9

Metal-Oxide-Semiconductor junction

9.1 Introduction

METAL semiconductor junctions are characterized by non-linear conductivity and non-linear capacitance. The introduction of an insulating layer between the metal and the semiconductor eliminates the conductivity and it gives rise to a capacitor where the non linear effects of the capacitance can be fully studied and exploited under any bias. In such a junction the state of the charges in the semiconductor can be modulated by the applied potential. This effect is known as *Field Effect* and it is of paramount importance in electronics because it is at the basis of the family of *Field Effect Transistors*.
The application of a bias voltage to a metal-insulator-metal capacitor creates an accumulation of electrons at the surface of one plate and a depletion of electrons on the other plate. These layers of charges are confined in ideal planes at the metal-insulator interfaces. When one of the plates is replaced by a semiconductor the effect of the potential is more complex and it results in the behaviour of the field effect devices.
Metal-insulator-semiconductor structures are a distinctive signature of silicon technology. The obvious insulator for silicon is its natural oxide (silicon oxide: SiO_2), thus the junction is called metal-oxide-semiconductor (MOS).
The oxidation of the crystalline silicon typically occurs in an atmosphere of either oxygen or water vapor kept at a temperature between $850°C$ and $1100°C$. The involved chemical reactions are :

$$Si_{solid} + O_{2(gas)} \rightarrow SiO_2$$
$$Si_{solid} + 2H_2O \rightarrow SiO_2 + 2H_2$$

The layer of oxide is made consuming the surface silicon atoms. In order to create a thick oxide film the oxygen has to diffuse through the formed oxide layer to reach and oxidise the buried silicon atoms. The diffusion of oxygen is favoured by the high temperature. Clearly, as the oxide becomes thicker the diffusion towards the silicon becomes less probable and the growth of the oxide becomes slower. When the process stops, a silicon oxide layer is formed on top of the silicon substrate. Since each silicon oxide unit contains one atom of silicon and two atoms of oxygen, the oxide layer is more thick than the consumed silicon layer. Approximately, 46% of the total height of the silicon oxide layer occupies the pristine silicon layer. Since the oxide is formed consuming the surface atoms of silicon, the silicon surface is displaced towards the interior of the crystal removing most of the

surface defects. Then the silicon close to the oxide is less defected than the original surface.
Thicker oxide layers, necessary for insulation purposes, can be formed with a chemical vapour deposition (CVD) process through the reaction in air of two gases: silane (SiH_4) and oxygen (O_2). The formed oxide molecules condense on the solid surface heated at few hundreds of centigrades.
The combination $Si - SiO_2$ is the most favourable case of a junction between different materials in terms of interface defects. This has an extraordinary consequence in the properties of the metal-oxide-semiconductor field effect transistors (MOSFETs).
In this chapter, the properties of the ideal MOS structure shown in Figure 1 are studied. The electric contact of the semiconductor is an ideal ohmic contact. The following analysis is valid for the combination of any metal insulator and semiconductor; clearly, in many practical cases, the defects at the interface between oxide and semiconductor may make very distant the theory from the real behaviour.

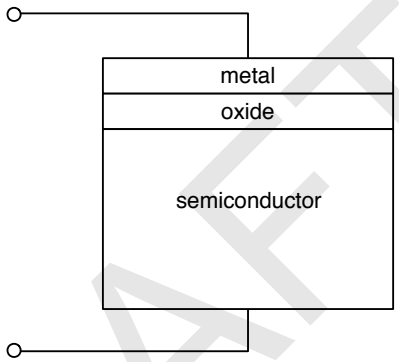

Fig. 9.1. *Structure of a ideal MOS structure.*

9.2 Band diagram and electrostatic quantities at the equilibrium

To study the MOS at the equilibrium, let us consider a structure made of aluminum, silicon dioxide, and p-type silicon. The quantities necessary to draw the bands diagrams are the following (the subscripts s, m, and ox stand for semiconductor, metal and oxide).

affinities: $q\chi_s = 4.05\ eV$; $q\chi_{ox} = 0.45\ eV$
work functions: $q\Phi_s = 4.9\ eV$; $q\Phi_m = 4.1\ eV$
band gaps: $E_{Gs} = 1.1\ eV$; $E_{Gox} = 8\ eV$

The parameters of the oxide might require a little discussion. The small affinity indicates the easy polarisability of the insulators, namely a little energy is sufficient to add electrons to an insulator. The band gap is so large that at room temperature the conduction band is empty ($exp(E_{Gox}/kT) \approx 10^{-134}$!). Finally, the Fermi level lies in the band gap but, however, the large band gap makes impossible the inter bands transitions, and thus the position of the Fermi level is

irrelevant. As a consequence, the oxide does not contribute to the charges redistribution necessary to establish the equilibrium.

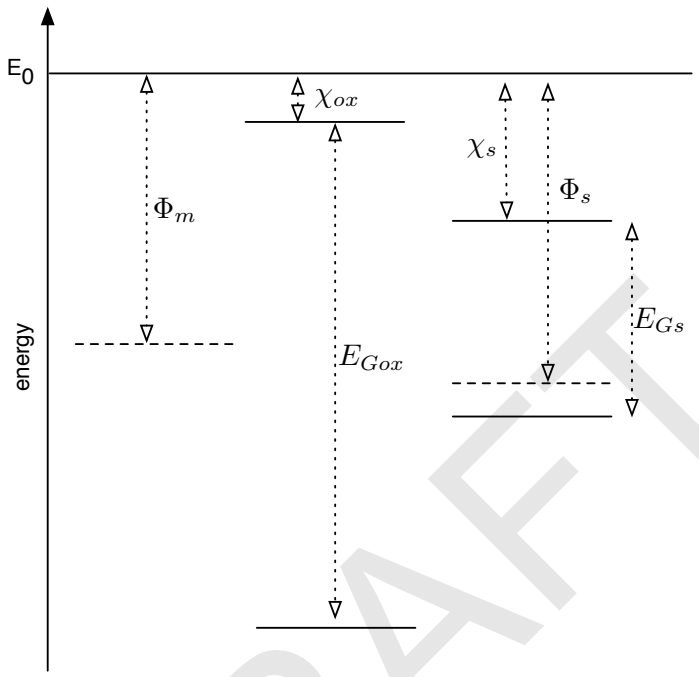

Fig. 9.2. *Band diagram of the three separated elements of the MOS .*

Figure 2 shows the bands diagrams of the insulated materials. Since the work function of the semiconductor is greater than the work function of the metal, the equilibrium is reached through a transfer of electrons from the metal to the semiconductor. Such a charge transfer is straightforward in a metal-semiconductor junction, but in this case, the two materials are kept separated by the insulator that should avoid any charge transfer. Then, the ideal structure shown in figure 1 should not allow the establishment of the equilibrium. However, although very large, the resistivity of the silicon oxide is not infinite and then the charges can find the way to transfer from one material to the other.
Due to the relocation of charges, a voltage drop appears at the sides of the oxide layer, and two perturbed regions are found: one in the metal and the other in the semiconductor in close proximity to the oxide. In the metal, the electrons transferred to the semiconductor leave a surface layer of positive charges and in the p-type semiconductor the electrons transferred from the metal recombine the holes forming a space charge region.
In figure 3 the equilibrium bands diagram is shown.

The built-in potential appears as the contact potential difference between the surfaces of the metal and the semiconductor. Due to the presence of the oxide, the built-in potential across the

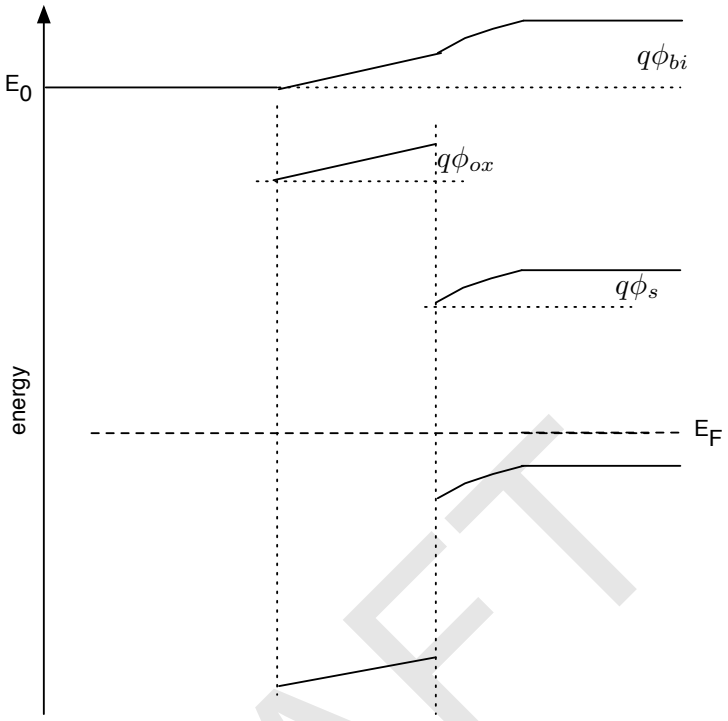

Fig. 9.3. *Equilibrium band diagram of the MOS structure.*

junction is split in two portions: one across the oxide and one across the depletion layer. Thus, the built-in potential is not the potential applied to the charges of the semiconductor.

$$q\phi_{bi} = q\Phi_m - q\Phi_s = q\phi_{ox} + q\phi_s \tag{9.1}$$

In order to determine the electric field and the potential at the equilibrium, it is necessary to make an assumption about the distribution of charges. As in all the previous cases, deep depletion and uniform doping hypothesis are assumed. Figure 4 shows the charge density distribution, where x_{ox} is the thickness of the oxide layer and x_d is the depth of the depletion layer. The charge neutrality condition holds, and the charge at the interface between metal and oxide is $Q = qN_A x_d$.

The electric field in the oxide ($\mathcal{E}_{ox}$) is constant and it is due to the surface charges on the metal.

$$\mathcal{E}_{ox} = \frac{Q}{\epsilon_{ox}} = \frac{qN_A x_d}{\epsilon_{ox}} \tag{9.2}$$

Across the interface between the oxide and the semiconductor the electric displacement vector is continuous: $D_{ox} = D_s \rightarrow \epsilon_{ox}\mathcal{E}_{ox} = \epsilon_s\mathcal{E}_s$, then the electric field at the surface of the semiconductor is:

$$\mathcal{E}_s(0) = \frac{\epsilon_{ox}}{\epsilon_s}\mathcal{E}_{ox} \tag{9.3}$$

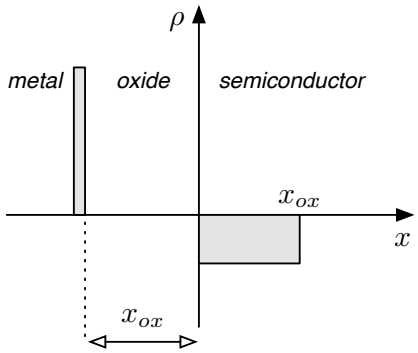

Fig. 9.4. *Charge density distribution at the equilibrium.*

in case of silicon-silicon dioxide, $\epsilon_{ox} = 3.9\ \epsilon_0$ and $\epsilon_s = 11.7\ \epsilon_0$ and the electric field in the oxide is about three times larger than the electric field at the surface of the silicon.
The electric field in the semiconductor is calculated from the Gauss law:

$$\int_{\mathcal{E}(0)}^{\mathcal{E}(x)} d\mathcal{E} = \int_0^x \frac{\rho}{\epsilon_s} dx = -\int_0^x \frac{qN_A}{\epsilon_s} dx \tag{9.4}$$

Replacing $\mathcal{E}(0)$ in the above relationship we obtain:

$$\mathcal{E}(x) = \frac{qN_A}{\epsilon_s}(x_d - x) \tag{9.5}$$

The electric field vanishes at x_d, and its behaviour is shown in figure 5.

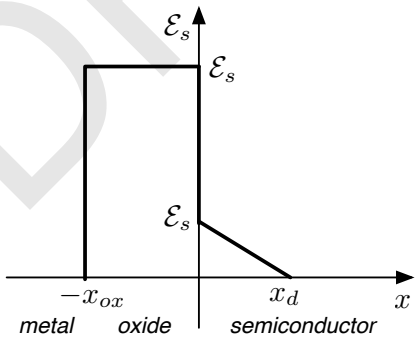

Fig. 9.5. *Behaviour of the electric field across the MOS. The proportion between $\mathcal{E}_s(0)$ and $\mathcal{E}_{ox}$ is pertinent to the silicon/silicon dioxide case.*

The potential is calculated integrating the electric field. As shown in the bands diagram the potential partially drops in the oxide and partially in the semiconductor across the depletion layer. The integral of the respective electric fields provides:

$$\phi_{ox}(x) = -\frac{qN_A x_d}{\epsilon_{ox}}(x + x_{ox}) \tag{9.6}$$

$$\phi_s(x) = \phi_{ox}(0) - \frac{qN_A}{\epsilon_s}\left(x_d x - \frac{x^2}{2}\right) \tag{9.7}$$

The potential is shown in fig. 6

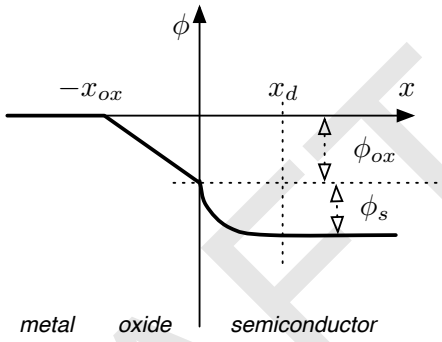

Fig. 9.6. *The potential in the MOS structure.*

The space charge region width is x_d, it can be calculated from $\phi_s = \phi_s(x_d)$

$$x_d = \sqrt{\frac{2\epsilon_s \|\phi_s\|}{qN_A}} \tag{9.8}$$

In the case of a MOS made of aluminum-silicon dioxide- p-type silicon, the built-in potential is $q\phi_{bi} = q\phi_{Si} - q\phi_{Al} = 0.8\ eV$.
The concentration of acceptors can be found from the work function, the affinity, and the energy gap:

$$N_A = p = n_i exp\left(\frac{E_i - E_F}{kT}\right) \tag{9.9}$$

where $E_i - E_F = q\phi_{Si} - (q\chi_{Si} + \frac{E_G}{2}) = 0.29\ eV$, then $N_A \approx 10^{15}\ cm^{-3}$
As an example let us calculate the dimensions of the depletion layer and the oxide layer when the built-in potential is equally distributed between oxide and semiconductor: $\phi_{ox} = \phi_{Si} = \frac{\phi_{bi}}{2} = 0.4\ eV$.
From eq. 9.8, $x_d = 865\ nm$ and from eq. 8.6 $x_{ox} = 114\ nm$. Thus, the oxide is less thick than the space charge region.

9.2.1 Relation between the potential and the charge carriers concentrations

The concentration of the charge carriers can be conveniently expressed as a function of the potential calculated as the difference between the Fermi level and the intrinsic Fermi level. Let us introduce the potential at the surface of the semiconductor: ϕ_s and the potential in the bulk ϕ_p . T In practice, as shown in fig. 7, the potential of fig. 6 is translated along the vertical axis so that the potential at the interface with the oxide is ϕ_s and potential in the bulk of the semiconductor is ϕ_p.
From these potentials, the concentration of electrons and holes at the surface and in the bulk are calculated:

$p_s = n_i exp(-\frac{q\phi_s}{kT})$
$p_b = N_A = n_i exp(-\frac{q\phi_p}{kT})$
$n_s = n_i exp(\frac{q\phi_s}{kT})$
$n_b = \frac{n_i^2}{N_A} = n_i exp(\frac{q\phi_p}{kT})$
Combining the two equations we obtain:

$$p_s = N_A exp\left(-\frac{q(\phi_s - \phi_p)}{kT}\right) \quad n_s = \frac{n_i^2}{N_A} exp\left(\frac{q(\phi_s - \phi_p)}{kT}\right) \tag{9.10}$$

The difference of potential between the interface and the oxide modulates the concentration of charges at the interface while in the bulk they remain constant.
Note that the above equations hold with any combination of materials, however, due to the absence of conductivity, the MOS allows for their whole exploitation. The potentials can be altered by the application of an external bias, then the concentration of charges at the interface with the oxide can be modulated by the application of suitable voltage drops between the metal and the semiconductor.

