CHAPTER 3

MOS TRANSISTOR

MOSFET module ⇩ 20-25M : Two terminal MOS Structure, MOSFET construction, Energy band diagrams under equilibrium and external bias, Threshold voltage, V-I and C-V characteristics, Channel length modulation, Short-channel effects, MOSFET Model.

The *MOS Field Effect Transistor* (MOSFET) is the fundamental building block of MOS and CMOS digital integrated circuits. Compared to the bipolar junction transistor (BJT), the MOS transistor occupies a relatively smaller silicon area, and its fabrication involves fewer processing steps. These technological advantages, together with the relative simplicity of MOSFET operation, have helped make the MOS transistor the most widely used switching device in LSI and VLSI circuits. In this chapter, we will examine the basic structure and the electrical behavior of nMOS (n-channel MOS), as well as pMOS (p-channel MOS) devices. The nMOS transistor is used as the primary switching device in virtually all digital circuit applications, whereas the pMOS transistor is used mostly in conjunction with the nMOS device in CMOS circuits. However, the basic operation principles of both nMOS and pMOS transistors are very similar to each other.

This chapter starts with a detailed investigation of the basic electrical and physical properties of *Metal Oxide Semiconductor* (MOS) systems, upon which the MOSFET structure is based. We will consider the effects of external bias conditions on charge distribution in the MOS system and on the conductance of free carriers. It will be shown that, in field effect devices, the current flow is controlled by externally applied electric fields, and that the operation depends only on the majority carrier flow between two device terminals. Next, the current-voltage characteristics of MOS transistors will be examined in detail, including physical limitations imposed by small device geometries and various second-order effects observed in MOSFETs. Note that these considerations will be particularly important for the overall performance of large-scale digital circuits built by using small-geometry MOSFET devices.

3.1. The Metal Oxide Semiconductor (MOS) Structure

We will start our investigation by considering the electrical behavior of the simple two-terminal MOS structure shown in Fig. 3.1. Note that the structure consists of three layers: The metal *gate electrode*, the insulating oxide (SiO_2) layer, and the p-type bulk semiconductor (Si), called the *substrate*. As such, the MOS structure forms a capacitor, with the gate and the substrate acting as the two terminals (plates) and the oxide layer as the dielectric. The thickness of the silicon dioxide layer is usually between 10 nm and 50 nm. The carrier concentration and its local distribution within the semiconductor substrate can now be manipulated by the external voltages applied to the gate and substrate terminals. A basic understanding of the bias conditions for establishing different carrier concentrations in the substrate will also provide valuable insight into the operating conditions of more complicated MOSFET structures.

Two-terminal MOS structure

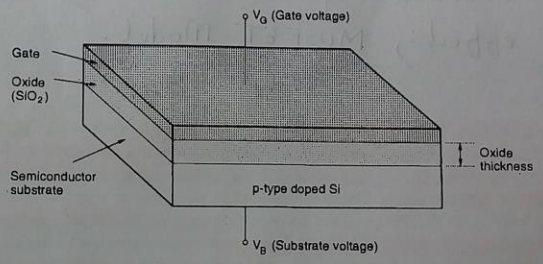


Figure 3.1. Two-terminal MOS structure.

Consider first the basic electrical properties of the semiconductor (Si) substrate, which acts as one of the electrodes of the MOS capacitor. The equilibrium concentrations of mobile carriers in a semiconductor always obey the *Mass Action Law* given by

$$n \cdot p = n_i^2 \tag{3.1}$$

Here, n and p denote the mobile carrier concentrations of electrons and holes, respectively, and n_i denotes the intrinsic carrier concentration of silicon, which is a function of the temperature T. At room temperature, i.e., $T = 300$ K, n_i is approximately equal to 1.45×10^{10} cm^{-3}. Assuming that the substrate is uniformly doped with an acceptor (e.g., Boron) concentration N_A, the equilibrium electron and hole concentrations in the p-type substrate are approximated by

$$n_{po} \cong \frac{n_i^2}{N_A}$$
$$p_{po} \cong N_A \tag{3.2}$$

The doping concentration N_A is typically on the order of 10^{15} to 10^{16} cm^{-3}; thus, it is much greater than the intrinsic carrier concentration n_i. Note that the bulk electron and hole concentrations given in (3.2) are valid in the regions farther away from the *surface*, where the semiconductor substrate and the oxide layer meet. The conditions on the surface, however, are far more significant for the electrical behavior and the operation of the MOS system, and we will discuss these conditions in more detail.

