

LOW POWER VLSI CIRCUITS AND SYSTEMS (15A04802)

LECTURE NOTES

B.TECH

IV-YEAR& II-SEM

Prepared by:

A.Mounika, Assistant Professor

Department of Electronics & Communication Engineering



ANNAMACHARYA INSTITUTE OF TECHNOLOGY & SCIENCES

(Approved by A.I.C.T.E., New Delhi & Affiliated to J.N.T. University, Ananthapur)
Venkatapuram (Village), Karakambadi (Post), Mangalam Road, Renigunta (Mandal),
Tirupati-517 520, Chittoor Dist., A.P. Ph: (0877) 2285608

COURSE MATERIAL

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR
B. Tech IV-II Sem. (ECE)

L T P C

3 1 0 3

15A04802 LOW POWER VLSI CIRCUITS AND SYSTEMS

Course Outcomes :

After completion of this subject, students will be able to

- Under stand the concepts of velocity saturation, Impact Ionization and Hot Electron Effect
- Implement Low power design approaches for system level and circuit level measures.
- Design low power adders, multipliers and memories for efficient design of systems.

UNIT I

Introduction, Historical background, why low power, sources of power dissipations, low power design methodologies.

MOS Transistors: introduction, the structure of MOS Transistor, the Fluid model, Modes of operation of MOS Transistor, Electrical characteristics of MOS Transistors, MOS Transistors as a switch.

UNIT II

MOS Inverters: introduction, inverter and its characteristics, configurations, inverter ratio in different situations, switching characteristics, delay parameters, driving parameters, driving large capacitive loads.

MOS Combinational Circuits: introduction, Pass-Transistor logic, Gate logic, MOS Dynamic Circuits.

UNIT III

Sources of Power Dissipation: introduction, short-circuit power dissipation, switching power dissipation, glitching power dissipation, leakage power dissipation.

Supply voltage scaling for low power: introduction, device features size scaling, architecture-level approaches, voltage scaling, multilevel voltage scaling, challenges, dynamic voltage and frequency scaling, adaptive voltage scaling.

UNIT IV

Minimizing Switched Capacitance: introduction, system-level approaches, transmeta's Crusoe processor, bus encoding, clock gating, gated-clock FSMs, FSM state encoding, FSM Partitioning, operand isolation, precomputation, logic styles for low power.

UNIT V

Minimizing Leakage Power: introduction, fabrication of multiple threshold voltages, approaches for minimizing leakage power, Adiabatic Logic Circuits, Battery-Driven System, CAD Tools for Low Power VLSI Circuits.

TEXT BOOKS

1. Ajit. Pal, Low power VLSI Circuits and systems, springer
2. Sung Mo Kang, Yusuf Leblebici, CMOS Digital Integrated Circuits, Tata Mcgrag Hill.
3. Neil H. E. Weste and K. Eshraghian, Principles of CMOS VLSI Design, 2nd Edition, Addison Wesley (Indian reprint).
4. A. Bellamour, and M. I. Elmasri, Low Power VLSI CMOS Circuit Design, Kluwer Academic Press, 1995.
5. Anantha P. Chandrakasan and Robert W. Brodersen, Low Power Digital CMOS Design, Kluwer Academic Publishers, 1995.

REFERENCES

1. Kaushik Roy and Sharat C. Prasad, Low-Power CMOS VLSI Design, Wiley-Interscience, 2000.
-

UNIT-1

Introduction to Low Power VLSI

1.1 Introduction

- ❖ **VLSI-Very Large Scale Integration-** Very-large-scale integration (VLSI) is the process of creating an integrated circuit (IC) by combining hundreds of thousands of transistors or devices into a single chip.
- ❖ Design for low power has become nowadays one of the major concerns for complex, very-large-scale-integration (VLSI) circuits.
- ❖ **Micron Technology** ==> 1 μ m, 2 μ m, 3 μ m, etc
- ❖ **Sub-Micron Technology** ==> 0.8 μ m, 0.6 μ m, 0.35 μ m 0.25 μ m etc
- ❖ **Deep Sub-Micron Technology** ==> 0.18 μ m, 0.13 μ m
- ❖ **Nanotechnology** ==> 90nm, 65nm etc

1.2 Historical Background

- ❖ **Moore's law-**Component density in an IC would double every 18 months.
- ❖ **Evolution of IC Technology**

Table 1.1 Evolution of IC Technology

Year	Technology	Number of Components	Typical Product
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 μ p, RAM, ROM
1980–1984	Very-large-scale integration	20,000–50,000	DSPs, RISC processors, 16-bit, 32-bit μ P
1985–	Ultra-large-scale integration	> 50,000	64-bit μ p, dual-core μ P

MUX-Multiplexer, μ P-Microprocessor, RAM-Random-Access Memory, ROM - Read-Only Memory, DSP-Digital Signal Processor, RISC-Reduced Instruction Set Computer

- ❖ **Landmark Years of Semiconductor Industry**
- ✓ 1947: Invention of transistor by **William Shockley** 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.

1.3 Why Low Power?

❖ Important issue in the present day VLSI circuit realization

- ✓ Increasing Transistor Count
- ✓ Higher Speed of Operation
- ✓ Greater Device Leakage Currents

❖ Packaging and Cooling Cost

- ✓ Contemporary high performance processor consume heavy power
- ✓ Cost associated with packaging and cooling such devices is prohibitive
- ✓ Low power methodology to be used to reduce cost of packaging and cooling

❖ Reliability

- ✓ Every 10°C rise in temperature roughly doubles the failure rate

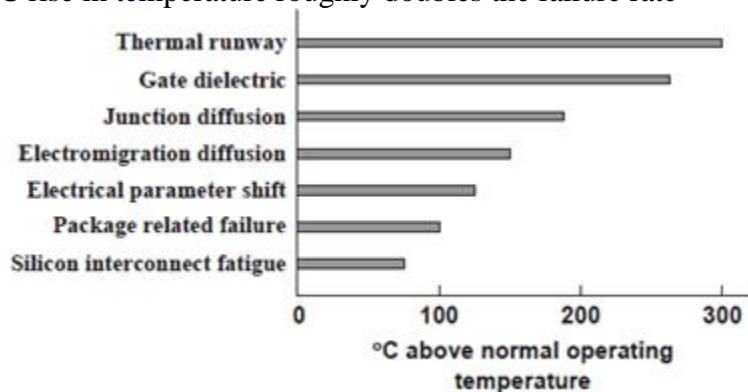


Fig. 1.1 Different Failure Mechanisms against Temperature

❖ Environment

- ✓ According to an estimate of the US Environmental Protection Agency (EPA), 80 % of the power consumption by office equipment is due to computing equipment and a large part from unused equipment.
- ✓ Power is dissipated mostly in the form of heat.
- ✓ The cooling techniques, such as air conditioner, transfer the heat to the environment.
- ✓ To reduce adverse effect on environment, efforts such as EPA’s Energy Star program leading to power management standard for desktop and laptops has emerged.

1.4 Sources of Power Dissipations

1.4.1 POWER and ENERGY

Power is the instantaneous power in the device, while energy is the integration of power with time. Figure 1.2 illustrates the difference between power energy. For example, in Fig. 1.2, 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.

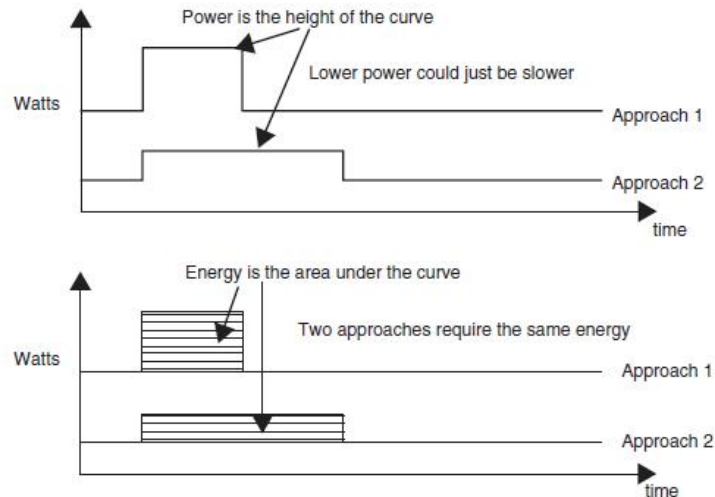


Fig. 1.2 Power versus Energy

1.4.2 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.

1.4.3 Types of Power Dissipations

- ❖ Dynamic power is the power consumed when the device is active, that is, when the signals of the design 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. 1.3.

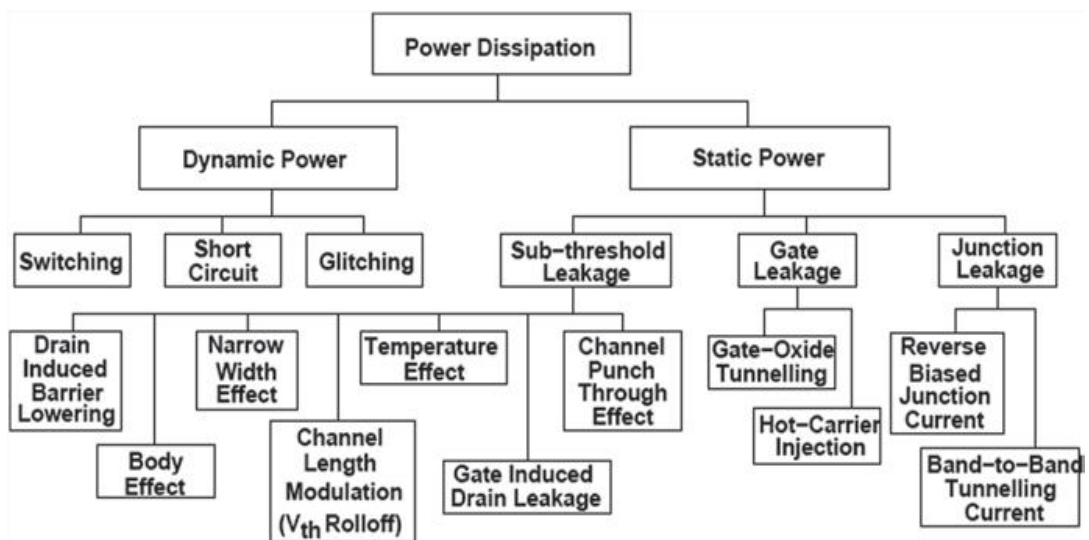


Fig. 1.3 Types of power dissipation

1.4.4 Dynamic Power Dissipation

- ❖ 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
 - ✓ Glitching Power

1.4.4.1 Switching Power Dissipation

- ❖ The first and primary source of dynamic power consumption is the Switching power dissipation occurs due the power required to charging and discharging of the output capacitance on a gate. Figure 1.4 illustrates switching power for charging a capacitor.

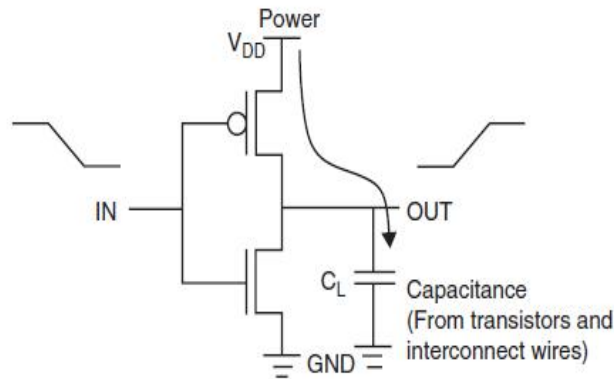


Fig. 1.4 Dynamic (switching) power.

The energy per transition is given by

$$\text{Energy/Transition} = \frac{1}{2} \times C_L \times V_{dd}^2 \quad (1.1)$$

Where C_L is the load capacitance and V_{dd} is the supply voltage

Switching power is therefore expressed as:

$$P_{switch} = \frac{\text{Energy}}{\text{Transition}} \times f = C_L \times V_{dd}^2 \times P_{trans} \times f_{clock} \quad (1.2)$$

Where f is the frequency of transitions, P_{trans} is the probability of an output transition and f_{clock} is the frequency of the system clock

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

$$P_{totalswitch} = C_L \times V_{dd}^2 \times P_{trans} \times f_{clock} + \sum \alpha_i \times C_i \times V_{dd} \times (V_{dd} - V_{th}) \times f_{clock} \quad (1.3)$$

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

1.4.4.2 Short-Circuit Power Dissipation

In addition to the switching power, short-circuit power also contributes to the dynamic power. Figure 1.5 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).

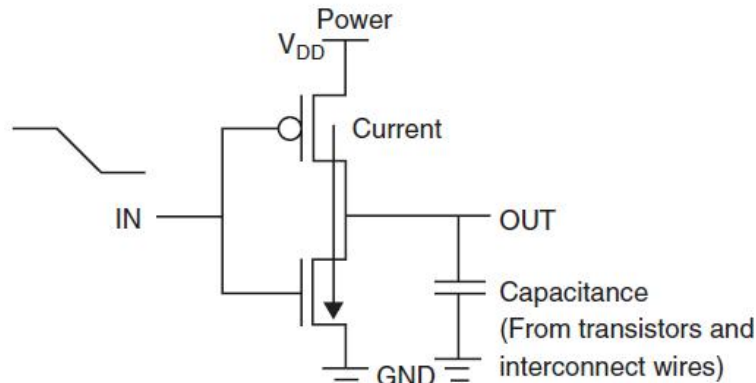


Fig. 1.5 Short-circuit current or crowbar current.

The expression for short-circuit power is given by

$$P_{shortcircuit} = t_{sc} \times V_{dd} \times I_{peak} \times f_{clock} = \frac{\mu \epsilon_{ox} W}{12LD} \times (V_{dd} - V_{th})^3 \times t_{sc} \times f_{clock} \quad (1.4)$$

- ✓ Where t_{sc} is the rise/fall time duration of the short-circuit current
- ✓ I_{peak} is the total internal switching current (short-circuit current plus the current to charge the internal capacitance)
- ✓ μ is the mobility of the charge carrier
- ✓ ϵ_{ox} is the permittivity of the silicon dioxide
- ✓ W is the width
- ✓ L is the length
- ✓ 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.

1.4.4.3 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 occur 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.

1.4.5 Static Power Dissipation

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.

Figure 1.6 shows several leakage mechanisms that are responsible for static power dissipation. Here, I_1 is the reverse-bias p–n junction diode leakage current, I_2 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, I_3 is the sub-threshold leakage current between the source and the drain when the gate voltage is less than the threshold voltage (V_{th}), I_4 is the oxide tunneling current due to reduction in the oxide thickness, I_5 is the gate current due to hot carrier injection of electrons (I_4 and I_5 are commonly known as I_{GATE} leakage current), I_6 is the gate-induced drain leakage current due to high field effect in the drain junction, and I_7 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. 1.7

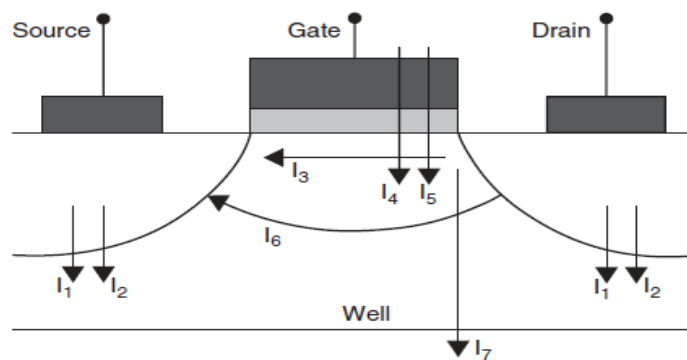


Fig. 1.6 Leakage currents in an MOS transistor.

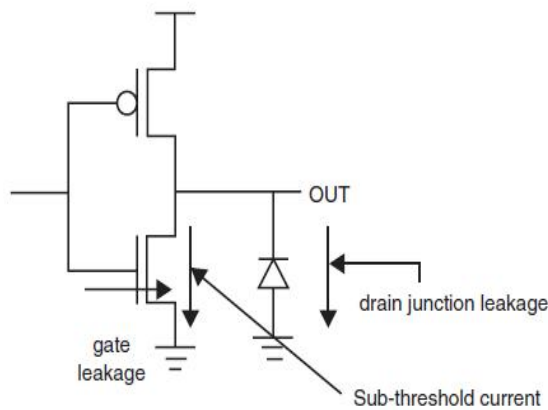


Fig. 1.7 Leakage currents in a CMOS inverter.

- ✓ 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**
- **Channel punch through current**

1.5.1 Low Power Design Methodology

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. Various approaches can be used at different level of design hierarchy.

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 alternative 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 issues such as cost complexity, performance, and power consumption. From the behavioral 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 logic-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 encoding 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 submicron 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 vector 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.

MOS Transistor

2.1 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.
- ✓ 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 (SiO₂)*.
- ✓ 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.

2.2 Structure of MOS Transistors

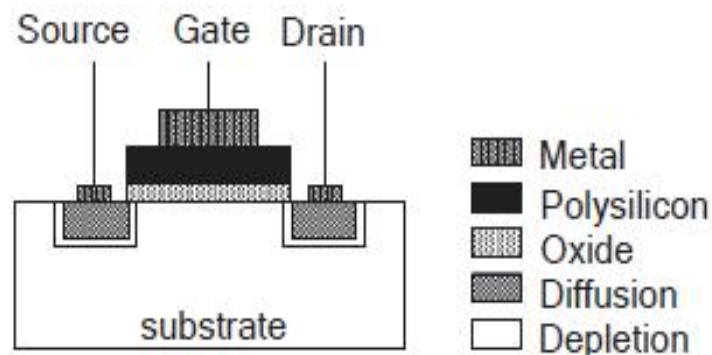


Fig. 2.1 Structure of an MOS transistor

- ✓ The structure of an MOS transistor is shown in Fig. 2.1. On a lightly doped substrate 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.

2.2.1 nMOS and pMOS Transistors

- ✓ 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 substrate 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.

2.2.2 nMOS Enhancement-Mode and Depletion-Mode Transistors

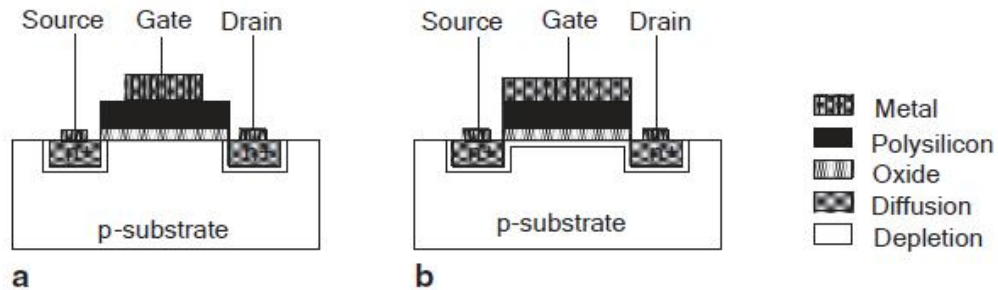


Fig. 2.2 **a** nMOS enhancement-mode transistor. **b** nMOS depletion-mode transistor

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. In order to make a useful device, there must be suitable means to establish some 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. 2.2a.

Enhancement-mode nMOS transistor:

- ✓ 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 between the source and drain. Current can flow between the source and drain through this conducting path.

Depletion-mode nMOS transistor:

- ✓ 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. 2.2b.
- ✓ Here, 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.

For example, consider the case when the substrate is lightly doped in p-type and the channel region implanted with n-type of impurity. This leads to the formation of an nMOS depletion-mode transistor. In both the cases, the current flow between the source and drain can be controlled by varying the gate voltage and only one type of charge carrier, that is, electron or hole takes part in the flow of current. That is the reason why MOS devices are called unipolar devices, in contrast to bipolar junction

transistors (BJTs), where both types of charge carriers take part in the flow of current. Therefore, by using the MOS technology, four basic types of transistors can be fabricated—nMOS enhancement type, nMOS depletion type, pMOS enhancement type, and pMOS depletion type. Each type has its own pros and cons. It is also possible to realize circuits by combining both nMOS and pMOS transistors, known as Complementary MOS (CMOS) technology. Commonly used symbols of the four types of transistors are given in Fig. 2.3.

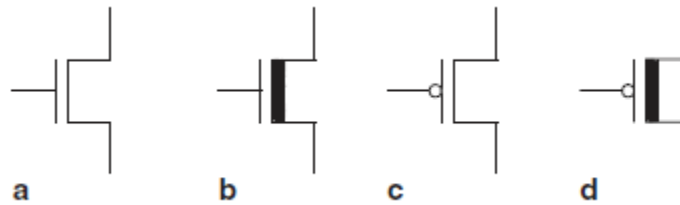


Fig. 2.3 a nMOS enhancement. b nMOS depletion. c pMOS enhancement. d pMOS depletion-mode transistors

2.3 FLUID MODEL

- ✓ The Fluid model 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).

The model is based on two simple ideas:

- 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
- 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.

2.3.1 The MOS Capacitor

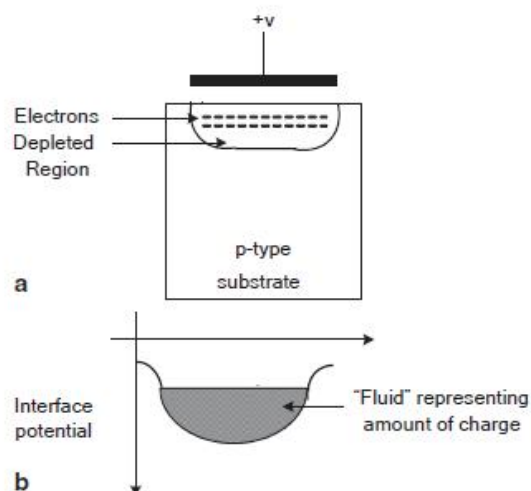


Fig. 2.4 a An MOS capacitor. b The fluid model

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 2.4a.

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 surface potential under the MOS electrode. In the presence of inversion charge, the surface potential is shown in Fig. 2.4b 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.

2.3.2 The MOS Transistor

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. 2.5a.

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 potential 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 these controlled with the help of the gate voltage. The potential at the channel region is shown by the dotted lines of Fig. 2.5b. 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. 2.5b. 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. 2.5b. 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.

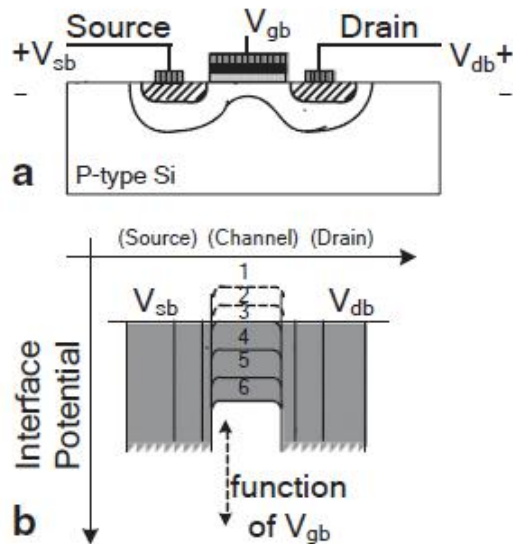


Fig. 2.5 a An MOS transistor. **b** The fluid model

❖ **Active, linear or unsaturated and Saturation Region**

When the source and drain are biased to different potentials ($V_{db} > V_{sb}$), there will be a difference in the potential levels. Let us consider two different situations. In the first case, the drain voltage is greater than the source voltage by some fixed value, and the gate voltage V_{gb} is gradually increased from 0 V. Figure 2.6 shows different situations. Initially, for $V_{gb} = 0$ V, the potential level in the channel region is above the potential level of either of the source and drain regions, and the source and drain are isolated. Now, if the gated voltage is gradually increased, first, the gate region potential reaches the potential of the source region. Charge starts moving from the source to the drain as the gate voltage is slightly increased. The rate of flow of charge moving from the source to the drain region, represented by the slope of the interface potential in the channel region, keeps on increasing until the gate region potential level becomes the same as that of the drain potential level. In this situation, the device is said to be operating in an *active, linear, or unsaturated* region. If the gate voltage is increased further, the width of the channel between the source and drain keeps on increasing, leading to a gradual increase in the drain current.

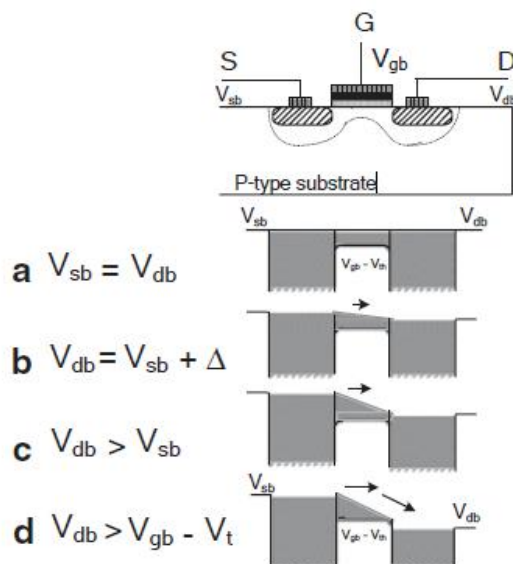


Fig. 2.6 The fluid model of an MOS transistor

Let us consider another case when the gate voltage is held at a fixed value for a heavily turned-on channel. To start with, the drain voltage is the same as that of the source voltage, and it is gradually increased. Figure 2.6a shows the case when the source and drain voltages are equal. Although the path exists for the flow of charges, there will be no flow because of the equilibrium condition due to the same level. In Fig. 2.6b, a small voltage difference is maintained by externally applied voltage level. There will be continuous flow of charge resulting in drain current. With the increase in voltage difference between the source and drain, the difference in the fluid level increases, and the layer becomes more and more thin, signifying faster movement of charges. With the increasing drain potential, the amount of charge flowing from the source to drain per unit time increases. In this situation, the device is said to be operating in an *active, linear, or unsaturated* region. However, there is a limit to it. It attains a maximum value, when the drain potential $V_{db} = (V_{gb} - V_t)$. Further increase in drain voltage does not lead to any change in the rate of charge flow. The device is said to be in the *saturation* region. In this condition, the drain current becomes independent of the drain voltage, and it is fully determined by the gate potential.