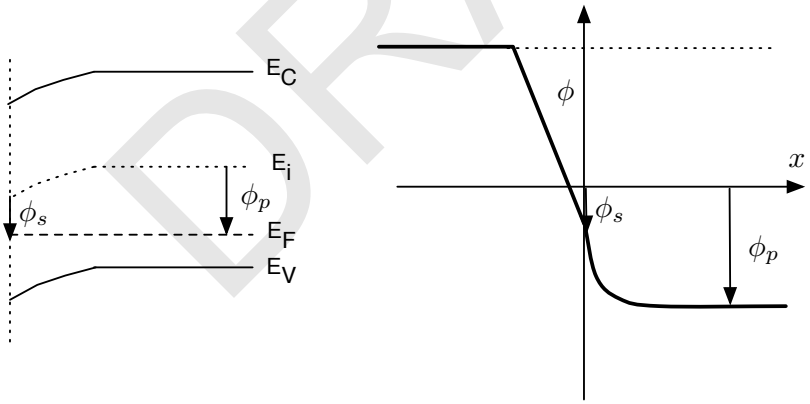

Fig. 9.7. *Surface and bulk potentials in the semiconductor of the MOS. Left: the potentials are defined as the difference between the Fermi level and the intrinsic Fermi level. Right: The potentials rescaled in order to evidence the potentials defined respect to the concentration of charge carriers.*

9.3 The MOS under bias

In order to study the modulation of the charges at the interface oxide-semiconductor let us consider a MOS capacitor biased by a variable d.c. voltage supply (V_A). For sake of simplicity, let us consider, as usual, the semiconductor grounded. The applied voltage is distributed between the oxide and the depletion layer.
At $V_A < 0$ the applied voltage reduces the bands bending in the semiconductor, then the concentration of holes at the interface increases respect to the equilibrium. This behaviour is similar to a forward bias Schottky but in this case due to the presence of the oxide no current can be observed.
At a particular value of the applied voltage the depletion layer is cancelled, and the concentration of holes at the interface is equal to the concentration in bulk. This is the *flat-band* condition, and the voltage at which it appears is called flat-band voltage (V_{FB}). In this situation the applied voltage cancels the effects of the junction, $V_{FB} = \phi_{bi}$.
Beyond the flat-band condition, at $V_A < V_{FB}$ the bands bend in the opposite direction and instead of being depleted, the interface region becomes to be overpopulated of majority charges. This condition is called accumulation. It corresponds to the space charge region in a ohmic Schottky contact (see section 3.4).
At $V_A > V_{FB}$ the equilibrium band bending is strengthened: the depletion layer becomes more wide, and the interface is more depleted of majority charges and more populated of minority charges. As the applied voltage grows, the concentration of electrons increases and the concentration of holes decreases. It is important to remind that since the current is zero, the semiconductor is always at the equilibrium and the mass-action law is always valid.
As V_A increases two important conditions are met: the intrinsic condition and the strong inversion. The first occurs when the Fermi level at the interface is equal to the intrinsic Fermi level ($E_{Fs} = E_{is}$), and then the concentration of holes and electrons are equal to the intrinsic semiconductor. values. Beyond the intrinsic condition the interface region is still a depletion layer but the material is now inverted, in the sense that although the semiconductor is doped with acceptors, the concentration of electrons is greater than the concentration of holes.
The discrepancy between electrons and holes increases until the difference between intrinsic and Fermi level at the surface is in absolute value equal to the condition in the bulk ($\|E_i - E_F\|_s = \|E_i - E_F\|_b$) in this situation: $n_S = p_b$.
The cases above described are reassumed in table 1:

The applied voltage modifies the density of the charges at the silicon-silicon oxide interface where the semiconductor may experience all the possible conditions. It is important to note that, since no current flows in the device, under the applied voltage the junction is still at thermal equilibrium and the applied voltage is simply added to the built-in potential. Thus, to bias is equivalent to change the work function of the metal, and in practice the metal becomes a sort of a virtual conductor whose work function is gradually changed.
The absence of current implies that the Fermi level in the semiconductor is constant and uniform, and the applied voltage modifies the band bending in the interface region. The potential in the semiconductor is still described as the difference between the Fermi level and the intrinsic Fermi level. The intrinsic Fermi level follows the band bending while the Fermi level remains constant.
Figure 8 visualises the behaviour of the potential under the whole interval of V_A. .

At strong inversion the total charge in semiconductor at the interface is the sum of the mobile electrons and the fixed acceptors ($Q = Q_n + Q_A$). The concentration of electrons is given by the eq. 9.10. Therefore, due to the exponential function, the concentration becomes extremely sensitive

Table 9.1. MOS cases at different applied voltages

V_A	sign of ϕ_s	relation between ϕ_s and ϕ_b	condition	holes concentration	electrons concentration
$< -V_{FB}$	-	$\|\phi_s\| > \|\phi_p\|$	accumulation	$p_s > N_A$	$n_s < \frac{N_i^2}{N_A}$
$-V_{FB}$	-	$\|\phi_s\| = \|\phi_p\|$	flat band	$p_s = N_A$	$n_s = \frac{N_i^2}{N_A}$
0	-	$\|\phi_s\| < \|\phi_p\|$	depletion at the equilibrium	$p_s < N_A$	$n_s > \frac{N_i^2}{N_A}$
> 0	-	$\|\phi_s\| < \|\phi_p\|$	depletion	$p_s < N_A$	$n_s > \frac{N_i^2}{N_A}$
$>> 0$	null	$\|\phi_s\| = 0$	intrinsic	$p_s < n_i$	$n_s > n_i$
$>>> 0$	+	$\|\phi_s\| < \|\phi_p\|$	weak inversion	$p_s < n_i$	$n_s > n_i$
$>>>> 0$	+	$\|\phi_s\| = \|\phi_p\|$	onset of strong inversion	$p_s < \frac{N_i^2}{N_A}$	$n_s = N_A$
$>>>>> 0$	+	$\|\phi_s\| > \|\phi_p\|$	strong inversion	$p_s < \frac{N_i^2}{N_A}$	$n_s \geq N_A$

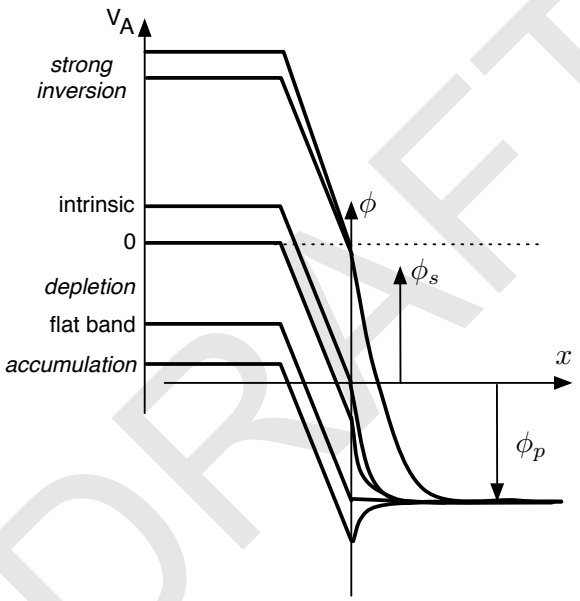

Fig. 9.8. *Potential profile in the MOS as a function of the applied voltage. Different MOS conditions are found along the applied voltage axis. Note that V_A is shifted in order to evidence ϕ_s and ϕ_p in the semiconductor.*

to the variations of ϕ_s. In practice, when the surface potential reaches the inversion, the potential remains almost fixed to ϕ_s. As an example, an increase of only 58 mV results in 10 times the concentration of electrons. The large sensitivity of the exponential function forces the argument to a fixed value, this is analog to the voltage drop across a forward biased diode.
Positive V_A corresponds to a forward bias of the depletion layer, that as a consequence, becomes always more deep. At strong inversio the potentials at the surface and in the bulk are equal: $\phi_s = -\phi_p$. Then, since the surface potential is fixed by the exponential and the bulk pot ential does not change because of the polarization, the depletion layer reaches the maximum extension.
$\phi_{Si} = -\phi_p - \phi_p = -2\phi_p = \frac{qN_A}{2\epsilon_s} x^2_{dmax}$ From which x^2_{dmax} is calculated

$$x_{dmax} = \sqrt{\frac{4\epsilon_s \|\phi_p\|}{qN_A}} \tag{9.11}$$

The charge associated to the maximum depletion layer is:

$$Q_{dmax} = -qN_A x_{dmax} = -\sqrt{4qN_A\epsilon_s\|\phi_p} \tag{9.12}$$

To draw the distribution of the total charge at the interface it is convenient to consider the band diagram at strong inversion (see fig. 9). The fixed charge is uniformly distributed in the depletion layer while the inversion charge is localised in the region where the Fermi Level lies above the intrinsic Fermi Level. The inversion region occupies a narrow region close to the interface while the depletion layer is spread for a wider distance. Finally, Figure 10 shows the behaviour of the concentrations of holes and electrons at the interface with the oxide.

It is important to remark that the above discussion is valid at the equilibrium when a d.c. voltage is applied. Any change in V_A elicits a transitory distribution of the charges that evolve towards the novel equilibrium conditions. The applied voltage can create accumulations of electrons or holes at the interface, and in principle, accumulation layers of electrons or holes should have the same behavior. However, in a p-type silicon, the electrons are always the minority charges and then their generation is unfavored process. This property is manifested when the concentration of electrons is required to increase as it happens when a a.c. signal is applied to the MOS. The manifestation of these phenomena is clearly visible in the C/V curve.

9.4 The C/V curve

The variable charges conditions at the surface of the semiconductor are visible in the total capacitance of the MOS. Specifically, the capacitance of the MOS is a function of the applied voltage and the C/V curve is an experimental method to visualise the processes that have been described in the previous section.
The C/V curve has been introduced in section 2.3.1 where it provided a method to experimentally determine both the doping concentration and the built-in potential in a metal-semiconductor junction.
In metal-semiconductor junctions the large conductivity under forward bias hides the capacitance and then the C/V curve is limited to the reverse bias condition. In the MOS the oxide layer avoid any conductivity and then the capacitance can be measured with any applied voltage.
The experimental set-up for the C/V curve has been shown in figure 2.8. The bias voltage determining the MOS condition and then the capacitance value is fixed by the d.c. voltage supply V_G.

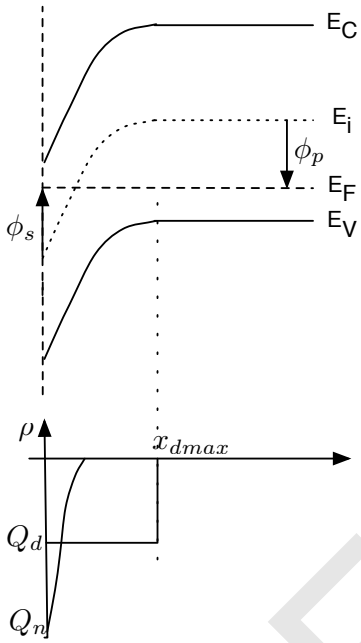

Fig. 9.9. *Band diagram in strong inversion and related distribution of mobile and fixed charge.*

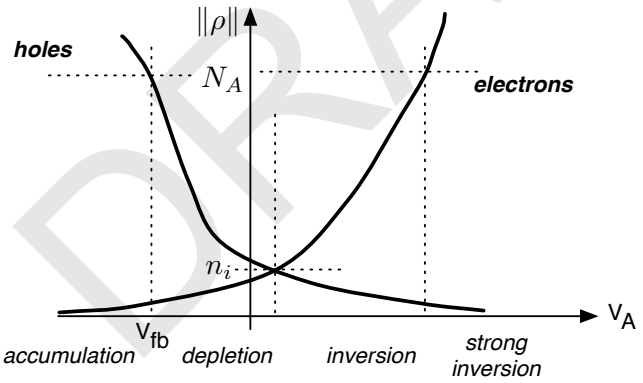

Fig. 9.10. *Modulation of absolute value of the density of holes and electrons as a function of the applied voltage.*

In order to measure the capacitance a small a.c. signal (v_t) is added in series. The amplitude of the a.c. signal is much smaller than the d.c. signal, so that it does not interfere with the MOS condition. We will see later, that some features of the C/V curve depend on the frequency of the a.c. signal.
The MOS is formed by a straightforward capacitance associated to the oxide layer, and an additional capacitance due to the semiconductor. The most obvious capacitive effect in the semiconductor is that related to the depletion layer. The depletion layer exists in an interval of V_G between the flat-band voltage and the inversion voltage. Beyond these values additional capacitive effect exists and they are going to be discussed in this section. V_G is applied across the metal and the bulk of the semiconductor, then it is distributed between the two capacitances which form a series of two capacitors whose total capacitance is:

$$\frac{1}{C_{tot}} = \frac{1}{C_{ox}} + \frac{1}{C_S} \tag{9.13}$$

Where C_S is the capacitance in the semiconductor and C_{ox} is the capacitance of the oxide layer. As reported the Table 1, the interface condition changes as a function of V_G, while the capacitance of the oxide remains constant:

$$C_{ox} = \frac{\epsilon_{ox}}{x_{ox}} \tag{9.14}$$

where x_{ox} is the thickness of the oxide.
It is convenient to begin the description of the C/V curve from the flat band condition when $V_G = -Vfb$. The flat band voltage is the difference of the work functions of the metal and the semiconductor: $V_{fb} = \phi_m - \phi_s$.
In flat band, the band bending induced by the work function difference is canceled by the applied voltage, then the concentration of electrons and holes is constant everywhere in the semiconductor and equal to the equilibrium condition: $p = N_A$ and $n = n_i^2/N_A$.
Although the depletion layer does not exist in this condition the semiconductor shows an additional capacitive effect that is, in some sense, intrinsic of the semiconducting material and it does not depend on the junction. In case of a capacitor made by metal plates, the capacitance is only due to the insulator because the charge modulation in the metal is confined to the surface in contact with the dielectric. Indeed, the electric field in the metal is null and then the voltage drops at the surface. In a semiconductor the situation is different, the voltage perturbs a non negligible volume of the semiconductor. Then the modulation of charge due to the applied voltage takes place at a distance from the surface with the dielectric. This is equivalent to an additional capacitor in series to the oxide capacitor.
To calculate the capacitance is necessary to evaluate the amount of charges created by the applied voltage. For this scope let us consider the Poisson equation and the perturbation that the applied voltage induces to the flat-band condition. The semiconductor is P-type doped, then the total charge is limited to the mobile holes and the fixed acceptors and the Poisson equation is

$$\frac{d^2\phi}{dx^2} = -\frac{\rho(x)}{\epsilon_s} = -\frac{q}{\epsilon_s}(p - N_A) \tag{9.15}$$

It is worth to remind that the majority charges are mobile in space and also in time. The *time mobility* is ensured by the fact that only a portion of the acceptor atoms are ionised, then out of equilibrium the concentration of holes can still change.
In order to calculate the capacitance, let us consider ϕ the portion of the modulating voltage(v_t)

that is applied to the semiconductor from the interface with the oxide to the bulk.Then ϕ is the potential with respect to the potential in the bulk (ϕ_p). This potential elicits a perturbation of the concentration of the holes:
$p = N_A exp(-\frac{q\phi}{kT})$.
Replacing the above equation in the Poisson equation we get:

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon_s} N_A \left[exp\left(-\frac{q\phi}{kT}\right) - 1 \right] \tag{9.16}$$

Considering the v_t is a small perturbation and ϕ is just a part of it, the exponential can be replaced by the first order Taylor series.

