



**VEMU Institute of Technology,
P.Kothakota,Chittoor,AP.**



15A04802-Low Power VLSI Circuits& Systems

**Prepared by
Dr.G.ELAIYARAJA.,M.E.Ph.D
Professor/ECE**

COURSE OUTCOMES

C421.1	Explain low power design methodologies and MOS transistor electrical characteristics
C421.2	Analyze the MOS inverter configurations and MOS combinational circuits
C421.3	Discuss the sources of power dissipation and voltage scaling approaches for low power
C421.4	Explain the minimizing of switched capacitance using various approaches.
C421.5	Analyze various approaches to minimize the leakage power

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

- Understand 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.



Low Power VLSI Circuits and Systems

Unit-1

Unit-1

- Introduction
- Historical Background
- Why Low Power
- Sources of Power Dissipations
- Low Power Methodologies

MOS Transistors

- Introduction
- The Structure of MOS Transistor
- The Fluid Model
- Modes of Operations of MOS Transistor
- Electrical Characteristics of MOS Transistors
- MOS Transistors as a Switch

Introduction

- Design for **low power** has become nowadays one of the major concerns for complex, **very-large-scale-integration (VLSI) circuits**.
- Deep **submicron technology**, from **130 nm** onwards, poses a new set of design problems related to the power consumption of the chip.
- **Tens of millions of gates** are nowadays being implemented on a relatively small die, leading to a **power density** and **total power dissipation** that are at the limits of what **packaging, cooling, and other infrastructure** can support.
- As technology has shrunk to **90 nm and below**, the leakage current has increased dramatically, and in some **65-nm designs**, leakage power is nearly as large as dynamic power.

Introduction

- So it is becoming impossible to increase the **clock speed** of high-performance chips as technology shrinks and the **chip density** increases, because the **peak power consumption** of these chips is already at the **limit** and cannot be increased further.
- Also, the **power density** leads to reliability problems because the mean time **to failure decreases with temperature.**
- Besides, the **timing degrades** and the **leakage currents** increase with **temperature.**

Introduction

- For **battery-powered devices** also, this **high on-chip power density** has become a significant problem, and techniques are being used in these devices from software to architecture to implementation level to alleviate this problem as much as possible like **power gating and multi-threshold** libraries.
- Some other techniques being used nowadays are using different **supply voltages** at different blocks of the design according to the performance requirements, or **voltage scaling techniques**

Introduction

- Moreover, aggressive device size scaling used to achieve high performance leads to increased variability due to **short-channel and other effects**.
- This, in turn, leads to variations in process parameters such as, **L_{eff} , N_{ch} , W , T_{ox} , V_t , etc.**
- Performance parameters such as **power** and **delay** are significantly affected due to the variations in **process parameters and environmental/ operational (V_{dd} , temperature, input values, etc.) conditions**.
- For designs, due to variability, the design methodology in the future **nanometer VLSI circuit** designs will essentially require a paradigm shift from deterministic to probabilistic and statistical design approach.

Historical Background

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

Historical Background

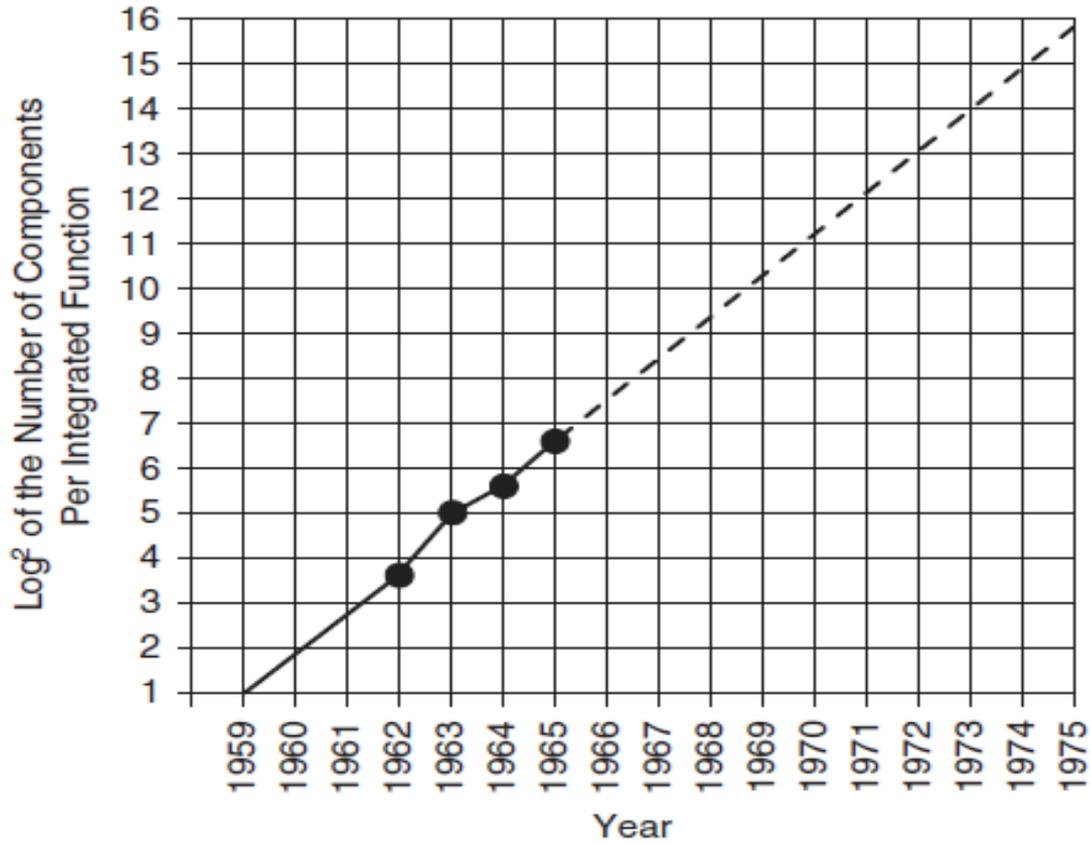
- The tremendous success of the transistor led to vigorous research activity in the **field of microelectronics**.
- Later, **Shockley founded a semiconductor industry**. Some of his colleagues joined him or founded semiconductor industries of their own.
- **Gordon Moore, a member of Shockley's team, founded Fairchild and later Intel.**
- Research engineers of Fairchild developed the first planar transistor in the late 1950s, which was the key to the development of integrated circuits (ICs) in 1959.
- Planar technology allowed realization of a complete electronic circuit having a number of devices and interconnecting them on a single silicon wafer.

Historical Background

- Within few years of the development of ICs, **Gordon Moore, director, Research and Development Laboratories, Fairchild Semiconductor**, wrote an article entitled “**Cramming More Components onto Integrated Circuits**” in the April 19, 1965 issue of the *Electronics Magazine*.
- *He was asked to predict what would happen over the next 10 years in the semiconductor component industry.*
- Based on the very few empirical data, he predicted that by 1975, it would be possible to cram as many as **65,000 components** onto a **single silicon chip** of about one fourth of a square inch.

Historical Background

- Moore's law based on his famous prediction



Historical Background

- Again after 30 years, Moore compared the actual performance of two kinds of **devices—random-access memories (RAM) and microprocessors**.
- Amazingly, it was observed that both kinds traced the slope fairly closely to the revised 1975 projection.
- Moore's law acted as a driving force for the spectacular development of IC technology leading to different types of products

Historical Background

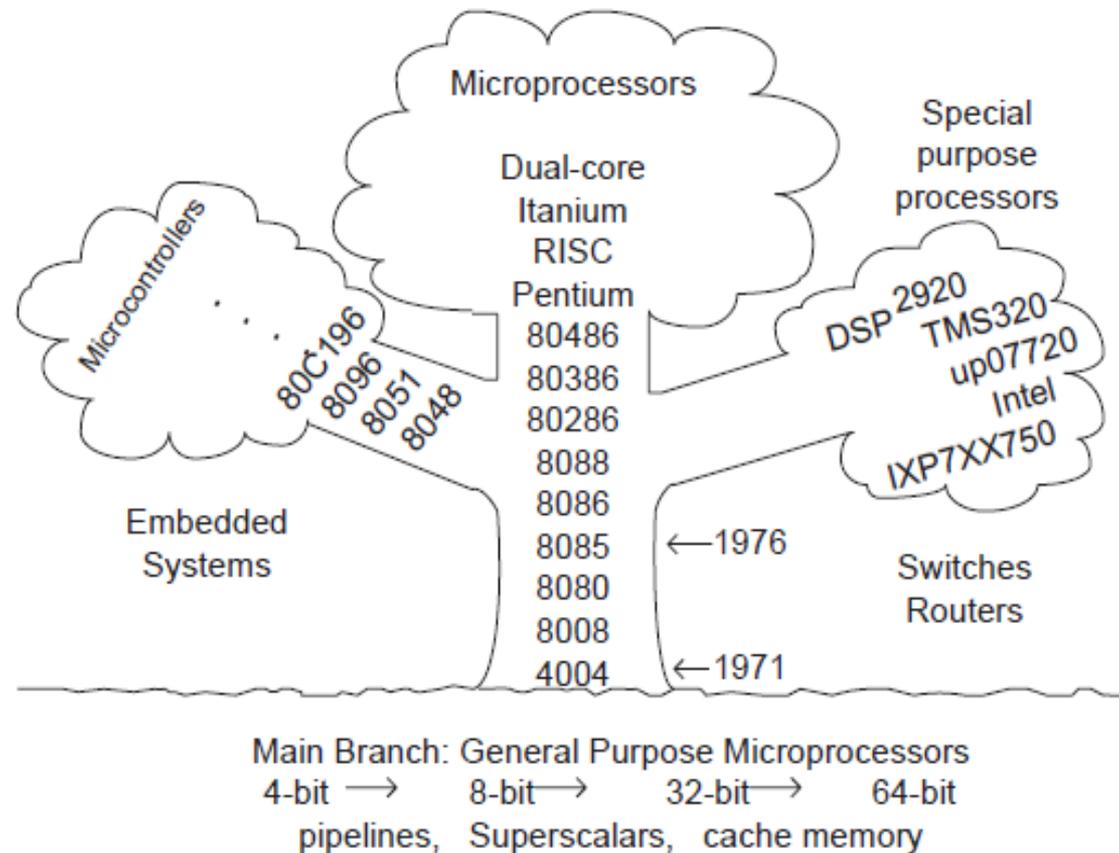
- Evolution of IC Technology

Year	Technology	Number of components	Typical products
1947	Invention of transistor	1	–
1950–1960	Discrete components	1	Junction diodes and transistors
1961–1965	Small-scale integration	10–100	Planner devices, logic gates, flip-flops
1966–1970	Medium-scale integration	100–1000	Counters, MUXs, decoders, adders
1971–1979	Large-scale integration	1000–20,000	8-bit 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