The energy band diagram of the p-type substrate is shown in Fig. 3.2. The band-gap between the *conduction band* and the *valence band* for silicon is approximately 1.1 eV. The location of the *equilibrium Fermi level* E_F within the band-gap is determined by the doping type and the doping concentration in the silicon substrate. The *Fermi potential* ϕ_F, which is a function of temperature and doping, denotes the difference between the intrinsic Fermi level E_i, and the Fermi level E_F.

$$\phi_F = \frac{E_F - E_i}{q} \tag{3.3}$$

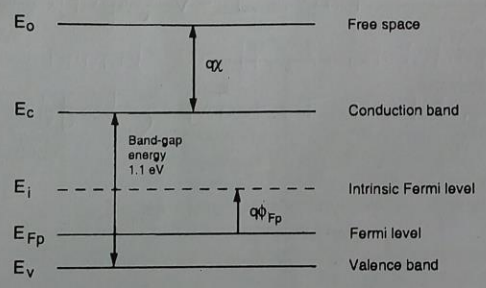


Figure 3.2. Energy band diagram of a p-type silicon substrate.

For a p-type semiconductor, the Fermi potential can be approximated by

$$\phi_{Fp} = \frac{kT}{q} \ln \frac{n_i}{N_A} \tag{3.4}$$

whereas for an n-type semiconductor (doped with a donor concentration N_D), the Fermi potential is given by

$$\phi_{Fn} = \frac{kT}{q} \ln \frac{N_D}{n_i} \tag{3.5}$$

Here, k denotes the Boltzmann constant and q denotes the unit (electron) charge. Note that the definitions given in (3.4) and (3.5) result in a positive Fermi potential for n-type

material, and a negative Fermi potential for p-type material. We will use this convention throughout the text. The *electron affinity* of silicon, which is the potential difference between the conduction band level and the vacuum (free-space) level, is denoted by $q\chi$ in Fig. 3.2. The energy required for an electron to move from the Fermi level into free space is called the *work function* $q\Phi_S$, and is given by

$$q\Phi_S = q\chi + (E_c - E_F) \tag{3.6}$$

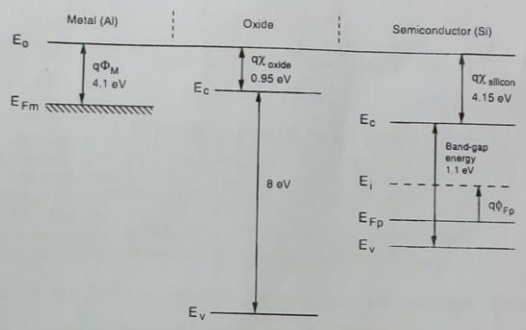
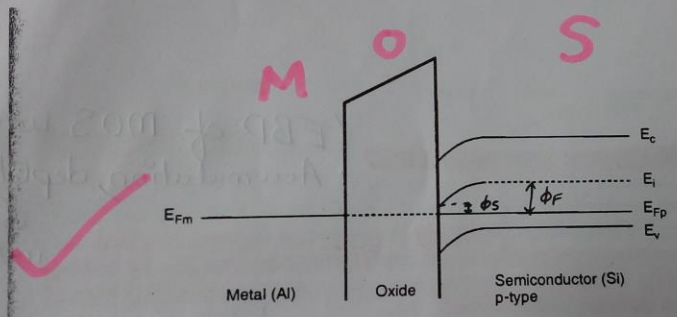


Figure 3.3. Energy band diagrams of the components that make up the MOS system.

The insulating silicon dioxide layer between the silicon substrate and the gate has a large band-gap of about 8 eV and an electron affinity of about 0.95 eV. On the other hand, the work function $q\Phi_M$ of an aluminum gate is about 4.1 eV. Figure 3.3 shows the energy band diagrams of metal, oxide, and semiconductor layers in a MOS system as three separate components.