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon_s} N_A \left(1 - \frac{q\phi}{kT} - 1\right) = \frac{q^2 N_A}{\epsilon_s}\phi \tag{9.17}$$

That can be written as:

$$\frac{d^2\phi}{dx^2} = -\frac{\phi}{L_D^2} \tag{9.18}$$

Where: $L_D = \sqrt{\frac{\epsilon_s KT}{q^2 N_A}}$ is the Debye length of the semiconductor.
The generic solution of eq. 18 is a sum of positive and negative exponential. The boundary conditions are: $\phi(0) = \phi_{ts}$; and $\phi(\infty) = 0$ and the solution is limited to the negative exponential, namely the potential decays towards the bulk of the semiconductor:

$$\phi(x) = \phi_{ts} exp\left(-\frac{x}{L_D}\right) \tag{9.19}$$

The applied a.c. signal creates in the semiconductor a charge modulation that expires at about four times the Debye length from the surface. The modulated charge is calculated directly from the Poisson equation:

$$\rho(x) = -\epsilon_s \frac{d^2\phi}{dx^2} = -\epsilon_s \frac{\phi_{ts}}{L_D^2} exp\left(-\frac{x}{L_D}\right) \tag{9.20}$$

Then the total charge related to the a.c.signal is:

$$Q = \int_0^\infty -\epsilon_s \frac{d^2\phi}{dx^2} dx = -\frac{\epsilon_s}{L_D}\phi_{ts} \tag{9.21}$$

Finally, the capacitance of the semiconductor in the flat band condition can be calculated as:

$$C_s^{fb} = \| \frac{dQ}{d\phi_t s} \| = \frac{\epsilon_s}{L_D} \tag{9.22}$$

The capacitance of a semiconductor in flat-band condition is equivalent to the capacitance of a metal-insultator-metal where the dielectric constant of the insulator is ϵ_s and the conductors are separated by L_D. Note that L_D depends on the concentration of the charge carriers, then it depends on doping. More doped is the material shorter is the Debye length and larger is the capacitance.
The total capacitance of the MOS at $V_G = V_{fb}$ is the series of the two capacitors:

$$\frac{1}{C_{tot}^{fb}} = \frac{1}{C_{ox}} + \frac{1}{C_s^{fb}} = \frac{x_{ox}}{\epsilon_{ox}} + \frac{L_D}{\epsilon_s} \tag{9.23}$$

As an example, let us consider a p-type silicon with $N_A = 10^{15}\ cm^{-3}$ and a layer of silicon dioxide of thickness $x_{ox} = 150\ nm$. Then $L_D = 130\ nm$, $C_{ox} = 2.30 \cdot 10^{-8}\ \frac{F}{cm^2}$ and $C_s^{fb} = 7.98 \cdot 10^{-8}\ \frac{F}{cm^2}$. The total capacitance is: $C_{tot}^{fb} = 1.78 \cdot 10^{-8}\ \frac{F}{cm^2}$. In this case $V_{fb} \approx -0.8V$.
When $V_A < V_{fb}$ the interface region enters in the accumulation mode; namely, the concentration of holes at the surface exceeds the concentration of holes in the bulk. In order to describe the behaviour of C_s we can consider that the Debye length is inversely proportional to the square of the carriers concentration. Then with respect to the flat-band condition since the holes concentration increases the Debye length decreases and then the capacitance associated to the semiconductor becomes larger.
The concentration of holes is: $p = N_a exp(-\frac{q(\phi_s - \phi_p)}{kT})$. At flat-band $\phi_s - \phi_p = 0$. Following the above mentioned example we observe that at $\phi_s - \phi_p = 0.1\ V$ the semiconductor capacitance is: $C_s = 54.9 \cdot 10^{-8}\ \frac{F}{cm^2}$ and then the total capacitance is $C_{tot} = 2.21 \cdot 10^{-8}\ \frac{F}{cm^2}$. If the applied voltage doubles to $0.2\ V$ the value of the capacitances become $C_s = 373 \cdot 10^{-8}\ \frac{F}{cm^2}$ and $C_{tot} = 2.28 \cdot 10^{-8}\ \frac{F}{cm^2}$. Eventually, the total capacitance converges to the oxide capacitance.
On the other hand, for $V_G > V_{fb}$ the interface becomes depleted of holes. The capacitance associated to the depletion layer has been calculated in chapter 2, then the capacitance of the semiconductor is: $C_s = \frac{\epsilon_s}{x_d}$ where x_d is the depth of the depletion layer that depends on the voltage drop between the surface and the bulk of the semiconductor:

$$x_d = \sqrt{\frac{2\epsilon_s \|\phi_s - \phi_p\|}{qN_A}} \tag{9.24}$$

As V_G grows into positive values, the absolute difference between $\phi_s - \phi_p$ increases and also x_d increases. Thus, the capacitance of the semiconductor becomes small and so the total capacitance. Always following the above example, at $\|\phi_s - \phi_p\| = 0.1\ V$ we have $x_d = 35\ nm$, $C_s = 2.87 \cdot 10^{-8}\ \frac{F}{cm^2}$, and $C_{tot} = 1.27 \cdot 10^{-8}\ \frac{F}{cm^2}$.
The depletion layer expands with the applied voltage until the threshold of strong inversion is met. In this condition we have seen in previous section that in practice the depletion layer reaches its maximum extension:

$$x_{dmax} = \sqrt{\frac{4\epsilon_s \|\phi_p\|}{qN_A}} \tag{9.25}$$

and the total capacitance reaches the smallest value.

$$\frac{1}{C_{tot}^{fb}} = \frac{1}{C_{ox}} + \frac{1}{C_s^{fb}} = \frac{x_{ox}}{\epsilon_{ox}} + \frac{x_{dmax}}{\epsilon_s} \tag{9.26}$$

The largest extension of the depletion layer in the above example is $x_{dmax} = 622\ nm$, then $C_{skin} = 1.66 \cdot 10^{-8}\ \frac{F}{cm^2}$, and $C_{tot,min} = 0.96 \cdot 10^{-8}\ \frac{F}{cm^2}$.
Until the strong inversion threshold, the interface is populated by the fixed charge defined by x_d and by the electrons. The population of electrons is rapidly growing but until the inversion the fixed charge dominates and then the capacitance is dominated by the depletion layer capacitance.
At the inversion and beyond, the layer of electrons close to the interface cannot be neglected. However, in order to affect the capacitance it is necessary that the charge is modulated by the a.c.

voltage.
The inversion layer is formed by minority charges that, due to the applied bias, become, in a narrow interface region, more numerous than holes . However, the minority charge character of the electrons is still valid and, in particular, the generation of electrons in the depletion layer requires a non negligible time. The typical time for the minority charges generation is of the order to tenths of seconds. An explicit calculation of the time necessary to create and modulate the inversion layer of electrons is explicitly calculated in the next section,
The limited speed of variation of the charges in the inversion layer makes the capacitance strongly dependent on the frequency of the probing a.c. signal (v_t). At frequencies of few Hertzs the variation of the signal occurs in a time interval that is compatible with the generation time of electrons. Then, at low frequency the capacitance in the inversion is again equal to C_{ox} indicating that at the inversion an accumulation layer, made of electrons instead of holes, exists.
This behaviour looks similar to the accumulation condition met at $V_G < V_{fb}$. However in inversion the situation is different because the depletion layer still exists and it may respond to the variable v_t. However, due to the large concentration of electrons in the inversion layer, the Debye length in inversion layer is much shorter than x_d and then the applied potential vanishes in the inversion layer and it does not perturbate the depletion layer charge. Eventually, the total capacitance is the oxide capacitance.
On the other hand, if the frequency of v_t is larger than few Hertzs, the applied signal changes too fast respect to the rate of generation of electrons. Then, in spite of the large concentration of electrons, the charges of the inversion layer are insensitive to the applied signal and the capacitance is only contributed by the depletion layer. In this condition, the presence of inversion layer is only reflected by the fact that the capacitance reaches the minimum value and is insensitive to any increase of V_G.
The total C/V curve is illustrated in Figure 11.

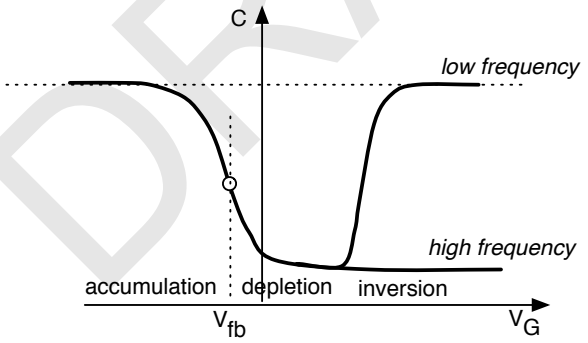

Fig. 9.11. *The MOS C/V curve.*

In order to better understand the different observed values of the capacitance, it might be useful to visualise in figure 12 the distribution of charges whose concentration is modulated by the applied a.c. voltage in the different cases.

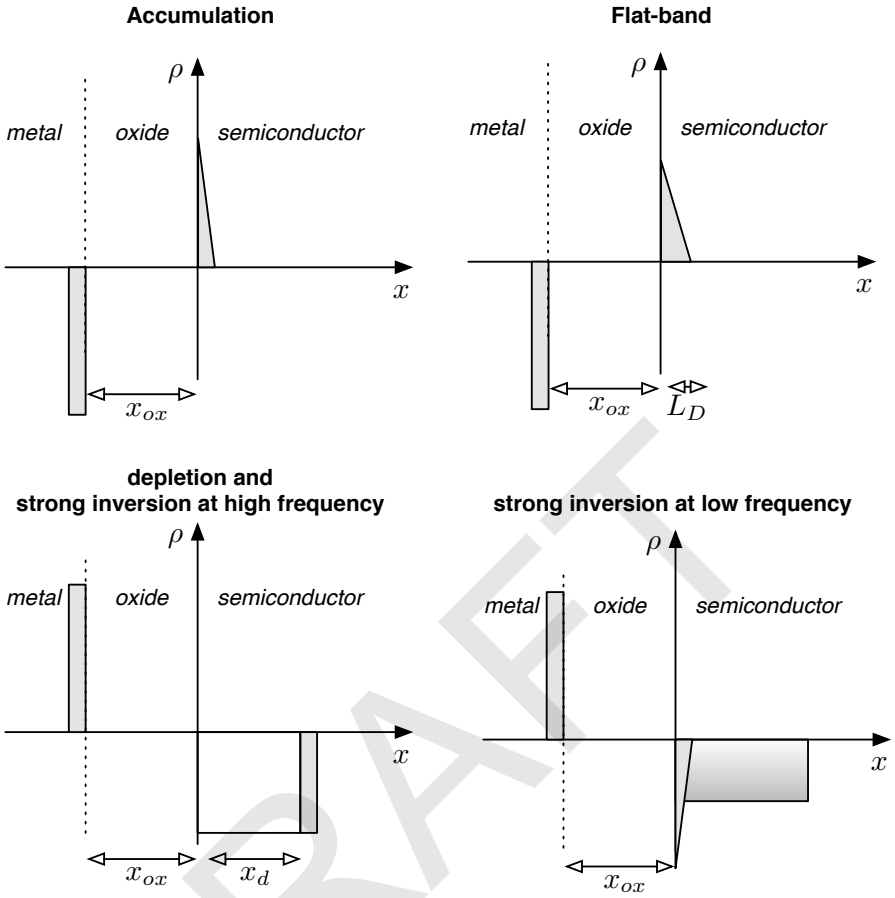

Fig. 9.12. *The different cases of the C/V curves corresponds to the different locations of the MOS where the charge modulation occurs.*

The previous discussion about the C/V curve have been carried assuming that the d.c. signal is stepwise varied and the capacitance is measured keeping V_G constant. Then, enough time is left to form the inversion layer.
An alternative mode consists in applying V_G as a ramp. In this way, there is not time to generate the equilibrium concentration of minority charges and the inversion layer is not formed. The absence of the inversion layer does not limit the expansion of the depletion layer which may continue to grow even beyond x_{dmax}, and the total capacitance continue to decrease. This condition is called *deep depletion*.
Figure 14 shows the two modality of voltage application. The complete C/V curve is shown in figure 15.

Finally, the C/V curve is a invaluable tool to investigate the defects of the oxide layer and the interface traps. During the oxidation process the oxide layer might incorporate impurity either from

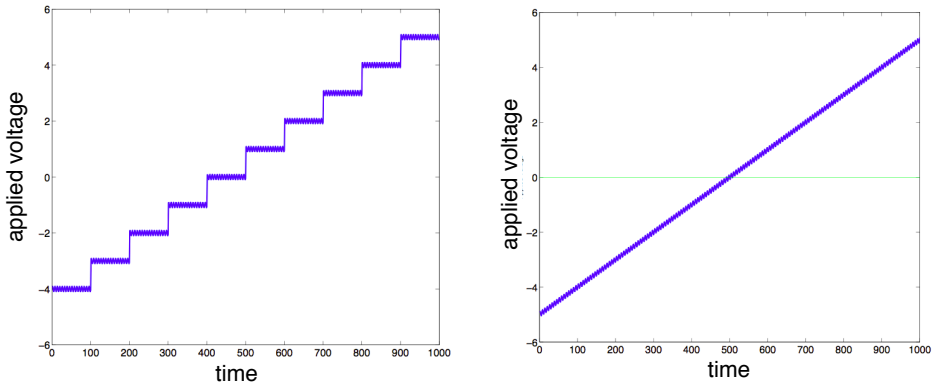

Fig. 9.13. *The two different modes of application of the applied voltage in a MOS C/V curve experiment. Left) the stepwise voltage allows for the formation of the inversion layer and the depletion capacitance reaches a maximum value. Right) the ramp if sufficiently fast does not allows for the formation of an inversion layer, then the depletion capacitance continues to decreases and so the total capacitance of the MOS.*

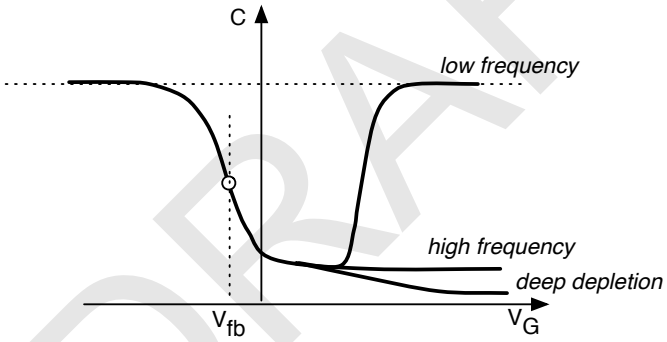

Fig. 9.14. *The complete MOS C/V curve complemented with the deep depletion case.*

the silicon or from the atmosphere. The impurity can be charged. So the oxide can be populated by a concentration of charges that can be both fixed or mobile since the ions can slow migrate, with a very low mobility, through the oxide.
The charges inside the oxide create a further potential added to the built-in potential. The presencer of charges in the oxide are macroscopically manifested by the shift of the flat-band voltage towards either positive or negative values accordingly to the sign of the total charge.
Interface states are due to the rearrangement of the silicon atoms at the interface with the oxide. As we have seen in chapter 2, these states may sequestrate electrons and holes subtracting charges to the inversion and accumulation layers respectively. The presence of the surface states makes the

C/V less sharp than expected around the flat-band voltage whose value remains unchanged. Figure 16 shows both the cases.

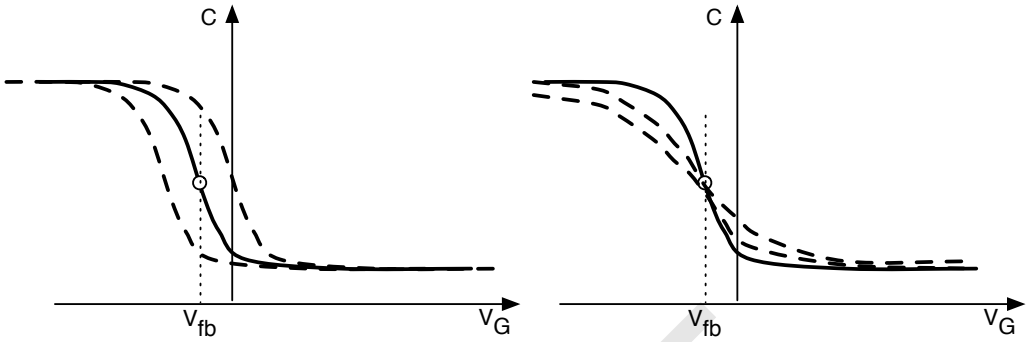

Fig. 9.15. *C/V curves in case of charges in the oxide (left) and interface states (right) with respect to the theoretical ideal curve where V_{fb} is the difference of the work functions of the metal and the semiconductor.*

9.4.1 Inversion layer charges generation

To evaluate the time necessary to generate the charges in the inversion layer, let us consider the case of a MOS made with a P-type semiconductor. The MOS is initially in depletion condition and let us suppose that at the time t=0 the MOS is suddenly biased at the Inversion voltage. The inversion layer is formed by electrons, generated in the depletion layer, that reach the interface region. This situation corresponds to the generation current in a reverse bias PN junction with the important difference that the current flows as the depletion layer increases from x_d (at $V_G = 0$) until to $x_d = x_{d0}$ corresponding to the new equilibrium established by the applied voltage. During the transition the depletion layer as well the concentration of electrons are a function of time . The generation current (eq. 5.64), does not explicitly depends on the applied voltage but only on the difference between x_d and x_{d0}:

$$J_G = \frac{dQ_n}{dt} = -\frac{qn_i(x_d - x_{d0})}{2\tau_0} \tag{9.27}$$

The relationship between the charges on the gate and the inversion layer charge generation rate can be written considering that the total charge in the semiconductor is $Q_G = -(Q_n - qN_A x_d)$. Note that Q_g is also the total charge on the metal. Replacing x_d from this equation we obtain:

$$\frac{dQ_n}{dt} = -\frac{-qn_i}{2\tau_0}\left(\frac{Q_g + Q_n}{qN_A} - x_{d0}\right) \tag{9.28}$$

The above equation can be rearranged as:

$$Q_n + \frac{2N_A\tau_0}{n_i}\frac{dQ_n}{dt} = -(Q_G - qN_A x_{d0}) \tag{9.29}$$

Introducing $\tau_n = \frac{2N_a}{n_i}\tau_0$ the equation becomes:

$$\frac{dQ_n}{dt} + \frac{Q_n}{\tau_n} = -\frac{Q_G - qN_A x_{d0}}{\tau_n} \tag{9.30}$$

The solution of the above equation with the boundary condition: $Q_n(t=0) = 0$ is:

$$Q_n = -(Q_G - qN_A x_{d0})\left[1 - exp\left(-\frac{t}{\tau_n}\right)\right] \tag{9.31}$$

Thus, the equilibrium charge in the inversion layer is reached through an exponential behavior whose time constant is τ_n. Note that the generation time does not depend only on the time scale of the generation/recombination processes but it is proportional to the density of doping then the generation rate of minority charges decreases with the doping. In case $\tau_0 = 1\ \mu s$ 0.1and $N_A = 10^{16}\ cm^{-3}$; $\tau_n = 0.13\ s$. This is the time scale necessary to form, and even to modulate, the charge in the inversion region.
The charge in the inversion layer follows the modulation of the applied a.c. voltage only if the frequency of the signal does not exceed $1/\tau_n$. In the above example where $\tau_n = 0.13\ s$ this frequency corresponds to about 8 Hz.
The time of formation/modulation of the inversion layer is strongly decreased in those cases where the minority charges (electrons in this case) are provided by a different sources. Noteworth cases are the nearby N-type semiconductors, such as the drain and source contacts of a MOSFET (see chapter 10), and the photo generated electrons as it happens in optical photodetectors.

