Unit I

1.1 Introduction

Design for low power has become nowadays one of the major concerns for complex, very-large-scale-integration (VLSI) circuits. Deep submicron technology, from 130 nm onwards, poses a new set of design problems related to the power consumption of the chip. Tens of millions of gates are nowadays being implemented on a relatively small die, leading to a power density and total power dissipation that are at the limits of what packaging, cooling, and other infrastructure can sup-port. As technology has shrunk to 90 nm and below, the leakage current has increased dramatically, and in some 65-nm designs, leakage power is nearly as large as dynamic power. So it is becoming impossible to increase the clock speed of high-performance chips as technology shrinks and the chip density increases, because the peak power consumption of these chips is already at the limit and cannot be increased further. Also, the power density leads to reliability problems because the mean time to failure decreases with temperature. Besides, the timing degrades and the leakage currents increase with temperature. For battery-powered devices also, this high on-chip power density has become a significant problem, and techniques are being used in these devices from software to architecture to implementation level to alleviate this problem as much as possible like power gating and multi-threshold libraries. Some other techniques being used nowadays are using different supply voltages at different blocks of the design according to the performance requirements, or voltage scaling techniques.

1.2 Historical Background:

The invention of transistor by William Shockley and his colleagues at Bell Laboratories, Murray Hills, NJ, ushered in the “solid state” era of electronic circuits and systems. Within few years after the invention, transistors were commercially available and almost all electronic systems started carrying the symbol “solid state,” signifying the conquest of the transistor over its rival—the vacuum tube. Smaller size, lower power consumption, and higher reliability were some of the reasons that made it a winner over the vacuum tube. About a decade later, Shockley and his colleagues, John Bardeen and Walter Brattain, of Bell Laboratories were rewarded with a Nobel Prize for their revolutionary invention.

The tremendous success of the transistor led to vigorous research activity in the field of microelectronics. Later, Shockley founded a semiconductor industry. Some of his colleagues joined him or founded semiconductor industries of their own. Gordon Moore, a member of Shockley’s team, founded Fairchild and later Intel. Re-search engineers of Fairchild developed the first planner transistor in the late 1950s, which was the key to the development of integrated circuits (ICs) in 1959. Planner technology allowed realization of a complete electronic circuit having a number of devices and interconnecting them on a single silicon wafer. Within few years of the development of ICs, Gordon Moore, predicted that by 1975, it would be possible to cram as many as 65,000 components onto a single silicon chip of about one fourth of a square inch.

Later in the year 1975, Moore revisited that his 10-year-old forecast of 65,000 components was on the mark. However, he revised his prediction rate from 1 year to 18 months, that is, the component density would double every 18 months. This became known as *Moore’s law*.

Moore’s law acted as a driving force for the spectacular development of IC technology leading to different types of products. Based on the scale of integration, the IC technology can be divided into five different categories, as summarized in Table 1.1. The first half of the 1960s was the era of small-scale integration (SSI), with about ten planner transistors on a chip.

The SSI technology leads to fabrication of gates and flip-flops. In the second half of the 1960s, counters, multiplexers, decoders, and adder were fabricated using the medium-scale integration (MSI) technology having 100–1000 components on a chip. The 1970s was the era of large-scale integration (LSI) technology with 10,000–20,000 components on a chip producing typical products like 8-bit microprocessor, RAM, and read-only memories (ROM). In the 1980s, VLSI with about 20,000–50,000 components led to the development of 16-bit and 32-bit microprocessors. Beyond 1985 is the era of ultra-large-scale

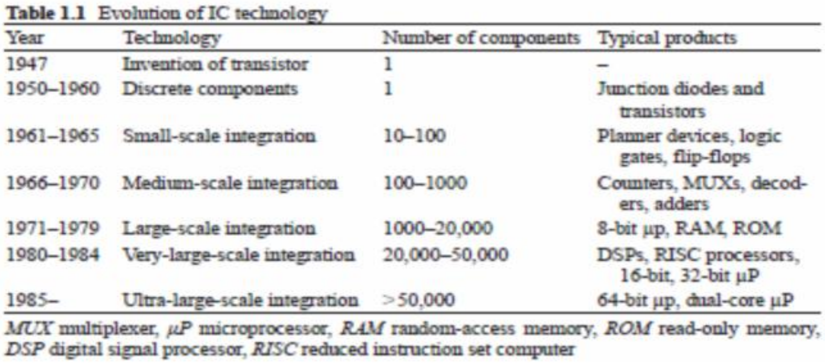


Table 1.1 Evolution of IC technology

Year	Technology	Number of components	Typical products
1947	Invention of transistor	1	–
1950–1960	Discrete components	1	Junction diodes and transistors
1961–1965	Small-scale integration	10–100	Planner devices, logic gates, flip-flops
1966–1970	Medium-scale integration	100–1000	Counters, MUXs, decoders, adders
1971–1979	Large-scale integration	1000–20,000	8-bit μp, RAM, ROM
1980–1984	Very-large-scale integration	20,000–50,000	DSPs, RISC processors, 16-bit, 32-bit μP
1985–	Ultra-large-scale integration	>50,000	64-bit μp, dual-core μP

MUX multiplexer, *μP* microprocessor, *RAM* random-access memory, *ROM* read-only memory, *DSP* digital signal processor, *RISC* reduced instruction set computer integration (ULSI) with more than 50,000 devices on a chip, which led to the fabrication of digital signal processors (DSPs), reduced instruction set computer (RISC) processor, etc.

In 1971, Intel marketed an IC with the capability of a general-purpose building block of digital systems. It contained all the functionality of the central processing unit (CPU) of a computer. The chip was code named as 4004. It was a 4-bit CPU. Later on, this device was given the name “microprocessor.” Thus, the micropro-cessor—“the CPU on a chip”—was born. The huge success of this chip led to the development of 8008 and 8085, the most popular 8-bit microprocessors, by Intel. In the past three decades, the evolution tree of microprocessors has grown into a large tree with three main branches as shown in Fig. 1.2.

The main branch in the middle represents the general-purpose microprocessors, which are used to build computers of different kinds such as laptops, desktops, workstations, servers, etc. The fruits of this branch have produced more and more powerful CPUs with processing capability of increased number of bits starting from 4-bit processors to the present-day 64-bit processors. Moreover, the clock rates increased from few megahertz to thousands of megahertz, and many advanced architectural features such as pipelining, superscalar, on-chip cache memory, dual core, etc. Computers built using the present-day microprocessors have the capability of mainframe computers of the 1980s and 1990s. Figure 1.3 shows the series of microprocessors produced by Intel in the past three-and-a-half decades conform-ing to Moore’s law very closely. It may be noted that the first microprocessor had only 2200 transistors and the latest microprocessors are having more than a billion transistors.

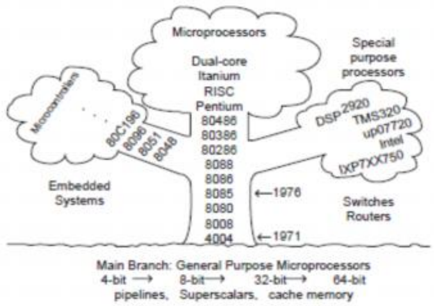


Fig. 1.2 Evolution tree of microprocessor. *RISC* reduced instruction set computer, *DSP* digital signal processor

The left branch represents a new breed of processors, known as microcontrollers. A microcontroller can be considered a “computer on a chip.” Apart from the CPU, other subsystems such as ROM, RAM, input/output (I/O) ports, timer, and serial port are housed on a single chip in a microcontroller. The CPUs of the microcontroller are usually not as powerful as general-purpose microprocessors. Microcontrollers are typically used to realize embedded systems such as toys, home appliances, intelligent test and measurement equipment, etc.

The branch on the right side represents special-purpose processors such as DSP microprocessors (TMS 320), network processors (Intel PXA 210/215), etc. These special-purpose processors are designed to enhance performance of special applications such as signal processing, router and packet-level processing in communication equipment, etc.

With the increase in the number of transistors, the power dissipation also kept on increasing as shown in Fig. 1.4. This forced the chip designers to consider low power as one of the design parameters apart from performance and area. In the following section, we shall focus on the importance of low power in IC design.

