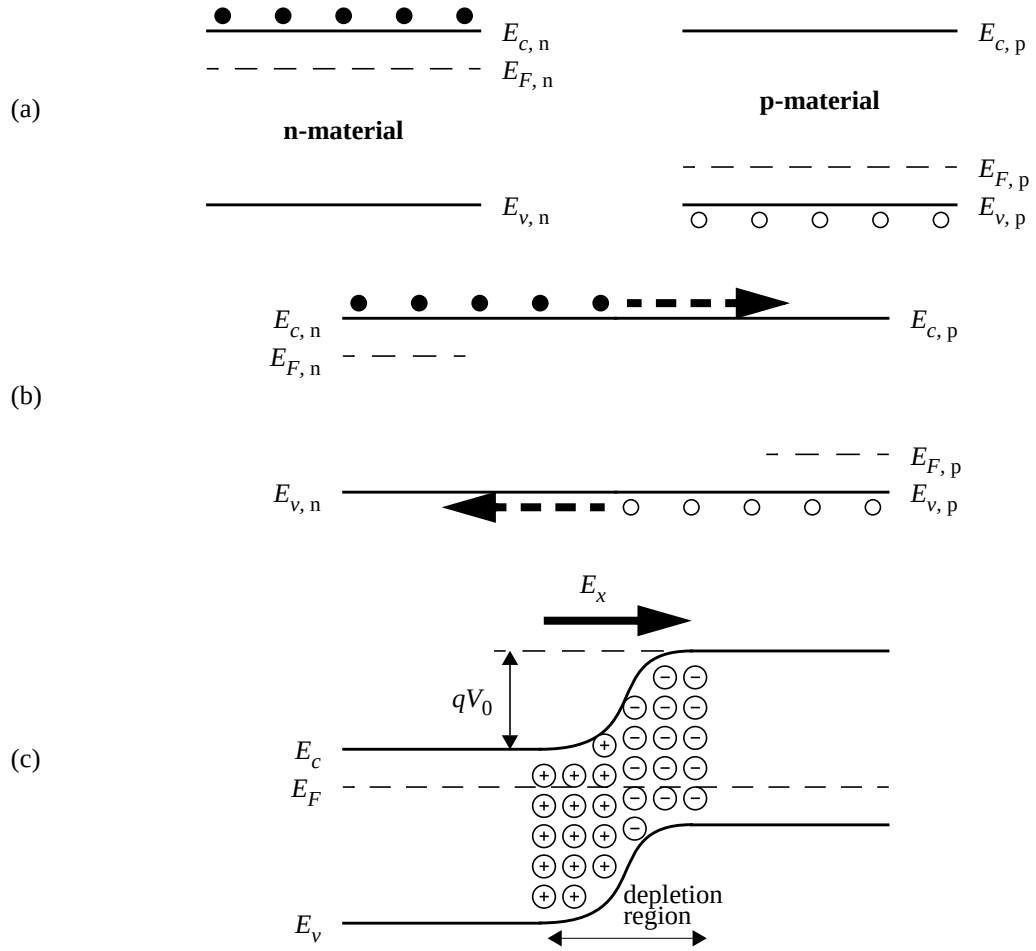


Fig. 11: Different phases of a thought experiment where a p-n junction is created.

13.1. The Contact Potential (*Sec. 5.2.1*)

What about the contact potential, V_0; with all our knowledge in basic semiconductor physics can we calculate its value? Looking at the p-n junction, the electron concentration in the p-type material is

$$n_{\mathrm{p}} = N_c\, e^{\frac{(E_{F,\,\mathrm{p}} - E_{c,\,\mathrm{p}})}{kT}}, \quad (2)$$

1. Here the liberty to indicate the presence of ions inside the energy band diagram has been taken - naturally there cannot exist ions in an energy band diagram, but only in sketches of the material itself.
2. Subscript indices n and p in plain text differ from italic *n* and *p* as they state type of material, not *carrier*.

whereas the electron concentration in the n-type material is

$$n_{\mathrm{n}} = N_c\, e^{\frac{(E_{F,\mathrm{n}} - E_{c,\mathrm{n}})}{kT}}.$$

But $qV_0 = E_{c,\mathrm{p}} - E_{c,\mathrm{n}}$ according to Fig. 11(c), and thus we can write

$$qV_0 = \left[E_{F,\mathrm{p}} - kT\,\ln\left(\frac{n_{\mathrm{p}}}{N_c}\right)\right] - \left[E_{F,\mathrm{n}} - kT\,\ln\left(\frac{n_{\mathrm{n}}}{N_c}\right)\right].$$

The junction is at equilibrium and then $E_{F,\mathrm{p}} = E_{F,\mathrm{n}}$ which cancel each other in the equation above. After simplification we hence get

$$V_0 = \frac{kT}{q}\,\ln\left(\frac{n_{\mathrm{n}}}{n_{\mathrm{p}}}\right). \tag{38}$$

In the n-material electrons are majority carriers and, thus, $n_{\mathrm{n}} = N_d$. But n_{p}, how do we go about to find that? In the p-material the holes are majority carriers and we will therefore know N_a which is equal to p_{p}. At equilibrium we can make use of the mass action's law in the p-material, $n_{\mathrm{p}}\, p_{\mathrm{p}} = n_i^2$, which consequently provides us with n_{p}. The final expression for the contact potential is

$$V_0 = \frac{kT}{q}\,\ln\left(\frac{N_a N_d}{n_i^2}\right).$$

Using $kT/q = 0.0259$ V and typical values for Si at room temperature; $N_a = 10^{18}$ cm^{-3}, $N_d = 10^{15}$ cm^{-3} and $n_i = 10^{10}$ cm^{-3}; we get $V_0 = 0.78$ V.

13.2. A First Encounter with the Law of Junction

Eq. (38) from the previous section can be written as

$$\frac{n_{\mathrm{n}}}{n_{\mathrm{p}}} = e^{\frac{qV_0}{kT}}.$$

This relationship is in fact similar to another expression, an expression we will derive right now from Eq. (28), which says:

$$np = n_i^2\, e^{\frac{(F_n - F_p)}{kT}}. \tag{39}$$

The meaning of Eq. (39) is that a displacement of the quasi-Fermi levels leads to an exponential growth in carriers. As one carrier type is majority carrier and does not have its density changed that much from equilibrium, the minority carrier type is the one that grows exponentially.

How can the quasi-Fermi levels be displaced? Well, simply by applying an external bias, V_A, on the p-n junction. As will be shown again in later sections (remember to take a look at Fig. 15), if the junction becomes biased we have

$$qV_A = F_n - F_p. \tag{40}$$

Substituting Eq. (40) into Eq. (39) yields

$$np = n_i^2 e^{\frac{qV_A}{kT}}, \tag{41}$$

which is often referred to as the law of junction.

LECTURE 6

Ex. 6: The Law of Junction - An IMPORTANT Example

Assignment:

A Si p-n junction is kept in an environment where $T = 300$ K and, thus, $n_i = 10^{10}$ cm^{-3}. Furthermore $N_d = 10^{16}$ cm^{-3}. Calculate the minority carrier hole concentration at the edge of the depletion region of the junction when a forward bias of $V_A = 0.60$ V is applied.

Solution:

We have the law of the junction in Eq. (41)

$$np = n_i^2 e^{\frac{qV_A}{kT}},$$

which can be rewritten as

$$p = \frac{n_i^2}{n} e^{\frac{qV_A}{kT}}.$$

Here we obviously are talking about the holes as minority carriers, i.e. we are talking about the n-side of the junction. Moreover, n is fairly constant as it represents the majority carriers, and not only is it constant but it is also equal to N_d - we assume low-level injection and complete ionization. We now can write

$$p_{\mathrm{n}} = p_{0,\mathrm{n}} e^{\frac{qV_A}{kT}}.$$

At equilibrium, in the n-type boundary to the depletion region we have

$$p_{0,\mathrm{n}} = \frac{n_i^2}{N_d} = \frac{(10^{10})^2}{10^{16}} = 10^4 \text{ cm}^{-3}.$$

Consequently,

$$p_{\mathrm{n}} = p_{0,\mathrm{n}} e^{\frac{qV_A}{kT}} = 10^4 e^{\left(\frac{0.60}{0.0259}\right)} = 1.15\times10^{14} \text{ cm}^{-3},$$

which indicates quite an increase in carrier number.

14. Analysis of the Depletion Region of the p-n Junction (*Sec. 5.2*)

In the subsections of Sec. 14 we will always assume that the semiconductors have a uniform cross-sectional area.

14.1. Review of Poisson's Equation

From the electrostatics we have to import an important equation, the Poisson equation, that states a relationship between the divergence of the electric field, $\nabla \cdot \bar{E}$, and the charge density, ρ:

$$\nabla \cdot \bar{E} = \frac{\rho}{\varepsilon_r \varepsilon_0}.$$

In the context of a junction with a depletion region, we prefer to use the space charge, i.e. charge times the difference of the number of positive charges and the number of negative charges inside a certain volume. Since all our analyses will take place in one-dimensional space, we can write Poisson's equation for our requirements as

$$\frac{dE_x}{dx} = \frac{q}{\varepsilon}(p - n + N_d - N_a).$$

Here space charge **density** has been used in accordance with the original Poisson equation.

14.2. The Electric Field in the Depletion Region (*Sec. 5.2.3*)

Using the depletion approximation means that there exist no free carriers inside the depletion region, while the charge density outside the depletion region is taken to be identically zero. Formally we thus have

$$\frac{dE_x}{dx} = \frac{q}{\varepsilon}(N_d - N_a), \text{ for } -x_n \le x \le x_p, \tag{42}$$

and

$$\frac{dE_x}{dx} = 0, \text{ for } x < -x_n \text{ and } x > x_p.$$

These equations are visualized for a step junction[1] in Fig. 12.

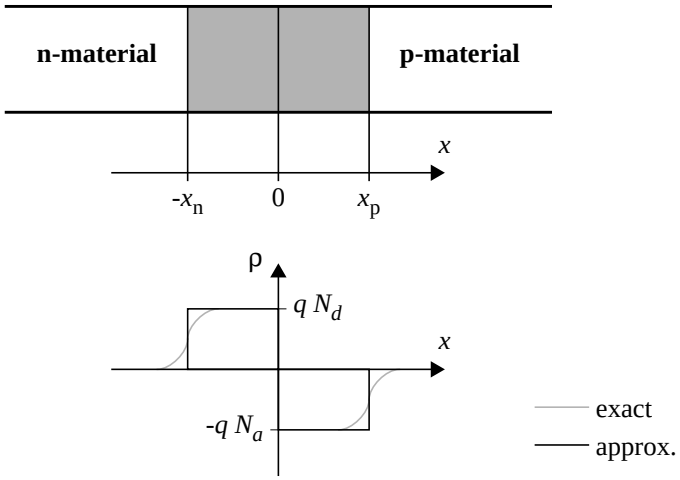


Fig. 12: Depletion region and charge density as functions of *x* in a step junction.

In the figures of these Lecture notes, the materials of the junctions are ordered such that the electric field increases with increasing *x*, in order provide another view than that of the main textbook.

Now, integrating both sides of Eq. (42) for $-x_n \le x \le 0$ and $0 \le x \le x_p$, respectively,

1. This is just one of many different junction types. Another common junction approximation is the graded junction in *Sec. 5.6.4*.

gives

$$\int_0^{E_x(x)} dE' = \int_{-x_n}^{x} \frac{qN_d}{\varepsilon}\, dx' \Rightarrow E_x(x) = \frac{qN_d}{\varepsilon}\,(x_n + x) \text{ for } -x_n \le x \le 0, \tag{43}$$

and

$$\int_{E_x(x)}^{0} dE' = \int_{x}^{x_p} -\frac{qN_a}{\varepsilon}\, dx' \Rightarrow E_x(x) = \frac{qN_a}{\varepsilon}\,(x_p - x) \text{ for } 0 \le x \le x_p. \tag{44}$$

Here the integral of E_x is only bounded by the conditions that $E_x(-x_n) = E_x(x_p) = 0$. In $x = 0$ the electric field has to be continuous and thus

$$E_x(0) = \frac{qN_d}{\varepsilon}\, x_n = \frac{qN_a}{\varepsilon}\, x_p \Rightarrow N_d x_n = N_a x_p \text{ [1]}. \tag{45}$$

This equation gives us an idea on how doping levels affect the depletion region. With heavy doping on one side and light doping on the other, obviously the depletion region on the side with heavy doping is quite narrow which is in contrast to the lightly-doped side where the depletion region penetrates much farther. This is important because most p-n junctions are asymmetrically doped, indicated as p^+-n or n^+-p where the + superscript means heavy doping[2].

14.3. The Potential in the Depletion Region and the Depletion Region Width (*Sec. 5.2.3*)

In the main textbook Streetman finds a clever way to formulate the total width of the depletion region. Let us instead make an instructive detour based on a calculation of the potential:

Solving the potential according to Eq. (24) yields

$$V_x = -\int E_x\, dx.$$

The boundaries of this equation are, by virtue of the definition of x and the ordering of n and p, simply $V_x(-x_n) = V_0$ and $V_x(x_p) = 0$. Applied to Eq. (43) and Eq. (44), this allows us to describe the potential anywhere in the depletion region as a function of the electric field. First, we have

$$\int_{V_0}^{V_x(x)} dV' = -\int_{-x_n}^{x} \frac{qN_d}{\varepsilon}\,(x_n + x')\, dx' \text{ for } -x_n \le x \le 0,$$

which leads to

$$V_x(x) - V_0 = -\frac{qN_d}{2\varepsilon}\,(x_n + x)^2 \text{ for } -x_n \le x \le 0.$$

Then we have

---

1. Yes, $E_x(0)$ is positive in these notes. This is due to the definition of x going from n to p.
2. To be called asymmetrical doping the heavily-doped side has to have a doping density many orders of magnitude larger than the lightly-doped side.

$$\int_{V_x(x)}^{0} dV' = -\int_{x}^{x_p} \frac{qN_a}{\varepsilon} (x_p - x')\, dx' \text{ for } 0 \le x \le x_p .$$

which after simplification of the integral yields

$$-V_x(x) = -\frac{qN_a}{(-2)\varepsilon} (x_p - x')^2 \Big|_x^{x_p} \Rightarrow V_x(x) = \frac{qN_a}{2\varepsilon} (x_p - x)^2 \text{ for } 0 \le x \le x_p .$$

The potential, just as much as the electric field, has to be continuous in $x = 0$. Thus,

$$V_x(0) = V_0 - \frac{qN_d}{2\varepsilon} x_n^2 = \frac{qN_a}{2\varepsilon} x_p^2 ,$$

which in conjunction with, for example, $x_n = \frac{N_a}{N_d} x_p$ derived from Eq. (45) leads to

$$V_0 - \frac{qN_d}{2\varepsilon} \left(\frac{N_a}{N_d} x_p\right)^2 = \frac{qN_a}{2\varepsilon} x_p^2 ,$$

which can be simplified into

$$x_p = \sqrt{\frac{2\varepsilon}{q} \frac{N_d}{N_a (N_a + N_d)} V_0} .$$

Now we can, thanks to Eq. (45), readily find x_n as

$$x_n = \frac{N_a}{N_d} x_p = \sqrt{\frac{2\varepsilon}{q} \frac{N_a}{N_d (N_a + N_d)} V_0} . \tag{46}$$

The total width of the depletion region W in equilibrium is $x_n + x_p$, or rather

$$W = \sqrt{\frac{2\varepsilon}{q} \frac{1}{N_a + N_d} V_0} \cdot \left(\sqrt{\frac{N_a}{N_d}} + \sqrt{\frac{N_d}{N_a}}\right) = \sqrt{\frac{2\varepsilon}{q} \frac{1}{N_a + N_d} V_0} \cdot \left(\frac{N_a + N_d}{\sqrt{N_a N_d}}\right),$$

which also can be written as

$$W = \sqrt{\frac{2\varepsilon}{q} \frac{N_a + N_d}{N_a N_d} V_0} . \tag{47}$$

What happens with the width in non-equilibrium? The total width of the depletion region depends on the amount of applied bias over the junction, V_A. Practically all applied bias voltage over a p-n diode, shown in Fig. 13, drops over the depletion region, since the contacts to the supply voltage wires are ohmic[1] in nature and the unperturbed n- and p-materials, the so-called quasi-neutral regions, conduct current well[2] - at least small currents. The depletion region on the other hand is highly resistive, lacking free carriers as it does. Eq. (47) can therefore be for-

1. See *Sec. 5.7.3*.
2. Yes, for current to flow, there has to be some voltage across the quasi-neutral regions.

mulated as

$$W = \sqrt{\frac{2\varepsilon}{q}\frac{N_a + N_d}{N_a N_d}(V_0 - V_A)}$$

to take applied bias into account. It is obvious that the higher potential V_A, the more narrow the depletion region becomes. Conversely, a lowering of V_A leads to a wider depletion region.

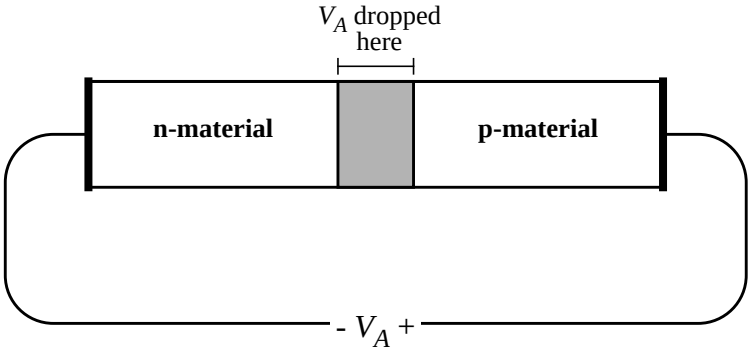


Fig. 13: Application of external voltage over a p-n diode, where +V_A connects to the p-material.

NOTE that this expression only is valid for $V_A \leq V_0$. However, when the expression for the width fails because of a large applied voltage, the current has become huge. The consequence of large currents through a diode is that the quasi-neutral regions will not act as such good conductors anymore, and thus we cannot neglect the voltage drops over these regions when V_A approaches V_0. Long before $V_A = V_0$ the diode is burning ...

Furthermore, under the assumption that we have a p^+-n junction diode, i.e. $N_a \gg N_d$, W can be simplified to

$$W = \sqrt{\frac{2\varepsilon}{q}\frac{1}{N_d}(V_0 - V_A)}, \tag{48}$$

since the large N_a cancels out. It should be **NOTE**d that the characteristics of the p-n junction are determined by the **lightly** doped side in an asymmetrically doped junction.

14.4. Depletion Region Capacitance in an Asymmetric p-n Junction (*Secs 5.5.4 - 5.5.5*)

The depletion region can act as the dielectric medium between two plates in a plate capacitor.

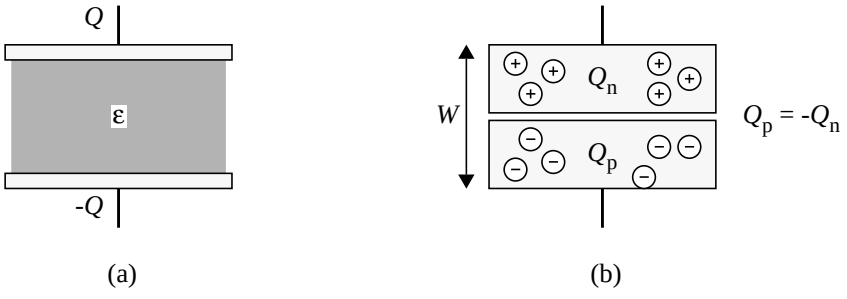


Fig. 14: (a) A conventional plate capacitor. (b) A “depletion region capacitor”.

However, a big difference from an ordinary plate capacitor is that the charged ions are distribution within the dielectric medium in the "depletion region capacitor", as shown in Fig. 14. Despite this, it can (and will soon) be shown that the capacitance in the p-n junction can be written as that of a plate capacitor.

Usually we define

$$C = \frac{Q}{V_A},$$

happily unaware of that the foundation for this expression is that the charge Q is varying linearly with the voltage V_A. Let us continue with the assumption that we have a p^+-n junction and therefore only concern ourselves with the n-side. Then, in the n-side of the depletion region we can define the space charge[1] from Sec. 14.1 as

$$Q_n = \rho \cdot Volume = qN_d \cdot A\, x_n.$$

With Eq. (46) adjusted to an applied bias and $N_a \gg N_d$, the space charge is

$$Q_n = qN_d \cdot A\sqrt{\frac{2\varepsilon}{q}\frac{1}{N_d}(V_0 - V_A)} = A\sqrt{2\varepsilon q\, N_d\,(V_0 - V_A)}.$$

Clearly the charge is not linearly dependent on the voltage and thus the capacitance has to be expressed as

$$C = \left|\frac{dQ_n}{d(V_0 - V_A)}\right|,$$

where the voltage difference is the voltage applied over the depletion area. Finally,

$$C = A\frac{1}{2}\sqrt{\frac{2\varepsilon q\, N_d}{(V_0 - V_A)}} = \varepsilon\frac{A}{\sqrt{\frac{2\varepsilon}{q}\frac{1}{N_d}(V_0 - V_A)}} = \varepsilon\frac{A}{W}$$

is achieved, an expression for a plate capacitor.

A p-n diode used as a voltage-dependent capacitor is called a varactor. Usually, it is used with $V_A < 0$ V, which leads to negligible d-c currents in the diode which is a behavior similar to a conventional capacitor. How currents are created in the p-n junction will be unveiled in Sec. 15.

14.5. Critical Field in a p^+-n Junction

Previously we derived an equation for the electric field in the depletion region. The maximum (absolute) value of the electric field was in $x = 0$, exactly in between the two types of materials. Let us now assume that we have a p^+-n junction; then we can write

$$E_x(0) = E_{x,max} = \frac{qN_d}{\varepsilon}x_n.$$

1. There is of course a space charge Q_p associated with the p-side, $Q_p = -Q_n$.

We can also review the expression for the width of the depletion region of a p^+-n junction which was derived shortly after the electric field was analyzed:

$$W = \sqrt{\frac{2\varepsilon}{q}\frac{1}{N_d}(V_0 - V_A)}.$$

However, in an asymmetrical p^+-n junction $W = x_n$. Thus, we can merge the two previous equations into

$$E_{x,max} = \frac{qN_d}{\varepsilon}\sqrt{\frac{2\varepsilon}{q}\frac{1}{N_d}(V_0 - V_A)}.$$

Let us now consider the maximal electric field a junction can take before experiencing a so-called breakdown, $E_{critical}$. Knowing the critical field allows for calculating the applied voltage it takes in the reverse direction, V_{BR}, to have a breakdown:

$$E_{critical} = \sqrt{\frac{2q}{\varepsilon}N_d(V_0 + V_{BR})} \Rightarrow V_{BR} = \frac{\varepsilon}{2q}\frac{1}{N_d}E_{critical}^2 - V_0.$$

We can notice that if we have a certain critical field; an increase in doping would give a decrease in breakdown voltage. Read *Sec. 5.4.2* in the main textbook on avalanche breakdown.

15. Currents in the p-n Junction (*Sec. 5.3*)

When a p-n diode is forward biased, as shown in Fig. 15, the difference between energy bands

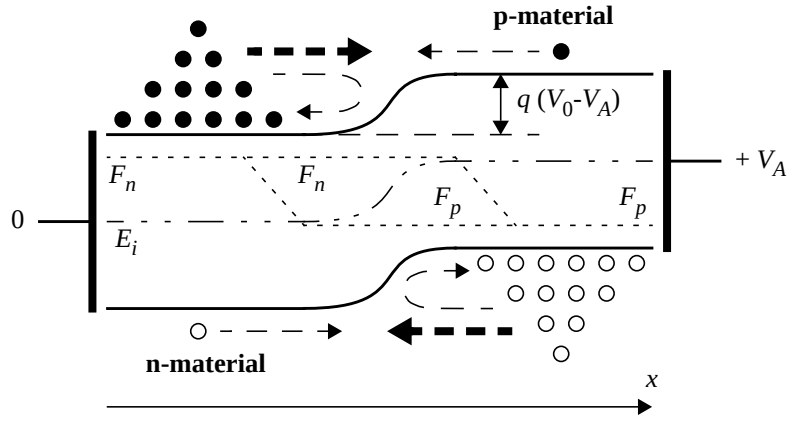


Fig. 15: A forward-biased p-n junction. (Forward always means that the highest potential is at the p-side.) The quasi-Fermi levels and the intrinsic Fermi level are indicated.

across the depletion region is reduced. The electric field is also reduced, but **NOTE**; **it will not change orientation**. Thus, the drift current existing in the p-n junction will take place in the opposite direction to what we might expect as people quite competent in basic electronics - the current in a forward-biased diode is supposed to go from the p-side to the n-side. No, let us leave the drift current for a while, it is all the same restricted in size since there are few minority carriers on each side of the depletion region and these carriers are the only that can be pushed by the electric field. The current in the forward-biased junction is in fact due to diffusion.

Illustrated in Fig. 15 is the decreasing number of majority carriers with increasing energies. There are quite few carriers having kinetic energy enough to surmount the potential difference but those, which have that energy, have the ability to diffuse into the depletion region.

At equilibrium the diffusion and the drift currents are exactly balancing each other, but when we apply an external forward bias voltage, we facilitate the diffusion by lowering the potential barrier thereby allowing for majority carriers at lower energies to diffuse into the depletion region.

15.1. Deriving a Current Equation for the p-n Junction (*Sec. 5.3*)

Eq. (41) provided the law of junction, which relates carrier densities to a potential as

$$np = n_i^2 e^{\frac{qV_A}{kT}},$$

under the assumption that **no recombination or generation take place** in the region over which the potential is applied and that low-level injection[1] prevails. We can thus write the carrier densities at the edges of the depletion region as

$$n_{\mathrm{p}}(x_{\mathrm{p}}) = n_{0,\,\mathrm{p}}(x_{\mathrm{p}})\, e^{\frac{qV_A}{kT}}$$

and

$$p_{\mathrm{n}}(-x_{\mathrm{n}}) = p_{0,\,\mathrm{n}}(-x_{\mathrm{n}})\, e^{\frac{qV_A}{kT}}$$

The excess carrier densities are therefore simply $\delta n = n - n_0$ and $\delta p = p - p_0$ or

$$\delta n_{\mathrm{p}}(x_{\mathrm{p}}) = n_{0,\,\mathrm{p}}(x_{\mathrm{p}}) \left(e^{\frac{qV_A}{kT}} - 1 \right)$$

and

$$\delta p_{\mathrm{n}}(-x_{\mathrm{n}}) = p_{0,\,\mathrm{n}}(-x_{\mathrm{n}}) \left(e^{\frac{qV_A}{kT}} - 1 \right).$$

Now, the excess densities, assuming no R-G processes in the depletion region, are defined at the boundaries of the depletion region. Taking, for example, electrons going from n to p: Out of $N_d(-x_{\mathrm{n}})$ majority carriers some electrons diffuse over the depletion region and become $\delta n_{\mathrm{p}}(x_{\mathrm{p}})$ minority carriers. Some of these minority carriers continue to diffuse deeper into the p-material, of course suffering from recombination as there are many holes present in this material (as opposed to the depletion region). Now we will describe the diffusion process in the quasi-neutral regions.

