1

INTRODUCTION TO FIELD-EFFECT TRANSISTORS

COPYRIGHTED MATERIAL

We are living in an era of information technology where smartphones, smart watches, and smart technology have become an inevitable part of our lives. You might have observed a drastic improvement in the performance of these smart devices. For instance, the shift from single core processors to multicore processors, the increase in CPU's frequency from few MHz to several GHz, the increase in the RAM from few MB to several GB, and so on. All these factors have led to a tremendous increase in the performance of these computing devices. The smart devices found in every household nowadays have a performance metric comparable to the earlier supercomputers. For instance, the Apple watch has twice the processing power of a 1985 Cray-2 supercomputer [1]. In addition, the device size has also shrunk significantly and the focus in the research and development of computing devices has shifted toward mobile devices. Moreover, the functionality per device has also increased considerably. For instance, the present day smartphones not only have processing capabilities of a supercomputer but can also perform the functions of a good quality camera, a Wi-Fi dongle, an X-BOX gaming system, and so on. To summarize, every other person in this modern era has access to low-cost, high-performance gadgets.

Have you ever wondered what drives the "smartness" and the supercomputing capabilities of all the smart technology gadgets? Let us try to understand this from a human body–gadget analogy. Just like the human body is composed of cells as the building block, the electronic gadgets are made up of transistors. In human body, the

Junctionless Field-Effect Transistors: Design, Modeling, and Simulation, First Edition.
Shubham Sahay and Mamidala Jagadesh Kumar.
© 2019 by The Institute of Electrical and Electronics Engineers, Inc. Published 2019 by John Wiley & Sons, Inc.

JWBS243-c01 JWBS243-Sahay December 21, 2018 16:46 Printer Name: Trim: 6in × 9in

cells are grouped together to perform a particular function and form an organ. Therefore, the efficiency and the number of different functions that can be performed by the body depends exclusively on these cells. Similarly, the transistors act like a switch and are wired together in a chip (which is similar to the organ from body–gadget analogy) in a specific manner to enable a particular function. The larger the number of transistors in a gadget, the more the number of functions it can perform. The research and development in the field of transistors has driven this "smart" revolution. It is indeed very interesting how such small chunks of silicon chips drive our lives.

1.1 TRANSISTOR ACTION

But what exactly is a transistor? The word transistor was given by its first inventors: Shockley, Brattain, and Bardeen in 1947 [2–5]. At that time, no one would have wondered that this discovery (which actually was an accident) would be driving the lives of common people for generations to come. The transistors are often conceived as a device where the resistance between two terminals may be controlled by the current/voltage at the third terminal. Therefore, transistor refers to any three-terminal device where the current (or voltage) between two terminals may be controlled by the action of voltage (or current) at the third terminal.

In the subsequent sections, we shall see how the most common transistors work from both a qualitative approach and an energy band diagram perspective. The bipolar junction transistors (BJTs) dominated the semiconductor industry until late 1970s. Although BJTs are still used in the high-frequency circuits such as in radio frequency circuits, the throne is captured by the metal-oxide-semiconductor field-effect transistors (MOSFETs) and they continue to drive the semiconductor industry even today. Therefore, we shall discuss the MOSFETs in detail in the next section.

Transistors such as MOSFETs act as switches in the integrated circuits. However, it may be noted that the MOSFETs are not ideal switches (which are expected to consume no power when switched-OFF and deliver a high current instantaneously when switched-ON). The MOSFETs exhibit a small leakage current and, therefore, consume power from the supply even when they are switched-OFF. This power consumption is termed as the static power dissipation (P_{s}) given as

$$P_{\mathrm{S}} = V_{\mathrm{DD}} \cdot I_{\mathrm{OFF}} \tag{1.1}$$

where V_{DD} is the supply voltage and I_{OFF} is the leakage current that flows through the transistor when the switch is turned off. Furthermore, the MOSFETs also consume a significant power when switched from the ON-state to OFF-state or vice versa. This power consumption also depends on the frequency of switching of the MOSFETs and is termed as the dynamic power dissipation (P_{d}) given as

$$P_D = V_{\mathrm{DD}}^2 C_{\mathrm{L}} f \alpha \tag{1.2}$$

where V_{DD} is the supply voltage, f is the frequency of operation and α is the switching probability, which simply tells us that the MOSFET is not switched in each cycle, and C_L is the load capacitance. In a wired network of MOSFETs, a MOSFET drives another MOSFET. Therefore, in most cases C_L is the input capacitance of the MOSFET. The interested readers are requested to refer [5] for more details.

Until recent past, the focus of the researchers all over the world was to miniaturize the dimensions of the MOSFETs so as to increase the number of MOSFETs per chip, which would not only reduce the area enabling mobile devices but also increase the number of operations that may be performed by a single chip. Scaling the MOSFET dimensions also reduces the input capacitance and increases its capability (current) to drive another MOSFET in the wired chip network and helps to achieve large frequency of operation due to fast charging of C_L. Although the drive current of MOSFET increases with scaling, the OFF-state current also increases drastically due to the short-channel effects that are triggered by MOSFET gate length scaling. The increase in the OFF-state current results in a significant static power dissipation. While the dynamic power dissipation was a major concern for the researchers until recent past, the scaling trends suggest that the static power dissipation would eventually surpass the dynamic power dissipation if the conventional MOSFETs are scaled aggressively.

A high static power consumption means that the MOSFETs would draw a significantly large power from the supply even when it is switched-OFF. Therefore, the chip would drain the battery or the power source even when the functionality provided by the chip is not being utilized. This is detrimental to the performance of computing devices especially for the hand-held devices like smartphones, which have a limited supply available in the form of a battery. Furthermore, the static power dissipation also heats up the chip and degrades the performance of the gadgets which are designed for room temperature operation. Of course, every consumer wants to have a smart device with an unlimited battery or power supply with no heating effects. To reduce the power dissipation, we can reduce the supply voltage as evident from equations (1.1) and (1.2). However, there lies a fundamental limitation on the MOSFETs which is inherent to the very physics of the device. The current in a MOSFET cannot increase by more than a ten-fold when the input voltage is raised by 60 mV. This limitation is due to the Maxwell–Boltzmann distribution of electrons in matter and is often referred to as the "Boltzmann tyranny." The application of MOSFET as a switch requires that the ON-state to OFF-state current ratio be high so that these states are easily distinguishable ($\sim 10^4$ to 10^6). To achieve an ON-state to OFF-state current ratio of a million, the variation of the input voltage, and therefore the supply voltage, needs to be at least equal to $60 \times \log(10^6) = 360$ mV. This limitation simply implies that if we have an extremely scaled supply voltage, the ratio of the ON-state current to the OFF-state current of the transistor would be very low and the MOSFET would cease to act like a switch. Therefore, the Boltzmann limit hinders the use of the conventional MOSFETs as a switch for ultralow supply voltages.

As a result, the conventional MOSFETs cannot cater the need of yielding an area and power-efficient chip with multiple functionalities. Moreover, scaling the conventional MOSFETs also requires a large investment from the manufacturing point of view. Therefore, present day research focuses on design of a low-cost and highly scalable MOSFET with minimum power dissipation. As you would have noted, the research and development in this context has gradually shifted from an area-driven perspective to a power-driven scenario.

This chapter will help to develop a basic understanding of the conventional MOSFETs. After a subtle discussion of the various modes of operation of these devices, Section 1.3 describes how basic circuits can be formed using MOSFETs. Section 1.3.2 focuses on different types of power dissipations reported earlier in the introduction.