10

Field Effect Transistors

10.1 Introduction

METal-oxide-semiconductor junctions are the core of a family of three terminals device where the conductivity of the inversion layer is modulated by the voltage applied to the MOS structure. Such devices are called Field Effect Transistors (FETs).
The basic concept of FETs is rather straightforward. It is based on the evidence that the mobile charges can be either accumulated or depleted, on the side of the conductor, by the application of a voltage in the orthogonal direction respect to the current flow.
The idea of the FET, as an alternative to triodes, arose in the twenties. The first patent was filed in 1925, while a following patent in 1934 provided a more detailed description of the device. The concept attracted W. Shockley that in 1948 developed a prototype that did not properly work because of the too high density of defects at the oxide-semiconductor interface. This failure prompted Shockley to direct his attention towards the PN junction, and the first working FET was fabricated only in 1957, at the Bell Lab., using silicon. This achievement was made possible by the maturity of the semiconductors technology that allowed the fabrication of a sufficiently clean oxide-semiconductor interface. The silicon FET takes the name of Metal Oxide Semiconductor Field Effect Transistor (MOSFET). First commercial devices were produced at the Bell Lab after 1960.
In the first part of this chapter the basic working mechanisms of silicon MOSFETs are illustrated. The second part of the chapter is dedicated to the further development of field effect devices mostly based on materials characterised by large mobilities such as GaAs and other III-V materials..

10.2 Channel charge modulation

Before to discuss the properties of the MOSFETs it is necessary to evaluate the relationship between the charge in the inversion layer, the channel of MOSFETs, and the voltage applied to the MOS (V_G).
The discussion considers a MOS made by a p-type semiconductor biased in strong inversion.
In this condition the depletion layer reaches its maximum extension

$$x_{dmax} = \sqrt{\frac{2\epsilon_s(2\|\phi_p\|)}{qN_A}} \tag{10.1}$$

The charge stored in the depletion layer is:

$$Q_d = -\sqrt{2\epsilon_s q N_A (2\|\phi_p\|)} \tag{10.2}$$

In order to write the relationship between (V_G) and the total charge in the semiconductor let us consider the potential profile between the gate and the bulk of the semiconductor.

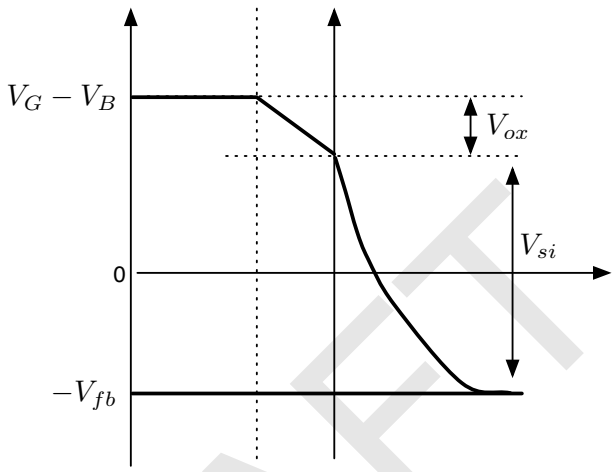

Fig. 10.1. *Distribution of the potential in the MOS structure in the strong inversion.*

The flat band voltage ($V_G = V_{fb}$) corresponds to the situation where the potential drop in the semiconductor is null. V_{fb} is negative and it corresponds to the difference between the work function of the metal gate and the semiconductor.
To calculate the charge in the semiconductor let us start to consider the electric field in the oxide:

$$\mathcal{E}_{ox} = \frac{V_{ox}}{x_{ox}} = \frac{V_G - V_B - V_{fb} - V_{si}}{x_{ox}} \tag{10.3}$$

At the interface oxide-semiconductor, the continuity of the electric displacement requires that: $\epsilon_{ox}\mathcal{E}_{ox} = \epsilon_{si}\mathcal{E}_{si0}$. Where $\mathcal{E}_{si0}$ is the electric field at the interface on the silicon side. Replacing $\mathcal{E}_{ox}$ with the equation 10.4, the electric field at the surface of the silicon is:

$$\epsilon_{si}\mathcal{E}_{si0} = \frac{\epsilon_{ox}}{x_{ox}}(V_G - V_B - V_{fb} - V_{si}) = C_{ox}(V_G - V_B - V_{fb} - V_{si}) \tag{10.4}$$

The total electric field in the semiconductor depends, thanks to the Gauss theorem, on the charges in the semiconductor:

$$\mathcal{E}_{sib} - \mathcal{E}_{si0} = \frac{Q_n + Q_d}{\epsilon_s} \tag{10.5}$$

where $\mathcal{E}_{sib}$ is the electric field in the bulk, Q_n are the mobile electrons and Q_d are the fixed acceptors. Since the electric field in the bulk is null, the field at the interface is $-\mathcal{E}_{si0} = \frac{Q_n+Q_d}{\epsilon_s}$ from which the mobile charge is calculated:

$$Q_n = -\epsilon_s \mathcal{E}_{si0} - Q_d \tag{10.6}$$

Replacing the electric field at the interface with eq. 10.4 and Q_d with eq. 10.2 (note that Q_d is negative) we get:

$$Q_n = -C_{ox}[V_G - V_B - V_{fb} - 2\|\phi_p\|] + \sqrt{2\epsilon_s q N_A (2\|\phi_p\|)} \tag{10.7}$$

The mobile charge at the interface is made of electrons, from eq. 10.7 this charge exists when Q_n is negative, and this happens only if the first term of eq. 10.7 is greater than the second term. The condition $Q_n = 0$ is particularly important because it defines the onset of the channel formation. The voltage V_G at which this condition is met is the *threshold voltage* (V_T). Setting $Q_n = 0$ in equation 10.8 leads to:

$$V_T = V_{fb} + 2\|\phi_p\| + \frac{1}{C_{ox}}\sqrt{2\epsilon_s q N_A (2\|\phi_p\|)} \tag{10.8}$$

Considering that $(V_G - V_B) - V_{fb} = V_{ox} + V_{si}$. At strong inversion $V_{si} = 2\|\phi_p\|$. Then $V_G = V_T$ is equivalent to:

$$V_T = V_{fb} + 2\|\phi_p\| + V_{ox} \tag{10.9}$$

Comparing eq. 10.8 and 10.9 we find that the voltage applied to the oxide is responsible for the charge of the depletion layer ($V_{ox} = Q_d/C_{ox}$).
Given the definition of V_T, a concise expression for the charge in the channel is obtained:

$$Q_n = -C_{ox}(V_G - V_T) \tag{10.10}$$

This fundamental relationship states that the charge, and then the conductivity, of the channel is controlled by the voltage applied to the MOS. It is worth to note that all the features of the MOS participate to the threshold voltage which depends on the the difference of work functions of metal and semiconductor (V_{fb}), the doping concentration (ϕ_p), and the thickness of the oxide (C_{ox}).

10.3 Metal Oxide Semiconductor Field Effect Transistor

The equation 10.10 describes the control, via the voltage applied to the gate, of the concentration of the electrons in the inversion layer. These charges form a thin layer of mobile electrons surrounded on one side by the oxide and on the other side by a space charge region. The layer of charges are then confined between two insulating regions. The charges can be kept in movement if two suitable electrodes are provided at the edges of the layer. These electrodes are two n^+ regions implanted at the edges of the inversion layer. Eventually, the device is a three electrodes structure where the inversion layer is turned into a conductive channel whose resistivity is controlled by the voltage applied to the MOS capacitor.
The electrodes at the edges of the channel are called drain and source, while the electrode applied

to the metal of the MOS is called gate. The metal electrode is separated from the p-type semiconductor by a thick oxide layer. In this way, the associated capacitance is so small to make negligible any field effect. The configuration contains a fourth electrode applied to the substrate and called body. The n^+ electrodes introduce two PN junctions between the electrodes and the bulk. In normal operation conditions $V_D - V_B > 0$ and $V_S - V_B > 0$, then both the junctions are reversely biased and only a negligible current flows towards the body contact.
Such a device is the Metal Oxide Semiconductor Field Effect Transistor (MOSFET).
It is interesting to observe that a conductive channel can be obtained accumulating either the majority or the minority charges. In both cases the conductivity between source and drain may be regulated by the gate voltage. However, the case of inversion is much more interesting. Mainly because the current in the inversion channel flows in a narrow strip confined between two insulators. As a consequence the conductivity between the contacts is only due to the channel.
In case of accumulation the source and drain contacts have to be ohmic contacts for the majority charges but also the bulk of the semiconductor (where $p = N_A$) participates to the conduction between the source and the drain. Since accumulation layer and bulk are in parallel if the semiconductor is much thicker than the accumulation layer the conductivity is always dominated by the bulk and the field effect is not observable. Such a configuration is becoming popular in the molecular electronic field where thin layers of molecular semiconductors are used.
Another important advantage of the use of the device in inversion condition is the automatic insulation from the rest of the substrate that is provided by the depletion layer. This is an important issue for the integration of more devices on the same substrate.
The MOSFET is used as a trans-amplifier, namely a voltage signal applied to the gate electrode is turned into a current signal at the drain-source contacts.
In the previous chapter we have seen that, in a MOS, the modulation of the charges in the inversion layer is a very slow process because the generation of minority charges occurs with a very low rate. However, in a MOSFET the charges in the channels can be varied at high speed. This difference is due to the n^+ wells that can act as electrons reservoirs that provide, with great efficiency, all the electrons necessary to modulate the charge in the channel.
Figure 2 illustrate the basic MOSFET configuration.

Most of the characteristics of the device comes from the difference between the work function of the metal and the semiconductor (V_{fb}), the doping of the semiconductor (ϕ_p) and the oxide thickness (C_{ox}). The combination of these three quantities defines two categories of MOSFETs called enhancement and depletion. In an enhancement MOSFET at $V_G = 0$ the channel is not formed . In a p-type semiconductor the channel is formed and the modulated at $V_G > 0$. The sign of V_G is inverted when a n-type semiconductor is considered. In a depletion MOSFET, the channel exists even at $V_G = 0$, then the gate voltage can increase, decrease or even shut-off the inversion layer.
In the following of this section the enhancement MOSFET is illustrated.
V_G sets the working condition: when $V_G < V_{fb}$ the interface is in accumulation, the path from source to drain forms a series of two opposite PN junctions (a configuration called back-to-back). In this condition the channel is analog to a BJT in cut-off, and the current from the source to the drain is the reverse current of the two PN junctions.
Increasing V_G we find the condition for which $V_{fb} < V_G < V_T$. In this interval, the channel region is depleted of mobile charges. As V_G increases a weak population of electrons is formed. This is not the inversion layer but a small current called sub-threshold current flows between the source and the drain .
Finally, when $V_G > V_T$ the strong inversion condition is met and the channel of electrons is com-

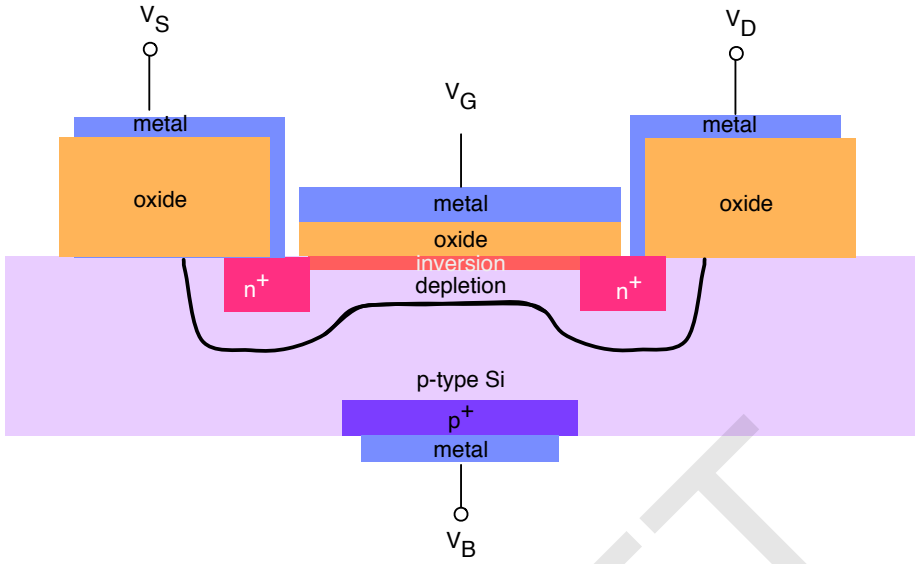

Fig. 10.2. *Basic configuration of a MOSFET.*

pletely formed. In strong inversion the depletion layer remains constant and any increase of V_G results in an increase of the electrons concentration in the channel.
Let us derive the relationship between the current flowing in the channel and the couple of voltages V_{DS} (applied to the channel) and V_G (applied to the gate). In principle, V_G sets the concentration of electrons and then the conductivity of the channel and V_{DS} drives the drift current. Actually the situation is a little more complicated because V_{DS} is distributed along the channel, and then, at each point of the channel an additional voltage $V_C(y)$ is found. This voltage is additive with respect to V_G and then it affects the charge density in the channel:

$$Q_n(y) = -C_{ox}(V_G - V_{fb} - 2\|\phi_p\| + V_C(y)) + \sqrt{2\epsilon_s q N_A(2\|\phi_p\| + V_C(y))} \tag{10.11}$$

Above the threshold voltage, the inversion channel is formed and the junction between the n^+ regions and the channel is negligible. Thus, the current is equivalent to the current flowing in a homogeneously doped semiconductor, and it is made by a drift and a diffusion component. The situation is analog to the transport in the base of a BJT in active zone where to decrease the transit time an electric field is created in the base either because of a graded doping or a graded band gap. In both cases, a limited voltage greater than twice the thermal voltage (56 mV) is sufficient to make the drift contribution greater that the diffusion current. It is straightforwards to suppose that since V_{DS} is greater than the thermal voltage, the drift current dominates over the diffusion current component.

In order to calculate the effect of V_{DS} on the charges in the channel let us assume that the potential along the coordinate channel (coordinate y) changes less rapidly than along the direction across the channel (coordinate x).

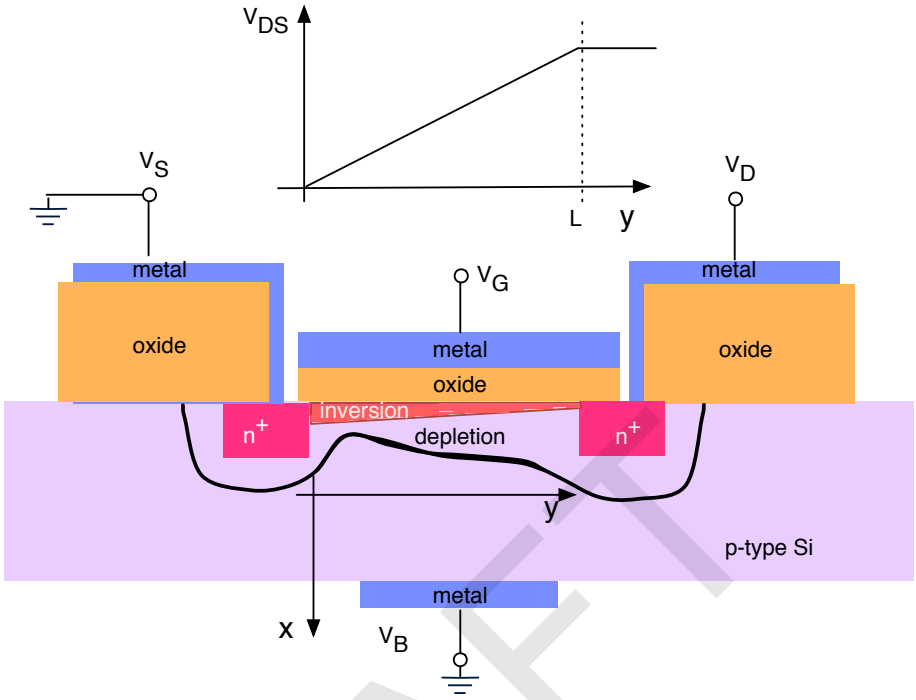

Fig. 10.3. *Behaviour of V_{DS} along the channel. A system of x y coordinates is settled to describe the behaviour of the charges along the channel and towards the bulk. Thanks to the voltage added to the channel the inversion layer under the oxide is not uniform, but it becomes more thin approaching the drain contact. For this reason, even if the mobile charge decreases the concentration of the charge is practically constant and the diffusion current can be neglected. Eventually, the current of the MOSFET, above threshold, can be described in terms of the only drift current.*

$$\|\frac{\partial \phi}{\partial y}\| \ll \|\frac{\partial \phi}{\partial x}\| \tag{10.12}$$

This condition is the *graded channel approximation*. Under this approximation, for each Δy the corresponding Δx is constant. In practice, the channel can be sliced in portions where the charge is constant.
Noteworthy, this condition is met when the length of the channel (L) is much greater than the depletion layer. The main consequence of the graded channel approximation is that the relationship $Q_n = -C_{ox}(V_G - V_T)$ is valid in each interval dy.
To calculate the distribution of V_{DS} it is necessary to calculate the profile of resistance of the channel. The differential resistance (dR) of a small slice of the channel (from y to dy) is : $dR(y) = \rho(y)\frac{dy}{A}$. The resistivity ρ depends on the concentrations of electrons at the position y

$$\rho(y) = \frac{1}{\sigma(y)} = \frac{1}{\mu_n Q(y)} \tag{10.13}$$

μ_n is the electrons mobility in the channel. The channel lies at the interface between the semiconductor and the oxide in a region overcrowded of defects with respect to the bulk. We can expect that the mobility at the interface is smaller than the mobility in the bulk. The actual value of the mobility depends on the conditions at which the device is grown and it cannot be theoretically calculated, an approximate estimation is $\mu_n \approx \frac{1}{2}\mu_{bulk}$ but μ_n is an unknown parameter of the device. The concentration of the charges that appears in eq. 9.12 is units of $C \cdot cm^{-3}$ while the charges concentration calculate by eq. 10.8 is in units of $C \cdot cm^{-2}$. It is easy to see that given a slice of the channel, its volume is Ady and the area below the oxide is wdy where w is the lateral dimension of the channel and and A is area of the slice. Eventually we have: $Q(y)A = Q_n(y)w$. The relationship between the volumetric and surface concentrations is shown in figure 4.