❖ **MOS Characteristics**

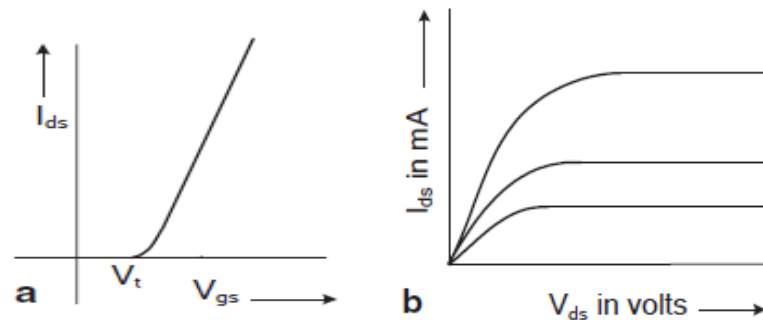


Fig. 2.7 (a) Drain Current (I_{ds}) Vs Gate Voltage (V_{gs}) (b) Voltage-Current Characteristic (V_{ds} Vs I_{ds})

To summarize this section, 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. 2.7a. 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. 2.7b.

2.4 Modes of Operation of MOS Transistors

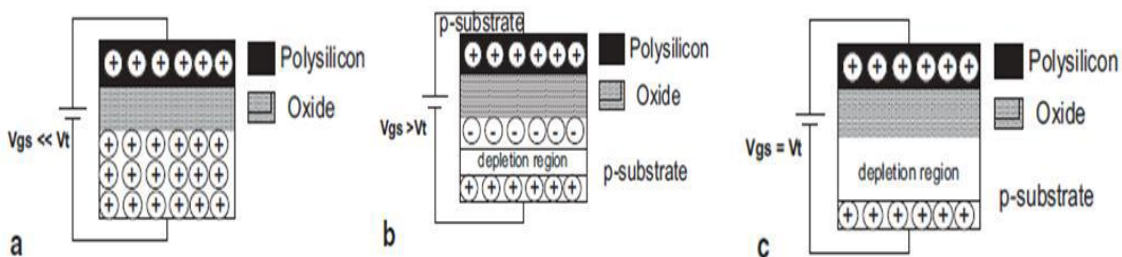


Fig. 2.8 a Accumulation mode, b depletion mode, and c inversion mode of an MOS transistor

Accumulation Mode: When the gate voltage is very small and much less than the threshold voltage. Fig. 2.8a 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*

Depletion Mode: As the gate voltage is increased, the holes are repelled from the SiO₂-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. 2.8b.

Inversion Mode: 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. 2.8c. The transistor is now said to be in *inversion mode*.

2.5 Electrical Characteristics of MOS Transistor

❖ Drain Source Current Expression for nMOS Enhancement Type Transistor

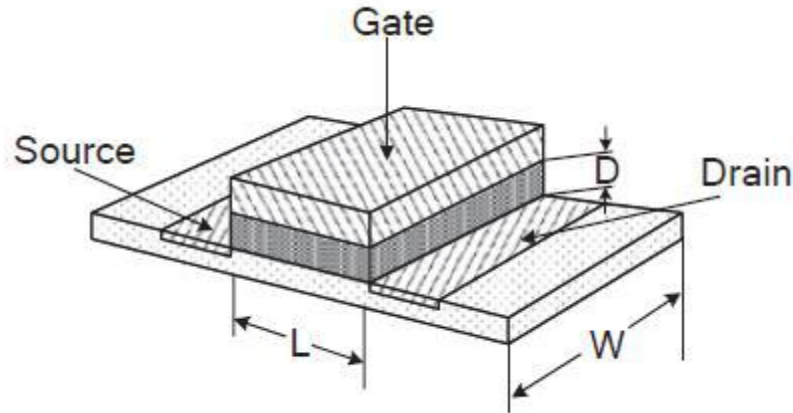


Fig. 2.9 Structural view of an MOS transistor

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:

- ✓ Cutoff region: This is essentially the accumulation mode, when there is no effective flow of current between the source and drain.
- ✓ Non-saturated region: This is the active, linear, or weak inversion mode, when the drain current is dependent on both the gate and the drain voltages.
- ✓ 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. 2.9, 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

$$I_{ds} = \text{Charge Induced in the Channel } (Q_c) / \text{Electron Transit Time } (t_n) \quad (2.1)$$

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 $C = \epsilon A / D$, 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 (SiO₂). For SiO₂, $\epsilon_{ox} = 3.9\epsilon_0$, where ϵ_0 is the permittivity of the free space.

$$\text{For MOS Transistor, Gate Capacitance, } C_G = \frac{\epsilon_{ox}WL}{D} \quad (2.2)$$

For the MOS transistor, $Q_c = C_G \cdot V_{eff}$ (2.3)

Where V_{eff} is the Effective gate voltage

Transit Time, $t_n = \text{Length of the Channel (L)} / \text{Velocity of Electron } (\tau_n)$ (2.4)

The Velocity, $\tau_n = \mu_n \cdot E_{ds}$, where μ_n is the Mobility of Electron (Typical value of $\mu_n = 650 \text{ cm}^2/\text{V}$ at Room temperature) and E_{ds} is the Drain to Source electric field due to the voltage V_{ds} applied between the drain and source, $E_{ds} = V_{ds}/L$.

$$\tau_n = \frac{\mu_n V_{ds}}{L} \text{ and } t_n = \frac{L^2}{\mu_n V_{ds}} \quad (2.5)$$

$$Q_c = \frac{WL\epsilon_{ox}}{D} V_{eff} \quad (2.6)$$

For Non-saturated Region

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

$$V_{eff} = (V_{gs} - V_t - V_{ds}/2) \quad (2.7)$$

Substituting equation (2.7) in Equation (2.6), we get

$$Q_c = \frac{WL\epsilon_{ox}}{D} \left[(V_{gs} - V_t) - \frac{V_{ds}}{2} \right] \quad (2.8)$$

Substituting the value of t_n and Q_c in equation (2.1), we get

$$I_{ds} = \frac{W\mu_n\epsilon_{ox}}{LD} \left[(V_{gs} - V_t) - \frac{V_{ds}}{2} \right] V_{ds} \quad (2.9)$$

$$I_{ds} = \frac{KW}{L} \left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2} \right] \text{ for } V_{gs} \geq V_t \text{ and } V_{ds} < V_{gs} - V_t \quad (2.10)$$

Where $K = \frac{\mu_n\epsilon_{ox}}{D}$

For Saturated Region

$$I_{ds} = \frac{W\mu_n\epsilon_{ox}}{DL} \frac{(V_{gs} - V_t)^2}{2} = K \frac{W}{L} \frac{(V_{gs} - V_t)^2}{2} \text{ for } V_{gs} \geq V_t \text{ and } V_{ds} \geq V_{gs} - V_t \quad (2.11)$$

For Cutoff Region

$$I_{ds} = 0 \text{ for } V_{gs} < V_t \quad (2.12)$$

✓ Threshold Voltage

$$V_t = V_{t0} + \gamma \sqrt{|-2\phi_b + V_{sb}|} - \sqrt{|2\phi_b|} = 0.4 + 0.82\sqrt{0.7 + V_{sb}} - \sqrt{0.7}$$

✓ Transistor Transconductance (g_m)

$$g_m = \frac{\delta I_{ds}}{\delta V_{gs}} / V_{ds} = \text{constant} = \frac{\mu_n \epsilon_{ins} \epsilon_o W}{D L} (V_{gs} - V_t)$$

✓ Figure of Merit

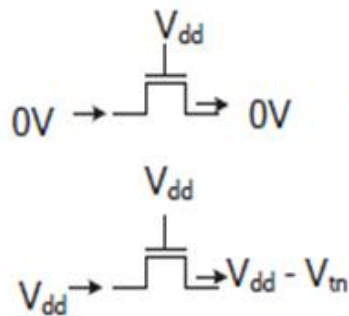
$$W_o = \frac{g_m}{C_g} = \frac{\mu_n}{L^2} (V_{gs} - V_t) = \frac{1}{t_{sd}}$$

✓ Body Effect

✓ Channel-Length Modulation

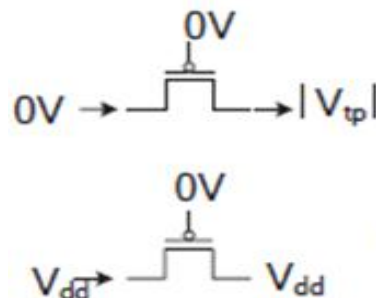
2.6 MOS Transistor as a Switch

❖ nMOS Pass Transistor



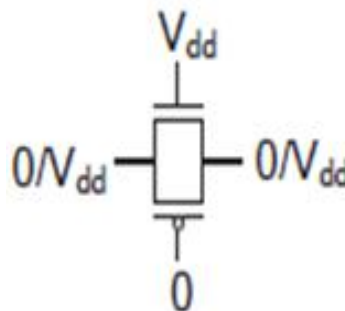
- ✓ nMOS transistor when used as a switch is OFF when $V_{gs} = 0\text{ V}$ and ON when $V_{gs} = V_{dd}$.
- ✓ $V_{in}=0\text{V}, V_{out}=0\text{V}$
- ✓ $V_{in}=5\text{V}, V_{out}=V_{dd}-V_{tn}$

❖ pMOS Pass Transistor



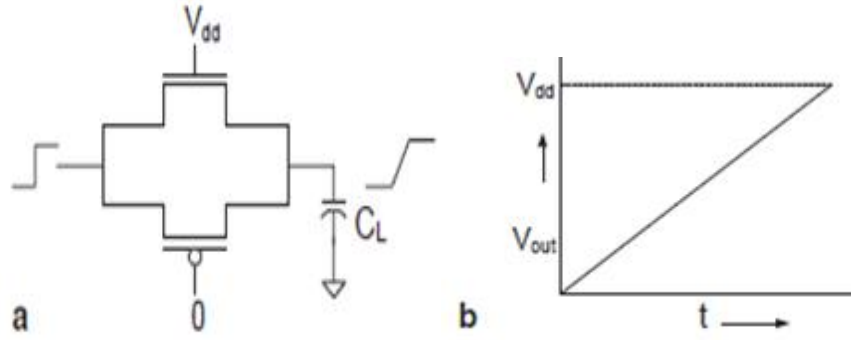
- ✓ PMOS transistor when used as a switch is ON when $V_{gs} = 0\text{ V}$ and OFF when $V_{gs} = V_{dd}$.
- ✓ $V_{in}=0\text{V}, V_{out}=|V_{tp}|$
- ✓ $V_{in}=+5\text{V}, V_{out}=+5\text{V}$

❖ Transmission Gate

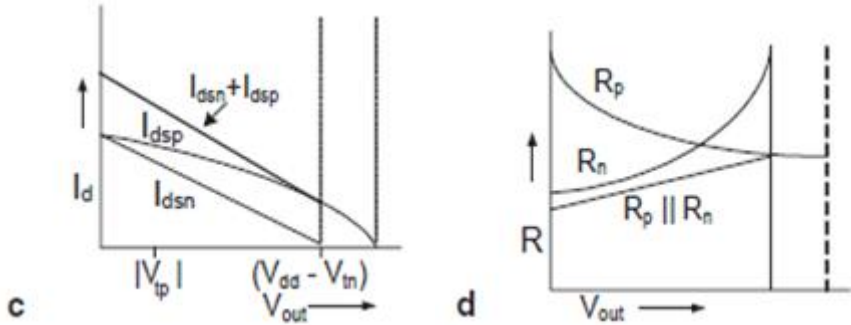


- ✓ One pMOS and one nMOS transistor can be connected in parallel with complementary inputs at their gates.
- ✓ This is known as Transmission Gate
- ✓ Both the devices are OFF when “0” and “1” logic levels are applied to the gates of the nMOS and pMOS transistors, respectively.
 - $V_{gsn}=0\text{V}$ and $V_{gsp}=+5\text{V}$, The Switch is OFF
- ✓ 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.
 - $V_{gsn}=+5\text{V}$ and $V_{gsp}=0\text{V}$, The Switch is ON
 - $V_{in}=0, V_{out}=0\text{V}$ and $V_{in}=+5\text{V}, V_{out}=+5\text{V}$

❖ Transmission gate Case I: Large Capacitive Load



(a) Output node charges from low-to-high level
 (b) The output voltage changing with time for different transitions.



(c) The drain currents through the two transistors as a function of the output voltage.
 (d) The equivalent resistances as a function of the output voltage

1. OUTPUT Node Changes from LOW-to-HIGH Level

Region I-Both nMOS and pMOS transistors are in SATURATION, $V_{out} < |V_{tp}|$

$$I_{dsn} = K_n \frac{W_n}{2L_n} (V_{dd} - V_{out} - V_{tn})^2$$

$$I_{dsp} = K_p \frac{W_p}{2L_p} (V_{dd} - |V_{tp}|)^2$$

$$R_{eqn} = \frac{V_{dd} - V_{out}}{I_{dsn}} = \frac{2L_n}{K_n W_n} \frac{(V_{dd} - V_{out})}{(V_{dd} - V_{out} - V_{tn})^2}$$

$$R_{eqp} = \frac{V_{dd} - V_{out}}{I_{dsp}} = \frac{2L_p}{K_p W_p} \frac{(V_{dd} - V_{out})}{(V_{dd} - |V_{tp}|)^2}$$

Region II-nMOS is in Saturation and pMOS in Linear, $|V_{tp}| < V_{out} < V_{dd} - V_{tn}$

$$I_{dsp} = K_p \frac{W_p}{L_p} \left[(V_{dd} - |V_{tp}|)(V_{dd} - V_{out}) - \frac{(V_{dd} - V_{out})^2}{2} \right]$$

$$R_{eqp} = \frac{1}{K_p W_p \left[2(V_{dd} - |V_{tp}|) - (V_{dd} - V_{out}) \right]}$$

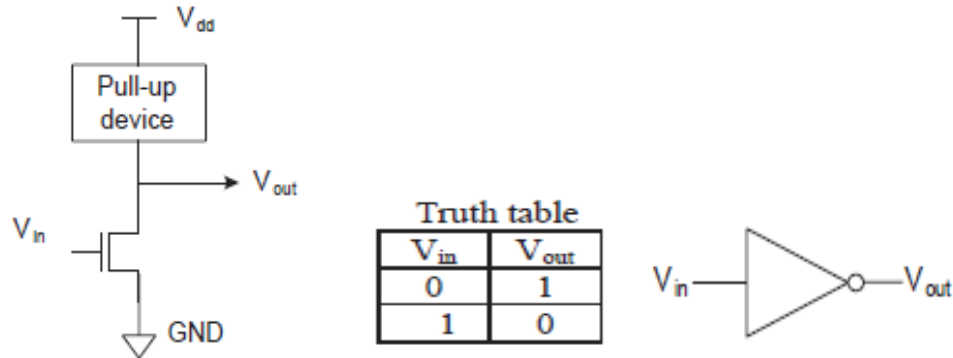
Region III-nMOS is in Cutoff and pMOS in linear, $V_{out} > V_{dd} - V_{tn}$

2. OUTPUT Node Changes from HIGH-to-LOW Level

UNIT-2

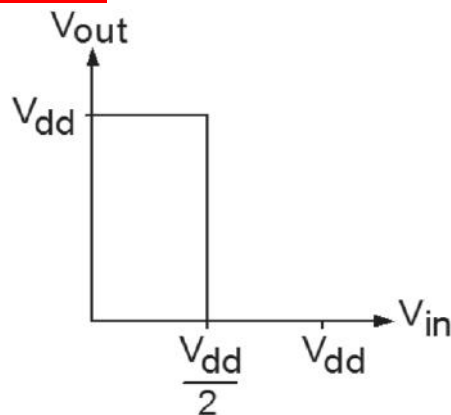
MOS Inverters

❖ Introduction to MOS Inverter:

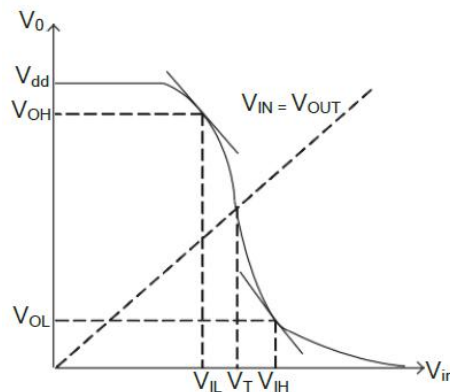


- ✓ An inverter can be realized with the source of an **nMOS enhancement transistor** connected to the ground, and the drain connected to the positive supply rail V_{dd} through a **pull-up device**
- ✓ The input voltage is applied to the gate of the nMOS transistor with respect to ground and output is taken from the drain.
- ✓ When the MOS transistor is ON, it pulls down the output voltage to the low level, and that is why it is called a **pull-down device**, and the other device, which is connected to V_{dd} , is called the **pull-up device**.

❖ MOS Inverter Characteristics



Ideal transfer characteristics of an inverter

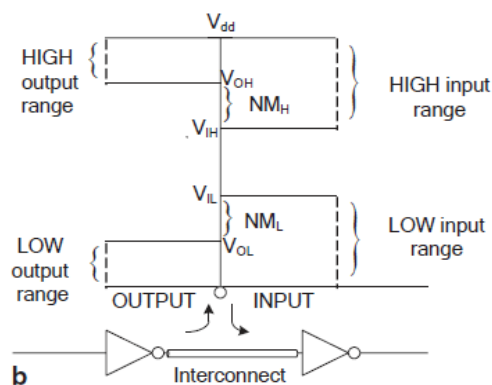


Various voltage levels on the transfer characteristics

- ✓ The output of an ideal inverter changes as the input of the inverter is varied from 0 V (logic level 0) to V_{dd} (logic level 1).
 - ✓ Initially, output is V_{dd} when the output is 0 V, and as the input crosses V_{dd}/2, the output switches to 0 V, and it remains at this level till the maximum input voltage V_{dd}.
 - ✓ This diagram is known as the input–output or *transfer characteristic* of the inverter.
 - ✓ The input voltage, V_{dd}/2, at which the output changes from high ‘1’ to low ‘0’, is known as *inverter threshold voltage*.
- V_{OH}-Maximum Output Voltage Level
V_{IL}-Maximum Input Voltage
V_{OL}-Minimum Output Voltage Level
V_{IH}-Minimum Input Voltage

❖ Noise Margins

It is defined as the allowable noise voltage on the input of a gate so that the output is not affected.



Low- and high-level noise margins

The *low-level noise margin* is defined as the difference in magnitude between the minimum low output voltage of the driving gate and the maximum input low voltage accepted by the driven gate.

$$NM_L = |V_{IL} - V_{OL}|$$

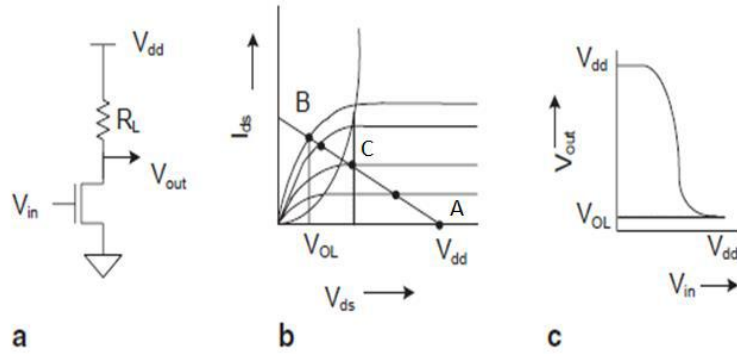
The *high-level noise margin* is defined as the difference in magnitude between the minimum high output voltage of the driving gate and the minimum voltage acceptable as high level by the driven gate:

$$NM_H = |V_{OH} - V_{IH}|$$

❖ MOS Inverter Configurations

- ✓ *Passive Resistive as Pull-up Device*
- ✓ *nMOS Depletion-Mode Transistor as Pull up*
- ✓ *nMOS Enhancement-Mode Transistor as Pull up*
- ✓ *The pMOS Transistor as Pull Up*
- ✓ *pMOS Transistor as a Pull Up in Complementary Mode*
- ✓ *Comparison of the Inverters*

❖ **MOS Inverter Configurations- Passive Resistive as Pull-up Device**



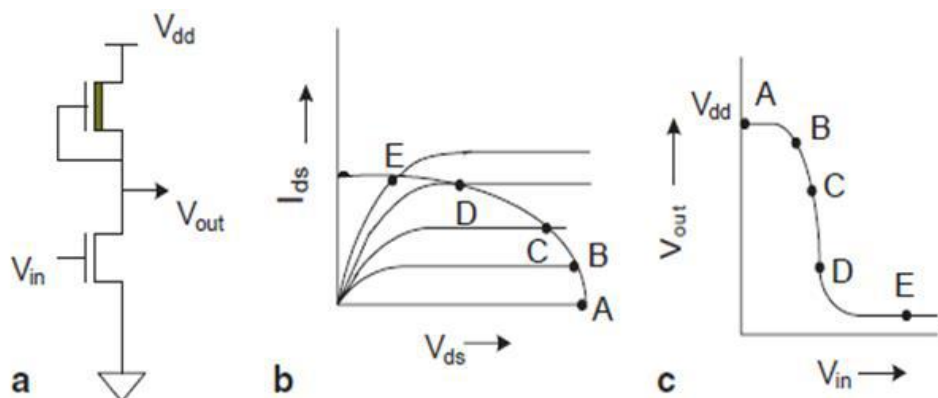
(a) An nMOS Inverter with Resistive Load; (b) Voltage–Current Characteristic; (c) Transfer Characteristic.

- ✓ At Point A, $V_{in}=0, V_{out}=V_{dd}$, nMOS in **Cut-off**
- ✓ At Point C, $V_{in}>V_t, 0 \leq V_{out} \leq V_{dd}$, nMOS in **Saturation**
- ✓ At Point B, $V_{in}=V_{dd}, V_{out}=V_{OL}=V_{dd} \cdot \frac{R_c}{R_c+R_L} \leq V_{tn}=0.2V_{dd}, R_L > 4R_c$, nMOS in **Linear**
- ✓ The resistive load can be fabricated by two approaches—using a diffused resistor approach or using an undoped poly-silicon approach.

Disadvantages

- ✓ Asymmetry in the ON-to-OFF and OFF-to-ON switching times
- ✓ Large Static power dissipation
- ✓ Requires a very large chip area
- ✓ Unsuitable for VLSI realization
- ✓ Strong High Output and Weak Low Output Level

❖ **MOS Inverter Configurations- nMOS Depletion-Mode Transistor as Pull up**



(a) nMOS inverter with depletion-mode transistor as pull-up device; (b) voltage current characteristic; (c) transfer characteristic.

- ✓ $V_{in}=0$ ($V_{in} < V_{tn}$), $V_{out}=V_{dd}$, $I_{ds}=0$, Point A, Pull-down device OFF, Pull-up Device in Linear
- ✓ $V_{in}>V_{tn}$, Point B, Pull-down device in Saturation, Pull-up Device in Linear
 - $I_{pd} = K_n \frac{W_{pd}}{2L_{pd}} (V_{in} - V_{tpd})^2$
 - $I_{pu} = K_n \frac{W_{pu}}{L_{pu}} \left[(V_{out} - V_{tpu}) - \frac{V_{out}}{2} \right] V_{out}$
- ✓ At Point C, Pull-down device in Saturation, Pull-up Device in Saturation
 - $I_{pd} = K_n \frac{W_{pd}}{2L_{pd}} (V_{in} - V_{tpd})^2$
 - $I_{pu} = K_n \frac{W_{pu}}{2L_{pu}} V_{tpu}^2$
- ✓ $V_{in}=V_{dd}$, Point E, Pull-down device in Linear, Pull-up Device in Saturation
 - $I_{pd} = \beta_{pd} \left(V_{in} - V_{tpd} - \frac{V_{OL}}{2} \right) V_{OL}$
 - $I_{pu} = \frac{\beta_{pd}}{2} V_{tpu}^2$

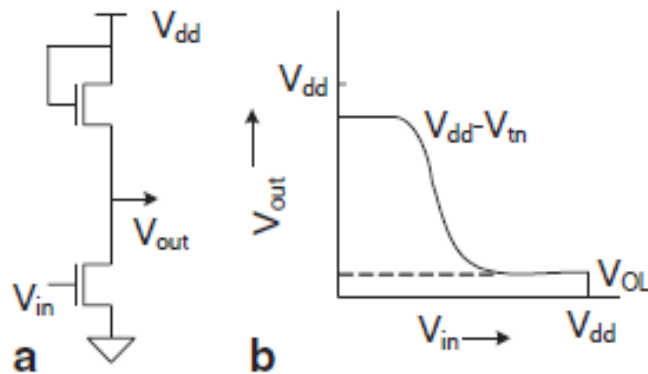
$$\text{Where, } \beta_{pd} = K_n \frac{W_{pd}}{L_{pd}} \text{ and } \beta_{pu} = K_n \frac{W_{pu}}{L_{pu}}$$

Equating the two currents and ignoring $V_{OL} / 2$ term, we get, $V_{OL} = \frac{1}{2k} \frac{V_{tpu}^2}{(V_{dd} - V_{tpd})}$,

$$\text{Where, } K = \frac{\beta_{pd}}{\beta_{pu}} = \frac{(W/L)_{pd}}{(W/L)_{pu}}$$

- The output is not ratioless, which leads to asymmetry in switching characteristics.
- There is static power dissipation when the output logic level is low.
- It produces strong high output level, but weak low output level.