1.2 METAL-OXIDE-SEMICONDUCTOR FIELD-EFFECT TRANSISTORS

To understand a MOSFET, we shall first get an in-depth understanding of a MOS capacitor (Fig. 1.1) which is the heart of a MOSFET, grasp the concept of "field–effect," and then discuss operation of MOSFETs. The MOS capacitor consists

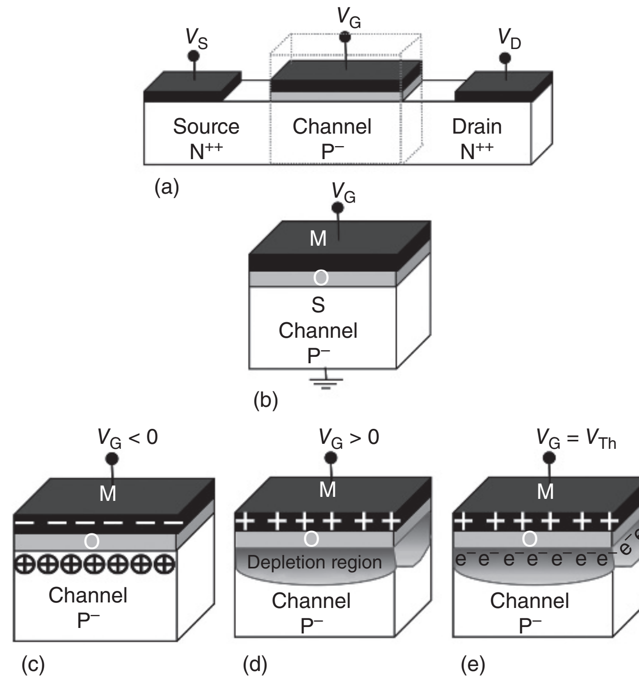

FIGURE 1.1 (a) Three-dimensional view of an n-MOSFET and (b) the MOS capacitor and the operation mode of a MOS capacitor in (c) accumulation regime, (d) depletion regime, and (e) inversion regime.

of three layers as the name suggests: metal-oxide-semiconductor. A thin insulating oxide layer is sandwiched between a metal and a semiconductor. Since the structure has a dielectric inserted between two conducting plates (assuming that the semiconductor is doped or at room temperature), the MOS structure is essentially a capacitor. The MOS capacitor with the p-type doped semiconductor is called a p-type MOS capacitor, whereas the MOS capacitor with an n-type doped semiconductor is called an n-type MOS capacitor.

The capacitance of the MOS structure can be controlled by the gate voltage just like the capacitance of a p–n junction is controlled by the applied bias. However, the range of capacitances exhibited by the MOSFET is large compared to the p–n junction capacitance.

At this point, we would also like to mention that the property of bulk atoms and surface atoms are different. Indeed, in the words of W. Pauli (who gave the Pauli exclusion principle), "God made the bulk; interfaces were invented by the devil" [7]. In MOS devices, all the charge dynamics occur at the surface. Therefore, silicon is the most preferred material for MOS devices as the Si–SiO_2 interface constitutes the best quality semiconductor–insulator interface. Silicon is also abundant on earth in the form of sand (silica).

1.2.1 "Field-Effect" and Operation Modes

To understand the different modes of operation of a MOS capacitor, it is essential to understand the concept of "field-effect" applied to the MOS devices. The field-effect simply means controlling the charge dynamics with the aid of an electric field. Now, let us look at how electric field controls charges in case of a p-type MOS capacitor shown in Fig. 1.1(b).

If we apply a negative voltage on the gate terminal with respect to silicon, an electric field will be generated across the insulator with a direction from the semiconductor to the metal. The applied negative potential on the gate can be conceptualized as depositing negative charges on the gate which attract the majority holes in the p–Si toward the Si–SiO_2 interface. Therefore, the holes would be accumulated at the Si–SiO_2 surface due to the application of a negative bias on the gate. From an electric field perspective, the holes move in the direction of the electric field and accumulate at the Si–SiO_2 interface. As a result, the effective carrier concentration at the interface is increased in the accumulation mode as shown in Fig. 1.1(c).

Now, if a positive voltage is applied to the gate terminal, the electric field direction across the insulator is reversed and points toward the semiconductor from the gate. A positive potential at the gate can also be conceptualized as depositing positive charges on the gate which repel the majority holes close to the Si–SiO_2 interface. The repelled holes move into the bulk leaving behind uncovered negative acceptor ions. Therefore, a depletion region is formed in the semiconductor in the vicinity of the Si–SiO_2 interface. In other words, the electric field pushes holes away from the interface to the bulk, increasing the depletion region width. This region of operation is called the depletion mode.

JWBS243-c01 JWBS243-Sahay December 21, 2018 16:46 Printer Name: Trim: 6in × 9in

Now, what happens if the positive voltage is increased even further? One may expect that the depletion region would continue to expand until it spans the entire semiconductor. However, this is not what happens since the minority electrons are also there in the bulk of the p-type semiconductor, which may provide negative charge for the electric field lines from the gate to terminate. Therefore, when we continue to increase the positive voltage, the depletion region increases until a maximum value and then the minority electrons move to the surface from the bulk and start accumulating to facilitate the termination of the electric field lines. From electric field perspective, the field becomes so strong that it pulls the minority electrons to the surface. The value of gate voltage at which the electron concentration at the surface becomes equal to the bulk doping concentration of the p-type semiconductor is called the threshold voltage. At the threshold voltage, the majority carriers change from holes to electrons at the surface. This phenomenon is called inversion, and the electron layer is called the inversion layer.

For the n-type MOS capacitor, accumulation of electrons takes place for positive gate voltages. Upon application of a negative gate voltage, the semiconductor is depleted first and then the depletion region reaches its maximum value. As the magnitude of the negative gate voltage is increased further, the minority holes start moving to the surface and, eventually, an inversion layer of holes is formed. Therefore, the characteristics of a p-type MOS capacitor are just complementary to the n-type MOS capacitor and hence do not require a detailed discussion.

Now, with this background, we shall discuss the structure of a MOSFET.

1.2.2 MOSFET as a Switch

A MOSFET consists of the MOS capacitor appended by the source and drain regions, which makes it a three-terminal device as shown in Fig. 1.1(a). The three terminals of the MOSFET are source (acts as a source of carriers), gate (controls the amount/concentration of carriers), and drain (acts as the sink for carriers). The source and drain are heavily doped, whereas the channel is lightly doped with a polarity opposite to that of source and drain. MOSFETs also utilize a fourth terminal called the body terminal. The body terminal is connected to the channel and is used to manipulate the electron conduction (and the threshold voltage) under special circumstances. Otherwise, it is normally grounded.

As discussed in Section 1.2.1, if a positive bias greater than the threshold voltage is applied, an inversion layer of electrons is formed at the surface of p-type silicon. Now, if a positive voltage is applied at the drain terminal, the electrons of the source would find a low-resistance conduction path via the inversion layer and flow into the drain. Therefore, the electrons would flow from the source to the drain region and MOSFET would act as a closed switch. However, when the applied bias is lower than the threshold voltage, the channel region remains depleted and offers a high-resistance path. Therefore, the electrons in the source find it difficult to reach the drain via the channel and the MOSFET behaves like a open switch. Therefore, the MOSFET acts like a switch which can be switched ON or OFF depending on

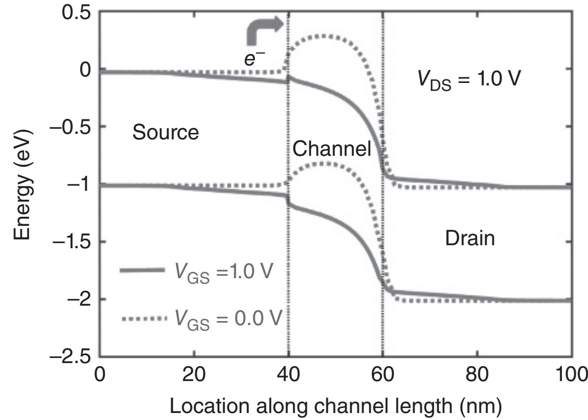

FIGURE 1.2 Energy band profiles of the MOSFET at ON-state ($V_{GS} = 1.0$ V) and OFF-state ($V_{GS} = 0.0$ V) showing that the gate voltage modulates the barrier height for source electrons.

the voltage applied to the gate. The drain to source current can, therefore, be controlled by the voltage applied to the gate. Since current through two terminals is being controlled by the voltage at the third terminal, the MOSFET is called a transistor, i.e. a resistor whose resistance may be controlled by the gate. As shown in Fig. 1.2, the gate voltage simply modulates the effective barrier height seen by the source electrons to move into the drain region through the channel.

1.2.3 Transfer Characteristics and Output Characteristics

At this point, we would like to introduce the concept of transfer characteristics and the output characteristics. The output characteristics refer to the relation between the output current (drain current in the case of a MOSFET) with the output voltage (drain voltage). Therefore, the output characteristics in a MOSFET are simply a plot between the drain current versus the drain voltage for a particular gate voltage as shown in Fig. 1.3(a). The drain current first increases linearly with the drain voltage and then gets saturated owing to the pinch-off of the channel region at the drain end. The inversion layer charge increases with increasing gate voltage leading to a larger drain current.

The relationship between the output current (drain current) and the input voltage (gate voltage) for a particular drain voltage is called the transfer characteristics. The drain current is very low below the threshold voltage (subthreshold regime) and is governed by diffusion of carriers from the source to the drain region. As the gate voltage increases above the threshold voltage, an inversion layer forms and the drain current increases significantly (Fig. 1.3(b)). If the transfer characteristics are plotted on a linear scale, the threshold voltage can be extracted by extrapolating the drain current after which the current starts increasing dramatically (Fig. 1.3(b)).

