Basics of Semiconductor Devices

Dinesh Sharma
EE Department, IIT Bombay

September 9, 2021

chapterBasics of Semiconductor Devices In this booklet, we review the fundamentals of Semiconductor Physics and basics of device operation. We shall concentrate largely on elemental semiconductors such as silicon or germanium, and most numerical values used for examples are specific to silicon.

1 Semiconductor fundamentals

A semiconductor has two types of mobile charge carriers: negatively charged *electrons* and positively charged *holes*. We shall denote the concentrations of these charge carriers by n and p respectively. The discussions in this booklet apply to elemental semiconductors (like silicon) which belong to group IV of the periodic table. We can intentionally add impurities from groups III and V to the semiconductor. These impurities are called dopants. Impurities from group III are called *acceptors* while those from group V are called *donors*. Each donor atom has an extra electron, which is very loosely bound to it. At room temperature, there is sufficient thermal energy present, so that the loosely bound electron breaks free from the donor, leaving the donor positively charged. This contributes an additional electron to the free charge carriers in the semiconductor, and a positive ionic charge at a *fixed* location in the semiconductor. Similarly, an acceptor atom captures an electron, thus producing a mobile hole and becoming negatively charged itself. A semiconductor without any dopants is called *intrinsic*. An unperturbed semiconductor must be charge neutral as a whole. If we denote the concentration of ionised donors by N_d^+ and the concentration of ionised acceptors by N_a^-, we can write for the net charge density at any point in the semiconductor as:

$$\rho = q(N_d^+ - N_a^- + p - n) \tag{1}$$

where q is the absolute value of the electronic charge. In an unperturbed semiconductor, ρ will be zero everywhere. Electrons and holes are generated thermally - the availability of energy equal to the band gap of the semiconductor results in the generation of an *electron - hole pair*. Simultaneously, electrons and holes can recombine to annihilate each other, giving out energy which is equal to the band gap of the semiconductor. Thus we have the reversible reaction:

$$\mathrm{e}^- + \mathrm{h}^+ \rightleftharpoons E_g$$

Where E_g is the band gap energy of the semiconductor.
Applying the law of mass action to the above reaction, we can write for the equilibrium concen-

tration of holes and electrons:

$$n \cdot p = \text{constant}$$

The above relation applies to doped as well as intrinsic semiconductors. But for an intrinsic semiconductor,

$$n = p \equiv n_i$$

Therefore, the constant in the equation connecting n and p must be n_i^2. Thus, for a semiconductor *in equilibrium*,

$$n \cdot p = n_i^2 \tag{2}$$

Since n and p are not independent, but are constrained by the above relation, we can define a single independent variable, the Fermi potential by

$$\Phi_{\mathrm{F}} \equiv \frac{K_B T}{q} ln \frac{p}{n_i} = \frac{K_B T}{q} ln \frac{n_i}{n} \tag{3}$$

Where $\mathrm{K_B}$ is the Boltzmann constant, T is the absolute temperature and q is the absolute value of the electronic charge. At room temperature, $K_B T/q$ is approximately 26 mV and n_i is of the order of $10^{10}/\mathrm{cm}^3$ for silicon. Now electron and hole concentrations are given by:

$$\begin{aligned} n &= n_i e^{-\frac{q\Phi_{\mathrm{F}}}{K_B T}} \\ p &= n_i e^{\frac{q\Phi_{\mathrm{F}}}{K_B T}} \end{aligned} \tag{4}$$

To simplify these relations, we define a dimensionless Fermi potential by:

$$u_F \equiv \frac{q\Phi_{\mathrm{F}}}{K_B T} = \ln(p/n_i) = \ln(n_i/n)$$

then:

$$\begin{aligned} n &= n_i e^{-u_F} \\ p &= n_i e^{u_F} \end{aligned} \tag{5}$$

Generally, a semiconductor will be doped with only one kind of impurity. A semiconductor doped with donors will have many more electrons than holes. This type of semiconductor is called N type, and electrons are the *majority* carriers in this type of semiconductor. Similarly, holes are the majority carriers in a semiconductor doped with acceptors and it is termed P type. If both types of dopants are present, the one present in higher concentration determines the 'type' of the semiconductor. The *net* doping is defined as the difference in the concentrations of the more abundant and the less abundant dopants.

In most practical cases, the ratio of majority to minority carriers is very high. The concentration of majority carriers is very nearly equal to the net dopant concentration. To take a typical example, consider P type silicon with boron concentration of 10^{16} atoms/cm^3. This gives:

$$\begin{aligned} p &= N_a = 10^{16}/\mathrm{cm}^3 \\ n &= n_i^2/p \approx 10^{20}/10^{16}/\mathrm{cm}^3 = 10^4/\mathrm{cm}^3 \\ p/n &\approx 10^{12}\ ! \end{aligned}$$

1.1 Band Diagrams

The above concepts are often visualised with the help of band diagrams. The arrangement of atoms in a semiconductor results in certain electron energies which are not permitted. Thus, the energy range is divided into bands of permitted energy values alternating with forbidden gaps.

The highest such band which is nearly filled with electrons is called the valance band. Unoccupied levels in this band correspond to holes. For stability, electrons seek the lowest energy level available. If a vacancy is available at a lower energy - an electron at a higher energy will drop to this level. The vacancy thus bubbles up to a higher level. Therefore, holes seek the highest *electron* energy available.

The band just above the valance band is called the conduction band. In a semiconductor, this is partially filled. Conduction in a semiconductor is caused by electrons in the conduction band (which are normally to be found at the lowest energy in the conduction band) and holes in the valance band - (found at the highest electron energy in the valance band).

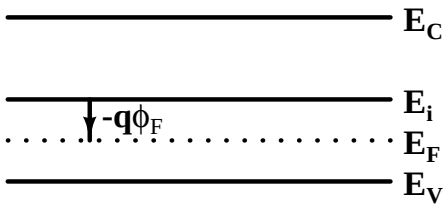


Figure 1: Semiconductor Bands

Band diagrams are plots of electron energies as a function of position in the semiconductor. Typically, the top of the valance band (corresponding to minimum hole energy) and the bottom of the conduction band are plotted. We can show the Fermi potential and the corresponding Fermi energy(= -qΦ_F) in the band diagram of silicon as a level in the band gap. We use the halfway point between the conduction and the valence band as the reference for energy and potential. When n = p = n_i, the Fermi potential is 0 (from eq. 3) and correspondingly, the Fermi energy lies at the intrinsic Fermi level halfway in the band gap. (Actually, this level can be slightly away from the middle of the band gap depending on the density of allowed states in the conduction and valance bands - but for now, we'll ignore this). When holes are the majority carriers, $\Phi_F$is positive and the Fermi energy (= -q Φ_F) lies below the mid gap level, as shown in figure 1. When electrons are the majority carriers, $\Phi_F$is negative, and the Fermi energy lies above the mid gap level.

1.2 A semiconductor in the presence of an electric field

In the presence of an electric field, the electrostatic potential is different at different positions. The energy of an electron has an extra component = $-q\phi$ where ϕ is the electrostatic potential. Consequently in the band diagram the conduction, valance and intrinsic levels are bent. In equilibrium, the Fermi level is still straight. (We shall see later that in the absence of a current, the slope of the Fermi level must vanish). Relations for n and p must now take the electrostatic potential as well as the Fermi potential into account and the electron and hole concentrations

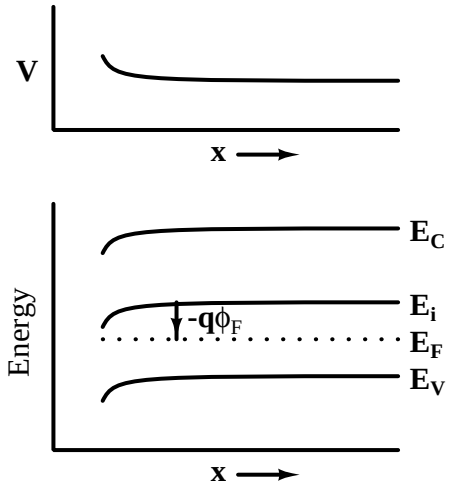


Figure 2: Potential distribution and Band Diagram in the presence of a field

are not uniform over the semiconductor. If we represent the concentrations of electrons and holes without any applied field by n_0 and p_0 respectively, then in the presence of a field (but in equilibrium),

$$\begin{aligned} n &= n_0 e^{\frac{q\phi}{K_B T}} \\ p &= p_0 e^{-\frac{q\phi}{K_B T}} \end{aligned} \tag{6}$$

where ϕ is the electrostatic potential.

If we define a dimensionless electrostatic potential by:

$$u \equiv \frac{q\phi}{K_B T} \tag{7}$$

we can write the above relations as:

$$\begin{aligned} n = n_0 e^{u} &= n_i e^{(u-u_F)} \\ p = p_0 e^{-u} &= n_i e^{-(u-u_F)} \end{aligned} \tag{8}$$

Since there is equilibrium, even though electron and hole concentration is not uniform, the product of n and p is still constant and equal to n_i^2 everywhere.

1.3 Non-equilibrium case

The above relations assume a semiconductor in equilibrium. It is possible to create excess carriers in the semiconductor over those dictated by equilibrium considerations. For example, if we shine light on a semiconductor, electron-hole pairs will be created. Since the value of n *as well as* that of p goes up, the np product will exceed n_i^2, till the equilibrium is restored after the light is turned off (by enhanced recombination). If the number of excess carriers is small compared to the majority carriers, we may assume that the carrier concentrations are still described by relations like those given above. However, the concentrations of electrons and holes are not constrained by relation(2) any more. Therefore, we cannot use the same value of u_F for describing electron as well as hole concentrations. We now have *separate* values of Φ_F for electrons and holes. These are

called *quasi* Fermi levels (or imrefs) for electrons and holes, $\Phi_{F_n}$and Φ_{F_p}, defined by the relations

$$\begin{aligned} n &= n_i e^{(u-u_{F_n})} \\ p &= n_i e^{-(u-u_{F_p})} \end{aligned} \tag{9}$$

Where u_{F_n} and u_{F_p} are the dimensionless versions of quasi Fermi levels $\Phi_{F_n}$and Φ_{F_p}defined as in equation(7)). The np product is now given by

$$np = n_i^2 e^{(u_{F_p}-u_{F_n})} \tag{10}$$

and is no longer constant. Because the number of additional carriers is assumed to be small compared to the majority carriers, the concentration of majority carriers and hence its quasi Fermi level is very close to the equilibrium value. The relative change in the concentration of minority carriers could, however, be large and consequently the minority carrier quasi Fermi level could be substantially different from the equilibrium Fermi level.