In the section on diffusion length, Sec. 10.2, we solved the diffusion equation for electrons at steady-state:

$$0 = D_n \frac{d^2}{dx^2}(\delta n) - \frac{\delta n}{\tau_n}$$

for an “infinitely” long bar. Now we can use the same equation for the quasi-neutral regions in

1. See Sec. 9.1.

the p-n junction diode, to find out how recombination affects the excess carrier concentrations, δn_{p} and δp_{n}.

Continuing to assume the electron case; the boundaries of this differential equation are $\delta n_{\mathrm{p}}(x_{\mathrm{p}})$, right on the boundary between the depletion region and the p-side, and 0 at the contact area of the p-side. We thus assume that the p-type quasi-neutral region is long compared to the diffusion length $L_n = \sqrt{D_n \tau_n}$, which means that all electrons recombine before reaching the p-side contact - this often is called the long-base diode assumption. The solution to the equation thus becomes

$$\delta n_{\mathrm{p}}(x) = \delta n_{\mathrm{p}}(x_{\mathrm{p}})\, e^{-\frac{(x-x_{\mathrm{p}})}{L_n}} = n_{0,\,\mathrm{p}}(x_{\mathrm{p}}) \left(e^{\frac{qV_A}{kT}} - 1 \right) e^{-\frac{(x-x_{\mathrm{p}})}{L_n}}.$$

We know from before that the current density due to diffusion depends on the carrier concentration gradient. Also from before, we know that in the step-junction model the doping in both materials is uniform and constant. Thus, the gradient of the carrier concentration is identical to the gradient of the excess carrier concentration, which we just described. Therefore, Eq. (30) can be applied here, yielding

$$J_{n,\,x} = qD_n \frac{dn}{dx} = qD_n \frac{d}{dx}\left[n_{0,\,\mathrm{p}}(x_{\mathrm{p}}) \left(e^{\frac{qV_A}{kT}} - 1 \right) e^{-\frac{(x-x_{\mathrm{p}})}{L_n}} \right]$$

which in evaluated form is written

$$J_{n,\,x} = -qD_n \frac{n_{0,\,\mathrm{p}}(x_{\mathrm{p}})}{L_n} \left(e^{\frac{qV_A}{kT}} - 1 \right) e^{-\frac{(x-x_{\mathrm{p}})}{L_n}}.$$

So what does this imply? The current decreases as we go farther away from the depletion region. But the total current must remain constant with distance from the junction in the steady-state case! The explanation to this “mystery” is that the hole current due to majority carriers increases as we move away from the junction. This hole current supplies the holes with which the minority carrier electrons recombine.

With an analogous treatment we find the solution to the hole diffusion equation in the n-side as

$$\delta p_{\mathrm{n}}(x) = \delta p_{\mathrm{n}}(-x_{\mathrm{n}})\, e^{\frac{(x+x_{\mathrm{n}})}{L_p}} = p_{0,\,\mathrm{n}}(-x_{\mathrm{n}}) \left(e^{\frac{qV_A}{kT}} - 1 \right) e^{\frac{(x+x_{\mathrm{n}})}{L_p}},$$

which allows us to define the current density in Eq. (31) as

$$J_{p,\,x} = -qD_p \frac{dp}{dx} = -qD_p \frac{p_{0,\,\mathrm{n}}(-x_{\mathrm{n}})}{L_p} \left(e^{\frac{qV_A}{kT}} - 1 \right) e^{\frac{(x+x_{\mathrm{n}})}{L_p}}.$$

To obtain an expression for the total current, $J_{t,\,x}$, we sum the minority-carrier components $J_{p,\,x}$ and $J_{n,\,x}$ at $-x_{\mathrm{n}}$ and x_{p}, respectively, as

$$J_{t,\,x} = -q\left[\frac{D_p}{L_p}\,p_{0,\,\mathrm{n}}(-x_\mathrm{n}) + \frac{D_n}{L_n}\,n_{0,\,\mathrm{p}}(x_\mathrm{p})\right]\left(e^{\frac{qV_A}{kT}} - 1\right). \tag{49}$$

Obviously the current moves opposite to the *x*-direction, which is reasonable considering what is illustrated in Fig. 12 and Fig. 15. Usually this expression is simplified into the ideal diode equation

$$J_{t,\,x} = -J_0\left(e^{\frac{qV_A}{kT}} - 1\right), \tag{50}$$

where J_0 is the magnitude of the reverse saturation current density predicted by this theory when a negative bias equal to a few kT/q is applied. This seems to indicate that when the diode is reverse biased the current - in electronics often referred to as the leakage current - is constant and is flowing in the *x*-direction, as defined in the diode. The reverse-biased scenario is depicted in Fig. 16, where the drift mechanism dominates over the diffusion. But even though the electric field increases significantly by the reverse bias applied, the drift current does not increase particularly much. This is due to the few minority carriers that are available for drift at each side of the depletion region. These few minority carriers come from the quasi-neutral material itself and they are due to thermal generation. After being generated the minority carriers diffuse around, some of them ending up at the boundary to the depletion region where they are attracted by the electric field.

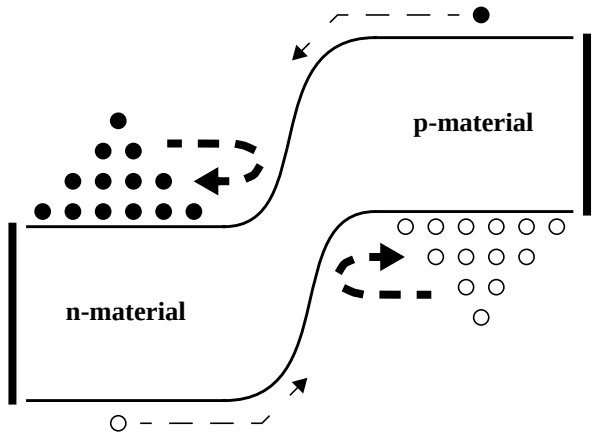


Fig. 16: A reverse-biased p-n junction.

LECTURE 7

Ex. 7: Thermal Response in a p-n Junction - An Example

Assignment:

Consider a Si p-n junction initially biased at 0.60 V at T = 300 K. Assume the temperature increases to T = 310 K. Calculate the change in the forward-bias voltage required to maintain a constant current through the junction.

Solution:

We have the following expression for the diode current density:

$$J_{t,\,x} = -q\left[\frac{D_p}{L_p}\,p_{0,\,\mathrm{n}}(-x_\mathrm{n}) + \frac{D_n}{L_n}\,n_{0,\,\mathrm{p}}(x_\mathrm{p})\right]\left(e^{\frac{qV_A}{kT}} - 1\right).$$

This can be slightly redefined if we consider that

$$p_{0,\,\mathrm{n}}(-x_\mathrm{n}) = \frac{n_i^{\,2}}{N_d}.$$

and

$$n_{0,\,\mathrm{p}}(x_\mathrm{p}) = \frac{n_i^{\,2}}{N_a}.$$

Then

$$J_{t,\,x} = -qn_i^{\,2}\left(\frac{D_p}{L_p}\,\frac{1}{N_d} + \frac{D_n}{L_n}\,\frac{1}{N_a}\right)\left(e^{\frac{qV_A}{kT}} - 1\right),$$

which, with forward-bias conditions, has a temperature dependence such as

$$J_{t,\,x} \propto n_i^{\,2}(T)\,e^{\frac{qV_A}{kT}},$$

if the dependence of the diffusion coefficients on temperature is neglected ("only" $\propto T$).

When we derived the mass action law we saw that

$$n_i^{\,2} = N_c N_v\,e^{\frac{-E_g}{kT}}$$

and thus, assuming the temperature dependence of N_c and N_v is not very big in comparison to the exponential function (it is "only" $\propto T^{3/2}$), we have the following relation:

$$J_{t,\,x} \propto e^{\frac{-E_g}{kT}}\,e^{\frac{qV_A}{kT}}.$$

With a constant current, despite the temperature change from T_1 = 300 K to T_2 = 310 K, we need to have a change from V_{A1} (= 0.60 V) to V_{A2} in applied bias according to:

$$e^{\frac{-E_g}{kT_1}} e^{\frac{qV_{A1}}{kT_1}} = e^{\frac{-E_g}{kT_2}} e^{\frac{qV_{A2}}{kT_2}} \Rightarrow \frac{-E_g}{kT_1} + \frac{qV_{A1}}{kT_1} = \frac{-E_g}{kT_2} + \frac{qV_{A2}}{kT_2}.$$

Thus, since E_g = 1.12 eV, we can solve for V_{A2}

$$V_{A2} = \left[\frac{T_2}{T_1}(qV_{A1} - E_g) + E_g\right]/q = 0.58 \text{ V}.$$

The change in applied bias has to be -20 mV to keep the current constant.

16. Recombination and Generation in the Depletion Region (*Sec. 5.6.2*)

The ideal diode equation is completely ideal, the derivation of Sec. 15 assumed that the diffusion across the depletion region was not disturbed by recombination or generation; a simplification far from reality. To compensate for the R-G processes taking place in the depletion region the ideal diode equation usually is modified by a constant such that the expression (with our notion of *x*-axis) looks like

$$J_x = -J_0\left(e^{\frac{qV_A}{\eta kT}} - 1\right).$$

The ideality factor, η, takes R-G processes into account in a way that is difficult to understand without further discussion. In short, when η is close to 1 diffusion dominates, while a η close to 2 indicates that recombination is dominating.

In Sec. 9 we did a thorough analysis of recombination and generation via traps. $U = R - G$ can be written as

$$U = \frac{np - n_i^2}{\tau_p(n + n_i) + \tau_n(p + n_i)}, \tag{51}$$

based on Eq. (32) and $E_t = E_i$ as in Sec. 9.1. We will now use this expression to analyze what really happens in the depletion region.

16.1. R-G Processes in the Depletion Region under Reverse Bias

For the case of reverse bias, U in the depletion region, as defined in Eq. (51), can be further simplified and evaluated by noting that $n \cong 0$ and $p \cong 0$ since all free carriers are swept out of the depletion region:

$$U = \frac{-n_i}{\tau_p + \tau_n},$$

where the sign indicates that in fact a generation process is going on. When generated, an electron and a hole will be swept with the electric field so that the generation current will go in the same direction as the reverse saturation current.

Now integrating over the depletion width W under the assumption that the generation is constant throughout the whole region yields

$$J_{gen} = q\int_0^W G\,dx = \frac{qn_iW}{\tau_p + \tau_n}.$$

Now, this is not completely uninteresting. This expression reveals that if a real p-n diode is reverse biased the total reverse current density is

$$J_{reverse} = J_0 + J_{gen} = J_0 + W\frac{qn_i}{\tau_p + \tau_n},$$

i.e. the total reverse current density is not constant, because the width *W* depends on the applied bias V_A.

In fact the expression for the reverse current density illustrates a number of things that are really important for modern electronics. Take for example the dependence on intrinsic carrier concentration,

$$J_{reverse} = J_0(n_i^2) + J_{gen}(n_i).$$

This shows that for small intrinsic carrier concentrations, the generation component becomes more important. Where do we encounter small intrinsic carrier concentrations? Well, in large bandgap materials, and at lower temperatures. Can this be of interest in other cases? - check out Example 8.

Furthermore, since we now have accounted for the two major mechanisms governing leakage current in semiconductors, we now have some insight into why large bandgap materials such as SiC would be nice to use as process technology for a dynamic memory. Don't we?

Finally, this expression **is** however uninteresting in so far as it does not explain the reason for including the ideality factor in the modified diode equation. No, the ideality factor is something that is due to the R-G processes under forward bias.

16.2. R-G Processes in the Depletion Region under Forward Bias

When the junction is forward biased, the carrier concentrations in Eq. (51) are not zero as in the case of reverse bias. There exist some free carriers inside the depletion region and these can recombine! Using the concept of quasi-Fermi levels, Eq. (26) and Eq. (27), to describe *U* we arrive at

$$U = \frac{n_i\, e^{\frac{(F_n - E_i)}{kT}} \cdot n_i\, e^{\frac{(E_i - F_p)}{kT}} - n_i^2}{\tau_p\left(n_i\, e^{\frac{(F_n - E_i)}{kT}} + n_i\right) + \tau_n\left(n_i\, e^{\frac{(E_i - F_p)}{kT}} + n_i\right)}\ (1).$$

The reader should by now be able to deduce $F_n - F_p = qV_A$ (2). Also, at point $x = 0$ at the "center" of the depletion region we have

1. Here it becomes apparent why asking students to mechanically memorize all formulae of semiconductor technology is completely futile. What really is interesting is what the qualitative discussion around this formula gives, not the actual values attained when used in a calculation.
2. Consult Fig. 15 if this deduction does not seem possible.

$$F_n - E_i = E_i - F_p = \frac{qV_A}{2}$$

and thus U_{max} can be written as

$$U_{max} = \frac{n_i\left(e^{\frac{qV_A}{kT}} - 1\right)}{\tau_p\left(e^{\frac{qV_A}{2kT}} + 1\right) + \tau_n\left(e^{\frac{qV_A}{2kT}} + 1\right)}.$$

Now two reasonable simplifications are made: First, the excess carrier lifetimes are replaced by an average lifetimes τ_0 and, secondly, the forward bias is supposed to be large enough to make the exponential terms dominate over unity. So now the expression becomes

$$U_{max} = \frac{n_i\, e^{\frac{qV_A}{kT}}}{2\tau_0\, e^{\frac{qV_A}{2kT}}} = \frac{n_i}{2\tau_0}\, e^{\frac{qV_A}{2kT}},$$

which forms the basis for an expression for the recombination current density at forward bias:

$$J_{rec} = \frac{qn_iW}{2\tau_0}\, e^{\frac{qV_A}{2kT}}.$$

Thus, the total forward current density, as defined in previous sections, can be written as

$$J = -J_0\left(e^{\frac{qV_A}{kT}} - 1\right) - \frac{qn_iW}{2\tau_0}\, e^{\frac{qV_A}{2kT}}.$$

NOTE that this is an approximation since J_{rec} is based on a constant U (U_{max} is only valid at $x = 0$) and on the fact that our expression for U only consider one single R-G center[1]. In any case, it serves well as an indicator on why there is an ideality factor in the modified diode equation.

Ex. 8: Generation Current - An Example

Assignment:

A reverse-biased p-n junction has an ideal reverse saturation current, I_0, according to *Eq. (5-36)* in the main textbook. Consider a p-n junction in Si at room temperature with the following specific data: $N_a = N_d = 10^{16}$ cm^{-3} and $\tau_p = \tau_n = 5\times10^{-7}$ s. There are three things we should investigate:

1. What is the ideal reverse saturation current density, J_0?

2. What is the generation current density, J_{gen}, when the p-n junction is reverse biased with 5 V?

1. For materials with more than one R-G center, the factor 2 in the denominator (prior to *kT*) would decrease slightly.

3. What can be said about the relative magnitudes of the two current densities?

Solution 1:

The ideal reverse saturation current density can be described as

$$J_0 = qn_i^2\left(\frac{D_p}{L_p}\frac{1}{N_d}+\frac{D_n}{L_n}\frac{1}{N_a}\right).$$

The diffusion length can be substituted $L = \sqrt{D\cdot\tau}$, leading to

$$J_0 = qn_i^2\left(\sqrt{\frac{D_p}{\tau_p}}\frac{1}{N_d}+\sqrt{\frac{D_n}{\tau_n}}\frac{1}{N_a}\right).$$

Since $N_a = N_d$ and $\tau_p = \tau_n$, we can rewrite the equation based on N (= $N_a = N_d$) and τ (= $\tau_p = \tau_n$)

$$J_0 = \frac{qn_i^2}{\sqrt{\tau}}\cdot\frac{1}{N}\left(\sqrt{D_p}+\sqrt{D_n}\right),$$

where the diffusion constants are obtained from *Fig. 3-23* (μ_n = 1000 cm^2/Vs, μ_p = 400 cm^2/Vs) being simply the mobility corresponding to actual carrier concentration followed by a compensation with Einstein's relationship such that

$$D = \frac{kT}{q}\cdot\mu.$$

The result is

$$J_0 = \frac{1.6\times10^{-19}\left(10^{10}\right)^2}{\sqrt{5\times10^{-7}}}\cdot\frac{1}{10^{16}}\left(\sqrt{10.4}+\sqrt{25.9}\right) = 1.87\times10^{-11}\ \text{A/cm}^2$$

Solution 2:

From Sec. 16.1 we know that

$$J_{gen} = \frac{qn_iW}{\tau_p+\tau_n}.$$

To describe this current we must know the depletion region width, *W*. From Sec. 14.4 we have

$$W = \sqrt{\frac{2\varepsilon}{q}\frac{N_a+N_d}{N_aN_d}\left(V_0-V_A\right)},$$

and here all parameters except V_0 are known. Way back in the course we encountered an important expression:

$$V_0 = \frac{kT}{q}\ln\left(\frac{N_aN_d}{n_i^2}\right),$$

which yields a contact potential of 0.69 V, with our values. As far as the expression of the width is concerned V_0 - V_A then becomes 5.69 V and

$$W = 1.22\times10^{-4}\ \text{cm}.$$

Finally we can calculate J_{gen}, and to our surprise (?) it is

$$J_{gen} = \frac{1.6\times10^{-19}\ 10^{10}\ 1.22\times10^{-4}}{10^{-6}} = 1.95\times10^{-7}\ \text{A/cm}^2.$$

Conclusion 3:

The generation current in the reverse-biased junction can certainly grow large!

17. Other Non-Ideal Effects in the p-n Junction

17.1. Reverse-Bias Breakdown (*Sec. 5.4*)

Zener diodes (less often referred to as breakdown diodes) are semiconductor devices whose function stems from a phenomenon called reverse-bias breakdown. Although referred to as breakdown, the large current that flows when the reverse voltage exceeds a certain value is a completely reversible process.

In a p-n diode biased in the reverse direction, at some voltage, V_{BR}, there is a sudden and dramatic increase in current through the diode. If the current is not limited in the external circuit which drives the diode, the diode will physically break down as well. The breakdown phenomenon is not one single phenomenon really, but there are two mechanisms in action - which of them is active, depends on the voltage applied - either the Zener effect or avalanching. Avalanching is typically the dominant mechanism, with the Zener effect only becoming important when both sides of the junction are heavily doped. The Zener effect is a mechanism governed by carrier tunneling over the forbidden band gap, while avalanching is carrier multiplication due to impact ionization[1].

It is possible to determine whether avalanche or Zener breakdown in a certain p-n diode is occurring by noting the temperature sensitivity of V_{BR}. Although not large the temperature variation of the two types of breakdown are of opposite sign. In the case of the Zener effect, V_{BR} decreases as the temperature increases since the number of valence-band electrons available for tunneling increases. The opposite behavior of breakdown voltage holds for avalanching, as the critical electric field is increased when the temperature increases. This is a much more pronounced behavior in comparison to the Zener effect's temperature dependence which sometimes is viewed as independent of temperature. The increase in the electrical field necessary for avalanching is due to the reduction of the mean-free path which is a result of the increase in the lattice scattering; the carriers don't have time to catch up much velocity until they experience lattice scattering and lose speed.

A mechanism related to breakdown behavior, but which seldom is labelled as one of the reverse-bias breakdown phenomena, is punch-through. This mechanism takes place when the depletion region extends an entire device region - in asymmetric p-n diodes punch-through takes place when the lightly doped side is fully depleted. If a continuously increasing reverse bias voltage is applied to a narrow-base[2] p-n junction diode, it is quite possible that it experiences punch-

1. The main textbook gives an adequate overview on these phenomena on pages 186-190.
2. If the width of the quasi-neutral region on the lightly doped side of the junction is comparable to or less than a diffusion length, the diode is called a narrow-base diode.

through long before the other breakdown mechanisms are brought into action[1]. Experimentally and in precise theoretical formulations, the diffusion current remains finite at the punch-through voltage in a narrow-base diode, provided the electric field inside the diode is insufficient to produce avalanche breakdown.

17.2. High-Level Injection, Ohmic Losses and Conductivity Modulation (*Secs 5.6.1, 5.6.3*)

The p-n junction current derivation of Sec. 15 relied on the assumption of low-level injection, which means that the excess majority (or minority) carriers are much fewer than the majority carriers due to doping. When V_A approaches V_0 the charge injection increases rapidly and the assumption of low-level injection may no longer be valid but high-level injection has to be assumed.

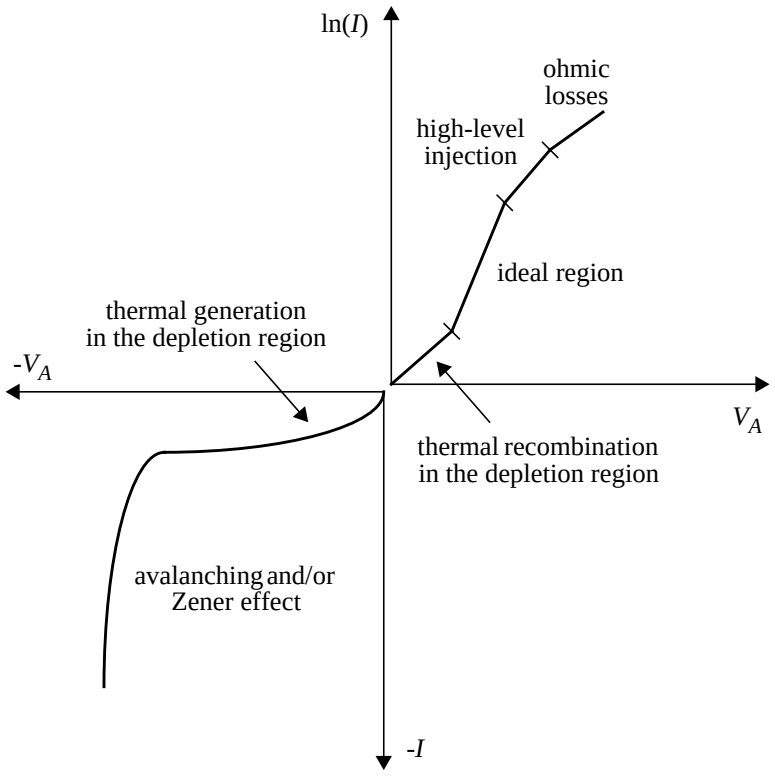


Fig. 17: Non-ideal *I-V* characteristic for a p-n junction.

In a p^+-n or n^+-p junction, high-level injection prevails when the minority carrier density at the depletion region edge on the lightly doped side approaches the doping concentration of that side; for Si typically an interval starting at a few tenths of a Volt below V_0.

It can be shown[2] that the ideal diode equation can be restated to take the injection level into account. The result of this is that **V_0 is the limiting forward voltage for the junction**. V_A across the diode can in fact be larger than V_0 but then there are ohmic losses in the quasi-neutral regions. Increasing the current through the quasi-neutral regions implies that the voltage across these regions increases, which in turn reduces the junction voltage. Reducing the junction

1. This is not true for Zener diodes, since they are carefully designed to have a well-characterized breakdown voltage caused by the Zener effect or avalanching.
2. *Sec. 5.6.1* in the main textbook.

voltage leads to less current through the device. Effectively, part of the applied voltage is wasted, a larger applied voltage is necessary to achieve the same level of current compared to the ideal, and the *I-V* characteristics slope over, which is manifested by a less than exponential current increase with increasing voltage.

In a case of high-level injection the conductivity starts to increase in the quasi-neutral regions. This gives rise to so-called conductivity modulation, which for high currents reduces the series resistance in the quasi-neutral regions. Since the conductivity is not varying greatly, this phenomenon is however mostly obscured by the ohmic losses.

17.3. Summary of Non-Ideal Effects

A non-ideal *I-V* characteristic for a p-n junction diode is shown in Fig. 17[1]. The effects discussed during the last sections are included in this characteristic.

18. DEVICES: p-n Junction Diodes (*Sec. 6.1*)

- Rectifying diodes[2].
- Zener (breakdown) diodes[3].
- Switching diodes[4].

Deliberately the discussion on the charge storage or diffusion capacitance in the p-n junction has been excluded from my lectures. I think the time is too limited. However, to understand the application of the switching diode, you need to browse through *Sec. 5.5.4*. Here, the depletion capacitance, discussed in Sec. 14.4, as well as the diffusion capacitance are dealt with. Also, Problem 6 in our training problems is a problem containing both types of capacitance.

- Varactor diodes[5].

19. Transistors

I wish to give a quite comprehensive treatment of transistors in this course. Otherwise, if obeying the trends of today, where a field-effect transistor called MOSFET dominates, I would essentially reject all transistor types except MOSFET. A clear indication on the trend is that the main textbook in its 5th edition has swapped, with respect to the 4th edition, the discussion on bipolar transistors and field-effect transistors (including MOSFET).

In this course too, the field-effect transistors are discussed prior to the bipolar transistor section, and the reason is **not** pedagogics, but rather a way of showing priority to the most commonly used device technology. We have one more reason to start with the field-effect transistors here, and that reason is to make space in the course for the laboratory exercises that are predominantly targeting MOSFETs. It is certainly good to have heard about field-effect transistors, before doing laboratory exercises on MOSFETs.

1. Compare to *Fig. 5-37* in the main textbook.
2. No use in repeating a good presentation: *Sec. 5.4.3*.
3. No use in repeating a good presentation: *Sec. 5.4.4*.
4. No use in repeating a good presentation: *Sec. 5.5.3*.
5. No use in repeating a good presentation: *Sec. 5.5.5*.

20. Field-Effect Transistors (*Ch. 6*)

There are two basic categories of transistors; the bipolar transistor[1] and the Field-Effect Transistors (FETs). The second category will be the main subject matter in this section.

In contrast to the few versions of bipolar transistors, there exist many transistor types based on field-effect behavior. The three main types are: Junction FETs (JFETs), MEtal-Semiconductor FETs (MESFETs)[2] and Insulated-Gate FETs (IGFETs)[3].

Field-effect behavior was already conceived in 1925 by J. E. Lilienfeld, a Polish-born physicist, in a patent. Independently of Lilienfeld, O. Heil, a German physicist, presented a similar but more detailed structure in a patent from 1935. The closest these first ideas are to modern FETs probably is to MESFET and IGFET structures, respectively.