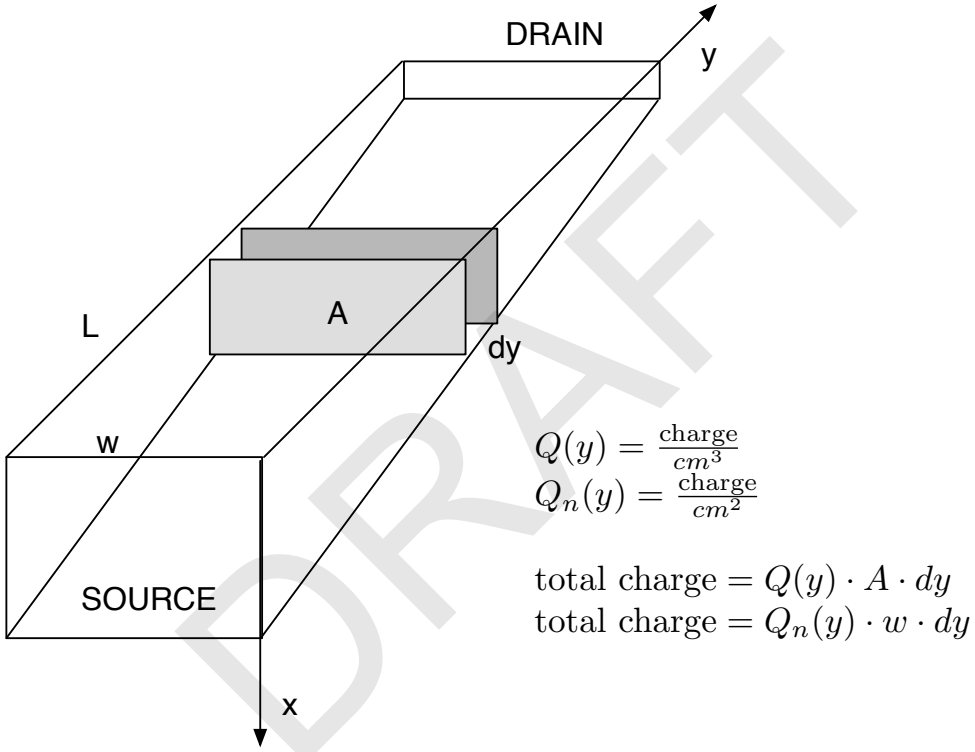

Fig. 10.4. *Behaviour of the channel between the source and drain contact. The relationship between the volume and surface charge concentrations is illustrated in a slice of the channel.*

the differential resistance of a slice dy of the channel at the position y is:

$$dR = \frac{1}{\mu_n Q(y)} \frac{dy}{A} = \frac{dy}{\mu_n Q_n(y) w} \tag{10.14}$$

The current in the channel (I_D) is a drift current that is equal to:

$$I_d = \frac{dV_C}{dR} \tag{10.15}$$

Obviously I_D along the channel is constant and independent on y.
The concentration of charges in the channel is given by equation 10.8 as a function of V_C, but dR is a function of y. Since V_C monotonically grows along y, the variables y and V_C can be interchanged replacing y with $V_C(y)$ and $Q_n(y)$ with $Q_n(V_C)$. The eq. 10.15 can be integrated in y from 0 to L, and in V_C from V_S to V_D.

$$w \int_{V_S}^{V_D} Q_n(V_C) dV_C = I_d \int_0^L \frac{dy}{\mu_n} \tag{10.16}$$

Assuming μ_n constant along the channel we obtain the drain current

$$I_d = \frac{\mu_n w}{L} \int_{V_S}^{V_D} Q_n(V_C) dV_C \tag{10.17}$$

where $Q_n = -C_{ox}[V_G - V_{fb} - 2\|\phi_p\| + V_C - V_B] + \sqrt{2\epsilon_s q N_A (2\|\phi_p\| + V_C)}$.
The calculus of the integral can be simplified neglecting the dependance of the charge in the depletion layer from V_C. Obviously, the size of the depletion layer changes along the channel; however, the uncertainty about the actual value of the mobility in the channel (μ_n) makes tolerable the approximations. In practice, it is important here to calculate a plausible functional relationship between the current and the voltages leaving to the experimental calibration the evaluation of the actual parameters of the device.
Considering the threshold voltage definition (see eq. 10.8), the charge can then be written as:

$$Q_n = -C_{ox}(V_G - V_T - V_C) \tag{10.18}$$

The mobile charge decreases as V_C increases than towards to the drain contact. As a consequence, V_{DS} is not equally distributed along the channel but it is more dense towards the drain contact.
The drain current is obtained integrating the eq. 10.17.

$$I_d = -\mu_n C_{ox} \frac{w}{L} \left[(V_G - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \tag{10.19}$$

The negative sign indicates that the current flows from the drain to the source.
This equation is called the long-channel MOSFET equation. It describe a family of parabolic functions $I_d vs. V_{DS}$ with V_G as parameter. Noteworthy, the parabolic shape implies the existence of a negative differential resistance tract.
Actually, when $V_C = V_G - V_T$ the charge in the channel becomes null and the model, that is based on the continuity of the charge in the channel, is no more valid. The condition of null charge implies that the resistance in that tract becomes infinite. In other words, at $V_{DS} = V_G - V_T$ the channel is interrupted.
This condition is met at the top of the parabolic function when the current reaches the maximum value.

$$I_{dmax} = -\mu_n C_{ox} \frac{w}{L} \frac{(V_G - V_T)^2}{2} \tag{10.20}$$

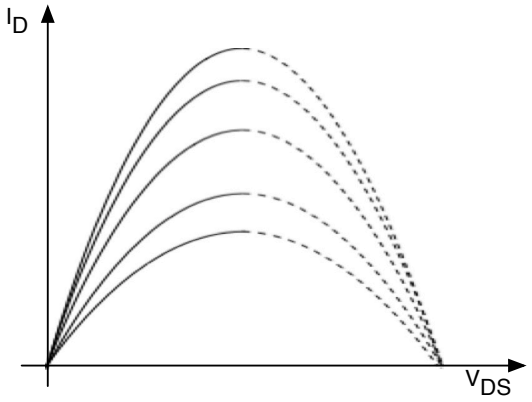

Fig. 10.5. *Drain current vs. source-drain voltage with the gate voltage as parameter. The model is correct until the maximum of the parabola is reached. After that the conductivity of a tract of the channel close to the drain contact becomes zero and the model is no more valid.*

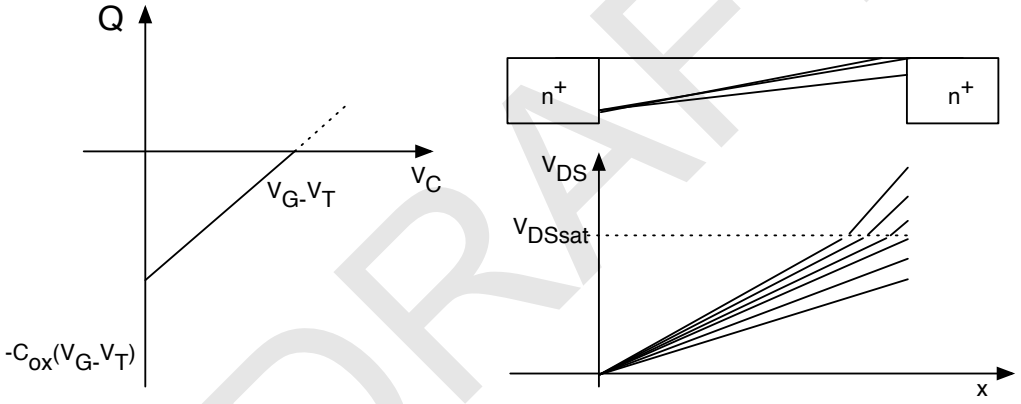

Fig. 10.6. *Left: Behaviour of the charge in the channel as a function of the offset voltage. Beyond $V_G - V_T$ the model predicts an impossible positive charge in the channel. Right: shape of the channel as V_{DS} increases and distribution of V_{DS} along the channel.*

The drain source voltage at which the maximum drain current is obtained is called the saturation voltage. It corresponds to the condition where Q_n is null at the drain contact. At $V_{DS} > V_{DSsat}$ any further increase of voltage drops at the border of the region where $Q_n = 0$. Then the rest of the channel remains still biased at V_{DSsat} then the current in the channels remains fixed to (I_{dmax}). The region depleted of mobile electrons is the *pinch-off* region. The pinch-off is very narrow and biased at $V_{DS} - V_{DSsat}$. Such a voltage across a small region gives rise to a large electric field than can drag the current from the end of the channel towards the drain contact.
The characteristics of the MOSFET are shown in figure 7. In the output characteristic, I_d vs. V_{DS}

is plotted, with V_G as a parameter. In figure 6 also the dependance of the I_{Dmax} from V_G is shown. The output characteristics is limited by current breakdown. At large V_{DS} the increased electric field confers to the accelerated electrons a kinetic energy sufficient to create by scattering electron-hole couple. This gives rise to an avalanche mechanism that elicits a sharp increase of the current similarly to the avalanche effect observed in the reverse biased PN junction.

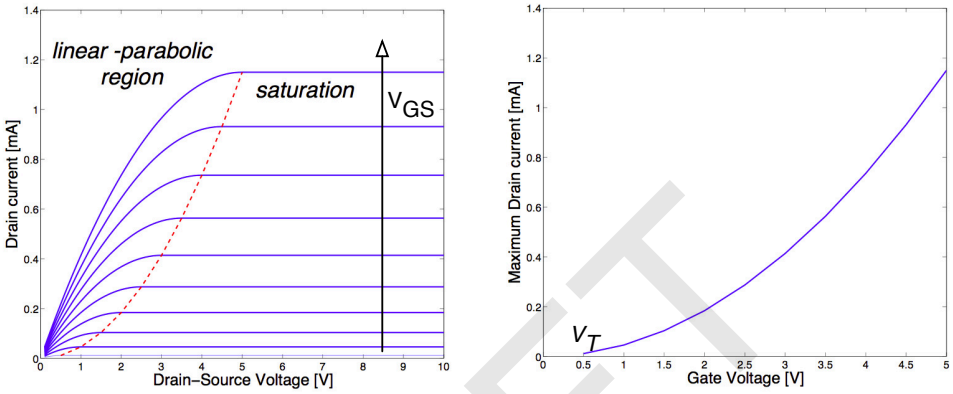

Fig. 10.7. *Left: MOSFET characteristics. The breakdown is not reported in the plot, but it has to be considered that at large V_{DS} the drain current undergoes a strong increase due to the avalanche effect. Right: dependence of the saturation current versus the gate voltage, note that the threshold voltage defines the onset of the current.*

In conclusion, the MOSFET experiences a number of different working conditions according to the combination of the voltages V_G and V_{DS}. These conditions encompass the cut-off, the sub-threshold condition, the linear-parabolic behaviour of the current (the so called triode region) and finally the saturation condition where the current does not depends on V_{DS}. Figure 10 reassumes all the conditions in the V_G - V_{DS} plane.

10.3.1 Channel length modulation

With the MOSFET in saturation regime, each further increase of V_{DS} falls across the pinch-off region. Then the size of the pinch-off region increases with V_{DS} and, as a consequence the channel of electrons decreases its length. Since The drain current is inversely proportional to the channel length, the increase of V_{DS} eventually induces an increase of the drain current.
This effect is similar to the Early effect in the BJT. It is interesting to note that in both BJT and MOSFET the working mechanisms of the devices avoids that the output characteristics are flat, namely that the devices behave as a ideal current source that is forbidden by the Kirchhoff network law.
The channel length modulation is sometimes expressed with a parameter λ that plays the role of the Early voltage in the BJT. The increase of current due to the channel length modulation is

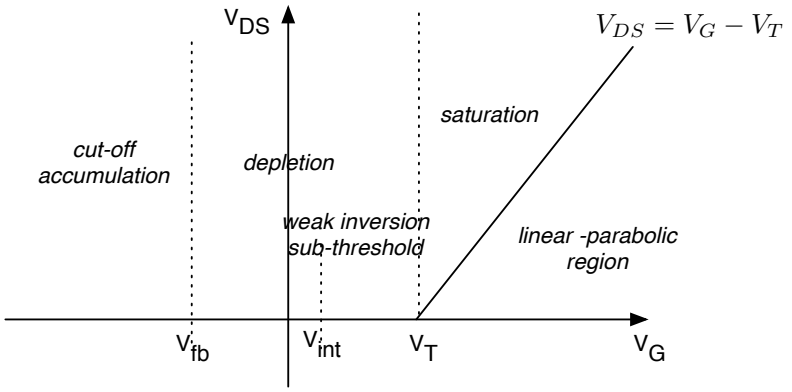

Fig. 10.8. *The different working conditions of the MOSFET depends on the combination of V_G and V_{DS}.*

expressed by the factor $[1 + \lambda \cdot (V_{DS} - (V_G - V_T))]$. In practice for large V_{DS} the saturation current may be written as:

$$I_{dsat} = (kV_G - V_T)^2(1 + \lambda V_{DS}) \tag{10.21}$$

The graphical meaning of lambda is given in figure 11.

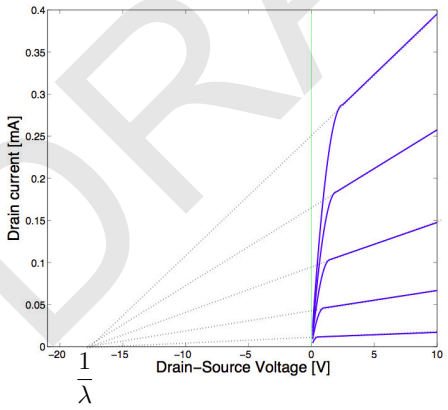

Fig. 10.9. *MOSFET characteristics with the channel length modulation considered. The parameter of channel modulation, sometimes called Early voltage after the nomenclature of BJT, is the origin of the slopes of the characteristics.*

10.3.2 Body effect

The body contact (V_B) provides the reference for the voltage applied to the device. In the previous derivation of the MOSFET characteristics it has been conveniently settled to zero. Actually, the body contact provides a further degree of freedom for the modulation of the threshold voltage. Considering the eq. 9.8 subtracting the case of $V_B \neq 0$ and $V_B = 0$ we get the change of threshold voltage due to the body voltage.

$$\Delta V_T = \frac{\sqrt{2\epsilon_s q N_A}}{C_{ox}} \left[\sqrt{2\|\phi_p\| + V_{SB}} - \sqrt{2\|\phi_p\|} \right] = \gamma \left[\sqrt{2\|\phi_p\| + V_{SB}} - \sqrt{2\|\phi_p\|} \right] \tag{10.22}$$

The quantity γ depends on the MOSFET parameters, and it is called the body effect parameter measured in units of $1/\sqrt{V}$.