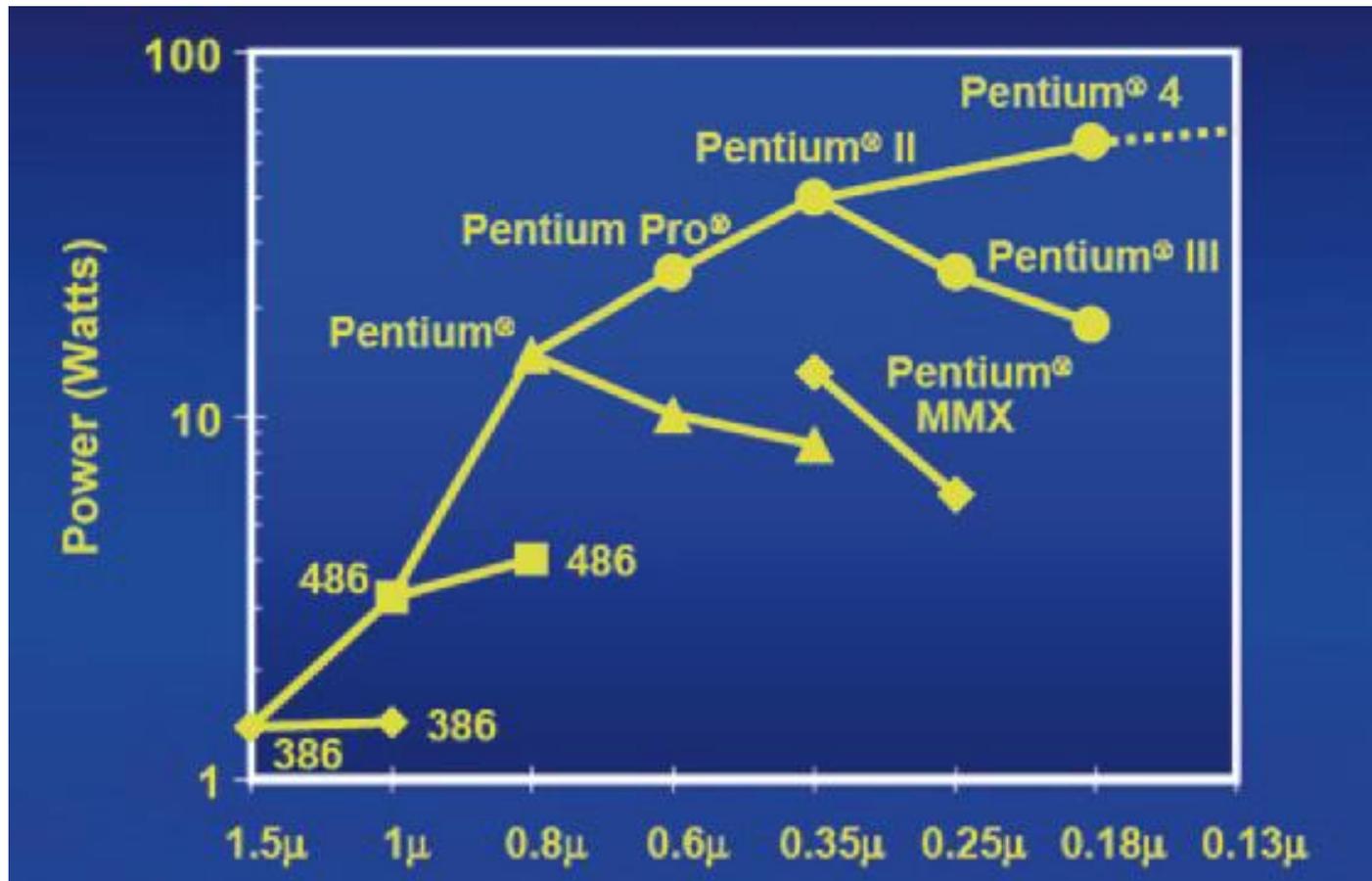
Historical Background

- Evolution tree of microprocessor. RISC reduced instruction set computer, DSP digital signal processor



Historical Background

- Power dissipation of Intel processors. (Source: Intel)



Historical Background

Landmark years of semiconductor industry

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

Why Low Power?

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

Increasing transistor count

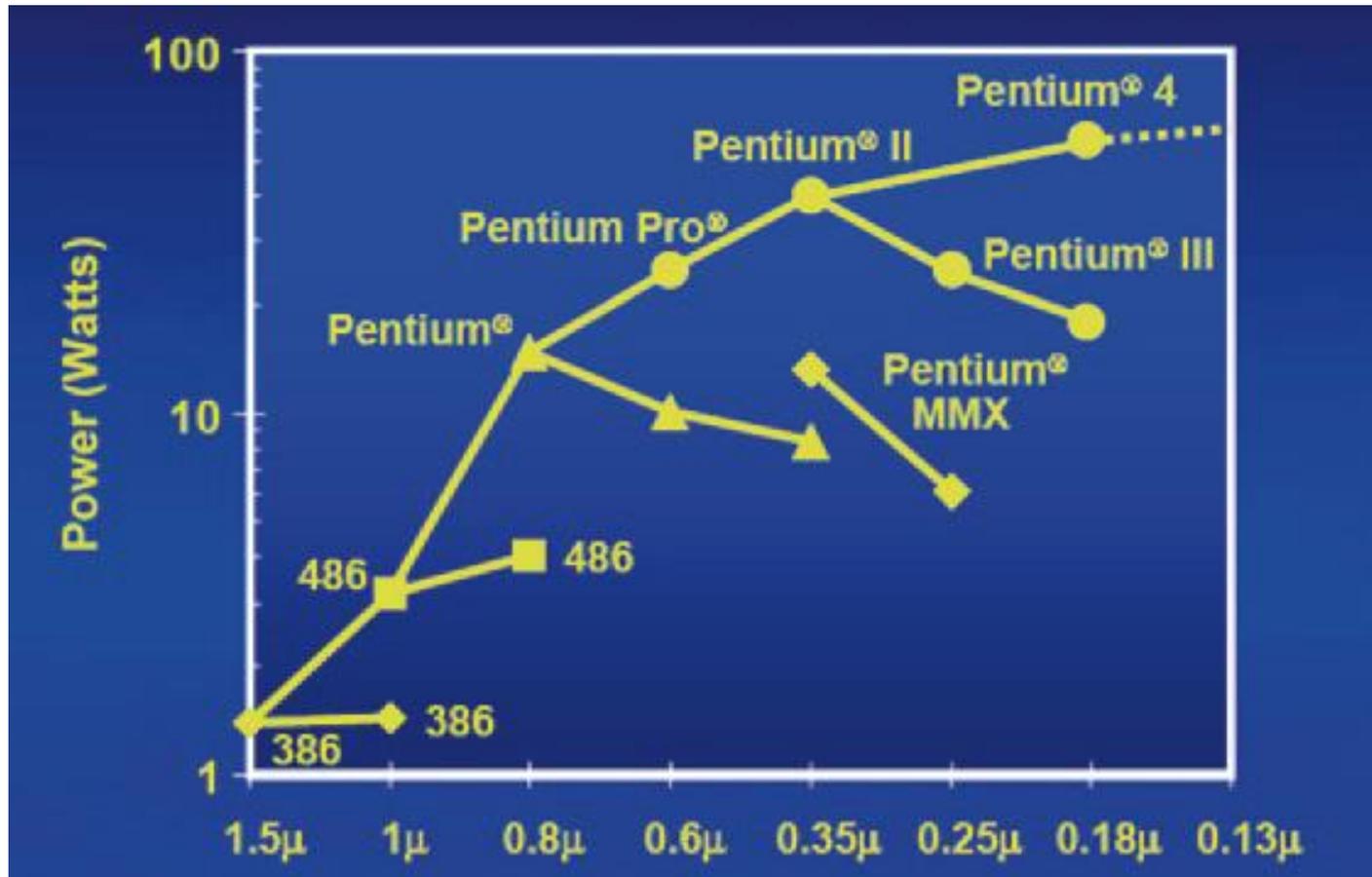
Higher speed of operation

Greater device leakage currents

- Increased process parameter variability due to **aggressive device size scaling** has created problems in **yield, reliability, and testing**.
- **Power consumption** is now considered one of the most important design parameters.

Why Low Power?