In effect the "paper" FETs from 1925 and 1935 were the first solid-state transistors ever proposed, thus approximately predating the bipolar transistor of ~1950 by 20-30 years. However, good semiconductor materials were not available at that time and technological immaturity in general retarded the development of field-effect structures for many years.

The JFET was the first modern-day field-effect device. It was conceptually invented by a certain William Shockley[4], a British-American physicist born 1910, and was made known in a paper from 1952, "A unipolar field-effect transistor". The first real JFET device was built by Dacey and Ross one year later. The word unipolar stems from the fact that only majority carriers are involved in the current transport in the JFET, which contrasts with Shockley's achievements some years back when inventing the bipolar transistor (Sec. 25). Traditionally the JFET has been used for high-impedance amplifiers, but nowadays it has been surpassed by a multitude of more efficient transistor structures.

The MESFET is an evolution of the JFET into a domain of structures that are suitable for ultra-high-frequency operation. The MESFET is simple to manufacture as it does not involve the process step of diffusion of a gate-channel junction as in the case of the JFET. The so-called Schottky contact[5] of the MESFET gate is much more easy to deal with in terms of shrinking the geometry. The advantages with small devices are obvious in that the transit time across the channel as well as capacitances are reduced.

Especially one thing about the MESFET should be **NOTE**d here: There is one particular type that is of great importance today, namely the MESFET that is based on GaAs. This is together with the Si Bipolar Junction Transistors (BJTs) the only competitors to the leading Si MOSFET technology in the area of integrated circuits (ICs) as they both present a strong case in terms of maximal frequencies. The highest f_T reported for GaAs MESFETs (~80 GHz) is approximately twice that of the best Si BJTs (~40 GHz).

In terms of circuit technologies, MESFETs in GaAs have been successful in analog applications such as radio-frequency and microwave circuits operating at frequencies on the order of 10

1. The most common bipolar transistor today is the Bipolar Junction Transistor (BJT).
2. Sometimes MESFETs are considered to be a sub category of JFETs!
3. Sometimes called Metal-Insulator-Semiconductor FET (MISFET).
4. Shockley is not to be confused with the Swiss physicist and engineer Walter Schottky, born 1886, who will soon be mentioned.
5. See *Sec. 5.7*.

GHz. Digital circuits based on GaAs MESFETs have one advantage in that they consume only a third of the power of what Emitter-Coupled Logic (ECL)[1] does. However, in the digital world of electronics of today including high-speed areas, which are defined a region around 1 GHz clock frequency, MOSFET-based circuits of so-called CMOS type are ruling. The reason is that the integration scale of GaAs is three orders of magnitude smaller than that of CMOS.

The course will focus on the most common transistor of today, the Metal-Oxide-Semiconductor FET (MOSFET), which is a representative of the IGFET category. Therefore, before we turn to MOSFETs, the principles behind the other two categories, JFETs and MESFETs, will be quickly reviewed by the aid of the two sketches in Fig. 18.

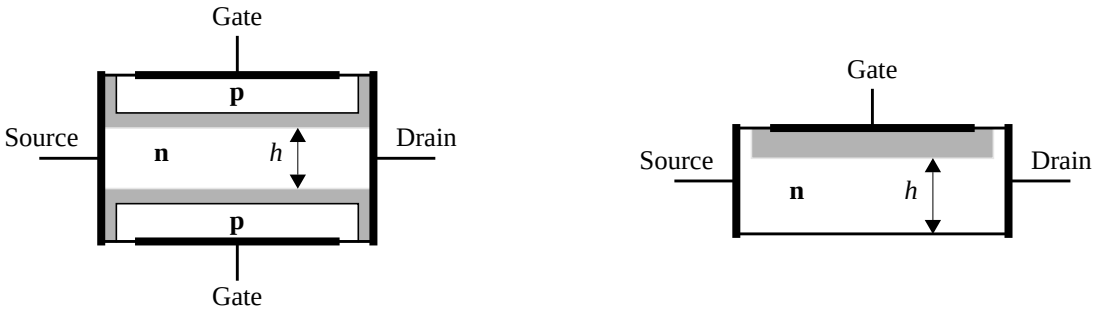


Fig. 18: (a) The basic JFET. (b) The basic MESFET.

By the application of an electric field on the gate in relation to what is known as the source, the channel conductivity in between drain and source can be controlled by change of effective channel width, *h* in Fig. 18. The difference between the two transistor types is that the JFET contains a p-n junction from gate to channel, while the MESFET has its gate tied to a piece of metal only, constituting a Schottky contact.

1. ECL is the fastest bipolar circuit technology available for ICs.

LECTURE 8

21. Metal-Oxide-Semiconductor (MOS) Field-Effect Transistors (*Sec. 6.4*)

21.1. The MOS Structure (*Sec. 6.4.1*)

The MOS designation in MOSFET is implicitly used only for the metal-SiO_2-silicon system, where the metal represents the gate, SiO_2 is the insulator between the gate and the semiconductor itself, and silicon is used as semiconductor (also, the bulk or the substrate of the transistor). Today metal is not used for gates inside chips, instead polycrystalline silicon, often referred to as poly or polysilicon by IC designers, is used for that purpose.

Another implicit convention is that MOSFET is abbreviated as MOS, except in courses like this. But quite unconventionally we adopt MOS from now on.

A sketch showing an n-type MOS is given in Fig. 19. The source and the drain are defined with respect to the potential on the transistor terminals. These two terminals are, in contrast to the BJT's emitter and collector, identical in doping and geometry.

The source is the terminal from which the majority carriers flows and the drain is the terminal where they are collected, i.e. $V_{DS} \geq 0$ for n-type MOS and $V_{DS} \leq 0$ for p-type MOS. Furthermore, here it is assumed that the substrate of the MOS is connected to the same potential as the source is[1].

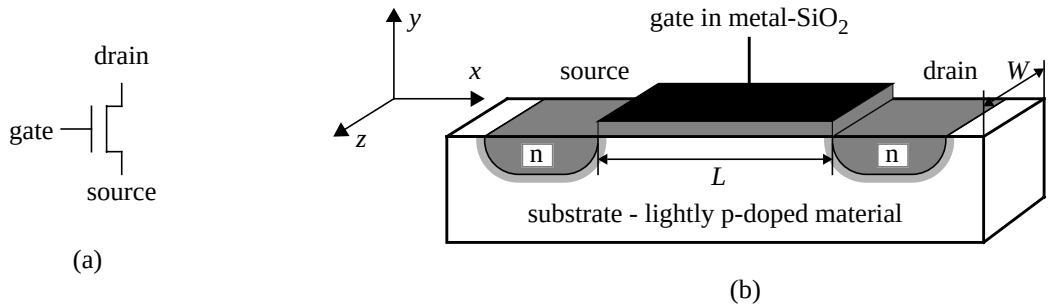


Fig. 19: (a) The "simplest" symbol for an n-type (enhancement-mode) MOS. (b) The structure of an n-type MOS, where an n-type channel can be created beneath the gate insulator. The width and length of the MOS, *W* and *L*, are two very important parameters.

The voltage at the gate terminal is usually defined as V_G. By convention, the source is the terminal used as reference for the gate voltage and V_G implicitly refers to the source potential. To make this clear, V_{GS} will be used from now on.

21.2. Controlling the n-Type MOS (*Sec. 6.4.2*)

Assuming the source, the drain and the substrate are connected to the same potential 0 V, applying a negative potential to the gate as in Fig. 20(a), leads to an accumulation of holes beneath the gate insulator. No current can flow between the source and the drain since the material in between lacks free electrons.

1. This is not true in IC design, because then the substrate mostly is connected to the supply voltage rails. This gives rise to the body effect, which is a modulation of the so-called threshold voltage, see Sec. 24.8.

When a moderate, positive voltage is applied at the gate the majority holes in the substrate will be pushed away from the region immediately beneath the gate, as in Fig. 20(b); a depletion region is formed. The transistor is now in the depletion operation region. Still no current can flow between the source and the drain.

Now, when V_{GS} is increased even further, electrons in the n-type drain and source regions will be attracted to the insulator-semiconductor interface region. An n-type material in the shape of a channel has been created, a channel through which electrons can go from the source to the drain, which is illustrated in Fig. 20(c). The transistor can now conduct current in the channel; the transistor is said to be in the inversion operation region.

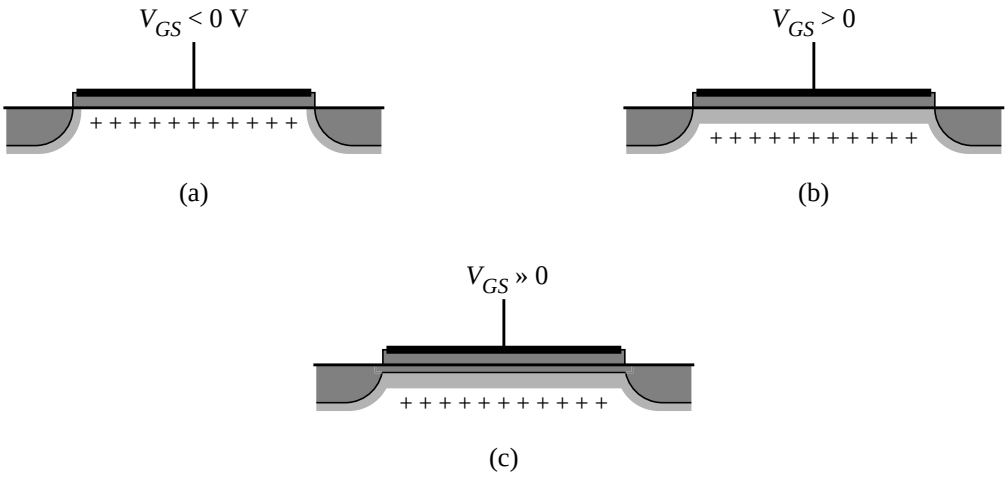


Fig. 20: Three operation regions depending on gate voltage: (a) Accumulation. (b) Depletion. (c) Inversion.

21.3. Enhancement- and Depletion-Mode MOS (*Secs 6.4.1, 6.5.5/page 299*)

Except from polarity differences, there exist two different kinds of MOS, namely enhancement- and depletion-mode MOS. In the enhancement-mode transistor there exist no channel at $V_{GS} = 0$, but V_{GS} has to be raised/(lowered)[1] to, at least, the threshold voltage V_T[2] in order for an efficient channel to be formed. In the depletion-mode MOS, on the other hand, a channel is already present at $V_{GS} = 0$, and V_{GS} rather has to be lowered/(raised) in order for the channel to vanish.

Today the enhancement-mode MOS is by far the most common of the two, but going back 15 years the depletion-mode MOS was as common as the enhancement-mode because it was central to the circuit techniques inside the ICs. A depletion-mode MOS, which has its gate and source tied together as one terminal, i.e. $V_{GS} = 0$, in effect forms a physically small resistor which has its other terminal in the drain; a resistor that is suitable for integration. For example the 8086, the microprocessor of the first PC launched in 1981, was based on a circuit technique called nMOS. This technique relies on n-type enhancement-mode MOS acting as electrical switches and n-type depletion-mode MOS acting as resistors connecting the switches to the supply voltage. Thus, the IC companies could use a simple chip process technology, since only n-type MOS was used.

1. The ordering indicates type of MOS: n-type/(p-type).
2. V_T is negative for p-type MOS.

If the mode is not explicitly given in future sections assume enhancement-mode MOS.

21.4. Pinch-Off in the MOS (*Secs 6.4.1, 6.5.10*)

Suppose we have V_{GS} so that the transistor is in the inversion operation region and that we now start to increase V_{DS}. The increasing voltage on the drain will counteract the electric field in the channel in the vicinity of the drain so that at some voltage, V_{DS}(sat), the channel reaches a state called pinch-off which is a kind of saturation. This phenomenon is shown in Fig. 21.

NOTE that the channel, although pinched off cannot completely vanish. Being a region with few carriers, and hence low conductance, the pinched-off section absorbs most of the voltage drop

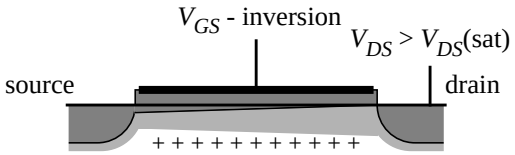


Fig. 21: Pinch-off in an n-type MOS.

in excess of V_{DS}(sat). Denoting the length of the pinched-off region with ΔL, given a long-channel device where $\Delta L \ll L$, the source-to-pinch-off region of the MOS ($L - \Delta L$) will be essentially identical in shape and will have the same endpoint voltages for all $V_{DS} > V_{DS}$(sat). When the shape of the conductive region and the potential applied across the region do not change, the current through the region must remain invariant.

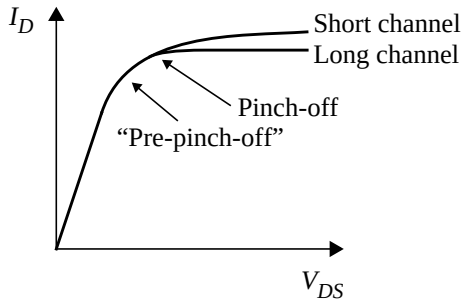


Fig. 22: I_D-V_{DS} characteristic for an n-type MOS at some V_{GS} leading to inversion. The saturation in I_D is due to pinch-off in the channel.

If we, however, have a short-channel MOS, where ΔL is comparable to L, the same voltage drop will appear across a shorter channel ($L - \Delta L$) and, as can be noted in Fig. 22, I_D[1] will increase slightly with increasing V_{DS} even after saturation. In an I_D-V_{DS} characteristic we can summarize the previous discussion with Fig. 22.

22. Analysis of the Long-Channel n-Type MOS (*Secs 6.4, 6.5*)

22.1. Band Diagrams and Energy Definitions in n-Type MOS Structures

The electron affinity is the energy difference between the vacuum energy level and the bottom edge of the conduction band; it is denoted χ. The work function is the difference between the

1. I_D is defined as flowing into the drain terminal, i.e. $I_D = - I_x$ as x is defined in Fig. 19.

vacuum energy level and the Fermi level; it is denoted Φ. We can make a few observations that will make the drawing of energy band diagrams in MOS systems easier: Throughout a MOS structure **1.** the vacuum energy level is continuous, **2.** the work function of the metal is constant, **3.** the electron affinities of the oxide and the semiconductors are constant, and **4.**, at equilibrium, the Fermi level is invariant.

In Fig. 23 the assembly of a MOS structure is shown, such that the proportions of energy levels are true to reality. **NOTE** that the Fermi level of the substrate is located below the intrinsic level, i.e. the **substrate** is of **p-type** and consequently the **MOS** is said to be of **n-type**.

In (a) the three parts are shown prior to being joined and here we have the following values: $\Phi_{me} = 4.10$ V in Al, $\chi_{ox} = 0.95$ V in SiO_2 and $\chi_{se} = 4.05$ V in Si (χ_{se} is slightly larger than 4 V also in Ge and GaAs).

The subscripts me, ox and se denote metal, oxide and semiconductor, respectively. **NOTE** the use of *italic* and plain text styles, plain style as usual indicates material type.

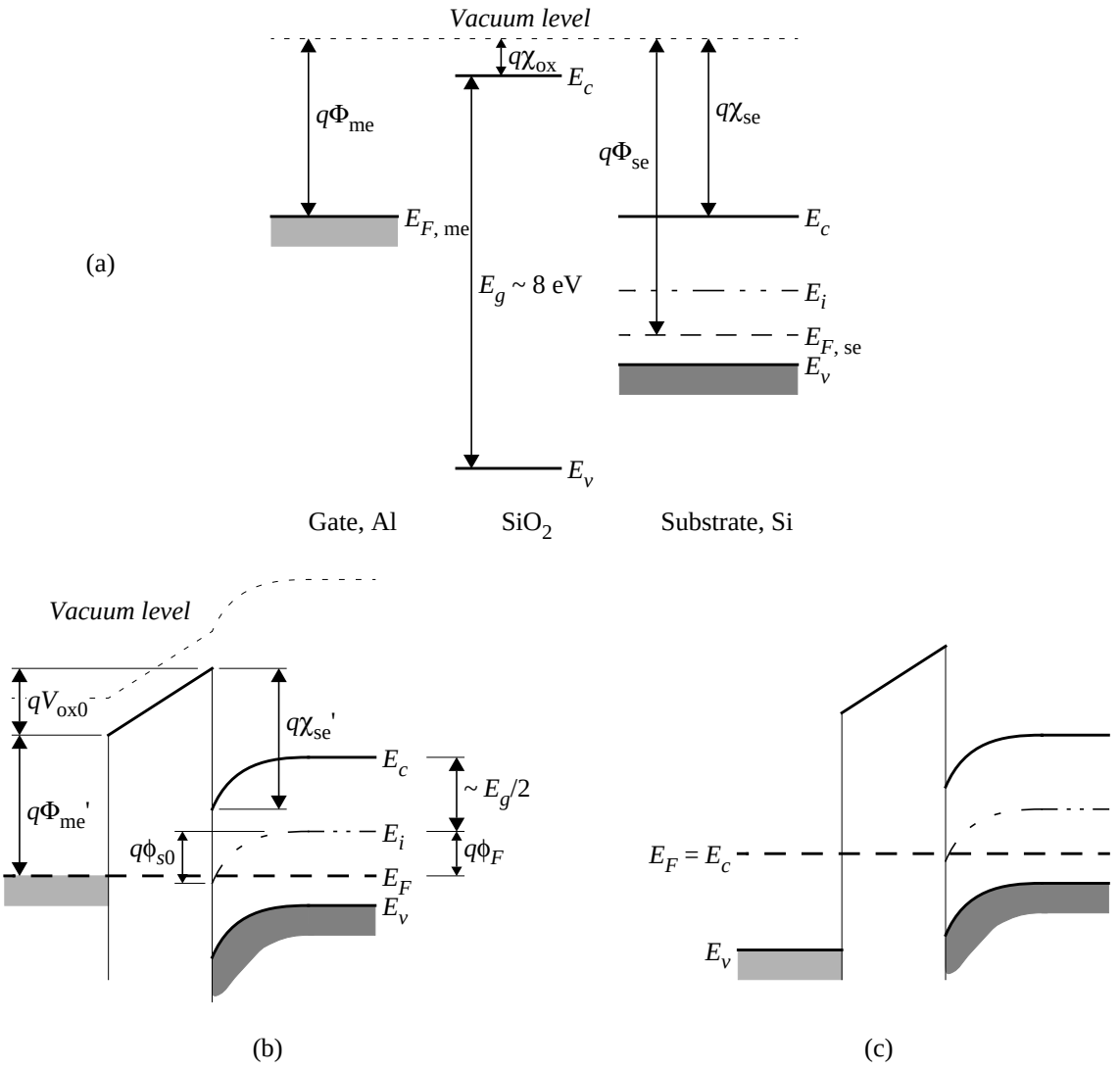


Fig. 23: (a) Energy levels in a MOS system prior to contact. (b) Energy-band diagram through the MOS structure in thermal equilibrium after contact. (c) Energy-band diagram through the MOS structure with a p-type substrate at zero gate bias for an n^+ polysilicon gate.

In Fig. 23(b) the three parts have been joined. We observe some new quantities that will assist

us in defining and analyzing important properties of the MOS structure.

We have χ_{se}' and Φ_{me}' which are the modified electron affinity of the semiconductor and the modified work function of the metal, respectively. They are differing from the earlier definitions in that they are related to the E_c **level of the insulator** as $\chi_{se}' = \chi_{se} - \chi_{ox}$ and $\Phi_{me}' = \Phi_{me} - \chi_{ox}$.

The potential drop across the oxide at zero applied bias simply is denoted V_{ox0}. Moreover, the surface potential for zero applied bias is defined as ϕ_{s0}[1] and corresponds to the amount of bending of the semiconductor band diagram[2]. Finally, the difference in intrinsic and extrinsic Fermi level of the semiconductor is given a new name relating to potential, the Fermi potential, ϕ_F[3].

It should be **NOTE**d that Φ_{se} is not used after the parts are joined, because this quantity varies as we go deeper and deeper into the substrate. This is due to the definition of Φ_{se} as the difference between the vacuum and the Fermi levels, which is increasing as we penetrate deeper into the substrate, see Fig. 23(b).

With reference to Fig. 23(b), since the Fermi level is invariant at equilibrium we can now write the following equation for the energy levels in this n-type MOS:

$$q\Phi_{me}' + qV_{ox0} = q\chi_{se}' + \frac{E_g}{2} - q\phi_{s0} + q\phi_F,$$

where the left side is defined at the metal-oxide interface and the right side, because of the definition of ϕ_{s0}, is defined at the oxide-semiconductor interface.

This can be rewritten as

$$\Phi_{ms} = -(V_{ox0} + \phi_{s0}) = \Phi_{me}' - \left(\chi_{se}' + \frac{E_g}{2q} + \phi_F\right),$$

where Φ_{ms}, consequently defined only from the metal-oxide interface to the oxide-semiconductor interface, is known as the metal-semiconductor work function difference for an n-type MOS.

N**OTE** that Φ_{ms} varies with the doping level of the substrate, because ϕ_F (for a p-type substrate) can be written on a similar form as Eq. (13) which stated that

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

namely as

$$\phi_F = \frac{kT}{q}\ln\left(\frac{N_a}{n_i}\right).$$

Also **NOTE** that Φ_{ms}, which is more easily observed in Fig. 23(a) as $\Phi_{ms} = \Phi_{me} - \Phi_{se}$, is almost always negative for the combination of Al and Si[4]. This is a reason as to why there is energy

1. Be careful not to confuse italic *s* (*surface potential*) with plain style se (semiconductor).
2. For an n-type MOS $q\phi_s = E_i$(bulk) - E_i(surface).
3. For an n-type MOS $q\phi_F = E_i - E_F$.

band bending already at equilibrium as shown in Fig. 23(b) - however, there may be more factors affecting the bending of the energy bands.

Ex. 9: Energy and Potential in the MOS - An Example

Assignment:

Below is a sketch of the energy-band diagram of an n-type Si MOS, i.e. the substrate is of p-type. The surface potential for zero applied bias is defined as ϕ_{s0} and corresponds to the amount

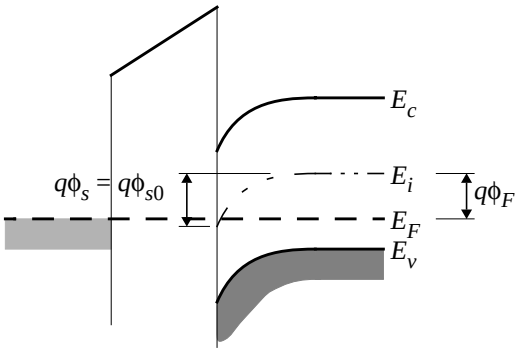


Fig. E.5: Energy-band diagram through the MOS structure in thermal equilibrium after contact.

of bending of the semiconductor band diagram - for an n-type MOS $q\phi_s = E_i$(bulk) - E_i(surface). Finally, the difference in intrinsic and extrinsic Fermi level of the semiconductor is given a new name relating to potential, ϕ_F - for an n-type MOS $q\phi_F = E_i - E_F$.

For the cases given in (a) and (b):

1. draw the energy-band diagram,

2. indicate the MOS transistor type and

3. indicate if the MOS is in the accumulation, the depletion or the inversion operation region.

a) $\dfrac{\phi_F}{kT/q} = 12$ and $\dfrac{\phi_s}{kT/q} = 12$ at room temperature.

b) $\dfrac{\phi_F}{kT/q} = 18$ and $\dfrac{\phi_s}{kT/q} = 36$ at room temperature.

Solution:

If we go back to Eq. (13) we have

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

which allows us to write ϕ_F as a function of the amount of doping when full ionization is assumed

4. An exception should be made for heavily doped n-type substrates; then, if $\Phi_{se} \rightarrow \chi_{se}$, Φ_{me} could become larger than Φ_{se}.

$$\phi_F = \frac{kT}{q} \ln\left(\frac{N_a}{n_i}\right).$$

ϕ_s, on the other hand, is a function of the applied bias on the gate.

Case (a):

1. Here we have

$$\frac{\phi_F}{kT/q} = 12,$$

which is evaluated into

$$q\phi_F = 12kT = 12 \cdot 0.0259 \text{ eV} = 0.31 \text{ eV} = 56\% \text{ of } \frac{E_g}{2}.$$

Also, simply

$$\phi_s = \phi_F.$$

An illustration of the energy bands is given in Fig. E.6(a).

2. ϕ_F is positive, i.e. this is an n-type MOS (it has a p-type substrate).

3. At the surface of the substrate, the Fermi level is coinciding with the intrinsic Fermi level, i.e. the MOS is in between depletion and inversion.

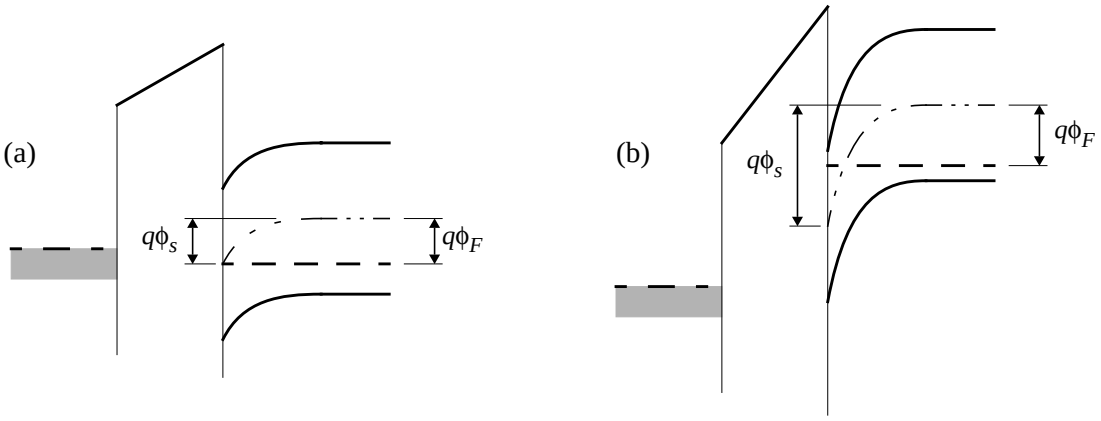


Fig. E.6: Drawn energy-band diagrams for cases (a) and (b) in the assignment.

Case (b):

1. Here we have

$$\frac{\phi_F}{kT/q} = 18,$$

that is,

$$q\phi_F = 18kT = 18 \cdot 0.0259 \text{ eV} = 0.47 \text{ eV} = 83\% \text{ of } \frac{E_g}{2}.$$

Now, simply

$$\frac{\phi_s}{kT/q} = 36$$

i.e.

$$q\phi_s = 36kT = 36 \cdot 0.0259 \text{ eV} = 0.93 \text{ eV}.$$

An illustration of the energy bands is given in Fig. E.6(b).