10.3.3 Subthreshold current

At $V_G < V_T$ the channel is not formed, the interface with the oxide is depleted of mobile charges, but the concentration of electrons is larger than the concentration of holes. In practice, immediately below the threshold the concentration of electrons is similar to the concentration found in a lightly doped semiconductor. As a consequence, a little current of electrons can still flow between the drain and the source. This current is the subthreshold current. Although small, it is important because even when the MOSFET is nominally switched off ($V_G < V_T$) a current still exists. This current is an important component of the leak current that tends to discharge the memory cells.
In this condition, the interface region is a dielectric layer much thicker than then depletion layer separating the n^+-type regions and the p-type substrate. Then most of the electric field generated by V_{DS} is distributed towards the substrate. This makes negligible the drift current and the diffusion is the dominant contribution to the subthreshold current.
The current is ruled by the charge concentration at the drain and source junctions. In subthreshold condition, the conduction band in the channel is rather distant from the Fermi level, and then the junctions with the n^+ regions are not neglibles, and ϕ_s is the equilibrium barrier between the channel and the highly doped contacts, under bias the barriers at the drain and source junctions are: $\phi_s - V_D$ and $\phi_s - V_S$ and the different barrier heights gives rise to a gradient of charges in the channel. Then the current due to V_{DS} is $j = -qD_n \frac{\partial n}{\partial y}$.
The concentrations of electrons at the edges of the channel (n_D at the drain and n_S at the source) are calculated as:

$$n_D = n_i exp\left(\frac{q(\phi_s - V_D)}{kT}\right);\ n_S = n_i exp\left(\frac{q(\phi_s - V_S)}{kT}\right) \tag{10.23}$$

In order to estimate the current let us assume a linear behaviour of the charges from the drain to the source.

$$j = -qD_n \frac{n_D - n_S}{L} = -q\frac{D_n}{L} n_i exp\left(\frac{q\phi_s}{kT}\right)\left[exp\left(-\frac{qV_D}{kT}\right) - exp\left(-\frac{qV_S}{kT}\right)\right] \tag{10.24}$$

Usually, $V_D \gg \frac{kT}{q}$, then the first exponential is negligible:

$$j = -q\frac{D_n}{L}n_i exp\left(\frac{q(\phi_s - V_S)}{kT}\right) \tag{10.25}$$

In strong inversion, ϕ_s is independent from V_G and it is equal to $-\phi_p$. But in sub-threshold condition, ϕ_s is proportional to V_G. We can then replace ϕ_s with V_G but introducing a factor η that takes into consideration the fact that ϕ_s is smaller than V_G. Then the sub threshold current can be written as:

$$j_{st} = -q\frac{D_n}{L}n_i exp\left(\frac{qV_{GS}}{\eta kT}\right) \tag{10.26}$$

The subthreshold current has an exponential dependence with V_{GS} while the current in inversion is proportional to V_{GS}^2. This difference is due to the different nature of the currents, being the inversion current a drift current and the sub threshold current a diffusion current. Note that the source contact is normally grounded, here V_S is maintained as a reference for the gate potential.

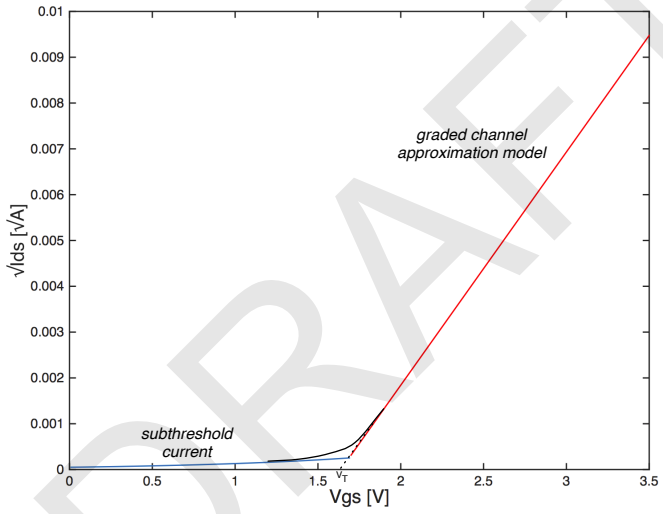

Fig. 10.10. *Behaviour of the root square of the saturation current versus the gate voltage. The intercept of the graded approximation model provides the threshold voltage. Below the threshold voltage the small sub-threshold current is plotted. The transition between the two regimes provides a deviation from both models around* V_T.

It is important to consider that the subthreshold current is actually present also when $V_{GS} > V_T$ but its amplitude is smaller than the drift current and then it is of course neglected. Indeed, the magnitude of the gradient is independent on the quantity of charges. We may have a large diffusion current with a small amount of charge and, on the other hand, a large drift current with a negligible gradient of charges.

10.3.4 The transit time

The response time of the MOSFET mainly depends on the time necessary to charge the depletion layer and on the transit time of the electrons in the channel. The time to charge the depletion layer has been treated in the PN junction while the transit time deserves a little more attention since it is peculiar of the MOSFET.
The velocity of the electrons is variable along the channel because the concentration is variable, then the variable velocity ensures that the drain constant is maintained the same all along the channel. The transit time can be calculated with the following integral:

$$T_{tr} = \int_0^L \frac{1}{v(y)} dy = -\int_0^L \frac{1}{\mu_n \mathcal{E}(y)} dy \tag{10.27}$$

where $\mathcal{E}(y) = \frac{dV_C(y)}{dy}$.
The function $V_C(y)$ can be calculated from the definition of the drain current:

$$\int_0^y I_d dy = -w\mu_n C_{ox} \int_0^{V_C(y)} (V_G - V_T - V_c) dV_C \tag{10.28}$$

from where $V_C(y)$ is extracted:

$$V_C(y) = V_G - V_T - \sqrt{(V_G - V_T)^2 - \frac{2I_d y}{w\mu_n C_{ox}}} \tag{10.29}$$

replacing the drain current with its saturation value (eq. 9.19) we obtain:

$$V_C(y) = (V_G - V_T) - (V_G - V_T)\sqrt{1 - \frac{y}{L}} \tag{10.30}$$

Then the electric field $\mathcal{E}(y)$ is:

$$\mathcal{E}(y) = \frac{dV_C(y)}{dy} = -\frac{V_G - V_T}{2} \frac{1}{\sqrt{1 - \frac{y}{L}}} \tag{10.31}$$

Replacing the electric field in eq. 10.27 and solving the integral the transit time is found:

$$T_{tr} = \frac{4}{3} \frac{L^2}{\mu_n (V_G - V_T)} \tag{10.32}$$

The transit time strongly depends on the channel length and on the mobility. The decrease of the channel length and the increase of the mobility are the possible actions to reduce the transit time, and then, the response time of the device.

10.4 Short channel MOSFET

In the previous section the so-called long channel model of the MOSFET has been derived. The long channel length condition ensures the graded channel approximation that is the basis for the derivation of the above model. Furthermore, the current has also been calculated assuming a constant mobility in the channel. We will see soon that this is one of the conditions that has to be

removed when a short channel MOSFET is considered.
The length of the channel is inversely proportional to the drain-source current (eq. 10.19) and to the transit time (eq. 10.32), therefore the decrease of the channel length is a optimal choice to increase the device performance. Actually, the reduction of the channel length has a number of additional consequences, not always positive, that has to be considered. We will see later that in order to maintain the performances the scaling of the channel length has be paralleled by a scaling of the other MOSFET parameters.
Hereafter, some of the consequences of channel length scaling are introduced.

10.4.1 Threshold voltage modulation

The n^+ doped regions that make the drain and source contacts create a PN junction with the p-type substrate. As a consequence, two additional depletion layers, over imposed to the depletion layer of the MOS, appear at the contact-substrate interfaces. In normal conditions the PN junctions are reverse biased than the additional depletion layer tends to expand towards the substrate and also towards the channel. In a long channel device, the size of these regions can be neglected with respect to the size of the depletion layer of the MOS. However, when the distance between drain and source is small the contributions of these two regions to the total depletion layer are no more negligible. In the depletion layers pertinent to the PN junction the acceptors charge is not controlled by V_G but rather by V_D and V_S.
As a consequence, the charge of the depletion layer controlled by V_G decreases. This charge is the last term of the threshold voltage (eq. 10.9), then if the depletion layer charge under control of V_G also the threshold voltage becomes smaller, of course with respect to a similar device where only the channel length is changed, as shown in the next equation:

$$V_T = V_{fb} - 2\|\phi_p\| + \frac{Q_d^*}{C_{ox}} \tag{10.33}$$

where Q_d^* is the portion of the charge of the depletion layer actually controlled by V_{GS}. Since the source is normally grounded the threshold voltage modulation occurs at the drain contact.
V_T has an exponential behaviour with the channel length. As shown in eq. 10.33 the decrease of Q_d^* can be compensated, or increased, by the oxide capacitance. Leaving untouched the materials, a reduction of the oxide thickness can properly counteract the threshold voltage decrease. As anticipated in the previous section, in order to maintain the device property the scaling of the length of the channel has to be complemented by the scaling of other quantities, in this case to maintain the same threshold voltage the oxide thickness has also to be reduced.

10.4.2 Drain Induced Barrier Lowering

The reduction of the distance between source and drain n^+ regions makes possible the merging of the depletion layers across the substrate. Figure 11 shows the band diagram from the source to the drain across the substrate. The drain contact is positively biased while the source and the substrate are kept grounded. In this condition, the drain-substrate junction is reverse biased the barrier becomes larger and the depletion layer is expanded. In a long channel device, this does not alter the conditions of the source-substrate junction, but in a short channel device, the depletion layer can become so large to influence the barrier of the source-substrate junction. As a consequence, an additional current of electrons from the source to the drain is found. This current is summed to

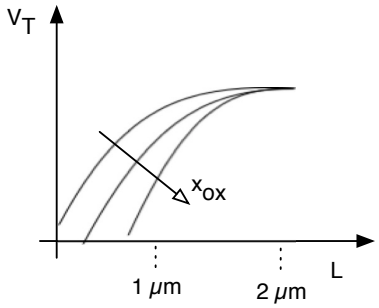

Fig. 10.11. *Behaviour of the threshold voltage with the channel length.*

the subthreshold current and depends on V_{DS}. This phenomenon is similar to the punchthrough of the base of a BJT. The Drain Induced Barrier Lowering can be counteracted increasing the doping of the substrate in order to limit the expansion of the depletion layer, however this reduces the mobility and then the current. Alternative solutions consist in making the substrate so thin to be completed depleted also at $V_{GS} = 0$.

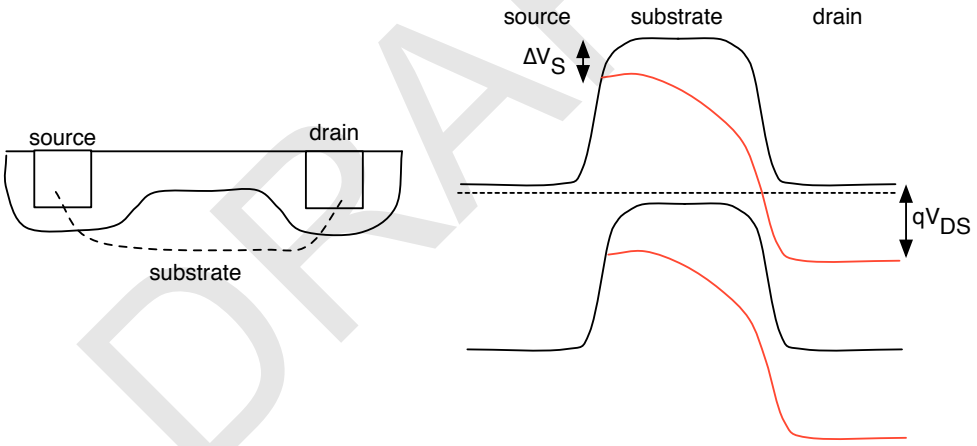

Fig. 10.12. *Scheme of the Drain Induced Barrier Lowering effect. Left: path from source to drain through the substrate; right: band diagram along the path, under reverse bias at the drain contact, the depletion layer expands through the substrate until to reduce the source to substrate barrier.*

10.4.3 Velocity saturation

When the channel is short the effects of the non linear mobility discussed in chapter 1 becomes manifested. Indeed, as L decreases the electric field increases and the velocity of the electrons

tends to reach the saturation value. The effect is favoured by the fact that since the mobility in the channel region is smaller with respect to the bulk, the saturation occurs at smaller electric fields. Of course also the saturation velocity itself is smaller. The typical values for the silicon are: $v_{sat}^n \approx 6 \div 10 \cdot 10^6 cm/s$; $v_{sat}^p \approx 4 \div 8 \cdot 10^6 cm/s$ for electrons and holes respectively. Of course the largest velocities are found in the bulk.
The main consequence of the non linear mobility is that the device may reach the saturation before the pinch-off condition and then at a V_{DS} smaller respect to that predicted by the long channel model
The drain current in the regime of non linear mobility is calculated considering the behaviour of the mobility with the electric field described in fig 1.22.
The relationship between the electric field and the velocity is given by the following piecewise relationship:

$$\begin{aligned} \mathcal{E} < \mathcal{E}_{sat} \quad & v = \frac{\mu_{eff}\mathcal{E}}{1 + \frac{\mathcal{E}}{E_{sat}}} \\ \mathcal{E} \geq \mathcal{E}_{sat} \quad & v = v_{sat} \end{aligned}$$

Where μ_{eff} is a parameter calculated from the saturation electric field:

$$\mathcal{E}_{sat} = 2\frac{v_{sat}}{\mu_{eff}} \tag{10.34}$$

In order to calculate the current let us consider:

$$I_d = qnAv = wQ_n v = wC_{ox}\left[V_G - V_T - V_c(y)\right]\frac{\mu_{eff}\mathcal{E}}{1 + \frac{\mathcal{E}}{E_{sat}}} \tag{10.35}$$

In the channel, the space coordinate progresses from 0 to y, and the voltage from 0 to V_{DS}, in this situation the electric field is directed from drain to source, and then it is negative. With a little algebra, the above equation is rearranged in :

$$I_d = \left(C_{ox}\mu_{eff} w\left[V_G - V_T - V_c(y)\right] - \frac{I_d}{\mathcal{E}_{sat}}\right)(-\mathcal{E}) \tag{10.36}$$

Considering that $\mathcal{E} = -dV_c/dy$ and integrating dy from 0 to L and dV_c from 0 to V_{DS} we have:

$$\int_0^L I_d dy = \int_0^{V_{DS}} C_{ox}\mu_{eff} w\left[V_G - V_T - V_c(y)\right] dV_c - \int_0^{V_{DS}} \frac{I_d}{\mathcal{E}_{sat}} dV_c \tag{10.37}$$

whose solution gives the short channel model of the MOSFET

$$I_d = -\frac{\mu_{eff}C_{ox}w}{L}\left[(V_G - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right]\frac{1}{1 + \frac{V_{DS}}{\mathcal{E}_{sat}L}} \tag{10.38}$$

Again, the negative sign means that the current flows from the drain to the source. The current equation is the same calculated for the long channel device divided by a correction term that takes into consideration the decreasing mobility. The calculation has been carried out using the same approximations of the long channel MOSFET; namely μ_{eff} is constant along the channel and the variation of the charge in the depletion layer is negligible. Nonetheless, this simplified equation is

sufficiently accurate to describe the behavior of the device.
At $\mathcal{E}_{sat} \gg \frac{V_{DS}}{L}$ the equation converges to the long channel model, hence we can consider equation 10.38 as the complete model for MOSFET current. The long channel condition is defined as:

$$L \gg \frac{V_{DS}}{\mathcal{E}_{sat}} \tag{10.39}$$

Considering $\mathcal{E}_{sat} = 4 \cdot 10^4$ V/cm and V_{DS}=5 V, the long channel model requires $L \gg 1.25\ \mu m$. A obsolete conditions for integrated MOSFET.
Due to the velocity saturation, the current reaches its saturation value at $V_{DS} = V_{Dsat}$ which is smaller than the value predicted by the long channel model. In practice, the saturation is a combined effect of the mobility reduction and the decrease of charge in the channel.
In saturation the charge in the inversion channel is :

$$Q_{nsat} = -C_{ox}(V_G - V_T - V_{Dsat}) \tag{10.40}$$

at which corresponds the saturation current

$$I_{dsat} = wC_{ox}(V_G - V_T - V_{Dsat})v_{sat} = wC_{ox}(V_G - V_T - V_{Dsat})\frac{\mu_{eff}\mathcal{E}_{sat}}{2} \tag{10.41}$$

V_{Dsat} occurs at the intersection of the saturated current (Eq. 10.41) and the short channel current model (Eq. 10.38).

$$\frac{\mu_{eff}C_{ox}w}{L}\left[(V_G - V_T)V_{Dsat} - \frac{V_{Dsat}^2}{2}\right]\frac{1}{1 - \frac{V_{Dsat}}{\mathcal{E}_{sat}L}} = wC_{ox}(V_G - V_T - V_{Dsat})\frac{\mu_{eff}\mathcal{E}_{sat}}{2} \tag{10.42}$$

From which V_{Dsat} is calculated:

$$V_{Dsat} = \frac{\mathcal{E}_{sat}L(V_G - V_T)}{\mathcal{E}_{sat}L + (V_G - V_T)} \tag{10.43}$$

At $\mathcal{E}_{sat}L \gg V_G - V_T$, V_{Dsat} coincides with that predicted by the long channel model. This gives a further condition for the long channel MOSFET:

$$L \gg \frac{V_G - V_T}{\mathcal{E}_{sat}L} \tag{10.44}$$

On the contrary, when $\mathcal{E}_{sat}L \ll V_G - V_T$, the saturation voltage is simply $V_{Dsat} = \mathcal{E}_{sat}L$, and the saturation current is:

$$I_{dsat} = wC_{ox}(V_G - V_T - \mathcal{E}_{sat}L)v_{sat} \tag{10.45}$$

Note that the saturation current increases as the channel length decreases.
In figure 14 the behaviour of the charge in the channel and the electrons velocity in both the models is illustrated.