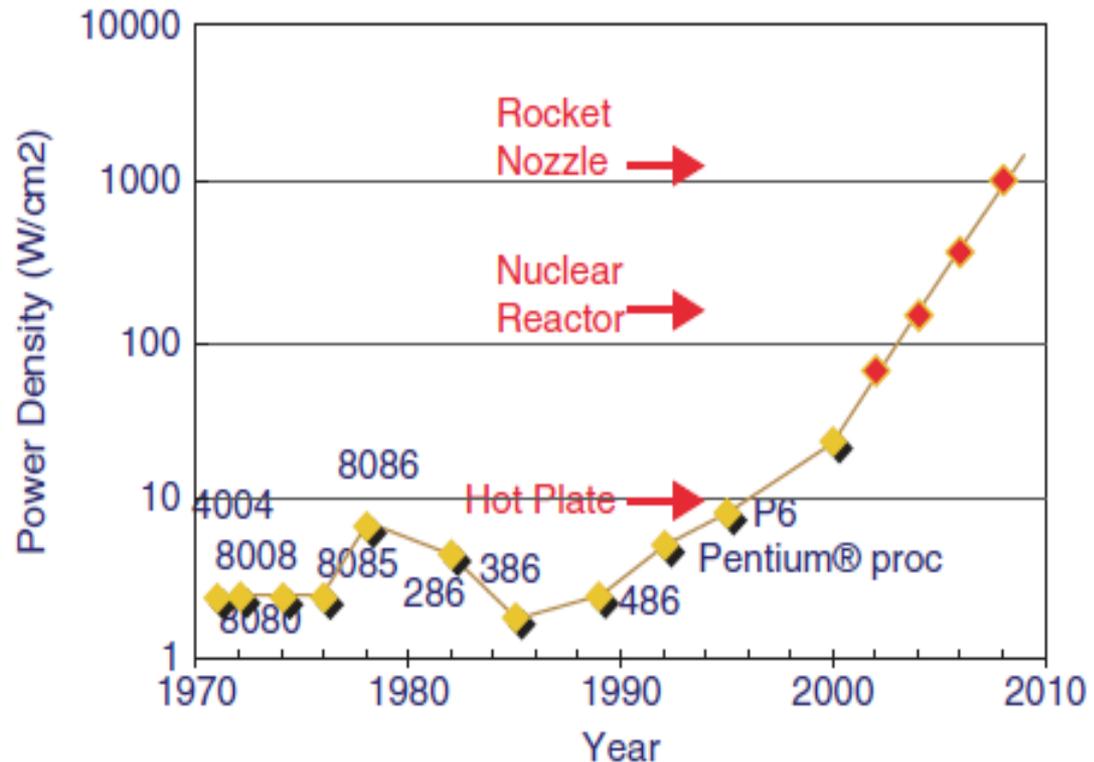
- Power dissipation of Intel processors. (Source: Intel)



Why Low Power?

- The magnitude of power per unit area known as **power density**

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



Why Low Power?

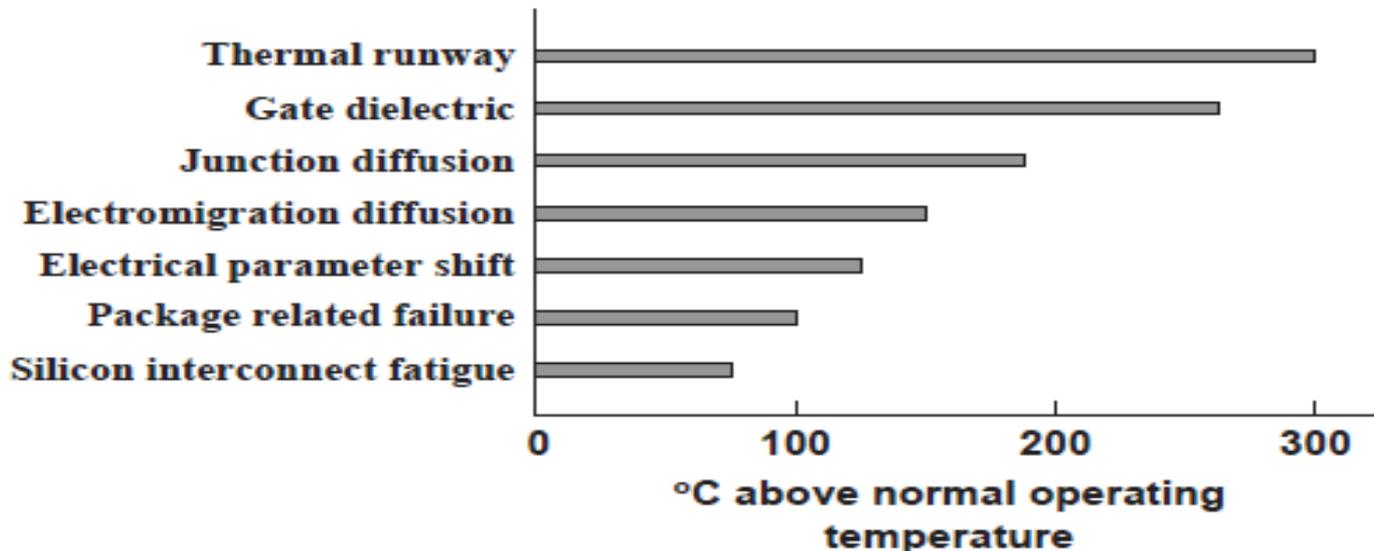
- To remove the heat generated by the device, it is necessary to provide suitable packaging and cooling mechanism.
- There is an escalation in the cost of packaging and cooling as the power dissipation increases.
- To make a chip commercially viable, it is necessary to reduce the cost of packaging and cooling, which in turn demands lower power consumption.

Why Low Power?

- Increased customer demand has resulted in proliferation of hand-held, battery operated devices such as **cell phone, personal digital assistant (PDA), palmtop, laptop, etc.**
- The growth rate of the portable equipment is very high.
- Moreover, users of cell phones strive for increased functionality (as provided by smart phones) along with long battery life.
- As these devices are battery operated, battery life is of primary concern.
- Unfortunately, the battery technology has not kept up with the energy requirement of the portable equipment.
- Commercial success of these products depends on size, weight, cost, computing power, and above all on battery life.
- **Lower power consumption** is essential to make these products commercially viable.

Why Low Power?

- It has been observed that reliability is closely related to the power consumption of a device.
- As **power dissipation** increases, the **failure rate** of the device increases because **temperature-related failures** start occurring with the increase in temperature



Onset temperatures of various failure

Different failure mechanisms against temperature

Why Low Power?

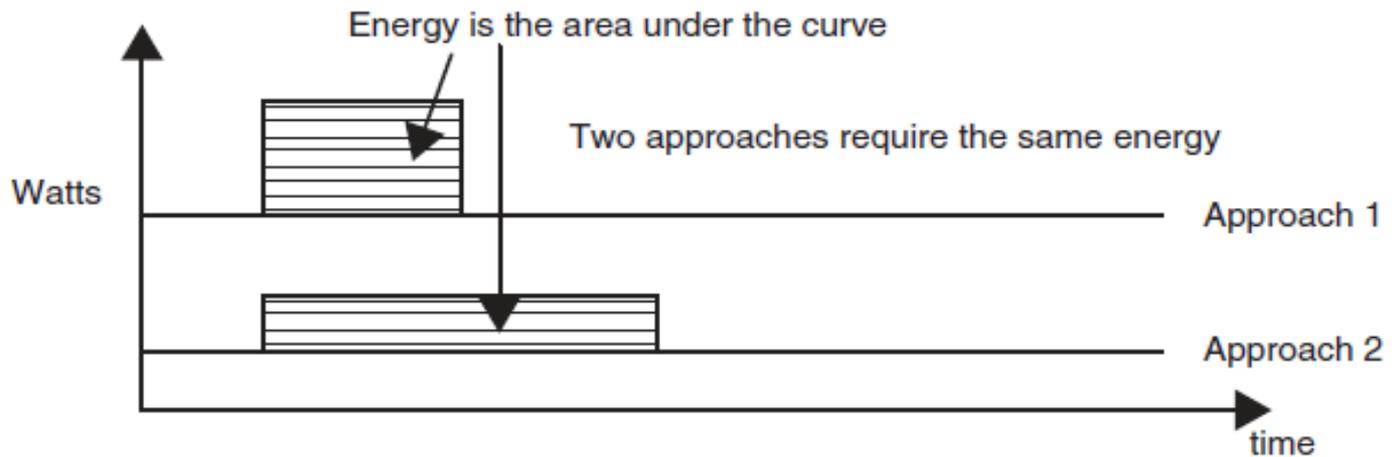
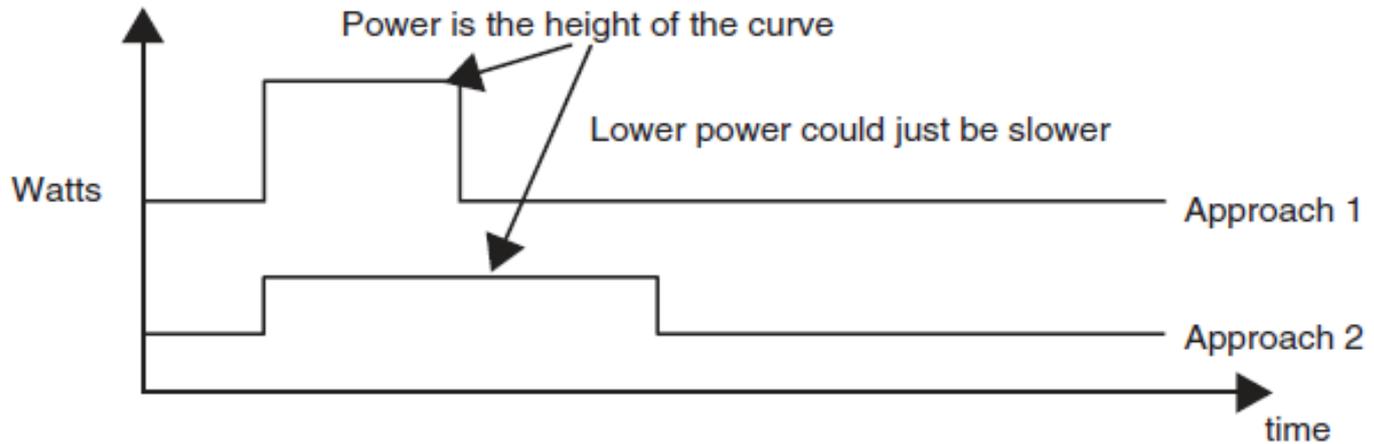
- It has been found that **every 10 °C rise in temperature roughly doubles the failure rate.**
- **Lower power dissipation** of a device is essential for reliable operation.
- According to an estimate of the **US Environmental Protection Agency (EPA)**, 80 % of the power consumption by office equipment 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 desktops and laptops has emerged.