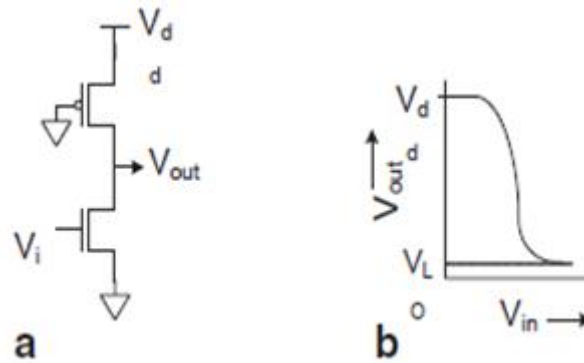
❖ **MOS Inverter Configurations- nMOS Enhancement-Mode Transistor as Pull up**



(a) nMOS inverter with enhance-mode transistor as a pull-up device; (b) transfer characteristic.

- ✓ When $V_{in}=0$, $V_{out}=V_{dd}-V_{tn}$, Pull-down OFF, Pull-up ON
- ✓ When $V_{in}=V_{dd}$, $V_{out}=V_{OL}$, Pull-down ON, Pull-up ON
- ✓ The output is **not ratioless**, which leads to asymmetry in switching characteristics.
- ✓ There is static power dissipation when the output level is low.
- ✓ It produces weak low and high output levels.

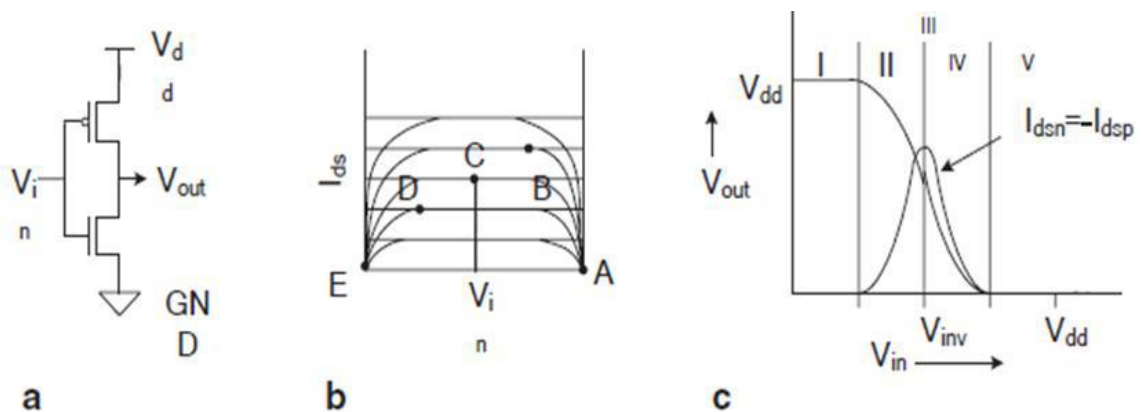
❖ **MOS Inverter Configurations- *p*MOS Enhancement-Mode Transistor as Pull up**



(a) A pseudonMOS inverter; (b) Transfer Characteristic.

- ✓ Functionally similar to a depletion-type nMOS load, it is called a ‘pseudo-nMOS’ inverter.
- ✓ When $V_{in}=0$, $V_{out}=V_{dd}$, Pull-down OFF, Pull-up ON
- ✓ When $V_{in}=V_{dd}$, $V_{out}=V_{OL}$, Pull-down ON, Pull-up ON

❖ **MOS Inverter Configurations- *p*MOS Transistor as a Pull Up in Complementary Mode**



(a) CMOS Inverter; (b) Voltage–Current Characteristic; and (c) Transfer Characteristic

- ✓ When the input voltage $V_{in} = 0V$, pull-up transistor ON, and the pull-down transistor OFF.
- ✓ When the input voltage $V_{in} = V_{dd}$, pull-up transistor OFF, and the pull-down transistor ON.
- ✓ No DC current flow between V_{dd} to ground.
- ✓ However, as the gate voltage is gradually increased from ‘0’ to ‘1’, the **pull-up transistor** switches from **ON to OFF** and the **pull-down transistor** switches from **OFF to ON**. Around the midpoint, **both transistors are ON** and **DC current flows between V_{dd} and ground**.

Region 1: $0 \leq V_{in} < V_{tn}$

- The pull-down transistor is off and the pull-up transistor is in the linear region as shown by a representative point ‘A’

- In this region, there is no DC current flow and output voltage remains close to V_{dd} .

Region 2: $V_{tn} < V_{in} < V_{inv}$

- The pull-down transistor moves into a **saturation region** and the pull-up transistor remains in the **linear region** as represented by point B, when the input is V_{IL} .
- Same current flows through both the devices, $I_{dsp} = -I_{dsn}$

$$I_{dsp} = -\beta_p \left[(V_{in} - V_{dd} - V_{tp})(V_O - V_{dd}) - \frac{1}{2}(V_O - V_{dd})^2 \right]$$

$$\text{where, } \beta_p = K_p \frac{W_p}{L_p}, V_{gsp} = V_{in} - V_{dd} \text{ and } V_{dsp} = V_O - V_{dd}$$

$$I_{dsn} = \beta_n \frac{(V_{in} - V_{tn})^2}{2}, \text{ Where } \beta_n = K_n \frac{W_n}{L_n} \text{ and } V_{gsn} = V_{in}$$

$$V_O = (V_{in} - V_{tp}) \sqrt{(V_{in} - V_{tp}) - 2 \left(V_{in} - \frac{V_{dd}}{2} - V_{tp} \right) V_{dd} - \frac{\beta_n}{\beta_p} (V_{in} - V_{tn})^2}$$

Equating, $I_{dsn} = -I_{dsp}$

$$\beta_n \frac{(V_{in} - V_{tn})^2}{2} = \frac{\beta_p}{2} \left[2(V_{in} - V_{dd} - V_{tp})(V_{out} - V_{dd}) - \frac{1}{2}(V_{out} - V_{dd})^2 \right]$$

Differentiating both sides with respect to V_{in} , we get

$$\frac{\beta_n}{2} (V_{in} - V_{tn}) = \beta_p \left[(V_{in} - V_{dd} - V_{tp}) \frac{dV_{out}}{dV_{in}} + (V_{out} - V_{dd}) - (V_{out} - V_{dd}) \frac{dV_{out}}{dV_{in}} \right]$$

$$\text{Substituting } V_{in} = V_{IL}, \frac{dV_{out}}{dV_{in}} = -1$$

$$\frac{\beta_n}{2} (V_{IL} - V_{tn}) = \beta_p \left[(V_{IL} - V_{dd} - V_{tp})(-1) + (V_{out} - V_{dd}) - (V_{out} - V_{dd})(-1) \right]$$

$$\frac{\beta_n}{2} (V_{IL} - V_{tn}) = \beta_p \left[-V_{IL} + V_{dd} + V_{tp} + V_{out} - V_{dd} + V_{out} - V_{dd} \right]$$

$$\frac{\beta_n}{2} (V_{IL} - V_{tn}) = \beta_p \left[-V_{IL} + V_{tp} + 2V_{out} \right]$$

$$\text{or } V_{IL} = \frac{(2V_{out} + V_{tp} - V_{dd} + \left(\frac{\beta_n}{\beta_p}\right) \cdot V_{tn})}{\left(1 + \left(\frac{\beta_n}{\beta_p}\right)\right)}$$

$$\text{For } \frac{\beta_n}{\beta_p} = 1 \text{ and } V_{out} \approx V_{dd}$$

$$V_{IL} = \frac{1}{8} (3V_{dd} + 2V_{tn})$$

Region 3: $V_{in} = V_{inv}$

At this point, both the transistors are in the saturation condition as represented by the point C.

$$V_{gs}^{pd} = V_{in}$$

$$V_{gs}^{pu} = V_{in} - V_{dd} = V_{inv} - V_{dd}$$

$$I_{dsn} = \frac{1}{2} K_n \frac{W_n}{L_n} (V_{inv} - V_{tn})^2$$

$$I_{dsp} = \frac{1}{2} K_p \frac{W_p}{L_p} (V_{inv} - V_{dd} - V_{tp})^2$$

Equating

$$\frac{\beta_n}{2} (V_{inv} - V_{tn})^2 = -\frac{\beta_p}{2} (V_{inv} - V_{dd} - V_{tp})^2$$

$$\frac{V_{inv} - V_{dd} - V_{tp}}{V_{inv} - V_{tn}} = -\sqrt{\frac{\beta_n}{\beta_p}}$$

$$V_{inv} \left(1 + \sqrt{\frac{\beta_n}{\beta_p}} \right) = V_{dd} + V_{tp} + V_{dd} + V_{tn} \sqrt{\frac{\beta_n}{\beta_p}}$$

$$V_{inv} = \frac{V_{dd} + V_{tp} + V_{dd} + V_{tn} \sqrt{\frac{\beta_n}{\beta_p}}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}$$

$$\text{for } \beta_n = \beta_p \text{ and } V_{tn} = -V_{tp}, V_{inv} = \frac{V_{dd}}{2}$$

Region 4: $V_{inv} < V_{in} \leq V_{dd} - |V_{tp}|$

The nMOS transistor moves from the saturation region to the linear region, whereas the pMOS transistor remains in saturation.

$$I_{dsn} = \beta_n \left[(V_{in} - V_{tn})V_o - \frac{V_o^2}{2} \right] \text{ and } I_{dsp} = -\beta_p (V_{in} - V_{dd} - V_{tp})^2$$

$$\text{Equating, } I_{dsn} = -I_{dsp}$$

$$\beta_n \left[(V_{in} - V_{tn})V_o - \frac{V_o^2}{2} \right] = \beta_p (V_{in} - V_{dd} - V_{tp})^2$$

$$\text{Equating, } I_{dsn} = I_{dsp}$$

$$\frac{\beta_n}{2} [2(V_{gsn} - V_{tn})V_{gsn} - V_{gsn}^2] = \frac{\beta_p}{2} (V_{gsp} - V_{tp})^2$$

Substituting, $V_{gsp} = -(V_{dd} - V_{in})$ and $V_{dsp} = -(V_{dd} - V_{out})$

$$\frac{\beta_n}{2} [2(V_{in} - V_{tn})V_{out} - V_{out}^2] = \frac{\beta_p}{2} (V_{in} - V_{dd} - V_{tp})^2$$

Differentiating both sides with respect to V_{in} , we get

$$\beta_n \left[(V_{in} - V_{tn}) \frac{dV_{out}}{dV_{in}} + V_{out} - V_{out} \left(\frac{dV_{out}}{dV_{in}} \right) \right] = \beta_p (V_{in} - V_{dd} - V_{tp})$$

$$\text{Substituting } V_{in} = V_{IH}, \frac{dV_{out}}{dV_{in}} = -1$$

$$V_{IH} = \frac{V_{dd} + V_{tp} + \left(\frac{\beta_n}{\beta_p}\right) \cdot (2V_{out} + V_{tn})}{1 + \left(\frac{\beta_n}{\beta_p}\right)}$$

$$\text{For } \frac{\beta_n}{\beta_p} = 1$$

$$V_{IH} = \frac{1}{8}(5V_{dd} - 2V_{tn})$$

For a symmetric inverter, $V_{IH} + V_{IL} = V_{dd}$

$$NM_L = V_{IL} - V_{OL} = V_{IL}$$

$$NM_H = V_{OH} - V_{IH} = V_{dd} - V_{IH} = V_{IL}$$

Region 5:

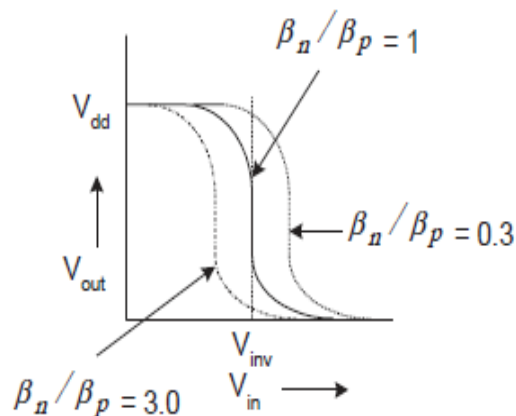
In this region, the pull-up pMOS transistor remains OFF and the pull-down nMOS transistor goes to deep saturation.

However, the current flow through the circuit is zero as the p transistor is OFF and the output voltage $V_O = 0$.

❖ Key features of the CMOS inverter

- ✓ It may be noted that unlike the use of nMOS enhancement- or depletion-mode transistor as a pull-up device, in this case, there is no current flow either for '0' or '1' inputs. So, there is no static power dissipation.
- ✓ Current flows only during the transition period. So, the static power dissipation is very small.
- ✓ Moreover, for low and high inputs, the roll of the pMOS and nMOS transistors are complementary; when one is OFF, the other one is ON. That is why this configuration is known as the complementary MOS or CMOS inverter.
- ✓ Another advantage is that full high and low levels are generated at the output.
- ✓ Moreover, the output voltage is independent of the relative dimensions of the pMOS and nMOS transistors. In other words, the CMOS circuits are ratioless.

❖ β_n/β_p Ratio:



- ✓ As we have mentioned earlier, the low- and high-level outputs of a CMOS inverter are not dependent on the inverter ratio.
- ✓ However, the transfer characteristic is a function of the β_n/β_p ratio.

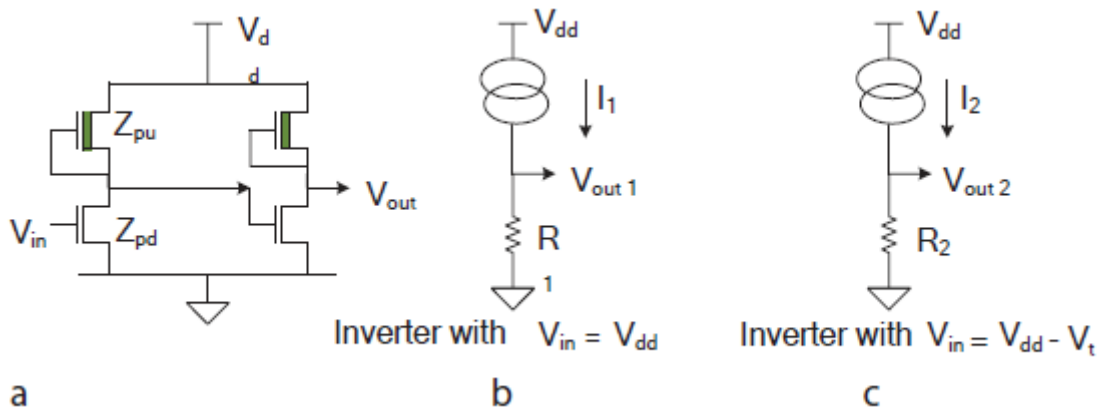
- ✓ The transfer characteristics for three different ratio values are plotted in Figure.
- ✓ Here, we note that the voltage at which the gate switches from high to low level (V_{inv}) is dependent on the β_n/β_p ratio.
- ✓ V_{inv} increases as β_n/β_p decreases.
- ✓ For a given process technology, the β_n/β_p can be changed by changing the channel dimensions, i.e., the channel length and width. Keeping L the same, if we increase W_n/W_p ratio, the transition moves towards the left and as W_n/W_p is decreased, the transition moves towards the right as show in Figure.

❖ **Comparison of the Inverters**

Inverters	V_{LO}	V_{HI}	Noise Margin	Power
Resistor	Weak	Strong	Poor for low	High
nMOS depletion	Weak	Strong	Poor for low	High
nMOS enhancement	Weak	Weak	Poor for low and high	High
Psuedo-nMOS	Weak	Strong	Poor for low	High
CMOS	Strong	Strong	Good	Low

❖ **Inverter ratio in Different Situations**

- ✓ An nMOS inverter driven by another inverter

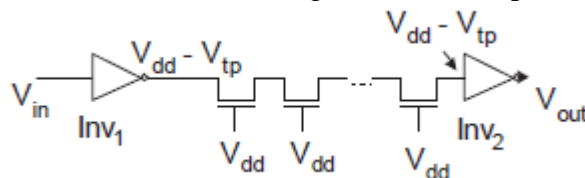


(a) An nMOS inverter driven by another inverter; (b) inverter with $V_{in} = V_{dd}$; and (c) inverter with $V_{in} = V_{dd} - V_t$

- ✓ Assuming $Z_{pd} = L_{pd} / W_{pd}$ and $Z_{pu} = L_{pu} / W_{pu}$, where Z is known as the aspect ratio of the MOS devices

$$\text{Inverter Ratio} = \frac{Z_{pu}}{Z_{pd}} = \frac{4}{1}$$

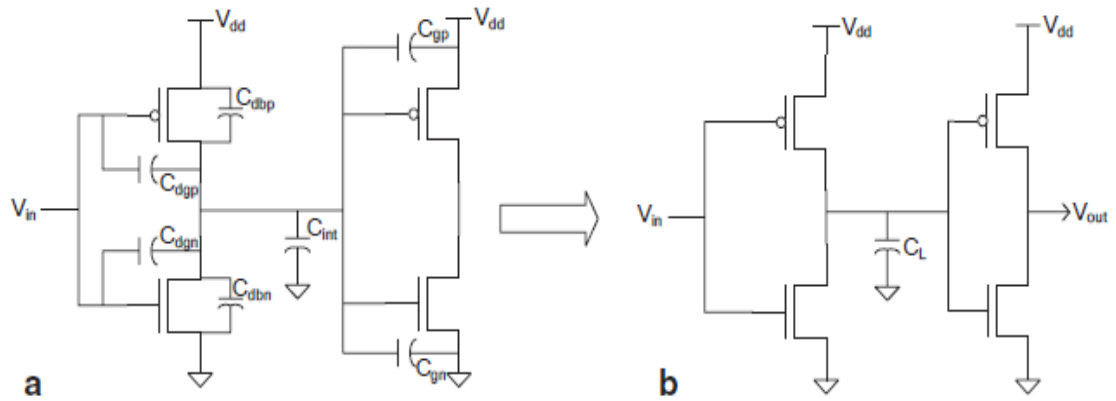
- ✓ An inverter driven through one or more passtransistors



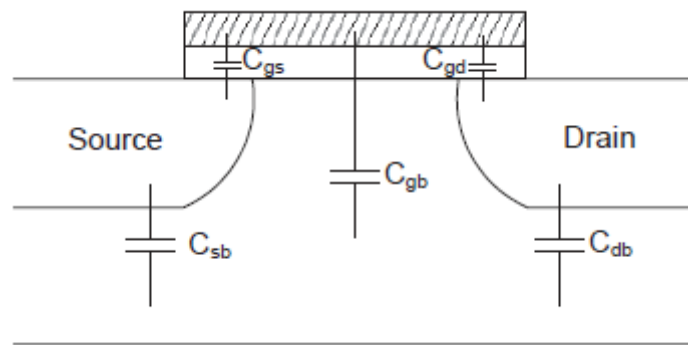
$$\text{Inverter Ratio} = \frac{Z_{pu2}}{Z_{pd2}} = \frac{4.0 Z_{pu1}}{2.5 Z_{pd1}} \text{ or } \frac{Z_{pu2}}{Z_{pd2}} \approx 2 \cdot \frac{Z_{pu1}}{Z_{pd1}} = \frac{8}{1}$$

$$\text{Inverter Ratio} = \frac{Z_{pu}}{Z_{pd}} \geq 8/1$$

❖ Switching Characteristics

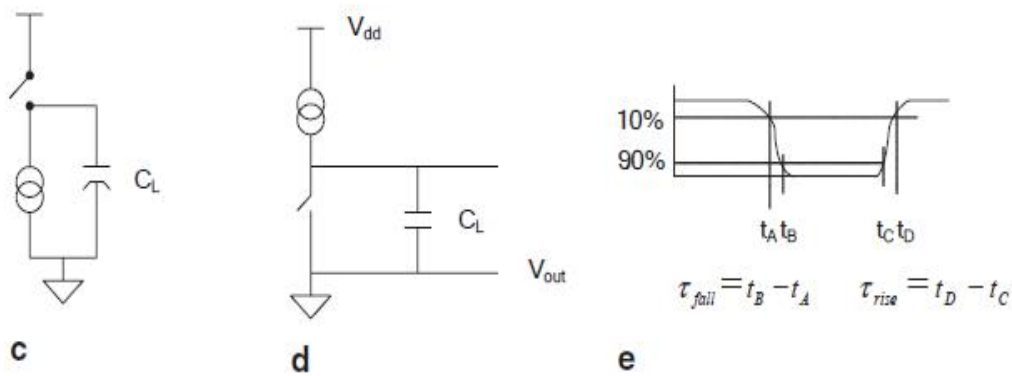
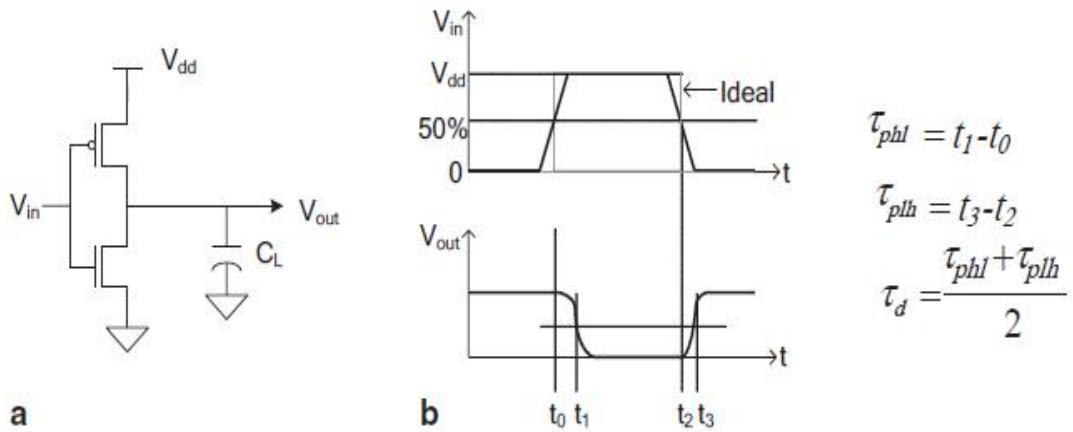


(a) Parasitic capacitances of a CMOS inverter. (b) CMOS Complementary metal-oxide-semiconductor



Internal Parasitic Capacitances of an MOS Transistor.

