**Introduction**

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.

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 (SiO2) 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 diagram of a rectangular object

Description automatically generated

Here, n and p denote the mobile carrier concentrations of electrons and holes.

The intrinsic carrier concentration, denoted by ni, refers to the number of free electrons and holes present within an intrinsic semiconductor per unit volume at a specific temperature. These free carriers are responsible for the material's electrical conductivity.



The bulk electron and hole concentrations given in (1) 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

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 EF within the band-gap is determined by the doping type and the doping concentration in the silicon substrate. The Fermi potential phiF, which is a function of temperature and doping, denotes the difference between the intrinsic Fermi level Ei, and the Fermi level Ef

A diagram of energy and electricity

Description automatically generated with medium confidence

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

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

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.

Electron affinity refers to the amount of energy released when an electron is added to a neutral atom or molecule in the gaseous state, forming an anion. In simpler terms, it's the willingness of an atom to accept an extra electron, is denoted by qX in Fig. 3.2. The energy required for an electron to move from the Fermi level into free space is called the work function qs, and is given by

A diagram of an electrical diagram

Description automatically generated

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 Os, is smaller in magnitude than the bulk Fermi potential OF

A diagram of different types of metal

Description automatically generated

**The MOS System under External Bias**

Assume that the substrate voltage is set at VB = 0, and let the gate voltage be the controlling parameter. Depending on the polarity and the magnitude of VG, three different operating regions can be observed for the MOS system:

* Accumulation
* Depletion
* Inversion

**Accumulation**

If a negative voltage VG 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.

A diagram of a circuit

Description automatically generated with medium confidence

**Depletion**

Now consider the next case in which a small positive gate bias VG 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 xd of this depletion region on the surface can easily be found as a function of the surface potential s. Assume that the mobile hole charge in a thin horizontal layer parallel to the surface is

A diagram of a diagram of a building

Description automatically generated with medium confidence



**Inversion**

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 Ei becomes smaller than the Fermi level EFP 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.

A diagram of a diagram of a radio wave

Description automatically generated with medium confidence

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 OF. 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, xdm, which remains constant for higher gate voltages.

**Structure and Operation of MOS Transistor (MOSFET)**

The basic structure of an n-channel MOSFET is shown in Fig. 3.8. This four-terminal device consists of a p-type substrate, in which two n+ diffusion regions, the drain and the source, are formed. The surface of the substrate region between the drain and the source is covered with a thin oxide layer, and the metal (or polysilicon) gate is deposited on top of this gate dielectric. The midsection of the device can easily be recognized as the basic MOS structure which was examined in the previous sections. The two n+ regions will be the current-conducting terminals of this device. Note that the device structure is completely symmetrical with respect to the drain and source regions; the different roles of these two regions will be defined only in conjunction with the applied terminal voltages and the direction of the current flow.

Diagram of a rectangular object with text and symbols

Description automatically generated

A conducting channel will eventually be formed through applied gate voltage in the section of the device between the drain and the source diffusion regions. The distance between the drain and source diffusion regions is the channel length L, and the lateral extent of the channel (perpendicular to the length dimension) is the channel width W. Both the charnel length and the channel width are important parameters which can be used to control some of the electrical properties of the MOSFET. The thickness of the oxide layer covering the channel region, tox, is also an important parameter.

A MOS transistor which has no conducting channel region at zero gate bias is called an enhancement-type (or enhancement-mode) MOSFET. If a conducting channel already exists at zero gate bias, on the other hand, the device is called a depletion-type (or depletion-mode) MOSFET. In a MOSFET with p-type substrate and with n+ source and drain regions, the channel region to be formed on the surface is n-type. Thus, such a device

with p-type substrate is called an n-channel MOSFET. In a MOSFET with n-type substrate and with p+ source and drain regions, on the other hand, the channel is p-type and the device is called a p-channel MOSFET.

A diagram of a circuit

Description automatically generated with medium confidence

The abbreviations used for the device terminals are: G for the gate, D for the drain, S for the source, and B for the substrate (or body). In an n-channel MOSFET, the source is defined as the n region which has a lower potential than the other n region, the drain. By convention, all terminal voltages of the device are defined with respect to the source potential. Thus, the gate-to-source voltage is denoted by VGS, the drain-to-source voltage is denoted by VDS, and the substrate-to-source voltage is denoted by VBS. Circuit symbols for both n-channel and p-channel enhancement-type MOSFETs are shown in Fig. 3.9. While the four-terminal symbolic representation shows all external terminals of the device, the simple three-terminal representation will also be used extensively.

Consider first the n-channel enhancement-type MOSFET shown in Fig. 3.8. The simple operation principle of this device is: control the current conduction between the source and the drain, using the electric field generated by the gate voltage as a control variable. Since the current flow in the channel is also controlled by the drain-,to-source voltage and by the substrate voltage, the current can be considered a function of these external terminal voltages.