2 The p-n diode

We shall analyse the abrupt pn junction, in reverse and forward bias. We assume that the doping

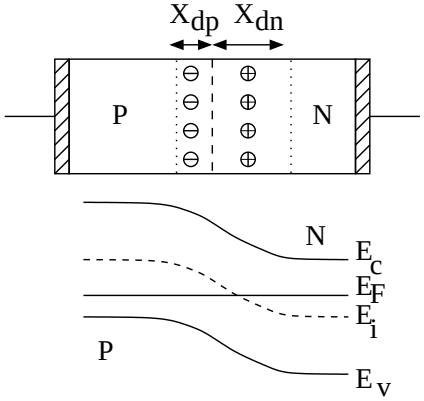


Figure 3: The abrupt p-n junction

density is constant and its value = N_a on the P side and N_d on the N side, changing abruptly at the metallurgical junction as shown. Because there is a strong concentration gradient for electrons and holes at the junction, there will be a diffusion current of holes towards the N side and of electrons towards the P side. As these carriers leave behind ionised dopants, small regions on either side of the junction acquire a charge. The P side, from where positively charged holes have left, (leaving behind negatively charge acceptor ions), acquires a negative potential. Similarly, the N side becomes positively charged. The regions from where mobile charges have left, are called depletion regions. The potential difference resulting from this charge redistribution (called the built-in voltage) opposes further diffusion of carriers. A dynamic equilibrium is reached when the drift current due to this potential difference and the diffusion current due to the concentration gradient become equal and opposite. In equilibrium, The electron as well as hole currents must be zero individually (principle of detailed balance). Writing the electron and hole current densities

as sums of their respective drift and diffusion current densities:

$$\begin{aligned} J_n &= nq\mu_n(-\frac{\partial\phi}{\partial x}) + qD_n\frac{\partial n}{\partial x} \\ J_p &= pq\mu_p(-\frac{\partial\phi}{\partial x}) - qD_p\frac{\partial p}{\partial x} \end{aligned} \tag{11}$$

From equation(9)

$$\begin{aligned} \frac{\partial n}{\partial x} &= n_i e^{(u-u_{F_n})}\frac{\partial}{\partial x}(u-u_{F_n}) \\ \frac{\partial p}{\partial x} &= n_i e^{(u_{F_p}-u)}\frac{\partial}{\partial x}(u_{F_p}-u) \end{aligned}$$

or

$$\begin{aligned} \frac{\partial n}{\partial x} &= n\frac{q}{K_BT}\frac{\partial}{\partial x}(\phi-\Phi_{F_n}) \\ \frac{\partial p}{\partial x} &= p\frac{q}{K_BT}\frac{\partial}{\partial x}(\Phi_{F_p}-\phi) \end{aligned}$$

Using Einstein relations ($\frac{q}{K_BT}D=\mu$), and Substituting in the relations for J_n and J_p,

$$\begin{aligned} J_n &= -nq\mu_n(\frac{\partial\phi}{\partial x}) + nq\mu_n\frac{\partial}{\partial x}(\phi-\Phi_{F_n}) \\ J_p &= -pq\mu_p(\frac{\partial\phi}{\partial x}) - pq\mu_p\frac{\partial}{\partial x}(\Phi_{F_p}-\phi) \end{aligned}$$

Which leads to

$$\begin{aligned} J_n &= -nq\mu_n\frac{\partial\Phi_{F_n}}{\partial x}; \\ J_p &= -pq\mu_p\frac{\partial\Phi_{F_p}}{\partial x}; \end{aligned} \tag{12}$$

When there is no flow of current, $\Phi_{F_n}=\Phi_{F_p}=\Phi_F$. according to the relations derived above, the derivative of $\Phi_F$must vanish everywhere for zero current. Thus, the Fermi level is constant and the same at the two sides of the junction. The Fermi potentials before being put in contact were:

$$\begin{aligned} \Phi_F = &\quad \frac{K_BT}{q}\ln(N_a/ni) && \text{P side} : x<0 \\ \Phi_F = &\quad -\frac{K_BT}{q}\ln(N_d/ni) && \text{N side} : x>0 \end{aligned}$$

The Fermi potential difference was, therefore, $\frac{K_BT}{q}\ln\left(\frac{N_dN_a}{n_i^2}\right)$. Since after being put in contact, the Fermi levels have equalised on the two sides, the built in voltage must be equal and opposite to this potential, taking the P side to a negative potential and the N side to a positive potential. We can write for the magnitude of the built in voltage:

$$V_{bi} = \frac{K_BT}{q}\ln\left(\frac{N_aN_d}{n_i^2}\right) \tag{13}$$

2.1 pn Diode in Reverse Bias

The diode is reverse biased when we apply a voltage such that the n side is more positive as compared to the p side. In this case, the applied voltage is in the same direction as the built-in field, which opposes the movement of majority carriers and widens the depletion regions on either side of the junction. We analyse the reverse biased diode by making the depletion approximation. We assume that in reverse bias, the depletion regions have zero carrier density, and the field is completely confined to depletion regions. Solving Poisson's equation in P region ($x < 0$) and the N region ($x > 0$)

$$\begin{aligned} \frac{\partial^2\phi}{\partial x^2} &= \frac{qN_a}{\epsilon_{si}} \quad \text{(for x} < 0) \\ \frac{\partial^2\phi}{\partial x^2} &= -\frac{qN_d}{\epsilon_{si}} \quad \text{(for x} > 0) \end{aligned}$$

Integrating with respect to x

$$\begin{aligned} \frac{\partial\phi}{\partial x} &= \frac{qN_a}{\epsilon_{si}}x + c1 \quad \text{(for x} < 0) \\ \frac{\partial\phi}{\partial x} &= -\frac{qN_d}{\epsilon_{si}}x + c2 \quad \text{(for x} > 0) \end{aligned}$$

where c_1 and c_2 are constants of integration, which can be evaluated from the condition that the field vanishes at the edge of the depletion regions at $-X_{dp}$ and at X_{dn}. This leads to

$$\begin{aligned} \frac{\partial\phi}{\partial x} &= \frac{qN_a}{\epsilon_{si}}(x + X_{dp}) \quad \text{(for x} < 0) \\ \frac{\partial\phi}{\partial x} &= -\frac{qN_d}{\epsilon_{si}}(x - X_{dn}) \quad \text{(for x} > 0) \end{aligned} \tag{14}$$

Since the value of the field must match at x = 0;

$$N_a X_{dp} = N_d X_{dn} \tag{15}$$

Integrating equation (14) once again with respect to x, we get

$$\begin{aligned} \phi &= \frac{qN_a}{\epsilon_{si}}\left(\frac{x^2}{2} + X_{dp}x\right) + c_3 \quad \text{(for x} < 0) \\ \phi &= -\frac{qN_d}{\epsilon_{si}}\left(\frac{x^2}{2} - X_{dn}x\right) + c_4 \quad \text{(for x} > 0) \end{aligned}$$

Where the constants of integration c_3 and c_4 can again be evaluated from the boundary conditions at $-X_{dp}$ and X_{dn}. If we require that the potential is 0 at $-X_{dp}$ and V at X_{dn},

$$\begin{aligned} c_3 &= \frac{qN_a}{2\epsilon_{si}}X_{dp}^2 \\ c_4 &= V - \frac{qN_d}{2\epsilon_{si}}X_{dn}^2 \end{aligned}$$

Substituting these values, we get:

$$\begin{aligned} \phi &= \frac{qN_a}{\epsilon_{si}}\left(\frac{x^2+X_{dp}^2}{2} + X_{dp}x\right) \quad \text{(for x} < 0) \\ \phi &= V - \frac{qN_d}{\epsilon_{si}}\left(\frac{x^2+X_{dn}^2}{2} - X_{dn}x\right) \quad \text{(for x} > 0) \end{aligned} \tag{16}$$

Since the potential at x = 0 should be continuous,

$$\frac{qN_a}{2\epsilon_{si}}X_{dp}^2 = V - \frac{qN_d}{2\epsilon_{si}}X_{dn}^2$$

$$\text{so, } V = \frac{q}{2\epsilon_{si}}(N_a X_{dp}^2 + N_d X_{dn}^2) \tag{17}$$

making use of equation (15), we can write

$$\begin{aligned} V &= \frac{qN_a X_{dp}^2}{2\epsilon_{si}N_d}(N_d + N_a) \\ &= \frac{qN_d X_{dn}^2}{2\epsilon_{si}N_a}(N_d + N_a) \end{aligned}$$

which leads to

$$\begin{aligned} X_{dp} &= \sqrt{\frac{2\epsilon_{si}V}{q(N_d + N_a)}\frac{N_d}{N_a}} \\ X_{dn} &= \sqrt{\frac{2\epsilon_{si}V}{q(N_d + N_a)}\frac{N_a}{N_d}} \end{aligned} \tag{18}$$

From which the total depletion width can be calculated as:

$$X_d \equiv X_{dp} + X_{dn} = \sqrt{\frac{2\epsilon_{si}V}{q(N_d + N_a)}}\left(\sqrt{\frac{N_d}{N_a}} + \sqrt{\frac{N_a}{N_d}}\right)$$

which gives

$$X_d = \sqrt{\frac{2\epsilon_{si}V}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)} \tag{19}$$

