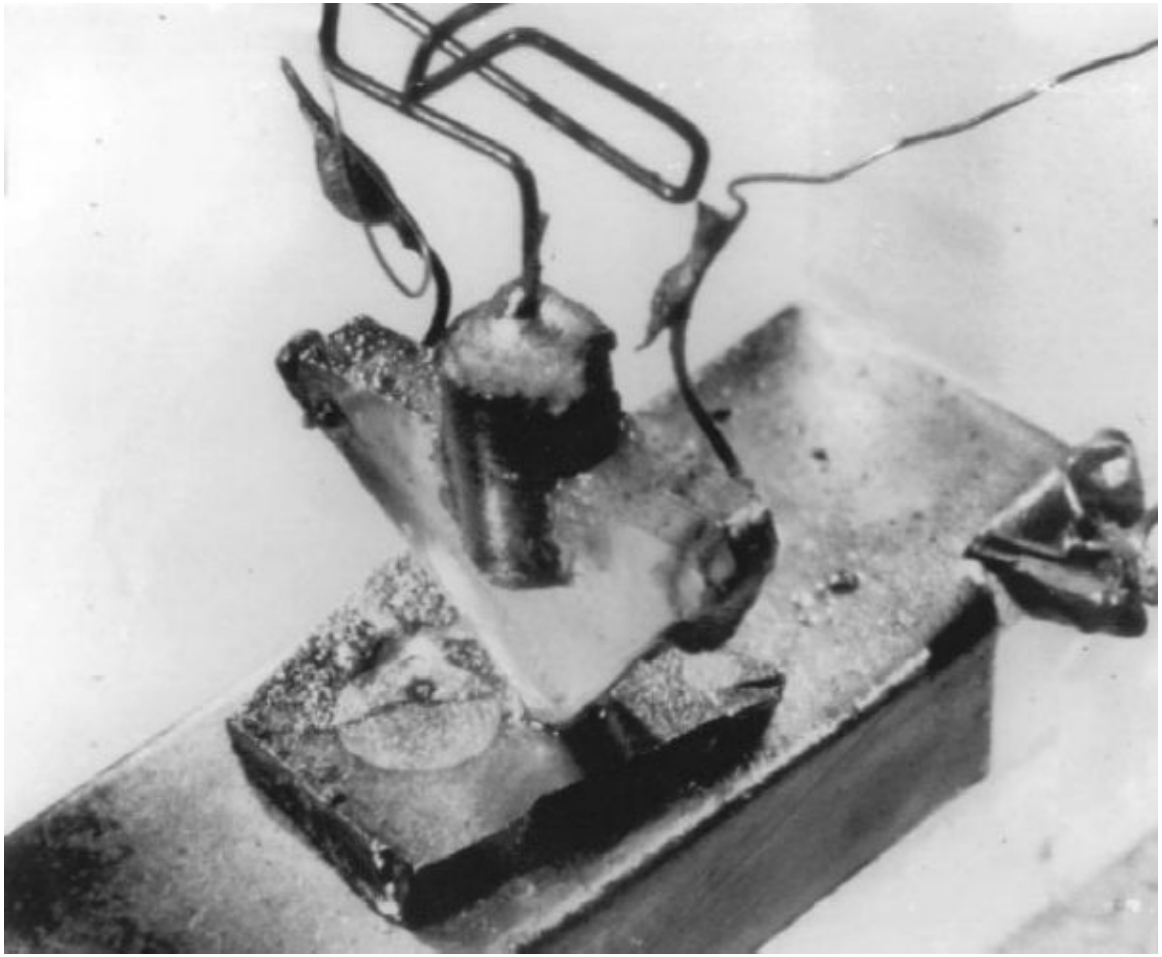
Metal-Insulator-Semiconductor and Metal-Insulator-Metal Structures. Part IV. Metal-Oxide-Semiconductor Structures

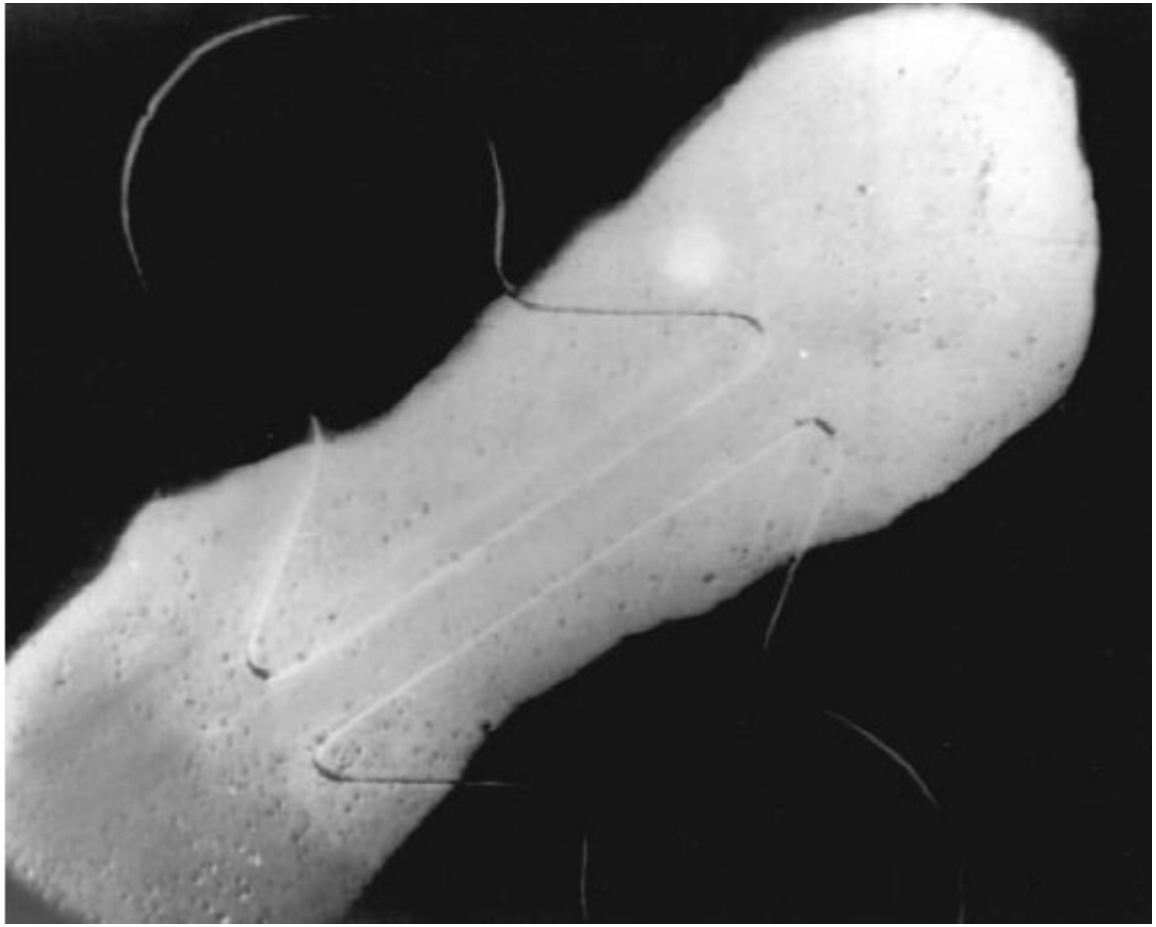
Alexander Gabovich, **KPI**,

Lecture 11

The first transistor



The first MOS field-effect transistor



Metal-oxide-semiconductor field-effect transistor (MOS or MOSFET)

The MOS transistor, also called MOSFET (**M**etal-**O**xide-Semiconductor **F**ield-**E**ffect **T**ransistor) or IGFET (**I**nsulated-**G**ate **F**ield-**E**ffect **T**ransistor) is the most widely used semiconductor device and is at the heart of every digital circuit. Without the MOSFET there would be no computer industry, no digital telecommunication systems, no video games, no pocket calculators and no digital wristwatches. MOS transistors are also increasingly used in analog applications such as switched-capacitor circuits, analog-to-digital converters, and filters.