Sources of Power Dissipations

- **Power Vs Energy**
- Although power and energy are used interchangeably in many situations.
- These two have different meanings and it is essential to understand the difference between the two, especially in the case of battery-operated devices.
- Power is the instantaneous power in the device, while energy is the integration of power with time.

Sources of Power Dissipations

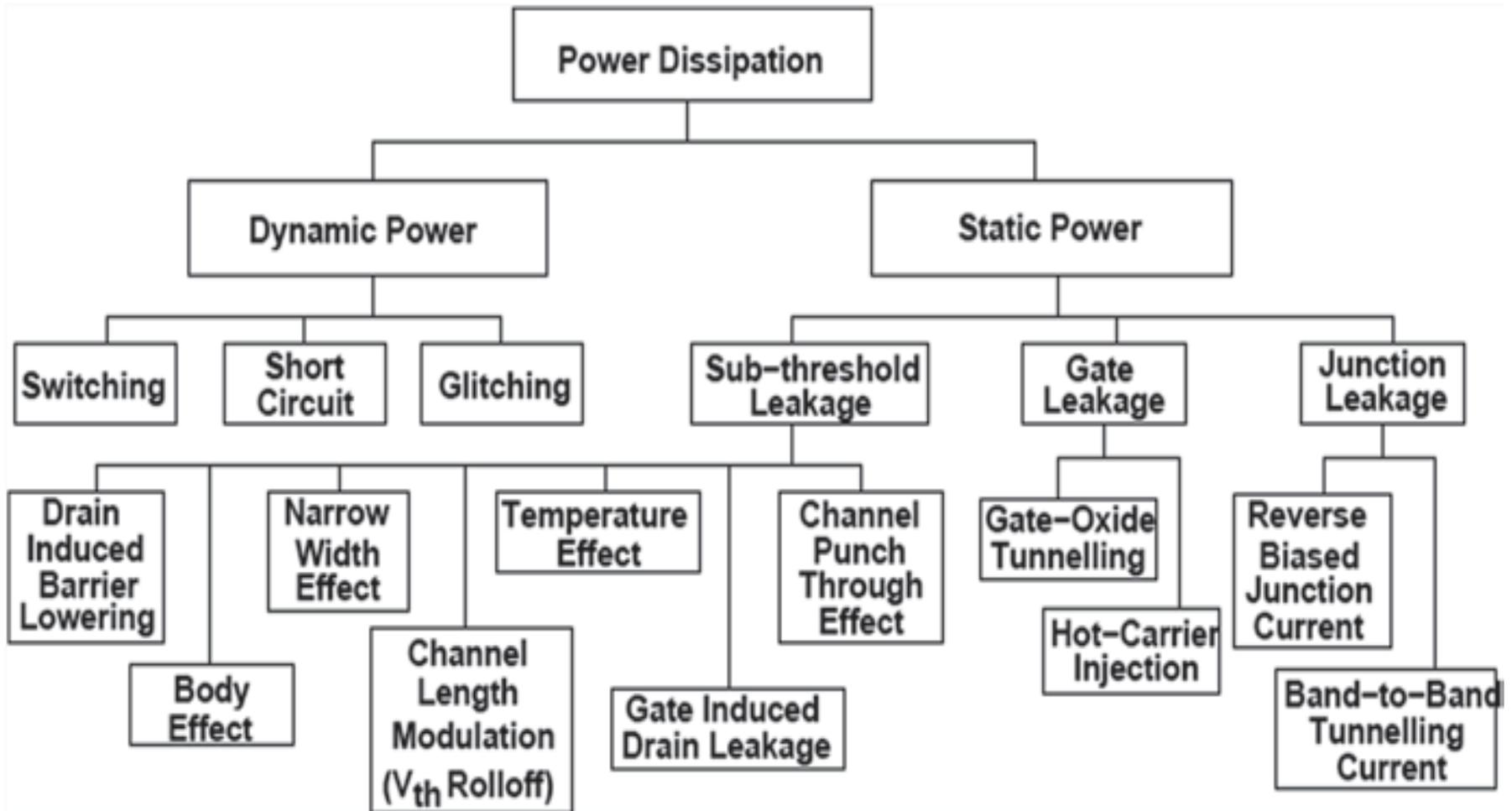
- Power Vs Energy



Sources of Power Dissipations

- 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.*

Sources of Power Dissipations



Sources of Power Dissipations

- **Dynamic Power**
 1. Switching Power
 2. Short-Circuit Power
 3. Glitching Power Dissipation
- **Static Power**

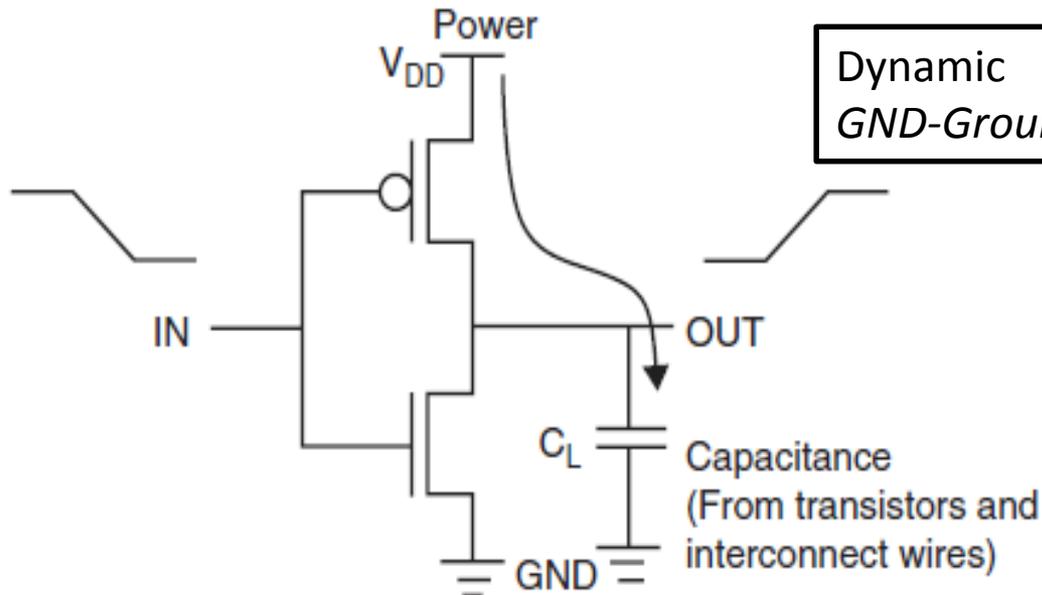
Sources of Power Dissipations

- **Dynamic Power**
- Dynamic power is the power consumed **when the device is active**, that is, when the signals of the design are changing values.
- It is generally categorized into three types: **switching power, short-circuit power, and glitching power**

Sources of Power Dissipations

- **Dynamic Power-Switching Power**
- The power required to charge and discharge the output capacitance on a gate.

SWITCHING POWER FOR CHARGING A CAPACITOR.