Now consider that the three components of the ideal MOS system are brought into physical contact. The Fermi levels of all three materials must line up, as they form the MOS capacitor shown in Fig. 3.1. Because of the work-function difference between the metal and the semiconductor, a voltage drop occurs across the MOS system. Part of this built-in voltage drop occurs across the insulating oxide layer. The rest of the voltage drop (potential difference) occurs at the silicon surface next to the silicon-oxide interface, forcing the energy bands of silicon to bend in this region. The resulting combined energy band diagram of the MOS system is shown in Fig. 3.4. Notice that the equilibrium Fermi levels of the semiconductor (Si) substrate and the metal gate are at the same potential. The bulk Fermi level is not significantly affected by the band bending, whereas the surface Fermi level moves closer to the intrinsic Fermi (mid-gap) level. The Fermi potential at the surface, also called *surface potential* ϕ_s, is smaller in magnitude than the bulk Fermi potential ϕ_F.



ϕ_F - bulk Fermi potential
ϕ_s - surface Fermi potential

Figure 3.4. Energy band diagram of the combined MOS system.

3.2. The MOS System under External Bias

We now turn our attention to the electrical behavior of the MOS structure under externally applied bias voltages. Assume that the substrate voltage is set at $V_B = 0$, and let the gate voltage be the controlling parameter. Depending on the polarity and the magnitude of V_G, three different operating regions can be observed for the MOS system: *accumulation*, *depletion*, and *inversion*.

If a negative voltage V_G is applied to the gate electrode, the holes in the p-type substrate are attracted to the semiconductor-oxide interface. The majority carrier concentration near the surface becomes larger than the equilibrium hole concentration in the substrate; hence, this condition is called carrier *accumulation* on the surface (Fig. 3.5). Note that in this case, the oxide electric field is directed towards the gate electrode. The negative surface potential also causes the energy bands to bend upward near the surface. While the hole density near the surface increases as a result of the applied negative gate bias, the electron (minority carrier) concentration decreases as the negatively charged electrons are pushed deeper into the substrate.

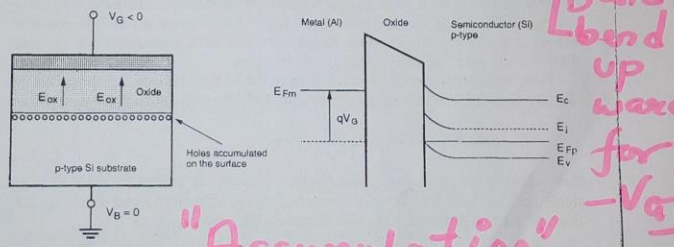
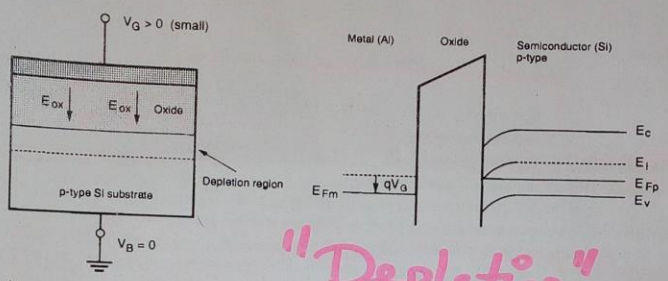


Figure 3.5. The cross-sectional view and the energy band diagram of the MOS structure operating in accumulation region.

Now consider the next case in which a small positive gate bias V_G is applied to the gate electrode. Since the substrate bias is zero, the oxide electric field will be directed towards the substrate in this case. The positive surface potential causes the energy bands to bend downward near the surface, as shown in Fig. 3.6. The majority carriers, i.e., the holes in the substrate, will be repelled back into the substrate as a result of the positive gate bias, and these holes will leave negatively charged fixed acceptor ions behind. Thus, a *depletion* region is created near the surface. Note that under this bias condition, the region near the semiconductor-oxide interface is nearly devoid of all mobile carriers.