The exponential progress of MOS technology is best illustrated by the evolution of the number of MOS transistors integrated in a single memory chip or single microprocessor, as a function of calendar year. Each memory cell of a dynamic random-access memory (DRAM) contains a MOS transistor and a capacitor. It can be observed from Figure 7.1 that there is a four-fold increase in the number of transistors in a DRAM every three years. This exponential growth of integration density with time is known as Moore's law.[1]

The integration density of memory circuits is about 5 to 10 times higher than that of logic circuits such as microprocessors because of the more repetitive layout of transistors in memory chips. The increase in integration density is essentially due to the reduction of transistor size. The first experimental 1-gigabit DRAMs were reported in 1995 [2] where 1-gigabit DRAM contains over a billion MOSFETs. About 400 of these chips can be fabricated on a single silicon wafer, 40 centimeters in diameter. Such a wafer, therefore, contains over 400,000,000,000 transistors. This number is equal to the number of stars in our galaxy... More MOSFETs have been fabricated during the last ten years than grains of rice have been harvested by humans since the dawn of mankind.

Metal-oxide-semiconductor field-effect transistor (MOS or MOSFET)

The first description of a device called IGFET dates back to the 1930's in patents by Lilienfeld and Heil.[3,4] Because of technological limitations the IGFET could not be successfully fabricated at that time. The first working MOS transistor was realized in 1960 by Kahng and Attala.[5] A few years later, the integrated circuit industry took off to reach incredible proportions and has become one of the leading industries worldwide.

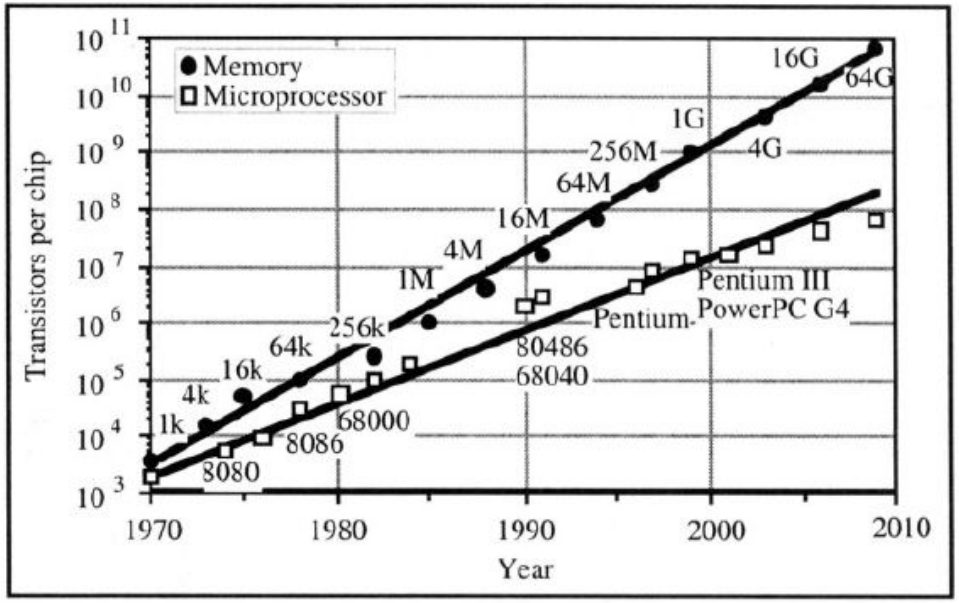


Figure 7.1: Actual and predicted evolution of circuit complexity in DRAMs and microprocessors.

Metal-oxide-semiconductor field-effect transistor (MOS or MOSFET)

There are two types of MOS transistors: the n-channel MOSFET, in which current flow is due to electron transport, and the p-channel MOSFET in which holes are responsible for current flow. A circuit containing only n-channel devices is produced by an nMOS process. Similarly, a pMOS process fabricates circuits that contain only p-channel transistors. Today the most commonly used technology is CMOS (Complementary MOS) in which both n-channel and p-channel transistors are fabricated. Here we will limit our analysis to n-channel devices. The current-voltage expressions describing a p-channel device can readily be derived from the n-channel equations, provided the appropriate changes of sign are made.

An n-channel MOS transistor is fabricated in a P-type semiconductor substrate, usually silicon. Two N-type diffusions are made in the substrate and the current flow will take place between these two diffusions. The diffusion with the lowest applied potential is called the "source" and the diffusion with the highest applied potential is called the "drain". Above the substrate, and between the source and the drain lies a thin insulating layer, usually silicon dioxide, and a metal electrode called "gate" (Figure 7.2). An electron-rich layer referred to as the "channel" can be created between the source and the drain underneath the gate insulator when a positive bias is applied to the gate. With appropriate voltages applied at the source and drain electrons can then flow from the source into the drain, through the channel. In a p-channel transistor an N-type substrate is used. The P-type drain is at a lower potential than the P-type source and the application of a negative bias to the gate enables the formation of a hole-enriched channel between source and drain. The metal-insulator-semiconductor structure is often referred to as a "MIS" structure, where the "I" stands for the insulator. When the insulator is an oxide, it is called a "MOS" structure.

Metal-oxide-semiconductor field-effect transistor (MOS or MOSFET)

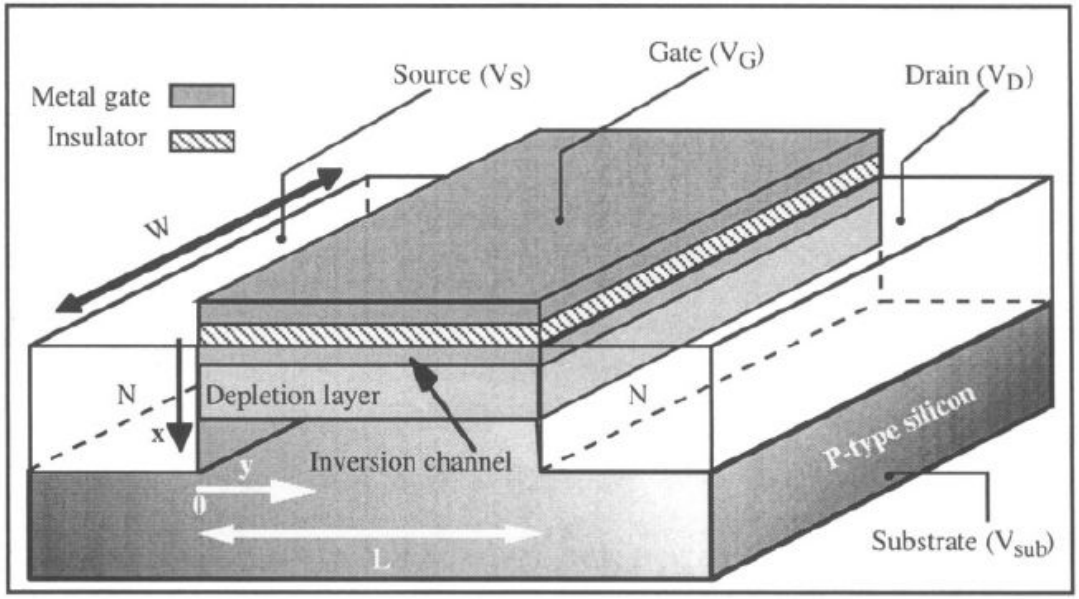


Figure 7.2: N-channel MOS transistor.