The voltage V in the above expressions is the total voltage across the junction. Since there is a reverse bias of V_{bi} for a zero applied voltage, that will add (in magnitude) to the applied reverse voltage. Using equation(13) we can write:

$$V = V_{bi} + V_{appl} = V_{appl} + \frac{K_B T}{q}\ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{20}$$

2.2 The pn diode in forward bias

If we apply an external voltage, such that the P side is made positive with respect to the N side, the applied voltage will reduce the built in voltage across the junction. The magnitude of the built-in voltage is such that it balances the drift and diffusion currents, resulting in zero net current. But if the voltage across the junction is reduced, a net current will flow through the diode. This is the forward mode of operation. Because of this flow of current, electrons are injected into the P side and holes into the N side. Consequently, the concentration of carriers is no longer at the equilibrium value. We denote the *equilibrium* value of electron and hole concentrations on P and N side by $n_{p_0}, n_{n_0}, p_{p_0}, p_{n_0}$ respectively. Since the majority carrier concentration in equilibrium is equal to the doping density, we have:

$$n_{n_0} \approx N_d, \qquad p_{p_0} \approx N_a \qquad \text{and} \qquad n_{p_0} = n_i^2/N_a, \qquad p_{n_0} = n_i^2/N_d$$

According to equation(10)

$$np = n_i^2 e^{(u_{F_p} - u_{F_n})}$$

As we make the potential of P type more positive compared to N type, the np product in forward bias is greater than n_i^2. From relations(12), we see that the change in quasi Fermi levels is small wherever the carrier concentration is high. Thus, we can assume that the quasi Fermi levels of the *majority* carriers at either side of the junction remain at their equilibrium values. Hence the voltage across the junction is given by

$$V = \phi_{F_p} - \phi_{F_n}$$

and therefore the *non-equilibrium* np product is given by

$$np = n_i^2 e^{\left(\frac{qV}{K_B T}\right)}$$

therefore,

$$\begin{aligned} n_p &= \frac{n_i^2}{p_p} e^{\left(\frac{qV}{K_B T}\right)} = n_{p_0} e^{\left(\frac{qV}{K_B T}\right)} \\ p_n &= \frac{n_i^2}{n_n} e^{\left(\frac{qV}{K_B T}\right)} = p_{n_0} e^{\left(\frac{qV}{K_B T}\right)} \end{aligned} \tag{21}$$

(22)

The continuity equation for any particle flow can be written as

$$\nabla.(\text{particle current density}) = -\frac{\partial}{\partial t}(\text{particle concentration})$$

Applying it to electron and hole currents in 1 dimension on the n side,

$$\begin{aligned} \frac{\partial}{\partial x}\left(\frac{J_n}{-q}\right) &= U \\ \frac{\partial}{\partial x}\left(\frac{J_p}{q}\right) &= U \end{aligned}$$

where U is the net recombination rate. Using relation(11), we have

$$\begin{aligned} \frac{\partial}{\partial x}\left(n_n \mu_n \frac{\partial \phi}{\partial x} - D_n \frac{\partial n_n}{\partial x}\right) &= U \\ \frac{\partial}{\partial x}\left(p_n \mu_p \frac{\partial \phi}{\partial x} + D_p \frac{\partial p_n}{\partial x}\right) &= U \end{aligned}$$

or

$$\begin{aligned} \mu_n \frac{\partial n_n}{\partial x}\frac{\partial \phi}{\partial x} + \mu_n n_n \frac{\partial^2 \phi}{\partial x^2} - D_n \frac{\partial^2 n_n}{\partial x^2} &= U \\ \mu_p \frac{\partial p_n}{\partial x}\frac{\partial \phi}{\partial x} + \mu_p p_n \frac{\partial^2 \phi}{\partial x^2} + D_p \frac{\partial^2 p_n}{\partial x^2} &= U \end{aligned}$$

Assuming the regions outside the small depletion regions to be charge neutral,

$$(n_n - n_{n_0}) \approx (p_n - p_{n_0})$$

We define ambipolar diffusion and lifetime by the relations

$$D_a \equiv \frac{n_n + p_n}{n_n/D_p + p_n/Dp} \tag{23}$$

$$\tau_a \equiv \frac{p_n - p_{n_0}}{U} = \frac{n_n - n_{n_0}}{U} \tag{24}$$

multiplying the electron continuity equation with $\mu_p p_n$ and the hole continuity equation with $\mu_n n_n$ and combining, we get

$$-\frac{p_n - p_{n_0}}{\tau_a} + D_a \frac{\partial^2 p_n}{\partial x^2} + \frac{n_n - p_n}{n_n/\mu_p + p_n/\mu_n} \frac{\partial p_n}{\partial x} \frac{\partial \phi}{\partial x} = 0 \tag{25}$$

If we make the low injection assumption ($p_n << n_n \approx n_{n_0}$), this reduces to

$$-\frac{p_n - p_{n_0}}{\tau_p} + D_p \frac{\partial^2 p_n}{\partial x^2} + \mu_p \frac{\partial p_n}{\partial x} \frac{\partial \phi}{\partial x} = 0 \tag{26}$$

In the neutral region, $\frac{\partial \phi}{\partial x}$ is zero, so the above simplifies further to

$$\frac{\partial^2 p_n}{\partial x^2} - \frac{p_n - p_{n_0}}{D_p \tau_p} = 0 \tag{27}$$

This can be solved with the boundary condition given by relation(21) and noting that $p_n = p_{n_0}$ at $x = \infty$ to give:

$$p_n - p_{n_0} = p_{n_0} \left(e^{\frac{qV}{K_B T}} - 1 \right) e^{\frac{x - x_n}{L_p}} \tag{28}$$

where

$$L_p \equiv \sqrt{D_p \tau_p} \tag{29}$$

Evaluating the hole current at X_{dn}, we get

$$J_p = -qD_p \frac{\partial p_n}{\partial x} = \frac{qD_p p_{n_0}}{L_p} \left(e^{\frac{qV}{K_B T}} - 1 \right) \tag{30}$$

Similarly, we can evaluate the electron current on the p side as

$$J_n = qD_n \frac{\partial n_p}{\partial x} = \frac{qD_n n_{p_0}}{L_n} \left(e^{\frac{qV}{K_B T}} - 1 \right) \tag{31}$$

which gives the total current density as

$$J = J_p + J_n = J_s \left(e^{\frac{qV}{K_B T}} - 1 \right) \tag{32}$$

$$\text{Where } J_s \equiv \frac{qD_p p_{n_0}}{L_p} + \frac{qD_n n_{p_0}}{L_n} \tag{33}$$

3 The MOS Capacitor

It is important to understand the MOS capacitor in order to understand the behaviour of the the MOS transistor. Before we describe the MOS structure, it is useful to review the basic electrostatics as applied to parallel plate capacitors. We shall then go on to analyse the MOS structure.

3.1 The Parallel Plate Capacitor

The parallel plate capacitor consists of two parallel metallic plates of area A, separated by an insulator of thickness t_i and dielectric constant ϵ. If we place a charge Q on the upper plate, it attracts charges of opposite sign in the bottom plate, while repelling charges of the same sign. If the bottom plate is connected to ground, the repelled charge flows to ground. Now the two

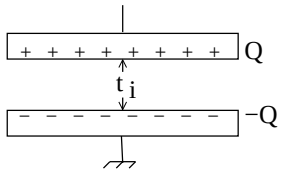


Figure 4: The parallel Plate capacitor

capacitor plates hold equal and opposite charge. This charge resides just next to the insulator on either side of it. This is true, *whatever the quantity or sign of charge* placed on the upper plate. The inducing and induced charge are always separated by the thickness of the insulator, t_i. Therefore this structure has a *constant* capacitance given by:

$$C_{\text{total}} = \frac{A\epsilon}{t_i}$$

Since there are no charges inside the dielectric, the electric field in the insulator is constant and the electrostatic potential changes linearly from one plate to the other.

3.2 The MOS capacitor

In a MOS capacitor, we replace the lower plate by a semiconductor. Unlike a metal, a semiconductor can have charges distributed in its bulk. For the sake of an example, let us consider a P type semiconductor (Si) doped to 10^{16}atoms /cm^3. As we know, holes outnumber electrons in this semiconductor by an extremely large factor. If we place a negative charge on the upper plate, holes will be attracted by this charge, and will accumulate near the silicon-insulator interface. This situation is analogous to the parallel plate capacitor and thus, the capacitance will be the same as that for a parallel plate capacitor. If, however, we place a positive charge on the upper plate, negative charges will be attracted by it and positive charges will be repelled. In a P type semiconductor, there are very few electrons. The negative charge is provided by the ionised acceptors after the holes have been pushed away from them. But the acceptors are fixed in their locations and cannot be driven to the edge of the insulator. Therefore, the distance between the induced and inducing charges increases - so the capacitance is lower as compared to the parallel plate capacitor. As more and more positive charge is placed on the upper plate, holes

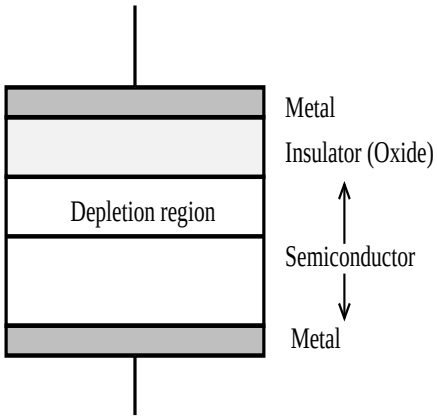


Figure 5: The MOS structure

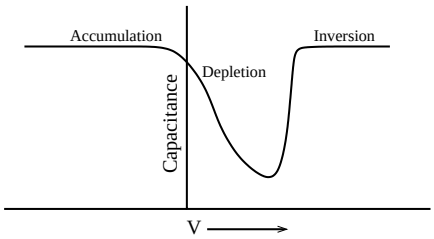


Figure 6: Low frequency capacitance for a MOS capacitor

from a thicker slice of the semiconductor are driven away, and the incremental induced charge is farther from the inducing charge. Thus the capacitance continues to decrease. This does not, however, continue indefinitely. We know from the law of mass action that as hole density reduces, the electron density increases. At some point, the hole density is reduced and electron density increased to such an extent that electrons now become the "majority" carriers near the interface. This is called *inversion*. Beyond this point, more positive charge on the upper plate is answered by more electrons in the semiconductor. But the electrons are mobile, and will be attracted to the silicon insulator interface. Therefore, the capacitance quickly increases to the parallel plate value.