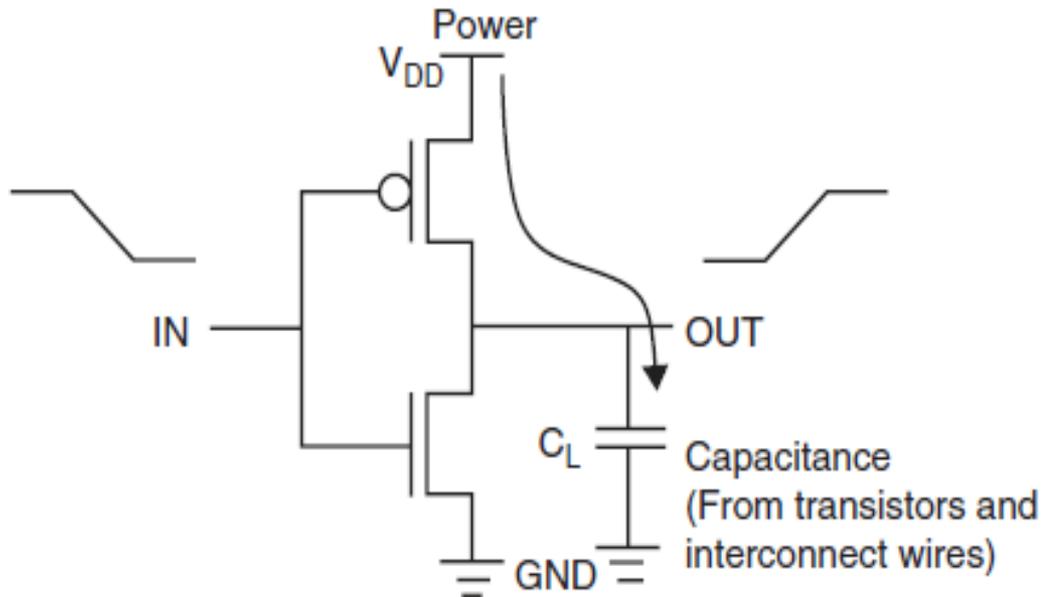
Dynamic (Switching) Power.
GND-Ground

Sources of Power Dissipations

- **Dynamic Power-Switching Power**
- The energy per transition is given by

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

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

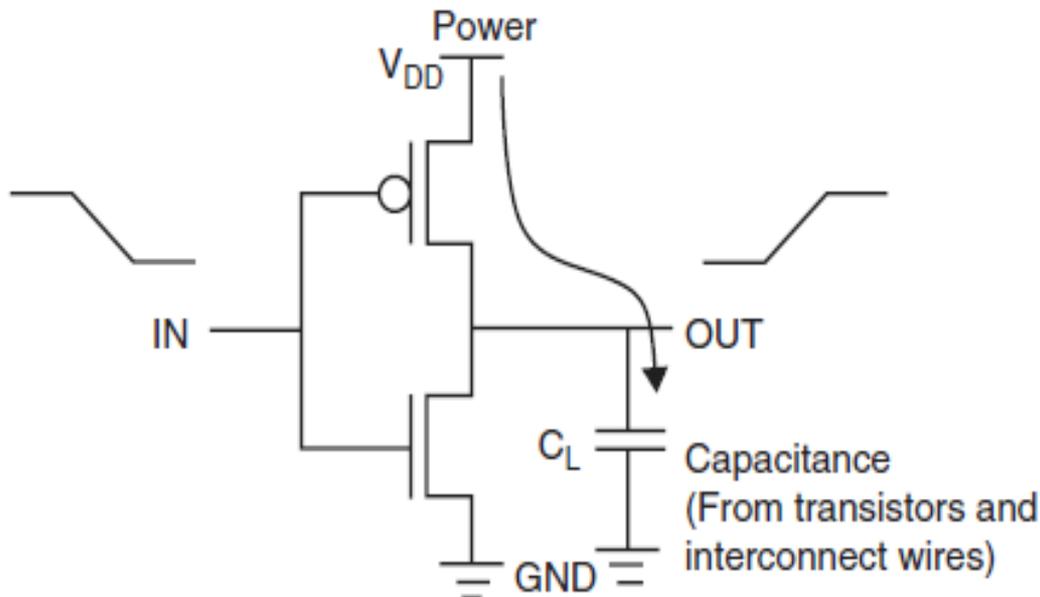


Sources of Power Dissipations

- **Dynamic Power-Switching Power**
- Switching power is therefore expressed as:

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

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.



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

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

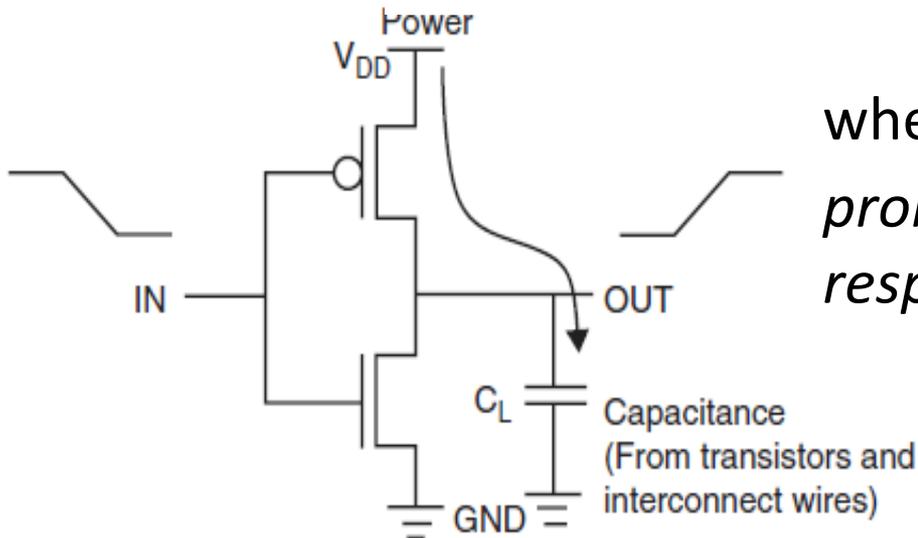
Sources of Power Dissipations

- **Dynamic Power-Switching Power**

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

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

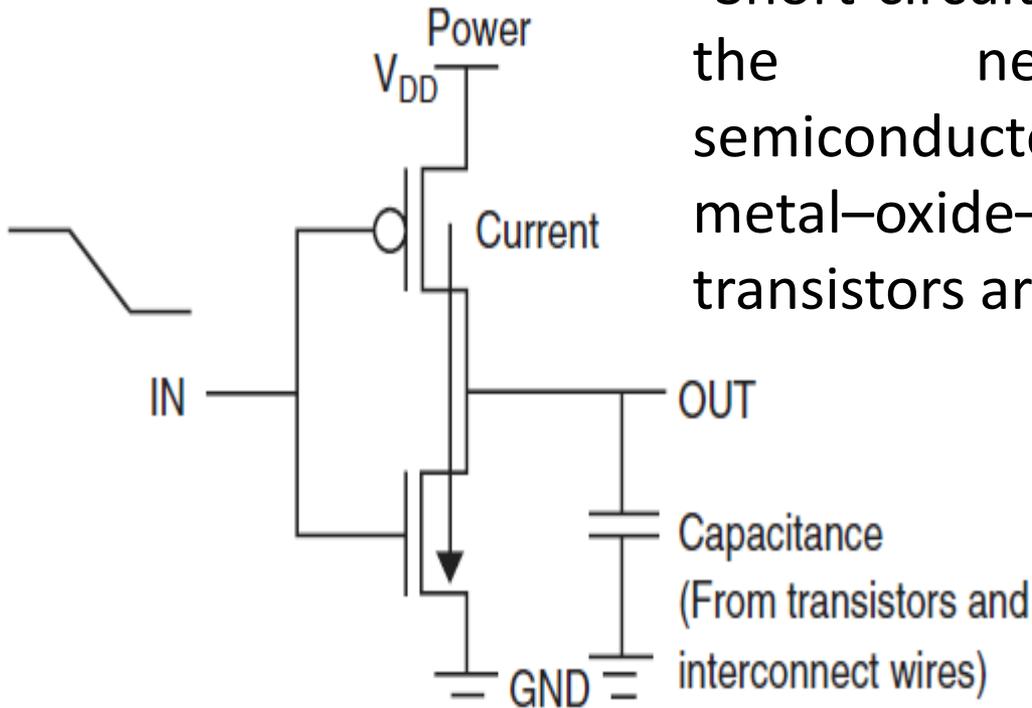


where α_i and C_i are the transition probability and capacitance, respectively, for an internal node i .

Sources of Power Dissipations

- **Dynamic Power-Short-Circuit Power**

Short-circuit current or crowbar current.



- Short-circuit currents occur when both the negative metal-oxide-semiconductor (NMOS) and positive metal-oxide-semiconductor (PMOS) transistors are ON.

Sources of Power Dissipations

- **Dynamic Power-Glitching Power Dissipation**

- The third type of dynamic power dissipation is the glitching power which arises due to finite delay of the gates.
- Since the dynamic power is directly proportional to the number of output transitions of a logic gate, glitching can be a significant source of signal activity and deserves mention here.
- Glitches often occur when paths with unequal propagation delays converge at the same point in the circuit.
- Glitches 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.

Sources of Power Dissipations

- **Static Power**

- **Static power dissipation** takes place as long as the device is powered ON, even when there are no signal changes.
- Normally in CMOS circuits, in the steady state, there is no direct path from *V_{dd}* to *GND* and so there should be no static power dissipation, but there are **various leakage current mechanisms** which are responsible for static power dissipation.
- Since the MOS transistors are not perfect switches, there will be leakage currents and substrate injection currents, which will give rise to static power dissipation in CMOS.

Sources of Power Dissipations

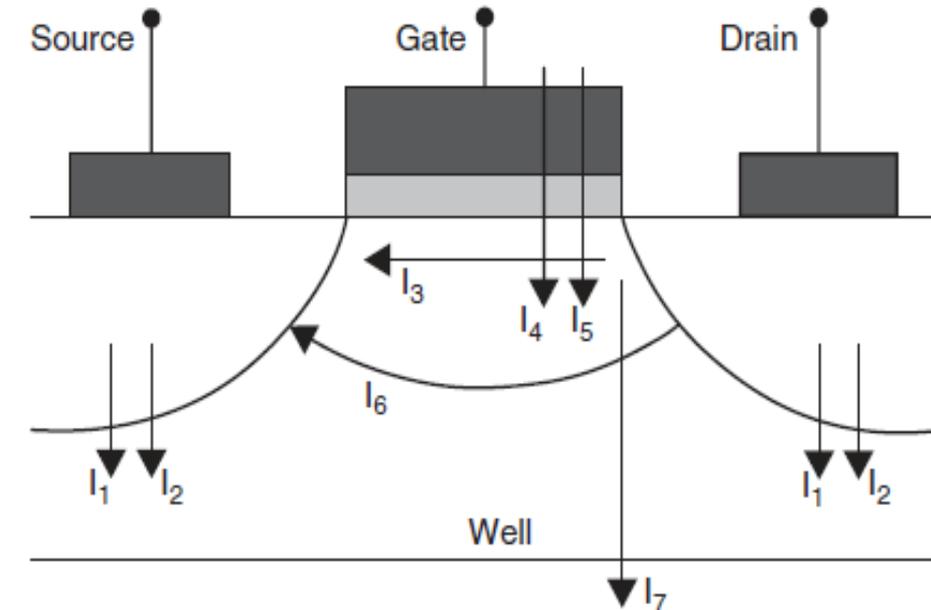
- **Static Power**

- Since the substrate current reaches its maximum for gate voltages near $0.4V_{dd}$ and *gate voltages are only transiently in this range when the devices switch*, the actual power contribution of substrate currents is negligible as compared to other sources of power dissipation.
- Leakage currents are also normally negligible, in the order of nano-amps, compared to dynamic power dissipation.
- But with deep submicron technologies, the leakage currents are increasing drastically to the extent that in 90-nm technology and thereby leakage power also has become comparable to dynamic power dissipation.