The basic operation of the n-channel MOSFET is the following. We will first consider the case where the gate voltage is equal to zero while the P-type substrate and the source are grounded ($V_{sub} = V_S = 0$). The drain is connected to a positive voltage source (V_D=5 volts, for instance). Since the source and the substrate are at the same potential there is no current flow in the source-substrate junction. The drain-substrate junction is reverse biased and except for a small negligible reverse leakage current no current flows in that junction either. Under these conditions there is no channel formation, and therefore, no current flow from source to drain.

Metal-oxide-semiconductor field-effect transistor (MOS or MOSFET)

In the second case a constant positive bias is applied to the gate. There is no gate current since the metal electrode is dielectrically insulated from the silicon. Because it is positively biased the gate electrode does, however, attract electrons from the semiconductor, and a thin, electron-rich layer forms under the gate insulator. These electrons are supplied by the source and the drain which, being N-type, are large reservoirs of electrons. The electron-rich layer underneath the gate is called "channel". The N-type source and the N-type drain are connected by the electron-rich channel, and current is now free to flow between source and drain. The effect of the gate voltage controlling the concentration of electrons in the semiconductor through the gate oxide is called "field effect". The bias on the gate creates an electric field which can either induce or inhibit the formation of an electron-rich region at the surface of the semiconductor. The terms "source", "drain", "channel" and "gate" come to mind quite naturally since the electrons originate at the source, flow through the channel and are finally collected by the drain, the whole process being controlled by the bias on the gate.

Metal-oxide-semiconductor field-effect transistor (MOS or MOSFET)

The current in the channel, from source to drain can, to aa first approximation, be estimated using Ohm's law. Using $V=IR$ in a small channel element having a length dy and a width W we obtain:

$$dV(y) = I\, dR(y) \qquad (7.1.1)$$

The channel resistance as a function of y is obtained from Equation 2.3.3 where the electron concentration in the channel per unit area (unit: cm^{-2}) results from integrating the electron concentration per unit volume (unit: cm^{-3}) over the thickness of the device:

$$dR(y) = \frac{dy}{q\,\mu_n\, W \int_0^{\infty} n(x,y)\, dx} \qquad (7.1.2)$$

where x is the depth in the silicon ($x = 0$ at the silicon/SiO_2 interface). Note that the electron charge per unit area in the channel element can be written as:

$$Q_n(y) = q \int_0^{\infty} n(x,y)\, dx \quad (C\ cm^{-2}) \qquad (7.1.3)$$

The formation of a channel occurs when the gate voltage is positive and sufficiently high. In practice, the channel forms if the gate voltage is larger than a given value called the "threshold voltage", noted V_{TH}. Considering that the Metal-Oxide-Semiconductor structure forms a parallel-plate capacitor, we can write:

$$Q_n(y) = C_{ox}\,(V_G - V_{TH} - V(y)) \qquad (7.1.4)$$

where C_{ox} is the capacitance of the gate oxide per unit area and $V(y)$ is the local potential in the channel element, which varies from $V(y=0) = V_S = 0$ near the source to $V(y=L) = V_D$ near the drain.

Metal-oxide-semiconductor field-effect transistor (MOS or MOSFET)

Introducing Equations 7.1.2 and 7.1.4 into Expression 7.1.1 we obtain:

$$I\,dy = \mu_n W C_{ox} (V_G - V_{TH} - V(y))\,dV \qquad (7.1.5)$$

Since $V_S = 0V$ and since the current I is constant from source to drain, the integration of Equation 7.1.5 yields:

$$I \int_0^L dy = \mu_n W C_{ox} \int_0^{V_D} (V_G - V_{TH} - V(y))\,dV \qquad (7.1.6)$$

$$\Downarrow$$

$$I = \mu_n C_{ox} \frac{W}{L}\left((V_G - V_{TH})\,V_D - \frac{V_D^2}{2}\right) \qquad (7.1.7)$$

Metal-oxide-semiconductor field-effect transistor (MOS or MOSFET)

If the local potential between source and drain, $V(y)$, becomes equal to or larger than $V_G - V_{TH}$ the formation of a channel can locally no longer be supported near the drain and the channel exists only between $y=0$ and a location y where $V(y) = V_G - V_{TH}$. In practice, that location is very close to L, and the current is obtained by replacing V_D by $V_G - V_{TH}$ in Expression 7.1.7. The current is then called the "saturation current" and noted I_{sat}. Saturation takes place when $V_D \geq V_G - V_{TH}$, and replacing V_D by $V_G - V_{TH}$ in Equation 7.1.7 we obtain:

$$I_{sat} = \mu_n C_{ox} \frac{W}{L} \frac{(V_G - V_{TH})^2}{2} \qquad (7.1.8)$$

Note that the current in saturation is no longer a function of the drain voltage and that the potential drop in the y-direction in the channel is fixed at a value equal to $V_G - V_{TH}$ in saturation.

In a p-channel MOSFET the source is at the highest potential and supplies holes to the channel. The holes are finally collected by the drain, which is at a lower potential than the source. In this case a negative bias relative to the substrate must be applied to the gate to create a hole-rich p-type channel.

A study of the metal-insulator-semiconductor structure, called the "MOS capacitor", will aid in the understanding of the detailed operation of the MOS transistor.

MOS capacitor

The MOS capacitor is comprised of a metal gate, an insulating oxide layer, and a semiconductor. The thickness of the oxide typically varies between 5 to 50 nanometers. The semiconductor chosen for the example of Figure 7.3 is P-type silicon, which corresponds to the substrate of an n-channel device (nMOS).

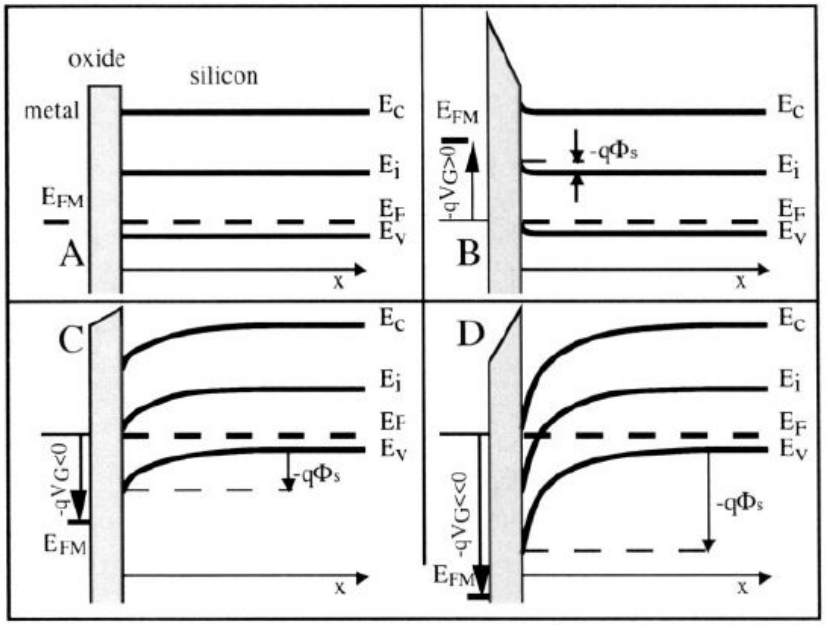


Figure 7.3: Energy bands in the MOS structure. [6] A: Flat energy bands, B: Accumulation, C: Depletion, D: Inversion

We will first consider the case of an hypothetical metal that has the same Fermi level as the silicon. When the structure is fabricated the Fermi level of the system is unique, and since the metal has the same Fermi level as the silicon, the band structure is that shown in Figure 7.3A. This condition is referred to as flat band for obvious reasons.