The thickness x_d of this depletion region on the surface can easily be found as a function of the surface potential ϕ_s. Assume that the mobile hole charge in a thin horizontal layer parallel to the surface is



"Depletion"

Figure 3.6. The cross-sectional view and the energy band diagram of the MOS structure operating in depletion mode, under small gate bias.

$$dQ = -q \cdot N_A \cdot dx \tag{3.7}$$

The *change* in surface potential required to displace this charge sheet dQ by a distance x_d away from the surface can be found by using the Poisson equation.

$$d\phi_s = -x \cdot \frac{dQ}{\varepsilon_{Si}} = \frac{q \cdot N_A \cdot x}{\varepsilon_{Si}} dx \tag{3.8}$$

Integrating (3.7) along the vertical dimension (perpendicular to the surface) yields

$$\int_{\phi_F}^{\phi_s} d\phi_s = \int_0^{x_d} \frac{q \cdot N_A \cdot x}{\varepsilon_{Si}} \, dx \tag{3.9}$$

$$\phi_s - \phi_F = \frac{q \cdot N_A \cdot x_d^2}{2\,\varepsilon_{Si}} \tag{3.10}$$

Thus, the depth of the depletion region is

$$x_d = \sqrt{\frac{2\varepsilon_{Si} \cdot |\phi_s - \phi_F|}{q \cdot N_A}} \tag{3.11}$$

cm

and the depletion region charge density, which consists solely of fixed acceptor ions in this region, is given by the following expression

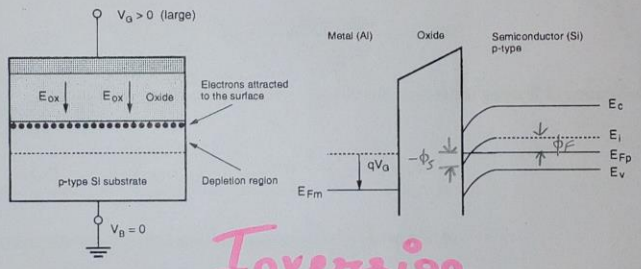
$$Q = -q \cdot N_A \cdot x_d = -\sqrt{2q \cdot N_A \cdot \varepsilon_{Si} \cdot |\phi_s - \phi_F|} \quad C/cm^2 \quad (3.12)$$

The amount of this depletion region charge plays a very important role in the analysis of threshold voltage, as we will examine shortly.

To complete our qualitative overview of different bias conditions and their effects upon the MOS system, consider next a further increase in the positive gate bias. As a result of the increasing surface potential, the downward bending of the energy bands will increase as well. Eventually, the mid-gap energy level E_i becomes smaller than the Fermi level E_{Fp} on the surface, which means that the substrate semiconductor in this region becomes n-type. Within this thin layer, the electron density is larger than the majority hole density, since the positive gate potential attracts additional minority carriers (electrons) from the bulk substrate to the surface (Fig. 3.7). The n-type region created near the surface by the positive gate bias is called the *inversion layer*, and this condition is called *surface inversion*. It will be seen that the thin inversion layer on the surface with a large mobile electron concentration can be utilized for conducting current between two terminals of the MOS transistor.

Inversion:

means pile up of minority carriers so that they dominate the "bulk" majority carriers at the interface.



Inversion

Figure 3.7. The cross-sectional view and the energy band diagram of the MOS structure in surface inversion, under larger gate bias voltage.

As a practical definition, the surface is said to be *inverted* when the density of mobile electrons on the surface becomes equal to the density of holes in the bulk (p-type) substrate. This condition requires that the surface potential has the same magnitude, but the reverse polarity, as the bulk Fermi potential ϕ_F. Once the surface is inverted, any further increase in the gate voltage leads to an increase of mobile electron concentration on the surface, but not to an increase of the depletion depth. Thus, the depletion region depth achieved at the onset of surface inversion is also equal to the maximum depletion depth, x_{dm}, which remains constant for higher gate voltages. Using the inversion condition $\phi_s = -\phi_F$, the maximum depletion region depth at the onset of surface inversion can be found from (3.11) as follows:

"Inversion condition"

$\phi_s = -\phi_F$

$$x_{dm} = \sqrt{\frac{2\varepsilon_{Si}|2\phi_F|}{qN_A}}$$

- The creation of a conducting surface inversion layer through externally applied gate bias is an essential phenomenon for current conduction in MOS transistors.