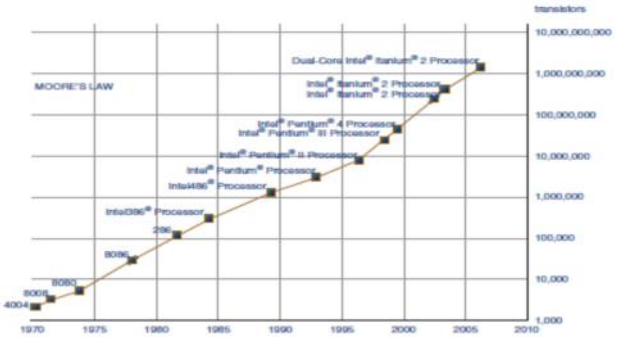


Fig. 1.3 Moore's law and the Intel microprocessors. (Source: Intel)

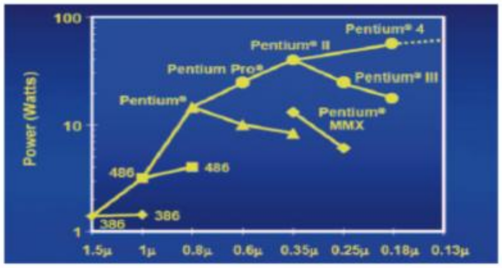
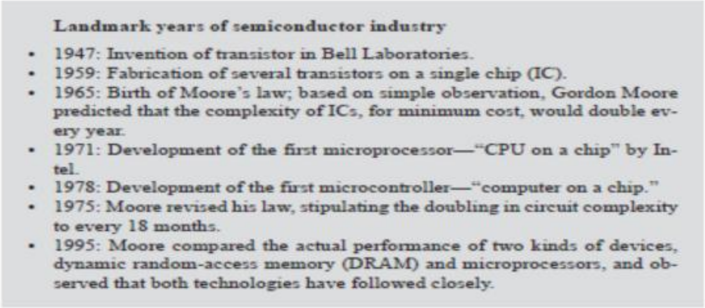


Fig. 1.4 Power dissipation of Intel processors. (Source: Intel)



Landmark years of semiconductor industry

- 1947: Invention of transistor in Bell Laboratories.
- 1959: Fabrication of several transistors on a single chip (IC).
- 1965: Birth of Moore's law; based on simple observation, Gordon Moore predicted that the complexity of ICs, for minimum cost, would double every year.
- 1971: Development of the first microprocessor—"CPU on a chip" by Intel.
- 1978: Development of the first microcontroller—"computer on a chip."
- 1975: Moore revised his law, stipulating the doubling in circuit complexity to every 18 months.
- 1995: Moore compared the actual performance of two kinds of devices, dynamic random-access memory (DRAM) and microprocessors, and observed that both technologies have followed closely.

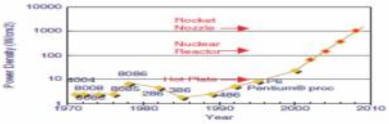
Why Low Power?

Until recently, performance of a processor has been synonymous with circuit speed or processing power, e.g., million instructions per second (MIPS) or million float-ing point operations per second (MFLOPS). Power consumption was of secondary concern in designing ICs. However, in nanometer technology, power has become the most important issue because of:

- Increasing transistor count
- Higher speed of operation
- Greater device leakage currents

Increased process parameter variability due to aggressive device size scaling has created problems in yield, reliability, and testing. As a consequence, there is a change in the trend of specifying the performance of a processor. Power consumption is now considered one of the most important design parameters. Among various reasons for this change in trend, some important reasons are considered below.

Fig. 1.5 Increasing power density of the very-large-scale-integration (VLSI) chip. (Source: Intel)



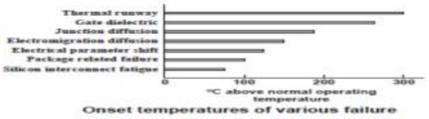


Fig. 1.6 Different failure mechanisms against temperature

In order to continuously improve the performance of the circuits and to integrate more and more functionality in the chip, the device feature size has to continuously shrink. Figure 1.4 shows the power dissipation of Intel processors. As a consequence, the magnitude of power per unit area known as power density is increasing as shown in Fig. 1.5. To make a chip commercially viable, it is necessary to reduce the cost of packaging and cooling, which in turn demands lower power consumption.

Increased customer demand has resulted in proliferation of hand-held, battery-operated devices such as cell phone, personal digital assistant (PDA), palmtop, laptop, etc. As these devices are battery operated, battery life is of primary concern. Unfortunately, the battery technology has not kept up with the energy requirement of the portable equipment. Commercial success of these products depends on size, weight, cost, computing power, and above all on battery life. Lower power consumption is essential to make these products commercially viable.

It has been observed that reliability is closely related to the power consumption of a device. As power dissipation increases, the failure rate of the device increases because temperature-related failures start occurring with the increase in tempera-ture as shown in Fig. 1.6. It has been found that every 10 ºC rise in temperature roughly doubles the failure rate. So, lower power dissipation of a device is essential for reliable operation.

Although power and energy are used interchangeably in many situations, these two have different meanings and it is essential to understand the difference be-tween the two, especially in the case of battery-operated devices. Figure 1.7 illustrates the difference between the two. Power is the instantaneous power in the device, while energy is the integration of power with time. For example, in Fig. 1.7, we can see that approach 1 takes less time but consumes more power than approach 2. But the energy consumed by the two, that is, the area under the curve for both the approaches is the same, and the battery life is primarily determined by this energy consumed.

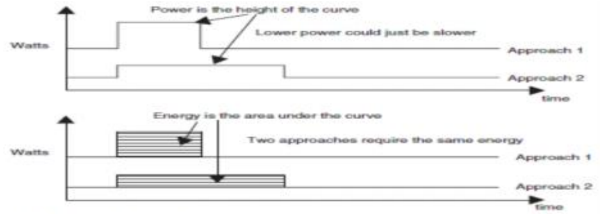

Power is the height of the curve
Lower power could just be slower
Watts
Approach 1
Approach 2
time
Energy is the area under the curve
Two approaches require the same energy
Watts
Approach 1
Approach 2
time

Fig. 1.7 Power versus energy

Sources of Power Dissipations

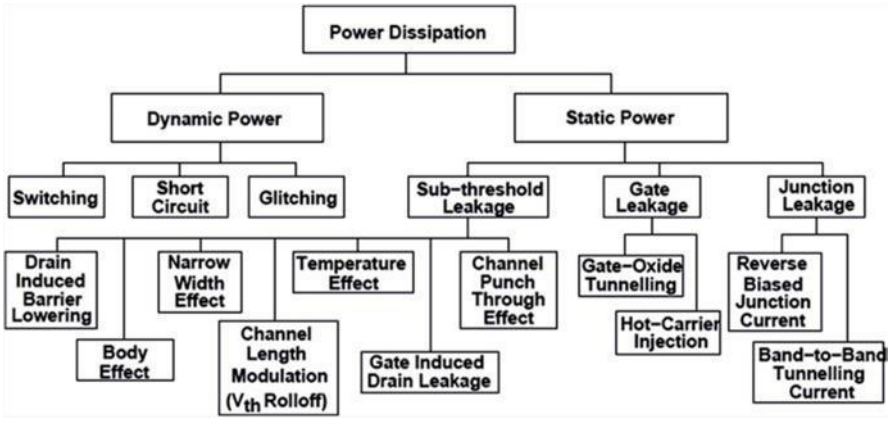


Fig. 1.8 Types of power dissipation

Power dissipation is measured commonly in terms of two types of metrics:

1. *Peak power*: Peak power consumed by a particular device is the highest amount of power it can consume at any time. The high value of peak power is generally related to failures like melting of some interconnections and power-line glitches.

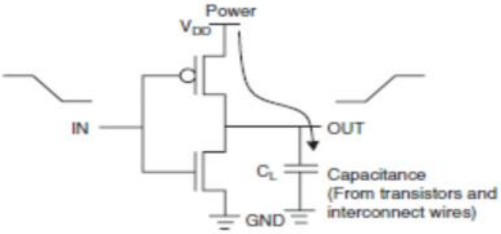
2. *Average power*: Average power consumed by a device is the mean of the amount of power it consumes over a time period. High values of average power lead to problems in packaging and cooling of VLSI chips.

In order to develop techniques for minimizing power dissipation, it is essential to identify various sources of power dissipation and different parameters involved in each of them. The total power for a VLSI circuit consists of dynamic power and static power. Dynamic power is the power consumed when the device is active, that is, when signals are changing values. Static power is the power consumed when the device is powered up but no signals are changing value. In CMOS devices, the static power consumption is due to leakage mechanism. Various components of power dissipation in CMOS devices can therefore be categorized as shown in Fig. 1.8.