MOS capacitor (Accumulation)

If a negative bias is applied to the metal gate while the silicon substrate is grounded the structure behaves like a parallel-plate capacitor where the two electrodes are the silicon and the metal, and the oxide is the insulator between them. The application of the bias gives rise to a negative charge on the gate. This is a surface charge in the metal, located at the metal-oxide interface. An equal charge of opposite sign appears at the surface of the silicon, at the silicon-oxide interface (Figure 7.3B). The charge in the silicon can also be considered a surface charge, as we will demonstrate next. Its thickness is approximately 10 nanometers. This thin, hole-rich layer is called an accumulation layer. The capacitance of the MOS structure in accumulation is that of a parallel-plate capacitor between the metal gate and the accumulation layer. Its value (in Farads per unit area) is equal to:

$$C = \frac{\varepsilon_{ox}}{t_{ox}} \equiv C_{ox} \quad \text{(unit: F cm}^{-2}\text{)} \qquad (7.2.1)$$

where ε_{ox} is the permittivity of silicon dioxide and t_{ox} is the thickness of the gate oxide. C_{ox} is called the gate oxide capacitance. The permittivity of SiO_2, ε_{ox}, is equal to $\kappa_{SiO2} \times \varepsilon_o$ where ε_o is the permittivity of vacuum, equal to 8.854×10^{-14} F/cm, and κ_{SiO2} is the dielectric constant of SiO_2, equal to 3.9.

MOS capacitor (Accumulation)

Thickness of the accumulation layer

A derivation of the accumulation layer thickness as a function of substrate doping concentration will show that the layer is very small and hence can be considered as a surface charge.[7] The distribution of the charge as a function of depth, x, can be found using Poisson's equation:

$$\frac{d^2\Phi(x)}{dx^2} = -\frac{\rho}{\varepsilon_{si}} = -\frac{q}{\varepsilon_{si}}\,(p - n + N_d - N_a) \qquad (7.2.2)$$

with:

$$p(x) = p_{po}\,exp\left(-\frac{q\Phi(x)}{kT}\right) = N_a\,exp\left(-\frac{q\Phi(x)}{kT}\right) \qquad (7.2.3)$$

and:

$$n(x) = n_{po}\,exp\left(\frac{q\Phi(x)}{kT}\right) = \frac{n_i^2}{N_a}\,exp\left(\frac{q\Phi(x)}{kT}\right) \qquad (7.2.4)$$

where p_{po} is the equilibrium hole concentration in the P-type material, n_{po} is the equilibrium electron concentration in the same material, and $\Phi(x)$ is the potential in the silicon as a function of depth. Far from the surface of the silicon the potential is equal to zero: $\Phi(x=\infty)=0$, which will be used as a boundary condition for Equation 7.2.2.

MOS capacitor (Accumulation)

In the hole accumulation layer formed in P-type material one can assume that $n<<p$ and that $N_d<<N_a$, thus Equation 7.2.2 can be rewritten as:

$$\frac{d^2\Phi(x)}{dx^2} = -\frac{\rho}{\varepsilon_{si}} = -\frac{q}{\varepsilon_{si}} N_a \left[exp\left(-\frac{q\Phi(x)}{kT}\right) - 1 \right] \qquad (7.2.5)$$

where ε_{si}, the permittivity of silicon is equal to $\kappa_{Si} \times \varepsilon_o$ where κ_{Si} is the dielectric constant of silicon ($\kappa_{Si} = 11.7$). In the accumulation layer the hole concentration is greater than the hole concentration due to doping concentration, and therefore, $p >> N_a$ in the accumulation layer. The following approximation can thus be used:

$$\frac{d^2\Phi(x)}{dx^2} \cong -\frac{qN_a}{\varepsilon_{si}} exp\left(-\frac{q\Phi(x)}{kT}\right) \qquad (7.2.6)$$

To integrate this equation we must first multiply both terms of the equation by $2\frac{d\Phi(x)}{dx}$, which yields:

$$2\frac{d\Phi(x)}{dx}\frac{d^2\Phi}{dx^2} = -2\frac{qN_a}{\varepsilon_{si}} exp\left(-\frac{q\Phi(x)}{kT}\right)\frac{d\Phi(x)}{dx}$$

or:

$$\frac{d}{dx}\left(\frac{d\Phi}{dx}\right)^2 = 2\frac{N_a kT}{\varepsilon_{si}}\frac{d}{dx}\left(exp\left(-\frac{q\Phi(x)}{kT}\right)\right) \qquad (7.2.7)$$

which can be rewritten:

$$\frac{d\mathcal{E}^2}{dx} = 2\frac{N_a kT}{\varepsilon_{si}}\frac{d}{dx}\left(exp\left(-\frac{q\Phi(x)}{kT}\right)\right) \qquad (7.2.8)$$

MOS capacitor (Accumulation)

Integrating from x to x_{acc}, where x_{acc} is the thickness of the accumulation layer, and noting that $\mathcal{E}(x=x_{acc}) = \Phi(x=x_{acc}) = 0$, since the silicon underneath the accumulation layer is neutral, one obtains:

$$\mathcal{E}^2(x) - 0 = \frac{2kTN_a}{\varepsilon_{si}}\left[\exp\left(-\frac{q\Phi(x)}{kT}\right) - 1\right]$$

$$= \frac{2}{L_D^2}\left(\frac{kT}{q}\right)^2\left[\exp\left(-\frac{q\Phi(x)}{kT}\right) - 1\right] \qquad (7.2.9)$$

with:

$$L_D = \sqrt{\frac{\varepsilon_{si}kT}{q^2N_a}} \qquad (7.2.10)$$

L_D is called the "Debye length". For example, L_D has a value of 40, 18 and 13 nanometers for doping impurity concentrations of 10^{16}, 5×10^{16} and 10^{17} cm^{-3}, respectively. Noting that $d\Phi(x)/dx > 0$, Equation 7.2.9 can be rewritten as follows:

$$\frac{d\Phi(x)}{\frac{kT}{q}\sqrt{\exp\left(-\frac{q\Phi(x)}{kT}\right) - 1}} = \frac{\sqrt{2}\,dx}{L_D} \qquad (7.2.11)$$

MOS capacitor (Accumulation)

The latter expression can be integrated using the following boundary conditions: $\Phi(x=x_{acc}) = 0$ and $\Phi(x=0) = \Phi s$ where Φs is the potential at the semiconductor surface and is called the "surface potential". Equation 7.2.11 can be rewritten as:

$$\frac{d\left(-\frac{q\Phi(x)}{2kT}\right)}{\sqrt{\exp\left(-\frac{q\Phi(x)}{kT}\right) - 1}} = -\frac{dx}{\sqrt{2}\,L_D} \qquad (7.2.12)$$

Numerator and denominator of the left-hand term are then multiplied by $\exp\left(\frac{-q\Phi(x)}{2kT}\right)$:

$$\frac{\exp\left(\frac{-q\Phi(x)}{2kT}\right) d\left(-\frac{q\Phi(x)}{2kT}\right)}{\exp\left(\frac{-q\Phi(x)}{2kT}\right)\sqrt{\exp\left(-\frac{q\Phi(x)}{kT}\right) - 1}} = -\frac{dx}{\sqrt{2}\,L_D} \qquad (7.2.13)$$

Changing variables and writing $u = \exp\left(\frac{-q\Phi(x)}{2kT}\right)$, one obtains:

$$\frac{du}{u\sqrt{u^2 - 1}} = -\frac{dx}{\sqrt{2}\,L_D} \qquad (7.2.14)$$

MOS capacitor (Accumulation)