Sources of Power Dissipations

• Static Power

Leakage currents in an MOS Inverter-
Several leakage mechanisms that are responsible for static power dissipation.

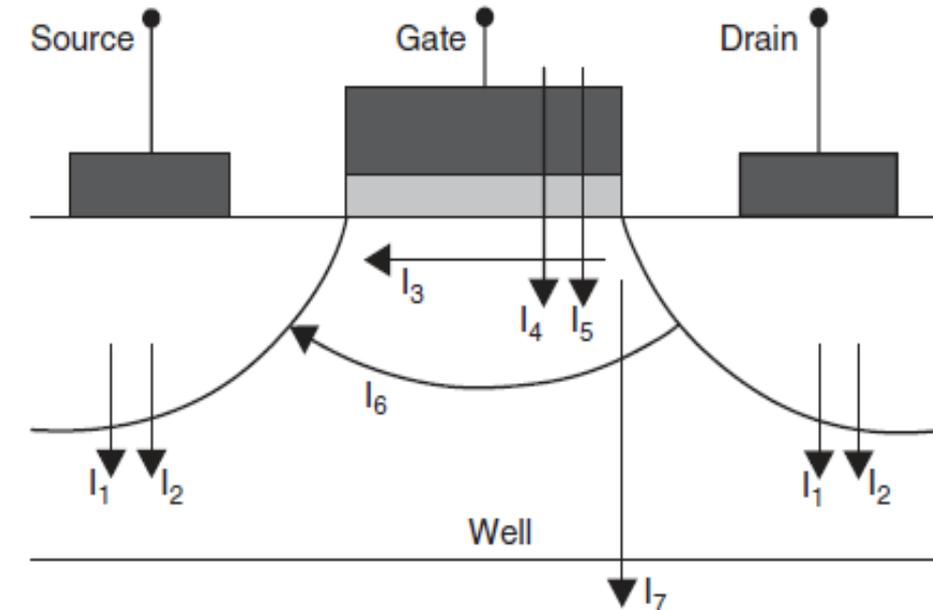


- ❖ **I_1** is the reverse-bias p-n junction diode leakage current
- ❖ **I_2** is the reverse-biased p-n junction current due to tunneling of electrons from the valence band of the p region to the conduction band of the n region
- ❖ **I_3** is the subthreshold leakage current between the source and the drain when the gate voltage is less than the threshold voltage (V_{th})

Sources of Power Dissipations

• Static Power

Leakage currents in an MOS Inverter-
Several leakage mechanisms that are responsible for static power dissipation.



❖ **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 IGATE leakage current)

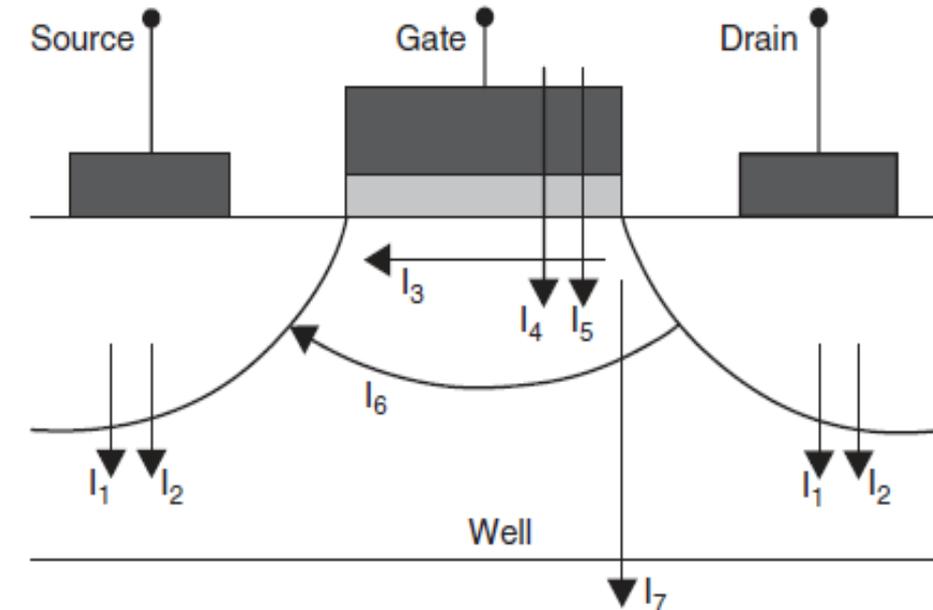
❖ **I_6** is the **gate-induced drain leakage current** due to high field effect in the drain junction

❖ **I_7** is the **channel punch through current** due to close proximity of the drain and the source in short-channel devices

Sources of Power Dissipations

- **Static Power**

Leakage currents in an MOS Inverter- Several leakage mechanisms that are responsible for static power dissipation.



These are generally categorized into four major types:

- ❖ **Subthreshold leakage**
- ❖ **Gate leakage**
- ❖ **Gate-induced drain leakage**
- ❖ **Junction leakage**

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**

Low-Power Design Methodologies

- **Low-power design methodology** needs to be applied throughout the design process starting from **system level to physical or device level** to get effective reduction of power dissipation in digital circuits based on MOS technology.
- Starting with the specifications the following steps are performed to get layout:
 - **System Specification** -SYSTEM-LEVEL Design
 - **Behavioral Description** -HIGH-LEVEL Synthesis
 - **Structural RTL Description** for Logic Synthesis
 - **Logic Level Net List** used that for Layout Synthesis and finally, you will get the layout for fabrication.
- So, throughout this design process you have to adopt low-power design methodology .

Low-Power Design Methodologies

- 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**.

Low-Power Design Methodologies

- 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.

Low-Power Design Methodologies

- 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.**

Low-Power Design Methodologies

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

Unit-1

- Introduction
- Historical Background
- Why Low Power
- Sources of Power Dissipations
- Low Power Methodologies

MOS Transistors

- Introduction
- The Structure of MOS Transistor
- The Fluid Model
- Modes of Operations of MOS Transistor
- Electrical Characteristics of MOS Transistors
- MOS Transistors as a Switch

Unit-1: MOS Fabrication Technology

- Introduction
- Basic Fabrication Processes
- nMOS Fabrication Steps

- *Wafer Fabrication*
- *Oxidation*
- *Mask Generation*
- *Photolithography*
- *Diffusion*
- *Deposition*

- CMOS Fabrication Steps

- *The n-Well Process*
- *The p-Well Process*
- *Twin-Tub Process*

- Latch-Up Problem and Its Prevention
- Short-Channel Effects

- *Use of Guard Rings*
- *Use of Trenches*

- *Channel Length Modulation Effect*
- *Drain-Induced Barrier Lowering*
- *Channel Punch Through*

- Emerging Technologies for Low Power

- *Hi-K Gate Dielectric*
- *Lightly Dopes Drain-Source*
- *Silicon on Insulator(SOI)*
- *FinFET*

Introduction

- **Metal–oxide–semiconductor (MOS) fabrication** is the process used to create the **integrated circuits (ICs)** that are presently used to realize electronic circuits.
- It involves multiple steps of **photolithographic and chemical processing steps** during which electronic circuits are gradually created on a **wafer** made of pure semiconducting material.
- **Silicon** is almost always used, but various compound semiconductors such as **gallium–arsenide** are used for specialized applications.

Introduction

- There are a large number and variety of **basic fabrication steps** used in the production of modern MOS ICs.
- The same process could be used for the fabrication of **n-type MOS (nMOS), p-type MOS (pMOS), or complementary MOS (CMOS) devices.**
- The **gate material** could be either **metal or poly-silicon.**
- The most commonly used substrate is either bulk **silicon or silicon on insulator (SOI).**
- In order to avoid the presence of parasitic transistors, variations are brought in the techniques that are used to isolate the devices in the wafer.

Basic Fabrication Processes

- Present day **very-large-scale integration (VLSI) technology** is based on **silicon**, which has bulk electrical resistance between that of a conductor and an **insulator**.
- That is why it is known as a **semiconductor material**.
- Its **conductivity** can be changed by several orders of magnitude by adding impurity atoms into the silicon crystal lattice.
- These impurity materials supply either **free electrons or holes**.

Basic Fabrication Processes

- The **donor** elements provide electrons and acceptor elements provide holes. Silicon having a majority of donors is known as **n-type**.
- On the other hand, silicon having a majority of acceptors is known as **p-type**.
- When n-type and p-type materials are put together, a junction is formed where the silicon changes from one type to the other type.
- Various semiconductor devices such as diode and transistors are constructed by arranging these junctions in certain physical structures and combining them with other types of physical structures, as we shall discuss in the subsequent sections.