Dynamic Power

Dynamic power is the power consumed when the device is active, that is, when the signals of the design are changing values. It is generally categorized into three types: switching power, short-circuit power, and glitching power, each of which will be discussed in details below.

Fig. 1.9 Dynamic (switching) power. *GND* ground



1.4.1.1 Switching Power

The first and primary source of dynamic power consumption is the switching pow-er, the power required to charge and discharge the output capacitance on a gate. Figure 1.9 illustrates switching power for charging a capacitor.

The energy per transition is given by

$$\text{Energy/transition} = \frac{1}{2} \times C_L \times V_{dd}^2$$

where C_L is the load capacitance and V_{dd} is the supply voltage. Switching power is therefore expressed as:

$$P_{switch} = \text{Energy}/\text{transition} \times f = C_L \times V_{dd}^2 \times P_{trans} \times f_{clock} ,$$

where f is the frequency of transitions, P_{trans} is the probability of an output transi-tion, and f_{clock} is the frequency of the system clock. Now if we take:

$$C_{switch} = P_{trans} \times C_L ,$$

then, we can also describe the dynamic power with the more familiar expression:

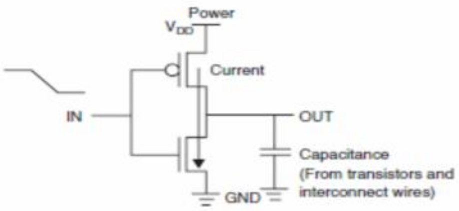
$$P_{switch} = C_{eff} \times V_{dd}^2 \times f_{clock} .$$

Switching power is not a function of transistor size, but it is dependent on switching activity and load capacitance. Thus, it is data dependent.

In addition to the switching power dissipation for charging and discharging the load capacitance, switching power dissipation also occurs for charging and dis-charging of the internal node capacitance. Thus, total switching power dissipation is given by

$$P_{\text{totalswitch}} = P_{\text{trans}} C_{\text{L}} \times V_{\text{dd}}^{2} \times f_{\text{clock}} + \sum \alpha_{\text{i}} \times C_{\text{i}} \times V_{\text{dd}} \times (V_{\text{dd}} - V_{\text{th}}) \times f_{\text{clock}} ,$$

Fig. 1.10 Short-circuit current or crowbar current. *GND* ground



where α_i and C_i are the transition probability and capacitance, respectively, for an internal node i.

Short-Circuit Power

In addition to the switching power, short-circuit power also contributes to the dynamic power. Figure 1.10 illustrates short-circuit currents. Short-circuit currents occur when both the negative metal–oxide–semiconductor (NMOS) and positive metal–oxide–semiconductor (PMOS) transistors are on. Let V_{tn} be the threshold voltage of the NMOS transistor and V_{tp} is the threshold voltage of the PMOS transistor. Then, in the period when the voltage value is between V_{tn} and V_{dd}–V_{tp}, while the input is switching either from 1 to 0 or vice versa, both the PMOS and the NMOS transistors remain ON, and the short-circuit current follows from V_{dd} to ground (GND).

The expression for short-circuit power is given by

$$P_{\text{shortcircuit}} = t_{\text{sc}} \times V_{\text{dd}} \times I_{\text{peak}} \times f_{\text{clock}} = \frac{\mu \cdot \varepsilon_{\text{ox}} \cdot W}{12LD} \times (V_{\text{dd}} - V_{\text{th}})^{3} \times t_{\text{sc}} \times f_{\text{clock}} ,$$

where t_{sc} is the rise/fall time duration of the short-circuit current, I_{peak} is the total in-ternal switching current (short-circuit current plus the current to charge the internal capacitance), μ is the mobility of the charge carrier, $\varepsilon_{ox}$is the permittivity of the silicon dioxide, W is the width, L is the length, and *D* is the thickness of the silicon dioxide.

From the above equation it is evident that the short-circuit power dissipation depends on the supply voltage, rise/fall time of the input and the clock frequency apart from the physical parameters. So the short-circuit power can be kept low if the ramp (rise/fall) time of the input signal is short for each transition. Then the overall dynamic power is determined by the switching power.

Glitching Power Dissipation

The third type of dynamic power dissipation is the glitching power which arises due to finite delay of the gates. Since the dynamic power is directly proportional to the number of output transitions of a logic gate, glitching can be a significant source of signal activity and deserves mention here. Glitches often occur when paths with unequal propagation delays converge at the same point in the circuit. Glitches oc-cur because the input signals to a particular logic block arrive at different times, causing a number of intermediate transitions to occur before the output of the logic block stabilizes. These additional transitions result in power dissipation, which is categorized as the glitching power.

1.4.2 Static Power

Static power dissipation takes place as long as the device is powered on, even when there are no signal changes. Normally in CMOS circuits, in the steady state, there is no direct path from V_{dd} to GND and so there should be no static power dissipation, but there are various leakage current mechanisms which are responsible for static power dissipation. Since the MOS transistors are not perfect switches, there will be leakage currents and substrate injection currents, which will give rise to static pow-er dissipation in CMOS. Since the substrate current reaches its maximum for gate voltages near $0.4V_{\mathrm{dd}}$ and gate voltages are only transiently in this range when the devices switch, the actual power contribution of substrate currents is negligible as compared to other sources of power dissipation. Leakage currents are also normally negligible, in the order of nano-amps, compared to dynamic power dissipation. But with deep submicron technologies, the leakage currents are increasing drastically to the extent that in 90-nm technology and thereby leakage power also has become comparable to dynamic power dissipation.

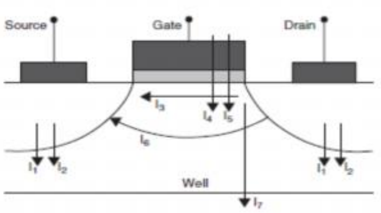


Fig. 1.11 Leakage currents in an MOS transistor. *MOS* metal–oxide–semiconductor [5]

Figure 1.11 shows several leakage mechanisms that are responsible for static power dissipation. Here, I1 is the reverse-bias p–n junction diode leakage current, I2 is the reverse-biased p–n junction current due to tunneling of electrons from the valence band of the p region to the conduction band of the n region, I3 is the subthreshold leakage current between the source and the drain when the gate volt-age is less than the threshold voltage (V_{th}), I4 is the oxide tunneling current due to reduction in the oxide thickness, I5 is the gate current due to hot carrier injection of electrons (I4 and I5 are commonly known as IGATE leakage current), I6 is the gate-induced drain leakage current due to high field effect in the drain junction, and I7 is the channel punch through current due to close proximity of the drain and the source in short-channel devices.

These are generally categorized into four major types: sub-threshold leakage, gate leakage, gate-induced drain leakage, and junction leakage as shown in Fig. 1.12. Apart from these four primary leakages, there are few other leakage currents which also contribute to static power dissipation, namely, reverse-bias p–n junction diode leakage current, hot carrier injection gate current, and channel punch through cur-rent.

Low-Power Design Methodologies

Low-power design methodology needs to be applied throughout the design process starting from system level to physical or device level to get effective reduction of power dissipation in digital circuits based on MOS technology [2–4]. Various ap-proaches can be used at different level of design hierarchy. Before venturing to do this, it is essential to understand the basics of MOS circuits and the way these are fabricated. So, we have started with fabrication technology in Chap. 2. The sub-sequent three chapters introduce MOS transistor, followed by MOS inverters, and then complex MOS combinational circuits. Chapter 6 introduces various sources of power dissipation in details. As the most dominant component has quadratic dependence and other components have linear dependence on the supply voltage, reducing the supply voltage is the most effective means to reduce dynamic power consumption. Unfortunately, this reduction in power dissipation comes at the expense of performance. It is essential to devise suitable mechanism to contain this loss in performance due to supply voltage scaling for the realization of low-power high-performance circuits. The loss in performance can be compensated by using suitable techniques at the different levels of design hierarchy; that is physical level, logic level, architectural level, and system level. Techniques like device feature size scaling, parallelism and pipelining, architectural-level transformations, dynamic voltage, and frequency scaling.

Apart from scaling the supply voltage to reduce dynamic power, another alter-native approach is to minimize the switched capacitance comprising the intrinsic capacitances and switching activity. Choosing which functions to implement in hardware and which in software is a major engineering challenge that involves is-sues such as cost complexity, performance, and power consumption. From the be-havioral description, it is necessary to perform hardware/software partitioning in a judicious manner such that the area, cost, performance, and power requirements are satisfied. Transmeta's Crusoe processor is an interesting example that demonstrated that processors of high performance with remarkably low power consumption can be implemented as hardware–software hybrids. The approach is fundamentally software based, which replaces complex hardware with software, thereby achieving large power savings.