Posing $u = \frac{1}{\cos(\theta)} = \sec(\theta)$, the latter equation becomes:

$$\frac{du}{u\sqrt{u^2 - 1}} = d\theta = -\frac{dx}{\sqrt{2}\,L_D} \Rightarrow \theta = C - \frac{x}{\sqrt{2}\,L_D} \qquad (7.2.15)$$

where C is an integration constant. We can conclude that:

$$\exp\left(\frac{-q\Phi(x)}{2kT}\right) = u = \sec\left(C - \frac{x}{\sqrt{2}\,L_D}\right) \qquad (7.2.16)$$

and, therefore:

$$\Phi(x) = -\frac{kT}{q} \ln\left\{\sec^2\left[C - \frac{x}{\sqrt{2}\,L_D}\right]\right\} \qquad (7.2.17)$$

The integration constant, C, can be related to the surface potential, Φ_S, by the following relationship:

$$\Phi(x=0) = \Phi_S \Rightarrow C = \cos^{-1}\left(\exp\left(\frac{q\Phi_S}{2kT}\right)\right)$$

Finally we find that the thickness of the accumulation layer, x_{acc}, can be found using the condition that $\Phi(x=x_{acc}) = 0$:

$$x_{acc} = \sqrt{2}\,L_D \cos^{-1}\left(\exp\left(\frac{q\Phi_S}{2kT}\right)\right)$$

MOS capacitor (Accumulation)

The thickness of the accumulation layer, x_{acc}, can thus vary between 0 and $\frac{\sqrt{2}}{2}\pi L_D$, depending on the accumulation charge (Figure 7.4).

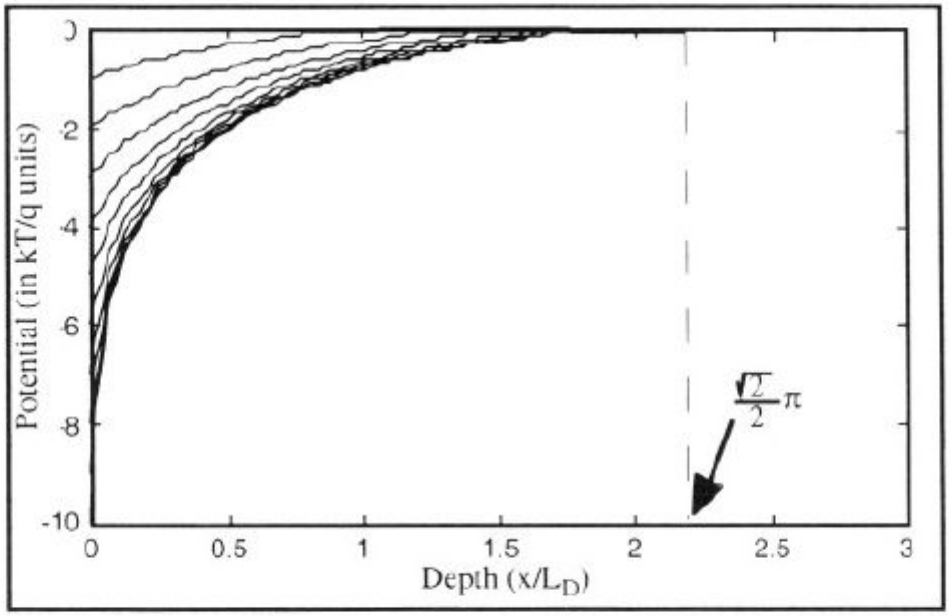


Figure 7.4: Potential (normalized to kT/q) in an accumulation layer (holes in P-type silicon) as a function of depth (normalized to L_D) for different values of surface potential.

MOS capacitor (Accumulation)

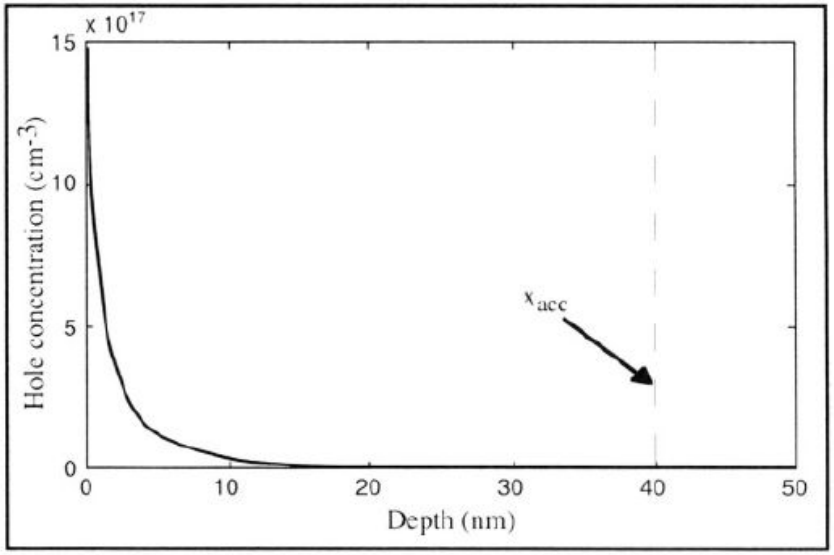


Figure 7.5: Hole concentration profile in an accumulation layer. The substrate doping concentration, N_a, is equal to 10^{16} cm^{-3} and the surface potential, Φ_S, is equal to -5kT/q

The hole concentration is an exponential function of the potential. Therefore, the charge density increases very rapidly close to the surface and most of the accumulation charge is concentrated within a depth much smaller than x_{acc} (Figure 7.5). Therefore, the charge in the accumulation layer can be considered as a surface charge. One can also consider that the surface potential, Φ_S, is very small. It actually is slightly negative and in practice reaches only a few *-kT/q* (*kT/q* is equal to 25.9 mV at room temperature).

MOS capacitor (Accumulation)

The application of a negative bias V_G on the gate gives rise to a negative surface charge Q_G in the metal at the metal-oxide interface. The accumulation charge in the semiconductor, Q_{acc}, is equal to Q_G, with opposite sign ($Q_{acc} = -Q_G$). Integrating Poisson's equation (Expression 7.2.2) from $x = 0$ to $x = +\infty$ we obtain:

$$Q_{acc} = \int_0^{\infty} \rho(x)\, dx = = q \int_0^{\infty} (p - n + N_d - N_a)\, dx = -\varepsilon_{si}\, \mathcal{E}(0)$$

Within the accumulation layer ($0 < x < x_{acc}$) we assumed that $n<<p$, $N_d<<N_a$ and $p>>N_a$ while the silicon underneath the accumulation layer is neutral ($\rho = 0$ for $x > x_{acc}$). The charge in the semiconductor is, therefore, equal to the accumulation charge Q_{acc}. Using Equation 7.2.9 evaluated for $\mathcal{E}(x=0)$ the expression for the accumulation charge is:

$$Q_{acc} = -\varepsilon_{si} \frac{\sqrt{2}}{L_D} \frac{kT}{q} \sqrt{exp\left(-\frac{q\Phi(x=0)}{kT}\right) - 1}$$

MOS capacitor (Accumulation)

The exact value of the surface potential $\Phi_S \equiv \Phi(x=0)$ is related to the applied gate voltage V_G in the following way. V_G is equal to the potential drop across the oxide, V_{ox}, added to the potential drop Φ_S within the semiconductor:

$$V_G = \Phi_S + V_{ox} = \Phi_S - \frac{Q_{acc}}{C_{ox}}$$

or

$$V_G = \Phi_S - \frac{\varepsilon_{si}}{C_{ox}} \frac{\sqrt{2}}{L_D} \frac{kT}{q} \sqrt{exp\left(-\frac{q\Phi_S}{kT}\right) - 1} \qquad (7.2.18)$$

The magnitude of the surface potential, Φ_S, is very small (only a few $\frac{-kT}{q}$), even for large applied negative gate voltage values. Since the accumulation charge Q_{acc} has a negligible thickness it can be considered as a surface charge and the approximation previously given for the capacitance of the MOS structure:

$$C = C_{ox} \equiv \frac{\varepsilon_{ox}}{t_{ox}} \qquad (7.2.19)$$

holds for the MOS structure in accumulation.