- ✓ Equivalent lumped capacitance $C_L = C_{dgn} + C_{dgp} + C_{dbn} + C_{dbp} + C_{int} + C_{gn} + C_{gp}$
- ✓ Estimation of load capacitance
- ✓ Delay-Time Estimation



Fall time delay

$$t_{pfl} = \frac{V_{dd}}{\beta_n} \times \frac{C_L}{(V_{dd} - V_{tn})^2}$$

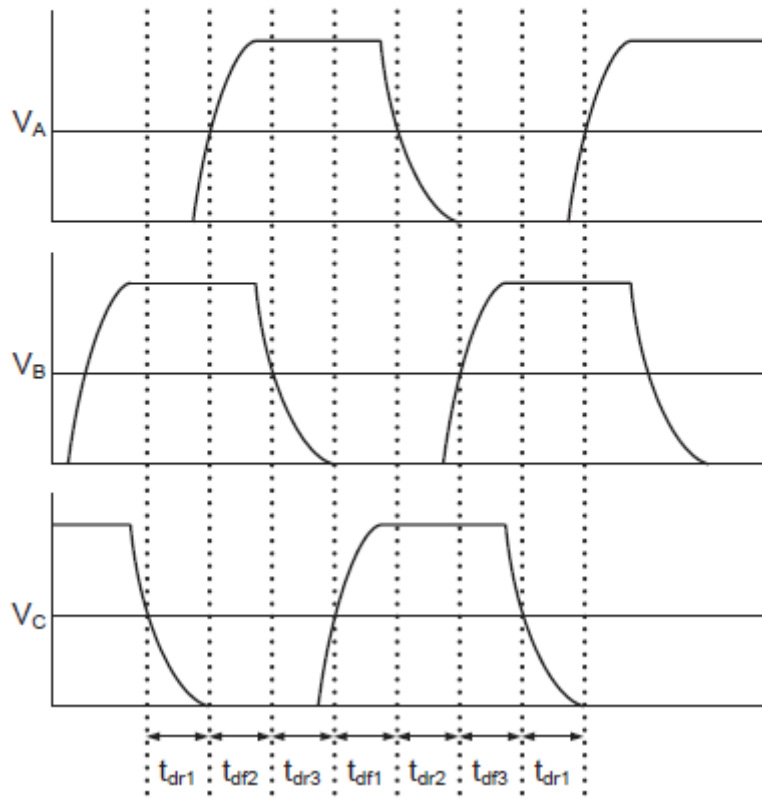
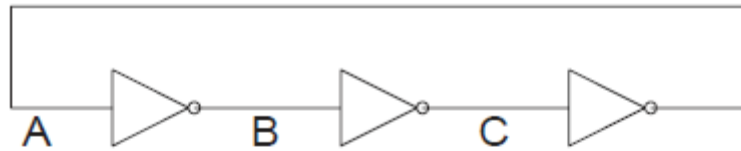
Rise time delay

$$t_{plh} = \frac{C_L}{\beta_p} \times \frac{1}{V_{dd} \left(1 - \left|\frac{V_{tp}}{V_{dd}}\right|\right)^2}$$

Delay time: By taking average of the rise and full delay time

$$t_d = \left[\frac{L_n}{K_n W_n} + \frac{L_p}{K_p W_p} \right] \frac{C_L}{V_{dd} \left(1 - \frac{V_t}{V_{dd}}\right)^2}$$

✓ Ring-Oscillator



Output waveform of a three-stage ring oscillator

- ✓ The time period can be expressed as the sum of the six delay times

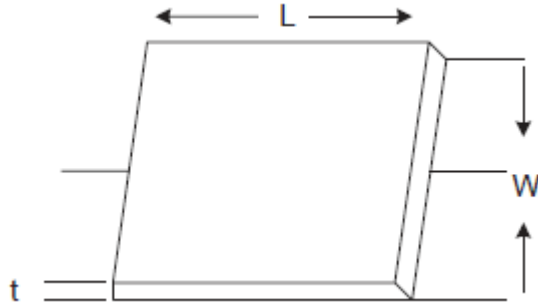
$$T = t_{ph1} + t_{ph2} + t_{ph3} + t_{ph4} + t_{ph5} + t_{ph6}$$

$$T = 6t_d = 2.3 t_d$$

- ✓ For an n-stage inverter, the Time Period $T = 2.n.t_d$, Frequency of oscillation $f = 1/2nt_d$ or $t_d = 1/2nf$
- ✓ Used for on-chip clock generation
- ✓ It does not provide a stable or accurate clock frequency due to dependence on temperature and other parameters
- ✓ To generate stable and accurate clock frequency, an off-chip crystal is used to realize a crystal oscillator

❖ Delay Parameters

- ✓ Various parameters such as Resistance and Capacitance of the transistors along with wiring and parasitic capacitances
- ✓ **Resistance Estimation**



One slab of conducting material

$$R_{AB} = \frac{\rho L}{t \cdot W} = \frac{\rho L}{A} \Omega, \text{ where } A \text{ is the cross section area}$$

Consider the case in which $L=W$, then

$$R_{AB} = \frac{\rho}{t}$$

$= R_s \Omega$, where R_s is defined as the resistance per square or the sheet resistance

$$I_{ds} = \beta \left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2} \right] =$$

Assuming $V_{ds} \ll (V_{gs} - V_t)$, $I_{ds} = \beta(V_{gs} - V_t) \cdot V_{ds}$,

$$R_c = \frac{V_{ds}}{I_{ds}} = \frac{1}{\beta(V_{gs} - V_t)}$$

$$R_c = \frac{1}{\mu C_g (V_{gs} - V_t)} \frac{L}{W} = K \left(\frac{L}{W} \right), \text{ where } K = \frac{1}{\mu C_g (V_{gs} - V_t)}$$

- K May take the value between 1000 to 3000 Ω/sq .
- Sheet Resistance(Ohm/Sq.) of different conductors

Layer	Min.	Typical	Max.
Metal	0.03	0.07	0.1
Diffusion	10	25	100
Silicide	2	3	6
Poly-silicon	15	20	30
n-channel	-	10_4	-
P-channel	-	$2.5 \times 10_4$	-

✓ **Area Capacitance of Different Layers**

$$C = \frac{\epsilon_0 \epsilon_{ins} A}{D} \text{ Farads}$$

Where D is the thickness of the silicon dioxide

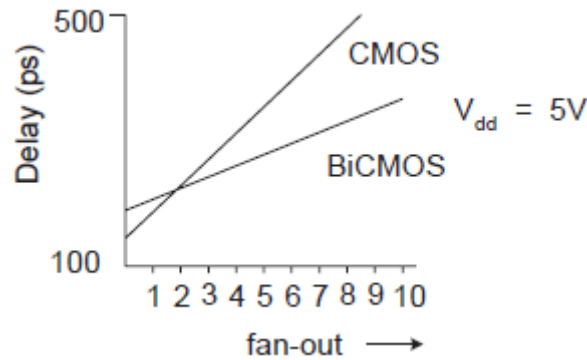
A is the Area of Place

ϵ_0 is the relative permittivity of SiO_2

$\epsilon_{ins} = 8.85 \times 10^{-14} \text{F.cm}$, permittivity of free space

Capacitance of different materials

Capacitance	Value of $\text{pF}/\mu\text{m}^2$	Relative Value
Gate to channel	4×10^{-4}	1
Diffusion	1×10^{-4}	0.25
Poly-Silicon	4×10^{-4}	0.1
Metal 1	0.3×10^{-4}	0.075
Metal 2	0.2×10^{-4}	0.50
Metal 2 To Metal	0.4×10^{-4}	0.15
Metal2 To Poly	0.3×10^{-4}	0.075



Delay of static CMOS and BiCMOS for different fan-out

✓ **Buffer Sizing**

The Minimum total delay is

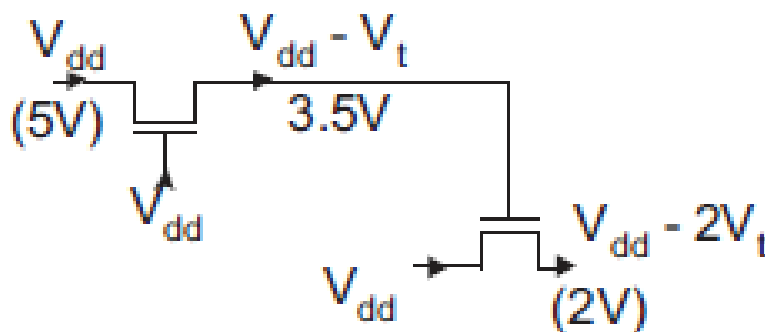
$$t_{min} = e\tau \ln \left[\frac{C_L}{C_g} \right]$$

MOS Combinational Circuits

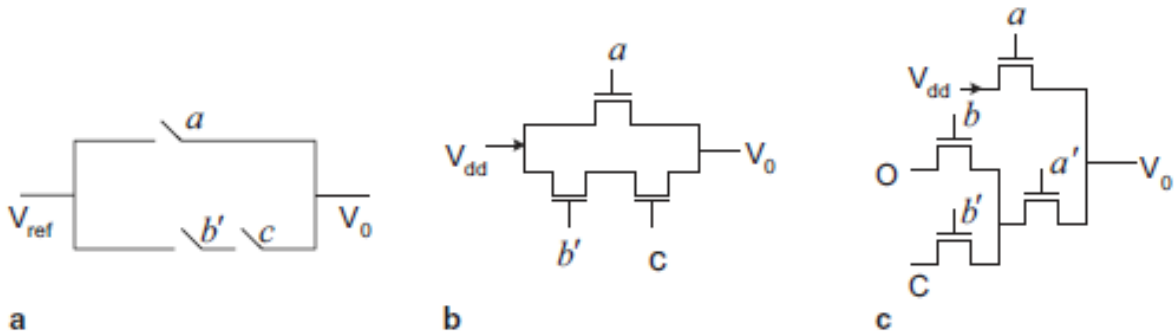
Introduction:

- ✓ There are two basic approaches of realizing digital circuits by metal–oxide–semiconductor (MOS) technology: **switch logic and gate logic**.
- ✓ A **switch logic** is based on the use of “*pass transistors*” or transmission gates, just like relay contacts, to steer logic signals through the device.
- ✓ On the other hand, **gate logic** is based on the realization of digital circuits using inverters and other conventional gates, as it is typically done in **transistor–transistor logic (TTL)** circuits.
- ✓ Moreover, depending on how circuits function, they can also be categorized into two types: **static and dynamic gates**.
- ✓ In case of **static gates**, no clock is necessary for their operation and the output remains steady for as long as the supply voltage is maintained.
- ✓ Dynamic circuits are realized by making use of the information storage capability of the intrinsic capacitors present in the MOS circuits.

❖ Pass-Transistor logic:



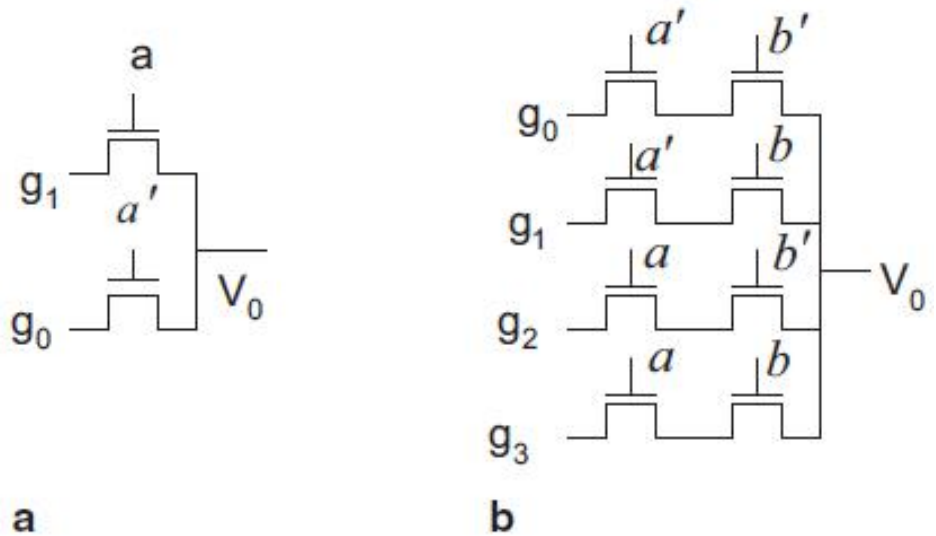
Pass-transistor Output Driving Another Pass-transistor Stage



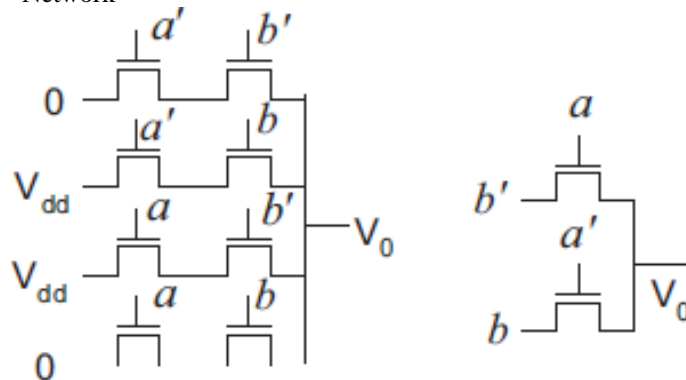
(a) Relay logic to realize $f = a + b'c$.
 (b) Pass-transistor network corresponding to relay logic.
 (c) Proper pass-transistor network for $f = a + b'c$

✓ **Realizing Pass-Transistor Logic**

- Realization of 2-to-1 and 4-to-1 Multiplexer
- $f(a,b) = g_0a'b' + g_1a'b + g_2ab' + g_3ab$



(a) A 2-to-1 Multiplexer. (b) A 4-to-1 Multiplexer Circuit using Pass-transistor Network



(a) Multiplexer realization of $f = a'b + ab'$
 (b) Minimum Transistor Pass-Transistor Realization of $f = a'b + ab'$

✓ **Advantages and Disadvantages**

- (a) Ratioless:
- (b) Powerless:
- (c) Lower area

However, pass-transistor logic suffers from the following disadvantages:

1. When a signal is steered through several stages of pass transistors, the delay can be considerable.
2. There is a voltage drop ($V_{out} = V_{dd} - V_{tn}$) as we steer the signal through nMOS transistors. This reduced level leads to high static currents at the subsequent output inverters and logic gates. In order to avoid this, it is necessary to use additional hardware known as the *swing restoration logic* at the gate output.
3. Pass-transistor structure requires complementary control signals. Dual-rail logic is usually necessary to provide all signals in the complementary form. As a consequence, two MOS networks are again required in addition to the swing restoration and output buffering circuitry.

✓ **Pass Transistor Logic families**

- Complementary Pass-Transistor Logic(CPL)
- Swing-Restored Pass-Transistor Logic(SRPL)
- Double Pass-Transistor Logic(DPL)
- Single-Rail Pass-Transistor Logic(SRPL)

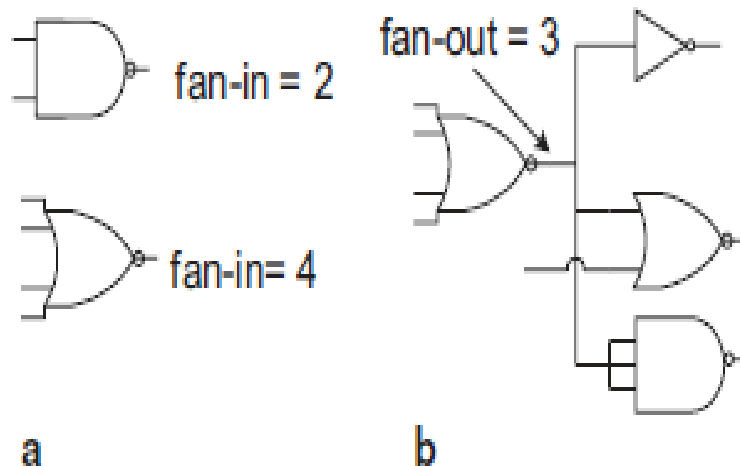
Qualitative Comparisons of the Logic Styles

Logic Style	#MOS Networks	Output Driving	I/O Decoupling	Swing Restorations	#Rails	Robustness
CMOS	2n	Med/ good	Yes	No	Single	High
CPL	2n+6	Good	Yes	Yes	Dual	Medium
SRPL	2n0+4	Poor	No	Yes	Dual	Low
DPL	4n	Good	Yes	No	Dual	High
LEAP	n+3	Good	Yes	Yes	Single	Medium
DCVSPG	2n+2	Medium	Yes	No	Dual	Medium

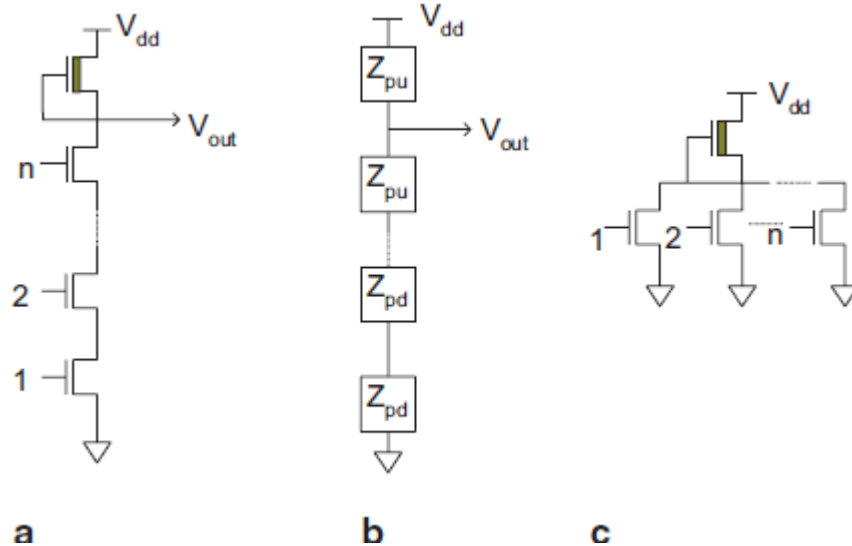
❖ **Gate logic:**

✓ **Fan-In and Fan-Out**

- Fan-in is the number of signal inputs that the gate processes to generate some output.
- The fan-out is the number of logic inputs driven by a gate



✓ **nMOS NAND and NOR Gates**



(a) n -input nMOS NAND gate; (b) equivalent circuits; and (c) n -input nMOS NOR gate

- For nMOS NAND Gate,

- Aspect ratio $\frac{Z_{pu}}{Z_{n_{pd}}} \geq \frac{4}{1}$

- Requires a considerably larger area than those of the corresponding nMOS inverter

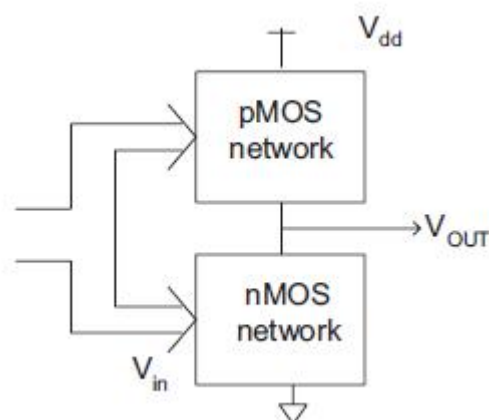
- For nMOS NOR Gate,

- Aspect ratio of the pull-up to any pull down transistor will be the same as that of an inverter, irrespective of the number of inputs of the NOR gate

- Requires a reasonable area because of pull-up transistor geometry is not affected by the number of inputs.

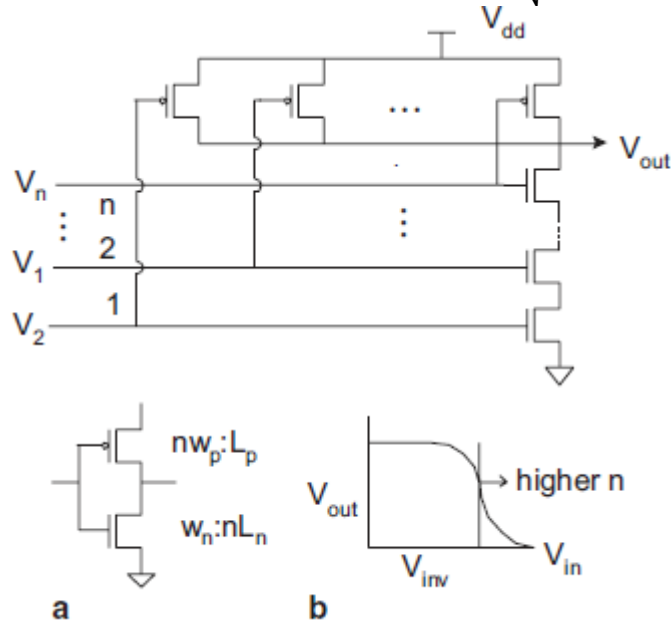
- Use of NOR gate in circuit realization is preferred compared to that of NAND gate, when there is a choice.

✓ **CMOS Realization-General CMOS network**



- **CMOS NAND Gates- n -input CMOS NAND gate**

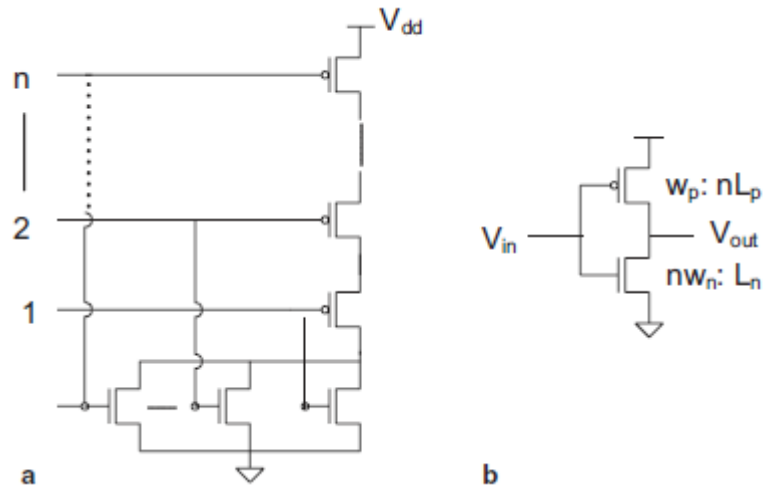
$$V_{inv} = \frac{V_{dd} + V_{dd} + \frac{V_{th}}{n} \sqrt{\frac{\beta_n}{\beta_p}}}{1 + \frac{1}{n} \sqrt{\frac{\beta_n}{\beta_p}}}$$



(a) Equivalent Circuit of n -input Complementary MOS (CMOS) NAND Gate; and (b) Transfer Characteristics of n -input CMOS NAND Gate

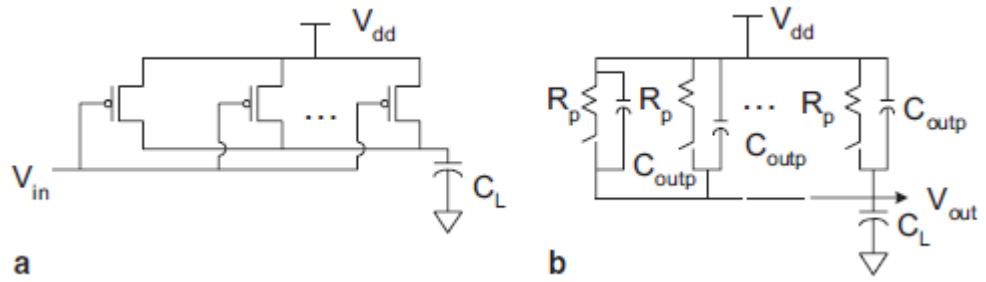
○ **CMOS NOR Gates**

$$V_{inv} = \frac{V_{dd} + V_{dd} + nV_{th} \sqrt{\frac{\beta_n}{\beta_p}}}{1 + n \sqrt{\frac{\beta_n}{\beta_p}}}$$



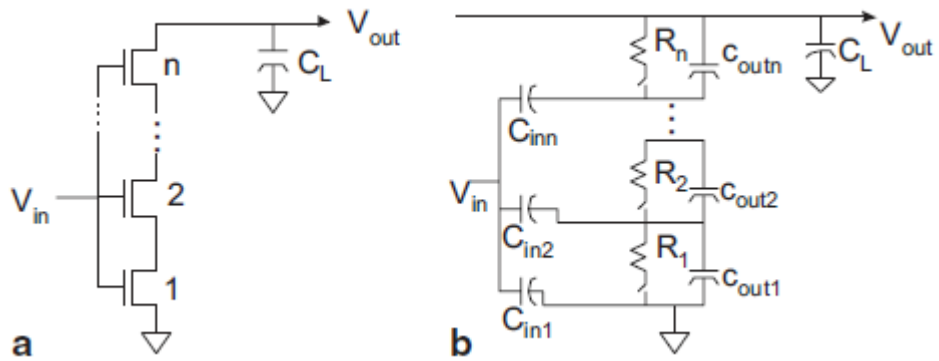
(a) n -input Complementary MOS (CMOS) NOR gate and (b) The Equivalent Circuit

✓ **Switching Characteristics**



(a) Pull-up transistor tied together with a load capacitance; and (b) equivalent circuit
Intrinsic time constant

$$t_{dr} = \frac{R_n}{n} (n C_{outp}) + \frac{R_p}{n}$$



(a) Pull-down transistors along with load capacitance C_L , and (b) equivalent circuit
Fall time

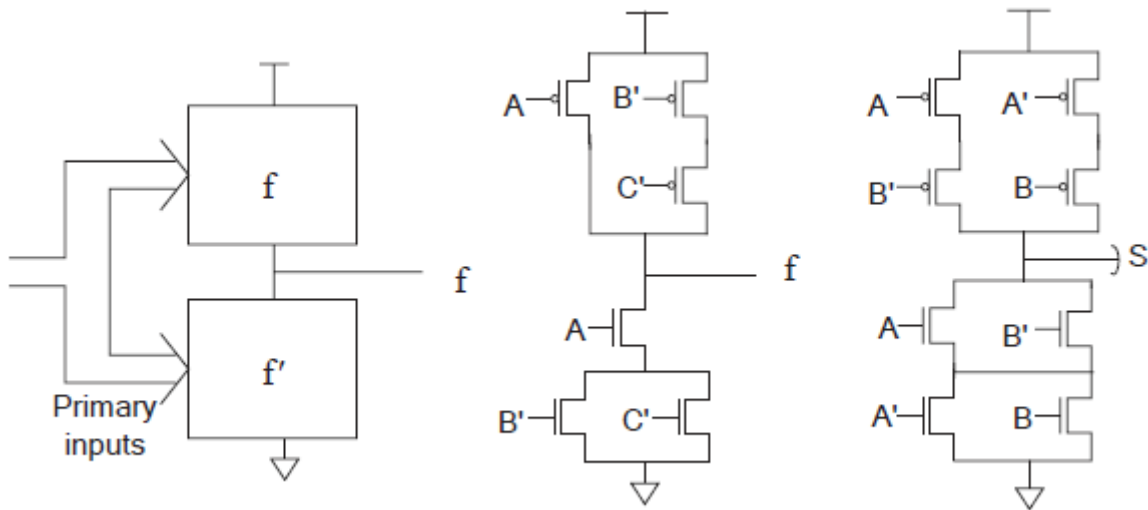
$$t_{df} = nR_n \left(\frac{C_{outn}}{n} + n C_{outp} + C_L \right) + 0.35R_p C_{inn} (n - 1)^2$$

✓ **CMOS NOR Gate**

$$t_{df} = \frac{R_n}{n} (n C_{outp}) + C_L$$

$$t_{dr} = nR_n \left(\frac{C_{outp}}{n} + n C_{outn} + C_L \right) + 0.35R_p C_{inn} (n - 1)^2$$

✓ **CMOS Complex Logic Gates**



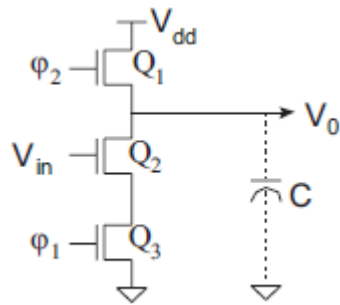
- (a) Realization of a function f by complementary MOS (CMOS) gate;
 (b) Realization of $f = A' + BC$;
 (c) Realization of $S = A'B + AB'$

❖ **MOS Dynamic Circuits:**

- ✓ In static circuits, the output voltage levels remain unchanged as long as inputs are kept the same and the power supply is maintained.
- ✓ nMOS static circuits have two disadvantages:
 - They draw static current as long as power remains ON, and they require larger chip area because of “ratioed” logic.
- ✓ These two factors contribute towards slow operation of nMOS circuits.
- ✓ Although there is no static power dissipation in a full-complementary CMOS circuit, the logic function is implemented twice, one in the pull-up p-network and the other in the pull-down n-network.
- ✓ Due to the extra area and extra number of transistors, the load capacitance on gates of a full-complementary CMOS is considerably higher.
- ✓ As a consequence, speeds of operation of the CMOS and nMOS circuits are comparable.
- ✓ The CMOS not only has twice the available current drive but also has twice the capacitance of nMOS.
- ✓ The trade-off in choosing one or the other is between the lower power of the CMOS and the lower area of nMOS (or pseudo nMOS).
- ✓ In MOS circuits, the capacitances need not be externally connected.
- ✓ Excellent insulating properties of silicon dioxide provide very good quality gate-to-channel capacitances, which can be used for information storage.
- ✓ The advantage of low power of full-complementary CMOS circuits and smaller chip area of nMOS circuits are combined in dynamic circuits leading to circuits of smaller area and lower power dissipation.
- ✓ MOS dynamic circuits are also faster in speed.
- ✓ However, these are not free from disadvantages.
- ✓ Like any other capacitors, charge stored on MOS capacitors also leak.
- ✓ To retain information, it is necessary to periodically restore information by a process known as *refreshing*.
- ✓ There are other problems like *charge sharing* and *clock skew* leading to hazards and races.