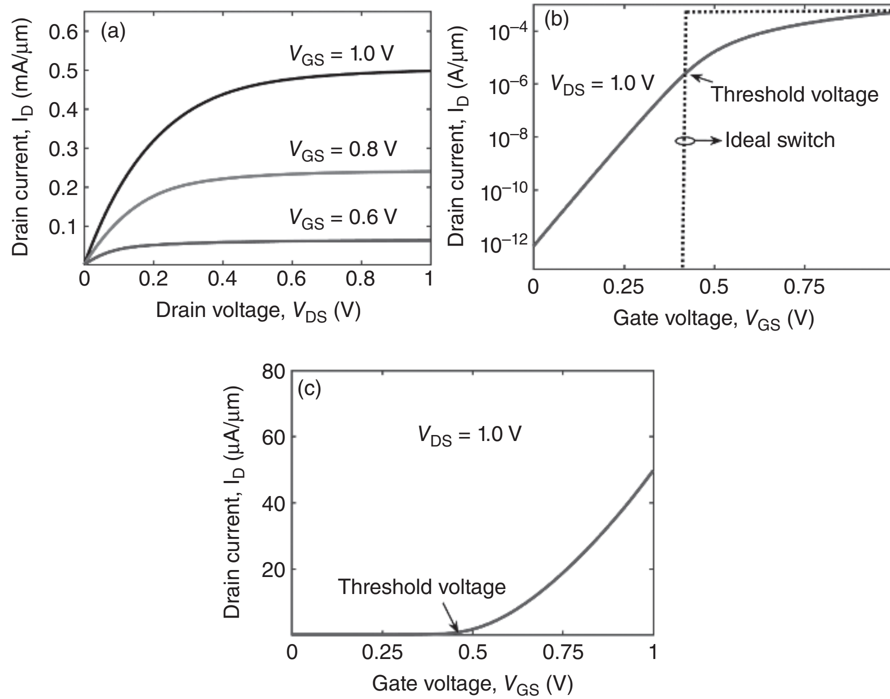

FIGURE 1.3 (a) Output characteristics and transfer characteristics of the MOSFET in (b) log scale and (c) linear scale.

Several methods have been proposed for determination of the threshold voltage. Some of these methods include finding the zeroes of the double derivative of the transfer characteristics (which is equivalent to finding maxima of the derivative), whereas others rely on finding the exact surface potential-gate voltage relationship by solving the Poisson equation in the channel region and equating the surface potential equal to twice the Fermi potential at threshold condition [8]. The Poisson equation simply relates the potential to the charge contained in any region and is defined as

$$\nabla^2 \varphi = -\frac{\rho}{\varepsilon_{\mathrm{Si}}} \tag{1.3}$$

Though equation (1.3) appears very simple, it is very difficult to solve analytically and requires numerical solvers or approximations as discussed in Chapter 7. The simplest approach to find the threshold voltage is the constant current method, which defines threshold voltage as the gate voltage for a particular constant drain current, generally $(W/L_g) \times 10^{-7}$ A [8]. This approach is simplest as it does not involve any modulation or extrapolation or numerical solver.

In addition, other important parameters can also be extracted from the transfer characteristics at a drain voltage of V_{DD}. The drain current corresponding to the $V_{GS} = V_{DS} = V_{DD}$ is termed as the ON-state current, and the drain current at

JWBS243-c01 JWBS243-Sahay December 21, 2018 16:46 Printer Name: Trim: 6in × 9in

$V_{GS} = 0$, $V_{DS} = V_{DD}$ is termed as the OFF-state current. Furthermore, the subthreshold swing can also be extracted from the transfer characteristics. There are two subthreshold swings for each transfer characteristics: a point subthreshold slope and an average subthreshold slope. The point subthreshold slope is simply the derivative of the transfer characteristics at a given gate voltage, whereas the average subthreshold slope is calculated by taking the mean of the point subthreshold slopes at different gate voltages ranging from the OFF-state voltage ($V_{GS} = 0$) to the threshold voltage ($V_{GS} = V_{Th}$) and is given as

$$SS_{avg} = \frac{V_{Th}}{\log{(I_D)}_{V_{Th}} - \log(I_{OFF})} \tag{1.4}$$

These are the parameters that are given as the design specification to a device designer. In general, there is a minimum ON-state current to OFF-state current ratio (I_{ON}/I_{OFF} $\sim 10^4$–10^6) and a maximum subthreshold swing, which a MOSFET must satisfy to be used in circuits.

With a background of essential physics of MOSFETs, we can now discuss the implementation of circuits with the help of MOSFET as a switch.

1.3 MOSFET CIRCUITS: THE NEED FOR COMPLEMENTARY MOS

If you remember from our discussion in Section 1.1, it is indeed these MOSFETs in the form of switches which are wired together in the integrated circuits to perform particular functions. Nearly every MOSFET in a circuit has to drive a load capacitance, which may correspond to input capacitance of other MOSFETs or an external load. Therefore, for circuit representation, we connect a load capacitance at the output terminal of the MOSFET. We chose the simplest circuit, i.e. an inverter, to give an insight into the digital circuit implementation using MOSFETs.

An inverter is a NOT gate which essentially inverts the input logic “0” into output logic “1” and vice versa. How can a MOSFET perform this action? Let us consider the case of an n-MOSFET with a load capacitance connected to the drain end (Fig. 1.4(a)), which represents the output load. If the output is initially at logic “1”, i.e. if the load capacitance is initially charged, then the voltage across this capacitor is essentially the drain voltage of the n-MOSFET. Now, if an input logic “0” (V_G = 0) is applied, since $V_{GS} < V_{Th}$ (assuming $V_{Th} \sim 0.2\ V_{DD}$), the n-MOSFET remains OFF and, hence, the output logic remains at “1”. However, if an input logic “1” (V_G = V_{DD}) is applied to the n-MOSFET, since $V_{GS} > V_{Th}$, the n-MOSFET turns on and current flows from the drain terminal to the source terminal. This current would drain the charges on the output load capacitance to the ground. This can also be viewed as discharging of the output load capacitance through the resistance of the n-MOSFET in the ON-state. Therefore, the load capacitance is discharged to the ground potential which corresponds to the output logic “0”. Therefore, the application of an input logic “0” leads to an output logic “1” and vice versa and the n-MOSFET acts like an inverter.

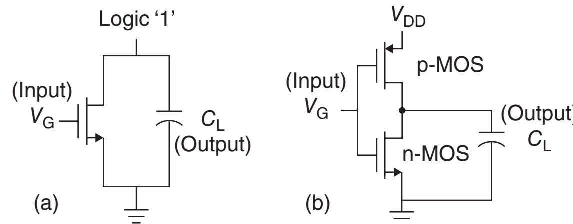

FIGURE 1.4 (a) n-MOSFET connected to a load capacitance, which is initially charged to logic "1", i.e., V_{DD} and (b) schematic view of a complementary metal-oxide-semiconductor (CMOS) inverter.

At this juncture, you may wonder what would happen in case the load capacitance was not charged initially? If the load capacitance is discharged, regardless of whether the input logic is "0" or "1" (n-MOSFET is OFF or ON), the output logic state remains at "0" because there is no means to charge the output load capacitance. Therefore, an n-MOSFET can only perform the inverter operation when the output logic is "1" and fails when the output logic is "0". Therefore, an n-MOSFET satisfies only half logic function and cannot be used alone for making even a simple inverter. Similarly, the p-MOSFET can only charge the load capacitance to logic "1" and works fine if output logic is initially "0" but fails when the output logic is initially "1". Therefore, even the p-MOSFET also satisfies only half logic function.

How can we make a complete inverter logic from MOSFETs if the n-MOSFET and p-MOSFET cannot individually perform the inverter operation? This calls for the need of complementary metal-oxide-semiconductor (CMOS) process. Since the p-MOSFET can charge the output to logic "1" whereas n-MOSFET can discharge the output to logic "0", both charging and discharging paths can be realized if they are used together as shown in Fig. 1.4(b). The p-MOSFET and n-MOSFET complement each other's logical function and can perform complete logic implementation only if used together. The circuit implementations in which both n-MOS and p-MOS are used together are known as CMOS circuits.

Now that we have recognized the importance of the CMOS process, let us analyze the working of a CMOS inverter and how a complete NOT operation is performed.