3.3 Quantitative Analysis

Consider a one dimensional representation of the MOS structures as shown in the figure below. The origin is assumed to be at the silicon-oxide interface and the positive x direction is into the bulk of silicon. Using a one dimensional analysis, we want to relate the semiconductor charge to the applied gate voltage. In a practical case, there is a potential difference between two dissimilar materials in contact. Also, the silicon - oxide interface will have some fixed charge sitting there. However, we consider the ideal case first - where there is no built in contact potential between the semiconductor and the metal, and there is no interface charge.

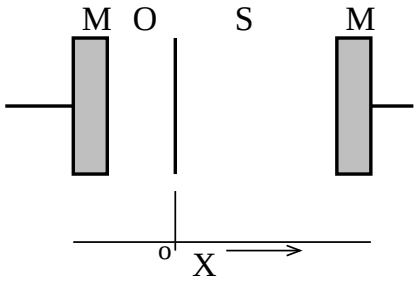
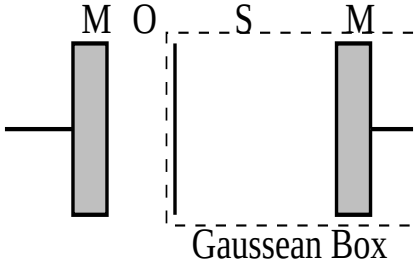


Figure 7: co-ordinate system used for analysis

3.3.1 Ideal Case

Let the back surface of Si be at zero potential and the voltage applied to the gate terminal be V_g. Let the electrostatic potential at any point x be denoted by $\phi(x)$ and let the potential at the silicon-oxide interface be ϕ_s.



We construct a Gaussian box passing through the interface and extending to $+\infty$. According to Gauss law, the integral of the outward pointing D vector around the box should be equal to the charge contained inside. The only boundary where D is non zero is the one passing through the interface. Therefore,

$$Area \times \epsilon_{ox} \frac{\phi_s - V_g}{tox} = \text{ Total Charge in silicon}$$

If we define Q_{si} to be the semiconductor charge *per unit area*, and C_{ox} to be the parallel plate capacitance *per unit area*, we get

$$V_g = \phi_s - \frac{Q_{si}}{C_{ox}}$$

Thus, the surface potential and the applied gate voltage can be related to each other. If the surface potential is known, we can evaluate the semiconductor charge by integrating the Poisson's equation in the semiconductor, once.

We can write the Poisson's equation in the semiconductor as

$$\nabla \cdot \mathbf{D} = \rho$$

or

$$-\epsilon_{si} \frac{\partial^2 \phi}{\partial x^2} = q(N_d^+ - N_a^- + p - n)$$

Since the electrostatic potential is dependent only on x, we can change partial derivatives to total derivatives.

$$-\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} = \frac{\mathrm{d}}{\mathrm{d}x}\left(-\frac{\mathrm{d}\phi}{\mathrm{d}x}\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\mathcal{E}\right)$$

where $\mathcal{E}$ is the electrostatic field. Changing the variable from x to ϕ.

$$-\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} = \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}x} = \left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)\frac{\mathrm{d}}{\mathrm{d}\phi}(\mathcal{E}) = -\mathcal{E}\frac{\mathrm{d}}{\mathrm{d}\phi}(\mathcal{E}) = -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\phi}\left(\mathcal{E}^2\right)$$

If we define

$$u \equiv \beta\phi \qquad \text{where } \beta \equiv \frac{q}{K_B T}$$

We get

$$-\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} = -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\phi}\left(\mathcal{E}^2\right) = -\frac{\beta}{2}\frac{\mathrm{d}}{\mathrm{d}u}\left(\mathcal{E}^2\right) \tag{34}$$

The right hand side of the Poisson's equation represents the charge density. In the absence of an applied voltage, this must be zero everywhere. Therefore,

$$q(N_d^+ - N_a^- + p_0 - n_0) = 0$$

where p_0 and n_0 represent the hole and electron density in the absence of an applied field. therefore,

$$N_d^+ - N_a^- = -(p_0 - n_0)$$

Substituting equation(34) and the above in the Poisson's equation,

$$-\frac{\beta\epsilon_{si}}{2}\frac{\mathrm{d}}{\mathrm{d}u}\left(\mathcal{E}^2\right) = q\left[p - p_0 - (n - n_0)\right]$$

so

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(\mathcal{E}^2\right) = -\frac{2qp_0}{\beta\epsilon_{si}}\left[\frac{p}{p_0} - 1 - \frac{n_0}{p_0}\left(\frac{n}{n_0} - 1\right)\right]$$

From equation(8)

$$n = n_0 e^u \qquad \text{and} \qquad p = p_0 e^{-u}$$

So,

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(\mathcal{E}^2\right) = -\frac{2qp_0}{\beta\epsilon_{si}}\left[e^{-u} - 1 - \frac{n_0}{p_0}(e^u - 1)\right]$$

This can be integrated from x = ∞ (where $\mathcal{E} = 0$ and u = 0) to x to give

$$\mathcal{E}^2 = \frac{2qp_0}{\beta\epsilon_{si}}\int_0^u \left[-e^{-u} + 1 + \frac{n_0}{p_0}(e^u - 1)\right]\mathrm{d}u$$

Integrating and putting limits at 0 and u, we get

$$\mathcal{E}^2 = \frac{2qp_0}{\beta\epsilon_{si}}\left[e^{-u} - 1 + u + \frac{n_0}{p_0}(e^u - 1 - u)\right]$$

Therefore

$$\mathcal{E} = \pm\sqrt{\frac{2qp_0}{\beta\epsilon_{si}}}\left[e^{-u} - 1 + u + \frac{n_0}{p_0}(e^u - 1 - u)\right]^{\frac{1}{2}}$$

And thus, the displacement vector D can be evaluated as:

$$D = \epsilon_{si}\mathcal{E} = \pm\sqrt{\frac{2qp_0\epsilon_{si}}{\beta}}\left[e^{-u} - 1 + u + \frac{n_0}{p_0}(e^u - 1 - u)\right]^{\frac{1}{2}} \tag{35}$$

This equation permits us to calculate D, given the value of u. In particular, the value of D at the surface (which is required for integration over the Gaussean box), can be evaluated from u_s.

In fact if u is very small, the exponentials in u can be expanded to second order. The first two terms cancel with 1 and u, leaving

$$D = \epsilon_{si}\mathcal{E} = \pm\sqrt{\frac{2qp_0\epsilon_{si}}{\beta}}\left[u^2/2 + \frac{n_0}{p_0}(u^2/2)\right]^{\frac{1}{2}}$$

$$\frac{\partial u}{\partial x} \simeq \mp\sqrt{\frac{q\beta p_0}{\epsilon_{si}}}\left(1 + \frac{n_0}{p_0}\right)u$$

This leads to exponential solutions for u with a characteristic length $L_D = \sqrt{\frac{\epsilon_{si}}{q\beta p_0}}$. This implies that small local perturbations in potential tend to decrease exponentially, with this characteristic length. This length is known as the extrinsic Debye Length.

For the p doped semiconductor under consideration,
$\frac{n_0}{p_0} << 1$, so $(1 + \frac{n_0}{p_0}) \simeq 1$. In this case, the characteristic length (known as the extrinsic Debye Length) is $L_D = \sqrt{\frac{\epsilon_{si}}{q\beta p_0}}$.
(In the intrinsic case, $n_0 = p_0$, so $1 + \frac{n_0}{p_0} = 2$).
Thus, in the intrinsic case, we get an additional factor of 2 in the denominator under the square root.)

By putting $u = u_s$ in eq. 35, we get the D vector at the surface. We construct a Gaussean box passing through the interface and enclosing the semiconductor (as described in section 3.3.1) The charge contained in the box is then the integral of the outward pointing D vector over the surface of the box. D is non zero only at the interface. The outward pointing D is along the negative x axis. Therefore by application of Gauss theorem,

$$\text{Sem. Charge} = \text{Area} \times (-D)$$

Hence the charge in the semiconductor *per unit area* is:

$$\begin{aligned} Q_{si} &= \mp\frac{\sqrt{2}\epsilon_{si}}{\beta L_D}\left[e^{-u_s} - 1 + u_s + \frac{n_0}{p_0}\left(e^{u_s} - 1 - u_s\right)\right]^{\frac{1}{2}} \qquad (36) \\ \text{where} \quad u_s &\equiv \beta\phi_s \\ \beta &\equiv \frac{q}{K_B T} \\ \text{and} \quad L_D &\equiv \sqrt{\frac{\epsilon_{si}}{q\beta p_0}} = \text{The Extrinsic Debye Length} \end{aligned}$$

Notice that Q_{si} is the charge in the semiconductor *per unit area*. In this treatment, we shall use symbols of the type Q and C with various subscripts to denote the corresponding charges and

capacitance values *per unit area*. Q_{si} consists of mobile as well as fixed charge. The mobile charge is contributed by holes when $u_s < 0$ and by electrons when $u_s > 0$ (for a P type semiconductor). As we shall see later, the mobile electron charge is substantial only when the positive surface potential exceeds a threshold value.