MOS capacitor (Depletion)

If a small positive bias is applied to the gate (Figure 7.3C) holes near the silicon surface are repelled by the gate. Because the acceptor doping atoms cannot move in the silicon lattice a negative charge appears underneath the gate oxide. Similarly a positive charge of equal magnitude can be found in the gate electrode, at the metal-oxide interface. The gate charge is a surface charge, but the charge in the silicon is not. It is a depletion charge which extends to a non-negligible depth into the silicon. The potential in the depletion region can be found integrating by Poisson's equation. Using $n<<p$ and $N_d<<N_a$ one can write:

$$\frac{d^2\Phi(x)}{dx^2} = -\frac{\rho}{\varepsilon_{si}} = -\frac{q}{\varepsilon_{si}}(p - N_a) = -\frac{q}{\varepsilon_{si}} N_a \left[exp\left(-\frac{q\Phi(x)}{kT}\right) - 1 \right] \quad (7.2.20)$$

The potential in the depletion region near the oxide/silicon interface is positive. Therefore, the exponent term of Equation 7.2.20 is small and can be neglected, which implies $p<<N_a$:

$$\Phi(x)>0 \Rightarrow exp\left(-\frac{q\Phi(x)}{kT}\right) \cong 0 \quad (7.2.21)$$

Using this approximation Equation 7.2.20 becomes:

$$\frac{d^2\Phi(x)}{dx^2} = -\frac{\rho}{\varepsilon_{si}} = \frac{qN_a}{\varepsilon_{si}} \quad (7.2.22)$$

MOS capacitor (Depletion)

This result is the *depletion approximation* which assumes that the charge density is constant and equal to $-qN_a$ in the depletion region. The depth up to which holes are repelled is called the depletion depth (or width) and noted x_d. Outside the depletion region the silicon is assumed to be neutral, such that $\rho(x)$, $\mathcal{E}(x)$ and $\Phi(x)$ are equal to zero for $x \geq x_d$. The potential in the silicon can be found by integrating the Poisson equation 7.2.22 with the following boundary conditions:

$$\Phi(x_d) = 0 \quad \text{and} \quad \frac{d\Phi(x_d)}{dx} = 0 \tag{7.2.23}$$

which yields:

$$\Phi(x) = \frac{qN_a}{2\varepsilon_{si}} (x - x_d)^2 \tag{7.2.24}$$

The surface potential at the oxide/silicon interface where $x=0$ is equal to:

$$\Phi_S = \Phi(x=0) = \frac{qN_a}{2\varepsilon_{si}} x_d^2 \tag{7.2.25}$$

Equation 7.2.25 can be used to evaluate the depletion depth expressed as a function of the surface potential:

$$x_d = \sqrt{\frac{2\varepsilon_{si}\Phi_s}{qN_a}} \tag{7.2.26}$$

MOS capacitor (Depletion)

The charge per surface area in the region from $x = 0$ to $x = x_d$, called "depletion charge" is equal to:

$$Q_d = - q N_a x_d = - \sqrt{2q\varepsilon_{si}N_a\Phi_s} \qquad (7.2.27)$$

The gate voltage, V_G, is equal to the potential drop across the oxide added to the potential variation in the semiconductor:

$$V_G = \Phi_s + V_{ox} = \Phi_s - \frac{Q_d}{C_{ox}} \qquad (7.2.28)$$

The capacitance of the structure can be calculated as follows:

$$C = \frac{dQ_G}{dV_G} = - \frac{dQ_d}{dV_G} = - \frac{dQ_d}{d(-\frac{Q_d}{C_{ox}} + \Phi_s)} = - \frac{dQ_d/d\Phi_s}{d(-\frac{Q_d}{C_{ox}} + \Phi_s)/d\Phi_s} = \frac{1}{\frac{1}{C_{ox}} + \frac{1}{C_D}} \qquad (7.2.29)$$

where

$$C_D = - \frac{dQ_d}{d\Phi_S} = \frac{\varepsilon_{si}}{x_d} \qquad (7.2.30)$$

The overall capacitance is thus the series association of the gate oxide capacitance and the depletion region capacitance, ε_{si}/x_d. The capacitance can also be expressed as a function of the gate voltage by rewriting expression 7.2.28 in the following way:

$$V_G = - \frac{Q_d}{C_{ox}} + \Phi_s = \frac{qN_a x_d}{C_{ox}} + \frac{qN_a}{2\varepsilon_{si}} x_d^2 \qquad (7.2.31)$$

x_d can be expressed as a function of the gate voltage:

$$x_d = - \frac{\varepsilon_{si}}{C_{ox}} + \sqrt{\left(\frac{\varepsilon_{si}}{C_{ox}}\right)^2 + \frac{2\varepsilon_{si}}{qN_a} V_G} \qquad (7.2.32)$$

MOS capacitor (Depletion, Inversion)

Substituting x_d into Equation 7.2.29 we obtain the capacitance as a function of the gate voltage:

$$C = \frac{C_{ox}}{\sqrt{1 + \frac{2C_{ox}^2 V_G}{qN_a\varepsilon_{si}}}} \qquad (7.2.33)$$

Inversion

If a larger positive voltage is applied to the gate the surface potential will continue to increase. The hole concentration near the surface decreases while the electron concentration increases, according to the following relationships:

$$p(x=0) = N_a \exp\left(-\frac{q\Phi_S}{kT}\right) \qquad (7.2.34)$$

and:

$$n(x=0) = \frac{n_i^2}{N_a} \exp\left(\frac{q\Phi_S}{kT}\right) \qquad (7.2.35)$$

Since $n = n_i \exp\left(\frac{E_F - E_i}{kT}\right)$, $p = n_i \exp\left(\frac{E_i - E_F}{kT}\right)$ and $pn = n_i^2$, the electron surface concentration is equal to the hole surface concentration ($n(0) = p(0) = n_i$) when E_i coincides with E_F at x=0. This happens when $\Phi_S = \Phi_F = \frac{kT}{q} \ln\left(\frac{N_a}{n_i}\right)$ (Figure 7.6).

MOS capacitor (Inversion)

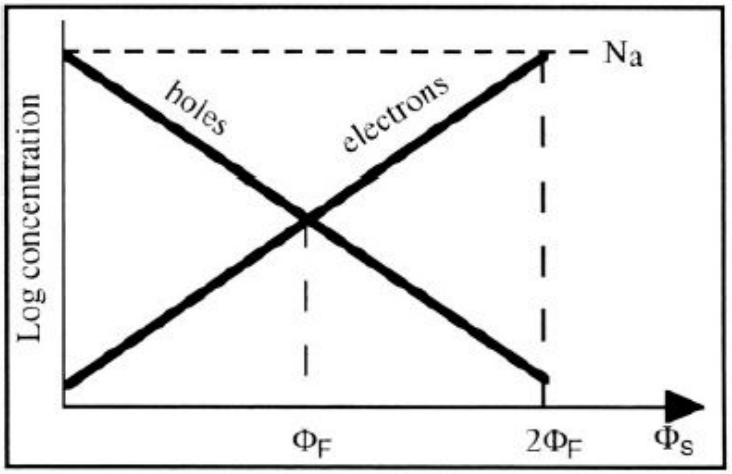


Figure 7.6: Hole and electron surface concentration as a function of Φ_S.

If the gate voltage is increased further the electron surface concentration increases up to a point where $n(x=0)$ becomes equal to $p_{po} = N_a$, which is the original hole concentration in the substrate. This happens when the band curvature at the surface ($x=0$) places E_i at an energy $q\Phi_F$ below E_F. In other words the band curvature is equal to $2(E_i - E_F)$ or:

$$\Phi_S = 2\,\Phi_F \qquad (7.2.36)$$

When this condition is met, the semiconductor surface is said to be in "strong inversion". For $\Phi_F \leq \Phi_S < 2\,\Phi_F$ the electron concentration is

larger than the hole concentration, and the surface is in weak inversion, while for $\Phi_S \geq 2\,\Phi_F$ it is in strong inversion (Figure 7.7).