2. ϕ_F is positive, i.e. this is an n-type MOS (it has a p-type substrate).

3. At the surface of the substrate, the Fermi level is as high above the intrinsic Fermi level as the Fermi level is below the intrinsic level in the bulk, i.e. the MOS has just entered strong inversion.

LECTURE 9

22.2. Threshold Voltage in n-Type MOS Structures

In Sec. 21.3 the mentioning of the threshold voltage was quite deliberately a bit vague - an "efficient" channel was formed when V_{GS} reached V_T.

Formally the threshold voltage is defined as the voltage needed to create a channel whose surface is as strongly n-type as the substrate, in which the channel is formed, is p-type. The definition of inversion remember, is when a channel of opposite type to the substrate is created, however, the channel may be very weak in that few free carriers exist in the channel. The requirement for the threshold voltage obviously does not necessarily hold at inversion, the channel may contain far too few carriers for that. No, we have go a bit further to the strong inversion operation region.

In terms of surface potential, ϕ_s, and Fermi potential, ϕ_F, referring to Fig. 23(b), the MOS is within the weak inversion region when

$$2\phi_F > \phi_s > \phi_F ,$$

while the strong inversion region is defined as

$$\phi_s \geq 2\phi_F .$$

Can we find a value of the threshold voltage for this n-type MOS? Applying a gate voltage V_{GS} leads to a shift in energy levels from equilibrium such that

$$V_{GS} = (V_{\text{ox}} - V_{\text{ox0}}) + (\phi_s - \phi_{s0}) = V_{\text{ox}} + \phi_s + \Phi_{\text{ms}} .$$

Since strong inversion occurs when $\phi_s = 2\phi_F$ we have

$$V_T = V_{\text{ox},\,T} + 2\phi_F + \Phi_{\text{ms}} , \tag{52}$$

where $V_{\text{ox},\,T}$ is the voltage across the oxide at the threshold voltage. In order to find $V_{\text{ox},\,T}$ we have to analyze the charges per unit area present in the MOS structure:

In the oxide there are almost always unwanted charges present, however advanced the fabrication process is. In the charge Q_{ox} we include contributions such as mobile ions (Na^+) which are a result from contamination in the fabrication process and oxide trapped charges which are due to imperfections in the oxide[1]. Also, Q_{ox} takes into account charges that are due to the oxide-semiconductor interface states, i.e. oxide fixed charges and interface trap charges[2]. The oxide charge is positive and we simplify the analysis by assuming this charge is sitting at the oxide-semiconductor interface.

In the semiconductor on the verge of inversion we only have one region in which there exist charges, the depletion region. This is based on neglecting the charge in the channel, which we can do since the channel is actually just formed ($V_{GS} = V_T$). Letting dr represent the depletion region, this charge is denoted Q_{dr}. Since we are dealing with an n-type MOS, we have a p-type substrate, and therefore Q_{dr} will be negative.

1. See *Sec. 6.4.3*.
2. See *Secs 5.7.4* and *6.4.3*.

As an extension, let us derive an expression for Q_{dr}, based on the depletion region width, which can be written in a way similar to Eq. (48)

$$W_{\mathrm{dr}} = \sqrt{\frac{2\varepsilon_{\mathrm{se}}}{q}\frac{1}{N_a}\phi_s},$$

and consequently, for W_{dr} ($W_{\mathrm{dr},\,T}$) at threshold,

$$W_{\mathrm{dr},\,T} = \sqrt{\frac{4\varepsilon_{\mathrm{se}}}{q}\frac{1}{N_a}\phi_F}. \tag{53}$$

At threshold we can write

$$Q_{\mathrm{dr}} = -qN_a\,W_{\mathrm{dr},\,T} = -qN_a\sqrt{\frac{4\varepsilon_{\mathrm{se}}}{q}\frac{1}{N_a}\phi_F} = -\sqrt{4q\varepsilon_{\mathrm{se}}\,N_a\,\phi_F}$$

Now, back on the track for finding V_T; the charge in the metal, Q_{me}, which has to compensate for the difference in Q_{dr} and Q_{ox} (since we don't want to have a current flowing through the insulator) is

$$Q_{\mathrm{me}} = \left|Q_{\mathrm{dr}}\right| - Q_{\mathrm{ox}}.$$

Thus, since the voltage across the oxide can be related to the charge on the metal and to the oxide capacitance per unit area as

$$V_{\mathrm{ox},\,T} = \frac{Q_{\mathrm{me}}}{C_{\mathrm{ox}}},$$

we arrive at

$$V_T = \frac{1}{C_{\mathrm{ox}}}\left(\left|Q_{\mathrm{dr}}\right| - Q_{\mathrm{ox}}\right) + 2\phi_F + \Phi_{\mathrm{ms}} \tag{54}$$

for an n-type MOS.

In a real MOS transistor the threshold voltage would be tuned by implanting donors or acceptors at the surface between oxide and semiconductor. In an n-type MOS, a donor implant with surface concentration per unit area, N_s, would shift V_T such that

$$\Delta V_T = -\frac{qN_s}{C_{\mathrm{ox}}}.^{(1)}$$

NOTE that our considerations on threshold voltage are only valid for long and wide MOS devices. In later sections, non-ideal effects due to smaller feature sizes are introduced.

22.3. The Current-Voltage Relationship for the n-Type MOS

Under the assumption that

1. the current in the channel is entirely due to drift,

1. See *Sec. 6.5.5.*

2. there is no current through the gate oxide,

3. $\frac{\partial E}{\partial y} \gg \frac{\partial E}{\partial x}$, i.e. the electric field in the *x*-direction as defined in Fig. 19 is constant,

4. any fixed oxide charge is an equivalent charge density at the oxide-semiconductor interface, and finally,

5. the carrier mobility in the channel is invariant of position.

The assumption in 3 is called the gradual channel approximation and the entire derivation has its name from that.

The current in the channel can be defined as

$$I_x = \int_y \int_z J_x \, dz \, dy$$

and the inversion layer electron charge per unit area as

$$Q_n = -\int_y qn(y) \, dy .$$

Applying the drift expression now yields

$$I_x = -W\mu_n Q_n E_x , \tag{55}$$

where I_x is the current in position *x* along the channel and *W* is the channel width, the result of integrating over *z*.

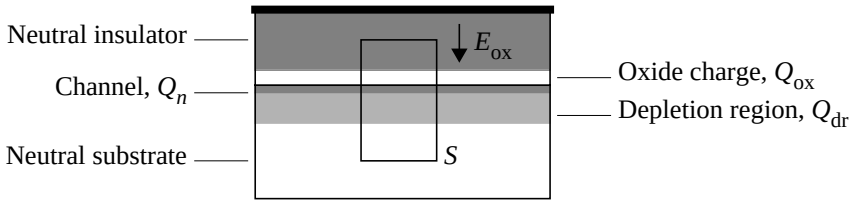


Fig. 24: Performing charge analysis in the channel based on Gauss' law.

We will now find an expression for Q_n: The use of Gauss' law helps us to formulate an expression for charges in the oxide-semiconductor interface. We have

$$\oint_S \varepsilon E_{normal} \, dS = Q_{tot}$$

for the closed surface *S*. Q_{tot} is the total charge enclosed by the surface while E_{normal} is the **outward** directed normal component of the electric field crossing the surface. Consider Fig. 24 in which the surface *S* is defined in the MOS structure. There is no *z*-component of the electric field so the end surfaces in the *x*-*y* plane do not contribute to the integral over *S*. Since the electric field in the *x*-direction is constant, E_x into the surface from the left and E_x out from the surface to the right cancel each other. So, only the *y*-direction is left and here we quickly observe that the bottom of the surface is inside the neutral substrate leaving only the top of the surface for the Gauss integral. Then we have

$$\oint_S \varepsilon E_{normal} \, dS = -\varepsilon_{ox} E_{ox} W \, dx .$$

Moreover the enclosed charge is

$$Q_{tot} = (Q_{ox} + Q_n + Q_{dr})W\ dx,$$

which together with the integral gives

$$-\varepsilon_{ox}E_{ox} = Q_{ox} + Q_n + Q_{dr}. \tag{56}$$

This looks fine, however we don't know E_{ox}. But this field should be possible to solve based on the voltage across the oxide and the oxide thickness, t_{ox}, by

$$E_{ox} = \frac{V_{ox}}{t_{ox}}$$

Here we have assumed that Q_{ox} is sitting at the oxide-semiconductor interface.

In Eq. (52) we formulated an expression for the threshold voltage. Using V_x as the potential in the channel at a point x along the channel length, we can derive a similar relationship:

$$V_{GS} - V_x = V_{ox} + 2\phi_F + \Phi_{ms},$$

which leads to

$$V_{ox} = V_{GS} - V_x - (2\phi_F + \Phi_{ms}).$$

The left hand side of Eq. (56) can now be evaluated as

$$-\varepsilon_{ox}E_{ox} = -\varepsilon_{ox}\frac{V_{ox}}{t_{ox}} = -\frac{\varepsilon_{ox}}{t_{ox}}\left[V_{GS} - V_x - (2\phi_F + \Phi_{ms})\right]$$

and consequently Eq. (56) becomes

$$-\frac{\varepsilon_{ox}}{t_{ox}}\left[V_{GS} - V_x - (2\phi_F + \Phi_{ms})\right] = Q_{ox} + Q_n + Q_{dr}.$$

Reorganizing for isolating Q_n (for use in Eq. (55)) leads to

$$Q_n = -C_{ox}\left\{V_{GS} - V_x - \left[2\phi_F + \Phi_{ms} - \frac{1}{C_{ox}}(Q_{ox} + Q_{dr})\right]\right\},$$

where Q_{dr} is negative. With the aid of Eq. (54)[1] we have

$$Q_n = -C_{ox}(V_{GS} - V_x - V_T).$$

So, we now have solved for Q_n. From Eq. (24) we also know that the electric field

$$E_x = -\frac{dV_x}{dx}$$

and we can consequently write Eq. (55) as

1. A simplification done here is to assume that Q_{dr} is constant, which is not true as the width of the depletion region varies with x-position in the channel when $V_{DS} > 0$.

$$I_x = -W\mu_n Q_n E_x = -W\mu_n C_{\text{ox}} (V_{GS} - V_x - V_T) \frac{dV_x}{dx}.$$

We can now integrate this equation over the length of the channel

$$\int_0^L I_x \, dx = -W\mu_n C_{\text{ox}} \int_{V_x(0)}^{V_x(L)} [(V_{GS} - V_T) - V_x] \, dV_x,$$

where, at the drain, $V_x(L) = V_{DS}$ and, at the source, $V_x(0) = 0$ V which yields

$$I_D = -I_x = \frac{W}{L} \frac{\mu_n C_{\text{ox}}}{2} [2(V_{GS} - V_T)V_{DS} - V_{DS}^2]. \tag{57}$$

This description of the drain current, defined as going from the drain to the source, is valid for $V_{GS} \geq V_T$ and $0 \leq V_{DS} \leq V_{DS}(\text{sat})$, the non-saturation or linear region. When the MOS has reached pinch-off however, the drain current becomes saturated and then remains essentially the same. The peak in I_D occurs at

$$V_{DS} = V_{GS} - V_T$$

and thus

$$V_{DS}(\text{sat}) = V_{GS} - V_T.$$

The drain current for $V_{DS} > V_{DS}(\text{sat})$, the saturation region, can therefore be written as

$$I_D = \frac{W}{L} \frac{\mu_n C_{\text{ox}}}{2} (V_{GS} - V_T)^2. \tag{58}$$

NOTE that the above equations for I_D not only hold for n-type enhancement-mode MOS but also for n-type depletion-mode MOS where V_T is negative.

Ex. 10: The MOS as a Resistance - An Example

Assignment:

A 5-μm wide, 1-μm long MOS for which the gate-oxide thickness is 80 nm (SiO_2), and the channel mobility $\mu_n = 600$ cm^2/Vs is to be used as a voltage-controlled resistor.

1. Calculate the free-electron density in the channel that is required for the MOS to present a resistance of 2.5 kΩ between the source and the drain at low values of V_{DS}.

2. Calculate the gate voltage in excess of the threshold voltage needed to produce the desired resistance under the conditions in (a).

Solution:

The resistance of the channel, R_{ds}, is related to the output conductance, g_{ds}, as

$$\frac{1}{R_{ds}} = g_{ds} = \frac{dI_D}{dV_{DS}}.$$

We are obviously in the linear region, as V_{DS} is small. Thus,

$$g_{ds} = \frac{d}{dV_{DS}}\left\{\frac{W}{L}\frac{\mu_n C_{\text{ox}}}{2}\left[2(V_{GS}-V_T)V_{DS}-V_{DS}^{\,2}\right]\right\},$$

which can be written as

$$g_{ds} = \frac{W}{L}\frac{\mu_n C_{\text{ox}}}{2}\left[2(V_{GS}-V_T)-2V_{DS}\right].$$

Since V_{DS} is small, we simplify the expression above, by using $V_{DS} = 0$, further into

$$g_{ds} = \frac{W}{L}\mu_n C_{\text{ox}}\,(V_{GS}-V_T).$$

As in the derivation of the square-law theory, the charge in the channel, Q_n, is represented by the capacitance and the voltage, such that

$$Q_n = -C_{\text{ox}}\,(V_{GS}-V_T).$$

In problem 1 we are seeking $\frac{Q_n}{-q}$, which simply can be found by the use of

$$\left(-\frac{W}{L}\mu_n\right)Q_n = \frac{1}{R_{ds}}$$

or rather

$$Q_n = -\frac{L}{W\mu_n R_{ds}}.$$

Then we get

$$\frac{Q_n}{-q} = \frac{L}{Wq\mu_n R_{ds}} = \frac{1}{5\cdot 1.6\times10^{-19}\cdot 600\cdot 2500} = 8.32\times10^{11}\ \text{cm}^{-2}.$$

In problem 2 we look for $V_{GS}-V_T$, which now can be extracted from

$$V_{GS}-V_T = \frac{Q_n}{-C_{\text{ox}}},$$

where

$$C_{\text{ox}} = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}}.$$

With $\varepsilon_r = 3.9$ for SiO_2, we have $\varepsilon_{\text{ox}} = 3.45\times10^{-13}$ F/cm and $t_{\text{ox}} = 80\times10^{-9}$ m $= 80\times10^{-7}$ cm, and thus

$$V_{GS}-V_T = \frac{L\,t_{\text{ox}}}{W\varepsilon_{\text{ox}}\mu_n R_{ds}} = 3.09\ \text{V}.$$

23. Performance of the MOS

23.1. The Transconductance of the MOS

We will now derive an expression for g_m, the transconductance for the MOS. The transconductance for the MOS is defined as

$$g_m = \frac{dI_D}{dV_{GS}}.$$

Using I_D according to Eq. (57) and Eq. (58) yields

$$g_m = \frac{W}{L}\,\mu_n C_{\text{ox}}\,V_{DS} \text{ in the non-saturation region,}$$

and

$$g_m = \frac{W}{L}\,\mu_n C_{\text{ox}}\,(V_{GS}-V_T) \text{ in the saturation region.} \quad (59)$$

Obviously, the transconductance increases linearly with V_{DS} but is independent of V_{GS} in the non-saturation region. In the saturation region on the other hand, the transconductance is a linear function of V_{GS} and independent of V_{DS}. **NOTE** that these quantities are valid in long-channel MOS only.

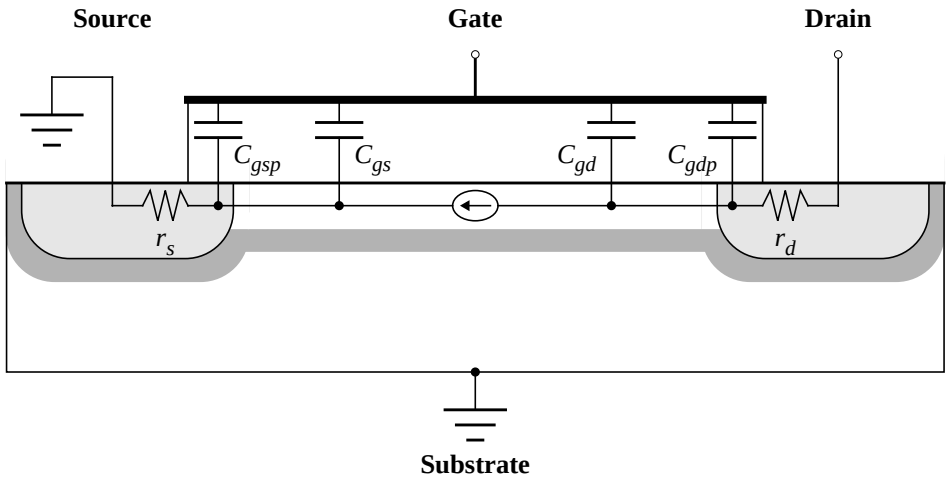


Fig. 25: Inherent resistances and capacitances in the n-type MOS.

23.2. Frequency Limitations and the Cut-Off Frequency of the MOS

There are two basic frequency-limiting factors in the MOS. The first factor is the channel transit time. If we have carriers travelling at their saturation drift velocity, v_{sat}, then the transit time is

$$\tau_t = \frac{L}{v_{sat}}.$$

Assume reasonable values such as $v_{sat} = 10^7$ cm/s and $L = 1$ μm; then $\tau_t = 10$ ps, which translates into a maximum frequency of 100 GHz. This is much larger than the typical frequency response of a MOS.

No, the transit time is not the limiting factor, but rather the capacitances present in the MOS structure are. In Fig. 25 there is a sketch on what parameters could show up in the MOS structure - this sketch is not fully complete, but is adapted to our purposes. C_{gs} and C_{gd} denote the capacitances which are due to the interactions between charges at the gate and the source, and charges at the gate and the drain, respectively. Likewise, C_{gsp} and C_{gdp} are symbols for the parasitic capacitances that are due to the physical overlap between, on one hand, the gate and, on the other hand, the source and the drain, respectively. Associated with the source and the drain terminals, we also have to take into account the resistances r_s and r_d, respectively.

A high-frequency small-signal equivalent circuit of a MOS, based on Fig. 25, is shown in Fig. 26. The parameters C_{gst} and C_{gdt} are the total gate-to-source and gate-to-drain capacitances, respectively. As a high frequency is assumed, r_s and r_d are neglected, but instead R_L, the load resistance, is included.

In this equivalent circuit we have at the input gate node

$$I_i = j\omega\, C_{gst}\, V_{gs} + j\omega\, C_{gdt}\, (V_{gs} - V_{ds})$$

and at the output drain node, summing all outward currents,

$$\frac{V_{ds}}{R_L} + g_m V_{gs} + j\omega\, C_{gdt}\, (V_{ds} - V_{gs}) = 0\,.$$

V_{ds} can be eliminated from the I_i equation by inserting the equation above solved for V_{ds}. The result is

$$I_i = j\omega\left[C_{gst} + C_{gdt}\left(\frac{1 + g_m R_L}{1 + j\omega R_L C_{gdt}}\right)\right]V_{gs}\,.$$

Normally, $\omega R_L C_{gdt}$ is much less than unity; therefore we may neglect the $(j\omega R_L C_{gdt})$ term in

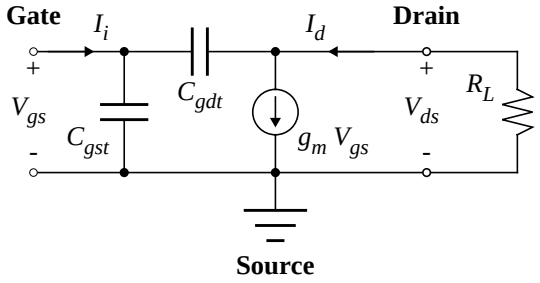


Fig. 26: A high-frequency small-signal equivalent circuit of a common-source n-type MOS.

the denominator. Thus, we can simplify I_i into

$$I_i = j\omega[C_{gst} + C_{gdt}\,(1 + g_m R_L)]V_{gs}$$

and this is a very important expression because it clearly shows us that the gate-to-drain capacitance can become a significant factor in the input impedance since it is multiplied by the transconductance, the transistor gain. Usually, the expression for I_i is written as

$$I_i = j\omega(C_{gst} + C_M)V_{gs}\,,$$

where C_M is the Miller capacitance. In Fig. 27 the inclusion of a Miller capacitance is illustrated.

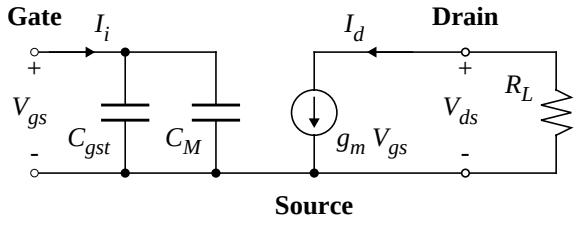


Fig. 27: Modified equivalent circuit with Miller capacitance.

Let us consider a MOS operating in the saturation region: Then we have pinch-off close to the drain and consequently C_{gd} is essentially zero. However, C_{gdp} will be constant and independent of applied voltages as it is due to the physical geometry of the device. So even though C_{gdp} may be insignificant itself, as it effectively acts as a capacitance many times larger due to the gain it has a serious effect on performance.

Remaining in the saturation region, we will now derive an expression for the cut-off frequency, f_T, a most useful figure of merit. The cut-off frequency is defined as the frequency, where the current gain of the device, having zero load impedance, has fallen to unity. However, in the case of the MOS, cut-off frequency will rely on an ideal MOS and therefore the applicability somewhat limited.

In Fig. 27 we have already defined I_i so we also need to have an expression for I_d to be able to calculate f_T. This is however simple because the ideal drain current is

$$I_d = g_m V_{gs}.$$

Thus, the magnitude of the current gain is

$$\left|\frac{I_d}{I_i}\right| = \frac{g_m}{2\pi f\ (C_{gst} + C_M)}$$

and the cut-off frequency hence can be defined as

$$f_T = \frac{g_m}{2\pi\ (C_{gst} + C_M)}.$$

In the ideal MOS, the parasitic capacitances are zero since there are no overlaps between the gate and either of the source and the drain. Even if the MOS was not ideal, since R_L is zero, according to the definition of cut-off frequency as having zero load impedance, C_{gdp} would have had a limited influence, in fact it is usually smaller than C_{gs}. Furthermore, C_{gd} will be essentially zero as the MOS is biased in the saturation region, so from Fig. 25 only C_{gs} will remain and this capacitance is roughly equal to $C_{\text{ox}}\,W\,L$.

In Eq. (59) we derived an expression for g_m in saturation,

$$g_m = \frac{W}{L}\,\mu_n C_{\text{ox}}\,(V_{GS} - V_T).$$

At last, we can now write an elegant function to describe f_T:

$$f_T = \frac{\frac{W}{L}\mu_n C_{ox}(V_{GS}-V_T)}{2\pi(C_{ox}WL)} = \frac{\mu_n(V_{GS}-V_T)}{2\pi L^2}.$$

Despite we had to sacrifice some accuracy[1] to achieve this expression, there is one thing that now is well illustrated; the dependence of the channel length on the cut-off frequency. As MOS devices are continuously scaled down in dimensions the cut-off frequency is increasing, leading to ever faster electronics in the future.

1. Mainly the contribution from parasitic capacitances will greatly reduce the practical f_T.

LECTURE 10

24. Non-Ideal Effects in the MOS (*Sec. 6.5*)

A significant number of general non-ideal effects will be reviewed in the following.

24.1. The Bulk-Charge Theory

In Sec. 22.3 a vital assumption for the derivation was that Q_n was constant. This is obviously not true as the channel is not constant in size but depends on the applied voltages. If care is taken, a varying Q_n leads to an expression such as the one that describes e.g. the non-saturation region:

$$I_D = \frac{W}{L}\frac{\mu_n C_{\text{ox}}}{2}\left[2(V_{GS}-V_T)V_{DS}-V_{DS}{}^2-\eta\right],$$

where

$$\eta = \frac{8}{3}\frac{qN_a W_{\text{dr},T}}{C_{\text{ox}}}\phi_F\left[\left(1+\frac{V_{DS}}{2\phi_F}\right)^{3/2}-\left(1+\frac{3V_{DS}}{4\phi_F}\right)\right]$$

and $W_{\text{dr},T}$ is given by Eq. (53). In contrast to the following sections on the non-ideal effects, the mathematical expressions here are not central to the course, but serve only the purpose to complete the picture for those who are interested.

N_a denotes the doping density of the substrate (or the bulk), and the use of this quantity is in fact the only reason for showing these complicated expressions: When the substrate is lightly doped the ordinary drain current equation in Eq. (57), the square-law theory works fine, but when the substrate is increasingly doped the expressions above, the so-called bulk-charge theory has to be used. As N_a is increased the drain current according to the bulk-charge theory is lower than for the square-law theory.

24.2. Subthreshold Operation (*Sec. 6.5.7*)

The ideal current-voltage relationships in Eq. (57) and Eq. (58) are not valid for $0 < V_{GS} < V_T$, the so-called subthreshold region. They predict that there can be no current transport in the MOS for $V_{GS} < V_T$, which is wrong. The reason for the subthreshold current to appear is quite simple: The MOS is experiencing weak inversion for

$$2\phi_F > \phi_s > \phi_F$$

and when this is fulfilled there exists a channel beneath the oxide interface, a tiny channel that can conduct current although $V_{GS} < V_T$.

In the subthreshold region, the conduction mechanism is not primarily based on drift but rather on diffusion. This is due to the potential barrier that exists between the source and the channel; at weak inversion only some of the electrons inside the source of an n-type MOS have enough energy to diffuse across the barrier, while at strong inversion the barrier is so small that the exponential dependence is lost[1]. A current-voltage expression at subthreshold operation,

1. At strong inversion, the junction between the source and the channel is more like an ohmic contact, about which you can read in *Sec. 5.7.3*.

can be written roughly as

$$I_D \propto e^{\frac{qV_{GS}}{kT}}\left(1-e^{-\frac{qV_{DS}}{kT}}\right),$$

and here we **NOTE** that if V_{DS} is just somewhat larger than kT/q, the subthreshold current is independent of V_{DS}.

In the research field of digital electronics the issue of currents in the subthreshold region is important, as digital circuits based on MOS transistors rely on having transistors either conducting (ON-switch) or isolating (OFF-switch). If the subthreshold region is considered, MOS transistors are no more perfect switches, since e.g. $V_{GS} < V_T$ still can cause an n-type MOS to conduct current. Leakage currents will appear in switched-OFF MOS and so-called dynamic circuits, which are based on charge storage on capacitances, will have their operation severely degraded.