Basic Fabrication Processes

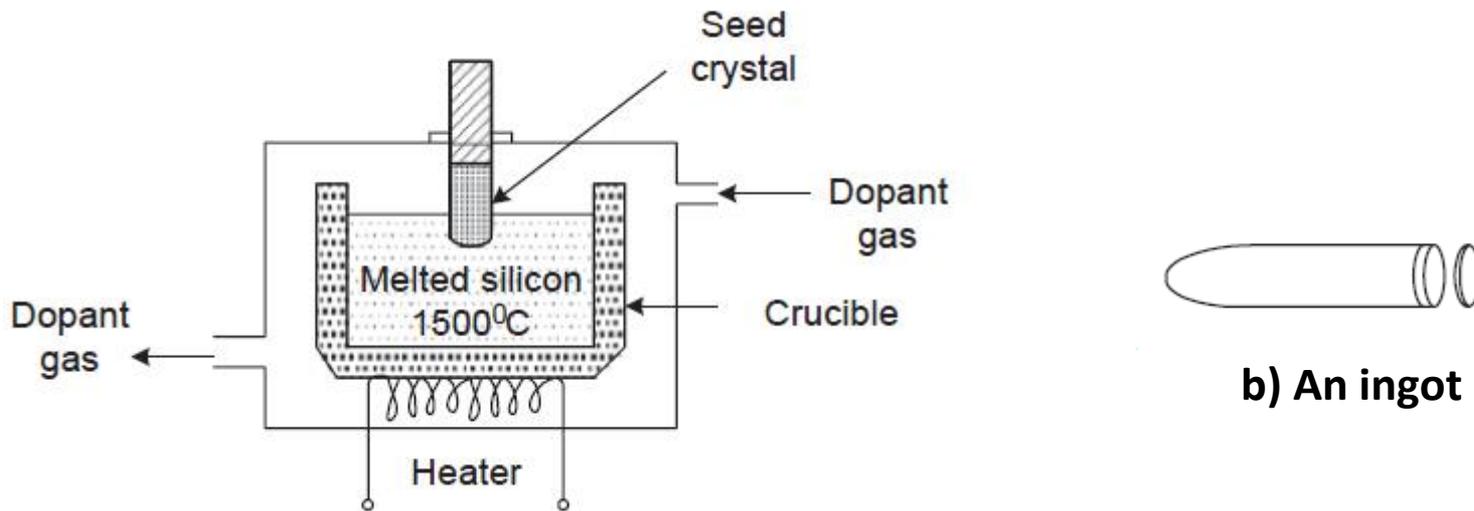
- *Wafer Fabrication*
- *Oxidation*
- *Mask Generation*
- *Photolithography*
- *Diffusion*
- *Deposition*

Basic Fabrication Processes-*Wafer Fabrication*

- The MOS fabrication process starts with a thin wafer of silicon.
- The raw material used for obtaining silicon wafer is sand or silicon dioxide.
- Sand is a cheap material and it is available in abundance on earth.
- However, it has to be purified to a high level by reacting with carbon and then crystallized by an epitaxial growth process.
- The purified silicon is held in molten state at about 1500 °C, and a seed crystal is slowly withdrawn after bringing in contact with the molten silicon.
- The atoms of the molten silicon attached to the seed cool down and take the crystalline structure of the seed.

Basic Fabrication Processes-*Wafer Fabrication*

- While forming this crystalline structure, the silicon is lightly doped by inserting controlled quantities of a suitable doping material into the crucible.
- The set up is for wafer fabrication to produce nMOS devices is shown in Figure.

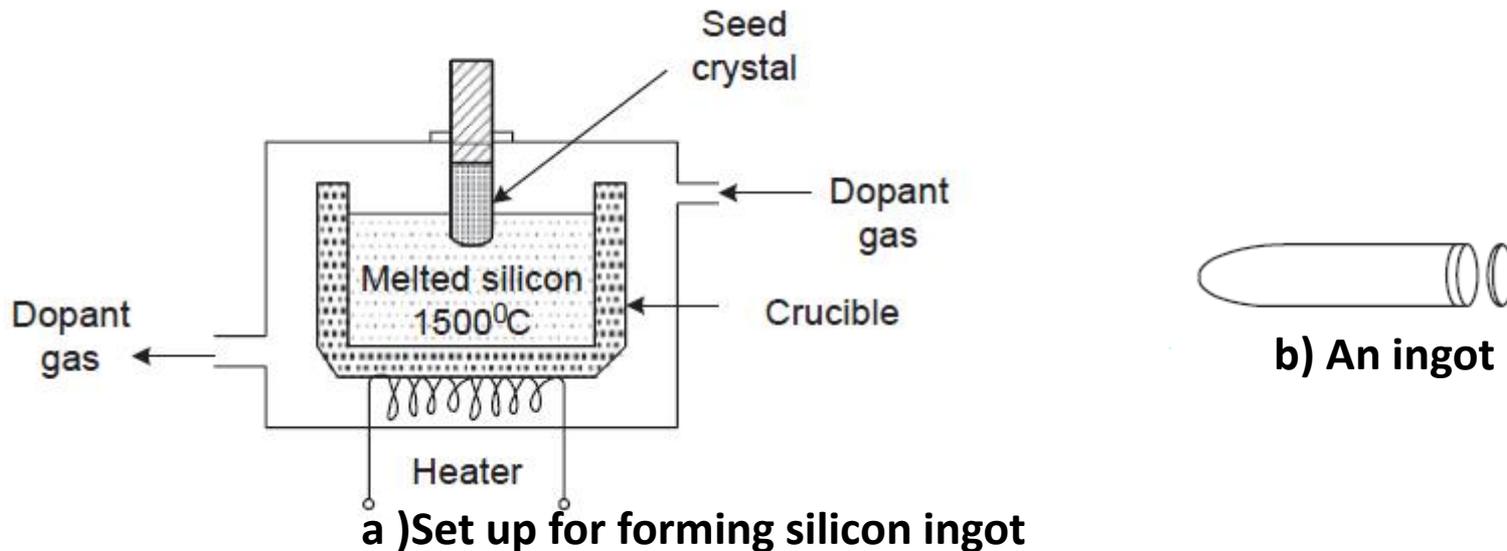


a) Set up for forming silicon ingot

b) An ingot

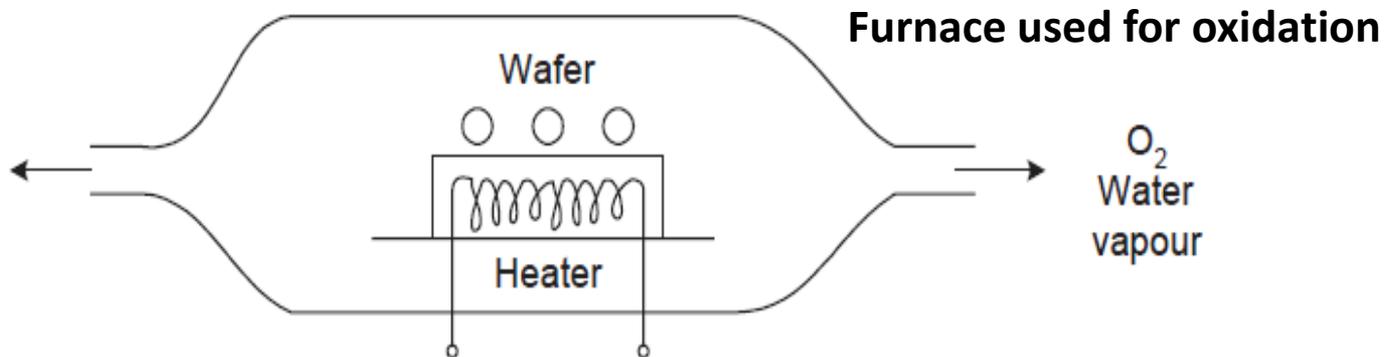
Basic Fabrication Processes-*Wafer Fabrication*

- Here, boron may be used to produce p-type impurity concentration of 10^{15} cm³ to 10^{16} per cm³.
- It gives resistivity in the range of 25–2 Ω cm. After the withdrawal of the seed, an “ingot” of several centimeters length and about 8–10 cm diameter as shown in Figure is obtained.
- The ingot is cut into slices of 0.3–0.4 mm thickness to obtain wafer for IC fabrication.



Basic Fabrication Processes-*Oxidation*

- Silicon dioxide layers are used as an insulating separator between different conducting layers.
- It also acts as mask or protective layer against diffusion and high-energy ion implantation.
- The process of growing oxide layers is known as oxidation because it is performed by a chemical reaction between oxygen (dry oxidation), or oxygen and water vapor (wet oxidation) and the silicon slice surface in a high temperature furnace at about 1000 °C as shown in Figure.



Basic Fabrication Processes-*Mask Generation*

- To create patterned layers of different materials on the wafer, masks are used at different stages.
- Masks are made of either inexpensive green glass or costly low expansion glass plates with opaque and transparent regions created using photographic emulsion, which is cheap but easily damaged.
- Other alternative materials used for creating masks are iron oxide or chromium, both of which are more durable and give better line resolution, but are more expensive

Basic Fabrication Processes-*Mask Generation*

- A mask can be generated either optically or with the help of an electron beam.
- In the optical process, a reticle, which is a photographic plate of exactly ten times the actual size of the mask, is produced as master copy of the mask.
- Transparent and opaque regions are created with the help of a pattern generator by projecting an image of the master onto the reticle.
- Special masking features such as parity masks and fiducials are used on the reticle to identify, align, and orient the mask.

Basic Fabrication Processes-*Mask Generation*

- Master plates are generated from reticles in a step-and-repeat process by projecting an image of the reticle ten times reduced onto the photosensitized plate to create an array of geometrical shapes in one over the entire plate.
- Fiducials are used to control the separation between exposures and align the reticle images relative to one another.
- This process has the disadvantage that if there is a defect on the reticle, it is reproduced on all the chips.
- The step-and-repeat process not only is slow but also suffers from alignment problems and defect propagation due to dust specks.
- The electron beam mask generation technique overcomes these problems.

Basic Fabrication Processes-*Mask Generation*

- Using this technique, several different chip types can be imprinted on the same set of masks.
- The main disadvantage of this approach is that it is a sequential technique.
- A better alternative is to use the soft X-ray photolithographic technique in which the entire chip can be eradicated simultaneously.
- This technique also gives higher resolution.
- These master plates are usually not used for mask fabrication. Working plates made from the masters by contact printing are used for fabrication.
- To reduce turnaround time, specially made master plates can be used for wafer fabrication.