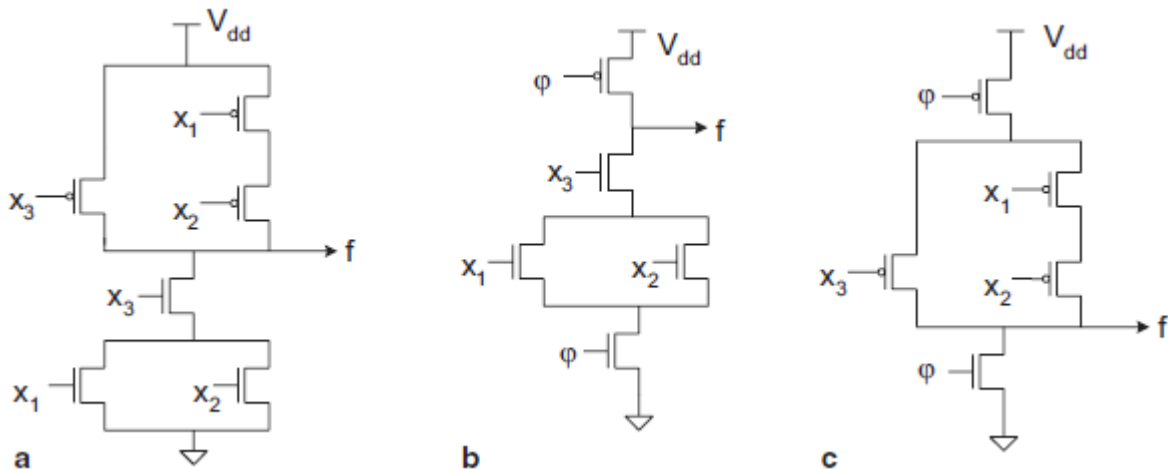
✓ **Single-Phase Dynamic Circuits**

(a) Two-phase clock; and (b) A Two-Phase Clock Generator Circuit



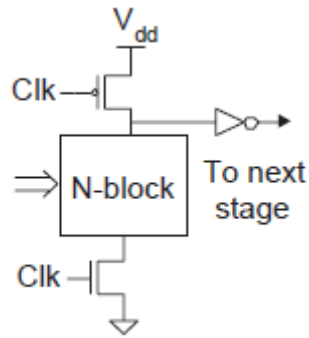
Two-Phase n-type MOS (nMOS) Inverter

✓ **CMOS Dynamic Circuits**



Realization of function $f = x_3(x_1 + x_2)$ using (a) static complementary MOS (CMOS), (b) dynamic CMOS with n-block, and (c) dynamic CMOS with p-block

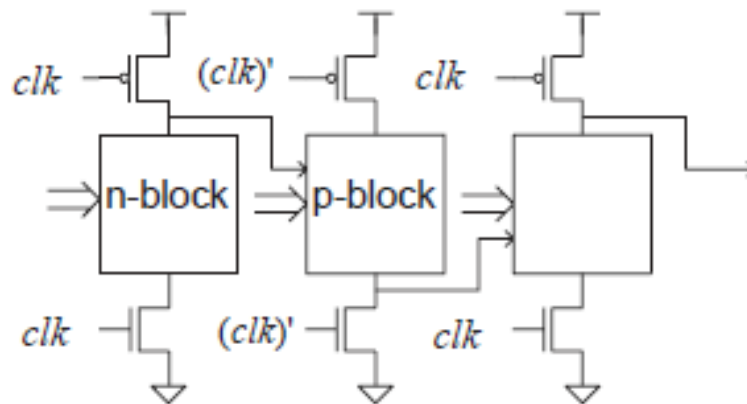
- ✓ **Advantages and Disadvantages**
- ✓ The number of transistors required for a circuit with fan-in N is $(N + 2)$, in contrast to $2N$ in case of state CMOS circuit.
- ✓ Not only dynamic circuits require $(N + 2)$ MOS transistors but also the load capacitance is substantially lower than that for static CMOS circuits.
- ✓ This is about 50 % less than static CMOS and is closer to that of nMOS (or pseudo nMOS) circuits.
- ✓ But, here full pull-down (or pull-up) current is available for discharging (or charging) the output capacitance.
- ✓ Therefore, the speed of the operation is faster than that of the static CMOS circuits.
- ✓ Moreover, dynamic circuits consume static power closer to the static CMOS.
- ✓ Therefore, dynamic circuits provide superior (area-speed product) performance compared to its static counterpart.
- ✓ For example, a dynamic NOR gate is about five times faster than the static CMOS NOR gate.
- ✓ The speed advantage is due to smaller output capacitance and reduced overlap current.
- ✓ Disadvantages
 - **Charge Leakage Problem**
 - **Charge Sharing Problem**
 - **Clock Skew Problem**
- ✓ **Domino CMOS Circuits**



Domino CMOS circuits have the following advantages:

- Since no DC current path is established either during the pre-charge phase or during the evaluation phase, domino logic circuits have lower power consumption.
- As n-block is only used to realize the circuit, domino circuits occupy lesser chip area compared to static CMOS circuits.
- Due to lesser number of MOS transistors used in circuit realization, domino CMOS circuits have lesser parasitic capacitances and hence faster in speed compared to static CMOS.

✓ **NORA Logic**



UNIT-3

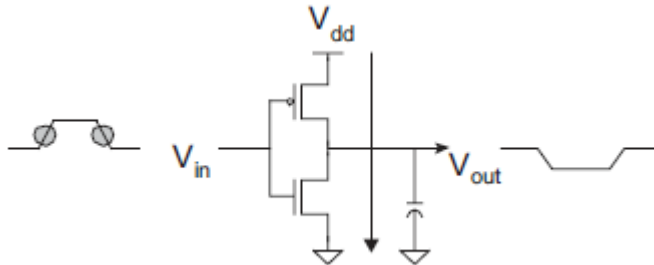
Sources of Power Dissipation

❖ Introduction:

- ✓ Static Power Dissipation
- ✓ Dynamic Power Dissipation

❖ Short-circuit Power Dissipation

- ✓ Short-circuit power dissipation occurs when both the nMOS and pMOS networks are ON.
- ✓ This can arise due to slow rise and fall times of the inputs



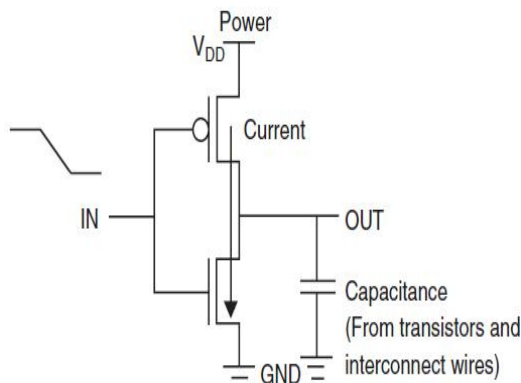
Transition

Short-circuit Power Dissipation During Input

$$P_{sc} = V_{dd} I_{mean} = \frac{\beta}{12} (V_{dd} - 2V_t)^3 \tau f = \frac{\beta}{12} V_{dd}^3 \left(1 - 2 \frac{V_t}{V_{dd}}\right)^3 \tau f$$

❖ Switching Power Dissipation

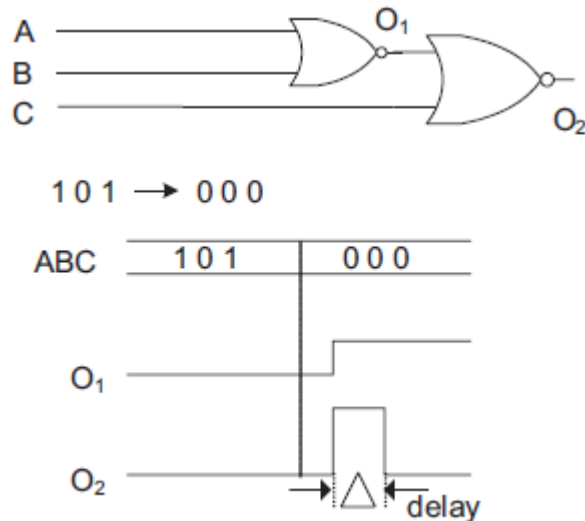
- ✓ As the input and output values keep on changing, capacitive loads at different circuit points are charged and discharged, leading to power dissipation.
- ✓ This is known as *switching power* dissipation



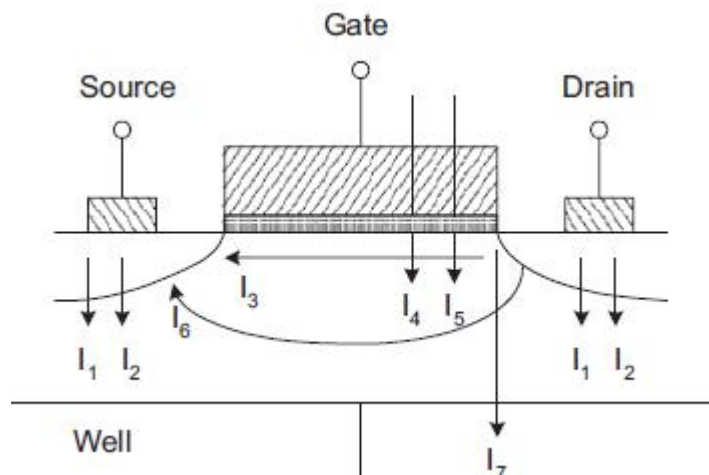
$$P_{shortcircuit} = t_{sc} \times V_{dd} \times I_{peak} \times f_{clock} = \frac{\mu \epsilon_{ox} W}{12LD} \times (V_{dd} - V_{th})^3 \times t_{sc} \times f_{clock}$$

❖ Glitching Power Dissipation

- ✓ Due to a finite delay of the logic gates, there are spurious transitions at different nodes in the circuit.
- ✓ Apart from the abnormal behavior of the circuits, these transitions also result in power dissipation known as glitching power dissipation.



❖ Leakage Power Dissipation



- I_1 is the *reverse-bias p-n junction diode leakage current*;
- I_2 is the reverse-biased p-n junction current due to *tunneling* of electrons from the valence bond of the *p* region to the conduction bond of the *n* region;
- I_3 is the *subthreshold leakage current* between the source and the drain when the gate voltage is less than the threshold voltage V_t ;
- I_4 is the *oxide-tunneling current* due to a reduction in the oxide thickness;
- I_5 is gate current due to *hot-carrier injection* of electrons;
- I_6 is the *GIDL current* due to a high field effect in the drain junction;
- I_7 is the *channel punch-through current* due to the close proximity of the drain and the source in short-channel devices.

❖ Static power dissipation occurs due to various leakage mechanisms.

- Reverse-bias p-n junction diode leakage current
- Reverse-biased p-n junction current due to the tunneling of electrons from the valence bond of the *p* region to the conduction bond of the *n* region, known as band-to-band-tunneling current

- Subthreshold leakage current between source and drain when the gate voltage is less than the threshold voltage V_t . Various mechanisms which affect the subthreshold leakage current are:
 1. Drain-induced barrier lowering (DIBL)
 2. Body effect
 3. Narrow-width effect
 4. Effect of channel length and V_{th} roll-off
 5. Effect of temperature
- Oxide-tunneling current due to a reduction in the oxide thickness
- Gate current due to a hot-carrier injection of electrons
- Gate-induced drain-leakage (GIDL) current due to high field effect in the drain junction
- Channel punch-through current due to close proximity of the drain and the source in short-channel devices

❖ Supply Voltage Scaling for Low Power

- ✓ *Static Voltage Scaling (SVS)* In this case, fixed supply voltages are applied to one
 - ✓ or more subsystems or blocks.
- ✓ *Multilevel Voltage Scaling (MVS)* This is an extension of the SVS, where two or few fixed discrete voltages are applied to different blocks or subsystems.
- ✓ *Dynamic Voltage and Frequency Scaling (DVFS)* This is an extension of the MVS, where a large number of discrete voltages are applied in response to the changing workload conditions of the subsystems.
- ✓ *Adaptive Voltage Scaling (AVS)* This is an extension of the DVFS, where a close-loop control system continuously monitors the workload and adjusts the supply voltage.

❖ Device Features Size Scaling

○ **S-Scaling Factor**

- ✓ *Constant-Field Scaling*

Constant-field scaling of the device dimensions, voltages, and doping densities

<i>Quantity</i>	<i>Before Scaling</i>	<i>After Scaling</i>
Channel length	L	$L' = L/S$
Channel width	W	$W' = W/S$
Gate oxide thickness	t_{ox}	$t_{ox}' = t_{ox}/S$
Junction depth	x_j	$x_j' = x_j/S$
Power supply voltage	V_{dd}	$V_{dd}' = V_{dd}/S$
Threshold voltage	V_{T0}	$V_{T0}' = V_{T0}/S$
Doping Densities	N_A N_D	$N_A' = N_A/S$ $N_D' = N_D/S$

Effects of constant-field scaling on the key device parameters

<i>Quality</i>	<i>Before Scaling</i>	<i>After Scaling</i>
Gate Capacitance	C_g	$C_g' = C_g/S$
Drain Current	I_D	$I_D' = I_D/S$
Power Dissipation	P	$P' = P/S^2$
Power Density	$P/Area$	$P'/Area'$
Delay	t_d	$t_d' = t_d/S$
Energy	$E = P \cdot t_d$	$E' = (1/S^3) \cdot E$

✓ *Constant-Voltage Scaling*

Constant-voltage scaling of the device dimensions, voltages, and doping densities

<i>Quantity</i>	<i>Before Scaling</i>	<i>After Scaling</i>
Channel length	L	$L' = L/S$
Channel width	W	$W' = W/S$
Gate oxide thickness	t_{ox}	$t_{ox}' = t_{ox}/S$
Junction depth	x_j	$x_j' = x_j/S$
Power supply voltage	V_{dd}	$V_{dd}' = V_{dd}$
Threshold voltage	V_{to}	$V_{to}' = V_{to}$
Doping Densities	N_A N_D	$N_A' = N_A \cdot S_2$ $N_D' = N_D \cdot S_2$

Effects of constant-voltage scaling on the key device parameters

<i>Quality</i>	<i>Before Scaling</i>	<i>After Scaling</i>
Gate Capacitance	C_g	$C_g' = C_g/S$
Drain Current	I_D	$I_D' = I_D/S$
Power Dissipation	P	$P' = P \cdot S$
Power Density	$P/Area$	$P'/Area' = S^3 P/Area$
Delay	t_d	$t_d' = t_d/S^2$

✓ *Short-Channel Effects*

Short-channel effects arise when channel length is of the same order of magnitude as depletion region thickness of the source and drain junctions or when the length is approximately equal to the source and drain junction depths.

❖ **Architecture-level Approaches**

- Architectural-level refers to register-transfer-level (RTL), where a circuit is represented in terms of building blocks such as adders, multipliers, read-only memories (ROMs), register files, etc..
- High-level synthesis technique transforms a behavioral-level specification to an RTL-level realization.
- It is envisaged that low-power synthesis technique on the architectural level can have a greater impact than that of gate-level approaches.
- Possible architectural approaches are: parallelism, pipelining, and power management.

✓ *Parallelism for Low Power*

Impact of parallelism on area, power, and throughput

<i>Parameter</i>	<i>Without Vdd Scaling</i>	<i>With Vdd Scaling</i>
Area	2.2X	2.2X
Power	2.2X	0.227X
Throughput	2X	1X

$$P_{par} \approx 0.277P_{ref}$$

✓ *Multi-Core for Low Power*

Power in multi-core architecture

<i>Number of Cores</i>	<i>Clock in MHz</i>	<i>Core Supply Voltage</i>	<i>Total Power</i>
1	200	5	15.0
2	100	3.6	8.94

4	50	2.7	5.20
8	25	2.1	4.5

✓ *Pipelining for Low Power*

Impact of pipelining on area, power, and throughput

<i>Parameter</i>	<i>Without Vdd Scaling</i>	<i>With Vdd Scaling</i>
<i>Area</i>	<i>1.15X</i>	<i>1.15X</i>
<i>Power</i>	<i>2.0X</i>	<i>0.28X</i>
<i>Throughput</i>	<i>2X</i>	<i>1X</i>

$$P_{pipe} \approx 0.28P_{ref}$$

✓ *Combining Parallelism with Pipelining*

$$P_{parpipe} \approx 0.1125P_{ref}$$

Impact of parallelism and pipelining on area, power, and throughput

<i>Parameter</i>	<i>Without Vdd Scaling</i>	<i>With Vdd Scaling</i>
<i>Area</i>	<i>2.5X</i>	<i>2.5X</i>
<i>Power</i>	<i>5.0X</i>	<i>0.1125X</i>
<i>Throughput</i>	<i>4X</i>	<i>1X</i>

❖ **Voltage Scaling Using High-Level Transformations**

- ✓ For automated synthesis of digital systems, high-level transformations such as dead code elimination, common sub-expression elimination, constant folding, in-line expansion and loop unrolling are typically used to optimize the design parameters such as the area and throughput.
- ✓ These high-level transformations can also be used to reduce the power consumption either by reducing the supply voltage or the switched capacitance.

❖ **Multilevel Voltage Scaling**

- ✓ A number of studies have shown that the use of multiple supply voltages results in the reduction of dynamic power from less than 10 % to about 50 %, with an average of about 40 %.
- ✓ It is possible to use more than two, say three or four, supply voltages.
- ✓ However, the benefit of using multiple Vdd saturates quickly.
- ✓ Extending the approach to more than two supply voltages yields only a small incremental benefit.
- ✓ The major gain is obtained by moving from a single Vdd to a dual Vdd.
- ✓ It has been found that in a dual-Vdd/single-Vt system, the optimal lower Vdd is about 60–70 % of the original Vdd.
- ✓ The optimal supply voltage depends on the threshold voltage Vt of the MOS transistors as well.

❖ **Challenges**

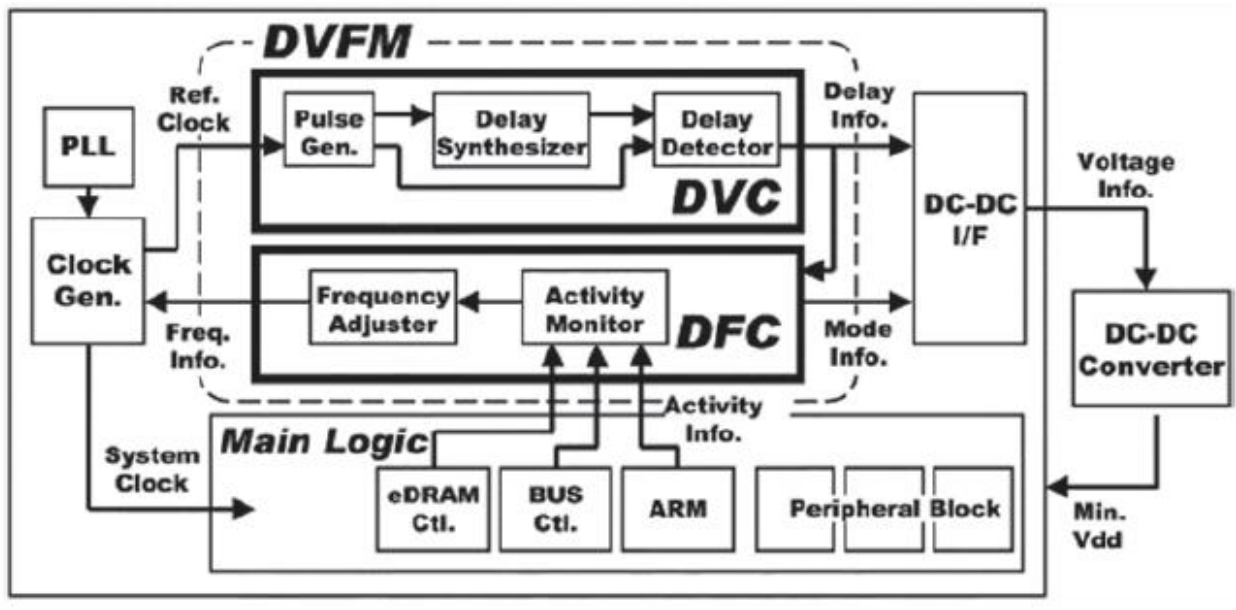
- ✓ Voltage Scaling Interfaces
- ✓ Converter Placement
- ✓ Floor Planning, Routing, and Placement
- ✓ Static Timing Analysis
- ✓ Power-Up and Power-Down Sequencing
- ✓ Clock Distribution
- ✓ Low-Voltage Swing

❖ Dynamic Voltage and Frequency Scaling

- ✓ DVFS has emerged as a very effective technique to reduce CPU energy.
- ✓ The technique is based on the observation that for most of the real-life applications, the workload of a processor varies significantly with time and the workload is bursty in nature for most of the applications.
- ✓ The energy drawn for the power supply, which is the integration of power over time, can be significantly reduced.
- ✓ This is particularly important for battery-powered portable systems.

❖ Adaptive Voltage Scaling

A better alternative that can overcome this limitation is the adaptive voltage scaling (AVS) where a close-loop feedback system is implemented between the voltage scaling power supply and delay-sensing performance monitor at execution time.



UNIT IV

Minimizing Switched Capacitance

❖ System-level Approaches

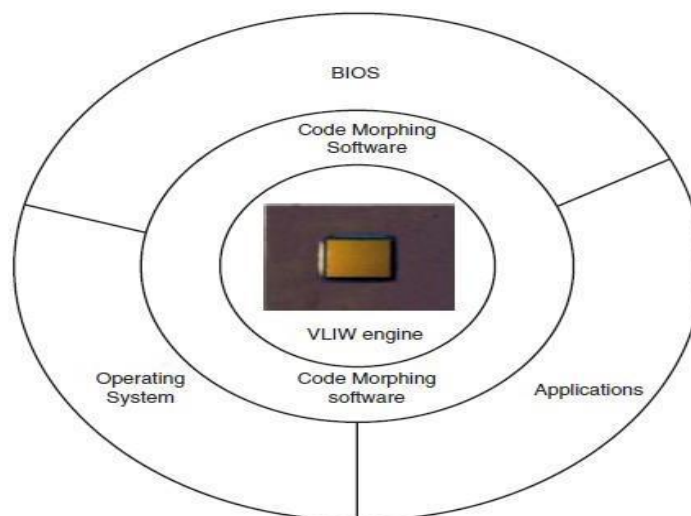
- ✓ It is well known that the same functionality can be either realized by hardware or by software or by a combination of both.
- ✓ The hardware-based approach has the following characteristic:
 - Faster
 - Costlier
 - Consumes more power
- ✓ On the other hand the software-based approaches the following characteristics:
 - Cheaper
 - Slower
 - Consumes lesser power

❖ Transmeta's Crusoe Processor

- ✓ 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.