24.3. Channel-Length Modulation (*Sec. 6.5.10*)

In Fig. 21 the saturation current as a function of V_{DS} was shown. When a long channel was assumed the MOS exhibited a constant I_D after pinch-off despite V_{DS} continued to increase, but for the short channel I_D increased with increasing V_{DS} after pinch-off as well. The variation in effective channel length is often referred to as channel-length modulation. To account for the dependence on V_{DS} during saturation a channel-length modulation factor, λ, usually is employed, such that

$$I_D = \frac{W}{L}\frac{\mu_n C_{ox}}{2}(V_{GS}-V_T)^2(1+\lambda V_{DS}).$$

24.4. Surface Scattering (*Sec. 6.5.3*)

The mobility in the channel of a MOS is reduced due to an effect called surface scattering or sometimes, in circuit-related texts, mobility degradation. As carrier motion in a MOS takes place in a surface inversion layer, the gate-induced electric field in **the vertical direction** acts so as to accelerate the carriers toward the surface. But when approaching the surface, the carriers are repelled by localized coulombic forces in the semiconductor surface. If fixed oxide charges are present at the oxide-semiconductor interface the mobility is reduced even further.

There are several different models to describe the mobility degradation with respect to the vertical electric field. The most popular model is the following:

$$\mu_{eff} = \frac{\mu_0}{\left(1+\frac{E_y}{E_{y0}}\right)^{\nu}},$$

where E_y is the actual vertical field, whereas E_{y0} and ν are two parameters obtained empirically.

24.5. Velocity Saturation (*Secs 6.5.3, 6.5.4*)

The mobility can be reduced for another reason than the one given in Sec. 24.4. As was discussed in Sec. 7.4, large electric fields **along the channel** can cause a reduction in the mobility, since the carriers cannot absorb much more energy when reaching a certain kinetic energy. Already in that discussion, it was revealed that the hot-carrier effect is a problem in today's short-

channel MOS transistors.

Sec. 24.4 describes the phenomenon usually called mobility degradation (at circuit level). We must distinguish between the two different effects that degrade mobility[1], thus, the name chosen for the hot-carrier effect at circuit level is usually velocity saturation.

Among all of the non-ideal effects that are discussed in this section, velocity saturation and subthreshold conduction are the most important ones. Velocity saturation has a very dominant effect on the *I-V* characteristics of modern-day devices. In fact Eq. (58), which was derived for a long-channel saturated case, can be shown to far from accurately describe short-channel devices, i.e. MOS with $L < 1$ μm. Thus, now we will derive an expression for I_D in an n-type MOS in the presence of velocity saturation:

The electron saturation velocity, v_{sn}, relates to the critical electric field as

$$E_{sn} \approx \frac{v_{sn}}{\mu_n}.$$

Once the electric field at the drain side of the channel, where the field is the highest, exceeds E_{sn}, the electron velocity saturates. Based on Eq. (55), where essentially the following is given

$$I_x = W\mu_n Q_n E_x,$$

the field at the drain side of the channel can be described as

$$E_x(\text{drain}) = \frac{I_D}{W\mu_n Q_n(V_{DS})}.$$

Now assuming that velocity saturation is manifest at the drain, i.e. $E_x(\text{drain}) = E_{sn}$, $V_{DS} = V_{SAT}$ and $I_D = I_{SAT}$, we get

$$E_{sn} = \frac{I_{SAT}}{W\mu_n Q_n(V_{SAT})}.$$

In Sec. 22.3, we derived an expression for the charge, $Q_n(V_x)$. Looking at absolute values we have

$$Q_n = C_{\text{ox}}\,(V_{GS} - V_x - V_T),$$

which allows us to evaluate E_{sn}

$$E_{sn} = \frac{I_{SAT}}{W\mu_n C_{\text{ox}}(V_{GS} - V_T - V_{SAT})}.$$

We would like to express I_{SAT} as a function of known parameters, but since V_{SAT} also is an unknown we need a second equation linking I_{SAT} and V_{SAT}. For this purpose we use Eq. (57), which is the current-voltage relation in the linear region and which should hold on the limit to saturation:

1. To be fair, only surface scattering is directly influencing the mobility. The hot-carrier effect goes beyond the simple notion of mobility.

$$I_D(V_{DS} \to V_{SAT}) = I_{SAT} = \frac{W}{L}\frac{\mu_n C_{ox}}{2}\left[2(V_{GS}-V_T)V_{SAT} - V_{SAT}^2\right].$$

Solving the two above equations, which is quite tiresome, gives us an expression for I_{SAT}:

$$I_{SAT} = \frac{\frac{W}{L}\mu_n C_{ox}(V_{GS}-V_T)^2}{1+\sqrt{1+\left(\frac{V_{GS}-V_T}{E_{sn}\cdot L}\right)^2}}. \tag{60}$$

This should be compared to Eq. (58), which is the *I-V* equation for the saturated MOS[1],

$$I_D = \frac{W}{L}\frac{\mu_n C_{ox}}{2}(V_{GS}-V_T)^2.$$

An analysis of the I_{SAT} expression, with the form of I_D, would be something like

$$I_D \propto (V_{GS}-V_T),$$

that is, under velocity saturation the dependence on V_{GS} is significantly weakened.

24.6. Source and Drain Series Resistances

In short-channel devices the channel resistance is smaller than in long-channel devices, and thus the relative impact of resistance in source and drain regions is larger in short-channel devices. If we let V_{GS} and V_{DS} denote the internal voltages, whereas V_{gs} and V_{ds} are the external voltages, we can define a MOS according to Fig. 28.

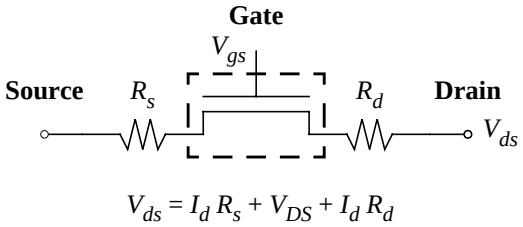


Fig. 28: Equivalent circuit of MOSFET with parasitic source and drain resistances. V_{DS} is the internal voltage across drain and source (at the border of the dashed box).

Now we can define

$$V_{GS} = V_{gs} - I_d \cdot R_s, \tag{61}$$

and

$$V_{DS} = V_{gs} - I_d \cdot (R_s + R_d).$$

Substituting Eq. (61) into Eq. (60) yields the following:

1. Don't mix "saturation" and "saturation": Saturation in Eq. (58) implies only pinch-off, whereas saturation recently has been used to denote velocity saturation. However, they are related: Velocity saturation happens when the drain bias is too high, and that happens when the transistor is in saturation mode!

$$I_{SAT} = \frac{g_{ch} \cdot (V_{gs} - V_T)}{1 + g_{ch}R_s + \sqrt{1 + 2g_{ch}R_s + \left(\frac{V_{gs} - V_T}{E_{sn} \cdot L}\right)^2}},$$

where

$$g_{ch} = \frac{\frac{W}{L}\,\mu_n C_{\mathrm{ox}}(V_{gs} - V_T)}{1 + \left[\frac{W}{L}\,\mu_n C_{\mathrm{ox}}(V_{gs} - V_T)\right](R_s + R_d)}.$$

24.7. Threshold Voltage Variation - the Hot-Carrier Effect Revisited (*Sec. 6.5.9*)

Another result from carriers travelling at high velocities, is that they might leave the semiconductor and tunnel into the insulator, creating fixed charges in the insulator as well as interface states between the semiconductor and the insulator. This effect is referred to as the hot-carrier charging effect. This causes a change in threshold voltage; typically electrons as carriers increase V_T in n-type MOS. Since this process is continuously going on, the device performance is degraded over time.

If showing up unexpectedly, the process of the hot-carrier charging effect is certainly devastating. However, the same mechanism is the key of function for programmable technology based on different sorts of EPROM. To program such ROMs, it is required to “heat up” the carriers so that they tunnel into the gate oxide, in which certain floating volumes of conducting material accept the carriers. The extra charge in the oxide turns on the corresponding device. Erasing of each memory device is done by energizing the carriers of the floating gates, either by optical excitation (UV-light) or electric energy, such that they acquire energy enough to leave the floating gate[1].

24.8. Threshold Voltage Variation - the Body Effect[2] (*Sec. 6.5.6*)

In all of our analyses so far, the substrate has been connected to the source potential. In practice, MOS transistors inside a chip are often having their substrates tied to the supply voltages (V_{DD} for the p-type and ground for the n-type MOS). Stacking several transistors in a series connection will then lead to a difference in source and substrate potential, V_{SB}, for those transistors far away from the supply voltage rail, see Fig. 29(a). The appearance of a $V_{SB} > 0$ gives rise to the so-called body effect, which affects V_T in an n-type MOS such that

$$\Delta V_T = \frac{\sqrt{2q\varepsilon N_a}}{C_{\mathrm{ox}}}\left(\sqrt{2\phi_F + V_{SB}} - \sqrt{2\phi_F}\right),$$

where $\Delta V_T = V_T(V_{SB} > 0) - V_T(V_{SB} = 0)$. N_a and ϕ_F are as usual the substrate doping and the difference E_i - E_F in the substrate, respectively. In summary, the change in V_T is positive for an n-type MOS.

We would not analyze the body effect unless it was of any practical interest. In the first example,

1. Read more on this in the main textbook on pages 470-473.
2. Body, bulk or substrate are all used interchangeably.

where a stack of MOS at some conditions gave a bias voltage on the source of the top n-type MOS, the impact of the body effect was limited. The effect is much more commonplace and important in a structure, which is a key circuit for low-power bus designs.

In Fig. 29(b) a source-follower[1] in CMOS is shown. Since power consumption of CMOS today mostly depends on the size of voltage swing and nodal capacitance, the power savings can be large when applying a local reduction of swing where it really counts, e.g. on a bus with lots of capacitance. Without considering how the input signals are organized, here we simply view the source-follower as a driver of a long bus, such that the high logic level is V_{DD}-V_T, instead of V_{DD}[2]. To get a feeling for how serious the body effect might be here, in the 0.35-μm chip manufacturing process that currently is used in the research at Electronic devices we

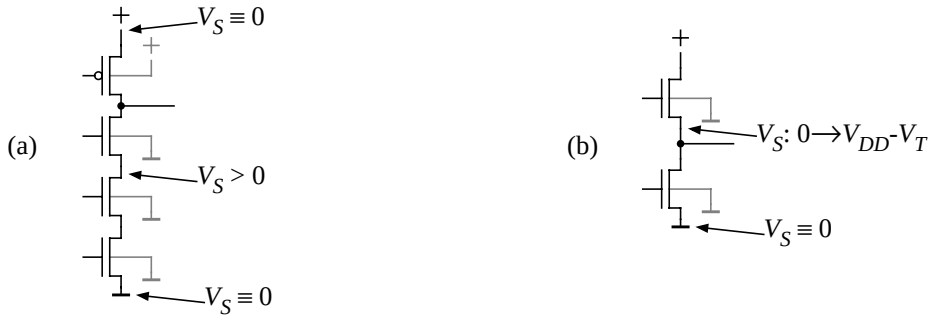


Fig. 29: CMOS structures, see Sec. 34. (a) Stacking of n-type MOS transistors, yielding $V_{SB} > 0$ for top n-type transistors. (b) Source-follower structure for low-swing buses. The transistor at the top has its source tied to the output, which means that V_{SB} can vary between 0 to V_{DD}-V_T V.

have $V_{SB}(\text{max}) = V_{DD}$-$V_T \approx 3.3$-$0.8 \approx 2.5$ V, which gives a $\Delta V_T(\text{max})$ of around 0.4 V[3]. That is, the threshold voltage can vary as much as 50%, depending on source potential.

24.9. Threshold Voltage Variation - the Short-Channel Effect (*Sec. 6.5.11*)

As the channel length is getting shorter and shorter, the depletion regions around the source and the drain will become comparable in size with that of the depletion region beneath the gate. This is important, since we have assumed in the derivation of the I_D-V_{DS} relationships that the space charge in the channel is controlled by the gate voltage only. According to Fig. 30(a), in short-channel devices the space charge density beneath the gate can also be a function of the potential on the other two terminals.

Now, the short-channel effect, this discrepancy from the ideal case, can be accounted for by making a modification to V_T. A derivation of $\Delta V_T = V_T(\text{short channel}) - V_T(\text{long channel})$ yields

$$\Delta V_T = -\frac{qN_a W_{\text{dr},T}}{C_{\text{ox}}} \frac{r_j}{L}\left(\sqrt{1+\frac{2W_{\text{dr},T}}{r_j}}-1\right),$$

where r_j is the depth of the source and the drain junctions. **NOTE** that V_T is reduced with smaller

1. Source-follower: the source (from which majority carriers are emitted) follows the output signal.
2. Why an n-type MOS only can pull the voltage up to V_{DD}-V_T on the output is explained in Sec. 34.1.
3. Now, finding ΔV_T would call for an iteration, as $\Delta V_T = 0.4$ V indicates that V_T no longer is 0.8 V but 1.2 V.

channel lengths.

The quantity r_j represents a kind of radius, which stems from the assumption that the lateral junction distance under the gate is equal to the vertical junction depth. This is a simplification that is more valid for diffused junctions rather than for ion implanted junctions[1].

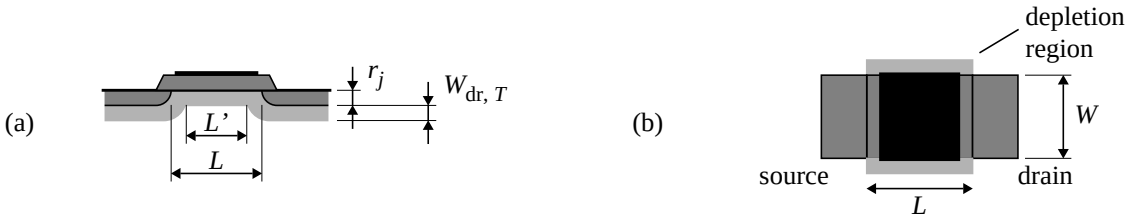


Fig. 30: (a) A sketch illustrating short-channel considerations in a MOS. (b) A top-view sketch of a MOS illustrating considerations of a narrow-channel MOS.

A second assumption made is that the depletion region width is the same everywhere, $W_{dr,\,T}$, which is not quite true. If a drain voltage is applied, the depletion region at the drain side widens, which makes the threshold voltage a function of drain voltage as well; this is sometimes called drain-induced barrier lowering (DIBL).

24.10. Threshold Voltage Variation - the Narrow-Channel Effect (*Sec. 6.5.11*)

Quite confusing, also the narrow-channel effect is important to the derivation of V_T. In contrast to a short channel, where L is small, in a narrow channel W is small.

The narrow-channel effect is due to the influence of the two depletion regions which are always present outside the edges of the gate in the z-direction, as shown in Fig. 30(b). For large W these regions are not of importance, but as the width decreases the threshold voltage will start to increase at some point, due to the narrow-channel effect.

24.11. Breakdown in the MOS

If the drain voltage increases significantly, a short-channel MOS may approach a punch-through effect similar to what was discussed in Sec. 17.1 for the p-n junction. When punch-through is reached, the drain-to-substrate depletion region extends completely across the channel region to the source-to-substrate depletion region. Then the potential barrier between the source and the drain is fully eliminated and a very large drain current would exist.

Fig. 31: Parasitic bipolar action in an n-type MOS.

Punch-through is not the only breakdown that can happen. Prior to a punch-through condition is reached, the drain voltage can be so high that the reverse-biased diode (constituted by the drain-substrate depletion region) may start to avalanche. The current brought into the substrate from this diode continues to the substrate (body) contact, where it is drained out. Since the sub-

1. The actual fabrication techniques will be discussed in a later chapter.

strate has a certain resistance there will be a certain voltage drop between the source region and the substrate contact. When the current from the avalanche process has reached a certain size, there is a drop of say 0.6 V, which now turns on the diode between the substrate and the source regions. This diode is in forward-bias and therefore the parasitic bipolar transistor in Fig. 31, which always is present in the MOS structure, will start to conduct current.

There is a third type of breakdown that can take place in the MOS, oxide breakdown. When the electric field across SiO_2 becomes around 6×10^6 V/cm breakdown occurs, and this time the breakdown is **irreversible** and **catastrophic**. Since a safety margin of three is used for the oxide breakdown in MOS integrated circuit processes, we can sketch on the maximum supply voltage for a particular oxide thickness. Say that the thin oxide thickness is 500 Å, which is reasonable, then we have

$$V_{DD,\,max} = \frac{1}{3}\cdot 6\times10^{6}\cdot 500\times10^{-8} = 10\ \text{V}.$$

Indeed oxide breakdown is one of the determining factors for the supply voltage in a modern chip.

LECTURE 11

25. Bipolar Transistors - a First Encounter (*Sec. 7*)

The bipolar transistor has three separately doped regions and two p-n junctions, sufficiently close together so that interactions can take place between the junctions. Much of the theory of the p-n junction in the previous sections can be used in the analysis of the bipolar transistors.

25.1. Background

Already in 1874 the metal-semiconductor contact was known due to the work by Braun. This knowledge evolved into so-called point-contact diodes, where a metallic whisker touched a semiconductor surface. In 1906 a silicon-based point-contact diode was taken out as a patent. By 1935, selenium rectifiers and silicon point-contact diodes were available for use as radio detectors. With the development of the advanced radar at MIT, the need for detector diodes and mixers increased and during this time methods of achieving high-purity silicon and germanium were developed.

Doping of semiconductors was a known method to enhance conductivity since the end of the 1930s and subsequent progress had been made in the design of the point-contact diodes. History has it, that the bipolar transistor was invented when two such diodes were put together in 1947. On the day before Christmas Eve 1947 two proud researchers at the Bell telephone laboratories showed the world, embodied in a couple of Bell managers, the first ever solid-state transistor in operation, a primitive kind of bipolar transistor that was called point-contact transistor or type A transistor. In fact, this transistor was accidently discovered during research on field-effect transistors. A fantastic achievement the invention (or discovery) of the bipolar transistor was, and its impact on our lives has been more than tremendous.

It is worth noting that the p-n junction diode replaced the point-contact structure in the 1950s probably much due to the advancements in the design of BJTs.

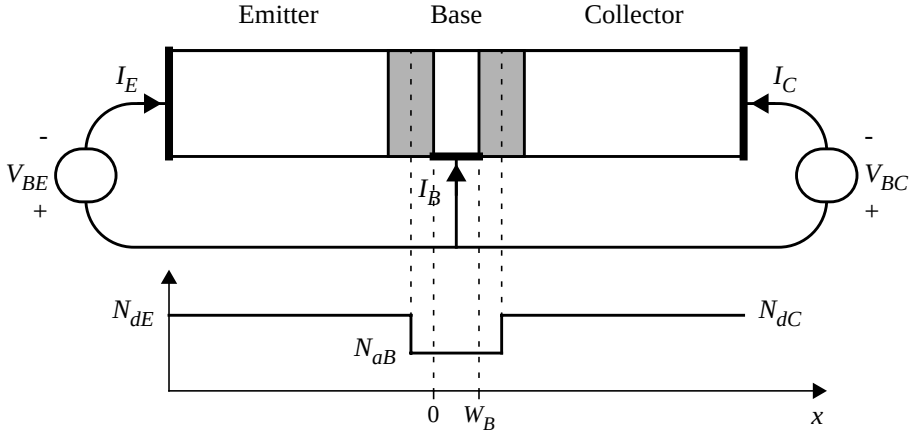


Fig. 32: An npn bipolar transistor.

The function of the bipolar transistor was so difficult to comprehend it was not until considerable time after the invention it was fully understood. The key problem was to understand that the injection of minority carriers was the fundament of the function - anyone heard of minority carriers? The bipolar transistor, and the struggle to understand it, should be compared to the field-effect transistor whose function was, by and large, already described in 1925, however, due to fabrication problems it was not possible to create such a transistor in practice until the 60s and by then advanced types of the bipolar transistor had become **the** transistor.

The three inventors of the bipolar transistor, William Shockley, John Bardeen, an American physicist born 1908, and Walter Brattain, an American physicist born 1902, were honored with the Nobel Prize for Physics in 1956. In retrospect, it seems Shockley was the ingenious driving force, while Bardeen provided mathematical excellence and Brattain was the practical wizard, when they did their pioneering work at Bell laboratories.

In 1949 Shockley presented a prototype transistor in a Bell journal, this was the first conception of the modern bipolar junction transistor. The prototype had abrupt junctions and constant base doping, such as the transistor shown in Fig. 32, which is not the case in modern BJTs. The particular material configuration in Fig. 32 is said to be an npn-type bipolar transistor. **NOTE** that by definition always the currents at the three terminals are flowing **into** the device and that the cross-sectional area *A*, as usual, is assumed to be uniform throughout the device.

25.2. A Sketch on the Current through the npn Bipolar Transistor (*Sec. 7.7.6*)

Under all bias conditions there will be very few holes flowing between the two p-n junctions in any direction, since the hole flow from either n-type region into the p-type base is very small. Thus, in the base region we can write

$$J_p = 0 = pq\mu_p E_x - qD_p \frac{dp}{dx} \Rightarrow E_x = \frac{D_p}{p\mu_p}\frac{dp}{dx} = \frac{kT}{q}\frac{1}{p}\frac{dp}{dx}$$

and

$$J_n = nq\mu_n E_x + qD_n \frac{dn}{dx} = nq\mu_n\left(\frac{kT}{q}\frac{1}{p}\frac{dp}{dx}\right) + qD_n\frac{dn}{dx}.$$

Thus, using Einstein's relationship again yields

$$J_n = \frac{n}{p}qD_n\frac{dp}{dx} + qD_n\frac{dn}{dx} = \frac{qD_n}{p}\left(n\frac{dp}{dx} + p\frac{dn}{dx}\right) = \frac{qD_n}{p}\frac{d}{dx}(pn),$$

which can be rearranged as

$$J_n\int_{x_1}^{x_2}\frac{p}{qD_n}\,dx = \int_{x_1}^{x_2}\frac{d}{dx}(pn)\,dx = p(x_2)\,n(x_2) - p(x_1)\,n(x_1).$$

The law of the junction in Eq. (41) allows us to describe a carrier product in terms of the applied bias, V_A:

$$np = n_i^2\, e^{\frac{qV_A}{kT}}.$$

Integrating over the base, which is assumed to have constant doping N_{aB} over the entire non-depleted width W_B, yields

$$J_{n,x} = \frac{p(W_B)\,n(W_B) - p(0)\,n(0)}{\int_0^{W_B}\frac{p}{qD_n}\,dx} = \frac{qD_n\,n_i^2}{N_{aB}\,W_B}\left(e^{\frac{qV_{BC}}{kT}} - e^{\frac{qV_{BE}}{kT}}\right). \qquad (62)$$

where we also have assumed a constant D_n over the entire base, which is a nice but not quite true simplification. The factor in the denominator $N_{aB}\ W_B$ represents the total doping of the base layer, and it is so important that it has its own name, the Gummel number.

We have now described the electron current flowing between the emitter and the collector. Apart from its usefulness in Sec. 27.2, this expression is interesting from several other views:

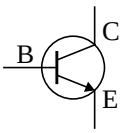
1. Obviously a short base leads to a high current. Thus, in practical situations the base is always minimized in width.

2. It is also obvious that if both junctions are reverse biased no current will flow. If, on the other hand, either V_{BE} or V_{BC} is positive and greater than kT/q, $J_{n,\,x}$ will be a sensitive function of the most positive voltage. The value of kT/q (the thermal voltage) at room temperature is 0.0259 V, which means that for an increase of 0.0259 V on the applied voltage, the change of current will be around 2.72. How's that for an exponential device?

25.3. Modes of Operation

In contrast to the p-n diode, since more than one current and one voltage are involved in the operation of the BJT, the device characteristics are inherently multidimensional. For the description to be tractable it is necessary to focus on the currents, voltages and polarities of primary interest in a particular application. This is accomplished by specifying the basic circuit configuration in which the device is connected and the biasing or operational mode. With reference to the npn transistor in Fig. 33, we can differentiate between four different modes of operation; active mode (or forward active mode), saturation mode, cut-off mode and inverted mode (or inverted active mode).

The active mode is the most common biasing arrangement, it is widely used in different kinds of amplifiers. The saturation and cut-off modes are exploited in digital circuits, but since the bipolar digital technologies are less used today than during the 70s, the 80s, and not to mention the 90s, I consider them to be of significantly less interest than the active mode. Since



V_{BC} negative V_{BE} positive	V_{BC} positive V_{BE} positive	V_{BC} negative V_{BE} negative	V_{BC} positive V_{BE} negative
(a)	(b)	(c)	(d)

Fig. 33: An npn bipolar transistor in the four different modes of operation. (a) Active mode, where the base-emitter junction is forward biased and the base-collector junction is reverse biased. (b) Saturation, where both junctions are forward-biased. (c) Cut-off, where both junctions are reverse biased. (d) Inverted mode, where the base-emitter junction is reverse biased and the base-collector junction is forward biased.

there is a big difference in doping of collector and emitter in a real device, the active and inverted modes are far from identical in behavior. Moreover, because the transistor at manufacturing is optimized with respect to the active mode, the inverted mode is almost only of academic interest.

26. Function of the Bipolar Transistor

26.1. Current Transport in the Active-Mode Scenario (*Sec. 7.1*)

The amplification behavior in the active mode is fairly simple to understand if the p-n diode

function was fully understood. We remember that when the p-n diode was reverse biased, hardly any current could flow from n to p because there were very few minority carriers present at the edges of the depletion region. If there was a way to insert, for example, excess electrons on the

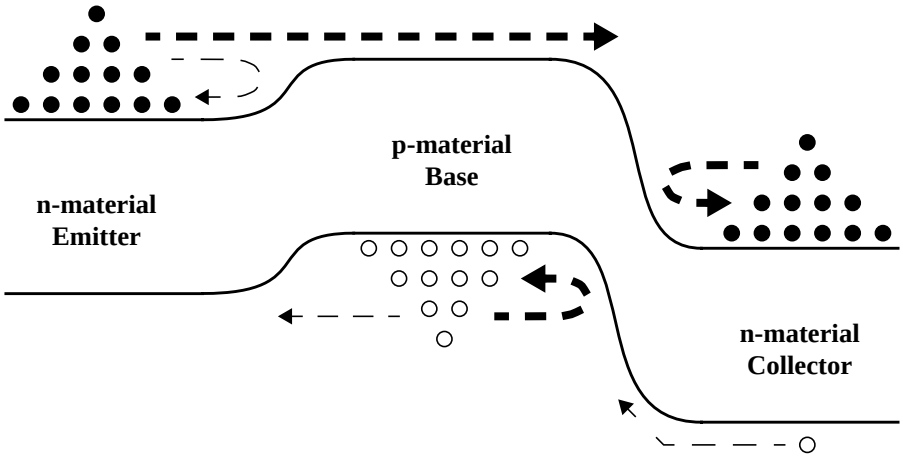


Fig. 34: Minority-carrier injection in a reverse-biased base-collector junction.

p-side of the depletion region these would be pulled by the electric field into the n-side and an electron current would flow. **NOTE** that the insertion of minority carriers on only one of the sides of the p-n diode implies that the current consists (primarily) only of either holes or electrons, **not both**.