The fixed charge is contributed by the depletion charge when the surface potential is positive. The depletion charge per unit area can be calculated by the depletion formula.

$$Q_{depl} = -qN_aX_d = \sqrt{2qN_a\epsilon_{si}\phi_s} \qquad (\phi_s > 0)$$

A somewhat more accurate expression for depletion charge accounts for slightly lower charge density at the edge of the depletion region by subtracting K_BT/q from ϕ_s.

$$Q_{depl} = -qN_aX_d = \sqrt{2qN_a\epsilon_{si}(\phi_s - K_BT/q)} \qquad (\phi_s > K_BT/q) \tag{37}$$

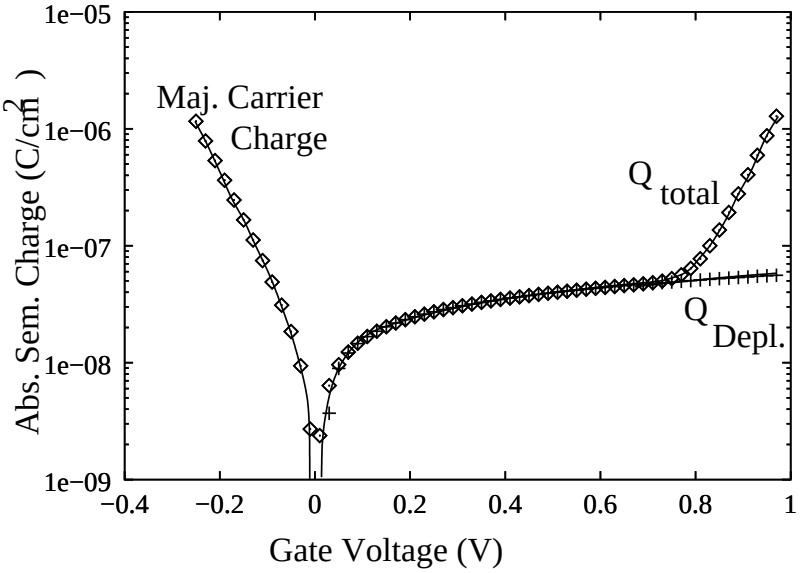


Figure 8: semiconductor charge as a function of surface potential

Calculated values for the total semiconductor charge per unit area (ie. inclusive of depletion and mobile charge) and just the depletion charge per unit area have been plotted in figure 8 for a P type semiconductor doped to $10^{16}/\text{cm}^3$. For small positive surface potential, the total semiconductor charge contains only depletion charge. However, beyond a surface potential near $2\Phi_F$, the total charge exceeds the depletion charge very rapidly. This additional charge is due to mobile minority carriers (in this case, electrons).

3.3.2 Practical case

A practical MOS structure will differ from the ideal case assumed above in a few respects. There is a built-in potential difference between the metal used and Si, due to the difference between their work functions. This shifts the relationship between $V_g$and ϕ_s. Also, there is a fixed oxide charge which resides essentially at the silicon-oxide interface. Thus, the total charge in the Gaussian

box includes this fixed charge and the semiconductor charge. These two non-idealities can be accounted for by modifying the relationship between $V_g$and $\phi_s$to be

$$V_g = \Phi_{ms} + \phi_s - \frac{Q_{si} + Q_{ox}}{C_{ox}} \tag{38}$$

Where Φ_{ms} is the metal to semiconductor work function difference.

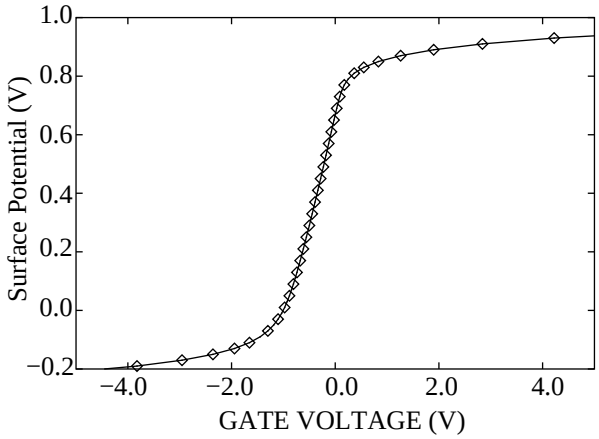


Figure 9: Surface potential as a function of gate voltage

Figure 9 shows the surface potential as a function of applied voltage for a MOS capacitor with oxide thickness of 22.5 nm, substrate doping of 10^{16}/cc, oxide charge of 4×10^{10}q and aluminium as the gate metal. The surface potential changes quite slowly as a function of gate voltage in the accumulation and inversion regions.

The absolute value of semiconductor charge has been plotted as a function of applied gate voltage

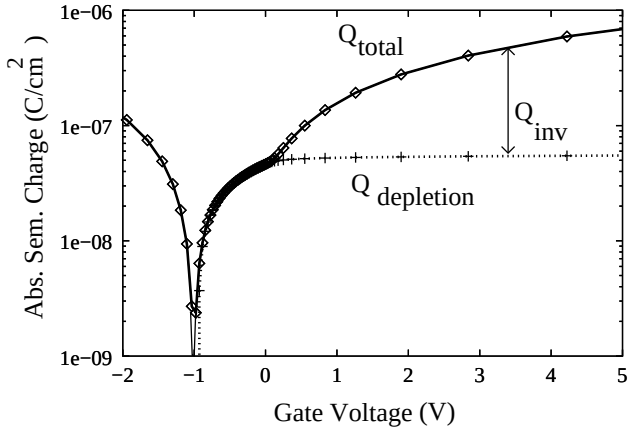


Figure 10: Semiconductor charge as a function of gate voltage

in figure 10. (The charge is actually negative for positive gate voltages). As one can see, for small positive gate voltages, the entire semiconductor charge is depletion charge. As the voltage exceeds a threshold voltage, the total charge becomes much larger than the depletion charge. The excess charge is provided by mobile electron charges. This is the *inversion* region of operation, where electrons become the majority carriers near the surface in a p type semiconductor. Notice that the depletion charge is practically constant in this region. This region begins when the surface potential exceeds $2\Phi_{\mathrm{F}}$.

4 The MOS Transistor

Inversion converts a p type semiconductor to n type at the surface. We can use this fact to construct a transistor. We place semiconductor regions strongly doped to N type on either side of a MOS capacitor made using P type silicon. Now if we try to pass a current between these

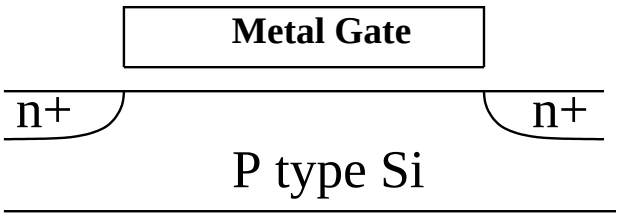


Figure 11: A MOS Transistor

two N regions when inversion has not occurred, we encounter series connected NP and PN diodes on the way. Whatever the polarity of the voltage applied to pass current, one of these will be reverse biased and practically no current will flow.

However, after inversion, the intervening P region would have been converted to N type. Now there are no junctions as the whole surface region is n type. Current can now be easily passed between the two n regions. This structure is an n channel MOS transistor. pMOS transistors can be similarly made using P regions on either side of a MOS capacitor made on n type silicon. When current flows in an n channel transistor, electrons are supplied by the more negative of the two n+ contacts. This is called the source electrode. The more positive n+ contact collects the electrons and is called the drain. The current in the transistor is controlled by the metal electrode on top of the oxide. This is called the gate electrode.

4.1 I-V characteristics of a MOS transistor

A quantitative derivation of the current-voltage characteristics of the MOS device is complicated by the fact that it is inherently a two dimensional device. The *vertical* field due to the gate voltage sets up a mobile charge density in the channel region as seen in figure 10. The *horizontal* field due to source-drain voltage causes these charges to move, and this constitutes the drain current. Therefore, a two dimensional analysis is required to calculate the transistor current, which can be quite complex. However, reasonably simple models can be derived by making several simplifying assumptions.

4.2 A simple MOS model

We make the following simplifying assumptions:

- The vertical field is much larger than the horizontal field. Then, the resultant field is nearly vertical, and the results derived for the 1 dimensional analysis for the MOS capacitor can be used to calculate the point-wise charge density in the channel. This is known as the *gradual channel* approximation. Accurate numerical simulations have shown that this approximation is valid in most cases.

- The source is shorted to the bulk.

- The gate and drain voltages are such that a continuous inversion region exists all the way from the source to the drain.
- The depletion charge is constant along the channel.
- The total current is dominated by drift current.
- The mobility of carriers is constant along the channel.