In order to start current flow between the source and the drain regions, however, we have to form a conducting channel first. The simplest bias condition that can be applied to the n-channel enhancement-type MOSFET is shown in Fig: 3.10. The source, the drain, and the substrate terminals are all connected to ground. A positive gate-to-source voltage VGS is then applied to the gate in order to create the conducting channel underneath the gate. With this bias arrangement, the channel region between the source and the drain diffusions behaves exactly the same as for the simple MOS structure we examined in Section 3.2. For small gate voltage levels, the majority carriers (holes) are repelled back into the substrate, and the surface of the p-type substrate is depleted. Since the surface is devoid of any mobile carriers, current conduction between the source and the drain is not possible.

A diagram of a device

Description automatically generated

Now assume that the gate-to-source voltage is further increased. As soon as the surface potential in the channel region reaches - F surface inversion will be established, and a conducting n-type layer will form between the source and the drain diffusion regions (Fig. 3.11). This channel now provides an electrical connection between the two n+ regions, and it allows current flow, as long as there is a potential difference between the source and the drain terminal voltages (Fig. 3.12). The bias conditions for the onset of surface inversion and for the creation of the conducting channel are therefore very significant for MOSFET operation.

The value of the gate-to-source voltage VGS needed to cause surface inversion (to create the conducting channel) is called the threshold voltage V. Any gate to-source voltage smaller than V70 is not sufficient to establish an inversion layer; thus, the MOSFET can conduct no current between its source and drain terminals unless VGS >VT . For gate-to-source voltages larger than the threshold voltage, on the other hand a larger number of minority carriers (electrons) are attracted to the surface, which ultimately contribute to channel current conduction. Also note that increasing the gateto-source voltage above and beyond the threshold voltage will not affect the surface potential and the depletion region depth. Both quantities will remain approximately constant and equal to their values attained at the onset of surface inversion.

**MOSFET Operation: A Qualitative View**

The basic structure of the n-channel MOS (nMOS) transistor built on a p-type substrate was shown in Fig. 3.8. The MOSFET consists of a MOS capacitor with two p-n junctions placed immediately adjacent to the channel region that is controlled by the MOS gate. The carriers, i.e., electrons in an nMOS transistor, enter the structure through the source contact (S), leave through the drain (D), and are subject to the control of the gate (G) voltage. To ensure that both p-n junctions are reverse-biased initially, the substrate potential is kept lower than the other three terminal potentials. We have seen that when 0 < VGS < V, the gated region between the source and the drain is depleted; no carrier flow can be observed in the channel. As the gate voltage is increased beyond the threshold voltage (VGS > Vr), however, the mid-gap energy level at the surface is pulled below the Fermi level, causing the surface potential Os to turn positive and to invert the surface (Fig. 3.12). Once the inversion layer is established on the surface, an n-type conducting channel forms between the source and the drain, which is capable of carrying the drain current. Next, the influence of drain-to-source bias VDS and different modes of drain current flow will be examined for an nMOS transistor with VGS > V70. At VDS = 0, thermal equilibrium exists in the inverted channel region, and the drain current ID is equal to zero (Fig. 3.14(a)). If a small drain voltage VDS > 0 is applied, a drain current proportional to VDS will flow from the source to the drain through the conducting channel. The inversion layer, i.e., the channel, forms a continuous current path from the source to the drain. This operation mode is called the linear mode, or the linear region. Thus, in linear region operation, the channel region acts as a voltage-controlled resistor. The electron velocity, in the channel for this case is usually much lower than the drift velocity limit. Note that as the drain voltage is increased, the inversion layer charge and the channel depth at the drain end start to decrease. Eventually, for VDS = VDSA7, the inversion charge at the drain is reduced to zero, which is called the pinch-off point (Fig. 3.14i?).

A diagram of a diagram

Description automatically generated

Beyond the pinch-off point, i.e., for VDS > VDSAT, a depleted surface region forms adjacent to the drain, and this depletion region grows toward the source with increasing drain voltages. This operation mode of the MOSFET is called the saturation mode or the saturation region; For a MOSFET operating in the saturation region, the effective channel length is reduced as the inversion layer near the drain vanishes, while the channel-end voltage remains essentially constant and equal to VDSAT(Fig. 3.14(c)). Note that the pinched-off (depleted) section of the channel absorbs most of the excess voltage drop (VDS - VDSAT) and a high-field region forms between the channel-end and the drain boundary. Electrons arriving from the source to the channel-end are injected into the drain-depletion region and are accelerated toward the drain in this high electric field, usually reaching the drift velocity limit. The pinch-off event, or the disruption of the continuous channel under high drain bias, characterizes the saturation mode operation of the MOSFET.

The influence of these operating conditions upon the external (terminal) current- voltage characteristics of the MOS transistor will be examined in the following section.