In the short channel MOSFET the saturation current depends linearly from $V_G - V_T$ while in a long channel MOSFET the dependence is quadratic. Furthermore the saturation condition occurs at smaller V_{DS}.

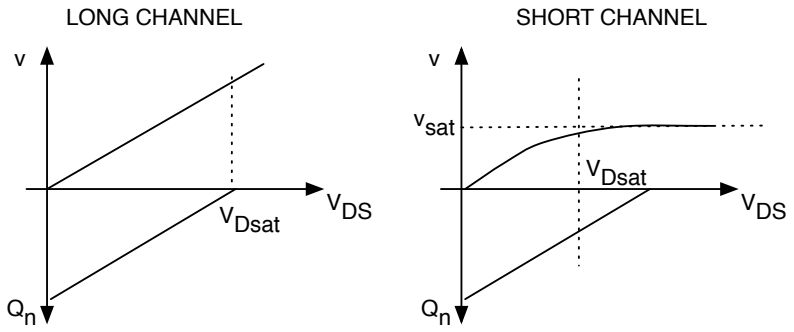

Fig. 10.13. *Behaviour of channel charge and velocity in long and short channel MOSFETs. In case of a short channel MOSFET, the saturation is given by a combination of mobility decrease and charge reduction.*

As the channel length approaches zero the saturation voltage also V_{Dsat} also becomes small. This because when L is very small, each applied voltage may provide an electric field greater than the saturation electric field. On the contrary, In the long channel MOSFET the current diverges as L tends to zero. Actually, this effect is expected in any conductor and it provides a necessity for the saturation of the velocity. The largest possible current is obtained at $L = 0$ it is given by:

$$I_{dmax} = wC_{ox}(V_G - V_T)v_{sat} \tag{10.46}$$

The ratio between the saturation current and the largest current is the ideality factor (α) which provides a measure of the actual current respect to its largest theoretical value:

$$\alpha = \frac{I_{dsat}}{I_{dmax}} = \frac{(V_G - V_T) - V_{Dsat}}{V_G - V_T} \tag{10.47}$$

for instance with $x_{ox} = 40\ nm;\ L = 1\ \mu m;\ V_{Dsat} = 13\ V; V_G - V_T = 4.3\ V$ the ideality factor is $k = 0.72$. Any other reduction of the dimension of the device can achieve no more than 30 % of increase of current. Actually we have also to consider that the threshold voltage changes with L.

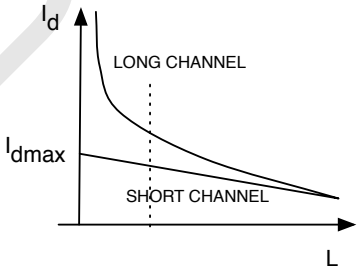

Fig. 10.14. *Behaviour of the drain current with L in the long and short MOSFET models.*

10.4.4 Transit time

The transit time can be calculated using the same approach of the long channel case. In saturation condition, the result for a short channel gives a linear dependence of the transit time from the channel length.

$$T_{tr} = \frac{4}{3}\frac{L}{v_{sat}} \tag{10.48}$$

This has to be compared with the case of the long channel model (eq. 9.30) where the transit time depends quadratically from the channel length.
Then in short channel MOSFET the decrease of transit time, and the increase of the cut-off frequency, is less dependent from the dimension of the device.
This means that with silicon MOSFETs, without considering many other important effects, the performance of the devices cannot be indefinitely increased by the dimension reduction. In particular the extension of the cut-off frequency towards larger frequencies requires additional strategies and in particular the introduction of semiconductors with larger mobility such as the gallium arsenide. However, since the interface with the insulators of these materials are worse than the case of silicon a different kind of field effect device has to be introduced.

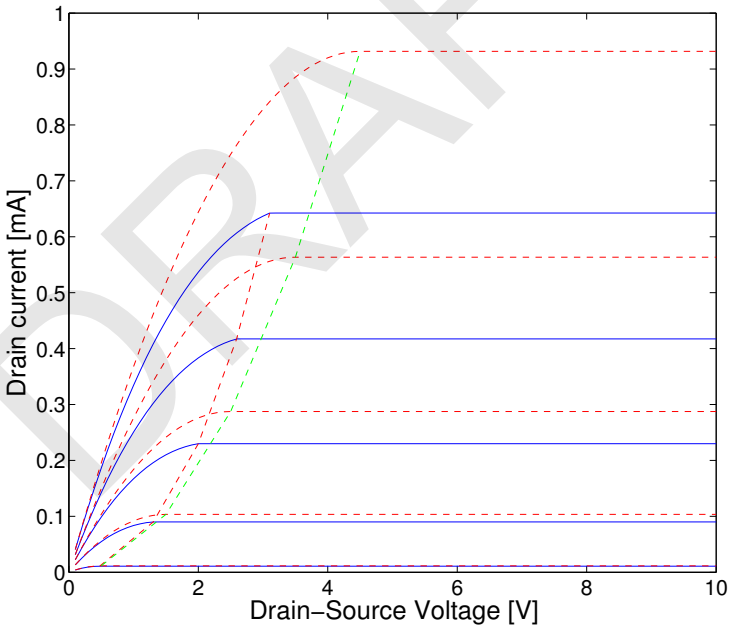

Fig. 10.15. *Comparison between short channel model (continuous line) and long channel model (dotted line). The lines marking the onsets of saturation in both the cases are also plotted.*

10.4.5 Scaling

Thanks to the fundamental use of the MOSFET in electronics, in particular in digital electronics, the reduction of the dimensions of the MOSFET prompted the great technological advances of microelectronics. The self-insulation properties of MOSFETs with respect to the rest of the substrate and the relatively easy implementation of complementary MOSFET based on n-type and p-type silicon (CMOS technology) are at the basis of the great quest for integration of always more complex and miniaturised devices.
The reduction of the dimensions of the MOSFETs has been dramatic during the last decades, and they are reassumed in the so-called Moore's law. Gordon Moore gave different version of its law. In 1965, when he was the director of research and development at Fairchild Semiconductor, he observed that the complexity for minimum component costs was increasing at a rate of roughly a factor of two per year. In 1975 he stated that the circuit density-doubling would occur every 24 months. This last sentence was nicknamed as Moore's law. In simple terms the law fixes a general and constant trend about the doubling of the number of transistors in integrated circuits. In spite of the changes of technologies the law is still quite valid nowadays. The Moore's law is very evident considering the increase of performances of devices directly related to the number of devices such as the solid-state memories and the image sensors whose capacities increased in spite of a reduction of their prices.
As discussed in the previous sections, the scaling of the MOSFET can induce a number of non ideal behaviors that can be adequately corrected with a proper scaling of the dimensions involving not only the size reduction of the gate length and width but also a proper change of the other quantities such as the oxide thickness and the depletion layer widths which also implies scaling of the semiconductor doping.
The device can be scaled choosing to maintain constant either the electric field or the voltage. A constant electric field avoids the decrease of mobility due to the large electric field but it requires a reduction of the voltage making the device no more compatible with the existing circuits. On the other hand, a constant voltage scaling exposes the device to the risk of mobility degradation. In table 1 the scaling rule for constant voltage are shown.

Table 10.1. Scaling rule preserving either the voltage or the electric field, the scaling factor is k

	parameter	constant voltage scaling	constant field scaling
Scaling assumptions	Gate length	1/k	1/k
	Gate width	1/k	1/k
	Oxide thickness	1/k	1/k
	Semiconductor doping	k	k
Derived quantities	Electric field	k	1
	Oxide capacitance	k	k
	Transit time	$1/k^2$	$1/k^2$
	Voltage	1	1/k
	Current	k	1/k
	Power	k	$1/k^2$

10.5 CMOS configuration

The MOSFET described in the previous sections was made on a p-type substrate where the channel is made of electrons. Of course the complementary configuration can be fabricated, namely on a n-type substrate and a channel made of holes. The different devices are indicated as n-mos and p-mos respectively. The characteristics of the p-mos are opposite to the n-mos. For instance the threshold voltage is negative and the a more negative gate voltage is necessary to activate the channel. In practice the channel formation conditions are $V_G - V_T > 0$ for a n-mos and $V_G - V_T < 0$ for a p-mos, with the consideration that V_T in case of p-mos is negative.

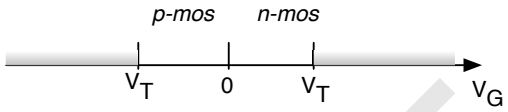

Fig. 10.16. *The conditions on V_G to achieve the conductive channel in p-mos and n-mos.*

The two complementary devices can be formed on the same substrate where a well contrarily doped with respect to the substrate is created. The well hosts the second device as shown in figure 10.18. This opportunity gives rise to a composed device called CMOS (complementary MOS). The more interesting use of CMOS is the inverter function that is the main building block of logic circuits.

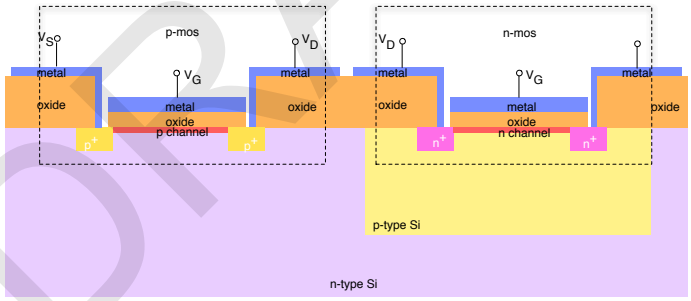

Fig. 10.17. *Constructive scheme of the CMOS structure.*

Figure 10.19 shows the CMOS circuit and the connections of the physical device. The gate voltages of the two transistors $V_G^{nmos} = V_{in}$ and $V_G^{pmos} = V_{in} - V_D$. As V_{in} increases from ground to V_{DD} the charge at the oxide-semiconductor interface changes, in particular considering the conditions at which conductive channels are formed in the two devices (see fig. 10.17) for small values of V_{in} only the pmos transistor conducts currents while at higher values only the nmos transistor conducts. Ideally, when both the transistors have the same V_T and the sub threshold current is negligible the current between V_{DD} and ground is always zero. Actually, since the threshold voltages are different and the sub threshold current is greater than zero a small current actually flows as

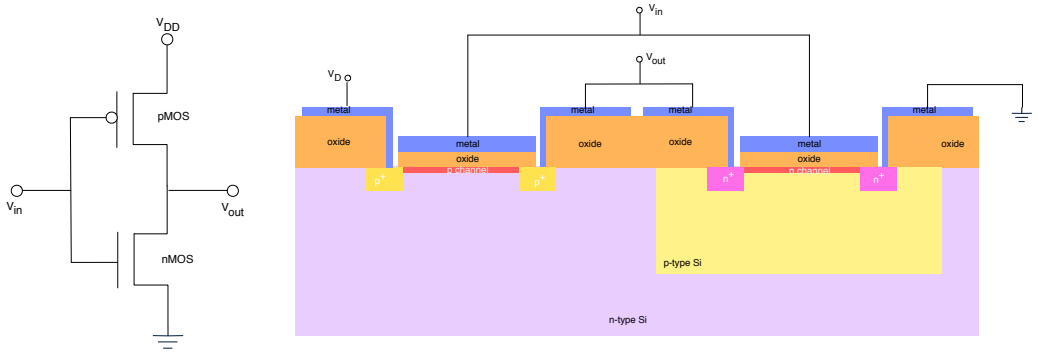

Fig. 10.18. *Inverter function made with complementary MOSFETs and the physical implementation with a CMOS structure.*

V_{in} goes from low to high states. Figure 10.19 shows the gate voltages versus V_{in} and the transfer function of the inverter circuit. It is interesting to note that, considering the inverter transfer function of figure 10.20, a large amplification factor, defined as dV_{out}/dV_{in} is achieved in the short interval where the current flows in the inverter.

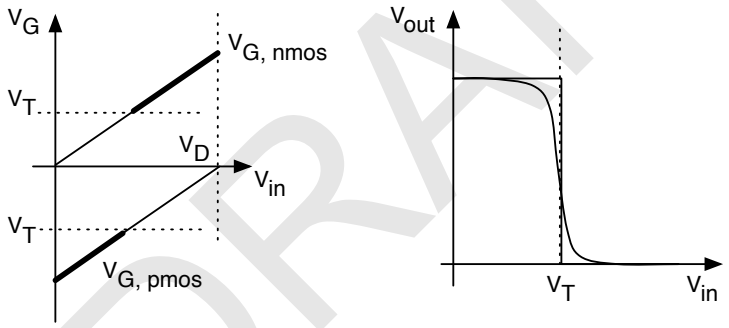

Fig. 10.19. *Left) gate voltages as a function of the input voltage. Thicker lines show the open channel conditions. Right) inverter transfer function, deviations from ideality results in a gradual change of the output voltage around the transition voltage.*

The CMOS structure contains a sequence of regions n and p that forms a combination of npn and npn bipolar transistors. Due to the reduced dimension of the MOSFETs in the CMOS structure, the base regions of the two transistors are sufficiently short to allow for the transistor effect to take place. As shown in figure 10.21, the two transistors are connected with a base-collector loop that may give rise to a positive feedback that results in a steadily growing current from power supply to the ground. In figure 10.21 the large current can be initiated by a sudden increase of the reverse current at the drain-bulk contact in the pmos. Such an increase can be due a energy release provided, for instance, by a background emitted ionizing particle. This disruptive phenomenon is called latch-up

and it has to be avoided in practical devices. A method to avoid the latch-up is the addition of recombination centers in the substrates, for instance through the injection of gold atoms, in order to increase the recombination, and then to reduce the amplification of the transistors. Another method is the formation of a deep oxide trench in order to physically insulate the two transistors.

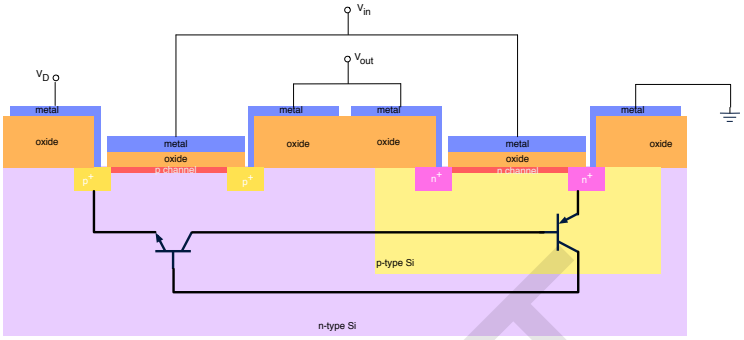

Fig. 10.20. *Parasite pop and npn transistors in the CMOS forms a positive feedback loop between the collector and the base terminals. Then a small increase of current at the emitter of the pnp transistor may give rise to a large current that can destroy the component.*

10.6 Metal Semiconductor Field Effect Transistor (MESFET)

The scaling of dimensions is a powerful method to improve the performance of MOSFETs in terms of density of devices, levels of current and operative frequency. However, to further enhance the response time of the device extending the operation towards high frequencies it is necessary to replace the semiconductor with a material with larger mobility.
Such materials are those made with atoms of the III and the V group of the periodic table (note that silicon and germanium belongs to the IV group). Among then gallium arsenide (GaAs) is the most widely used.
The design of devices based on III-V elements is affected by some technological limitations in particular due to the lack of a natural oxide whose interface has the same characteristics of the silicon dioxide / silicon system. The MOSFET configuration with these materials leads to a channel so defected that the performances of transistor are not adequate for a practical exploitation. Due to this limitations, the devices based on III-V semiconductors are designed following a different architecture where the oxide layer is omitted and the gate is formed by a direct metal-semiconductor junction (a Schottky diode) that, during the operation, is reverse biased. Such a device is called Metal Semiconductor Field Effect Transistor (MESFET) and in spite of the non null gate current it allows to combine the field effect transistor principle with the high electrons mobility of III-V semiconductors to achieve a high frequency operation.
The velocity vs. electric field of GaAs and the others III-V semiconductors has been discussed in chapter 5. It is characterised by a maximum velocity and a negative differential tract that converges towards the saturation velocity. The mobility of electrons is about $\mu_{GaAs} \approx 8000 \frac{cm^2}{Vs}$ compared to

the mobility in silicon that is $\mu_{Si} \approx 1400 \frac{cm^2}{Vs}$. On the other hand, the mobility of holes is lower in GaAs with respect to silicon. This asymmetry hinders the possibility to develop complementary devices as in the CMOS technology.
GaAs has a wider band gap with respect to silicon ($E_{gap} = 1.24\ eV$) therefore the intrinsic concentration of electrons is also smaller ($n_i = 9 \cdot 10^6\ cm^{-3}$). As a consequence the intrinsic GaAs is less conductive than the intrinsic silicon. This is an interesting property that is exploited in the design of devices.
The MESFET configuration is shown in figure 10.22. A n-doped GaAs is formed above a substrate of intrinsic material, the gate contact is evaporated metal (as in the Schottky diode) and two n^+ regions form the drain and the gate contacts.