1.3.1 CMOS Inverter

The schematic of a CMOS inverter is shown in Fig. 1.4(b). When the input voltage is close to 0 V and low, for n-MOSFET, $V_{GS} < V_{Th}$ and it does not conduct. However, for the p-MOSFET, $|V_{GS}| > |V_{Th}|$. Therefore, the p-MOSFET turns ON and conducts. The load capacitance gets charged to V_{DD} via the p-MOSFET. Therefore, the output logic becomes "1". Similarly, when the input voltage becomes high, i.e. close to V_{DD}, for the n-MOSFET, $V_{GS} > V_{Th}$ and it conducts while for the p-MOSFET, $|V_{GS}| < |V_{Th}|$ and it remains switched OFF. Therefore, the load capacitance discharges to output logic "0" via the n-MOSFET. Hence, the inverter action is realized.

1.3.2 Power Dissipation in CMOS Inverter

A CMOS inverter takes in current from the supply only when both n-MOSFET and p-MOSFET are ON simultaneously, resulting in a path from the supply to the ground. Therefore, the current flows in a CMOS inverter only when the input voltage is close to 0.5 V_{DD} [6].

Ideally, we assume that the transistors do not consume any current when they are in the OFF-state. However, from our discussion in Section 1.1, we know that even below threshold voltage, the current is not equal to zero and a finite subthreshold leakage current flows through the MOSFETs. This leads to a power dissipation even when the input and output states of the CMOS inverter remain idle. This is called static power dissipation given by equation (1.1).

Also, every time the CMOS inverter output switches from “0” to “1”, the load capacitor gets charged by V_{DD}. Therefore, the charge deposited on the load capacitance is $V_{\mathrm{DD}}C_{\mathrm{L}}$. Now, the energy taken from the supply is simply $C_{\mathrm{L}}V_{\mathrm{DD}}^2$. However, when a capacitor is charged with a voltage V_{DD}, the energy stored in the capacitor is only 0.5 $C_{\mathrm{L}}V_{\mathrm{DD}}^2$. Therefore, out of the total $C_{\mathrm{L}}V_{\mathrm{DD}}^2$ energy taken from the supply, only half is stored in the load capacitance. Since the law of conservation of energy explicitly says that energy can neither be created nor be destroyed, where did the half of the energy go? Actually, the capacitor gets charged via the p-MOSFET, which acts as a resistor, and this half energy is dissipated as heat across this resistor. Now, when the inverter output switches from “1” to “0”, capacitor discharges through the n-MOSFET and the energy stored in the capacitor is dissipated through the n-MOSFET as heat. Hence, in every cycle of switching from “1” to “0” and back from “0” to “1”, a power equal to $C_{\mathrm{L}}V_{\mathrm{DD}}^2$ is dissipated in the CMOS inverter. Since power is dissipated only when the inverter switches, this power dissipation is called the dynamic power dissipation. The dynamic power dissipation may be generalized for any CMOS circuit as shown in equation (1.2).

1.4 THE NEED FOR CMOS SCALING

In Section 1.3.2, we analyzed the different power consumption mechanisms in the CMOS inverter. You may wonder what can be done to reduce the power consumption? The static power dissipation depends on the supply voltage V_{DD} and the OFF-state leakage current. Therefore, a reduction in the supply voltage or the leakage current can reduce static power dissipation. Since the digital circuits, for example, the microprocessor runs at a dramatically high frequency (~GHz range), the contribution from the dynamic power dissipation is most significant in the total power dissipation. The dynamic power dissipation depends on α, f, V_{DD}, and C_{L}. The parameter α depends on the functionality of the digital circuit and cannot be altered. The frequency of operation needs to be increased for faster computation speed. As a result, the power dissipation can only be lowered by reducing V_{DD} and C_{L}.

However, reducing the supply voltage reduces the ON-state to OFF-state current ratio ($I_{\mathrm{ON}}/I_{\mathrm{OFF}}$) due to the fundamental Boltzmann limit on subthreshold swing. For

feasible switching operation, the I_{ON}/I_{OFF} has to be at least 10^4. Therefore, the supply voltage cannot be reduced significantly.

The load capacitance C_L is essentially the input capacitance of similar logic circuits. C_L can be minimized by scaling down the area of MOSFETs. This calls for the need of CMOS scaling. Scaling the length of the MOSFET not only reduces the input capacitance but also increases the speed of the transistor as the ON-state current varies inversely with the gate length. Since the output capacitances are essentially charged and discharged, a higher drive current will increase the rate of charging or discharging. Therefore, scaling the MOSFET gate length leads to a reduced capacitance and facilitates high-frequency operation while increasing the number of MOSFETs and functionalities in a given chip area. CMOS scaling seems to be the best method to achieve a power efficient and high-speed multifunctionality device.

There are two ways in which CMOS scaling can be performed. These scaling techniques are categorized depending upon whether the supply voltage is scaled along with the channel length and width. A constant electric field scaling rule or the full scaling rule implies that the supply voltage is scaled by the same ratio as that of the length and the width. In the fixed-voltage scaling rule, the voltage is not scaled along with the length and width. The constant electric field scaling rule, which is also referred to as the Dennard's scaling rule, is followed in the industry, and the impact of the scaling factor (S) on various parameters is summarized in Table 1.1 [9].

TABLE 1.1 Scaling Factor for Different Parameters Utilizing Dennard's Scaling Rule [9]

Parameter	Scaling factor	Scaled value considering $S = \sqrt{2}$	Relative change
Channel length (L_g) Channel width (W) Gate oxide thickness (t_{OX}) Supply voltage (V_{DD})	$1/S$	0.7	30% ↓ 😊
Gate capacitance (C_{gg}) ~ $\left(\frac{WL_g}{t_{ox}}\right)$ Depletion region thickness (x_j) Intrinsic delay (τ) ~ $\left(\frac{C_{gg}V_{DD}}{I_{eff}}\right)$	$1/S$	0.7	30% ↓ 😊
Power dissipation (P_D) ~ (V_{DD}•I_{eff}) Area (A) ~ (W•L_g)	$1/S^2$	0.5	50% ↓ 😃
Power-delay product ~ (P_D• τ)	$1/S^3$	0.35	65% ↓ 😁
Electric field ~ $\left(\frac{V_{DD}}{t_{ox}}\right)$ Power density ~ $\left(\frac{P_D}{A}\right)$	1	1	0% = 😐
Frequency (f) ~ $\left(\frac{1}{\tau}\right)$	S	1.4	40% ↑ 😊

1.5 MOORE'S LAW

At this point, we would also introduce the famous law of CMOS scaling introduced by Gordon Moore, which says that the number of transistors in a chip will double after every one and a half years (18 months). The CMOS industry followed this famous Moore's law for more than 40 years with the help of CMOS scaling [10]. However, the dimensions of the scaled MOSFETs gradually became comparable to the depletion region widths at the source–channel and channel–drain interfaces. Such MOSFETs in which the channel length approaches the source–channel or channel–drain depletion region width are known as short-channel MOSFETs. The MOSFET electrostatics, which we have discussed until now, can be extended to the short-channel MOSFETs with slight modifications, which arise due to the effects that originate only when devices are scaled to the short-channel regime. These short-channel effects are discussed in Section 1.7.1.

1.6 KOOMEY'S LAW

Another parameter for estimating the efficiency of the microprocessors apart from the number of transistors in a chip is the number of computations it performs per unit power consumption. This is a more fundamental property since it relates to the energy efficiency of the microprocessors. This parameter is evaluated as the number of computations performed by the microprocessor every kilo-watt-hour of power consumed by it when it is operating at its peak output frequency.

Interestingly, because of a reduction in the power dissipation and improvement in the operating speed owing to scaling, even the number of computations performed per unit energy consumption follows the same trend as the number of transistors per chip. Even this parameter has nearly doubled every 18 months [11]. This observation was first made by Jonathan G. Koomey and hence came to be known as the Koomey's law. However, unlike the deviation from the Moore's law owing to the short-channel effects, the Koomey's law has remained intact even in the post 2010 scenario and the computing efficiency with respect to power has been doubling every 18 months. Since the physical basis for Koomey's law is more centric to today's energy-efficient computing systems including Internet of things (IoT), servers, and big data systems, it is expected to last longer than the Moore's law.

Now, in the subsequent section, we will discuss about the challenges while scaling the MOSFETs.