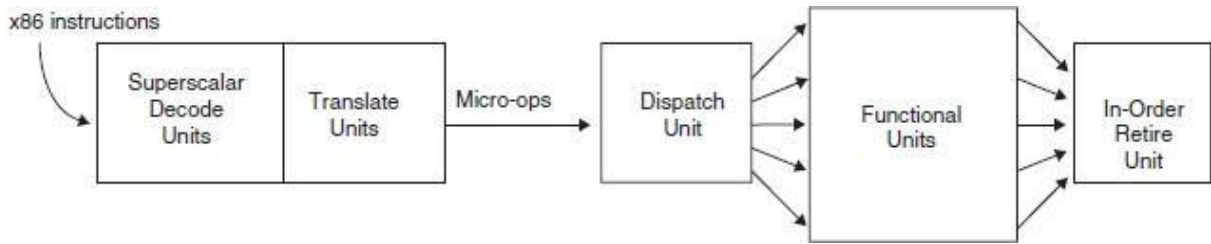
❖ Role of Code Morphing Software

- ✓ The Code Morphing software mediates between x86 software and the VLIW engine.
- ✓ It is fundamentally a dynamic translation system.
- ✓ A program that translates instructions from one instruction set architecture to another instruction set architecture.
- ✓ Here, x86 code is compiled into VLIW code of the Crusoe processor.
- ✓ Code Morphing software insulates x86 programs from the hardware engine's native instruction set.
- ✓ The code morphing software is the only program that is written directly for the VLIW processor.



❖ Superscalar Architecture and VLIW Architecture

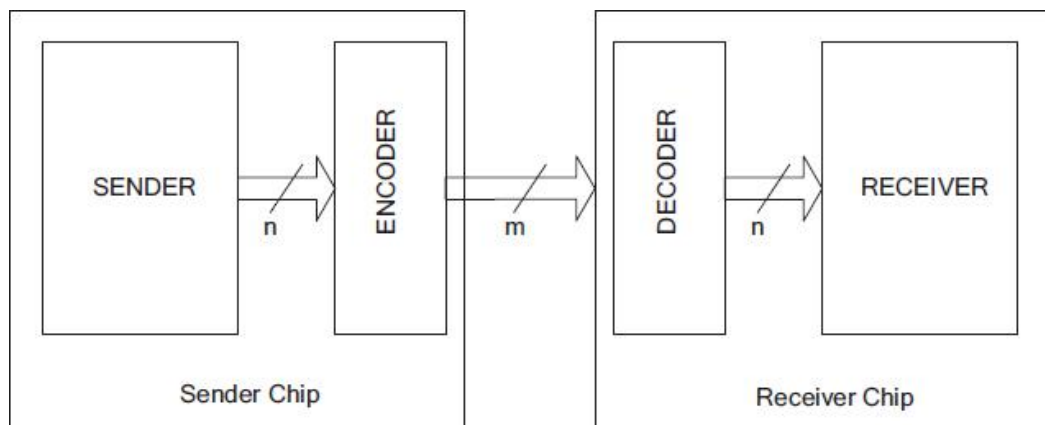
- ✓ Both superscalar and VLIW architectures implement a form of parallelism called instruction-level parallelism within a single processor.
- ✓ VLIW incurs lower power dissipation
- ✓ Superscalar out-of-order architecture



❖ Basic Concept of Bus Encoding

- ✓ Switching activity can be reduced by coding the address bit before sending over the bus.
- ✓ This is done introducing sample to sample correlation such that total number of bit transitions is reduced.
- ✓ Similarly, communicating data bits in an appropriately coded form can reduce the switching activity.
 - *Gray Coding*
 - *One-Hot Coding*
 - *Bus-Inversion Coding*
 - *T0 Coding*

❖ *Encoder and decoder blocks to reduce switching activity*



- ✓ 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 suitably encode the data before sending over the I/O interface.
- ✓ A decoder is used to get back the original data at the receiving end
- ❖ *Switching activity of a modulo-7 counter using binary and Gray codes for state encoding*

The reduction in the number of bit transitions for the two types of coding is given below:

Binary code	Transitions	Gray code	Transitions
000		000	
001	1	001	1
010	2	011	1
011	1	010	1
100	3	110	1
101	1	111	1
110	2	101	1
	2		2
Total	12		8

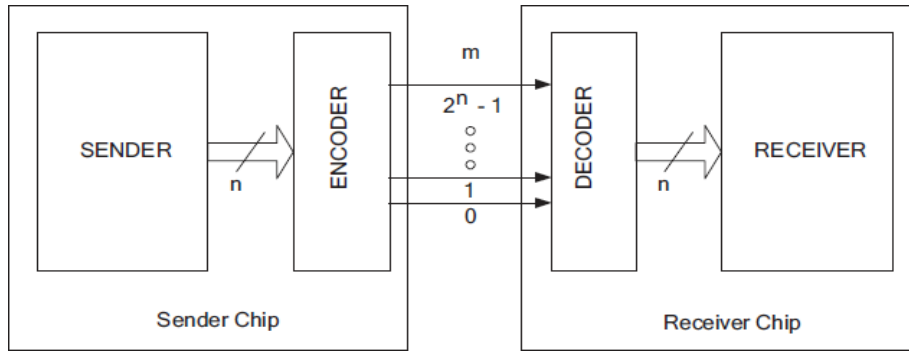
❖ *Gray Coding*

- ✓ A gray code sequence is a set of numbers in which adjacent numbers have only one bit difference.
- ✓ On the other hand, the number of transitions vary from 1 to n ($n/2$ on the average) as shown in the adjacent table.

Decimal Value	Binary Code	Gray Code
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111
6	0110	0101
7	0111	0100
8	1000	1100
9	1001	1101
10	1010	1111
11	1011	1110
12	1100	1010
13	1101	1011
14	1110	1001
15	1111	1000

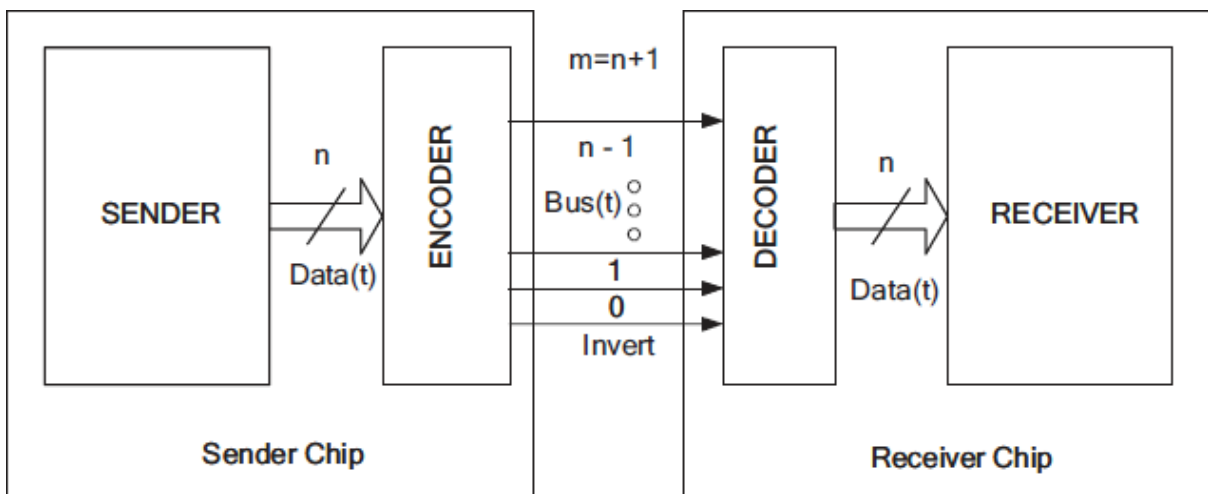
- ✓ The power dissipations of the bus driver decreases because of the reduction of switching activity

❖ One-Hot Coding



- ✓ Both encoder and decoder are memory-less
- ✓ The most important advantage of this approach is that the number of transitions for transmission of any pair of data words one after the other is two: one 0-to-1 and one 1-to-0.
- ✓ The reduction in dynamic power consumption can be computed
- ✓ Although one-hot encoding provides a large reduction in switching activity, the number of signal lines increases exponentially ($2n$) with n . For example, for $n = 8$, the number of signal lines is $m = 256$, and reduction in switching activity is 75 %.

❖ Bus-Inversion Coding



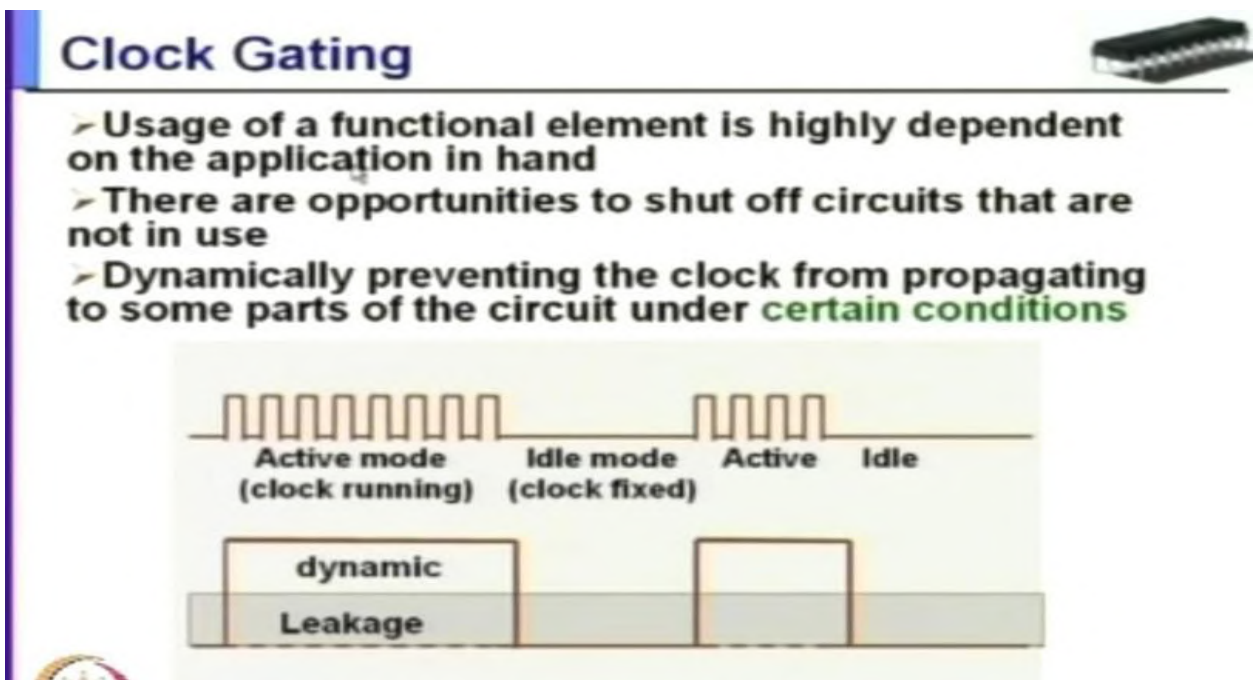
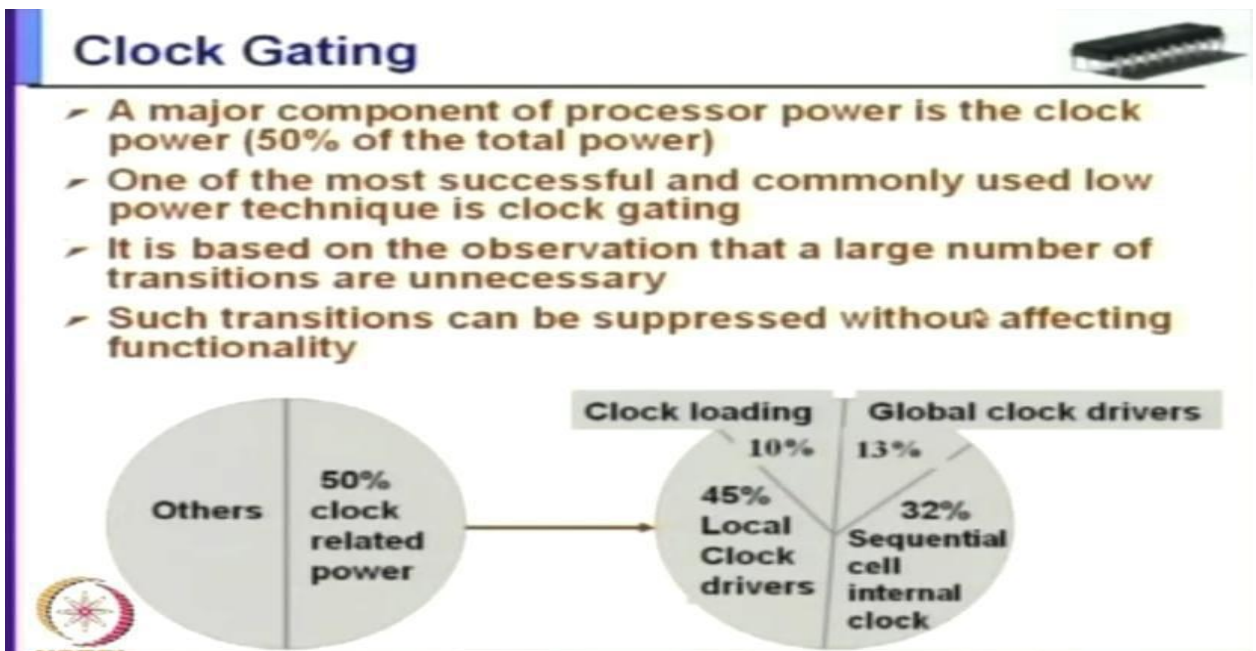
- ✓ Another redundant coding scheme is bus-inversion coding, which requires only one redundant bit i.e., $m = n + 1$ for the transmission of data words.
- ✓ It may be noted that this approach is not applicable to address busses.

❖ T0 Coding

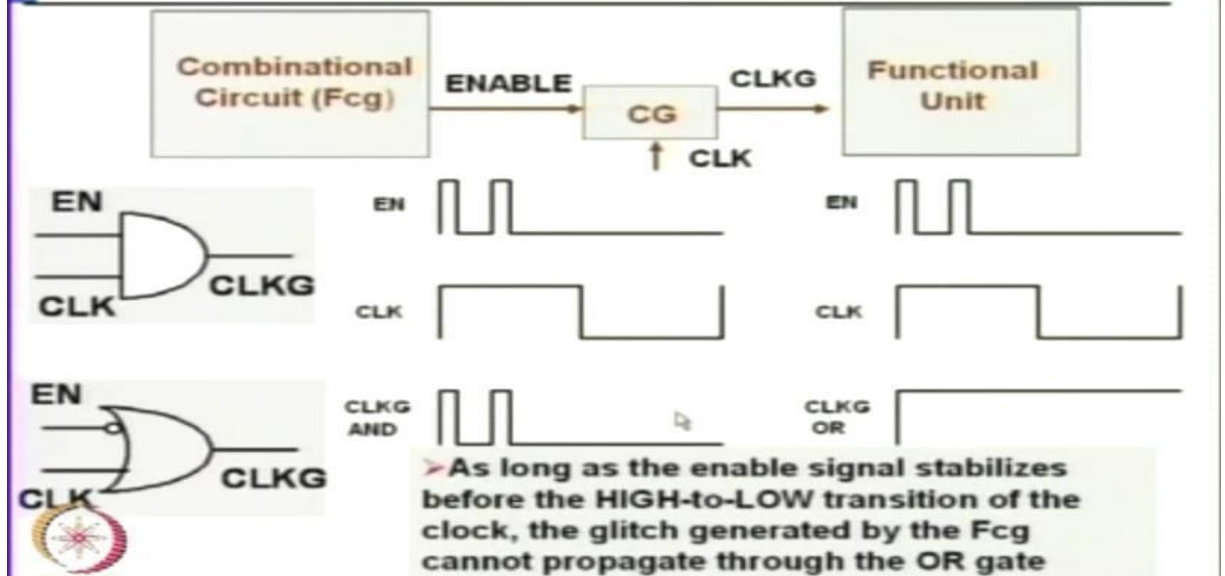
- ✓ In T0 encoding, after sending the first address, the same address is sent for infinite streams of consecutive addresses.
- ✓ The receiver side is informed about it by sending an additional bit known as increment (INC) bit.
- ✓ However, if the address is not consecutive, then the actual address is sent.
- ✓ The T0 code provides, zero transition property for infinite streams of consecutive addresses.

❖ Basic Concept of Clock Gating

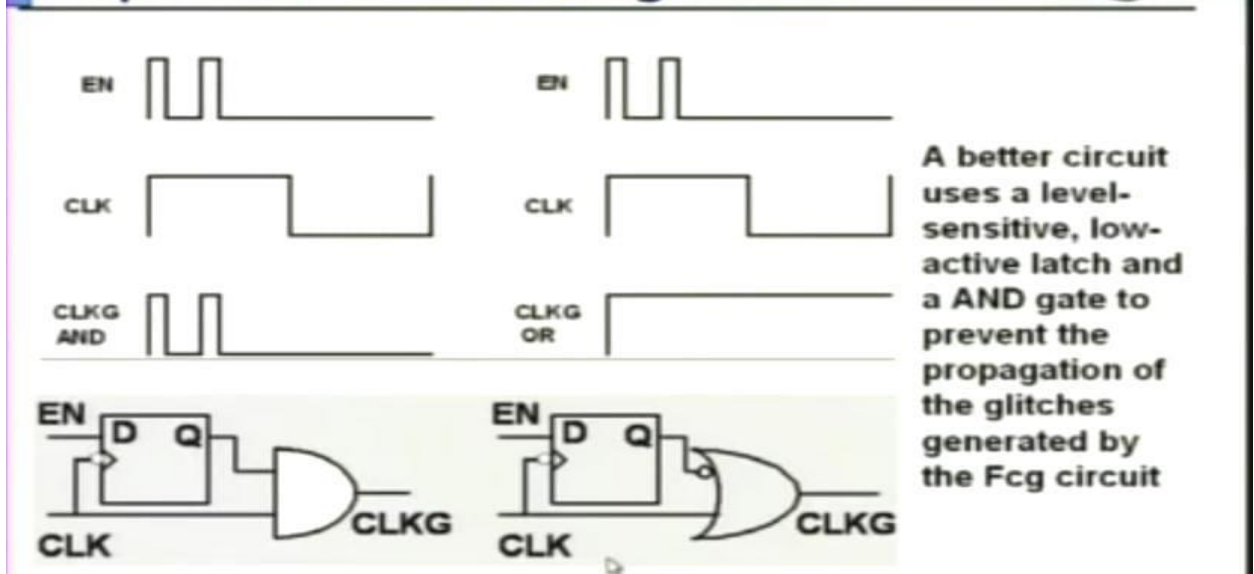
- ✓ It has been observed that a major component of processor power is the clock power (50% of the total power).
- ✓ So, there is scope for large reduction of power dissipation by using suitable technique to remove a large number of unnecessary transitions.
- ✓ Such transitions can be suppressed without affecting functionality.
- ✓ One of the most successful and commonly used low power technique is clock gating.



Clock Gating Principle



Improved Clock Gating Circuit



❖ Three levels of clock gating granularity

- ✓ Module-level clock gating: Large reduction in power but there is limited opportunity.
- ✓ Register-level clock gating: There is more opportunity compared to module level clock gating, but lesser reduction of power.
- ✓ Cell-level clock gating: Provides many more opportunities and it lends itself to automated insertion and can result in massively clock gated designs 📁

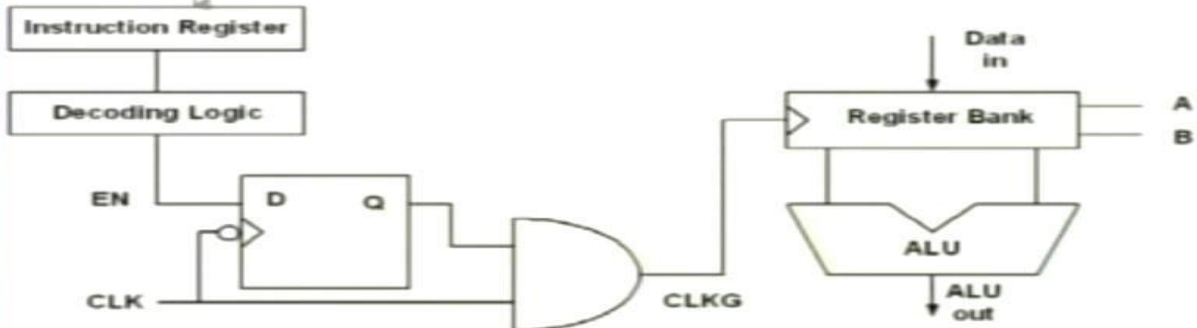
Module-level Clock Gating



Module-level Clock Gating

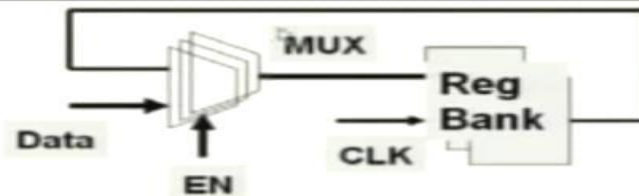


- **Example:** The register bank is clock-gated to prevent unnecessary loading of operands to the ALU when load/store instruction is executed



- Memory bank can be clock-gated in case of ALU operations in a similar manner

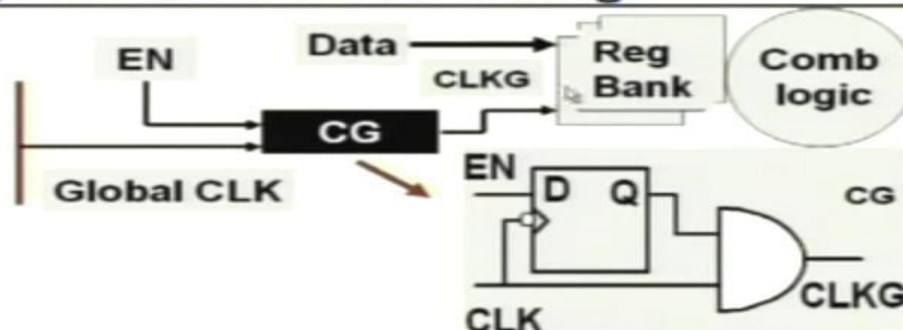
Register-level Clock Gating



- The clock to a single register or a set of registers is gated
- A synchronous load-enabled register bank typically implemented using clocked D flip-flops and a recirculating multiplexer
- The register gets the clock in all the cycles even when no new data is loaded
- Gating condition for registers may be obtained through symbolic analysis of the combinational circuit that feeds it



Register-level Clock Gating



- The register does not get the clock in the cycles when no new data is loaded
- Elimination of the MUX saves power
- The penalty of the clock gating circuit can be amortized over a large number of registers
- The power saving per clock-gate is much less, but there is much more scope than module-level gating



❖ *Clock Gating Granularity Challenges*

Although CG helps to reduce dynamic power dissipation, it introduces several challenges in the application-specific integrated circuit (ASIC) design flow. Some of the important issues are as follows:

- Clock latency
- Effect of clock skew
- Clock tree synthesis
- Physical CG
- Testability concern

Clock Latency



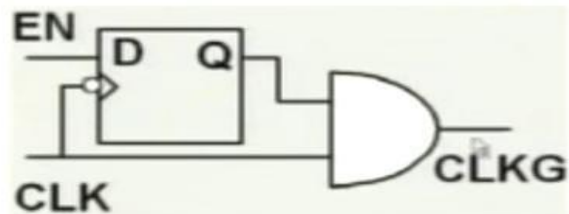
- When a single clock gate drives a large number of registers, it may not have enough drive strength, requiring clock tree at the output
- The enable signal must be ready before the clock arrives at the gating logic
- This must be addressed during synthesis
- Clock latency at the clock gate must be smaller than that at the registers and the difference is the delay of the clock gate
- Either a worst-case estimate of clock-tree synthesis delay or restriction the fan-out of the clock gates



Effect of clock skew



- ❑ Clock skew between the latch and the AND gate may lead to glitches at the gated-clock output



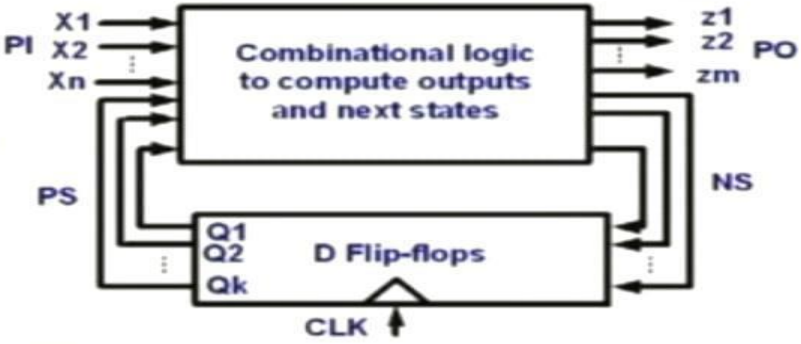
- It is essential to ensure that glitches are not introduced, which may lead to circuit malfunction due to spurious loading of registers
- The clock-skew between the latch and the AND gate should be less than the clock-to-output delay of the latch
- Relative placement of the latch and the AND gate imposes stringent constraints on the clock-tree synthesis tool



❖ Basic Concept of Gated-clock FSMs

- ✓ There are conditions when the next state and output values do not change (idle condition).
- ✓ Clocking the circuit during this idle condition leads to unnecessary wastage of power.
- ✓ The clock can be stopped, if the idle conditions can be detected.
- ✓ This saves power both in the combinational circuit as well as the registers/latches.