MOS capacitor (Inversion)

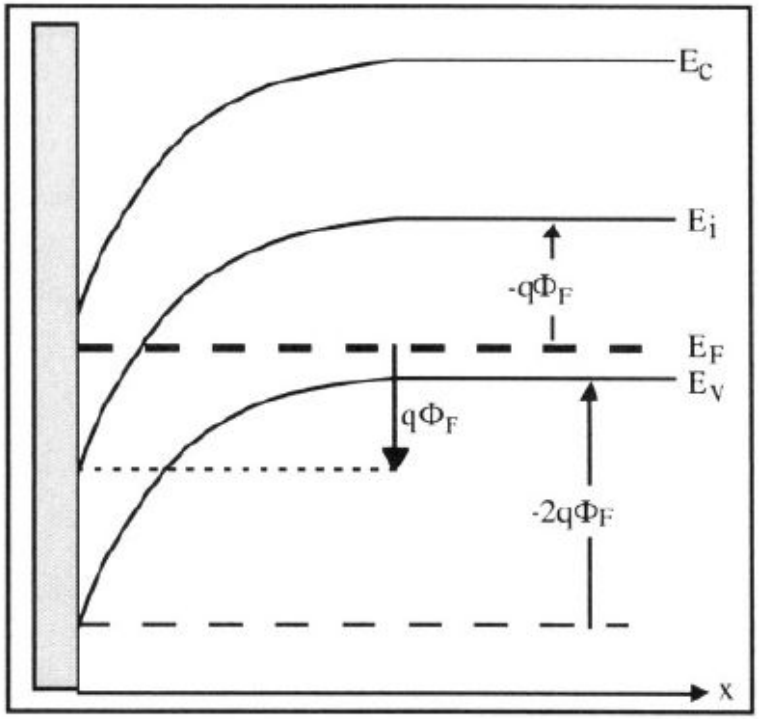


Figure 7.7: Energy band curvature under strong inversion at the surface.

The inversion layer is rich in electrons, and therefore, a good conductor. The MOS capacitor consists of two conducting electrodes (the metal gate and the inversion layer at the silicon surface). As in the case of accumulation, the capacitance of the MOS structure is once again equal to C_{ox}.

MOS capacitor (Inversion)

When an inversion layer is formed electrons are locally majority carriers at the surface. Any subsequent increase in gate voltage increases the electron concentration in the inversion layer, and a larger inversion charge, Q_{inv}, is produced. However, the thickness of the inversion layer remains very small. Its actual thickness is similar to that of an accumulation layer (derived in Section 7.2.1). The electron charge in an inversion layer can, therefore, be considered as a surface charge. As in the case of an accumulation layer the inversion charge depends exponentially on the surface potential ($Q_{inv} \propto exp\left(\frac{q\Phi_S}{kT}\right)$). When the gate voltage is increased beyond inversion formation the surface potential, Φ_S, increases only very slightly above $2\Phi_F$ and for all practical purposes one can assume that $\Phi_S = 2\Phi_F$ when an inversion layer is present, regardless of the gate voltage. Therefore, the depth of the depletion region is given by Equation 7.2.26 where $\Phi_S = 2\Phi_F$:

$$x_{dmax} = \sqrt{\frac{4\varepsilon_{si}\Phi_F}{qN_a}} \qquad (7.2.37)$$

MOS capacitor (Inversion)

Since the semiconductor is P-type one may wonder where the electrons in the inversion layer come from. They are produced by thermal generation, which is a rather slow process at room temperature. They can also be produced by external generation (if a light source is present, for example). If the semiconductor is in the dark and at cryogenic temperature the inversion layer may never form.

Figure 7.8 shows the capacitance of an MOS capacitor as a function of the applied gate bias. Such a curve is often called a capacitance curve, or C-V curve. Different types of measurements can be made, each of these probing a different aspect of the device properties.

In a first measurement the gate voltage is slowly ramped from negative to positive values, and a small ac signal is superimposed to this quasi-dc bias. The small signal is used to measure the value of the capacitance at the various dc gate biases. Different curves can be obtained for a given device depending on the frequency of the ac signal.

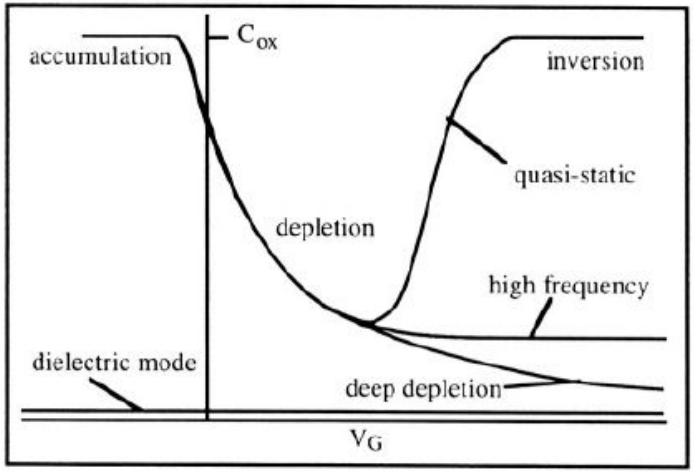


Figure 7.8: C-V curves of a MOS capacitor on a P-type substrate. [9]

MOS capacitor (Inversion)

In summary the following rules will be used to describe the relationships between the charge on the metal gate and the charge in the accumulation, depletion and inversion layers (Figure 7.9):

$$-Q_G = Q_{acc} \quad \text{(accumulation)} \quad (7.2.38a)$$

$$-Q_G = Q_d \quad \text{(depletion)} \quad (7.2.38b)$$

$$-Q_G = Q_d + Q_{inv} \quad \text{(inversion)} \quad (7.2.38c)$$

$-Q_G$ = charge at the backside contact of the sample (dielectric mode) (7.2.38d)

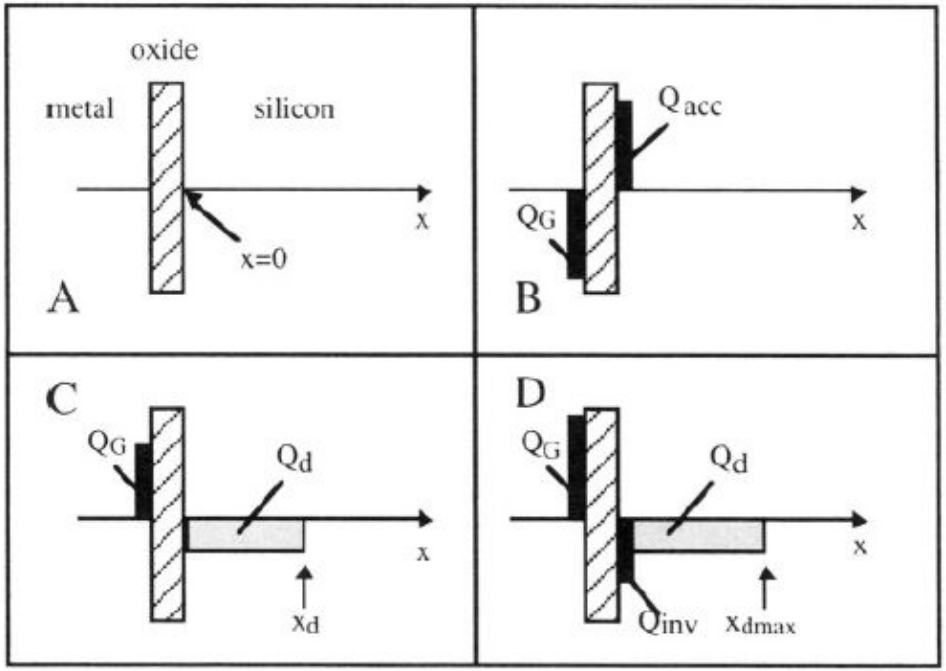


Figure 7.9: Charges in the MOS structure:
A: Flat band, B: Accumulation, C: Depletion, D: Inversion