1.7 CHALLENGES IN SCALING THE MOSFET

1.7.1 Short-Channel Effects

In Section 1.5, we discussed that if the channel length of a MOSFET is comparable to the length of the source–channel and channel–drain depletion regions, it is termed as the short-channel MOSFET [12]. You may wonder how much exactly is the depletion

JWBS243-c01 JWBS243-Sahay December 21, 2018 16:46 Printer Name: Trim: 6in × 9in

region width at the source–channel or channel–drain interface? Typically, the doping concentration of source/drain region is $N_D = 10^{20}$ cm^{-3} and that of the channel is $N_A = 10^{16}$ cm^{-3}. The expression for depletion region width (x_{dep}) for a one-sided p–n junction with doping levels similar to MOSFET is

$$x_{dep} = \frac{2\varepsilon_{Si} V_{bi}}{qN_A} \tag{1.5}$$

$$\text{where} \quad V_{bi} = \frac{kT}{q} \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right), \tag{1.6}$$

Solving this expression, the typical depletion region width at the source–channel or channel–drain interface comes out to be ∼350 nm. As the channel lengths are scaled in this regime, several new physical phenomena arise and degrade the performance of MOSFETs. Therefore, in this section, we would give a brief overview of few short-channel effects, which dominate the performance of the MOSFETs in this ultrashort-channel length regime. Interested readers are directed to [13–15] for more detailed analysis of the short-channel effects.

1.7.1.A Threshold Voltage Roll-Off In a MOSFET, the depletion region of the heavily doped source/drain region protrudes into the p-channel and depletes it at the source–channel and channel–drain interface. As the gate length of the MOSFET is scaled to the short-channel regime, the source/drain-induced depletion region widths become a significant proportion of the overall channel length. The charge dynamics of the source/drain induced depletion region in the channel is no longer controlled solely by the "field-effect" of the gate electrode. Therefore, the effective channel charge that may be controlled by the gate electrode reduces significantly. As a result, the amount of gate voltage required to invert the channel region reduces considerably as compared to an undepleted MOS capacitor of similar dimensions. This reduction in the channel charge and the threshold voltage required to invert the channel with gate length scaling is known as threshold voltage roll-off [16, 17].

We already know that the subthreshold current, which contributes to the OFF-state leakage current, is due to diffusion of carriers from source to drain region and varies exponentially with the gate overdrive voltage, i.e., $V_{GS} - V_{Th}$. A lower V_{Th} simply means that the subthreshold leakage current would increase significantly. Therefore, the threshold voltage roll-off due to gate length scaling increases the OFF-state leakage current exponentially. It may be noted that while the increase in the ON-state current due to channel length scaling is linear, the OFF-state current increases exponentially. This leads to a significant reduction in the I_{ON}/I_{OFF} with channel length scaling.

1.7.1.B Drain-Induced Barrier Lowering As shown in Fig. 1.2, the application of a gate voltage simply modulates the source to channel barrier height and alters the

injection of electrons from source to drain region in a MOSFET. The source to channel barrier height in an ideal "long" channel MOSFET is controlled exclusively by the gate voltage. However, when the channel length is scaled to the short-channel regime, the channel–drain depletion region may interact with the source–channel depletion region. As a result, in the short-channel MOSFETs, the application of a drain voltage not only reduces the electron energy level in the drain region but the drain electric field also couples through the depletion regions and reduces the source to channel barrier height. This reduction in the source–channel barrier height with the drain voltage is not pronounced in the long-channel MOSFETs due to the absence of interaction between the depletion regions. However, in the short-channel MOSFETs, this drain-induced barrier lowering (DIBL) reduces the source to channel barrier height considerably and results in an increased leakage current [18–21]. Nowadays, one of the major challenges is to design a MOSFET with minimum drain–source coupling such that the source to channel barrier height is controlled exclusively by the gate voltage.

1.7.2 Hot Electron Effect

When the channel length is scaled, the lateral electric field estimated as V_{DS}/L_g increases significantly. You may wonder how much is the average lateral electric field in a typical MOSFET? For a MOSFET with gate length $L_g = 1\ \mu m$ and $V_{DS} = 5.0$ V, the average lateral electric field is 5 MV/m whereas for a MOSFET with $L_g = 100$ nm and $V_{DS} = 2.0$ V, the lateral electric field is 20 MV/m. As you can see, the magnitude of the electric field is very high for the short-channel MOSFETs. The channel electrons gain high momentum and kinetic energy due to this lateral field and collide with the atoms exchanging momentum and energy. The collision may result in generation of an electron–hole pair due to the impact ionization mechanism if the electrons transfer sufficient energy to the atoms.

The application of a gate voltage also creates a longitudinal field in the channel region. The impact-generated electrons are, therefore, attracted by the longitudinal electric field and may enter the oxide and knock out electrons even in the oxide region [22]. However, the energy required by these hot electrons to enter the oxide region is quite high (~3.1 eV). The hot electrons entering the oxide layer degrade the insulating capability of the oxide and lead to a current through the gate. This gate leakage current results in a reduced input impedance and is undesirable for a good transistor action. The hot electrons may accumulate inside the oxide layer as charges and affect the threshold voltage.

1.7.3 Gate-Induced Drain Leakage

For understanding the gate-induced drain leakage (GIDL), we would like to introduce the concept of tunneling at this juncture. Tunneling is a quantum-mechanical phenomenon. The fact that light not only behaves as a particle but also as a wave was established by the classical experiments of Thomas Young [23]. Electrons show a similar behavior and undergo diffraction as conceived by Davisson and

JWBS243-c01 JWBS243-Sahay December 21, 2018 16:46 Printer Name: Trim: 6in × 9in

Germer during their famous experiment, which gained them the noble prize for physics [23].

Now, as you may recall from freshman course on physics, the waves penetrate through the objects in their path unlike particles. Upon impinging on any object, some part of the wave is reflected while the remaining portion is transmitted. Depending on the object, there exists a finite probability that the wave would be transmitted through it. You would also remember that the larger the thickness of the object, the more the interaction between the wave and the object and the consequent attenuation (a decrease in the amplitude) in the output wave that comes out after penetrating through the object would also be significantly higher. Therefore, the transmission probability depends on the thickness of the object which poses itself as a barrier for the wave propagation.

For finding the exact solution for the transmission probability, one needs to solve the Schrödinger equation:

$$\left[-\frac{\hbar^2}{2m^*}\nabla^2 - q\varphi\right]\psi_i = E_i\psi_i \tag{1.7}$$

which gives the wave function (ψ_i). The probability of finding electrons is simply obtained by multiplying the wave function with its complex conjugate yielding the square of the magnitude of the wave function. The probability of finding an electron outside a potential barrier of a given height and length can be easily found by solving the Schrödinger equation in the different regions. This famous potential well problem is available in all standard text books [13–15].

Now, we shall talk about the implication of the wave nature of electrons in a reverse biased p–n junction. Let us take the case when the semiconductor is lightly doped on the p-side and the n-side. Figure 1.5(a) shows the energy band profiles of a p–n junction when the p and n sides are symmetrically doped to 10^{15} cm^{-3}. Let us

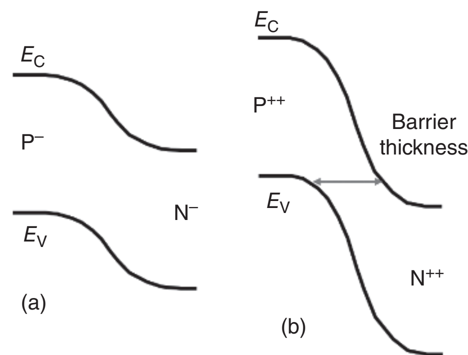

FIGURE 1.5 Energy band profile of symmetrically doped p–n junctions with (a) p and n doping concentration = 10^{15} cm^{-3} and (b) p and n doping concentration = 10^{20} cm^{-3}.

focus our attention on an electron in the valence band of the p-side. To become a free electron, jump to the conduction band and move to the n-side, the electron sees a potential barrier height equal to the band gap.

Now, we shall take the case of a heavily doped symmetrical p–n junction with doping equal to 10^{20} cm^{-3}. In this case also, let us focus our attention to an electron in the valence band on p-side. For this electron to become free and contribute to current conduction, it sees a potential barrier height equal to the band gap. However, the thickness of the potential barrier created (see Fig. 1.5(b)) is extremely thin as the depletion region width is significantly low for heavily doped p–n junctions.

Although this electron in the valence band may not get enough energy to surmount the potential barrier and move into the conduction band on the p-side, considering the wave nature of this electron, there is a finite probability that it may tunnel through the barrier since the barrier width is extremely small. The extremely thin barrier width offered by the ultrathin depletion region facilitates propagation of this electron wave directly into the conduction band on the n-side and become free. This phenomenon of transmission of electrons through potential barriers is called band-to-band tunneling (BTBT). However, BTBT is not significant in the first case since the potential barrier thickness is large. Therefore, it is not only the potential barrier height, which governs the conduction of electrons due to the drift and diffusion mechanism, but also the potential barrier width, which governs the conduction through the BTBT mechanism. Both these phenomena must be accounted for while analyzing the transport in any device.