Channel length modulation

**MOSFET Scaling and Small-Geometry Effects**

The reduction of the size, i.e., the dimensions of MOSFETs, is commonly referred to as scaling. It is expected that the operational characteristics of the MOS transistor will change with the reduction of its dimensions. Also, some physical limitations eventually restrict the extent of scaling that is practically achievable. There are two basic types of size-reduction strategies: full scaling (also called constant-field scaling) and constant voltage scaling.

Scaling of MOS transistors is concerned with systematic reduction of overall dimensions of the devices as allowed by the available technology, while preserving the geometric ratios found in the larger devices. The proportional scaling of all devices in a circuit would certainly result in a reduction of the total silicon area occupied by the circuit, thereby increasing the overall functional density of the chip. To describe device scaling, we introduce a constant scaling factor S > 1.

All horizontal and vertical dimensions of the large-size transistor are then divided by this scaling factor to obtain the scaled device. The extent of scaling that is achievable is obviously determined by the fabrication technology and more specifically, by the minimum feature size.

We consider the proportional scaling of all three dimensions by the same scaling factor S. Figure 3.24 shows the reduction of key dimensions on a typical MOSFET, together with the corresponding increase of the doping densities.

A diagram of a rectangular object with text and symbols

Description automatically generated

The primed quantities in Fig. 3.24 indicate the scaled dimensions and doping densities. It is easy to recognize that the scaling of all dimensions by a factor of S > 1 leads to the reduction of the area occupied by the transistor by a factor of S2. To better understand the effects of scaling upon the current-voltage characteristics of the MOSFET, we will examine two different scaling options in the following sections

**Full Scaling (Constant-Field Scaling)**

This scaling option attempts to preserve the magnitude of internal electric fields in the MOSFET, while the dimensions are scaled down by a factor of S. To achieve this goal, all potentials must be scaled down proportionally, by the same scaling factor. Note that this potential scaling also affects the threshold voltage V.0 Finally, the Poisson equation describing the relationship between charge densities and electric fields dictates that the charge densities must be increased by a factor of S in order to maintain the field conditions. Table 3.2 lists the scaling factors for all significant dimensions, potentials, and doping densities of the MOS transistor.

This significant reduction of the power dissipation is one of the most attractive features of full scaling. Note that with the device area reduction by S2 discussed earlier, we find the power density per unit area remaining virtually unchanged for the scaled device. Finally, consider the gate oxide capacitance defined as Cg = WL COX' It will be shown later in Section 3.6 that charging and discharging of this capacitance plays an important role in the transient operation of the MOSFET. Since the gate oxide capacitance C is scaled down by a factor of S, we can predict that the transient characteristics, i.e., the charge-up and charge-down times, of the scaled device will improve accordingly. In addition, the proportional reduction of all dimensions on-chip will lead to a reduction of various parasitic capacitances and resistances as well, contributing to the overall performance improvement. Table 3.3 summarizes the changes in key device characteristics as a result of full (constant-field) scaling

A table with text and symbols

Description automatically generated

**Constant-Voltage Scaling**

While the full scaling strategy dictates that the power supply voltage and all terminal voltages be scaled down proportionally with the device dimensions, the scaling of voltages may not be very practical in many cases. In particular, the peripheral and interface circuitry may require certain voltage levels for all input and output voltages, which in turn would necessitate multiple power supply voltages and complicated level shifter arrangements. For these reasons, constant-voltage scaling is usually preferred over full scaling.

A table with black text

Description automatically generated with medium confidence

In constant-voltage scaling, all dimensions of the MOSFET are reduced by a factor

of S, as in full scaling. The power supply voltage and the terminal voltages, on the other hand, remain unchanged. The doping densities must be increased by a factor of s2 in order to preserve the charge-field relations. Table 3.4 shows the constant-voltage scaling of key dimensions, voltages, and densities. Under constant-voltage scaling, the changes in

device characteristics are significantly different compared to those in full scaling, as we will demonstrate.

To summarize, constant-voltage scaling may be preferred over full (constant-field) scaling in many practical cases because of the external voltage-level constraints. It must be recognized, however, that constant-voltage scaling increases the drain current density and the power density by a factor of S3. This large increase in current and power densities may eventually cause serious reliability problems for the scaled transistor, such as electromigration, hot-carrier degradation, oxide breakdown, and electrical over-stress. As the device dimensions are systematically reduced through full scaling or constant-voltage scaling, various physical limitations become increasingly more prominent, and ultimately restrict the amount of feasible scaling for some device dimensions. Consequently, scaling may be carried out on a certain subset of MOSFET dimensions in many practical cases. Also, the simple gradual channel approximation (GCA) used for the derivation of current-voltage relationships does not accurately reflect the effects of scaling in smaller-size transistors.