MOS capacitor (Inversion)

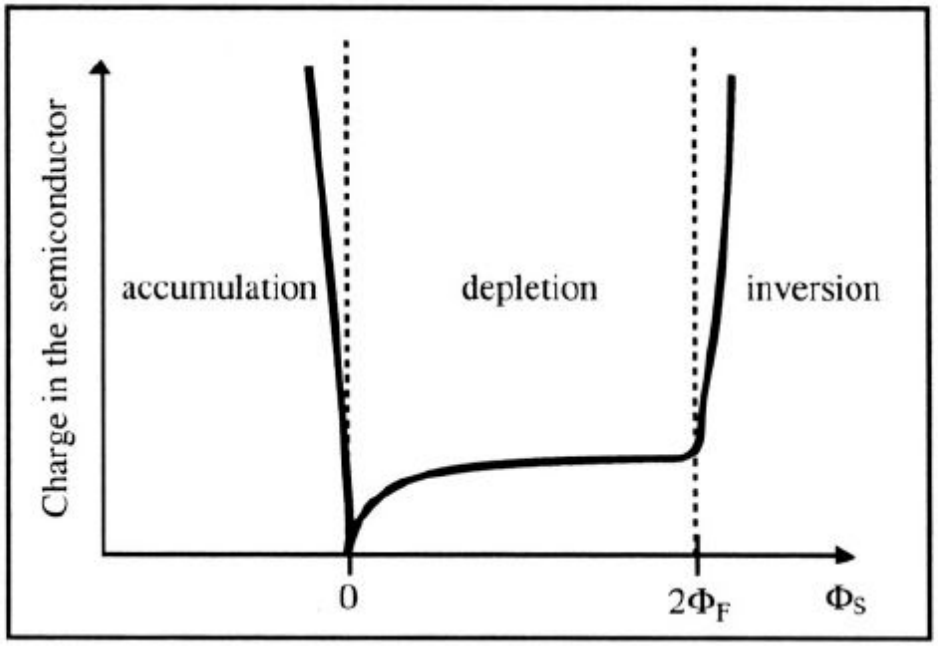


Figure 7.10: Absolute value of the charge in the semiconductor as a function of the surface potential.[10]

The total charge in the semiconductor can be plotted as a function of the surface potential (Figure 7.10). Accumulation and inversion charges are exponential functions of the surface potential, while the depletion charge varies as a square root.

MOS capacitor (Threshold voltage)

7.3. Threshold voltage

The threshold voltage of a MOS transistor is the voltage that must be applied on the gate to form an inversion layer. It depends on several device parameters which will be described next.

7.3.1 Ideal threshold voltage

In a MOS transistor the gate voltage is equal to sum of the potential drops in the semiconductor and the oxide. If one assumes that the back of the semiconductor is grounded, one can write:

$$V_G = \Phi_S + \frac{Q_G}{C_{ox}} \qquad (7.3.1)$$

where Q_G is equal to the positive charge on the gate electrode. An equal amount of negative charge exists in the semiconductor, comprised of ionized impurities in the depletion zone, and free electrons at the oxide/silicon interface a inversion. If we assume that the charge due to the free electrons is much smaller than the charge due to ionized impurities when the inversion layer starts being formed then Equation 7.3.1 can be written as:

$$V_G = 2\Phi_F - \frac{Q_d}{C_{ox}} = 2\Phi_F + \frac{\sqrt{4q\varepsilon_{si}N_a\Phi_F}}{C_{ox}} \equiv V'_{THo} \qquad (7.3.2)$$

V'_{THo} is called the "ideal threshold voltage" and it is measured with respect to the source. In this definition of the threshold voltage both the source and the substrate are grounded.

MOS capacitor (Flat-band voltage)

Equalization of the Fermi levels

We have so far assumed that the Fermi level of the metal gate was equal to that of the silicon. In practice this is not the case. In modern devices the gate material is not an actual metal, but heavily doped polycrystalline silicon, also called polysilicon. The doping concentration used for that material is so high ($\approx 10^{20}$ cm^{-3}) that it can be considered as a metal, for all practical purposes. Let us first consider the metal and the semiconductor separately. The energy which is necessary to extract an electron with an energy E_{FM} from the metal is called the "work function", $q\Phi_m$. Similarly, the work function in the semiconductor is noted $q\Phi_{sc}$.

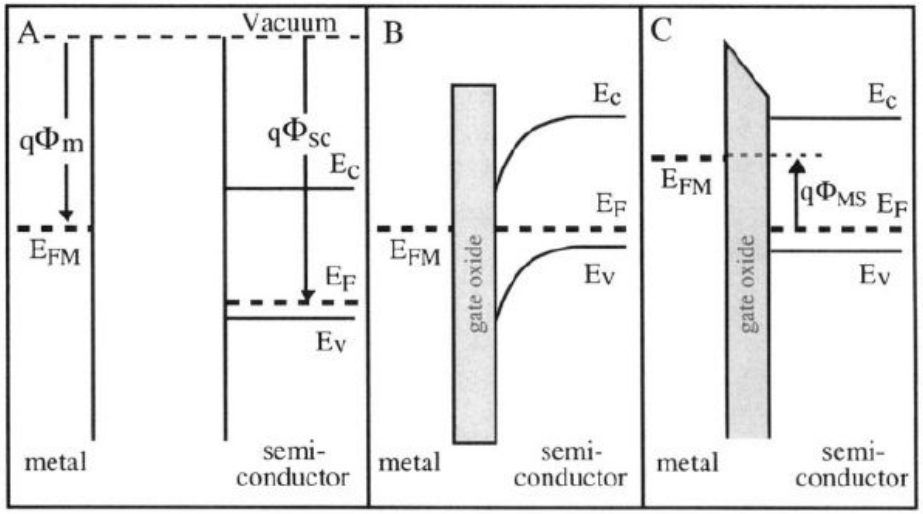


Figure 7.11: Energy bands in the MOS structure. A: The metal and the semiconductor are taken separately; B: Alignment of the Fermi levels with no applied bias, C: A bias equal to Φ_{ms} is applied to the gate.[11]

MOS capacitor (Flat-band voltage)

When the two materials are put together to form the MOS structure, the Fermi levels align, and the charge transfer resulting from this process curves the energy bands in the semiconductor, near the semiconductor-oxide interface (Figure 7.11). To recover to a flat-band condition a voltage must be applied to the gate. This voltage is equal to the difference of the work functions between the two materials, called the "work function difference", and is noted Φ_{ms}:

$$\Phi_{ms} = \Phi_m - \Phi_{sc} = \frac{E_F - E_{FM}}{q} \qquad (7.3.3)$$

Flat-band voltage

The "flat-band voltage" is the voltage that must be applied to the gate to bring the semiconductor energy bands to a flat level. Platband is achieved by applying a gate voltage which compensates for 1) differences in work functions of the semiconductor and the gate electrode, 2) the presence of charges in the oxide, and 3) interface traps. The sum of all these effects is found by adding Expressions 7.3.3, 7.3.5, and 7.3.6:

$$V_{FB} = \Phi_{ms} + V_Q + V_{it} = \Phi_{ms} - \frac{Q_{ox}}{C_{ox}} + \frac{2qN_{it}\Phi_F}{C_{ox}} \qquad (7.3.7)$$