However, the above analysis may arise a question in the minds of the readers: Will BTBT occur whenever the potential barrier is thin? To the dismay of readers, the answer is no. The quantum mechanical perspective of electron motion is quite different from the classical Newtonian perspective. According to the Bohr's theory, the electrons may occupy only specific energy “states” at discrete energy “levels.” Without getting into details of the theory, we would like to mention that the electrons traverse only when they find an empty energy state corresponding to the energy level of that particular electron. This is quite intuitive and simply states that the electrons would tunnel only when they find empty energy states at that particular energy level. In case II shown earlier, the valence band electrons of p-side can easily tunnel into the conduction band of n-side since the conduction band contains energy states which are largely empty at room temperature. However, had this band alignment been between valence band of n- and p-side, no tunneling would have taken place since the valence band is nearly filled and hardly contains any empty states.

Let us focus our attention on the GIDL phenomenon in the MOSFETs. The process of fabrication of MOSFETs induces an inherent gate-on drain overlap region. The source/drain ion implantation process is performed after gate polysilicon (gate metal) deposition in the self-aligned gate CMOS process. In addition to the lateral straggle of ion-implanted dopant atoms, the thermal annealing process to reduce the defects and activate the dopants in the heavily doped source/drain regions also leads to lateral diffusion of dopant atoms. Therefore, the dopant atoms from source and drain

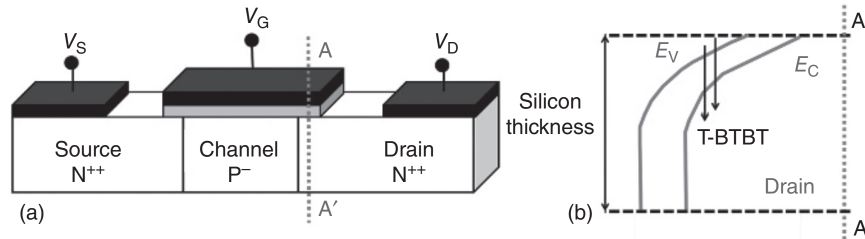

FIGURE 1.6 (a) Three-dimensional view of the fabricated MOSFET after the thermal annealing process and (b) energy band profiles along the cutline A-A′.

regions diffuse under the gate region. This leads to the gate-on source/drain overlap architecture, which exhibits a reduced channel length as compared to the defined gate length.

Now, let us consider the case of an n-MOSFET as shown in Fig. 1.6(a). When a positive drain voltage is applied and the gate voltage is negative ($V_{GS} \leq 0.0\,\mathrm{V}$), the drain region in the gate-on drain overlap area is depleted. If we analyze the band diagram along a vertical cutline as shown in Fig. 1.6(b), we can clearly observe that there is a large band bending in the drain region under the gate. Also, the band bending is sufficient to align the filled valence band of the drain region close to the surface with the empty conduction band of the drain region away from the surface. This band alignment facilitates BTBT of electrons within the drain region from the depleted surface to the bulk. This phenomenon is named transverse BTBT since this tunneling takes place in a direction perpendicular to the electron flow which is from the source to the drain. This BTBT significantly increases the leakage current of MOSFETs and is detrimental to the device performance [24–38].

Had drift-diffusion been the only mechanism governing the drain current of MOSFETs, the drain current would reduce with negative gate voltage due to the absence of carriers available in the channel region for current conduction. However, GIDL is the reason for an increase in the current in conventional MOSFETs for negative gate voltages [24–38].

Since GIDL originates due to gate-on drain overlap, it may be mitigated by properly designing the drain profile so that there is an exact alignment of the gate with the drain. Furthermore, the use of a lightly doped drain (LDD) close to the channel region may be beneficial for reduction of GIDL since the depletion region width would be larger in the LDD region [39]. A larger depletion region width would increase the tunneling width reducing the BTBT.

1.7.4 Direct Source to Drain Tunneling

In the preceding section, we discussed about GIDL in MOSFETs which arises due to BTBT. The tunneling takes place between the valence band and the conduction band in the gate-on drain overlap region. This kind of tunneling, which takes place

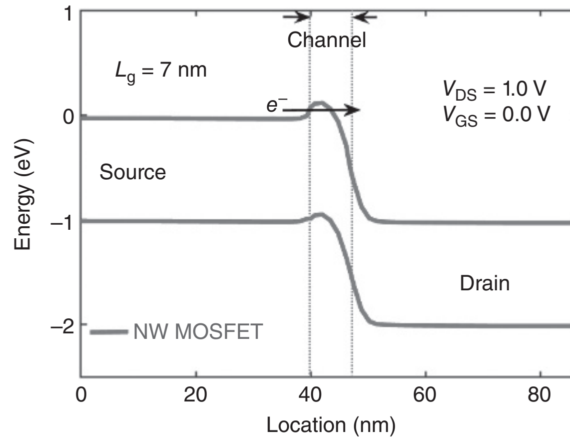

FIGURE 1.7 Energy band profile of a nanowire MOSFET with a gate length of 7 nm.

between two different energy bands (conduction to valence or vice versa), is termed as interband tunneling. However, there is yet another tunneling process that may take place between same type of bands, i.e. between conduction bands or valence bands. This tunneling phenomenon is termed as intraband tunneling.

If we carefully observe the band diagram of an ultrashort-channel ($L_g < 10$ nm) MOSFET in the OFF-state as shown in Fig. 1.7, we will notice that although the source–channel potential barrier is sufficient to prevent electrons from surmounting the barrier, the potential barrier width is significantly small owing to the ultrashort-channel length. The density of states increases with the energy level in the conduction band. However, the energy states at higher energy level in the conduction band have a negligible probability of occupancy. As a result, the conduction band electrons on the source side see empty energy states in the conduction band of the drain region. The ultrathin potential barrier, therefore, facilitates a tunneling of source conduction band electrons into the conduction band of the drain since both the conditions listed for BTBT are satisfied. This source to drain intraband tunneling is referred to as direct source to drain tunneling (DSDT) and is the most severe short-channel effect increasing the OFF-state current in ultrashort-channel ($L_g < 10$ nm) MOSFETs. The DSDT does not allow the ultrashort-channel MOSFETs to turn off. Reducing DSDT is a major challenge while designing FETs for the sub-10-nm regime [40–42].

One of the approaches proposed to mitigate DSDT is to use a different crystal orientation of silicon for active device layer. For instance, <211> silicon has a high effective carrier mass. Therefore, if <211> silicon is used in the active device layer of the MOSFETs, the tunneling probability and the consequent tunneling current would be reduced significantly. Therefore, DSDT can be mitigated by appropriately selecting the orientation or the channel material. Architectural-level designs that somehow increase the attenuation of the energy bands at the source–channel and channel–drain

interface like gate sidewall spacer or gate underlap architectures may also reduce the DSDT.

1.7.5 Boltzmann Tyranny

For any device to work as a switch, we need a very steep transition in the current from the OFF-state to the ON-state. However, the thermionic injection over the source–channel barrier height in a MOSFET imposes a fundamental limit on the rate of change of current with the applied gate voltage as discussed in the beginning of the chapter. Let us discuss the inversion layer charge below the threshold voltage to gain a better insight into this limitation.

The inversion layer charge (Q_n) can be found by integrating the inversion layer charge density (n), which is a function of position:

$$Q_n = \int n(y)dy \tag{1.8}$$

Now, $n(y)$ may be obtained from the surface potential as

$$n(y) = N_C e^{\frac{(E_f - E_C(y))}{kT}} \tag{1.9}$$

where N_C is the density of states in the conduction band. The conduction band energy may be expressed in terms of the surface potential as

$$E_C(y) = E_C(\text{bulk}) - q\varphi_S(y) \tag{1.10}$$

where φ_S is the surface potential. Utilizing equations (1.10) in (1.9), we obtain

$$n(y) = N_C e^{\frac{(E_f - E_C(\text{bulk}))}{kT}} e^{\frac{q\varphi_S(y)}{kT}} = n_B e^{\frac{q\varphi_S(y)}{kT}} \tag{1.11}$$

where n_B is the minority carrier concentration given by n_i^2/N_A.