Figure 12 shows the co-ordinate system used for evaluating the drain current. The x axis points

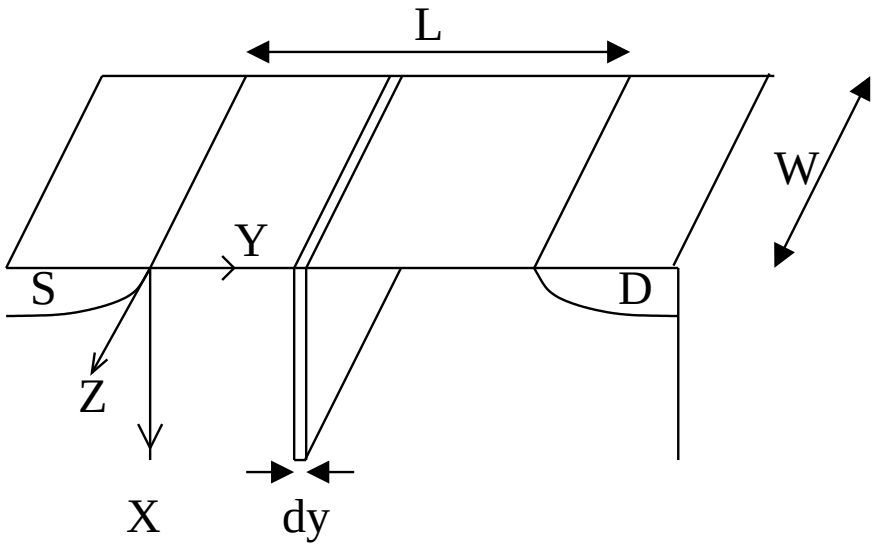


Figure 12: Coordinate system used for analysing the MOS transistor

into the semiconductor, the y axis is from source to the drain and the z axis is along the width of the transistor. The origin is at the source end of the channel. We represent the channel voltage as V(y), which is 0 at the source end and V_d at the drain end. We assume the current to be made up of just the drift current. Since we are carrying out a quasi 2 dimensional analysis, all variables are assumed to be constant along the z axis. Let n(x,y) be the concentration of mobile carriers (electrons for an n channel device) at the position x,y (for any z). The drift current density at a point is

$$\begin{aligned} J &= \text{no. of carriers} \times \text{charge per carrier} \times \text{velocity} \\ &= n(x,y) \times (-q) \times \mu \times \left(-\frac{\partial V(y)}{\partial y}\right) \\ &= \mu n(x,y) q \frac{\partial V(y)}{\partial y} \end{aligned}$$

Integrating the current density over a semi-infinite plane at the channel position y (as shown in the figure 12) will then give the drain current.

$$I_d = \int_{x=0}^{\infty} \int_{z=0}^{W} \mu n(x,y) q \frac{\partial V(y)}{\partial y} dz dx$$

Since there is no dependence on z, the z integral just gives a multiplication by W. Therefore,

$$I_d = \mu W q \int_{x=0}^{\infty} n(x,y) \frac{\partial V(y)}{\partial y} dx$$

the value of n(x,y) is non zero in a very narrow channel near the surface. We can assume that $\frac{\partial V(y)}{\partial y}$ is constant over this depth. Then,

$$I_d = \mu W q \frac{\partial V(y)}{\partial y} \int_{x=0}^{\infty} n(x,y) dx$$

but $q \int_{x=0}^{\infty} n(x,y)dx = -\mathrm{Q_n}(y)$ where $\mathrm{Q_n}$(y) is the electron charge per unit area in the semiconductor at point y in the channel. ($\mathrm{Q_n}$(y) is negative, of course). therefore

$$I_d = -\mu W \frac{\partial V(y)}{\partial y} \mathrm{Q_n}(y) \tag{39}$$

Integrating the drain current along the channel gives

$$\begin{aligned} \int_0^L I_d dy &= -\mu W \int_0^L \mathrm{Q_n}(y) \frac{\partial V(y)}{\partial y} dy \\ I_d \times L &= -\mu W \int_0^{V_d} \mathrm{Q_n}(y) \mathrm{d}V(y) \\ \text{So, } I_d &= -\mu \frac{W}{L} \int_0^{V_d} \mathrm{Q_n}(y) \mathrm{d}V(y) \end{aligned}$$

We now use the assumption that the surface potential due to the vertical field saturates around $2\Phi_\mathrm{F}$if we are in the inversion region. Therefore, the total surface potential at point y is V(y) + 2 Φ_F. Now, by Gauss law and continuity of normal component of D at the interface,

$$C_{ox}\left(V_g - \Phi_{\mathrm{MS}} - \phi_s\right) = -\left(Q_{si} + Q_{ox}\right)$$

therefore,

$$-Q_{si} = C_{ox}\left(V_g - \Phi_{\mathrm{MS}} - V(y) - 2\Phi_\mathrm{F} + Q_{ox}/C_{ox}\right)$$

However,

$$Q_{si} = \mathrm{Q_n} + \mathrm{Q_{depl}}$$

So

$$\begin{aligned} -\mathrm{Q_n}(y) &= -Q_{si}(y) + \mathrm{Q_{depl}} \\ &= C_{ox}\left(V_g - \Phi_{\mathrm{MS}} - V(y) - 2\Phi_\mathrm{F} + (Q_{ox} + \mathrm{Q_{depl}})/C_{ox}\right) \end{aligned}$$

We have assumed the depletion charge to be constant along the channel. Let us define

$$\mathrm{V_T} \equiv \Phi_{\mathrm{MS}} + 2\Phi_\mathrm{F} - \frac{(Q_{ox} + \mathrm{Q_{depl}})}{C_{ox}}$$

then

$$-\mathrm{Q_n}(y) = C_{ox}(V_g - \mathrm{V_T} - V(y))$$

and therefore,

$$\begin{aligned} I_d &= \mu C_{ox}\frac{W}{L}\int_0^{V_d}(V_g - \mathrm{V_T} - V(y))\mathrm{d}V(y) \\ &= \mu C_{ox}\frac{W}{L}[(V_g - \mathrm{V_T})V_d - \frac{1}{2}V_d^2] \end{aligned} \tag{40}$$

This derivation gives a very simple expression for the drain current. However, it requires a lot of simplifying assumptions, which limit the accuracy of this model.
If we do not assume a constant depletion charge along the channel, we can apply the depletion formula to get its dependence on V(y).

$$\mathrm{Q_{depl}} = -\sqrt{2\epsilon_{si}qN_a(V(y) + 2\Phi_\mathrm{F})}$$

then,

$$-\mathrm{Q_n} = C_{ox}\left(V_g - \Phi_\mathrm{MS} - V(y) - 2\Phi_\mathrm{F}\right) + Q_{ox} - \sqrt{2\epsilon_{si}qN_a(V(y) + 2\Phi_\mathrm{F})}$$

which leads to

$$\begin{aligned} I_d = \mu C_{ox}\frac{W}{L}\Bigg[&\left(V_g - \Phi_\mathrm{MS} - 2\Phi_\mathrm{F} + \frac{Q_{ox}}{C_{ox}}\right)V_d - \frac{1}{2}V_d^2 \\ &- \frac{2}{3}\frac{\sqrt{2\epsilon_{si}qN_a}}{C_{ox}}\left((V_d + 2\Phi_\mathrm{F})^{3/2} - (2\Phi_\mathrm{F})^{3/2}\right)\Bigg] \end{aligned}$$

This is a more complex expression, but gives better accuracy.

4.3 Modeling the saturation region

The treatment in the previous section is valid only if there is an inversion layer all the way from the source to the drain. For high drain voltage, the local vertical field near the drain is not adequate to take the semiconductor into inversion. Several models have been used to describe the transistor behaviour in this regime. The simplest of these defines a saturation voltage at which the channel just pinches off at the drain end. The current calculated for this voltage by the above models is then supposed to remain constant at this value for all higher drain voltages. The pinch-off voltage is the drain voltage at which the channel just vanishes near the drain end. Therefore, at this point the gate voltage $\mathrm{V_g}$ is just less than a threshold voltage above the drain voltage $\mathrm{V_d}$. Thus, at this point,

$$V_{dsat} = V_g - V_T$$

The current calculated at $\mathrm{V_{dsat}}$ will be denoted as $\mathrm{I_{dss}}$. Thus,

$$I_{dss} = \mu C_{ox}\frac{W}{L}[(V_g - \mathrm{V_T})^2 - \frac{1}{2}(V_g - \mathrm{V_T})^2]$$

for the simple transistor model. Thus

$$I_{dss} = \frac{1}{2}\mu C_{ox}\frac{W}{L}(V_g - \mathrm{V_T})^2 \tag{41}$$

The drain current is supposed to remain constant at this $\mathrm{V_d}$ independent value for all drain voltages $> V_g - V_T$.

4.3.1 Early Voltage approach

Assuming a constant current in the saturation region leads to an infinite output resistance. This can lead to exaggerated estimates of gain from an amplifier. Therefore, we need a more realistic model for the transistor current in the saturation region. One of these is a generalisation of the model proposed by James Early for bipolar transistors. This model is not strictly applicable to MOS transistors. However, due to its numerical simplicity, it is often used in compact models for circuit simulation.

A geometrical interpretation of the Early model states that the drain current increases linearly in the saturation region with drain voltage, and if saturation characteristics for different gate voltages are produced backwards, they will all cut the drain voltage axis at the same (negative) drain voltage point. The absolute value of this voltage is called the Early Voltage V_E.

The current equations in saturation mode now become:

$$\begin{aligned} I_{dss} &\equiv I_d(V_g, V_{dss}) \\ I_d &= I_{dss}\frac{V_d + V_E}{V_{dss} + V_E} \qquad \text{For } V_d > V_{dss} \end{aligned} \tag{42}$$

Any model can be used for calculating the drain current for $V_d < V_{dss}$. The value of V_{dss} will be determined by considerations of continuity of the drain current and its derivative at the changeover point from linear to saturation regime. For example, if we use the simple model described in eq. 40,

$$\begin{aligned} \frac{\partial I_d}{\partial V_d} &= \mu C_{ox}\frac{W}{L}\left(V_g - V_T - V_d\right) \qquad \text{For } V_d \le V_{dss} \\ \text{And} \qquad \frac{\partial I_d}{\partial V_d} &= \frac{I_{dss}}{V_{dss} + V_E} \qquad \text{For } V_d \ge V_{dss} \\ \text{Where} \qquad I_{dss} &\equiv \mu C_{ox}\frac{W}{L}\left[\left(V_g - V_T\right)V_{dss} - \frac{1}{2}V_{dss}^2\right] \end{aligned}$$

On matching the value of $\frac{\partial I_d}{\partial V_d}$ on both sides of V_{dss}, we get

$$V_{dss} = V_E\left(\sqrt{1 + \frac{2\left(V_g - V_T\right)}{V_E}} - 1\right)$$

In practice, V_E is much larger than $V_g - V_T$. If we expand the above expression, we find that to first order the value of V_{dss} remains the same as the one used in the simple model - that is, $V_g - V_T$. Expansion to second order gives

$$V_{dss} \simeq \left(V_g - V_T\right)\left(1 - \frac{V_g - V_T}{2V_E}\right) \tag{43}$$

4.3.2 Simulation Model

Since the value of V_{dss} does not change substantially from the ideal saturation case, a simpler approach can be tried. The drain current is calculated using the ideal saturation model and its value is multiplied by a correction factor $= (1 + \lambda V_d)$ in saturation *as well as* in linear regime. This automatically assures continuity of I_d and its derivative. λ is a fit parameter, whose value is $\approx 1/V_E$. This approach is used in SPICE, a popular circuit simulation program.