Now, minority-carrier injection **is** accomplished in the npn bipolar transistor. Here one of the p-n junctions (the base-emitter junction) in the forward-biased case supplies a current of electrons into the p-type base where they become minority carriers. The flow of minority carriers across the base is due to the large gradient in the minority carrier concentration, i.e. diffusion, which is due to the reverse-biased base-collector junction and the narrow base width. The reverse bias across the base-collector junction finally pulls the minority carriers into the collector by virtue of the drift mechanism. The process is summarized in Fig. 34.

The p-type base is as narrow as possible in order to minimize the recombination in the base and maximize the electron current as shown in Eq. (62). Also, to reduce the hole diffusion current from base to emitter[1], the emitter is heavily doped in comparison to the base. This is important to remember!

Let us have a pause and notice the following: In an npn transistor we have a current primarily composed of electrons. In a pnp transistor the current is primarily carried by holes. This difference has great impact on the performance; since the mobility and the diffusion constant for electrons are higher than for holes, the npn-type BJT is superior to the pnp-type in terms of speed.

Obviously, we have currents in the emitter and collector terminals, but what about the base current, what is that? Up until now, we have used the base terminal as a convenient reference for bias voltages, but this terminal is important in another way. When operated in the active mode the biases lead to an electron transport from emitter to collector, whereby the electrons traverse the base region. During the traversal some electrons are lost due to recombination. These lost ones are not many as the region is usually much more narrow than the diffusion length, but the

1. It is a bit difficult to motivate this requirement at this point - take my word for it, it leads to higher so-called current gain.

recombination process tiny as it is will nevertheless affect the space charge neutrality condition. Holes will disappear, after recombining with a few of the travelling electrons, and this has to be compensated for by a hole current into the base that replenishes the lost holes. Thus, a current into the base is necessary, a current which is marked I_{RB} in Fig. 35.

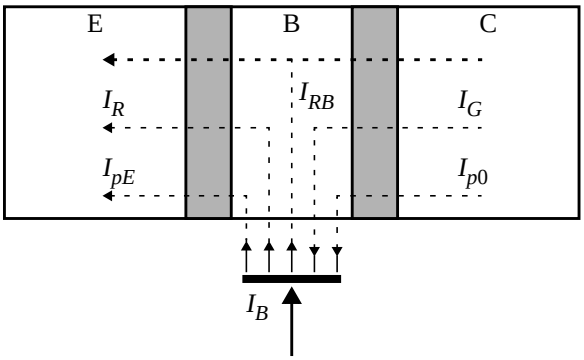


Fig. 35: Current components in an npn bipolar transistor operating in the active mode.

It is important also to recognize the impact on base current of several other processes:

First and foremost, since the base-emitter junction is forward biased, not only electrons will diffuse from the n-type emitter into the p-type base, but also some holes will diffuse in the opposite direction. This hole current, which is denoted I_{pE} in Fig. 35, has to be compensated for by an increase in base current. **In fact, in modern-day devices I_{pE} is larger than I_{RB}.** Increasing both emitter and base current with a constant, leads to a reduction in the ratio between emitter and base current which is the same as a reduction in current gain - an important performance parameter in BJTs. This ratio was maximized by using a heavily doped emitter and a lightly doped base.

Secondly, we have the current I_{p0} which is due to the reverse-biased base-collector junction. By the injection of minority carriers to the p-type base it is possible to increase the electron flow from base to collector, but the hole current in the opposite direction has not changed at all. The reverse saturation current stems from Eq. (49)

$$I_0 = qA\left[\frac{D_p}{L_p}\,p_{0,\mathrm{n}}(-x_\mathrm{n}) + \frac{D_n}{L_n}\,n_{0,\mathrm{p}}(x_\mathrm{p})\right] .$$

What happens in the reverse-biased junction between base and collector is that the base-emitter diode causes an increase in $n_{0,\mathrm{p}}(x_\mathrm{p})$, while $p_{0,\mathrm{n}}(-x_\mathrm{n})$ is left untouched. However small it is, the reverse saturation current due to holes drifting from the n-type collector, after being thermally generated inside the collector, into the p-type base reduces the base current that is needed for compensating the holes recombining with travelling electrons. (Of course, electrons too are drifting across the reverse-biased junction, from the base to the collector. This has, however, no relation to our discussion as the electrons are majority carriers in the base.)

For the third, we should pay attention to something that happens in the two depletion regions. In Sec. 16.1 and Sec. 16.2 we saw that in the depletion region in a reverse- and forward-biased junction there is a process of generation and recombination, respectively, going on. This means that in the active mode, electron-hole pairs are generated in the base-collector junction, while electron and holes are recombining in the base-emitter junction. Thus, there will exist currents I_G and I_R caused by these R-G processes, the first reduces the base current while the latter increases it.

27. Equivalent Circuit Models and BJT Performance

27.1. Amplification Performance Parameters of Bipolar Transistors (*Sec. 7.2*)

The mathematics introduced up until was included for the reason to give insight in some mechanisms present in the bipolar transistor. When the BJTs are used in a real circuit, however, designers tend to look for engineering formulae that relate physical characteristics to performance figures. In amplification situations obviously the amplification is of great interest to a designer so we have to take a look at such parameters.

In the case of an npn-type BJT, the proportionality factor B, the base transport factor, denotes the fraction of injected electrons which make it across the base to the collector

$$i_C = B\, i_{En}.$$

Furthermore, the emitter injection efficiency denotes how much of the total emitter current, including both holes I_{Ep} and electrons I_{En}, is made up of electrons

$$\gamma = \frac{i_{En}}{i_E}.$$

Based on these two factors we can quite quickly formulate a parameter that relates the three terminal currents according to the two most common circuit configurations, namely the common-base and the common-emitter circuits. The current transfer ratio or common-base current gain is

$$\frac{i_C}{i_E} = \frac{B\, i_{En}}{i_E} = B\gamma = \alpha \tag{63}$$

and the base-to-collector current amplification or common-emitter current gain is

$$\frac{i_C}{i_B} = \frac{B\, i_{En}}{i_{Ep} + (1-B)\, i_{En}} = \frac{B\, i_{En}}{(i_{Ep} + i_{En}) - B\, i_{En}} = \frac{\alpha}{1-\alpha} = \beta.$$

27.2. A Simplified Deduction of the Ebers-Moll Equations (*Sec. 7.5.1*)

The Ebers-Moll equations form the basis for one kind of equivalent circuit, in which a bipolar transistor is replaced by current sources and p-n diodes. These equations find use in, for example, circuit simulators like SPICE for establishing d-c operating conditions and BJT device characteristics. Usually the Ebers-Moll equivalent circuit is used in large-signal applications, while the hybrid-π equivalent circuit in Sec. 27.3 is used in amplification (small-signal) applications.

We will deduce the Ebers-Moll equations by **not** taking recombination into account. Now, we can use the law of junction to sketch the carrier densities as

$$\frac{n_{BE}}{n_{B0}} = e^{\frac{qV_{BE}}{kT}} = \frac{p_{EB}}{p_{E0}},$$

where, for example, n_{BE} and n_{B0} denotes the electron concentration at the side of the (p-type) base nearer the (n-type) emitter and the equilibrium electron concentration of the base, respectively.

Similarly, for the base-collector junction we can write

$$\frac{n_{BC}}{n_{B0}} = e^{\frac{qV_{BC}}{kT}} = \frac{p_{CB}}{p_{C0}}.$$

Based on the minority-carrier density graph in Fig. 36 and the equation for diffusion current density, the minority-carrier current densities, defined with respect to the *x*-axis, in the three respective regions[1] can now be written as

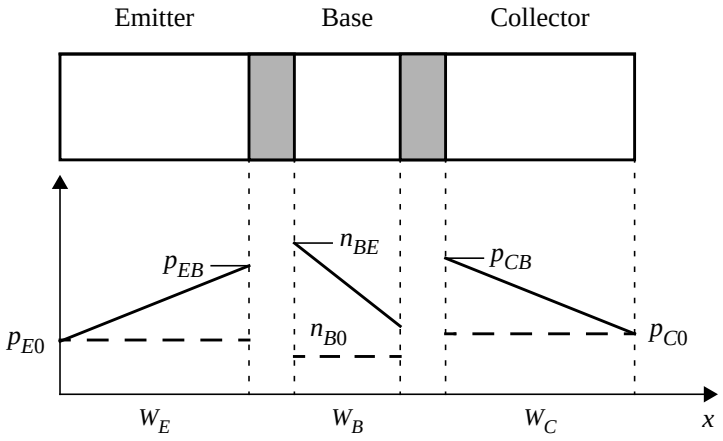


Fig. 36: The npn minority-carrier model for the Ebers-Moll analysis. The width of the base region, which is substantially below 1 μm in contemporary processes, is sufficiently smaller than the diffusion length L_n. Consequently, the minority carriers in the base region display a linear gradient since there is almost no recombination taking place in the base. For simplicity of the derivation, also the minority carrier distributions in the emitter and the collector are drawn as linear functions, although they in reality fall of exponentially. This simplification however does not alter the final structure of the Ebers-Moll equations.

$$J_{nB} = qD_n \frac{dn_B}{dx} = qD_n \frac{n_{BC} - n_{BE}}{W_B} = \frac{qD_n\, n_{B0}}{W_B}\left(e^{\frac{qV_{BC}}{kT}} - e^{\frac{qV_{BE}}{kT}}\right)^{(2)}, \tag{64}$$

$$J_{pE} = -qD_p \frac{dp_E}{dx} = -qD_p \frac{p_{EB} - p_{E0}}{W_E} = -\frac{qD_p\, p_{E0}}{W_E}\left(e^{\frac{qV_{BE}}{kT}} - 1\right),$$

and

$$J_{pC} = -qD_p \frac{dp_C}{dx} = -qD_p \frac{p_{C0} - p_{CB}}{W_C} = \frac{qD_p\, p_{C0}}{W_C}\left(e^{\frac{qV_{BC}}{kT}} - 1\right).$$

Since we assumed that no recombination takes place in the base (i.e., the linear description of $n(x)$ in the base) the electron current in the base must also be the electron current in both the emitter and in the collector. Thus, the emitter and the collector currents can be written as

$$I_E = -A\left[\frac{qD_p\, p_{E0}}{W_E}\left(e^{\frac{qV_{BE}}{kT}} - 1\right) + \frac{qD_n\, n_{B0}}{W_B}\left(e^{\frac{qV_{BE}}{kT}} - e^{\frac{qV_{BC}}{kT}}\right)\right]$$

and

1. Even though Fig. 36 is very specific, the equations will be general and valid for all operation modes.
2. Compare with Eq. (62) - any similarities?

$$I_C = -A\left[\frac{qD_p\, p_{C0}}{W_C}\left(e^{\frac{qV_{BC}}{kT}}-1\right)+\frac{qD_n\, n_{B0}}{W_B}\left(e^{\frac{qV_{BC}}{kT}}-e^{\frac{qV_{BE}}{kT}}\right)\right]^{(1)}.$$

Now it's time for some tedious algebra. To make the equations reasonable in size we first assign a new definition to the terms in front of the exponential expressions such that

$$K_B = A\,\frac{qD_n\, n_{B0}}{W_B}\ ,\ K_E = A\,\frac{qD_p\, p_{E0}}{W_E}\ \text{ and } K_C = A\,\frac{qD_p\, p_{C0}}{W_C}\ .$$

Now I_E and I_C can be rearranged as

$$I_E = -\left[(K_E+K_B)\left(e^{\frac{qV_{BE}}{kT}}-1\right)-\left(\frac{K_B}{K_C+K_B}\right)(K_C+K_B)\left(e^{\frac{qV_{BC}}{kT}}-1\right)\right]$$

and

$$I_C = -\left[(K_C+K_B)\left(e^{\frac{qV_{BC}}{kT}}-1\right)-\left(\frac{K_B}{K_E+K_B}\right)(K_E+K_B)\left(e^{\frac{qV_{BE}}{kT}}-1\right)\right].$$

Making a final adjustment in that we replace the *K*s again, brings us to

$$I_E = -\left[I_{ES}\left(e^{\frac{qV_{BE}}{kT}}-1\right)-\alpha_R\, I_{CS}\left(e^{\frac{qV_{BC}}{kT}}-1\right)\right]$$

and

$$I_C = -\left[I_{CS}\left(e^{\frac{qV_{BC}}{kT}}-1\right)-\alpha_F\, I_{ES}\left(e^{\frac{qV_{BE}}{kT}}-1\right)\right],$$

where I_{ES} and I_{CS} is the reverse saturation current for the emitter and the collector junction, respectively, and α_R and α_F is the common-base current gain[2] in the inverse-active mode and the common-base current gain in the forward-active mode, respectively. These are the Ebers-Moll equations and they equip us with the tool of an equivalent circuit. Before writing down the circuit, the equations are slightly modified:

Since

$$I_{ES}\left(e^{\frac{qV_{BE}}{kT}}-1\right) = \alpha_R\, I_{CS}\left(e^{\frac{qV_{BC}}{kT}}-1\right)-I_E$$

we can write

$$I_C = -\left[I_{CS}\left(e^{\frac{qV_{BC}}{kT}}-1\right)-\alpha_F\left\{\alpha_R\, I_{CS}\left(e^{\frac{qV_{BC}}{kT}}-1\right)-I_E\right\}\right].$$

1. Now we have switched to current and also to the flow convention of Fig. 32.
2. α was defined in Eq. (63).

By introducing I_{C0}, the saturation current for the collector when $I_E = 0$, which is defined as

$$I_{C0} = (1 - \alpha_F \alpha_R)\, I_{CS}$$

the final form of the collector current equation is

$$I_C = -\left[\alpha_F\, I_E + I_{C0}\left(e^{\frac{qV_{BC}}{kT}} - 1\right)\right].$$

Similarly, with the use of

$$I_{E0} = (1 - \alpha_F \alpha_R)\, I_{ES}$$

we can write

$$I_E = -\left[\alpha_R\, I_C + I_{E0}\left(e^{\frac{qV_{BE}}{kT}} - 1\right)\right]$$

and we now have all parameters we need for drawing a useful equivalent circuit for the transistor. The drawing is presented in Fig. 37.

Now, the Ebers-Moll equivalent circuit is not a perfect model, no it is far from that. There are a number of simplifications that are done in order to find a closed analytical expression, thus yielding an unrealistic behavior under some operational conditions. Another popular equivalent circuit model is the Gummel-Poon model[1], which was developed so as to avoid some of the Ebers-Moll's requirements for operational conditions. The Gummel-Poon model e.g. accounts for high-injection levels in a semi-empirical way. Still, we perhaps appreciate the Ebers-Moll model when we are about to use the model, since we only need to input three parameters in this model[2], whereas Gummel-Poon requires more than 30 parameters!

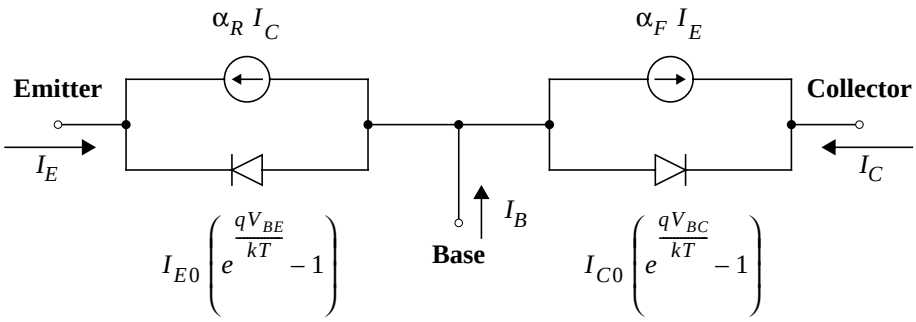


Fig. 37: The equivalent circuit based on the Ebers-Moll equations.

1. The first steps in obtaining the Gummel-Poon model were in fact taken in Sec. 25.2. More info on the Gummel-Poon model can be found in *Sec. 7.7.6* in the main textbook.
2. α_R, α_F, and only one saturation current.

LECTURE 12

27.3. The Hybrid-π Equivalent Circuit (*Sec. 7.8.1*)

For BJTs used in amplifying circuits there exist more than one equivalent circuit. The hybrid-π (or hybrid-pi) model is especially suited for lecturing purposes as it conveys some information on how dynamic performance parameters relate to the physical attributes of the transistor. The name π stems from how some resistances in one of the many versions of this type of equivalent circuit are organized. In Fig. 38(a) a hybrid-π equivalent circuit with minimized complexity is

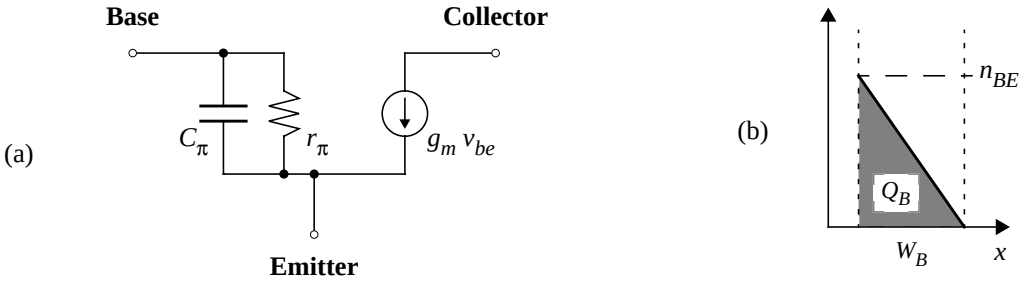


Fig. 38: (a) The first three elements of the hybrid-π equivalent circuit. (b) Density gradient and excess charge in the base region of an npn transistor.

shown(1). The base-emitter characteristics are important for performance while the base-collector is not mentioned anywhere, which is due to the independence of current on base-collector voltage. The potential difference of the reverse-biased junction has no impact on current as we have stated numerous times during this course(2).

The transconductance, g_m, is a kind of a-c correspondence to the common-emitter current gain. Similarly to the case of the MOS in Sec. 23.1, the transconductance of the BJT is defined as

$$g_m = \frac{dI_C}{dV_{BE}}$$

where, under the assumption that the BJT is in the active mode, the collector current consists of electrons due to diffusion in the base. I_C thus is on a similar form as Eq. (64), with the difference that $n_{BC} = 0$ and that I_C is defined with respect to the terminal, not to the x-axis:

$$I_C = -qA\,D_n\,\frac{0 - n_{BE}}{W_B} = \frac{qA\,D_n\,n_{B0}}{W_B}\,e^{\frac{qV_{BE}}{kT}}. \qquad (65)$$

Here we **NOTE** that evaluating g_m obviously leads to

$$g_m = \frac{q}{kT}\,I_C.$$

Now back to analyzing I_C; according to Fig. 38(b) we can describe the excess charge as a triangle

1. We avoid the extrinsic resistances and capacitances used in *Fig. 7-24* on page 366, but deal only with the fundamental, intrinsic properties of the BJT.
2. Of course, non-ideal effects, such as breakdown, can occur at the base-collector junction.

$$Q_B = \frac{1}{2} qA\, n_{BE}\, W_B,$$

from which

$$n_{BE} = \frac{2Q_B}{qA\, W_B},$$

can be inserted in Eq. (65) above resulting in

$$I_C = \frac{qA\, D_n}{W_B} n_{BE} = \frac{qA\, D_n}{W_B}\left(\frac{2Q_B}{qA\, W_B}\right) = \frac{Q_B}{(W_B{}^2/2D_n)}. \tag{66}$$

The triangle approximation of the base charge is a bit wrong as there is a time-varying contribution to the charge, i.e. n_{BE} depends on the small-signal behavior of the base-emitter voltage[1]. However, in the Lecture notes this time-varying contribution has not been included as it may hide the really essential parts of the discussion.

Now, since current is a rate of flow of charge, the final denominator of the previous expression is the time taken for Q_B to flow through the base, and is given the symbol τ_t, the minority-carrier transit time:

$$I_C = \frac{Q_B}{(W_B{}^2/2D_n)} = \frac{Q_B}{\tau_t}.$$

The excess charge can also be used for finding the base current in this analysis which takes place in an npn transistor. The d-c stored charge in the base is $I_B\, \tau_n$, where τ_n is the time an average excess hole supplied from the base spends in the base ensuring space charge neutrality during the lifetime of an average excess electron which traverses the base[2]. Thus, I_B is

$$I_B = \frac{Q_B}{\tau_n}.$$

Substituting Q_B with Eq. (66) yields

$$I_B = \frac{I_C\, W_B{}^2}{2D_n\tau_n},$$

while a subsequent insertion of I_C from Eq. (65) leaves us at

$$I_B = \left(\frac{qA\, D_n\, n_{B0}}{W_B}\, e^{\frac{qV_{BE}}{kT}}\right)\frac{W_B{}^2}{2D_n\tau_n} = \frac{qAW_B}{2\tau_n}\, n_{B0}\, e^{\frac{qV_{BE}}{kT}}.$$

By definition, the input resistance is

1. *Eq. (7-71)* indicates the relationship for a pnp transistor.
2. Now we are considering recombination in the base.

$$\frac{1}{r_\pi} = \frac{dI_B}{dV_{BE}} = \frac{q^2 A W_B}{2\tau_n}\frac{n_{B0}}{kT} e^{\frac{qV_{BE}}{kT}} = \frac{q}{kT} I_B,$$

and thus

$$r_\pi = \frac{kT}{q}\frac{1}{I_B}.$$

Both g_m and r_π can obviously be evaluated from the d-c operating conditions. We should pay attention to the following here: Since we have

$$I_C = \frac{kT}{q} g_m$$

and

$$I_B = \frac{kT}{q}\frac{1}{r_\pi},$$

we can use the identity

$$Q_B = I_B \tau_n = I_C \tau_t \quad (1)$$

to formulate a relationship between the transconductance and the input resistance

$$\frac{1}{r_\pi} = g_m \frac{\tau_t}{\tau_n}.$$

Now, finally in the discussion on the hybrid-π model we quickly derive the capacitance C_π as

$$C_\pi = \frac{dQ_B}{dV_{BE}} = \frac{d}{dV_{BE}}(I_C \tau_t) = \frac{d}{dV_{BE}}\left(\frac{qA D_n n_{B0}}{W_B} e^{\frac{qV_{BE}}{kT}}\right)\tau_t = \frac{q}{kT} I_C \tau_t.$$

27.4. High-Frequency Operation (*Sec. 7.7.1, 7.8.2*)

The bipolar transistor has today been outstripped by the MOSFET, in all application areas but the ones that call for extremely high operation frequencies or a good capability of driving loads. The high operation frequency that is achieved in a BJT, is due to the narrow base and the current-driven function, and an estimate of the maximal frequency can be obtained by analyzing the transit time through the transistor.

In Eq. (66) we made use of the minority-carrier transit time through the base and this is in fact the dominant delay for the carriers when going from emitter to collector. There is no delay inside the quasi-neutral regions of emitter and collector because the applied biases have an immediate impact all the way right up to the boundaries of the depletion regions. In a complete delay picture there are also delay contributions from the charging time of both the base-emitter junction capacitance(2) and the collector capacitance(3) and the transit time of the base-

1. **NOTE** that this also yields $\beta = I_C/I_B = \tau_n/\tau_t$.
2. The base-emitter junction capacitance is high because the junction is forward biased.
3. The base-collector junction capacitance is very small because the junction is reverse biased. However, there is a capacitance between collector and the so-called substrate, in which the transistor sits.

collector depletion region[1]. Thus, capacitance must be minimized for high operational frequencies and this is achieved by making the transistors small in size.

The amplification available from a BJT falls at high frequencies. Similarly to the high-frequency discussion on the MOS in Sec. 23.2, we can calculate the cut-off frequency, where the magnitude of the current gain in the common-emitter circuit with zero load impedance is unity:

$$f_T = \frac{1}{2\pi\,\tau_t} = \frac{1}{2\pi\left(\frac{W_B^{\,2}}{2D_n}\right)} = \frac{D_n}{\pi\,W_B^{\,2}}.$$

f_T can be increased by using a certain non-uniform base doping profile. The always-present built-in field in the base, which will be the result of such non-uniform doping, assists the motion of minority carriers such that the transit across the base is reduced significantly, with say a factor of 4. **NOTE** that the minority-carrier transit time across the base is similar to the channel transit time in the MOS.

Ex. 11: Frequency-Dependent Current Gain - An Example

Assignment:

Based on the hybrid-π equivalent circuit in Fig. 38(a), determine the frequency at which the small-signal current gain decreases to $1/\sqrt{2}$ of its low-frequency value. What is this frequency if C_π = 10 pF and r_π = 3 kΩ?

Solution:

Previously the transconductance was defined as

$$g_m = \frac{dI_C}{dV_{BE}},$$

with reference to Fig. E.7.

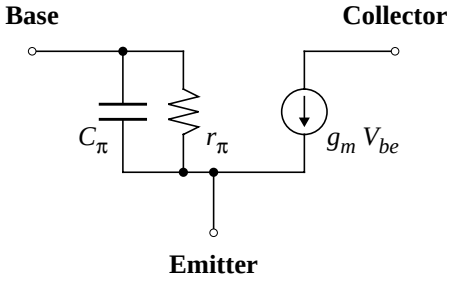


Fig. E.7: The first three elements of the hybrid-π equivalent circuit.

At very low frequency, C_π can be neglected and then we have

$$V_{be} = I_b\, r_\pi.$$

1. The base-collector depletion region is wide because the junction is reverse biased.

From the equivalent circuit we moreover have

$$I_c = g_m V_{be} = g_m r_\pi I_b$$

or rather

$$\beta_{lowf} = \left.\frac{I_c}{I_b}\right|_{\text{low frequency}} = g_m r_\pi .$$

Now we have a measure on current gain at a low frequency; let us look to higher frequencies.