In CMOS digital circuits, the switching activity can be reduced by algorithmic optimization, by architectural optimization, by use of suitable logic-style or by log-ic-level optimization. The intrinsic capacitances of system-level busses are usually several orders of magnitude larger than that for the internal nodes of a circuit. As a consequence, a considerable amount of power is dissipated for transmission of data over I/O pins. It is possible to save a significant amount of power reducing the number of transactions, i.e., the switching activity, at the processors I/O interface. One possible approach for reducing the switching activity is to use suitable encod-ing of the data before sending over the I/O interface. The concept is also applicable in the context of multi-core system-on-a-chip (SOC) design. In many situations the switching activity can be reduced by using the sign-magnitude representation in place of the conventional two's complement representation. Switching activity can be reduced by judicious use of clock gating, leading to considerable reduction in dynamic power dissipation. Instead of using static CMOS logic style, one can use other logic styles such as pass-transistor and dynamic CMOS logic styles or a suitable combination of pass-transistor and static CMOS logic styles to minimize energy drawn from the supply.

Although the reduction in supply voltage and gate capacitances with device size scaling has led to the reduction in dynamic power dissipation, the leakage power dissipation has increased at an alarming rate because of the reduction of threshold voltage to maintain performance. As the technology is scaling down from submi-cron to nanometer, the leakage power is becoming a dominant component of total power dissipation. This has led to vigorous research for the reduction of leakage power dissipation. Leakage reduction methodologies can be broadly classified into two categories, depending on whether it reduces *standby* leakage or *runtime* leakage. There are various standby leakage reduction techniques such as input vec-tor control (IVC), body bias control (BBC), multi-threshold CMOS (MTCMOS), etc. and runtime leakage reduction techniques such as static dual threshold voltage CMOS (DTCMOS) technique, adaptive body biasing, dynamic voltage scaling, etc.

Aggressive device size scaling used to achieve high performance leads to in-creased variability due to short-channel and other effects. Performance parameters such as *power* and *delay* are significantly affected due to the variations in process parameters and environmental/operational (V_{dd}, temperature, input values, conditions. For designs, due to variability, the design methodology in the future nanometer VLSI circuit designs will essentially require a paradigm shift from deterministic to probabilistic and statistical design approach. The impact of process variations has been investigated and several techniques have been proposed to optimize the performance and power in the presence of process variations

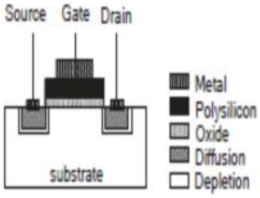
MOS Transistors

Introduction:-

The base semiconductor material used for the fabrication of metal–oxide–semiconductor (MOS) integrated circuits is silicon. *Metal*, *oxide*, and *semiconductor* form the basic structure of MOS transistors. MOS transistors are realized on a single crystal of silicon by creating three types of conducting materials separated by intervening layers of an insulating material to form a sandwich-like structure. The three conducting materials are: *metal*, *poly-silicon*, and *diffusion*. Aluminum as metal and polycrystalline silicon or poly-silicon are used for interconnecting different elements of a circuit. The insulating layer is made up of silicon dioxide (SiO_2). Patterned layers of the conducting materials are created by a series of photolithographic techniques and chemical processes involving oxidation of silicon, diffusion of impurities into the silicon and deposition, and etching of aluminum on the silicon to provide interconnection.

Structure of MOS Transistors

Fig. 3.1 Structure of an MOS transistor



The structure of an MOS transistor is shown in Fig. 3.1. On a lightly doped sub-strate of silicon, two islands of diffusion regions of opposite polarity of that of the substrate are created. These two regions are called *source* and *drain,* which are connected via metal (or poly-silicon) to the other parts of the circuit. Between these two regions, a thin insulating layer of silicon dioxide is formed, and on top of this a conducting material made of poly-silicon or metal called *gate* is deposited. There are two possible alternatives. The substrate can be lightly doped by either a p-type or an n-type material, leading to two different types of transistors. When the sub-strate is lightly doped by a p-type material, the two diffusion regions are strongly doped by an n-type material. In this case, the transistor thus formed is called an *nMOS transistor*. On the other hand, when the substrate is lightly doped by an n-type material, and the diffusion regions are strongly doped by a p-type material, a *pMOS transistor* is created.

The region between the two diffusion islands under the oxide layer is called the *channel* region. The operation of an MOS transistor is based on the controlled flow of current between the source and drain through the channel region. Based on the channel current to flow and control it, There are two possible ways to achieve this, which have resulted in *enhancement*- and *depletion-mode* transistors. After fabrication, the structure of an enhancement-mode nMOS transistor looks like Fig. 3.2a. In this case, there is no conducting path in the channel region for the situation $V_{gs} = 0$ V, that is when no voltage is applied to the gate with respect to the source. If the gate is connected to a suitable positive voltage with respect to the source, then the electric field established between the gate and the substrate gives rise to a *charge inversion* region in the substrate under the gate insulation, and a conducting path is formed be-tween the source and drain. Current can flow between the source and drain through this conducting path.

By implanting suitable impurities in the channel region during fabrication, prior to depositing the insulation and the gate, the conducting path may also be established in the channel region even under the condition $V_{gs} = 0$ V. This situation is shown in

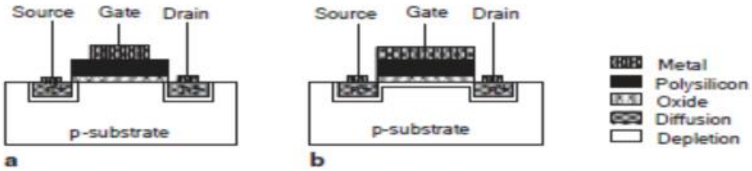
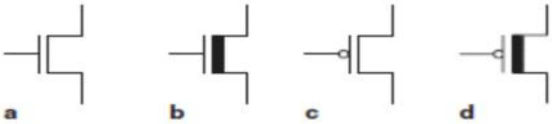


Fig. 3.2 **a** nMOS enhancement-mode transistor. **b** nMOS depletion-mode transistor

Fig. 3.3 **a** nMOS enhancement. **b** nMOS depletion. **c** pMOS enhancement. **d** pMOS depletion-mode transistors



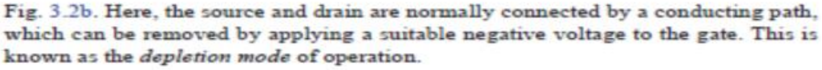


Fig. 3.2b. Here, the source and drain are normally connected by a conducting path, which can be removed by applying a suitable negative voltage to the gate. This is known as the *depletion mode* of operation.

The Fluid Model

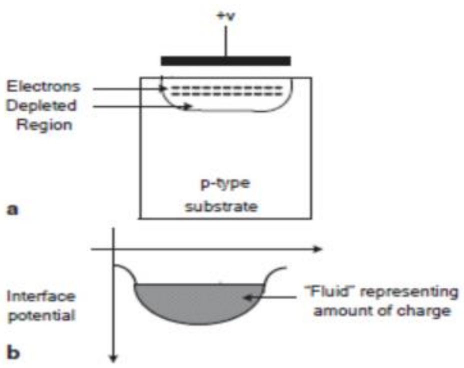


Fig. 3.4 **a** An MOS capacitor. **b** The fluid model

The *Fluid model* [1] is one such tool, which can be used to visualize the behavior of charge-controlled devices such as MOS transistors, charge-coupled devices (CCDs), and bucket-brigade devices (BBDs). Using this model, even a novice can understand the operation of these devices.

The model is based on two simple ideas: (a) Electrical charge is considered as fluid, which can move from one place to another depending on the difference in their level, of one from the other, just like a fluid and (b) electrical potentials can be mapped into the geometry of a container, in which the fluid can move around. Based on this idea, first, we shall consider the operation of a simple MOS capacitor followed by the operation of an MOS transistor.