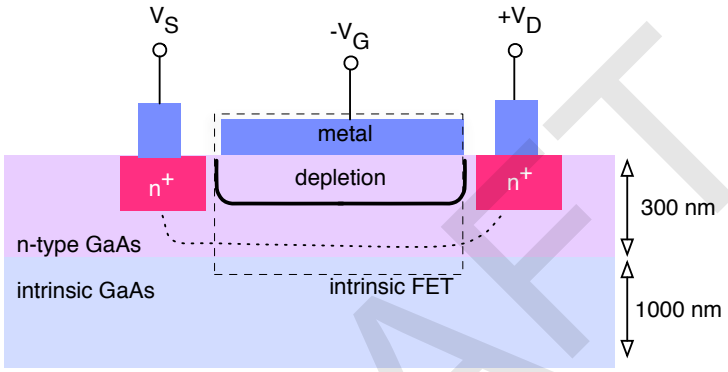

Fig. 10.21. *Schematic drawing of a MESFET. The dotted line indicates the current path from drain to source contacts. The dashed line indicates the intrinsic FET.*

The metal is chosen in order to form a rectifying Schottky contact (Schottky diode) whose characteristics have been described in chapter 2.
At the equilibrium, a depletion layer is formed at the interface between the metal and the doped semiconductor. The presence of the depletion layer shrinks the conductive region in the N-type GaAs between the drain and the source contact giving rise to a conductive structure analog to the MOSFET channel
The scope of the device is to control the drain-source current by the voltage applied to the gate contact. This is obtained when the the metal-semiconductor junction is under reverse bias. In this condition, the depletion layer becomes more wider with respect to the equilibrium. If the thickness of the doped semiconductor layer is small the conductivity is largely modulated by the applied voltage. Note that the device behaves like a MOSFET in depletion mode.
The great advantage with respect to the MOSFET is that the current flows in the bulk of the device at the interface between the doped and the intrinsic semiconductor. However, the current still flows in a doped material and then the mobility is lower with respect to the largest mobility that is expected in the intrinsic material.
The voltage between drain and source contacts is distributed along the channel as indicated by the dotted line in figure 10.22. The voltage acts a further reverse bias of the metal-semiconductor

junction contributing to increase the size of the depletion layer according to the profile shown in figure 10.23. The effect is similar to that observed in the MOSFET with the great different that, here, the current is formed by the majority charges without neither inversion nor accumulation and it flows in the bulk.
The n-type region is of the order of $10^{17}\ cm^{-3}$. Then the contact resistance between the drain and source contacts and the intrinsic FET structure (see fig. 17) is very small and V_{DS} is applied across the channel.

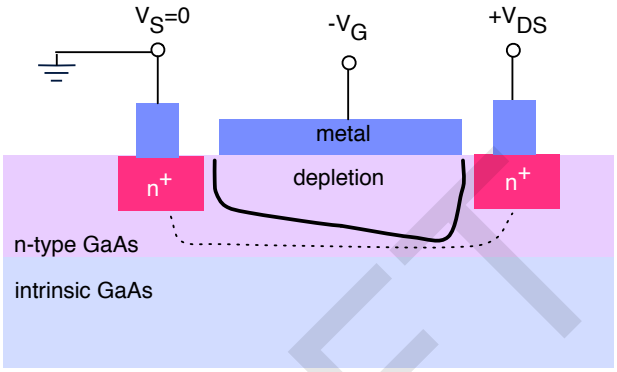

Fig. 10.22. *The MESFET under bias. Note that the gate voltage is negative and the depletion layer is deformed by the action of V_{DS}.*

The total voltage drop across the depletion layer is: $V_s = \phi_i - V_G + \phi_c(y)$. Where ϕ_i is the built-in potential of the metal-semiconductor junction, V_G is the gate potential and $V_c(y)$ is the voltage due to V_{DS}.
The depth of the depletion layer is (see eq. 2.15):

$$x_d(y) = \sqrt{\frac{2\epsilon_s}{qN_D}\left[\phi_i - V_G + \phi_c(y)\right]} \tag{10.49}$$

As in the MOSFET, the V_{DS} is distributed in the channel according to the conductivity. The conductivity is modulated by the depletion layer depth. Being t the thickness of the doped layer, the differential resistance in the tract dy of the channel is:

$$dR = \frac{1}{q\mu_n N_D}\frac{dy}{w(t - x_d y)} \tag{10.50}$$

where w is the transverse dimension of the device. The current $I_d = d\phi/dR$ is :

$$I_d = \frac{q\mu_n N_D w[t - x_d(y)]}{dy} d\phi \tag{10.51}$$

The above equation can be solved replacing x_d with equation 10.48 and integrating, as usual, dy from 0 to L and V_c from 0 to V_{DS}. The result provides the characteristics of the MESFET

$$I_d = \frac{q\mu_n N_D wt}{L}\left[V_{DS} - \frac{1}{t}\sqrt{\frac{2\epsilon_s}{qN_D}}\frac{2}{3}\left[(\phi_i - V_G + V_T)^{3/2} - (\phi_i - V_G)^{3/2}\right]\right] \tag{10.52}$$

The I/V curve is non linear with an exponent 3/2 with respect to the exponent 2 found in MOSFET.
The current reaches a maximum value when the channel touches the intrinsic layer. In this condition the model is no more valid. As in the MOSFET, the model predicts a physical impossible positive charge. Since the intrinsic layer is quasi-insulating the situation is analog to the pinch-off of the MOSFET. The same argument used in section 9.3 can be applied here to justify the fact that after the pinch-off condition the current maintains its maximum value that is ideally independent from V_{DS}.
The saturation condition is $x_{dsat}(V_{dsat}) = t$, and the saturation voltage is:

$$V_{dsat} = \frac{qN_D t^2}{2\epsilon_s} - \phi_i - V_G \tag{10.53}$$

The saturation voltage depends on the thickness of the doped layer and on the doping level (N_D and ϕ_i). The first two terms of the eq. 51 define a sort of threshold voltage for the MESFET. Differently from the MOSFET this voltage does not define the occurrence of the channel, but it is the value of V_G, at $V_{DS} = 0$, for which the channel disappears.
Replacing V_{dsat} in the equation 9.49 the relation between saturation current and V_G is: $I_{dsat} \approx (V_G - V_T)^{3/2}$.
The exact relationship of the saturation current is:

$$I_{dsat} = I_p\left[\frac{1}{3} - \left(\frac{\phi_i - V_G}{V_p}\right) + \frac{2}{3}\left(\frac{\phi_i - V_G}{V_p}\right)^{2/3}\right] \tag{10.54}$$

where $I_p = \frac{m_n q^2 N_D^2 t^3 w}{2\epsilon_s L}$ and $V_p = \frac{qN_D t^2}{2\epsilon_s}$.
The term I_p being inversely proportional to L contains the modulation of the current due to channel length and the same considerations done in the case of MOSFET about the channel length modulation apply also to the MESFET.

The MESFET structure exists also in silicon where it is called Junction Field Effect Transistor (JFET). Due to the non negligible conductivity of intrinsic silicon, the MESFET configuration cannot work in silicon. In case of silicon the device is designed in order to confine the conductivity at the centre of the semiconductor. The channel is created by a symmetric couple of PN junctions located at the both sides of the semiconductor. A drawing the of the JFET is shown in figure 20. Besides the different symmetry, the JFET equations are rather similar to the MESFET equations. The JFET allowed for the development of a field effect transistor avoiding the problems of mobility reduction occurring at the oxide/semiconductor interface. On the other hand, being a non planar device it is not suitable for integrated circuits where the planar architecture is a fundamental requirement for devices integration. The introduction of silicon-on-insulator technology makes possible the growth of thin doped silicon layer on an insulator and the fabrication of a transistor following the MESFET architecture.

10.7 High Electron Mobility Transistor: HEMT

The MESFET offers the possibility to exploit the high mobility of GaAs and similar materials. This makes possible to extend operative frequency up to the microwave region. The lack of a good

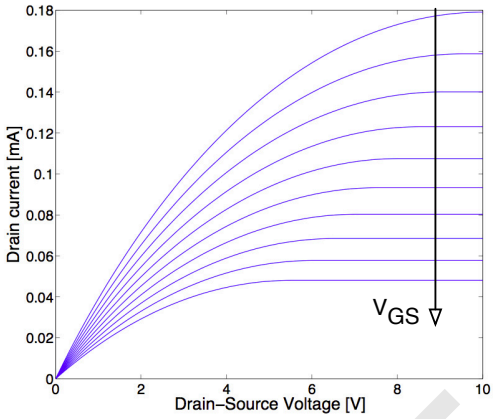

Fig. 10.23. *Output characteristics of a MESFET.*

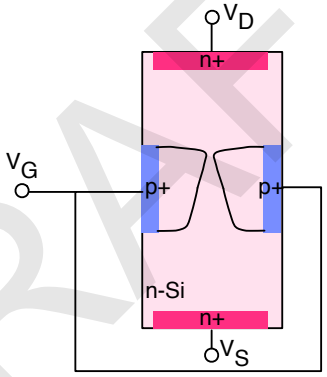

Fig. 10.24. *Structure of the Junction Field Effect Transistor. The gate contact is a PN junction with a heavily doped p-silicon in order to confine the depletion layer into the n-type semiconductor. Two gates are used to confine the channel in the bulk of the semiconductor.*

semiconductor-insulator interface prompted the design of a device where the channel is formed in the bulk of the doped semiconductor and the gate voltage simply modulates the channel width through the modulation of the depletion layer of the metal-semiconductor junction.
However, the performances are still limited by the fact that the current flows in a doped material where the mobility is lower than the mobility of the intrinsic semiconductor. The optimal solution could be to design a device where the channel is formed in an intrinsic material. Of course this is not straightforward because the concentration of electrons in the intrinsic material is rather small. In GaAs the concentration of intrinsic electrons is about four orders of magnitude smaller than in the silicon.

The solution to the problem is achieved using a heterojunction between the metal and the intrinsic substrate. As an example let us consider a device where the gate structure is formed by a sequence of metal-n-type $Al_xGa_{1-x}As$-GaAs. The band diagram of $Al_xGa_{1-x}As$ is modulated by the percentage of aluminium. In practice: $E_gap = 1.42 + 1.24 \cdot x$ and $q\chi = 4.07 - 1.1 \cdot x$ for $x \lesssim 0.5$. Where 1.42 eV and 4.07 eV are gallium arsenide energy gap and affinity respectively.
For instance at $x = 0.3$ the energy gap is $E_{gap} = 1.80eV$ and the affinity is $q\chi = 3.74eV$. The method to study heterojunctions have been discussed in chapter 7. In figure 21 an example of a band diagrams is shown. The metal is chosen in order to form a Schottky junction with the AlGaAs layer. Furthermore the AlGaAs layer is sufficiently thin to be completely depleted by the electrons transfer necessary to equilibrate the whole structure.

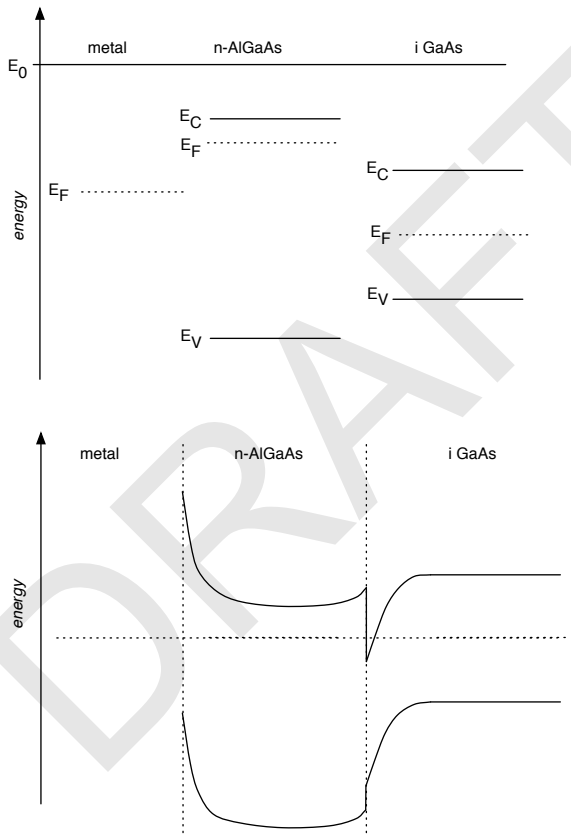

Fig. 10.25. *Band diagram before (above) and after (below) the equilibrium in metal - n-type AlGaAs and intrinsic GaAs structure. Since the Fermi level of the n-type AlGaAs lies above the conduction band of the intrinsic GaAs, at the equilibrium the conduction band of the GaAs lies below the Fermi level.*

The materials are chosen in order to have the Fermi level of the AlGaAs layer above the conduction band of the intrinsic GaAs. Applying the rules for the equilibrium band diagram drawing

outlined in chapters 1 and 7 the bottom of the spike at the interface lies below the Fermi level in a very narrow region close to the interface with the AlGaAs.
In this region, the GaAs becomes a degenerated semiconductor with a very large concentration of electrons. In a such narrow region the electrons form a bi-dimensional gas of free electrons, and this region is called 2-DEG.
The 2-DEG is a strip of high concentration of electrons connecting the drain and the source and embedded in the intrinsic GaAs. Then the electrons of the 2-DEG move with very high mobility. With this structure it is possible then to achieve large currents, very short transit time, and then very high operative frequencies. Such a device is called a High Electrons Mobility Transistor (HEMT).
The AlGaAs layer is typically highly doped and very narrow with the scope to obtain at the equilibrium a total depletion of the N-type region. The depletion of the AlgaAs is necessary in order to avoid that the 2DEG is shunted by a parallel conductor. In this condition, the AlGaAs behaves like an insulator between the metal and the 2-DEG. As a consequence the voltage applied to the metal can modulate the the 2-DEG depth greatly changing the concentration of electrons confined in the 2-DEG.
The 2-DEG electrons are hindered to move towards the metal by a large barrier, then even if their concentration is large, the current towards the metal is negligible and actually smaller than in the MESFET. However this leakage current, virtually null in the MOSFET, deteriorates the device performance.
Like the MESFET, the HEMT works in a depletion mode, the 2-DEG usually exists at $V_G = 0$. The application of a V_{DS} between the drain and the source results in a narrowing of the channel at the drain contact and in a consequent saturation of the current. Then the I_d vs. V_{DS} characteristics is of the same kind of those found in the other field effect devices.
The mobility in the 2-DEG is large but anyway is less than that found in the bulk of the intrinsic material. To this regard, we have to consider that the dimensions of the 2-DEG requires a quantum description of the electrons, and the solution of the Schrodinger equation results in a wave function that is partially evanescing into the AlGaAs region. This means that the mobility of the AlGaAs influences the mobility of the electrons in the 2-DEG. The mobility of the AlGaAs is reduced by the high doping. To avoid the influence of doping in the mobility a narrow layer of undoped AlGaAs, called spacer, is created between the GaAs and the N-AlGaAs. Furthermore, the strain due to the lattice mismatch between GaAs and AlGaAs contribute to deteriorate the performance of the device. Some technological issues have been developed to improve the lattice matching, such as the so-called pseudomorphic HEMT.
The leakage current can also be further increased reducing the gate contact without affecting the channel length. For the scope the gate contact is shaped as a mushroom.
Finally, it is important to mention that superior performance in terms of speed of operation is obtained when materials with larger mobility, such as the InGaAs, replace the GaAs substrate.

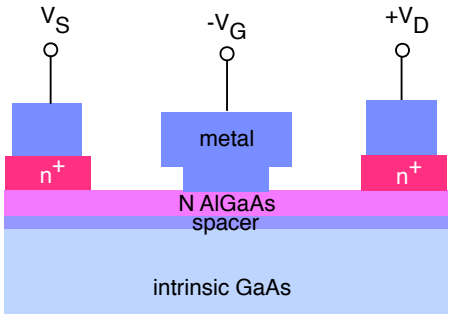

Fig. 10.26. *Schematic structure of a HEMT. Typical thickness of the various layers are of the following order of magnitude: intrinsic GaAs: 1000 μm; intrinsic AlGaAs (spacer): 5 nm; N-AlGaAs: 50 nm. The N-AlGaAs doping is of the order of* $10^{18} cm^{-3}$