5 MOS Device Scaling

Since the transistor current depends on W/L, it is interesting to see what happens if we reduce both W and L, keeping their ratio constant. We have to adjust other parameters as well, in order to ensure that the transistor works without problems.

Due to technological constraints, we cannot reduce lateral geometries without reducing layer thicknesses. To define finer lateral dimensions through etching etc., we need the layers to be thinner. Thus all dimensions, vertical or lateral, need to be scaled by the same factor. To ensure that higher fields in the device do not cause breakdown, we have to scale down all the voltages by the same factor as L. (This is known as constant field scaling).

We also need to scale depletion widths in the same ratio as W and L. This is essential in order to scale down the separation between transistors and to control channel length modulation due to drain voltage. This requires doping densities to be scaled up by the same factor as the one used to scale down W and L.

So we define a scaling factor S, and reduce W, L, junction depths and oxide thicknesses etc. by this factor. Doping densities need to be increased by factor S. All working voltages and the Threshold voltage V_T need to be scaled down by S. Once this scaling is done, we are interested in evaluating the impact on the circuit performance.

5.1 Consequences of Scaling

We assume classical or Constant Field scaling in the following.

Device Area: Since W reduces by $\downarrow S$ and L reduces by $\downarrow S$, the area reduces by $\downarrow S^2$.

Packing Density: For a given chip area, the number of devices which can be packed in this chip will go up by $\uparrow S^2$.

C_{ox}: The gate capacitance per unit area is given by ϵ/t_{ox}. Since t_{ox} scales down by $\downarrow S$, C_{ox} increases by $\uparrow S$. C_{ox} determines the transconductance, so this increase is good.

Load capacitance: All dimensions, including depletion widths have been scaled down by $\downarrow S$. Total capacitance $= \epsilon A/t$. Now A reduces by $\downarrow S^2$, while the dielectric thickness (be it oxide or depletion width) reduces by $\downarrow S$. The net effect is that total capacitance $= \epsilon$Area $\downarrow S^2/t \downarrow S$ reduces by $\downarrow S$.

Voltages: All voltages such as V_{DS}, V_{GS}, V_T etc. are scaled down by $\downarrow S$ to keep the field constant.

Drain current: I_{DS} is given by $\mu C_{ox}(W/L)f(V_{DS}, V_{GS}, V_T)$. Since all voltages are scaled down by $\downarrow S$ and f is a square function of voltages both in linear mode and saturation, f will scale down as $\downarrow S^2$. Thus,

$$I_{DS} = \mu C_{ox}(\uparrow S)(W \downarrow S/L \downarrow S)f(V_{DS}, V_{GS}, V_T)(\downarrow S^2)$$

So combining all dependences, $I_{DS} \downarrow S$.

Slew Rate: Slew rate is the rate of change of voltage at any node. Since $I = C\frac{dV}{dt}$, the slew rate goes as $I(\downarrow S)/C(\downarrow S)$. Thus slew rate remains unchanged with scaling.

Delay: Delay is given by the total voltage change divided by $\frac{dV}{dt}$. Since all voltages are scaled down by $\downarrow S$, while $\frac{dV}{dt}$ remains unchanged, the delay reduces as $\downarrow S$.

Static Power: It is given by $V \times I$. So it scales as $(\downarrow S)(\downarrow S)$, that is $\downarrow S^2$.

Dynamic Power: Dynamic power is given by $C_{total}V^2f$. This scales as $(\downarrow S)(\downarrow S^2)(\uparrow S)$. Thus dynamic power reduces as $\downarrow S^2$ even when the frequency of operation is increased by $\uparrow S$ to take advantage of shorter delays, which scale down by $(\downarrow S)$.

With all dimensions and voltages divided by the factor S(> 1), We can summarise the impact of using constant field Scaling as follows:

Device area	$\propto W \times L : (\downarrow S)(\downarrow S)$	$\downarrow S^2$
C_{ox}	ϵ_{ox}/t_{ox} : const$/(\downarrow S)$	$\uparrow S$
C_{total}	$\epsilon A/t : (\downarrow S^2)/(\downarrow S)$	$\downarrow S$
V_{DS}, V_{GS}, V_T	Voltages : $(\downarrow S)$	$\downarrow S$
I_d	$\mu C_{ox}(W/L)(\propto V^2)$: $(\uparrow S)$(const)$(\downarrow S^2)$	$\downarrow S$
Slew Rate $\frac{dV}{dt}$	$I/C_{total} : (\downarrow S)/(\downarrow S)$	*const.*
Delay	$V/\frac{dV}{dt} : (\downarrow S)/(const)$	$\downarrow S$
Static Power	$V \times I : (\downarrow S)(\downarrow S)$	$\downarrow S^2$
dynamic power	$C_{total}V^2f : (\downarrow S)(\downarrow S^2)(\uparrow S)$	$\downarrow S^2$
Power delay product	delay $\times$ power$(\downarrow S)(\downarrow S^2)$	$\downarrow S^3$
Power density	power/area : $(\downarrow S^2)/(\downarrow S^2)$	*const.*

Thus, scaling leads to:

- Improved packing density: $\uparrow S^2$
- Improved speed: delay $\downarrow S$
- Improved power consumption: $\downarrow S^2$

So, circuit performance improves dramatically with transistor scaling. This provides the motivation for making transistors as small as possible.

What are the limits on scaling?

These come from processing technology limitations, device limitations and circuit considerations such as reduced signal to noise ratio due to reduced supply voltages.

5.2 Moore's "Law"

In 1965, Gordon Moore, the co-founder of Fairchild Semiconductor as well as Intel, described a doubling every year in the number of components per integrated circuit. It is an observation of a trend and an empirical relationship – not a physical or natural law! However, given the

prominence of Gordon Moore, it is widely referred to as Moore's Law.

In 1975, Moore modified his observation for the rate of device scaling and predicted a doubling of device density every two years. It is remarkable that this trend has continued over several decades. It is only in the last decade that the rate of doubling has slowed down remarkably, as we hit several physical limits.

Device scaling started initially as an empirical observation. The theoretical basis of constant field device scaling was laid down in a landmark paper in 1974 from a group of scientists from IBM.

R.H. Dennard, F. H. Gaensslen, Hw A-Nien Yu, V. L. Rideout, E. Bassous, and A. R, LeBlanc, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions", IEEE Journal of Solid-State Circuits, Vol. SC-9, No. 5, pp. 256-268, 1974.

This is one of the most quoted papers in this field. I strongly recommend that you read it – for its contents, but also to learn from its style of technical writing.

5.3 The technology road map

The incredible rate of increase in circuit performance has been possible through careful planning. The semiconductor industry used Moore's prediction for setting specific targets for development in process technology, processing equipment and for research and development in critical areas of device Physics. The result of this planning was the creation of an International Technology Roadmap for Semiconductor Scaling – or ITRS. ITRS has been revised every year till recently. A new ITRS has not been issued after 2016.

It is hard to track and scale the optimum size of numerous structures on an Integrated circuit. As discussed earlier, it is common to describe feature sizes in units of a parameter called λ. Now sizes of various structures can be described in units of λ. As we scale technologies, we just scale the value of λ. Feature sizes remain the same in units of λ, which is convenient.

The smallest feature on a chip is the contact window. The value of λ is so defined that the smallest feature size is 2λ. The smallest registration rule – for example the extent to which a contact window must be inside a diffused region – is λ.

As a result of careful planning and the considerable financial rewards of improved MOS technology, feature sizes have been continually scaled. The table below gives the commonly used channel lengths by the year in various decades.

1971	10 μm	1974	6 μm	1977	3 μm				
1981	1.5 μm	1984	1 μm	1987	800 nm				
1990	600 nm	1993	350 nm	1996	250 nm	1999	180 nm		
2001	130 nm	2003	90 nm	2005	65 nm	2007	45 nm	2009	32 nm
2012	22 nm	2014	14 nm	2017	11 nm				

The scaling rate has slowed down after 2010. This is because feature sizes had already reached about 20 nm – about 3% of the wavelength of sodium light!

5.4 Demand from Processing Technology

Circuit performance improves dramatically with transistor scaling. This provides the motivation for making transistors as small as possible.
What demands does it place on processing technology?

- Scaling requires much higher resolution in defining geometries. Size of the finest patterns in the state of the art technologies is about 10nm. This is about a fiftieth of the wavelength of sodium light!
- Advanced photo-lithographic techniques need to be used to define such fine geometries. We need deep UV lithography and even XRay lithography to define such fine structures.
- Etching techniques have to be improved to define such fine structures. Dry etching using plasma or reactive ion etching is used rather than wet chemical etching to define such fine structures

5.5 Limits of scaling

Scaling is being limited now due to several reasons.

- We are reaching limits of resolution possible with photo-selective processes and etching etc.
- Traditional Device Physics is not valid any more for such small structures.

Remember, the lattice constant of Silicon is ≈ 0.5 nm. So there are as few as 20 atoms between source and drain of a 10 nm channel MOSFET. Clearly, conduction models based on statistics will not hold here.

Indefinite voltage scaling is not possible. If the voltage is scaled down drastically,

- signal to noise ratio will become poor,
- leakage currents will become dominant as KT/q has not been scaled and current equations of junctions involve qV/KT.
- System considerations such as interconnect delay will limit performance gain.
- At low voltages, supplying power requires higher currents. Feeding such high currents through IC pins becomes impractical.