The MOS Capacitor

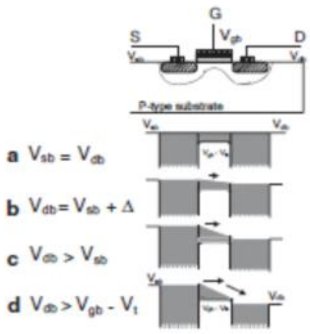
From the knowledge of basic physics, we know that a simple parallel-plate capacitor can be formed with the help of two identical metal plates separated by an insulator. An MOS capacitor is realized by sandwiching a thin oxide layer between a metal or poly-silicon plate on a silicon substrate of suitable type as shown in Fig 3.4a. As we know, in case of parallel-plate capacitor, if a positive voltage is applied to one of the plates, it induces a negative charge on the lower plate. Here, if a positive voltage is applied to the metal or poly-silicon plate, it will repel the majority carriers of the p-type substrate creating a depletion region. Gradually, minority carriers (electrons) are generated by some physical process, such as heat or incident light, or it can be injected into this region. These minority carriers will be accumulated underneath the MOS electrode, just like a parallel-plate capacitor. Based on the fluid model, the MOS electrode generates a pocket in the form of a surface potential in the silicon substrate, which can be visualized as a container. The shape of the container is defined by the potential along the silicon surface. The higher the potential, the deeper is the container, and more charge can be stored in it. However, the minority carriers present in that region create an inversion layer. This changes the surface potential; increase in the quantity of charge decreases the positive sur-face potential under the MOS electrode. In the presence of inversion charge, the surface potential is shown in Fig. 3.4b by the solid line. The area between the solid line and the dashed line shows not only the presence of charge but also the amount of charge. The capacity of the bucket is finite and depends on the applied electrode voltage. Here, it is shown that the charge is sitting at the bottom of the container just as a fluid would stay in a bucket. In practice, however, the minority carriers in the inversion layer actually reside directly at the silicon surface. The surface of the fluid must be level in the equilibrium condition. If it were not, electrons would move under the influence of potential difference until a constant surface potential is established. From this simple model, we may conclude that the amount of charge accumulated in an MOS capacitor is proportional to the voltage applied between the plates and the area between the plates.

The MOS Transistor Fluid Model

By adding diffusion regions on either side of an MOS capacitor, an MOS transistor is realized. One of the diffusion regions will form the *source* and the other one will form the *drain*. The capacitor electrode acts as the gate. The cross-sectional view of an MOS transistor is shown in Fig. 3.5a. We can use the fluid model to explain the behavior of MOS transistors.

To start with, we may assume that the same voltage is applied to both the source and drain terminals ($V_{db} = V_{sb}$) with respect to the substrate. This defines the poten-tial of these two regions. In the potential plot, the diffusion regions (where there is plentiful of charge carriers) can be represented by very deep wells, which are filled with charge carriers up to the levels of the potentials of the source and drain regions. The potential underneath the MOS gate electrode determines whether the two wells are connected or separated. The potential in the channel region can be controlled with the help of the gate voltage. The potential at the channel region is shown by the dotted lines of Fig. 3.5b. The dotted line 1 corresponding to $V_{gb} = 0$ is above the drain and source potentials. As the gate voltage is gradually increased, more and more holes are repelled from the channel region, and the potential at the channel region moves downward as shown by the dotted lines 2, 3, etc. In this situation, the source and drain wells are effectively isolated from each other, and no charge can move from one well to the other. A point is reached when the potential level at the gate region is the same as that of the source and diffusion regions. At this point, the channel region is completely devoid of holes. The gate voltage at which this happens is called the threshold voltage (V_t) of the MOS transistor. If the gate voltage is increased further, there is an accumulation of electrons beneath the SiO_2 layer in the channel region, forming an *inversion layer*. As the gate voltage is increased further, the potential at the gate region moves below the source and drain potentials as shown by the dotted lines 3 and 4 in Fig. 3.5b. As a consequence, the barrier between the two regions disappears and the charge from the source and drain regions spills underneath the gate electrode leading to a uniform surface potential in the entire region. By varying the gate voltage, the thickness of the inversion layer can be controlled, which in turn will control the conductivity of the channel as visualized in Fig. 3.5b. Under the control of the gate voltage, the region under it acts as a movable barrier that controls the flow of charge between the source and drain areas.

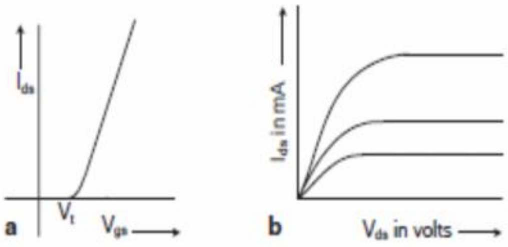
Fig. 3.6 The fluid model of an MOS transistor



we can say that an MOS transistor acts as a voltage-controlled device. The device first conducts when the effective gate voltage ($V_{gb}-V_t$) is more than the source voltage. The conduction characteristic is represented in Fig. 3.7a. On the other hand, as the drain voltage is increased with respect to the source, the current increases until $V_{db} = (V_{gb}-V_t)$. For drain voltage $V_{db} > (V_{gb}-V_t)$, the channel becomes pinched off, and there is no further increase in current. A plot of the drain current with respect to the drain voltage for different gate voltages is shown in Fig. 3.7b.

Modes of Operation of MOS Transistors

Fig. 3.7 **a** Variation of drain current with gate voltage. **b** Voltage–current characteristics



After having some insight about the operation of an MOS transistor, let us now have a look at the charge distribution under the gate region under different operating conditions of the transistor. When the gate voltage is very small and much less than the threshold voltage, Fig. 3.8a shows the distribution of the mobile holes in a p-type substrate. In this condition, the device is said to be in the *accumulation mode*. As the gate voltage is increased, the holes are repelled from the SiO_2–substrate interface and a depletion region is created under the gate when the gate voltage is equal to the threshold voltage. In this condition, the device is said to be in *depletion mode* as shown in Fig. 3.8b. As the gate voltage is increased further above the threshold voltage, electrons are attracted to the region under the gate creating a conducting layer in the p substrate as shown in Fig. 3.8c. The transistor is now said to be in *inversion mode*.

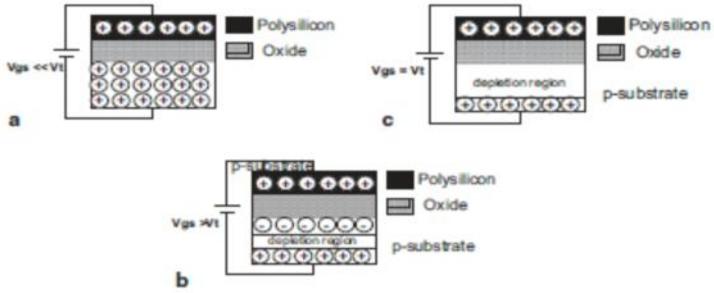


Fig. 3.8 **a** Accumulation mode, **b** depletion mode, and **c** inversion mode of an MOS transistor

Electrical Characteristics of MOS Transistors

The fluid model, presented in the previous section, gives us some basic understanding of the operation of an MOS transistor. We have seen that the whole concept of the MOS transistor is based on the use of the gate voltage to induce charge (inversion layer) in the channel region between the source and the drain. Application of the source-to-drain voltage V_{ds} causes this charge to flow through the channel from the source to drain resulting in source-to-drain current I_{ds}. The I_{ds} depends on two variable parameters—the gate-to-source voltage V_{gs} and the drain-to-source voltage $V_{ds.}$ The operation of an MOS transistor can be divided into the following three regions:

(a) *Cutoff region:* This is essentially the accumulation mode, when there is no effective flow of current between the source and drain.

(b) *Nonsaturated region:* This is the *active, linear,* or *weak inversion* mode, when the drain current is dependent on both the gate and the drain voltages.

(c) Saturated region: This is the strong inversion mode, when the drain current is independent of the drain-to-source voltage but depends on the gate voltage.

In this section, we consider an nMOS enhancement-type transistor and establish its electrical characteristics. The structural view of the MOS transistor, as shown in Fig. 3.9, shows the three important parameters of MOS transistors, the channel length *L*, the channel width *W*, and the dielectric thickness *D*. The expression for the drain current is given by

$$I_{ds} = \frac{\text{charge induced in the channel } (Q_c)}{\text{electron transit time } (t_n)}. \qquad (3.1)$$

Let us separately find out the expressions for Q_c and t_n.