General Model of a FSM



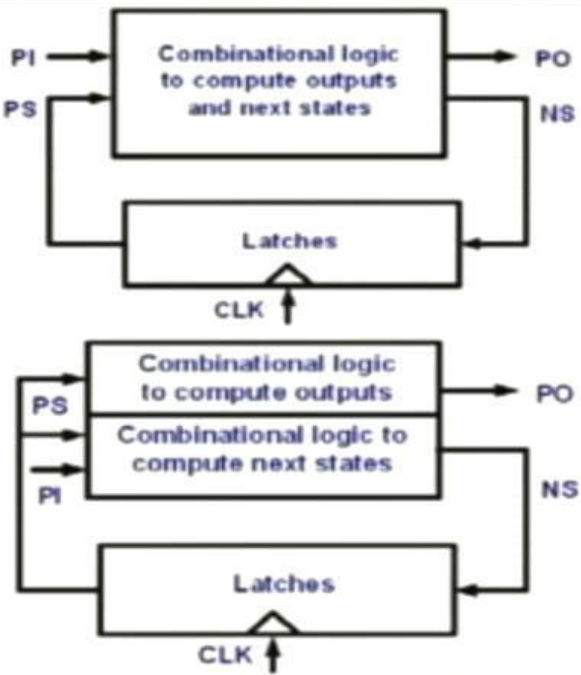
The diagram illustrates the general model of a FSM. It consists of two main blocks: 'Combinational logic to compute outputs and next states' and 'D Flip-flops'. The combinational logic block has multiple inputs labeled 'PI' (Primary Inputs) including x_1, x_2, \dots, x_n and a state input 'PS'. It produces multiple outputs labeled 'PO' (Primary Outputs) including z_1, z_2, \dots, z_m and a next state output 'NS'. The D Flip-flops block has multiple outputs labeled 'Q1, Q2, \dots, Qk' and a common clock input 'CLK'. The 'NS' output from the combinational logic is fed back into the 'PS' input of the combinational logic block. The 'Q' outputs from the flip-flops are fed back into the 'PS' input of the combinational logic block.

❑ FSMs are integral parts of digital systems

➤ FSM model: $(PI, PO, S, \delta, \lambda)$

- PI: Set of Inputs $\{x_1, x_2, \dots, x_n\}$
- PO: Set of Outputs $\{z_1, z_2, \dots, z_m\}$
- S: Set of states $\{s_1, s_2, \dots, s_p\}$
- δ : State Transition function $S^+ = \delta(S, X)$
- λ : Output transition function $Z = \lambda(S, X)$

Mealy and Moore Machines



The diagram shows two types of FSMs: Mealy and Moore machines. Both use 'Latches' for state storage, which are clocked by 'CLK'. In the Mealy machine, the combinational logic block takes 'PI' and 'PS' as inputs and produces 'PO' and 'NS' as outputs. The 'NS' output is fed back into the 'PS' input of the combinational logic block. In the Moore machine, the combinational logic is split into two blocks: one for computing outputs and one for computing next states. The 'PS' input is fed into both blocks. The 'PI' input is fed into the next state logic block. The output logic block produces 'PO' and the next state logic block produces 'NS'. The 'NS' output is fed back into the 'PS' input of both combinational logic blocks.

❑ Mealy machine:

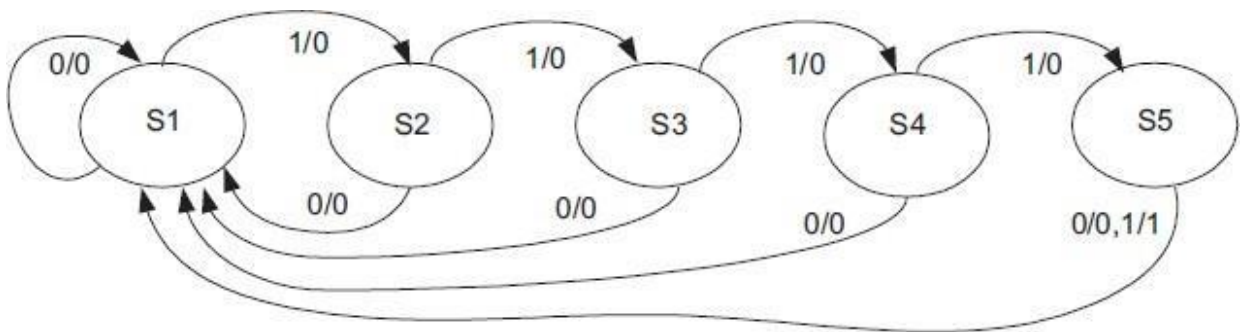
- Outputs are dependent on inputs and current state
- Output transition function $Z = \lambda(S, X)$

❑ Moore machine:

- Outputs are dependent only on current state
- Output transition function $Z = \lambda(S)$
- Inputs (PI) and outputs (NS) are applied in each cycle

❖ FSM State Encoding

- ✓ State encoding can be used to reduce power dissipation in an FSM
- ✓ In the state assignment phase of an FSM, each state is given a unique code.
- ✓ It has been observed that states assignment strongly influences the complexity of its combinational logic part used to realize the FSM.
- ✓ Traditionally state assignment has been used to optimize the area and delay of the circuit.
- ✓ It can also be used to reduce switching activity for the reduction of the dynamic power.
- ✓ State-transition diagram of the “11111” sequence detector



- ✓ State assignments using Gray code and binary code for sequence detector

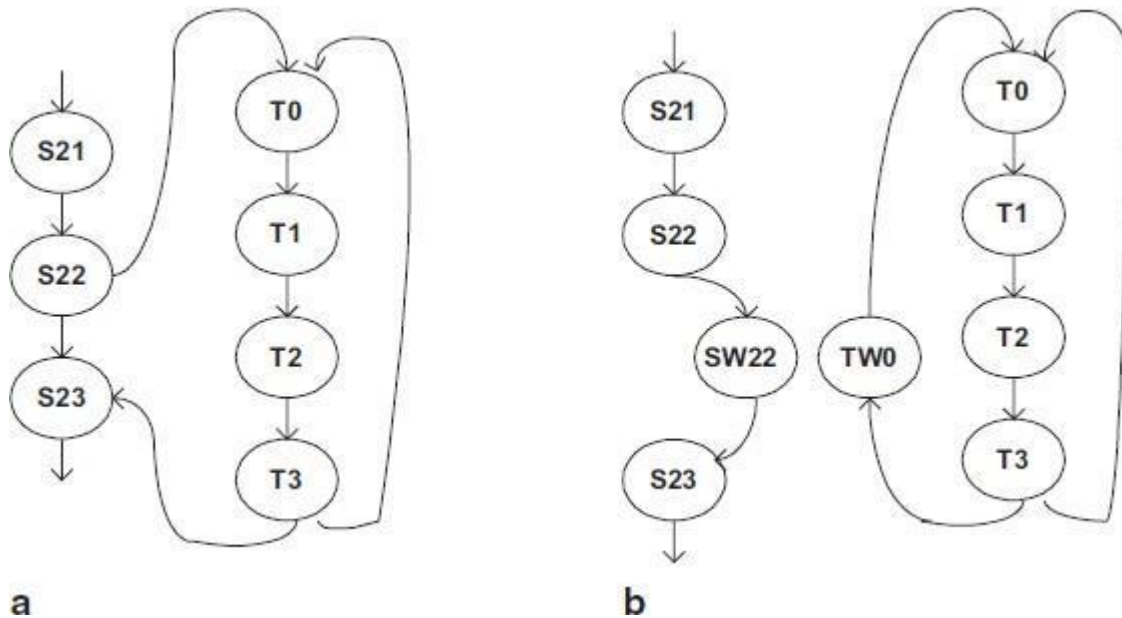
State	Encoding
S1	000
S2	111
S3	001
S4	110
S5	101

State	Encoding
S1	000
S2	001
S3	011
S4	010
S5	100

Transitions	Assignment-1	Assignment-2
S1→S1	0.0	0.0
S1→S2	1.5	0.5
S2→S1	1.5	0.5
S2→S3	1.0	0.5
S3→S1	0.5	1.0
S3→S4	1.5	0.5
S4→S1	1.0	0.5
S4→S5	1.0	1.0
S5→S1	2.0	1.0
Total	10.0	5.5

❖ **Basic Concept of FSM Partitioning**

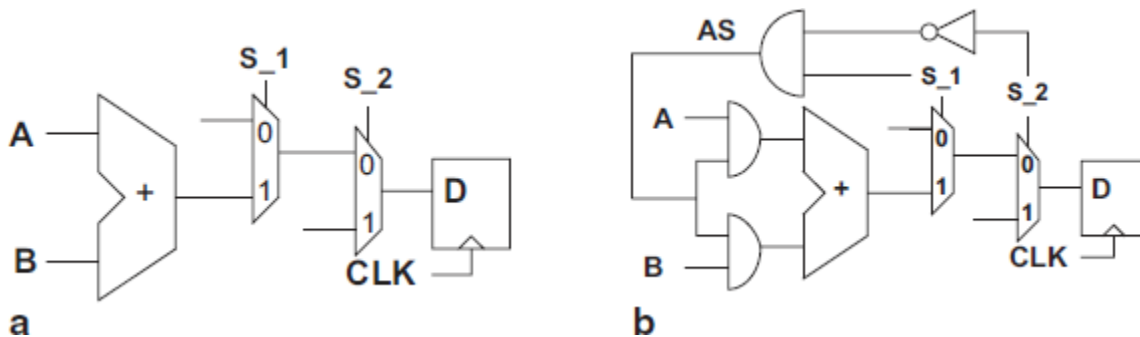
- ✓ The idea is to decompose a large FSM into a several smaller FSMs with smaller number of state registers and combinational blocks.
- ✓ Out of all the FSMs, only the active FSMs receive clock and switching inputs, and the others are idle and consume no dynamic power.
- ✓ This is the basic concept of reducing dynamic power by partitioning an FSM.



(a) An example finite-state machine FSM and (b) decomposed FSM into two FSMs

❖ **Operand Isolation**

- ✓ Operand isolation is a technique for power reduction in the combinational part of the circuit. Here the basic concept is to ‘shutoff’ logic blocks when they do not perform any useful computation.
- ✓ Shutting-off is done by not allowing the inputs to toggle in clock cycles when the output of the block is not used.
- ✓ In the following example, the output of the adder is loaded into the latch only when S_1 is 1 and S_2 is 0.
- ✓ So, input lines of the adder may be gated based on this condition, as shown in the diagram.



(a) An example circuit and (b) operand isolation.

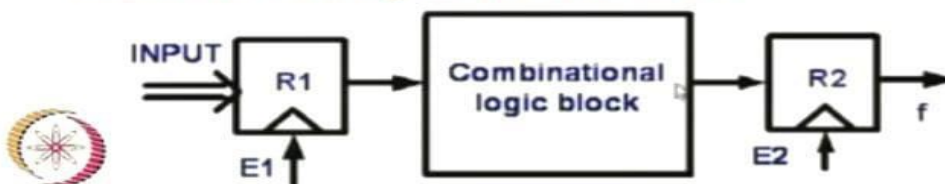
❖ **Precomputation**

- ✓ Precomputation is a technique in which selective computation of output values is done in advance using a much simpler circuit than the original circuit.
- ✓ Precomputed values are used to reduce the switching activity in the subsequent cycles.

Pre-Computation



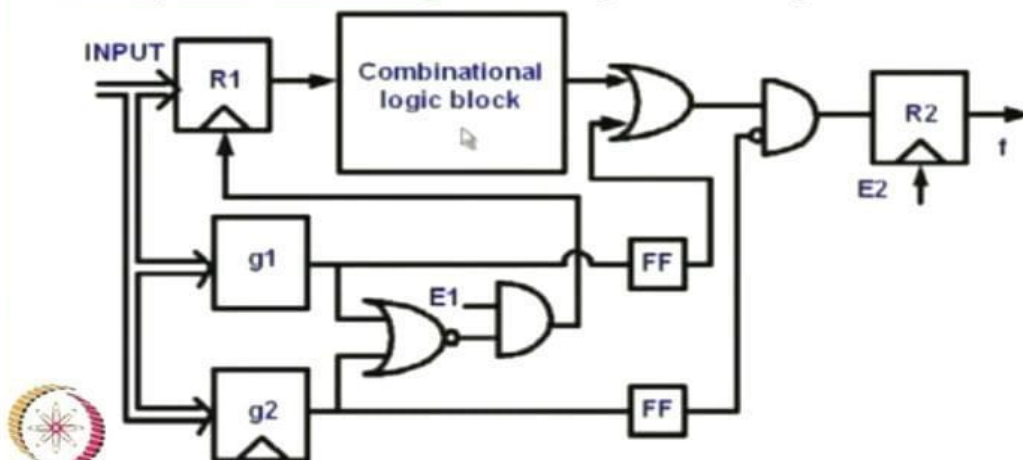
- ❑ Pre-computation is a technique in which selective computation of output values is done one or few cycles in advance using a much simpler circuit than the original circuit
- ❑ Pre-computed values are used to reduce switching activity in the subsequent cycles
- ❑ Consider a combinational logic block sandwiched between two registers R1 and R2



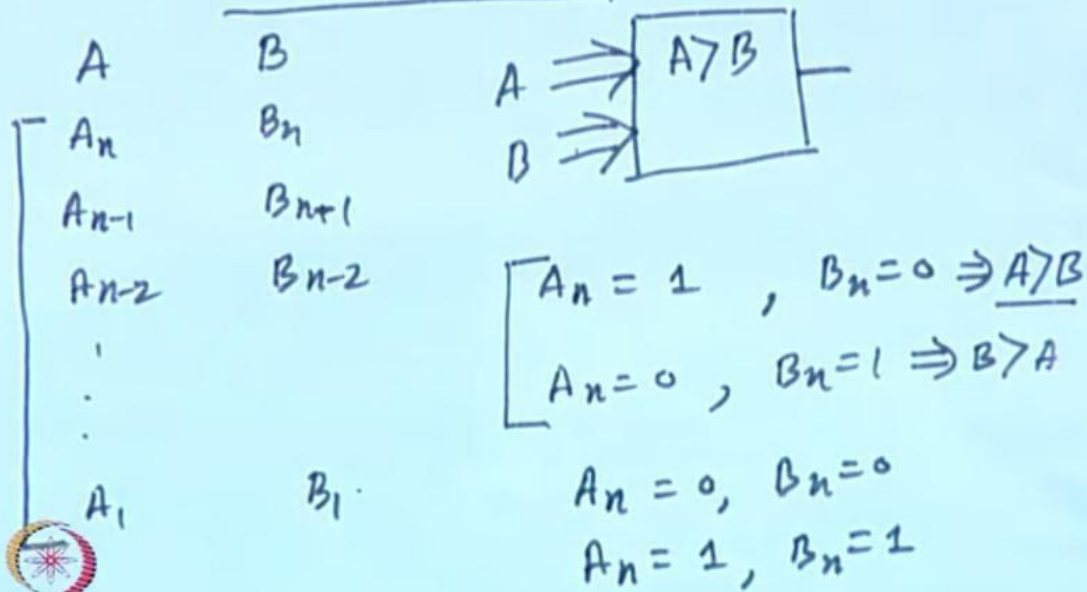
Pre-Computation



- $g1 = 1 \Rightarrow f = 1$ and $g2 = 1 \Rightarrow f = 0$
- During clock cycle n , if either $g1$ or $g2$ evaluates to 1, the register R1 is disabled from loading
- Complete disabling of all inputs take place



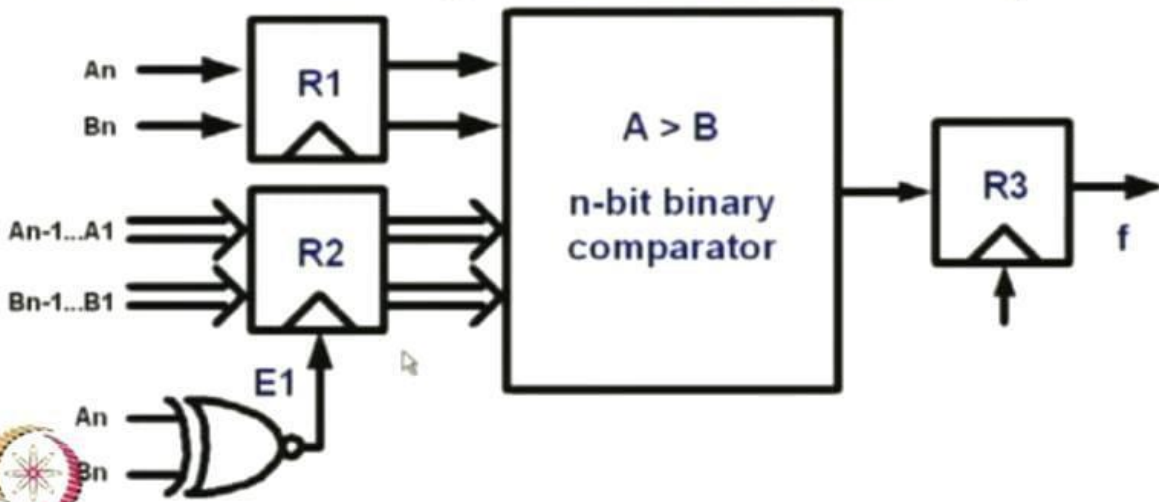
Comparatar



Example of Pre-Computation



- A simple and realistic example is a comparator
- Here the pre-computation function is $A_n \oplus B_n$, which can be realized using an XOR gate
- In this case disabling of a subset of inputs take place



LOGIC STYLES FOR LOW POWER

- There are two basic approaches to realize a digital circuit by metal–oxide–semiconductor (MOS) technology: gate logic and switch logic
- A gate logic is based on the implementation of digital circuits using inverters and other conventional gates such as NAND, NOR, etc.
- Moreover, depending on how circuits function, they can also be categorized into two types—**static and dynamic gates**.
- **Static CMOS Logic:**
- The static CMOS or full complementary circuits require two separate transistor networks: pull-up pMOS network and pull-down nMOS network, as shown in Fig.a, for the realization of digital circuits. Realization of the function $f = A + B_C$ is shown in Fig.b. Both the nMOS and pMOS networks are, in general, a series–parallel combination of MOS transistors. In realizing complex functions using full complementary CMOS logic, it is necessary to limit the number (the limit is in the range of four to six transistors) of MOS transistors in both the pull-up network (PUN) and pull-down network (PDN), so that the delay remains within acceptable limit.

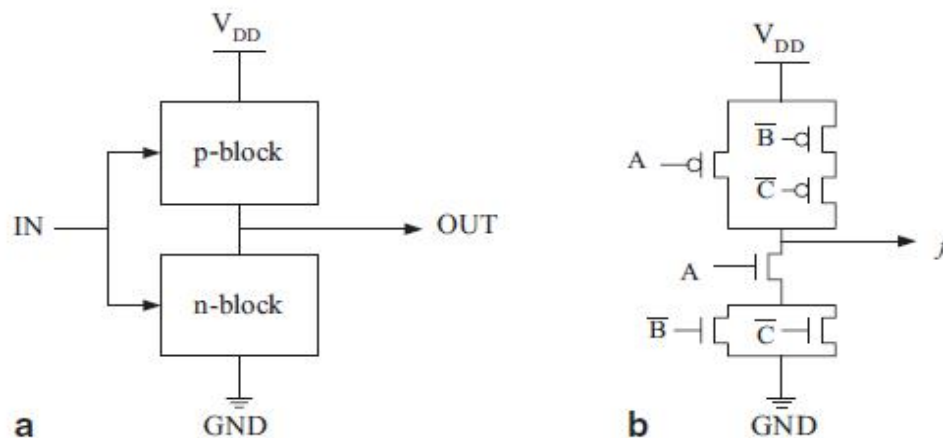


Fig. a Static complementary metal–oxide–semiconductor (CMOS) gate and b realization of $f = \bar{A} + B \cdot C$ with static CMOS gate

Advantages of Static CMOS Logic:

- Ease of fabrication
- Availability of matured logic synthesis tools and techniques
- Good noise margin
- Good robustness property against voltage scaling and transistor sizing
- Lesser switching activity
- No need for swing restoration

- Good I/O decoupling
- Easy to implement power-down circuit
- No charge sharing problem

Disadvantages of Static CMOS Logic

- Larger number of transistors (larger chip area and delay)
- Spurious transitions due to finite propagation delays from one logic block to the next, leading to extra power dissipation and incorrect operation
- Short-circuit power dissipation
- Weak output driving capability
- Large number of standard cells requiring substantial engineering effort for technology mapping

Dynamic CMOS Logic

- Dynamic CMOS circuits are realized based on pre-charge logic. There are two basic configurations of a dynamic CMOS circuit as shown in Fig.a and b. In the first case, an n-block is used as the PDN as shown in Fig. a. In this case, the output is charged to "1" in precharge phase, and in the valuation phase, the output either discharges to "0" through the PDN, if there is discharge path depending on the input combination. Otherwise, the output maintains the "1" state. In the second case, a pblock is used as PUN as shown in Fig.b. In the pre-charge phase, output is discharged to "0" level, and in the evaluation phase, it is charged to "1" level through the PUN, if there is charging path depending on the input combination. Otherwise, the output remains at "0" level.

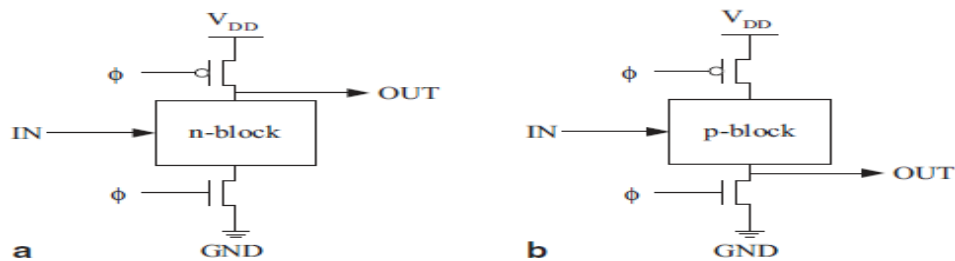


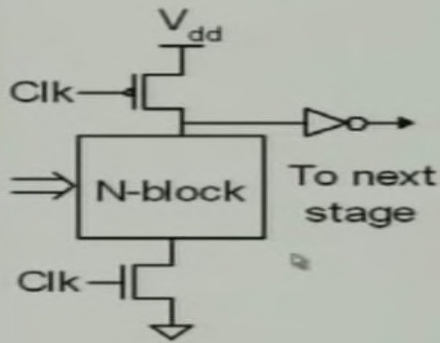
Fig. Dynamic complementary metal-oxide-semiconductor (CMOS) gate with a n-block and b p-block

NORA Logic

The diagram shows three blocks connected in series. The first block is an n-block with clock input clk and clock output clk' . The second block is a p-block with clock input clk' and clock output clk . The third block is an n-block with clock input clk and clock output clk' . Each block has a data input and a data output.

> NORA stands for NO RACE
 > By alternatively using p-blocks and n-blocks, the clock skew problem is overcome in NORA logic circuits
 Here both the n and p-blocks are in pre-charge phase when = 1 or ($clk = 0$)

Domino CMOS Circuits



- It consists of two distinct components:
- The first component is a conventional dynamic pseudo-nMOS gate
- The second component is a static inverting CMOS buffer
- **Advantages:**
- Lower power consumption
- Reduced chip area
- Higher speed of operation (only rising delay)
- No short-circuit power dissipation
- No glitching power dissipation

Advantages of Dynamic CMOS Logic

- Combines the advantage of low power of static CMOS and lower chip area of pseudo-nMOS
- The number of transistors is substantially lower compared to static CMOS, i.e., $N + 2$ versus $2N$
- Faster than static CMOS
- No short-circuit power dissipation occurs in dynamic CMOS, except when static pull-up devices are used to reduce charge sharing.
- No spurious transitions and glitching power dissipation, since any node can undergo at the most one power-consuming transition per clock cycle

Disadvantages of Dynamic CMOS Logic:

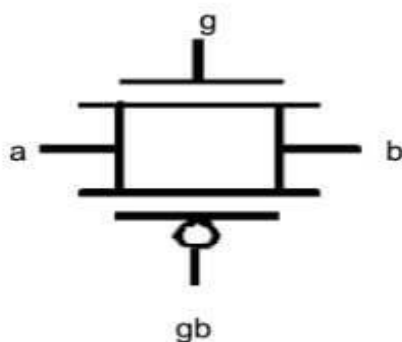
- Higher switching activity
- Not as robust as static CMOS
- Clock skew problem in cascaded realization
- Suffers from charge sharing problem
- Suffers from charge leakage problem requiring precharging at regular interval
- Difficult to implement power-down circuits
- Matured synthesis tools not available

INTRODUCTION OF PTL

- ▶ In electronics **pass transistor logic** (PTL) describes several logic families used in the design of integrated circuits.
- ▶ It reduces the count of transistors used to make different logic gates, by eliminating redundant transistors.
- ▶ Disadvantage that output levels are always lower than the input level.