Now, we may change the variable of integration in equation (1.8) from position to surface potential as

$$Q_n = \int n_B e^{\frac{q\varphi_S(y)}{kT}} d\varphi_S \left(\frac{dy}{d\varphi_S}\right) \tag{1.12}$$

Now, we assume that the electric field (E_S) within the inversion layer is constant because the entire inversion layer is located at the surface, we obtain:

$$Q_n = \frac{qn_B}{E_S} \int e^{\frac{q\varphi_S(y)}{kT}} d\varphi_S \tag{1.13}$$

JWBS243-c01 JWBS243-Sahay December 21, 2018 16:46 Printer Name: Trim: 6in × 9in

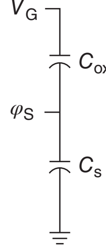

FIGURE 1.8 The effective capacitor divider network representation of the gate electrode, gate oxide, and the semiconductor channel.

Without getting into detailed mathematical analysis, we simply predict that the inversion layer charge depends exponentially on the surface potential. Now, how can we relate surface potential to the gate voltage? We introduce a simple capacitor divider principle to relate the gate voltage with the surface potential.

As shown in Fig. 1.8, the surface potential is simply the voltage that falls across the semiconductor capacitance (C_S). Therefore, the relation between φ_S and V_G can be given as

$$\varphi_S = V_G \left(\frac{C_{ox}}{C_{ox} + C_S} \right) = \frac{V_G}{m} \tag{1.14}$$

In the subthreshold regime, the semiconductor is essentially in the depletion mode. Therefore, the semiconductor capacitance can be approximated by a depletion capacitance. Using this simple relationship between φ_S and V_G, we can clearly see that the inversion layer charge density and, hence, the drain current varies exponentially with the gate voltage as

$$I_D \sim Q_n \sim e^{\frac{qV_G}{mkT}} \tag{1.15}$$

Now, we introduce the concept of the subthreshold swing, which is defined as the amount of voltage required to change the subthreshold current by one order of magnitude (one decade). For a conventional MOSFET, using equation (1.15), the subthreshold slope can be given as

$$\frac{dV_G}{d(\log_{10} I_D)} = \frac{2.3mkT}{q} \approx 60m \text{ mV/decade at } T = 300 \text{ K} \tag{1.16}$$

Now, since $m = 1 + C_S/C_{ox}$, and both C_S and C_{ox} are positive, the value of subthreshold swing is always more than 60 mV/decade. This limitation on the minimum value of the attainable subthreshold swing is called the Boltzmann tyranny. For changing the drain current by one order of magnitude, a voltage equal to 60 mV must be supplied at the gate even if C_S/C_{ox} is assumed to be negligible. This restricts a steep

transition from the OFF-state to the ON-state in a MOSFET. The lower the value of subthreshold swing, the steeper is the transition from the OFF-state to the ON-state, and the better is the switch. Therefore, most of the present-day research on MOSFETs is focused on reducing the subthreshold swing below 60 mV/decade. It may be noted that this limit arises out of the fundamental thermionic emission mechanism inherent to the MOSFETs as the current is inversely proportional to the exponential of the barrier height.

The lower limit of 60 mV/decade on the subthreshold swing of MOSFETs also hinders the scaling of their supply voltage. For effective operation of MOSFETs as a switch, the ON-state current to OFF-state current ratio should be sufficient (so that the states are easily distinguishable), at least 10^4. However, if a low supply voltage is utilized, the current swing between the ON-state and the OFF-state would be compromised owing to the limitation on the minimum attainable subthreshold swing. Therefore, the Boltzmann tyranny limits the efficiency of the conventional MOSFETs as switch at low supply voltages. To obtain high I_{ON}/I_{OFF} even at lower supply voltages which is the prime goal of the device designers, the subthreshold slope should be as low as possible. It should be close to 0 mV/decade for abrupt turning on of the FETs. Therefore, the quest for ultralow power steep subthreshold swing devices is underway.

1.7.6 Ultrasteep Doping Profile

The MOSFET contains two p–n junctions: one at the source–channel interface and the other at the channel–drain interface. The doping at the source–channel and channel–drain junction should change abruptly from a high value (typically $\sim 10^{20}$ cm^{-3}) at the source and drain regions to a low value (typically $\sim 10^{15}$ to 10^{17} cm^{-3}) with complimentary dopants in the channel region. Otherwise, the effective channel length would be significantly reduced as compared to the drawn channel length and this would increase the short-channel effects considerably. However, realizing such an ultrasteep doping profile is extremely difficult.

The ion-implantation process employed for doping the source/drain regions inherently leads to a stochastic distribution of dopant atoms (leading to lateral straggle) and creates defect centers as the dopant ions are bombarded onto the semiconductor film [43–45]. The typical value of the doping gradient between source/drain region and the channel region in MOSFETs doped using ion implantation ranges between 2 and 3 nm/decade [43–45]. This simply means that the doping can be changed, for instance, from 10^{20} cm^{-3} in the source/drain region to 10^{17} cm^{-3} in the channel region within 5–10 nm. Therefore, in any case 5–10 nm of the channel region would be consumed by the source/drain doping and unintentionally doped higher than the channel doping concentration. The effective channel length reduces by 5–10 nm from the drawn channel length due to the inability to form abrupt source/drain doping profiles.

Also, the dopant activation in the heavily doped source/drain regions requires a high-temperature annealing which in turn leads to a thermally assisted lateral diffusion of dopant atoms from source/drain regions into the channel region, minimizing

the possibility of realizing ultrasteep doping profiles [46]. This puts a complex constraint on the thermal budget as lateral diffusion is inevitable while annealing. Therefore, development of alternative doping techniques and ultrafast annealing systems is essentially required for realizing ultrasteep doping profiles.

1.8 CONCLUSION

In this chapter, we discussed the fundamentals of a MOSFET. We saw how MOSFETs work as switch and can be wired together to form circuits. We also analyzed the power dissipation in circuits made from CMOS. The need for scaling and the scaling rules were also discussed in detail. We saw how MOSFETs continued to shrink in size following the Moore's law. However, it is indeed the Koomey's law which is more fundamental. The various short-channel effects such as DIBL, threshold voltage roll-off, hot electron effects, DSDT, GIDL, and so on, which degrade the performance of the MOSFETs upon scaling were also examined carefully. A brief overview of the Boltzmann tyranny in MOSFETs was provided. This chapter lays the basis for the discussions presented in rest of the book. In the next chapter, we look at the different device architectures proposed to mitigate the challenges faced by the conventional MOSFETs. We also discuss the alternate conduction mechanisms, which may be useful to overcome the fundamental Boltzmann limit of the conventional MOSFETs.

REFERENCES

[1] Apple discussion forum [online]. Available: https://apple.stackexchange.com/questions/194367/does-the-apple-watch-have-more-processing-power-than-a-cray-2-supercomputer, Accessed Dec. 23, 2017.

[2] W. H. Brattain and B. John, "Three-electrode circuit element utilizing semi conductive materials," U.S. Patent 2524035, Oct. 1950.

[3] J. Bardeen and W. H. Brattain, "The transistor, a semiconductor triode," *Proc. IEEE*, vol. 86, no. 1, pp. 29–30, Jan. 1998.

[4] I. M. Ross, "The invention of the transistor," *Proc. IEEE*, vol. 86, no. 1, pp. 7–28, Jan. 1998.

[5] M. Riordan, L. Hoddeson, and C. Herring, "The invention of the transistor," *Rev. Mod. Phys.*, vol. 71, no. 2, pp. S336, Mar. 1999.

[6] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits*, Englewood Cliffs, NJ: Prentice-Hall, 2002.

[7] B. Jamtveit and P. Meakin, "Growth, dissolution and pattern formation in geosystems," in *Growth, Dissolution and Pattern Formation in Geosystems*, pp. 1–19, Springer, Dordrecht, the Netherlands, 1999.

[8] A. Ortiz-Conde, F. G. Sánchez, J. J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent MOSFET threshold voltage extraction methods," *Microelectron. Rel.*, vol. 42, no. 4, pp. 583–596, Apr. 2002.

[9] S. Saurabh and M. J. Kumar, *Fundamentals of Tunnel Field-Effect Transistors*, CRC Press (Taylor & Francis Group), Boca Raton, FL, 2016.

[10] Moore's law [online]. Available: https://en.wikipedia.org/wiki/Moore%27s_law#/media/File:Moore%27s_Law_Transistor_Count_1971-2016.png, Accessed Dec. 23, 2017.

[11] Intel Newsroom [online]. Available: http://download.intel.com/pressroom/pdf/computertrendsrelease.pdf, Accessed Dec. 23, 2017.