Increasing the frequency *f*, means that C_π **cannot** be neglected anymore. Expressing V_{be} has to be done with the inclusion of C_π.

$$V_{be} = I_b \left(\frac{r_\pi \frac{1}{j\omega C_\pi}}{r_\pi + \frac{1}{j\omega C_\pi}} \right) = I_b \left(\frac{r_\pi}{1 + j\omega r_\pi C_\pi} \right).$$

This gives in turn

$$I_c = g_m I_b \left(\frac{r_\pi}{1 + j\omega r_\pi C_\pi} \right) = I_b \left(\frac{\beta_{lowf}}{1 + j\omega r_\pi C_\pi} \right)$$

and

$$\beta_{highf} = \left.\frac{I_c}{I_b}\right|_{\text{high frequency}} = \frac{I_b \left(\frac{\beta_{lowf}}{1 + j\omega r_\pi C_\pi} \right)}{I_b} = \frac{\beta_{lowf}}{1 + j\omega r_\pi C_\pi} .$$

Thus,

$$\beta_{highf} = \frac{1}{\sqrt{2}} \beta_{lowf} \Rightarrow \frac{\beta_{lowf}}{1 + j\omega r_\pi C_\pi} = \frac{1}{\sqrt{2}} \beta_{lowf} ,$$

which is solved for *f*:

$$\sqrt{1 + (2\pi f \, r_\pi C_\pi)^2} = \sqrt{2} \Rightarrow f = \frac{1}{2\pi \, r_\pi C_\pi} .$$

Numbers inserted gives a frequency, f = 5.3 MHz.

28. Non-Ideal Effects in the Bipolar Transistor (*Sec. 7.7*)

The bipolar transistor shares several non-ideal effects with the p-n diode, such as avalanche breakdown and high-level injection. In the BJT there also exist unique non-ideal effects due to the more complex structure:

The effective width of the base is reduced when we decrease (the already negative) V_{BC}, i.e. increase the reverse bias across the base-collector junction. This effect is called base-width modulation or the Early effect[1]. It follows from Eq. (62) that the result of the Early effect is that

1. After James Early who did the discovery in 1952. Read more on this in *Sec. 7.7.2*.

the collector current increases rather than staying constant with increasing voltage on the collector. When the depletion region fills the base we reach punch-through, exactly as in the case of the p-n diode.

The physical structure of a fabricated BJT in conjunction with the need for a lightly doped base region implies a quite high resistance in the base, on the order of some hundreds of ohms (which is to be compared to the emitter which has a resistance of something like one ohm). When the base resistance is increasing a number of parameters will change: V_{BE} needed for a particular I_C increases, emitter crowding[1] becomes worse, and the common-emitter current gain, β, increases.

The emitter usually is quite heavily doped to increase the emitter injection efficiency, γ. However, as said in the context of degenerate semiconductors (Sec. 4.1) heavy doping can lead to bandgap narrowing which influences the emitter injection efficiency; γ begins to fall off rather than continuing to increase with increased emitter doping[2].

Another problem that has to do with reduced emitter injection efficiency is when we have high-level injection in the base, which comes about when V_{BE} is increased enough. Then the excess majority carriers are comparable in number with, or even greater than, the equilibrium majority carriers. When this condition holds, the diffusion of majority carriers from the base into the emitter will be significantly higher than at low-level injection and γ will consequently be lower.

NOTE that the two effects related to the emitter injection efficiency counteract the γ-promoting concept of heavily doped emitter and lightly doped base.

This a very pedagogical ending before we take a short look at the Heterojunction Bipolar Transistor (HBT).

29. The Heterojunction Bipolar Transistor (*Secs 5.8, 7.9*)

The main feature of the HBT is that it has an excellent emitter injection efficiency thanks to a base-emitter heterojunction that facilitates (in an npn case) electron diffusion from emitter to base, while blocking hole diffusion from base to emitter. This fact relaxes the condition for high emitter injection efficiency in the BJT, i.e. heavy emitter doping and light base doping. Thus, the doping levels can be steered toward other goals, such as reduced base resistance, by heavy base doping, and reduced junction capacitance, by light emitter doping.

From the condensed description given in most places with regard to HBTs one can be excused to think that the HBT device is nothing but a minor player. HBTs do, however, exist in "real" applications, such as the integrated InP (indium-phosphide) HBTs that Ericsson uses for 40 Gb/s de-multiplexing[3]. Also, SiGe HBTs are offered by IBM for high-frequency applications[4].

The HBT apparently can cope with high frequencies. The reason behind its high-speed

1. Read more on this in *Sec. 7.7.5*.
2. Read more on this in *Sec. 7.9*.
3. Read more about this in "InP-HBT Chip-Set for 40-Gb/s Fiber Optical Communication Systems Operational at 3 V" by M. Mokhtari, *et al*, in the *IEEE Journal of Solid-State Circuits*, vol. 32, no. 9, p. 1371, Sept. 1997.
4. http://www.research.ibm.com/sigetech/

operation is that the electrons going into the base originate from an emitter made of a high band-gap material, which suggests the carriers have high energy and thus high velocity. The cut-off frequency for silicon BJTs is of the order of 50-300 MHz, whereas f_T for high-speed BJTs could be 10-40 GHz. This is to be compared to a high-speed HBT, where f_T can be as large as 300 GHz!

LECTURE 13

30. Optoelectronic Devices

In 1873 the British engineer W. Smith discovered that the resistivity of selenium varied with light. That was possibly the first time people encountered the strange features of semiconductors, and it had to do with optoelectronic or photonic devices. Today, there are a number of devices that link photons - *the optical domain* - and charge carriers - *the electronics domain*. In these Lecture notes mainly the light-emitting diode and photodetectors will be discussed as they are only briefly mentioned in the main textbook[1]. Also, a brief perspective is given on solar cells.

Photoluminescence (*Sec. 4.2.1*) and cathodoluminescence are of course fascinating phenomena, but they do not really belong to device-oriented semiconductor technology. Charge-coupled devices (*Sec. 9.4*) are important as they form the basis for imaging in video cameras, but here Streetman offers a comprehensive treatment.

31. Electroluminescence - the Light-Emitting Diode (*Secs 4.2.2, 8.2*)

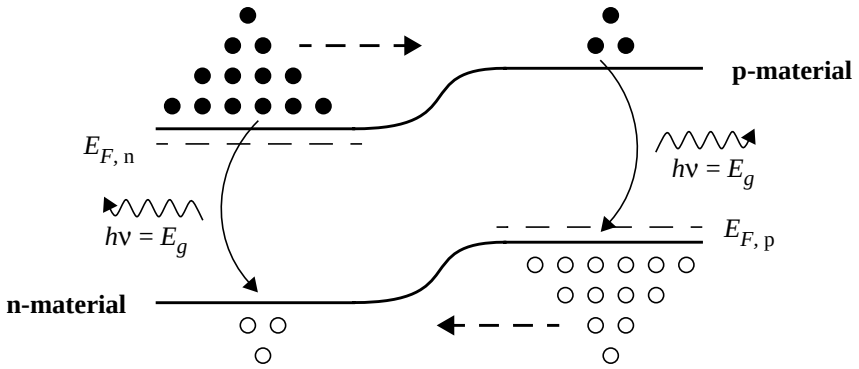


Fig. 39: A forward-biased p-n junction with recombination processes in action.

In fact, the phenomenon behind the Light-Emitting Diode (LED) was reported by Round already in 1907; then in the form of a silicon carbide (SiC)[2] point-contact rectifier. In a junction LED, photons of near-bandgap energy are generated by the process of electroluminescence, in which a large number of electrons, injected into a normally empty conduction band, recombine with holes in the valence band as shown in Fig. 39. Since we have understood the function of both p-n junctions and bipolar transistors, it is obvious that most of the recombination takes place at the edges of the depletion region and that the recombination falls of as an exponential function as the minority carrier electrons approach the right end of the diode in Fig. 39. Of course, current can also be transported by holes from the p-type to the n-type material, but that direction of current is usually minimized by asymmetrical doping[3].

The reason for having recombination taking place only in one of the two quasi-neutral regions is due to the physical structure of LEDs. These are often manufactured in a diffused

1. For you who are interested in in-depth coverage, attend the course in TFFY22 Optoelectronics, which is given in period 3.
2. Carborundum is another name for silicon carbide.
3. This consideration is very similar to that of emitter injection efficiency of a BJT.

structure, which is depicted in Fig. 40. As the n-type material is hidden within the diode, it is

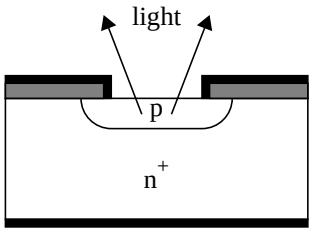


Fig. 40: Physical structure of surface-emitting LED manufactured by a diffusion process.

inefficient to allow for photon emission in the bulk of the diode. This is because the peak in the density distribution of electrons, $N_c(E)\,f(E)$, in the conduction band of a non-degenerate semiconductor occurs at an energy $kT/2$ above E_c [1]. This particular displacement of the maximum from the bandgap edge is easily found from a differentiation with respect to the energy of

$$N_c(E)\,f(E) = \left[\frac{4\pi\,(2m_{dn}^*)^{3/2}\,(E-E_c)^{1/2}}{h^3}\right] e^{\frac{-(E-E_F)}{kT}}.$$

The derivative is set to zero, and the following falls out:

$$\frac{1}{2}(E_{n_{max}}-E_c)^{-1/2} - \frac{1}{kT}\,(E_{n_{max}}-E_c)^{1/2} = 0,$$

or

$$E_{n_{max}} = E_c + \frac{kT}{2}.$$

Anyway, this has the effect that photons that are generated by band-to-band recombination in direct bandgap materials can be reabsorbed as they travel through the material because they possibly have an energy above the bandgap energy. This leads to a reduction in the radiative efficiency of the LED.

One remedy to reabsorption is to use impurities to form energy levels within the bandgap, i.e. R-G centers (Sec. 6.4). Emission due to recombination from R-G centers gives photons with less energy than the bandgap and few photons are reabsorbed. However, the R-G centers must be shallow in that they are not too far away from the bandgap edges, otherwise the emitted light may have too large a wavelength.

Another strategy to reduce reabsorption is to use heterojunction diodes, where materials are ordered according to variation in bandgap such that reabsorption is small (Fig. 41).

There exists an interesting application of the Heisenberg uncertainty principle (a famous part of the quantum mechanics, that was treated in Modern Physics) in relation to some LEDs with indirect bandgaps. As was told in Sec. 6.4, if we are seeking photon- instead of phonon-generating recombination, the material used has to either have a direct bandgap or contain impurities that can act as R-G centers. When a carrier is trapped in a recombination center it is spatially localized. We can write

1. See *Fig. 3-16* in the main textbook for an illustration.

$$\Delta x \cdot \Delta p > \frac{h}{2\pi},$$

where Δx is decreased due to the attraction to the impurity. The decrease in Δx, increases the spread in momentum, which is related to the *k*-space by

$$p = \frac{h}{2\pi} k.$$

Now a significant rate of transitions between the R-G centers and the band edges are allowed **without** the involvement of phonons. The change in momentum can be attributed to the uncertainty principle!

31.1. Some of the Most Common LEDs

Some of the most successful LED types for visible light are the following:

- **$GaAs_{0.6}P_{0.4}$**: This is a direct bandgap material which produces red light. LEDs based on this material were introduced already in the early 70s. The efficiency η[1] of this LED is about 0.2%.
- **$GaAs_{0.35}P_{0.65}$:N, $GaAs_{0.14}P_{0.86}$:N and GaP:N**: These are all indirect bandgap materials[2], but due to the introduction of nitrogen as a replacement for phosphorous in some lattice sites (denoted by **:N**), and thus the use of the uncertainty principle, they can produce orange-red, yellow and green light, respectively. For these materials η is 0.7%, 0.2% and 0.4%, respectively.
- **AlGaAs**: This is a direct bandgap material that can produce an intensive red light. LEDs based on this material are very efficient due to the perfection of lattice-matched layers[3] and the direct recombination. To avoid reabsorption a LED based on this material is grown epitaxially according to Fig. 41. An η between 4% and 16% can be obtained. The obvious application is e.g. as tail lights on bicycles or cars or as heel lights on athletic tennis shoes!

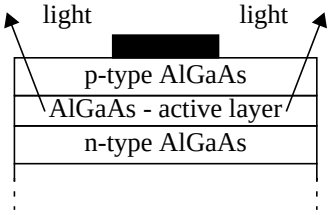


Fig. 41: Physical structure of an AlGaAs LED made by epitaxy. The active layer is a narrow region, with smaller bandgap than the layer above and below, to reduce reabsorption. To achieve the desired difference in bandgaps, the proportion of Al is larger at the expense of Ga in the confining layers.

- **SiC and GaN**: The first is an indirect bandgap IV-IV material with E_g = 2.9 eV whereas the second is a direct bandgap III-V material with E_g = 3.4 eV. These materials are both

1. η = photo power out/electrical power in.
2. $GaAs_{1-x}P_x$ is an indirect bandgap material for $x > 0.45$, but direct otherwise!
3. AlGaAs has almost the same lattice constant as GaAs has, and therefore the two materials are said to be lattice matched.

used for designing **blue** LEDs, a development done as late as during the 90s because of the material problems(1) associated with the wide bandgap materials that are needed for producing light with high frequencies(2). GaN is leading the race, because of the higher efficiency proved in practical LEDs, and a sketch of such a LED is given in Fig. 42.

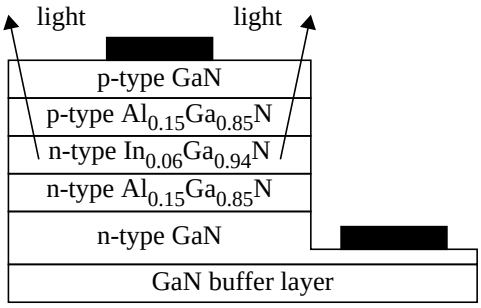


Fig. 42: A so-called double heterostructure based on GaN implementing a blue LED.

32. Photogeneration - Photodetectors and Solar Cells (*Secs 4.1, 8.1*)

The photoconductor in Sec. 12, which is fabricated in one piece of homogeneously doped material, is based on the principle that photons excite electrons so that excess carriers can enhance the conductivity (see Example 5). Thus, the photoconductor is a photodetector. However, its so-called dark current and the associated thermal noise(3) (see Example 2) make it unsuitable for high-performance communication applications.

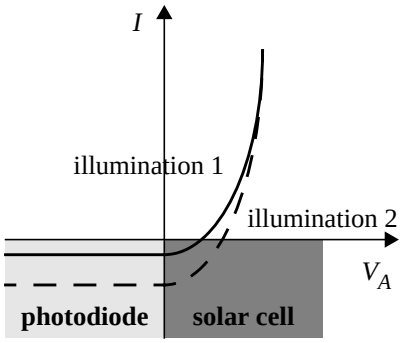


Fig. 43: Operational regions for a p-n junction. Illumination 2 is stronger than illumination 1.

Now, in this section, we pay our attention to devices that are based on junctions; on one hand, the photodetectors that are called photodiodes and, on the other hand, solar cells. They are in terms of principle of operation the same, but the actual fabrication of the devices is very different as they are targeted at different goals. For a photodiode only a narrow wavelength range centered at the optical signal wavelength is important, whereas for a solar cell, high spectral responses over a broad solar wavelength range are required. Also, photodiodes are small so as

1. The higher the bandgap, the higher is the melting points and the lower is their structural stability. Also, the higher bandgap materials have higher resistivity and cannot be easily doped to high levels.
2. Once I had a student in a course in basic physics who proposed the use of blue caps on red LEDs to achieve blue LEDs.
3. Also called Johnson noise.

to minimize junction capacitance, while solar cells are large-area devices. Finally, one of the most important figures of merit for photodiodes is the quantum efficiency, whereas the main concern for solar cells is the power conversion efficiency.

There is a big difference between the two devices in how they are used. The photodiode is always used under reverse bias, whereas the solar cell is unbiased and connected to some load impedance. In basic electronics courses teachers may have referred to photodiodes operating in different quadrants of the *I-V* characteristic of the p-n junction. In Fig. 43 the *I-V* characteristic is given, with proper indications on operational regions.

32.1. General Function of Photodiodes

When a photon excites a valence-band electron into the conduction band, a valence band hole is left behind. Thus, an electron-hole pair is created, leading to two carriers that are able to take part in current transport. This basic phenomenon is not new as we have discussed generation by optical means a number of times throughout this course. As shown in Fig. 44, photons with a

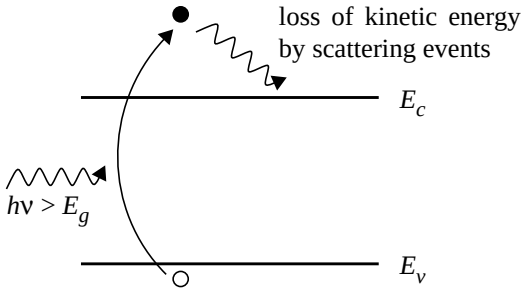


Fig. 44: Photon absorption.

frequency ν, where $h\nu > E_g$, will be absorbed.

NOTE that too high a photon energy leads to photon absorption close to the surface, where the probability for recombination is higher than in the bulk of the material. Thus, usually the bandgap is very selective in that it puts an upper and lower bound on the wavelength,

$$\lambda = \frac{c}{\nu} = \frac{hc}{h\nu} = \frac{1.24}{h\nu}\ \mu\text{m, where } h\nu \text{ is given in eV,}$$

that is detected in practice; the net result being a progressive reduction on the response with decreasing λ. The characteristic called spectral response is used to categorize photodiodes with respect to how they respond, in terms of diode current, to the wavelength of the incident light.

In connection to mentioning this characteristic, it is also important to notice the frequency response (the bandwidth) of a photodiode. This denotes how rapidly the detector can respond to a time-varying optical signal and relates to the carrier transit time of the diode.

The (quantum) efficiency η is here defined as the number of carriers collected to produce the photocurrent divided by the number of incident photons.

In Fig. 45 a sketch on a reverse-biased ordinary p-n junction diode is given. Here, the key to having an efficient photodiode is that the diode structure facilitates light transport into the junction and the quasi-neutral regions close to the junction. Photodiodes based on a p-n junction exist, but their performance is not particularly good in that the frequency response is not very large due to the time-consuming diffusion that takes place in the quasi-neutral regions[1]. A maximal

1. L_p and L_n usually are significantly larger than W.

bandwidth on the order of tens of MHz is what typically can be achieved in a p-n junction photodiode. No, the two dominating types of photodiodes[1] are the p-i-n and the avalanche photodiodes.

32.2. The p-i-n Photodiode

In a p-i-n[2] photodiode a high resistivity ("intrinsic") region is inserted between the p- and the n-type materials. Because of the low doping the i-region is totally depleted under zero bias or becomes depleted at small reverse biases. Furthermore, the heavy doping of the outer p- and n-type regions[3] causes the depletion widths in these regions to be very narrow. Thus, the depletion width inside the device is effectively equal to the i-region width, independent of bias. This implies that the photodiode, by controlling the width of the i-region, can be tailored to meet certain requirements on:

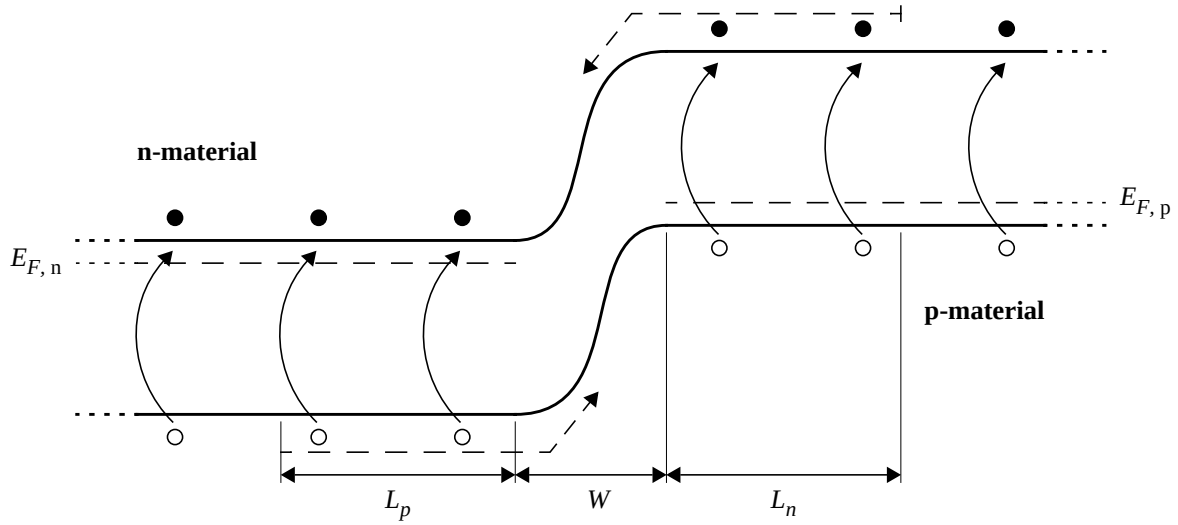


Fig. 45: A reverse-biased p-n junction where optical generation of carriers add to the reverse current if the carriers are generated within the depletion region or, at the most, a diffusion length away from it.

1. the frequency response, as a small *W* decreases the transit time[4] and thus increases the bandwidth,

2. the spectral response, by adjusting the width and thus the optimal wavelength according to the absorption expression in *Eq. (4-3)*[5]

$$I_t = I_0\, e^{-\alpha l},$$

1. Other types, outside these Lecture notes, are the phototransistor and the Metal-Semiconductor-Metal (MSM) photoconductor.
2. The i in p-i-n stands for intrinsic. However, the middle region is not made of an intrinsic material but rather of a lightly doped n-type material.
3. They are degenerately doped, in fact.
4. For very small *W*, the *RC* time constant of the diode limits the frequency response.
5. Choose *W* equal to the average penetration depth, i.e. $W = 1/\alpha(\lambda)$ for optimum, where λ is our preferred wavelength.

3. the efficiency $\eta^{(1)}$, as a large *W* allows for more photons to be absorbed.

NOTE that especially requirement 1 and 3 are conflicting, on the selection of *W*.

The p-i-n photodiodes find their application mainly as receivers in high-speed optical fiber communication, and then the frequency response is having the highest priority. Since the carriers travelling across the depletion region are driven by the drift mechanism, their velocity is much higher than the velocity of diffusing carriers in a p-n junction photodiode. We can assume that even for moderate reverse biases, the carriers drift across the i-region (of width ~ 1 μm) at their saturation velocity. There is, however, the effect of diffusion of carriers created outside the i-region, which can lower the frequency response, but this effect is minimized by fabricating the junction close to the illuminated surface.

32.3. The Avalanche Photodiode

For many applications, where very low levels of light are to be detected, it is desirable to use a detector with a large sensitivity. Large gains can be obtained in an avalanche photodiode (APD), which essentially is a reverse-biased p-n junction that is operated close to the breakdown voltage. Photogenerated carriers in the depletion region travel at their saturation velocities, and if they acquire enough energy from the electric field during such transit, an ionizing collision with the lattice can occur. Depending on the semiconductor material and device design, very large avalanche gains (~ 200 or more) can be achieved, and the APD therefore exhibits a very high sensitivity. A problem is that the avalanche process is stochastic, and thus the APD suffers from much higher noise than p-i-n photodiodes.

32.4. Solar Cells

Subsequent to the invention of the p-n junction diode in 1949, a large number of researchers tried to exploit its function in a number of areas. One of the most important advancements was the discovery of the solar cell in 1954, which was due to Chapin, Fuller and Pearson. Since then, solar cells have been developed and produced with polysilicon, cadmium telluride (CdTe) and GaAs. Today, more than 95% of the solar cells produced are Si based.

The conversion of radiation energy into electric energy is, in general, the photovoltaic effect and the most important photovoltaic device is the solar cell. The primary requirements for a material to be applicable to solar cells, is a bandgap matching the solar spectrum as well as high mobilities and lifetimes of charge carriers.

It is interesting to notice the influence of the bandgap energy on the conversion efficiency. Comparing the two most common materials, Si and GaAs, where E_g = 1.12 eV and 1.42 eV, respectively, gives the following:

Since photons with $h\nu > E_g$ are absorbed, all wavelengths below the cut-off wavelength λ_g, which is 1107 nm in Si and 873 nm in GaAs, are absorbed, which means that considering the solar spectrum, ~ 20% of the incident energy is wasted in Si, while ~ 35% is wasted in GaAs.

However, for $h\nu > E_g$, a portion of the photon energy adds to the kinetic energy of the photogenerated carriers and is eventually dissipated as heat, as shown in Fig. 44. Calculations show that ~ 40% of the absorbed photon energy is wasted in Si, while only ~ 30% is wasted in GaAs because of its larger bandgap.

1. If the photodetector has a gain larger than unity, the efficiency can become larger than 100%. This happens in photoconductors and avalanche photodiodes. The p-i-n photodiodes, however, is a unity-gain device.

Clearly, a trade-off exists between the two cited mechanisms, giving rise to an optimum bandgap where the energy conversion is at a maximum. Fortuitously, both Si and GaAs, which are the most advanced technologies, both are close to the theoretical maximum of the energy conversion.

The current through the solar cell under illumination is given by

$$I = I_L - I_0\left(e^{\frac{qV_A}{kT}} - 1\right),$$

where the current now is defined customary to the solar-cell field of engineering, i.e. the light-generated current, I_L, is assumed to be positive. When we use the solar cell in a short circuit and therefore $V_A = 0$, we get $I_{sc} = I_L$. When removing the load, we have an open-circuit arrangement, from which we can find V_{oc}:

$$I = 0 = I_L - I_0\left(e^{\frac{qV_{oc}}{kT}} - 1\right),$$

which can be solved with respect to V_{oc}:

$$V_{oc} = \frac{kT}{q}\ln\left(\frac{I_L}{I_0} + 1\right).$$

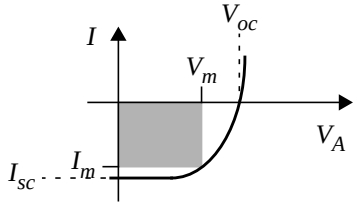


Fig. 46: Maximum power rectangle, $P_m = V_m \cdot I_m$.

In Fig. 46 the maximum power rectangle is drawn. When the operating point described by V_m and I_m is used, the solar cell delivers the maximum power, P_m. The ratio

$$FF = \frac{P_m}{I_{sc} \cdot V_{oc}},$$

is called the fill factor (why?). This allows us to define the power conversion efficiency η_s

$$\eta_s = \frac{I_{sc} \cdot V_{oc} \cdot FF}{P_{in}},$$

where P_{in} is the power of the incident solar radiation.

Probably the most important photovoltaic technology suitable for generating very large amounts of electricity is the a-Si:H technology, which denotes amorphous hydrogenated silicon. Even though typical efficiencies of a-Si solar cells are under 10%[(1)], these cells with large area

can be produced fairly inexpensively. The obtained values of the short-circuit current of these cells are on the order of 15 mA/cm^2, and typical open-circuit voltages are close to 900 to 950 mV with fill factors on the order of 0.7 to 0.75. An illustration of an a-Si:H solar cell with an amorphous hydrogenated silicon carbide window is given in Fig. 47.