INTRODUCTION CONT..

▶ PASS TRANSISTOR LOGIC



$g=0, gb=1$
Switch is open

$g=1, gb=0$
Switch is closed

So when $g=1$

If input is '0' then output will be strong '0'.

If input is '1' then output will be strong '1'

Advantages of PTL

- Lower area due to smaller number of transistors and smaller input loads.

- As the PTL is ratioless, minimum dimension transistor can be used. This makes pass-transistor circuit realization very area efficient.
- • No short-circuit current and leakage current, leading to lower power dissipation.

Disadvantages of PTL

- When a signal is steered through several stages of pass transistors, the delay can be considerable.
- There is a voltage drop as we steer signal through nMOS transistors. To overcome this problem, it is necessary to use swing restoration logic at the gate output.
- Pass-transistor structure requires complementary control signals. Dual-rail logic is usually necessary to provide all signals in complementary form.
- Double intercell wiring increases wiring complexity, and capacitance by a considerable amount.
- There is possibility of sneak path.
- Efficient techniques for the logic synthesis using dynamic CMOS and PTL have been developed. An overview of the existing approaches is discussed.

UNIT V

Minimizing Leakage Power

5.1 INTRODUCTION:

- Due to aggressive device-size scaling, the very-large-scale integration (VLSI) technology has moved from the **millimetre to nanometre** era by providing increasingly higher performance along the way.
- Performance improvement has been continuously achieved primarily because of the gradual **decrease of gate capacitances**.
- However, as the supply voltage must continue to scale with device-size scaling to maintain a constant field, the threshold voltage of the metal–oxide–semiconductor (V_{cc}/V_t) and hence performance.
- Unfortunately, the reduction of V_t leads to an exponential increase in the subthreshold leakage current.
- As a consequence, the leakage power dissipation has gradually become a significant portion of the total power dissipation.
- For example, for a 90-nm technology, the leakage power is 42 % of the total power and for a 65-nm technology, the leakage power is 52 % of the total power.
- This has led to vigorous research work to develop suitable approaches for leakage power minimization.

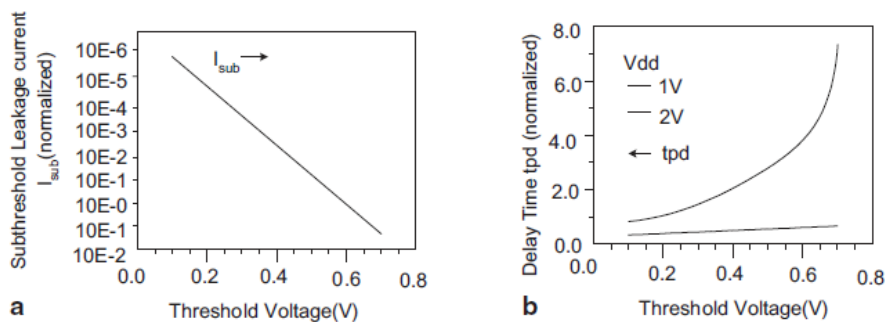


Fig. 5.1 Gate delay time (a) and sub-threshold leakage current (b) dependence on threshold voltage

- As the supply voltage is scaled down, the delay of the circuit increases. Particularly, there is a dramatic increase in delay as the supply voltage approaches the threshold voltage. This tends to limit the advantageous range of the supply voltage to a minimum of about twice the threshold voltage. The delay can be kept constant if the threshold voltage is scaled at the same ratio as the supply voltage; i.e. the ratio of V_t/V_{dd} is kept constant. Unfortunately, as the threshold voltage is scaled down, the sub-threshold leakage current increases drastically, as shown in Fig. 5.1a. Moreover, the delay increases with an increase in threshold voltage when the supply voltage is kept constant as shown in Fig. 5.1b.
- The threshold voltage is the parameter of importance for the control leakage power. As leakage power has an exponential dependence on the threshold voltage and all the leakage power reduction techniques are based on controlling the threshold voltage either statically or dynamically.
- The leakage power reduction techniques can be categorized into two broad types—**standby and run-time leakage**.

- When a circuit or a part of it is not in use, it is kept in the standby mode by a suitable technique such as **clock gating**. The clock gating helps to reduce the dynamic power dissipation, but leakage power dissipation continues to take place even when the circuit is not in use. There are several approaches such as transistor stacking, variable-threshold-voltage complementary metal–oxide–semiconductor (**VTCMOS**), and multiple-threshold-voltage complementary metal–oxide–semiconductor (**MTCMOS**), which can be used to reduce the leakage power when a circuit is in the standby condition.
- On the other hand, there are several approaches for the reduction of the leakage power when a circuit is in actual operation. These are known as **run-time leakage power reduction techniques**. It may be noted that run-time leakage power reduction techniques also reduce the leakage power even when the circuit is in standby mode. As leakage power is a significant portion of the total power, importance of run-time leakage power reduction is becoming increasingly important.
- Classification on leakage power reduction techniques is also possible based on whether the technique is applied at the time of fabrication of the chip or at run time. The approaches applied at fabrication time can be classified as **static** approaches. On the other hand, the techniques that are applied at run time are known as **dynamic** approaches.
- **Run-time leakage power reduction** based on multi-threshold-voltage CMOS (MTCMOS) has been addressed the power-gating technique to minimize leakage power.
 - ✓ Isolation Strategy
 - ✓ State Retention Strategy
 - ✓ Power-gating Controllers
 - ✓ Power Management Techniques
 - ✓ Dual-Vt Assignment Technique
 - ✓ Delay-constrained Dual-Vt Technique
 - ✓ Energy Constraint
 - ✓ Dynamic Vt Scaling Technique

5.2 FABRICATION OF MULTIPLE THRESHOLD VOLTAGES:

- The present-day process technology allows the fabrication of MOSFETs of multiple threshold voltages on a single chip.
- This has opened up the scope for using dual-Vt CMOS circuits to realize high-performance and low-power CMOS circuits.
- The basic idea is to use high-Vt transistors to reduce leakage current and low-Vt transistors to achieve high performance.
- Various fabrication techniques used for implementing multiple threshold voltages in a single chip as follows

5.2.1. Multiple Channel Doping:

The most commonly used technique for realizing multiple-VT MOSFETs is to use different channel-doping densities based on the following expression:

$$V_{th} = V_{tb} + 2\tau_B + \frac{\sqrt{2\varepsilon_{si}\cdot q\cdot Na(2\tau_B+V_{bs})}}{C_{ox}} \quad (5.1)$$

Where V_{tb} is the flat-band voltage, Na is the doping density in the substrate, and $\tau_B = kT / q(Lx (Na / x))$.

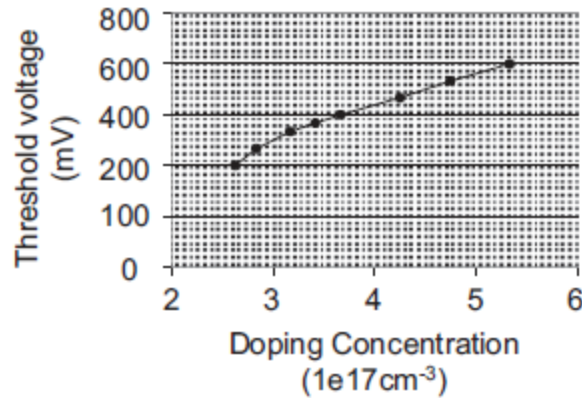


Fig. 5.2 Variation of threshold voltage with doping concentration

- Based on this expression, the variation of threshold voltage with channel-doping density is shown in Fig. 5.2.
- A higher doping density results in a higher threshold voltage.
- However, to fabricate two types of transistors with different threshold voltages, two additional masks are required compared to the conventional single- V_t fabrication process.
- This makes the dual- V_t fabrication costlier than single- V_t fabrication technology.
- Moreover, due to the non-uniform distribution of the doping density, it may be difficult to achieve dual threshold voltage when these are very close to each other.

5.2.2. Multiple Oxide CMOS:

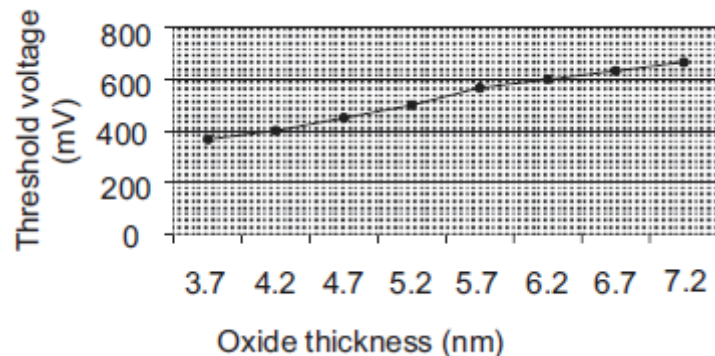


Fig. 5.3 Variation of threshold voltage with gate oxide thickness

The expression for the threshold voltage shows a strong dependence on the value of C_{ox} , the unit gate capacitance. Different gate capacitances can be realized by using different gate oxide thicknesses. The variation of threshold voltage with oxide thickness (t_{ox}) for a 0.25- μm device is shown in Fig. 5.3. Dual- V_{th} MOSFETs can be realized by depositing two different oxide thicknesses. A lower gate capacitance due to higher oxide thickness not only reduces subthreshold leakage current but also provides the following benefits:

a. Reduced gate oxide tunnelling because the oxide tunnelling current exponentially decreases with the increase in oxide thickness.

Reduced dynamic power dissipation due to reduced gate capacitance, because of higher gate oxide thickness. Although the increase in gate oxide thickness has the above benefits, it has some adverse effects due to an increase in short-channel effect. For short-channel devices as the gate oxide thickness increases, the aspect ratio (AR), which is defined by $AR = \text{lateral dimension} / \text{vertical dimension}$, decreases:

$$AR = \frac{L}{\left[t_{ox} \left(\frac{\epsilon_{si}}{\epsilon_{ox}} \right) \right]^{1/3} W_{dm}^{1/3} X_j^{1/3}},$$

Where ϵ_{si} and ϵ_{ox} are silicon and oxide permittivities, L , t_{ox} , W_{dm} , and X_j are channel length, gate oxide thickness, depletion depth, and junction depth, respectively.

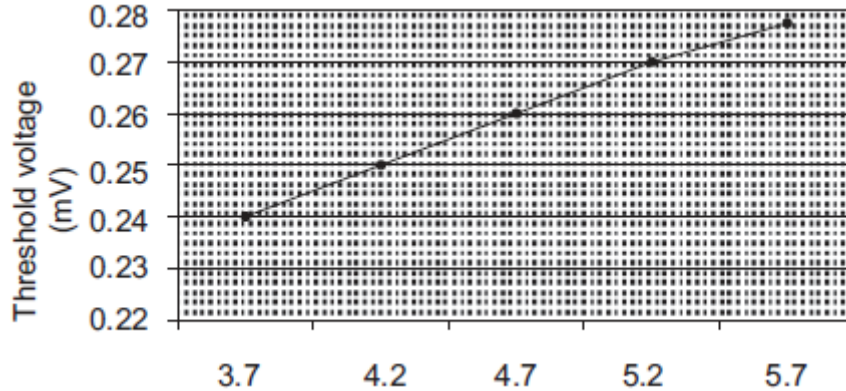


Fig. 5.4 Variation of threshold voltage with oxide thickness for constant Aspect Ratio(AR)

Immunity to the short-channel effect decreases as the AR value reduces. Figure 5.4 shows the channel lengths for different oxide thicknesses to maintain AR. A sophisticated process technology is required for fabricating multiple oxide CMOS circuits.

5.2.3. Multiple Channel Length:

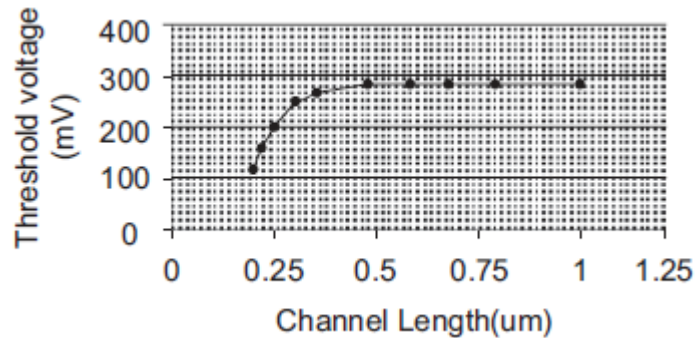


Fig. 5.5 Variation of threshold voltage with channel length

In the case of short-channel devices, the threshold voltage decreases as the channel length is reduced, which is known as V_{th} roll-off. This phenomenon can be exploited to realize transistors of dual threshold voltages. The variation of the threshold voltage with channel length is shown in Fig. 5.5. However, for transistors with feature sizes close to $0.1 \mu\text{m}$, halo techniques have to be used to suppress the short-channel effects. As the V_{th} roll-off becomes very sharp, it turns out to be a very difficult task to control the threshold voltage near the minimum feature size. For such technologies, longer channel lengths for higher V_{th} transistors increase the gate capacitance, which leads to more a dynamic power dissipation and delay.

5.2.4. Multiple Body Bias:

The application of reverse body bias to the well-to-source junction leads to an increase in the threshold voltage due to the widening of the bulk depletion region, which is known as *body effect*. This effect can be utilized to realize MOSFETs having multiple threshold voltages. However, this necessitates separate body biases to be

applied to different nMOS transistors, which means the transistors cannot share the same well. Therefore, costly triple-well technologies are to be used for this purpose.

Another alternative is to use silicon-on-insulator (SoI) technology, where the devices are isolated naturally. In order to get best of both the worlds, i.e. a smaller delay of low-V_t devices and a smaller power consumption of high-V_t devices, a balanced mix of both low-V_t and high-V_t devices may be used. The following two approaches can be used to reduce leakage power dissipation in the standby mode.

5.3 APPROACHES FOR MINIMIZING LEAKAGE POWER

- ✓ VTCMOS Approach
 - VTCMOS circuits make use of the body effect to reduce the subthreshold leakage current, when the circuit is in normal mode
- ✓ MTCMOS Approach
 - In this approach, MOSFETs with two different threshold voltages are used in a single chip.
 - It uses two operational modes—*active* and *sleep* for efficient power management.
- ✓ Dual-V_t Assignment Approach (DTCMOS)

VTCMOS and MTCMOS Approach

- ✓ In case of VTCMOS, basic principle is to adjust threshold voltage by changing substrate bias. Transistors initially have low V_{th} during normal operation and substrate bias is altered using substrate bias control circuit. The threshold is increased by using reverse body bias when the circuit is not in use. Effective in reducing leakage power dissipation in standby mode and it involved additional area and higher circuit complexity. So, it is a post-silicon approach.
- ✓ On the other hand, in case of MTCMOS approach MOS transistors of multiple threshold voltages are fabricated in which a power gating transistor is inserted in the stack between the logic transistors and either power or ground, thus creating a virtual supply rail or a virtual ground rail, respectively. The logic block contains all low-V_{th} transistors for fastest switching speeds while the switch transistors, header and footer, are built using high-V_{th} transistors to minimize the leakage power dissipation. So, it is a pre-silicon approach.

❖ Standby and Runtime Leakage Power

Standby leakage power dissipation takes place when the circuit is not in use, i.e. inputs do not change and clock is not applied.

On the other hand, runtime leakage power dissipation takes place when the circuit is being used

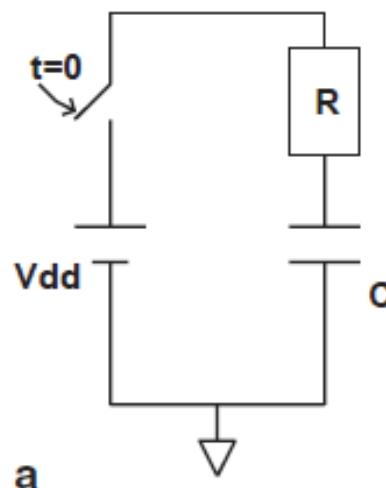
5.4 Adiabatic Logic Circuits

- ✓ Static complementary metal–oxide–semiconductor (CMOS) circuits are extremely successful in terms of market share because of many advantages such as lower power dissipation, reliable operation and availability of computer-aided design (CAD) synthesis tools.
- ✓ We have seen that all the circuit nodes make a rail-to-rail (0 and V_{dd}) transition for each switching event and the supply voltage V_{dd} remains constant.
- ✓ As a consequence, the output node makes a transition from 0 to V_{dd} with a load capacitance C_L, an energy of C_L V²_{dd} is drawn from the power supply.

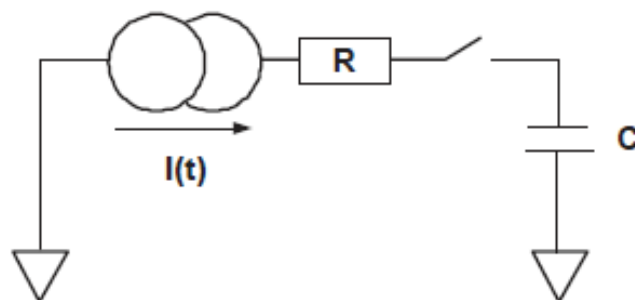
- ✓ Out of this, $1/2 C_L V_{dd}^2$ is stored in the capacitor and the remaining half is dissipated in the p-type metal–oxide–semiconductor (pMOS) network.
- ✓ Subsequently, when the output node switches from V_{dd} to 0, the energy that was stored in the capacitor is dissipated in the n-type metal–oxide–semiconductor (nMOS) network.
- ✓ The power dissipation that takes place because of these switching events is converted to heat, which is ultimately released to the environment. This has far-reaching consequences like global warming. To reduce power dissipation, the circuit designers can reduce the supply voltage, decrease the node capacitance or minimize the number of switching events. In the preceding chapters, we have discussed these approaches.
- ✓ A novel approach to achieve energy dissipation below this lower limit of $C_L V_{dd}^2$. This has resulted in a new class of circuits known as ‘adiabatic circuit’. Adiabatic switching is a circuit-level approach that has made it possible to realize the ultra-low-power computing applications without scaling the supply voltage. The term ‘adiabatic’ refers to the thermodynamic processes that exchange no heat with the environment. The electric charge transfer between various circuit nodes can be considered as the process. In adiabatic CMOS circuits, the energy consumption is minimized by slowing down the charge transport between the drain and source terminals of the metal–oxide–semiconductor field-effect transistor (MOSFET) switch and recovering the energy without dissipating as heat.

Conventional charging of a capacitor C through a resistor R

- ✓ Conventional charging of a capacitor leads to the dissipation of an energy of $1/2 C_L V_{dd}^2$

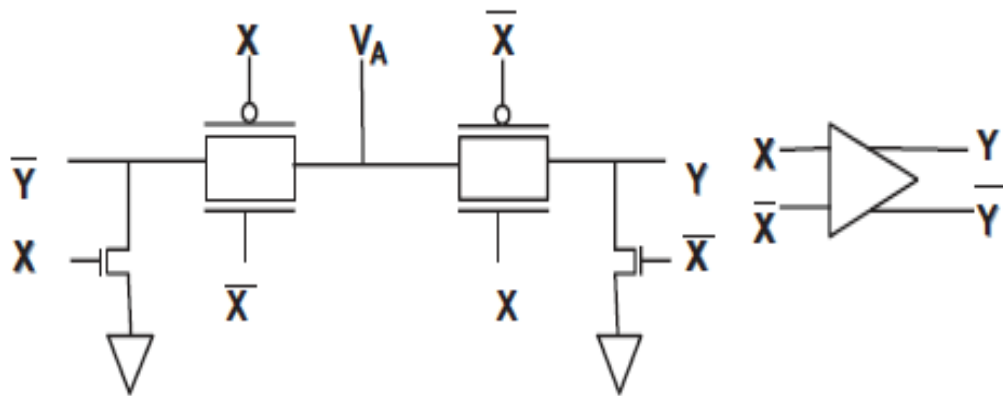


Adiabatic charging of a capacitor



$$E_{diss} = \left(\frac{RC}{T} \right) \cdot C \cdot V_c(T)^2$$

Adiabatic Amplification



Step 1: Input X and its complement are applied to the circuit, which remain stable in the following steps.

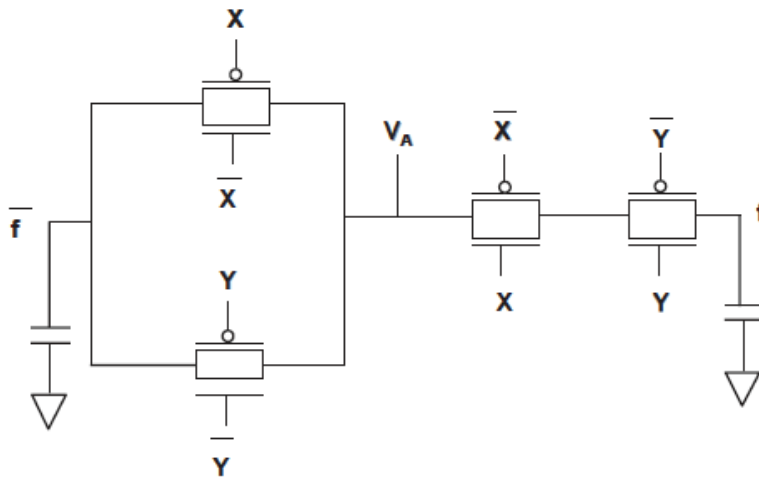
Step 2: The amplifier is activated by applying V_A , which is a slow ramp voltage from 0 V to V_{dd} .

Step 3: One of the two capacitors which is connected through the transmission gate is adiabatically charged to V_A and the other one is clamped to 0 V in transition time T .

Step 4: After the charging is complete, the output signal pair remains stable and can be used as inputs to the next stage of the circuit.

Step 5: The amplifier is de-energized by ramping the voltage from V_A to 0 V . In this step, the energy that was stored in C is transferred back to the power supply.

Adiabatic realization of the AND/ NAND gate



5.5 Battery-Driven System

- ✓ In recent years there large proliferation of portable computing and communication equipment, such as laptops, palmtops, cellphones, etc. and the growth rate of these portable equipment is very high.
- ✓ The complexity of these devices is also increasing with time, leading to larger energy consumption.
- ✓ 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.

- ✓ Moreover, the commercial success of these products depend on weight, cost and battery life.
- ✓ Low power design methodology is very important to make these battery-operated devices commercially viable.

Battery-Driven System Design

- ✓ Battery-driven system design involves the use of one or more of the following techniques:
 - ✓ Voltage and Frequency Scaling
 - ✓ Dynamic Power Management
 - ✓ Battery-Aware Task Scheduling
 - ✓ Battery Scheduling and Management
 - ✓ Static Battery Scheduling
 - ✓ Terminal Voltage-Based Battery
 - ✓ Discharge Current-Based Battery Scheduling
 - ✓ Battery-Efficient Traffic Shaping and Routing

Rate capacity effect for rechargeable batteries

- ✓ Dependency between the actual capacity and the magnitude of the discharge current depends on the availability of active region.
- ✓ When discharge rate is high, surface of the cathode gets coated with insoluble compound.
- ✓ This prevents access to many active areas and consequent reduction of actual capacity of the battery.
- ✓ As a result, a higher rate of discharge leads to a lower available capacity.
- ✓ This is known as Rate Capacity effect.

Recovery effect for rechargeable batteries.

- ✓ Availability of charge carriers D depends of the concentration of positively charged ions near the cathode.
- ✓ When heavy current is drawn, rate at which positively charged ions consumed at the cathode is more than supplied.
 - ✓ This improves as the battery is kept idle for some duration.
 - ✓ As a consequence, the battery voltage recovers in idle periods

❖ *Non-increasing profile effect' of a battery*

- ✓ It has been experimentally verified that if the tasks consuming higher power are scheduled first followed by tasks with decreasing power consumption, then energy available in the battery is larger compared to other schedules. This is known as non-increasing profile effect.

❖ *Basic steps of battery aware task scheduling*

- There are three steps. In the first step, an early deadline first (EDF) based schedule is made, provided the task dependencies are not violated.
- In the second step, the task schedule is modified by scheduling the tasks in the non increasing order of the current loads provided the deadlines and the task dependencies are not violated.

- In the third step, starting from the last task, the slack obtained at the end of the task is utilized to get the optimal pair of supply voltage and the body bias voltage.

❖ Reverse Body Biasing (RBB)

With the advancement of technology, as the process technology further gets lower, the energy due to static power becomes more significant, and the algorithm using RBB to reduce the leakage current provides larger saving in power dissipation.

5.6 CAD Tools for Low Power VLSI Circuits

Limitation

- ✓ In RTL coding there is no provision to use Multi-Vt , Multi-Vdd, Body biasing and power gating in RTL synthesis.
- ✓ So, the static power reduction techniques cannot be used.
- ✓ As supply voltage and the operating frequency are also not handled at the RTL level, the dynamic power can be reduced primarily by reducing the switching activity α . Commonly used techniques in RTL synthesis to reduce α are:
 - Bus encoding
 - Clock gating
 - FSM state assignment