With a voltage *V* applied across the plates, the charge is given by $Q = CV$, where

C is the capacitance. The basic formula for parallel-plate capacitor is $C = \frac{\varepsilon A}{D}$,

where ε is the permittivity of the insulator in units of F/cm. The value of ε depends on the material used to separate the plates. In this case, it is silicon dioxide (SiO_2). For SiO_2, $\varepsilon_{ox} = 3.9\varepsilon_0$, where ε_0 is the permittivity of the free space. For the MOS transistor, the gate capacitance

$$C_G = \frac{\varepsilon_{ox} WL}{D}. \quad (3.2)$$

Now, for the MOS transistor,

$$Q_c = C_G \cdot V_{eff},$$

where C_G is the gate capacitance and V_{eff} is the effective gate voltage.

$$\text{Now, the transit time, } t_n = \frac{\text{length of the channel } (L)}{\text{velocity of electron } (\tau_n)}. \quad (3.3)$$

The velocity, $\tau_n = \mu_n \cdot E_{ds}$, where μ_n is the mobility of electron and E_{ds} is the drain to the source electric field due to the voltage V_{ds} applied between the drain and source. Now, $E_{ds} = V_{ds}/L$.

So,

$$\tau_n = \frac{\mu_n V_{ds}}{L} \quad \text{and} \quad t_n = \frac{L^2}{\mu_n V_{ds}}. \quad (3.4)$$

Typical value of $\mu_n = 650\text{cm}^2/V$ (at room temperature).

The nonsaturated region: As the channel formation starts when the gate voltage is above the threshold voltage and there is a voltage difference of V_{ds} across the channel, the effective gate voltage is

$$V_{eff} = (V_{gs} - V_t - V_{ds}/2). \quad (3.5)$$

Substituting this, we get

$$Q_c = \frac{WL\varepsilon_{ox}}{D}\left[(V_{gs} - V_t) - \frac{V_{ds}}{2}\right]. \quad (3.6)$$

Now, the current flowing through the channel is given by

$$I_c = \frac{Q_c}{t_n}.$$

Substituting the value of t_n, we get

$$I_c = \frac{W\mu_n\varepsilon_{ox}}{LD}\left[(V_{gs} - V_t) - \frac{V_{ds}}{2}\right]V_{ds}. \quad (3.7)$$

Assuming $V_{ds} \le V_{gs} - V_t$ in the nonsaturated region and $K = \frac{\mu_n \varepsilon_{ox}}{D}$, we get

$$I_{ds} = \frac{KW}{L}\left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right]. \quad (3.8)$$

Now, the gate-channel capacitance based on parallel-plate capacitor model is

$$C_g = \frac{\varepsilon_{ins}\varepsilon_0 WL}{D} \quad \text{and} \quad K = \frac{C_g \mu_n}{WL}. \quad (3.9)$$

So, in terms of the gate-channel capacitance the expression for drain-to-source current can be written as

$$I_{ds} = \frac{C_g \mu_n}{L^2}\left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right]. \quad (3.10)$$

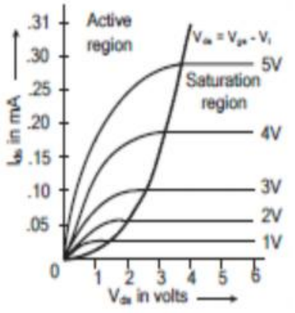


Fig. 3.10 Voltage–current characteristics of nMOS enhancement-type transistor

The Saturated Region As we have seen in the previous section, the drain current (I_{ds}) increases as drain voltage increases until the IR drop in the channel equals the effective gate voltage at the drain. This happens when $V_{ds} = V_{gs} - V_t$. At this point, the transistor comes out of the active region and I_{ds} remains fairly constant as V_{ds} increases further. This is known as saturation condition. Assuming $V_{ds} = V_{gs} - V_t$ for this region, the saturation current is given by

$$I_{ds} = K\frac{W}{L}\frac{(V_{gs} - V_t)^2}{2}$$

or

$$I_{ds} = \frac{C_g \mu_n}{2L^2}(V_{gs} - V_t)^2 = \frac{C_{ox} W \mu_n}{2L}(V_{gs} - V_t)^2 = \frac{\mu_n C_{ox}}{2}\frac{W}{L}(V_{gs} - V_t)^2. \quad (3.11)$$

It may be noted that in case of the enhancement-mode transistor, the drain-to-source current flows only when the magnitude exceeds the threshold voltage V_t. The I_{ds}−V_{ds} characteristic for an enhancement-type nMOS transistor is shown in Fig. 3.10.

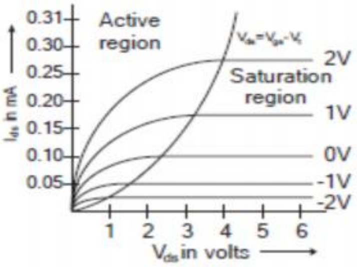
$$I_{ds} = 0 \quad \text{for } V_{gs} < V_t,$$

$$I_{ds}(\text{lin}) = \frac{\mu_n C_{ox}}{2}\frac{W}{L}\left(2(V_{gs} - V_t)V_{ds} - V_{ds}^2\right) \quad \text{for } V_{gs} \geq V_t \quad \text{and} \quad V_{ds} < V_{gs} - V_t,$$

$$I_{ds}(\text{sat}) = \frac{\mu_n C_{ox}}{2}\frac{W}{L}(V_{gs} - V_t)^2 \quad \text{for } V_{gs} \geq V_t \quad \text{and} \quad V_{ds} \geq V_{gs} - V_t.$$

Electrical characteristics of the nMOS enhancement-type transistor have been discussed above. In the depletion-type nMOS transistor, a channel is created by implanting suitable impurities in the region between the source and drain during fabrication prior to depositing the gate insulation layer and the poly-silicon layer. As a result, channel exists even when the gate voltage is 0 V. Here, the channel current can also be controlled by the gate voltage. A positive gate voltage increases the channel width resulting in an increase of drain current. A negative gate voltage decreases the channel Width leading to a reduced drain current. A suitable negative gate voltage fully depletes the channel isolating the source and drain regions. The characteristic curve, as shown in Fig. 3.11, is similar except the threshold volt-age, which is a negative voltage in case of a depletion-mode nMOS transistor. In a similar manner, the expression for drain current can be derived and voltage–current characteristics can be drawn for pMOS enhancement-mode and pMOS depletion-mode transistors.

Fig. 3.11 Voltage–current characteristics of nMOS depletion-type transistor



Threshold Voltage

One of the parameters that characterize the switching behavior of an MOS transistor is its threshold voltage V_t. As we know, this can be defined as the gate voltage at which an MOS transistor begins to conduct. Typical value for threshold voltage for an nMOS enhancement-type transistor is 0.2 V_{dd}, i.e., for a supply voltage of 5 V, V_{tn} = 1.0 V. As we have seen, the drain current depends on both the gate voltage and the drain voltage with respect to the source. For a fixed drain-to-source voltage, the variation of conduction of the channel region (represented by the drain current) for different gate voltages is shown in Fig. 3.11 for four different cases: nMOS deple-tion, nMOS enhancement, pMOS enhancement, and pMOS depletion transistors, as shown in Fig. 3.12a–d, respectively

.

The threshold voltage is a function of a number of parameters, including gate conductor material, gate insulation material, thickness of the gate insulator, doping level in the channel regions, impurities in the silicon–insulator interface and voltage between the source and substrate V_{sb}.

Moreover, the absolute value of the threshold voltage decreases with an increase in temperature at the rate of −2 mV/ C and − 4mV/ C for low and high substrate doping levels, respectively.

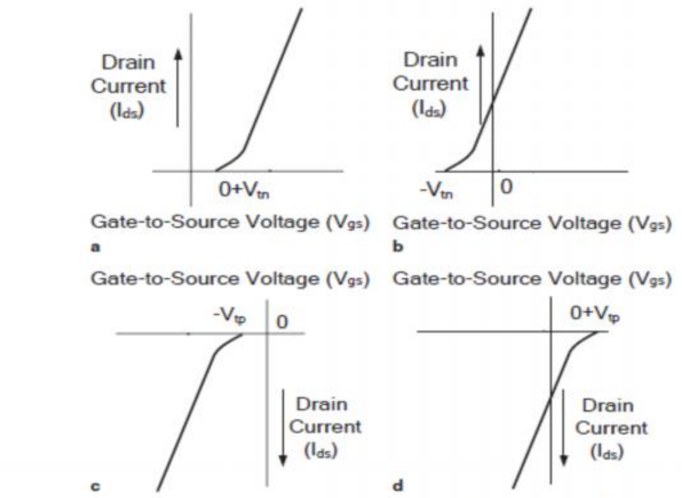
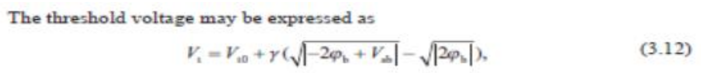


Fig. 3.12 Variation of drain current with gate voltage. **a** n-Channel enhancement. **b** n-Channel depletion. **c** p-Channel enhancement. **d** p-Channel depletion



The threshold voltage may be expressed as

$$V_t = V_{t0} + \gamma\left(\sqrt{\left|-2\varphi_b + V_{sb}\right|} - \sqrt{\left|2\varphi_b\right|}\right), \qquad (3.12)$$

where the parameter γ is the substrate bias coefficient, φ_b is substrate Fermi poten-tial and V_{sb} is the substrate-bias coffecient.