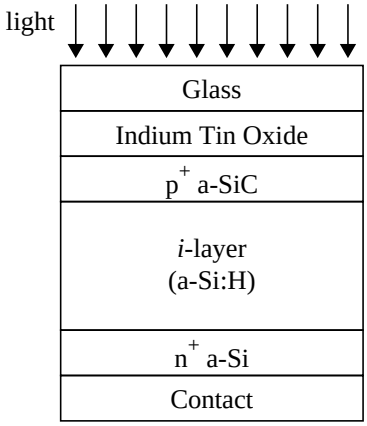


Fig. 47: Physical structure of an a-Si:H solar cell. The indium tin oxide acts as top contact, quite successfully as it is both transparent and has a high conductivity.

1. Maximum efficiencies in small-area a-Si solar cells have reached ~15%.

LECTURE 14

33. Fabrication and Integration

The main textbook is giving a lot of interesting facts on device fabrication, or device processing, that quickly can be summarized as:

Crystal growth, forming the single crystal material which is needed as substrate in the semiconductor devices, is covered in *Sec. 1.3*.

In *Sec. 1.4* Streetman tells us all we need to know about epitaxial growth, how to grow a thin material layer on top of the substrate. Different important epitaxial techniques such as Liquid-Phase Epitaxy (LPE), Vapor-Phase Epitaxy (VPE) and Molecular-Beam Epitaxy (MBE) are discussed. **NOTE** that the difference between the Czochralski process in *Sec. 1.3* and the epitaxial techniques is that in the latter ones the crystal can be grown below the melting point.

The essence of *Sec. 5.1* is fabrication processes targeting a p-n junction. It includes important things such as oxidation and doping. We have two options to introduce impurities in a single-crystal material; either a diffusion or an ion implantation process is used. The grown junction and the alloyed junction are mainly of historical interest only. Also, a brief discussion on lithography, etaching and metallization is provided.

In the short section of *Sec. 7.3*, the processing steps used in a BJT structure are shown in *Fig. 7-5*.

In *Ch. 9* finally you can find a quite good overview on integrated circuits. Of special interest is *Sec. 9.3.1*, which provides fabrication-related information on the MOS[1]. *Sec. 9.6* gives insight into testing, bonding and packaging of integrated circuits.

This course has up until now dealt with individual devices mainly. Of course, there exist many applications where so-called discrete (stand-alone) devices are used, especially in power applications. However, today the majority of the semiconductor devices are transistors integrated on chips or ICs. In the case of MOS ICs, resistors, capacitors and diodes are mostly implemented in the form of transistors since this is area efficient. This simplifies the manufacturing (processing) of the chips, in that only transistor-forming steps are required.

34. Integration - a Digital Perspective on the MOS

34.1. MOS and the Digital Signal Voltages

The symbol for the p-type MOS differs from the n-type symbol in that a ring is inserted on the gate plate. This ring indicates that the p-type MOS has, in certain respects, an inverted function in comparison to the n-type MOS. In Sec. 21.3 it was revealed that an n- and a p-type (enhancement-mode) MOS start to conduct when V_{GS} is above and below, respectively, the threshold voltage, V_T. Furthermore, V_T is positive and negative for the n- and the p-type MOS, respectively. Only in this section, they will be distinguishable by the notations V_{Tn} and V_{Tp}.

Out of the two digital levels, 0 or 1, a logical 1 is needed to turn on an n-type MOS, since we must create a positive potential difference between the gate and the source. Now, the discussion on the inverted function of the p-type MOS comes about because a logical 0 is needed to turn on a p-type MOS, since we must create a negative potential difference between the

1. If you want to learn more on fabrication of chips, you must attend TFYY33 Microchip Fabrication.

gate and the source. To illustrate, we assume that a logical 0 and 1 correspond to ground (V_{SS} = 0 V) and supply voltage (V_{DD} = 3 V), respectively.

The implication of the difference between n- and p-type MOS in digital circuit design is that n-type MOS conducts logical 0s well, whereas logical 1s are degraded by the transistor. In the same way, p-type MOS are preferred over n-type MOS when logical 1s are to be switched through the MOS. The reason we can assert this is the following:

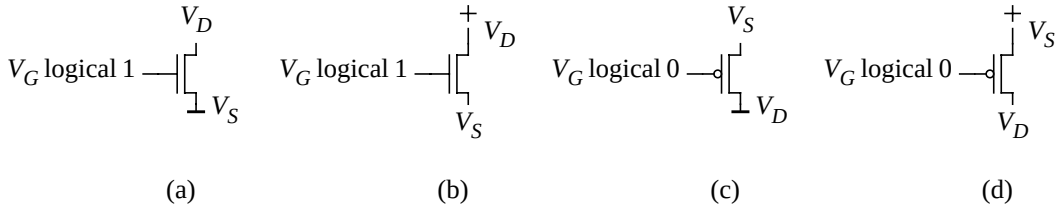


Fig. 48: Different ways of biasing n- and p-type MOS transistors in digital systems.

In Fig. 48(a) we have an n-type MOS that is supposed to transfer a logical 0 from the source to the drain[(1)], i.e. V_G corresponds to a logical 1 and V_{GS} = 3 V. The MOS conducts as long as $V_{GS} > V_{Tn}$, which is fulfilled for all possible voltages on the drain. The n-type MOS conducts logical 0s well.

In Fig. 48(b) we have an n-type MOS that is supposed to transfer a logical 1 from the drain to the source[(2)], i.e. V_G corresponds to a logical 1. The MOS conducts as long as $V_{GS} = V_G - V_S > V_{Tn}$, which is fulfilled when $V_S < V_G - V_{Tn} = 3\text{ V} - V_{Tn}$. This implies that the highest possible voltage the drain of the MOS can reach is V_{Tn} below the supply voltage in any system. The n-type MOS conducts logical 1s in a poor way, since a typical V_{Tn} (for V_{DD} = 3 V) is 0.8 V.

In Fig. 48(c) we have a p-type MOS that is supposed to transfer a logical 0 from the drain to the source, i.e. V_G corresponds to a logical 0. The MOS conducts as long as $V_{GS} = V_G - V_S < V_{Tp}$, which is fulfilled when $V_S > V_G - V_{Tp} = 0\text{ V} - V_{Tp}$. As V_{Tp} is negative, this implies that the lowest possible voltage the drain of the MOS can reach is $|V_{Tp}|$ above ground in any system. The p-type MOS conducts logical 0s in a poor way, since a typical V_{Tp} (for V_{DD} = 3 V) is -0.8 V.

Finally, in Fig. 48(d) we have a p-type MOS that is supposed to transfer a logical 1 from the source to the drain, i.e. V_G corresponds to a logical 0 and V_{GS} = -3 V. The MOS conducts as long as $V_{GS} < V_{Tp}$, which is fulfilled for all possible voltages on the drain. The p-type MOS conducts logical 1s well.

There is a way to form a quite efficient electronic switch by using both an n-type and a p-type MOS, a so-called transmission gate, Fig. 49. If C = 1, we have *out* = *in*, otherwise the output is

1. **From** the source **to** the drain is due to the definition of drain as being the terminal with the highest potential. Since one terminal is always grounded in a system where this is the lowest potential, this terminal will **always** be the source.

2. **From** the drain **to** the source is due to the definition of drain as being the terminal with the highest potential. Since one terminal is always connected to the supply voltage in a system where this is the highest potential, this terminal will **always** be the drain.

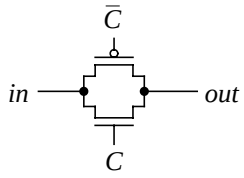


Fig. 49: A transmission gate.

floating. In Example 10 the resistance of a switched-ON MOS transistor was estimated. Considering that the resistance through the transmission gate, when conducting, is on the order of some kΩ, it would be of great interest to find a way to propagate signals without taking the risk that a signal is degraded from one logical level to another.

34.2. The Complementary MOS (CMOS) Circuit Technique

A unique feature of MOS-based circuits is that they can utilize the complementary function of n- and p-type MOS transistors. When one is switched ON, the other, in a proper complementary circuit, is switched OFF.

According to the discussion in Sec. 34.1, we are forced to use the n-type MOS for connection to V_{SS} and p-type MOS for connection V_{DD}. The simplest possible circuit we can devise would be the circuit in Fig. 50.

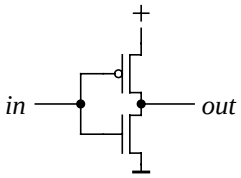


Fig. 50: A CMOS inverter.

When *in* = 1, the n-type MOS (but not the p-type MOS) is conducting, thus *out* = 0. Similarly, when *in* = 0, the p-type MOS (but not the n-type MOS) is conducting, thus *out* = 1. Thus, Fig. 50 presents the CMOS inverter.

Virtually no current flows in this circuit, when the signals are stable. No current passes through the gate to the channel, and no current passes from supply voltage to ground as one of the transistors is switched-OFF[1]. The implication of Sec. 34.1 is that the logical level on the input is almost perfectly transferred to the output in its inverted form, as the output node goes all the way up to V_{DD} for a logical output 1 and all the way down to V_{SS} for a logical output 0.

A bit more advanced CMOS circuits are presented in Fig. 51. It is obviously easier to implement inverted functions, as non-inverted need an extra inverter.

35. The MOS and the BJT - a Perspective on Fabrication and Integration

Since the BJT was the first transistor fabricated, it got off to a good start and dominated the IC business from the beginning.

A certain William Shockley moved to Palo Alto in 1956, to form a company of his

1. There is of course a small amount of current due to leakage. This is because the resistance across gate and semiconductor and the resistance across the channel in the MOS that is cut-off are both very high, but not infinitely high.

own. He had set his mind to improve the double-diffused bipolar transistor. He attracted many great engineers and scientists, but had problems in steering the company. Shockley was a tremendous engineer and scientist, but he was not as successful as a company manager.

One of the problems Shockley Semiconductor Laboratories was faced with, was the interconnections between different devices. And as Shockley himself selected germanium as target material, he disagreed with his staff that believed in a future silicon dominance, based on the advances made by a Dallas company called Texas Instruments. Due to the many controversies, most of his scientific staff left him and started a new company with funding from Sherman Fairchild - Fairchild Semiconductors was born.

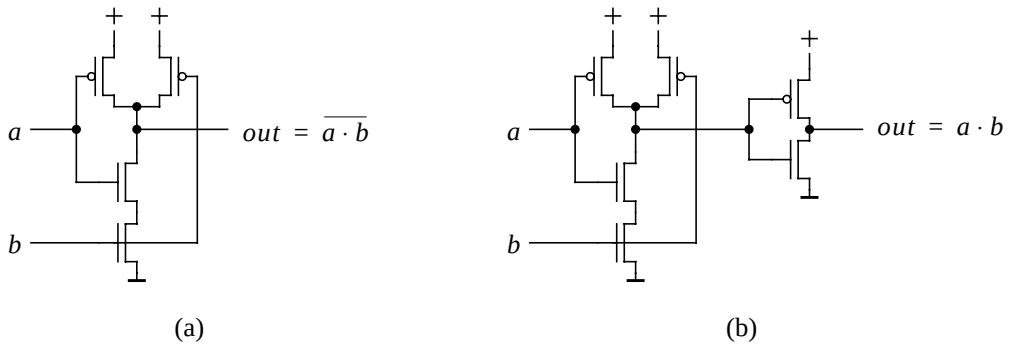


Fig. 51: (a) A CMOS NAND-gate. (b) A CMOS AND-gate.

In Dallas in 1958, the IC was invented by Jack Kilby at Texas Instruments. However, this IC lacked an efficient interconnection structure and therefore it was possible for Robert Noyce at Fairchild Semiconductors[1], to improve on this in 1959, claiming that his IC was the first true IC. Kilby and Noyce, through their respective companies, fought for ten years in a bitter way over who was to own the IC patent. Today, Kilby is viewed as the one who first conceived the idea of integrated circuits, while Noyce gets the credit for conceiving the modern IC in terms of interconnections.

So, BJTs were used for ICs; and circuit techniques such as RTL, DTL, TTL, I^2L and ECL were proposed. Today ECL (Emitter-Coupled Logic) and a variant of it, CML (Current-Mode Logic), hold strong positions; the first as a high speed circuit technique and the second as an area- and power-efficient replacement of ECL. The rest of the bipolar circuit techniques are only of historical interest[2].

Designers of electronics started to think about the power consumption during the end of the 60s, although no extremely severe limitation was associated with power (as it is today). Low power electronics was mainly brought forth as an important area for portable applications, such as pocket calculators, which literally was the first important application for ICs.

Digital circuit techniques based on MOS transistors were proposed for the simple reason that in a MOS the controlling terminal, the gate, is fully isolated from the channel. There is essentially no power consumption due to currents between the gate and the source/drain. Thus, the MOS can be categorized as a voltage-controlled transistor while the BJT, with its base cur-

1. Noyce went on to form a quite famous company, Intel, after his invention.
2. A bit arrogant perhaps; yes, you can buy several advanced versions of TTL (Transistor-Transistor Logic). Either that is because some designers think it is still 1978 or because they need backward compatibility for some reason. If you are used to a certain IC type, it is best to stay with it; experience gives reliability.

rent, is a current-controlled transistor.

The first MOS-based circuit technique of importance was nMOS, but it had some problems associated with power consumption as the logic circuits[1] draw current continuously between supply voltage and ground. In CMOS, however, improvements were possible since there was no short-circuit current between supply voltage and ground, as the n-type and p-type transistor nets complemented each other.

The low power consumption was the driving force behind the introduction of MOS-based ICs, but the total dominance it has earned since the middle of the 80s is mainly due to its suitability for integration. First and foremost, a higher integration density can be achieved for MOS transistors than for BJTs, since the MOS is less complex in its physical structure than the BJT is, a fact which considerably simplifies processing and improves the chip yield[2].

However, it should be noted that the length L of the MOS channel, horizontal as it is, is determined by the shortest distance that can be achieved without e.g. having problems with diffraction in the photolithographic process. The base width W_B of the vertical BJT, on the other hand, is determined by how well the diffusion or the ion implantation of the collector, the base and the emitter can be controlled. Since the base width is in the vertical direction, W_B can be made smaller than L (however the difference is not that big today). This is one of the reasons as to why a BJT inherently has a higher current drive capability and speed than a MOS.

As you recall from previous lectures, the BJT is an exponential device, as the collector current increases exponentially with V_{BE}. In contrast to the BJT, the drain current of the MOS at best has a square-law dependence on V_{GS}[3]. Thus, the BJT has higher gains than the MOS has for this reason. Coupled with the direct connection of the base terminal to the base and the capacitive connection of the gate to the channel, we have three reasons - geometry, *I-V* dependence, and control connection - for the BJT to outperform the MOS in terms of switching speed.

Now, not only speed and processing efficiency are used as parameters for comparison. When the two transistor types are integrated, a couple of parameters related to the circuits of transistor must be considered. The MOS-based CMOS circuit technique performs better than bipolar circuit techniques, because CMOS is quite immune to noise and tolerates a large interval of supply voltages, which the bipolar techniques do not.

36. Lowering the Supply Voltage

The current trend in integration of many transistors on a chip, is that power dissipation is becoming a limiting factor. To charge the output node of e.g. the inverter in Fig. 50, a node with capacitance C_L which originally carries no charge, to ΔV, energy according to

$$E = Q \cdot V_{DD} = (C_L \cdot \Delta V) \cdot V_{DD}$$

is required from the supply voltage source. This energy is transferred into heat during charging and discharging of the output node. Since power is energy over time, we have for a general circuit, where $\Delta V = V_{DD}$, that

1. Only n-type MOS transistors were used.
2. The percentage of operational chips.
3. With velocity saturation, unfortunately the exponent of 2 is replaced by say 1.3.

$$P_{sw} = f \cdot V_{DD}^{\ 2} \cdot \sum_i A_i \cdot C_i ,$$

where f is the system frequency, A_i is the probability that node i toggles 0→1 (or 1→0) and C_i is the load capacitance of node i, respectively.

From looking at the derivation of P_{sw}, the switching power[1], it's obvious that lowering the supply voltage would greatly reduce the power consumption. If we start lowering the supply voltage, also V_T has to be reduced, otherwise our circuits will not work well. Now let us take a look at how we can lower V_{DD} and V_T. However, first we notice that the turn-on V_{GS} for a bipolar transistor is dictated by the contact potential, thus, the supply voltage cannot be lowered below approximately 2 V. This is a huge disadvantage for bipolar technologies as the geometrical scaling will continue[2].

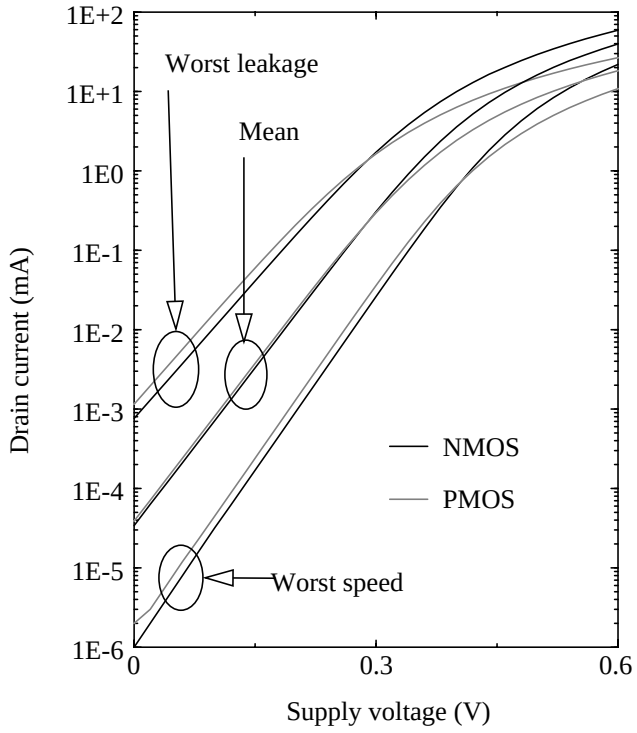


Fig. 52: Graph showing the subthreshold current as function of supply voltage in a 0.35-μm process.

Now we will try to understand the limitation to supply voltage reduction. We discussed subthreshold conduction in Sec. 24.2. For a MOS in the subthreshold region we have

$$I_D = I_0 \, e^{\frac{qV_{GS}}{mkT}} \left(1 - e^{-\frac{qV_{DS}}{kT}} \right),$$

where m is not equal to the ideality factor, η, from Sec. 16. Instead

1. Today the switching power is ~80% of the total power consumption of a chip.
2. The BJT has an advantage over the MOS, due to the turn-on potential's dependence on a material parameter such as the bandgap: BJTs are very efficient in some analog circuits, where we need to accurately match several transistors with respect to their *I-V* characteristics.

$$m = 1 + \frac{C_{\mathrm{dr}}}{C_{\mathrm{ox}}},$$

i.e., it is a measure on how much of the gate voltage is dropped over the depletion region, giving rise to a base voltage on the bipolar-like function of the subthreshold current. In a conventional process m is around 2, but for low-voltage processes it can reach 1.2.

Forming a CMOS inverter with both transistors in the subthreshold region, gives a maximum voltage gain of

$$G_{max} = \frac{e^{\frac{qV_{DD}}{2kT}} - 1}{m}$$

or

$$V_{DD} = \frac{2kT}{q} \ln(1 + m\ G_{max}).$$

For a robust circuit function G_{max} should be approximately 10, and with m = 1.5 this gives V_{DD} = 0.14 V.

When we use CMOS as circuit technology, we have a great advantage over BJT-technologies since the ratio of ON-current and OFF-current is large. We can write a relationship on this as

$$\frac{I_{d,\,ON}}{I_{d,\,OFF}} < e^{\frac{qV_{DD}}{mkT}}.$$

To make the ratio large enough for use in circuits, we have to select a ratio of 10^3. To keep the ratio at 10^3 or above we need, according to Fig. 52, $V_{DD} > 0.3$ V (assuming an m = 1.5).

A practical problem when a chip process is employed is that all sorts of parameters will vary, because the fabrication steps and the chemical processes are not perfectly stable and will not repeat the same results every time. We can describe the OFF-current as a function of threshold voltage, such that

$$I_{d,\,OFF} \sim e^{-\frac{qV_T}{mkT}}.$$

Thus, the OFF-current is very sensitive to process variations in the threshold voltage, so to have a margin of function, we need to add a ΔV_T. The worst case leakage must take into account the worst case V_T. The concept of controlling the threshold voltage, is referred to as threshold voltage control.

Not only is the control of the statistical spread of parameters important for leakage currents, but it also is necessary to control the delay of logical circuits. We note that delay is a function of V_T, such that

$$\text{Delay} \sim (V_{DD} - V_T)^{-1},$$

and thus the delay will vary dramatically at low supply voltages, unless the threshold is con-

trolled.

In practise, there are two different ways of controlling the threshold voltage: either process control is employed, or else so-called backbiasing is used. In the process control, the chip vendor has to include processing steps that control the threshold voltages by doping, etc. Today the variation on the threshold voltage can be kept within 25 mV. As far as it is possible, a kind of natural threshold should be used; then we assume both low substrate doping and low V_T adjustment doping. In the backbiasing, or substrate voltage control, case, we simply control the threshold voltage by actively using the body effect (Sec. 24.8).

Now, when we are faced with the reality of lower supply voltages, and hence lower threshold voltages, that was driven by a need to reduce power consumption, what did we get? Well, there are also problems with low V_T. In circuits where switching is not taking place that often, now the static power of leakage may start to dominate. Think of static memories! On the other hand, they can still be used, which cannot be said about dynamic memories where the dynamic, charge storing mechanism is dramatically degraded by leakage.

One way to make a compromise, is to use two different V_Ts. One set of low V_Ts is used in high speed, highly active logic, whereas the other set of V_Ts, high V_Ts, is used in slower and less active parts, in power switching and in SRAMs. Depending on application, a decision has to be made on process supply voltage and threshold voltages:

- In a desktop computer, we need very high performance, and thus we must have minimum power with no performance loss. However, we can accept a high standby power. Here probably, a low supply voltage and a low threshold voltage would be reasonable.
- In mobile equipment, heat and battery issues have to be considered together with performance. Here large standby power cannot be accepted, so dual V_Ts along with low supply voltage would be needed.
- For hand-held equipment, battery life is the primary parameter, making performance a secondary parameter. Just as before low supply voltage is natural, but now we have to focus on higher threshold voltages, either one high V_T or maybe dual V_Ts.

The benefits of a low voltage process are many more than the power consumption reduction. Since we get a stable V_T, when migrating to low voltage processes, we will have a good leakage and speed control of the circuits. Drain induced barrier lowering is reduced when V_{DD} is lowered, and then we can lower V_T further. Also, the subthreshold characteristics are good, i.e. the current is low, since

$$I_D = I_0 \, e^{\frac{qV_{GS}}{mkT}} \left(1 - e^{-\frac{qV_{DS}}{kT}} \right),$$

and m is small for low voltage processes.

If we sketch on the short-channel effects from Sec. 24, we have that

$$I_d(\text{low } V_{DD}) = F_v \cdot F_\mu \cdot I_d(\text{high } V_{DD}),$$

where F_v and F_μ are the gains from reducing the adverse effects of velocity saturation and mobility degradation, respectively, when going from “high” to “low” V_{DD}:

$$F_v = \frac{2}{1 + \sqrt{1 + \left(\frac{V_{DD} - V_T}{E_{sn} \cdot L}\right)^2}}$$

and

$$F_\mu = \frac{1}{\left(1 + \frac{E_y}{E_{y0}}\right)^\nu}.$$

If we consider a change of process supply voltage such that we decrease from 3 V to 1 V in a 0.25-μm process, we roughly double $F_v \cdot F_\mu$. Say furthermore that we can reduce V_T from 0.5 V to 0.3 V, then we get a speed improvement of 2.6. Together the improvement on the performance is a factor 5.2, to be compared to a nominal speed loss, by looking at the delay, of

$$\frac{\text{low } (V_{DD} - V_T)^{-1}}{\text{high } (V_{DD} - V_T)^{-1}} = \frac{3 - 0.5}{1 - 0.3} = 3.6 \text{ times.}$$

So, we lose some speed because of longer circuit delay, but wins it back by faster devices. Did we win anything? YES, we have reduced power by a factor of 9 ($3^2/1^2$).

In this example, we considered the most apparent mechanisms, but in fact low voltage also is advantageous in other ways. The reliability is improved for several reasons: Hot carriers (Sec. 24.7) are less likely to be created, since the maximum drain voltage is reduced. Also, we can avoid latch-up, a parasitic bipolar phenomenon which can occur in circuits where n-type and p-type MOS, as illustrated in Fig. 53.

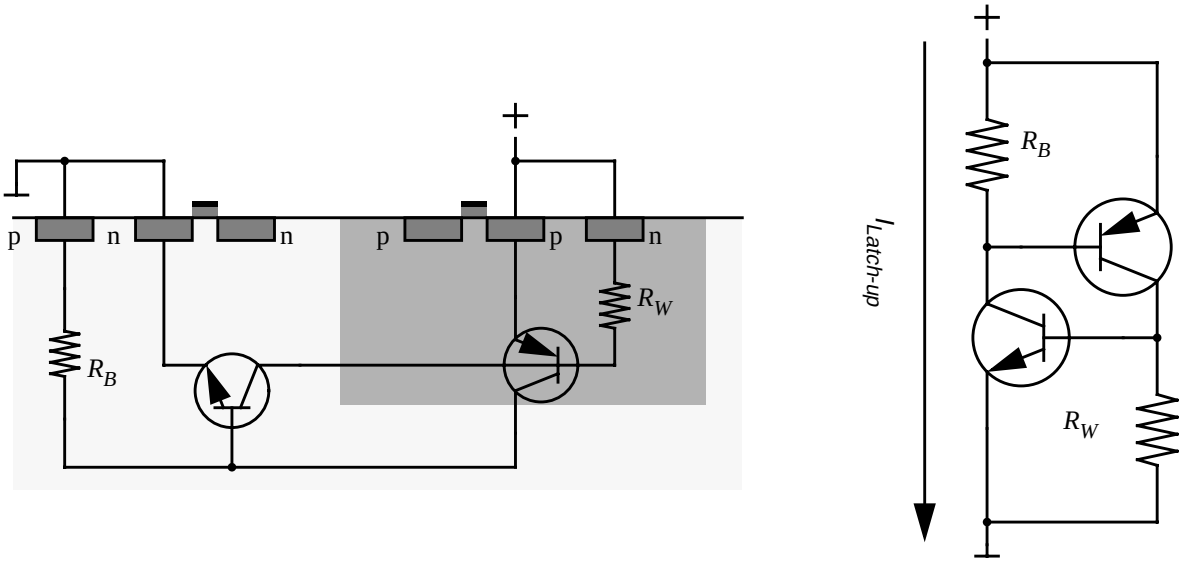


Fig. 53: Schematic of the latch-up phenomenon when taking place in an inverter.