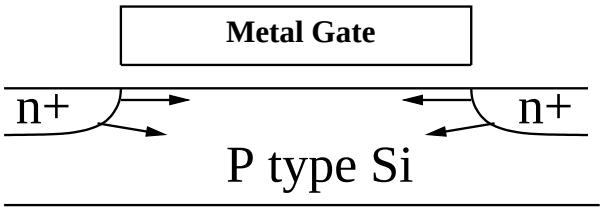


Figure 13: Sideways diffusion from source drain regions

5.6 Short Channel Effects

Several effects become prominent once transistor dimensions are made very small. The classical model for MOS transistors assumed a uniform doping in the channel region. However, because of heavy doping in the source/drain region, there is a sideways diffusion of impurities into the channel. If the channel length is quite short, the region of non-uniform doping becomes a large fraction of the channel length. This results in considerable deviations from the transistor model.

Threshold voltage of a long channel transistor is independent of channel length. As we scale down channel length, the threshold voltage becomes dependent on channel length. (Short channel effect on V_T).

Also, as the drain comes closer to the source, the field due to the drain channel junction reaches the source channel junction. This reduces the barrier to carrier injection from the source into the channel. This is known as Drain Induced Barrier Lowering or DIBL.

5.7 Narrow Channel Effects

As we scale down devices, channel widths as well as channel lengths are reduced. The threshold voltage becomes dependent on channel width as well as channel length for scaled down devices. This is because the depletion charge on the sides of the channel is no more negligible compared to the charge directly under the gate. For uniformly doped devices, V_T increases as the channel is made narrower. However, the dependence is more complex when doping is non-uniform.

6 Breakdown Phenomena

6.1 Avalanche Breakdown

The drain channel junction is reverse biased. In saturation region, there is high field region next to the gate. If the field exceeds some critical value, carrier multiplication will occur, leading to avalanche breakdown. Multiplication produces excess electron-hole pairs. Electrons are attracted towards the positively biased drain. Holes drift towards the source and constitute a "base current" for the parasitic lateral npn transistor.
Carrier multiplication near drain can result in a sharp increase in drain current. This is the avalanche breakdown of the transistor. It is also possible that the parasitic bipolar turns on, due

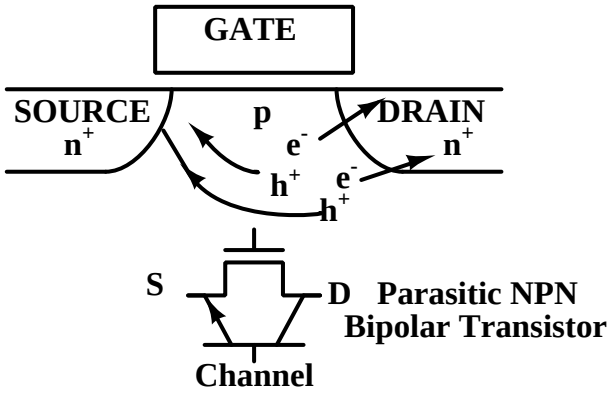


Figure 14: Avalanche breakdown at high fields

to the base current provided by the drifting holes from the drain junction, adding its current to the drain current. The additional current due to the bipolar action, combined with carrier multiplication near the drain can result in early breakdown of the transistor.

6.2 Punch Through

If the channel is very short, at high drain voltages, the depletion region due to the drain-substrate junction can reach the source. Due to the drain field, the source/substrate junction will get forward biased and will inject current into the channel, even if the gate voltage is below V_T. This is an extreme case of drain induced barrier lowering and results in heavy current even though the transistor is supposed to be 'OFF'. This is known as "Punch Through".

7 Parasitic Devices

7.1 Field transistors

As we make the devices needed for the desired circuit, several other devices get formed. The most common of these is the field transistor. A MOS like structure exists between unrelated diffusion areas due to metal lines crossing over unrelated diffusion areas.

7.2 Latch up due to parasitic pnpn structures

Fig.16 shows the lay out of an inverter. (As we shall learn later, this is a bad layout!). While the lay out does form an inverter as desired, it also forms a parasitic latch-up structure which can turn on, shorting V_{DD} to ground and destroying the IC due to the resulting heavy current.
A vertical pnp transistor is formed by

1. the p+ source of a pMOS transistor connected to V_{DD} (which becomes the emitter),
2. the n well (which becomes the base), and
3. the p substrate (which becomes the collector of this transistor).

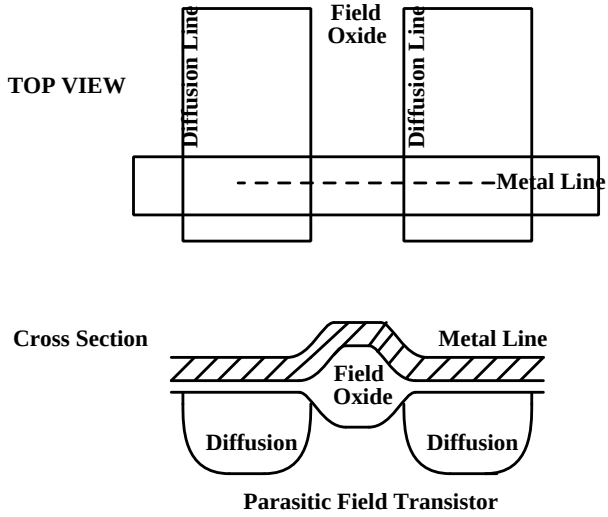


Figure 15: Parasitic Field transistor in MOS technology

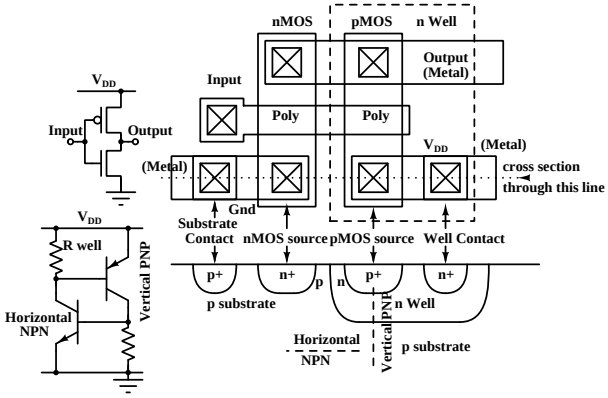


Figure 16: Formation of a parasitic latch up structure in an inverter

The n well is connected to V_{DD} through a resistive path, which represents the resistance of the n well to the well contact.
This can also be seen in Fig.17. A horizontal npn transistor is formed by

1. the n+ source of an nMOS transistor connected to ground (which becomes the emitter),
2. the p substrate, (which becomes the base), and
3. the n well, (which becomes the collector).

Since the collector of the npn and the base of the pnp are both formed by the n well, these two are connected. Similarly, the collector of the pnp and the base of the npn are formed by the p substrate, so these are connected too.

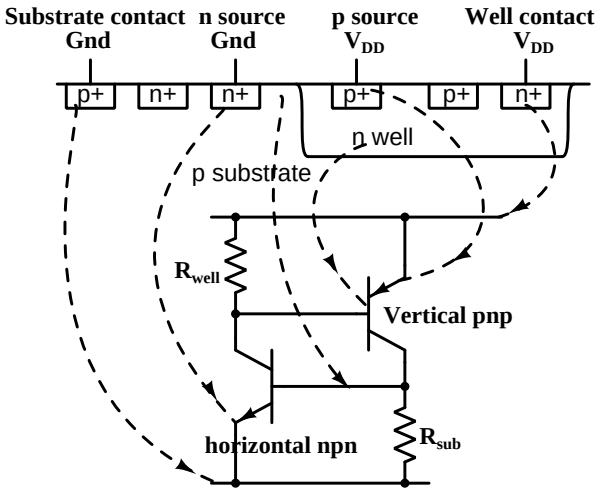


Figure 17: Circuit for the parasitic latch up circuit

From the equivalent circuit given in Fig.18, we can see that it forms a positive feedback circuit.

- An increase in the base current of the pnp will be amplified by its β_p and a large part of it will flow through the base emitter junction of the npn transistor.
- This part will be amplified by the β_n of the npn and a substantial part of it will go through the base emitter junction of the pnp transistor.

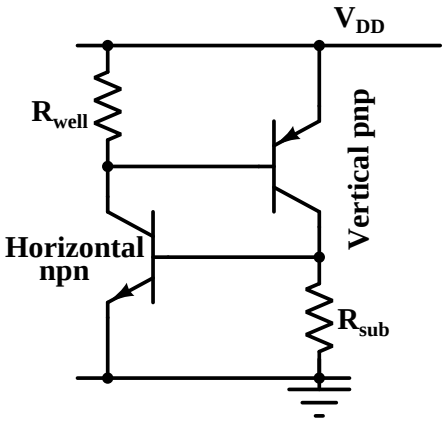


Figure 18: Positive feed back in the latch-up structure

If the product of the two amplification factors β_p and β_n and the current division ratios between the resistors and the base emitter junctions exceeds 1, the currents will keep increasing due to this feedback, till there is a dead short between V_{DD} and ground. This is called latch up.

7.3 Preventing Latch up

To prevent latch up, we must reduce the β of the parasitic bipolar transistors, and make sure that most of the collector current of either transistor is directed to the resistor and not to the base-emitter junction of the other transistor.

This can be done through process steps as well as through layout design rules.

The doping gradient of the n well should be made retrograde. (Doping should increase as we go deeper). This kills the current gain β_p of the pnp transistor.
The n well should have a guard ring connected to V_{DD}, which will collect any current which could form the base current of the pnp.

In layout, substrate and well contacts should be placed frequently, to reduce the value of R_{well} and $R_{substrate}$. n channel transistors should be placed far from the edge of the n well. This increases the base width of the npn transistor and kills its current gain. p channel transistors should also be placed far from the well edge and the n well should be deep to kill the gain of the pnp transistor.