The expression holds good for both n-channel and p-channel devices.

•The substrate Fermi potential φ_b is negative in nMOS and positive in pMOS.

•The substrate bias coefficient γ is positive in nMOS and negative in pMOS.

•The substrate bias voltage V_{sb} is positive in nMOS and negative in pMOS.

V_{t0} is the threshold voltage for $V_{sb}=0$.

$$
\begin{aligned}
\varphi_b &= \frac{KT}{q}\ln\left(\frac{n_i}{N_A}\right) = 0.026\ln\left(\frac{1.45\times10^{10}}{10^{16}}\right) = -0.35 \\
C_{ox} &= \frac{\varepsilon_{ox}}{t_{ox}} = \frac{3.97\times8.85\times10^{-14}}{500\times10^{-8}} = 7.03\times10^{-8}\,\mathrm{F/cm^2} \\
\gamma &= \frac{\sqrt{2q\varepsilon_{si}N_A}}{C_{ox}} = \frac{\sqrt{2\times1.6\times10^{-19}\times10^{16}\times11.7\times8.85\times10^{-14}}}{7.03\times10^{-8}} = 0.82 \\
V_t &= V_{t0} + \lambda\sqrt{\left|-2\varphi_b + V_{sb}\right|} - \sqrt{\left|2\varphi_b\right|} = 0.4 + 0.82\sqrt{0.7+V_{sb}} - \sqrt{0.7},
\end{aligned}
\tag{3.13}
$$

where q is the charge of electron, ε_{ox} is the dielectric constant of the silicon sub-strate, N_A is the doping concentration densities of the substrate (10^{16} cm^{-3}), and C_{ox} is the oxide capacitance, N_i is the carrier concentration of the intrinsic silicon (1.45×10^{10} cm^{-3}).

Transistor Trans-conductance g_m

Transconductance is represented by the change in drain current for a change in gate voltage for a constant value of drain voltage. This parameter is somewhat similar to β, the current gain of BJTs.

$$
g_m = \left.\frac{\delta I_{ds}}{\delta V_{gs}}\right|_{V_{ds}=\text{constant}} \tag{3.14}
$$

This can be derived from

$$
I_{ds} = \frac{Q_c}{t_{sd}} \quad \text{or} \quad \delta I_{ds} = \frac{\delta Q_c}{t_{sd}}, \tag{3.15}
$$

$$
t_{sd} = \frac{L^2}{\mu_n V_{ds}}. \tag{3.16}
$$

Thus,

$$
\delta I_{ds} = \frac{\delta Q_c}{L^2} V_{ds}\mu_n. \tag{3.17}
$$

But,

$$
\delta Q_c = C_g \delta V_{gs}.
$$

So,

$$\delta I_{ds} = \frac{\mu_n C_g}{L^2} V_{ds} \delta V_{gs} \tag{3.18}$$

$$\begin{aligned} &\text{or } g_m = \frac{\delta I_{ds}}{\delta V_{gs}} = \frac{C_g \mu_n V_{ds}}{L^2}, \\ &\text{in saturation } V_{ds} = (V_{gs} - V_t), \\ &\text{and substituting } C_g = \frac{\varepsilon_{ins}\varepsilon_0 WL}{D}. \end{aligned} \tag{3.19}$$

We get

$$g_m = \frac{\mu_n \varepsilon_{ins} \varepsilon_0}{D} \frac{W}{L} (V_{gs} - V_t). \tag{3.20}$$

3.5.3 Figure of Merit

The figure of merit W_0 gives us an idea about the frequency response of the device

$$\begin{aligned} W_0 &= \frac{g_m}{c_g} = \frac{\mu_n}{L^2}(V_{gs} - V_t) \\ &= \frac{1}{t_{sd}}. \end{aligned} \tag{3.21}$$

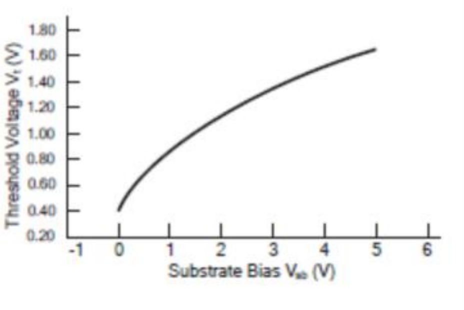
A fast circuit requires g_m as high as possible and a small value of C_g. From Eq. 3.23, it can be concluded that higher gate voltage and higher electron mobility provide better frequency response.

Body Effect

All MOS transistors are usually fabricated on a common substrate and substrate (body) voltage of all devices is normally constant. However, as we shall see in subsequent chapters, when circuits are realized using a number of MOS devices, several devices are connected in series. This results in different source potentials for different devices. It may be noted from Eq. 3.13 that the threshold voltage V_t is not constant with respect to the voltage difference between the substrate and the source of the MOS transistor. This is known as the substrate-bias effect or *body effect*. Increasing the V_{sb} causes the channel to be depleted of charge carriers, and this leads to an increase in the threshold voltage.

Using Eq. 3.13, we compute and plot the threshold voltage V_t as a function of the source-to-substrate voltage V_{sb}. The voltage V_{sb} will be assumed to vary between 0 and 5 V. The graph obtained is shown in Fig. 3.13.

Fig. 3.13 Variation of the threshold voltage as a function of the source-to-substrate voltage



The variation of the threshold voltage due to the body effect is unavoidable in many situations, and the circuit designer should take appropriate measures to over-come the ill effects of this threshold voltage variation.

Channel-Length Modulation

Simplified equations derived in Sect. 3.3 to represent the behavior of an MOS tran-sistor is based on the assumption that channel length remains constant as the drain voltage is increased appreciably beyond the onset of saturation. As a consequence, the drain current remains constant in the saturation region. In practice, however, the channel length shortens as the drain voltage is increased. For long channel lengths, say more than 5 µm, this variation of length is relatively very small compared to the total length and is of little consequence. However, as the device sizes are scaled down, the variation of length becomes more and more predominant and should be taken into consideration.

To have better insight of this phenomenon, let us examine the mechanisms of the formation of channel and current flow in an MOS transistor in different operat-ing conditions. Figure 3.14a shows the situation of an MOS transistor operating in the active or nonsaturation region ($0 < V_{ds} < V_{gs} - V_{tn}$) . In this mode, the inversion layer (i.e., channel)

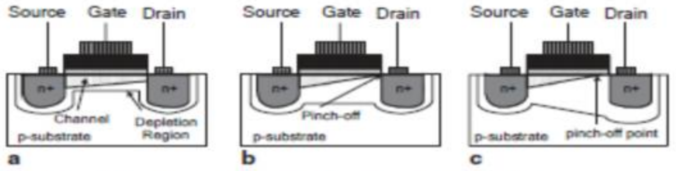
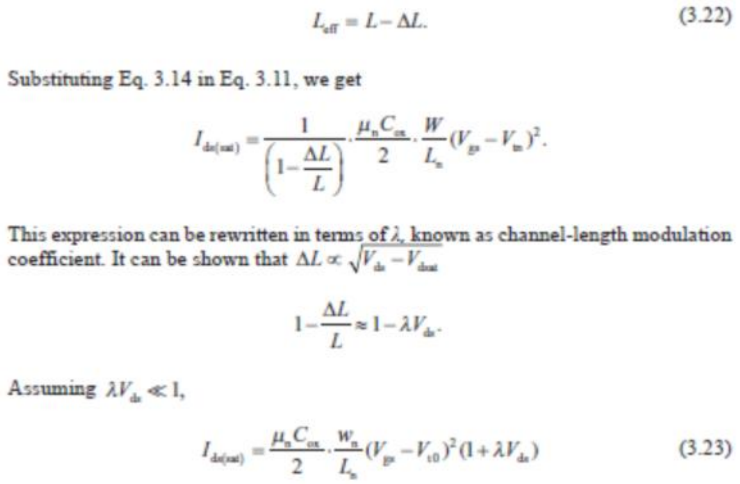