[12] A. Chaudhry and M. J. Kumar, "Controlling short-channel effects in deep submicron SOI MOSFETs for improved reliability: A review," *IEEE Trans. Dev. Mater. Rel.*, vol. 4, pp. 99–109, Mar. 2004.

[13] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York, 2013.

[14] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, John Wiley & Sons, Inc., Hoboken, NJ, 2007.

[15] B. G. Streetman and S. K. Banerjee, *Solid State Electronic Devices*, Pearson Education, 2016.

[16] P. K. Chatterjee and J. E. Leiss, "An analytic charge-sharing predictor model for submicron MOSFETs," in *Proc. IEDM*, pp. 28–33, 1980.

[17] P. K. Chatterjee, W. R. Hunter, T. C. Holloway, and Y. T. Lin, "The impact of scaling laws on the choice of n-channel or p-channel for MOS VLSI," *IEEE Electron Dev. Lett.*, vol. 1, no. 10, pp. 220–223, Oct. 1980.

[18] R. R. Troutman, "VLSI limitations from drain-induced barrier lowering", *IEEE J. Solid-State Circuits*, vol. 14, no. 2, pp. 383–391, Apr. 1979.

[19] T. Toyabe and S. Asai, "Analytical models of threshold voltage and breakdown voltage of short-channel MOSFET's derived from two-dimensional analysis," *IEEE Trans. Electron Devices*, vol. 26, no. 4, pp. 453–461, Apr. 1979.

[20] J. J. Barnes, K. Shimohigashi, and R. W. Dutton, "Short-channel MOSFETs in the punch through current mode," *IEEE J. Solid-State Circuits*, vol. 14, no. 2, pp. 368–375, Apr. 1979.

[21] R. T. Jerdonek, W. R. Bandy, and J. Birnbaum, "A model for the submicrometer n-channel deep-depletion SOS/MOSFET," *IEEE Trans. Electron Devices*, vol. 27, no. 8, pp. 1566–1570, Aug. 1980.

[22] S. Tam, F. C. Hsu, P. K. Ko, C. Hu, and R. S. Muller, "Hot-electron induced excess carriers in MOSFET's," *IEEE Electron Device Lett.*, vol. 3, no. 12, pp. 376–378, 1982.

[23] M. J. Kumar, R. Vishnoi, and P. Pandey, *Tunnel Field-effect Transistors (TFET): Modelling and Simulation*, John Wiley and Sons Ltd, West Sussex, UK, 2016.

[24] J. Fan, M. Li, X. Xu, Y. Yang, H. Xuan, and R. Huang, "Insight into gate-induced drain leakage in silicon nanowire transistors," *IEEE Trans. Electron Devices*, vol. 62, no. 1, pp. 213–219, Jan. 2015.

[25] J. Hur, B.–H. Lee, M.–H. Kang, D.–C. Ahn, T. Bang, S.–B. Jeon, and Y.–K. Choi, "Comprehensive analysis of gate-induced drain leakage in vertically stacked nanowire FETs: Inversion-mode vs. junctionless mode," *IEEE Electron Device Lett.*, vol. 37, no. 5, pp. 541–544, May 2016.

[26] S. Sahay and M. J. Kumar, "Physical insights into the nature of gate-induced drain leakage in ultrashort channel nanowire FETs," *IEEE Trans. Electron Devices*, vol. 64, no. 6, pp. 2604–2610, June 2017.

[27] S. Sahay and M. J. Kumar, “A novel gate-stack-engineered nanowire FET for scaling to the sub-10-nm regime,” *IEEE Trans. Electron Devices*, vol. 63, no. 12, pp. 5055–5059, Dec. 2016.

[28] S. Sahay and M. J. Kumar, “Spacer design guidelines for nanowire FETs from gate-induced drain leakage perspective,” *IEEE Trans. Electron Devices*, vol. 64, no. 7, pp. 3007–3015, July 2017.

[29] S. Sahay and M. J. Kumar, “Insight into lateral band-to-band-tunneling in nanowire junctionless FETs,” *IEEE Trans. Electron Devices*, vol. 63, no. 10, pp. 4138–4142, Oct. 2016.

[30] S. Sahay and M. J. Kumar, “Controlling L-BTBT and volume depletion in nanowire JLFETs using core-shell architecture,” *IEEE Trans. Electron Devices*, vol. 63, no. 9, pp. 3790–3794, Sept. 2016.

[31] S. Sahay and M. J. Kumar, “Diameter dependency of leakage current in nanowire junctionless field-effect transistors,” *IEEE Trans. Electron Devices*, vol. 64, no. 3, pp. 1330–1335, Mar. 2017.

[32] S. Sahay and M. J. Kumar, “Nanotube junctionless FET: Proposal, design, and investigation,” *IEEE Trans. Electron Devices*, vol. 64, no. 4, pp. 1851–1856, Apr. 2017.

[33] M. J. Kumar and S. Sahay, “Controlling BTBT induced parasitic BJT action in junctionless FETs using a hybrid channel,” *IEEE Trans. Electron Devices*, vol. 63, no. 8, pp. 3350–3353, Aug. 2016.

[34] S. Sahay, and M. J. Kumar, “Realizing efficient volume depletion in SOI junctionless FETs,” *IEEE J. Electron Devices Soc.*, vol. 4, no. 3, pp. 110–115, May 2016.

[35] S. Sahay and M. J. Kumar, “Symmetric operation in an extended back gate JLFET for scaling to the 65 nm regime considering quantum confinement effects,” *IEEE Trans. Electron Devices*, vol. 64, no. 1, pp. 21–27, Jan. 2017.

[36] A. K. Jain, S. Sahay, and M. J. Kumar, “Controlling L-BTBT in emerging nanotube FETs using dual-material gate,” *IEEE J. Electron Dev. Soc.*, vol. 6, pp. 611–621, June 2018.

[37] V. Nathan and N. C. Das, “Gate-induced drain leakage currents in MOS devices,” *IEEE Trans. Electron Devices*, vol. 40, no. 10, pp. 1888–1890, Oct. 1993.

[38] T. Hoffmann, G. Doornbos, I. Ferain, N. Collaert, P. Zimmerman, M. Goodwin, R. Rooyackers, A. Kottantharayil, Y. Yim, A. Dixit, K. De Meyer, M. Jurczak, and S. Biesemans, “GIDL (gate induced drain leakage) and parasitic Schottky barrier leakage elimination in aggressively scaled HfO_2/TiN FinFET devices,” in *IEDM Tech. Dig*., pp. 725–729, 2005.

[39] S. Ogura, P. J. Tsang, W. W. Walker, D. L. Critchlow, and J. F. Shepard, “Design and characteristics of the lightly doped drain–source (LDD) insulated gate field-effect transistor,” *IEEE Trans. Electron Devices*, vol. 27, no. 8, pp. 1359–1367, Aug. 1980.

[40] H. Kawaura, T. Sakamoto, and T. Baba, “Observation of source-to-drain direct tunneling current in 8 nm gate electrically variable shallow junction metal–oxide–semiconductor field-effect transistors,” *Appl. Phys. Lett.*, vol. 76, no. 25, pp. 3810–3812, June 2000.

[41] R. A. Vega and T. J. K. Liu, “Dopant-segregated Schottky source/drain double-gate MOSFET design in the direct source-to-drain tunneling regime,” *IEEE Trans. Electron Devices*, vol. 56, no. 9, pp. 2016–2026, Sept. 2009.

[42] W. S. Cho and K. Roy, “The effects of direct source-to-drain tunneling and variation in the body thickness on (100) and (110) sub-10-nm Si double-gate transistors,” *IEEE Electron Device Lett.*, vol. 36, no. 5, pp. 427–429, May 2015.

JWBS243-c01 JWBS243-Sahay December 21, 2018 16:46 Printer Name: Trim: 6in × 9in

[43] J. F. Gibbons, "Ion implantation in semiconductors—Part I: Range distribution theory and experiments," *Proc. IEEE*, vol. 56, no. 3, pp. 295–319, Mar. 1968.

[44] J. F. Gibbons, "Ion implantation in semiconductors—Part II: Damage production and annealing," *Proc. IEEE*, vol. 60, no. 9, pp. 1062–1096, Sept. 1972.

[45] S. Furukawa, H. Matsumura, and H. Ishiwara, "Theoretical considerations on lateral spread of implanted ions," *Jap. J. Appl. Phys.*, vol. 11, no. 2, pp. 134, Feb. 1972.

[46] J.-P. Colinge, C.-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White, A.-M. Kelleher, B. McCarthy, and R. Murphy, "Nanowire transistors without junctions," *Nature Nanotechnol.*, vol. 5, no. 3, pp. 225–229, Mar. 2010.