Fig. 3.14 **a** Nonsaturated region. **b** Onset of saturation. **c** Deep in saturation

formed under the influence of gate voltage provides a current path between the source and drain. As the drain voltage is increased from zero, the current flow increases linearly with the drain voltage, and the channel depth at the drain end also gradually decreases. Eventually at drain voltage $V_{ds} = V_{gs} - V_t$, the inversion charge and the channel depth reduces to zero as shown in Fig. 3.14b. This is known as the *pinch-off* point. As the drain voltage is increased further, a depletion region is formed adjacent to the drain, and the depletion region gradually grows with the increase in drain voltage. This leads to gradual shifting of the pinch-off point towards the source, thereby reducing channel length as shown in Fig. 3.14c. This effective channel length L_{eff} can be represented by



$$L_{eff} = L - \Delta L. \tag{3.22}$$

Substituting Eq. 3.14 in Eq. 3.11, we get

$$I_{ds(sat)} = \frac{1}{\left(1 - \frac{\Delta L}{L}\right)} \cdot \frac{\mu_n C_{ox}}{2} \cdot \frac{W}{L_n} (V_{gs} - V_{tn})^2.$$

This expression can be rewritten in terms of λ, known as channel-length modulation coefficient. It can be shown that $\Delta L \propto \sqrt{V_{ds} - V_{dsat}}$

$$1 - \frac{\Delta L}{L} \approx 1 - \lambda V_{ds}.$$

Assuming $\lambda V_{ds} \ll 1$,

$$I_{ds(sat)} = \frac{\mu_n C_{ox}}{2} \cdot \frac{W_n}{L_n} (V_{gs} - V_{t0})^2 (1 + \lambda V_{ds}) \tag{3.23}$$

The channel-length modulation coefficient λ has the value in the range of 0.02– 0.005 per volt. Taking into consideration the channel-length modulation effect, the voltage–current characteristic is shown in Fig. 3.15.

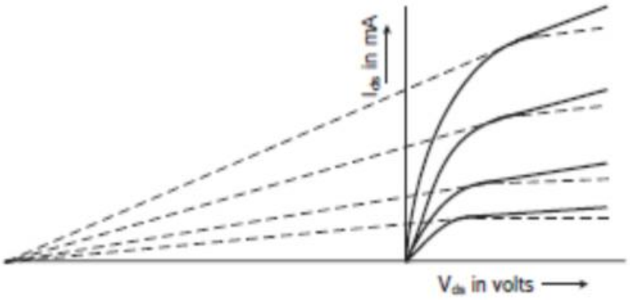


Fig. 3.15 Drain-current variations due to channel-length modulation

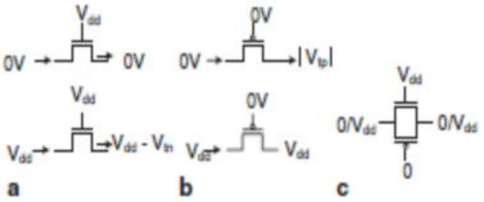


Fig. 3.16 **a** nMOS pass transistor. **b** pMOS pass transistor. **c** Transmission gate

3.6 MOS Transistors as a Switch

We have seen that in the linear region (when the drain-to-source voltage is small) an MOS transistor acts as a variable resistance, which can be controlled by the gate voltage. An nMOS transistor can be switched from very high resistance when the gate voltage is less than the threshold voltage, to low resistance when V_{gs} exceeds the threshold voltage $V_{t n}$. This has opened up the possibility of using an MOS tran-sistor as a switch, just like a relay. For example, an nMOS transistor when used as a switch is OFF when $V_{gs} = 0$ V and ON when $V_{gs} = V_{dd}$. However, its behavior as a switch is not ideal. When $V_{gs} = V_{dd}$, the switch turns on but the on resistance is not zero. As a result, there is some voltage drop across the switch, which can be ne-glected when it is in series with a large resistance. Moreover, if V_{dd} is applied to the input terminal, at the other end we shall get $(V_{dd}-V_{t n})$. This is because when output voltage is more than $(V_{dd}-V_{t n})$, the channel turns off, and it no longer functions as a closed switch as shown in Fig. 3.15a. However, a low-level signal can be passed without any degradation. The transistor used in the above manner is known as *pass transistor*. It may be noted that the roles of drain and source are interchangeable, and the device truly acts as a bilateral switch.

Similarly, a pMOS transistor can also be used as a switch. In this case, the mini-mum voltage that it can pass is V_{tp}, since below this value gate-to-source voltage will be higher than $-V_{tp}$ and the transistor turns off. This is shown in Fig. 3.16b. Therefore, a p-channel transistor passes a weak low-level signal but a strong high-level signal as shown below. Later, we shall discuss the use of pass transistors in realizing Boolean functions and discuss its advantages and disadvantages.

To overcome the limitation of either of the transistors, one pMOS and one nMOS transistor can be connected in parallel with complementary inputs at their gates. In this case, we can get both low and high levels of good quality of the output. The low level passes through the nMOS switch, and the high level passes through the pMOS switch without any degradation as shown in Fig. 3.16c. A more detailed discussion on transmission gates is given in the following subsection.

3.6.1 Transmission Gate

The transmission gate is one of the basic building blocks of MOS circuits. It finds use in realizing multiplexors, logic circuits, latch elements, and analog switches.

The characteristics of a transmission gate, which is realized by using one nMOS and one pMOS pass transistors connected in parallel, can be constructed by com-bining the characteristics of both the devices. It may be noted that the operation of a transmission gate requires a dual-rail (both true and its complement) control signal. Both the devices are off when “0” and “1” logic levels are applied to the gates of the nMOS and pMOS transistors, respectively. In this situation, no signal passes through the gate. Therefore, the output is in the high-impedance state, and the intrinsic load capacitance associated to the output node retains the high or low voltage levels, whatever it was having at the time of turning off the transistors. Both the devices are on when a “1” and a “0” prior to the logic levels are applied to the gates of the nMOS and pMOS transistors, respectively. Both the devices take part in passing the input signal to the output. However, as discussed below, their contribu-tions are different in different situations.

To understand the operation of a transmission gate, let us consider two situations. In the first case, the transmission gate is connected to a relatively large capacitive load, and the output changes the state from low to high or high to low as shown in Fig. 3.17.

Case I: Large Capacitive Load First, consider the case when the input has changed quickly to V_{dd} from 0 V and the output of the switch changes slowly from 0 V (V_{ss}) to V_{dd} to charge a load capacitance C_{L}. This can be modeled by using V_{dd} as an input and a ramp voltage generated at the output as the capacitor charges from V_{ss} to V_{dd}. Based on the output voltage, the operations of the two transistors can be divided into the following three regions:

Region I: As the voltage difference between the input and output is large, both nMOS and pMOS transistors are in saturation.

Region II: nMOS is in saturation and pMOS in linear for $V_{\mathrm{tp}} < V_{\mathrm{out}} < V_{\mathrm{dd}} - V_{\mathrm{tn}}$.

Region III: nMOS is in cutoff and pMOS in linear for $V_{\mathrm{out}} > V_{\mathrm{dd}} - V_{\mathrm{tn}}$.

Similarly, when the input voltage changes quickly from V_{dd} to 0 V and the load capacitance discharges through the switch.

Region I: Both nMOS and pMOS are in saturation for $|V_{out}| < V_{tp}$.

Region II: nMOS is in the linear region, and pMOS is in saturation for

$(Vdd - |Vtp|) < Vout < Vtn$.

Region III: nMOS is in the linear region, and pMOS is cutoff for $|V_{out}| < (V_{dd} - V_{tn})$.

As the current decreases linearly as voltage across the capacitor decreases from V_{dd} to 0 V. Note that the role of the two transistors reverses in the two cases.

Case II: Small Capacitive Load Another situation is the operation of the trans-mission gate when the output is lightly loaded (smaller load capacitance). In this case, the output closely follows the input. This is represented in Fig. 3.18a.

In this case, the transistors operate in three regions depending on the input voltage as follows:

Region I: nMOS is in the linear region, pMOS is cutoff for $V_{in} < |V_{tp}|$.

Region II: nMOS is in the linear region, pMOS linear for $V_{tp} < |V_{in}| < (V_{dd} - V_{tn})$.

Region III: nMOS is cutoff, pMOS is in the linear region for

$V_{in} > (V_{dd} - |V_{tn}|)$.

As the voltage difference between the transistors is always small, the transistors either operate in the nonsaturated region or are off.