Basic Fabrication Processes-*Photolithography*

- The photolithographic technique is used to create patterned layers of different materials on the wafer with the help of mask plates. It involves several steps.
- The first step is to put a coating of photosensitive emulsion called photo-resist on the wafer surface.
- After applying the emulsion on the surface, the wafer is spun at high speed (3000 rpm) to get a very thin (0.5–1 μm) and uniform layer of the photo-resist.
- Then the masking plate is placed in contact with the wafer in a precise position and exposed to the UV light.

Basic Fabrication Processes-*Photolithography*

- The mask plate, with its transparent and opaque regions, defines different areas. With negative photo-resist, the areas of the wafer exposed to UV light are polymerized (or hardened), while with positive photo-resist, the exposed areas are softened and removed.
- The removal of the unwanted photo-resist regions is done by a process known as development.
- Unexposed (negative) or exposed (positive) portions of the photoresist are chemically dissolved at the time of development.
- A low-temperature baking process hardens the subsequently remaining portion.

Basic Fabrication Processes-*Photolithography*

- To create the desired pattern, actual removal of the material is done by the *etching* process.
- The wafer is immersed in a suitable etching solution, which eats out the exposed material leaving the material beneath the protective photo-resist intact.
- The etching solution depends on the material to be etched out. Hydrofluoric acid (HF) is used for SiO₂ and poly-silicon, whereas phosphoric acid is used for nitride and metal.

Basic Fabrication Processes-*Photolithography*

- Another alternative to this wet chemical etching process is the plasma etching or ion etching.
- In this dry process, a stream of ions or electrons is used to blast the material away.
- Ions created by glow discharge at low pressure are directed to the
- target.
- Ions can typically penetrate about 800 Å of oxide or photo-resist layers, and thick layers of these materials are used as a mask of some area, whereas the exposed material is being sputtered away.
- This plasma technique can produce vertical etching with little undercutting.
- As a consequence, it is commonly used for producing fine lines and small geometries associated with high-density VLSI circuits.

Basic Fabrication Processes-*Photolithography*

- Finally, the photo-resist material is removed by a chemical reaction of this material with fuming nitric acid or exposure to atomic oxygen which oxidizes away the photo-resist.
- Patterned layers of different materials in engraved form are left at the end of this process

Basic Fabrication Processes-Diffusion

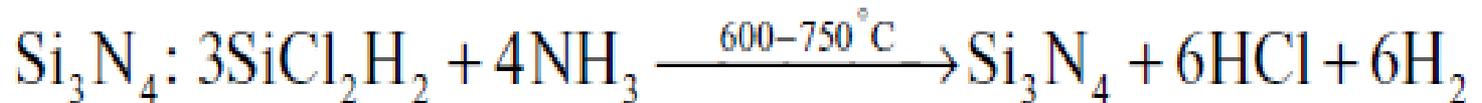
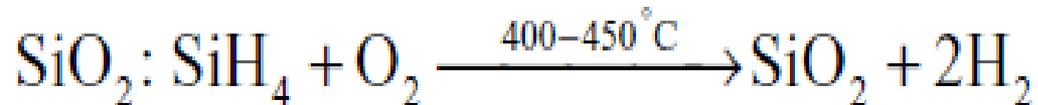
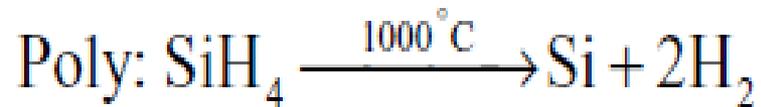
- After masking some parts of the silicon surface, selective *diffusion can be done in* the exposed regions.
- There are two basic steps: pre-deposition and drive-in.
- In the pre-deposition step, the wafer is heated in a furnace at 1000 °C, and dopant atoms such as phosphorous or boron mixed with an inert gas, say nitrogen, are introduced into it.
- Diffusion of these atoms takes place onto the surface of the silicon, forming a saturated solution of the dopant atoms and solid.
- The impurity concentration goes up with a temperature up to 1300 °C and then drops.
- The depth of penetration depends on the duration for which the process is carried out.
- In the drive-in step, the wafer is heated in an inert atmosphere for few hours to distribute the atoms more uniformly and to a higher depth.

Basic Fabrication Processes-Diffusion

- Another alternative method for diffusion is *ion implantation*.
- *Dopant gas* is first ionized with the help of an ionizer and ionized atoms are accelerated between two electrodes with a voltage difference of 150 kV.
- The accelerated gas is passed through a strong magnetic field, which separates the stream of dopant ions on the basis of molecular weights, as it happens in mass spectroscopy.
- The stream of these dopant ions is deflected by the magnetic field to hit the wafer.
- The ions strike the silicon surface at high velocity and penetrate the silicon layer to a certain depth as determined by the concentration of ions and accelerating field.
- This process is also followed by drive-in step to achieve uniform distribution of the ions and increase the depth of penetration.

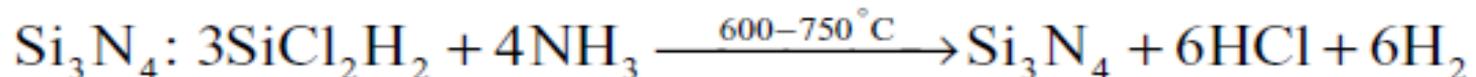
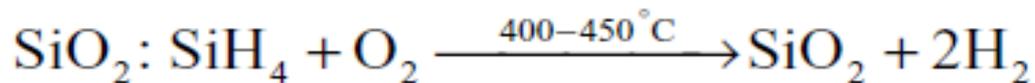
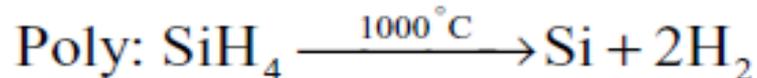
Basic Fabrication Processes-Deposition

- In the MOS fabrication process, conducting layers such as polysilicon and aluminium, and insulation and protection layers such as SiO₂ and Si₃N₄ are deposited onto the wafer surface by using the chemical vapor deposition (CVD) technique in a high-temperature chamber:



Basic Fabrication Processes-Deposition

- Poly-silicon is deposited simply by heating silane at about 1000 °C, which releases hydrogen gas from silane and deposits silicon.
- To deposit silicon dioxide, a mixture of nitrogen, silane, and oxygen is introduced at 400–450 °C.
- Silane reacts with oxygen to produce silicon dioxide, which is deposited on the wafer.
- To deposit silicon nitride, silane and ammonia are heated at about 700 °C to produce nitride and hydrogen.
- Aluminium is deposited by vaporizing aluminium from a heated filament in high vacuum.



CMOS Fabrication Steps

- There are several approaches for CMOS fabrication, namely, p-well, n-well, twintub, triple-well, and SOI.
- The n-well approach is compatible with the nMOS process and can be easily retrofitted to it.
- However, the most popular approach is the p-well approach, which is similar to the n-well approach.
- The twin-tub and silicon on sapphire are more complex and costly approaches.
- These are used to produce superior quality devices to overcome the *latch-up problem*, which is predominant in CMOS devices.

CMOS Fabrication Steps-**n-Well Process**

- The most popular approach for the fabrication of n-well CMOS starts with a lightly doped p-type substrate and creates the n-type well for the fabrication of pMOS transistors.
- Major steps for n-well CMOS process are illustrated as follows:
- **Step 1**
- **The basic idea behind the n-well process is the formation of an n-well or tub** in the p-type substrate and fabrication of p-transistors within this well.
- The formation of an n-well by ion implantation is followed by a drive-in step ($1.8 \times 10^{22} \text{ p cm}^{-2}$, 80 kV with 1150 °C for 15 h of drive-in).
- This step requires a mask (MASK 1), which defines the deep n-well diffusions.
- The n-transistor is formed outside the well.

CMOS Fabrication Steps-**n-Well Process**

Step 1

- The basic steps are mentioned below:
 - Start with a blank wafer, commonly known as a **substrate**, which is lightly doped.



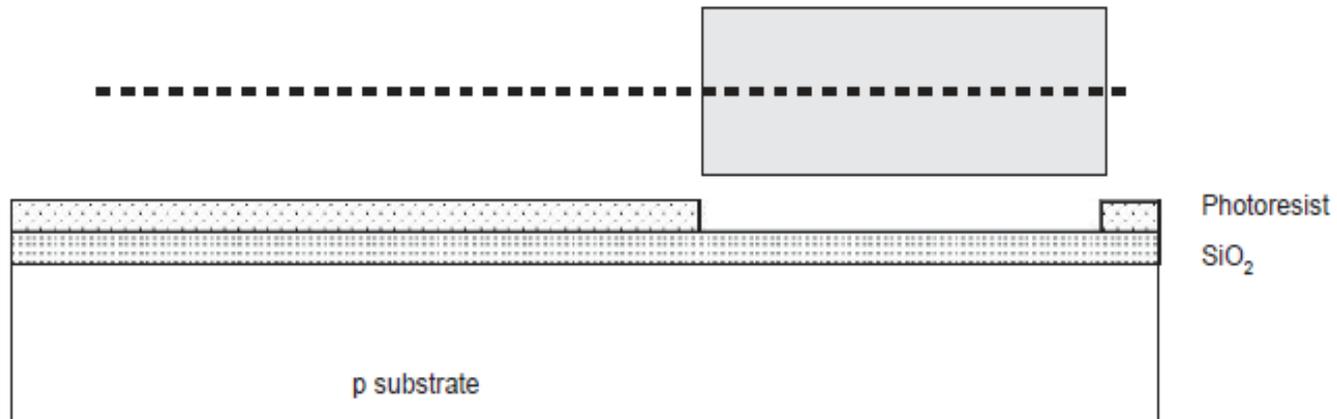
- Cover the wafer with a protective layer of SiO₂ (oxide) using the oxidation process at 900–1200 °C with H₂O (wet oxidation) or O₂ (dry oxidation) in the oxidation furnace.



CMOS Fabrication Steps-n-Well Process

Step 1

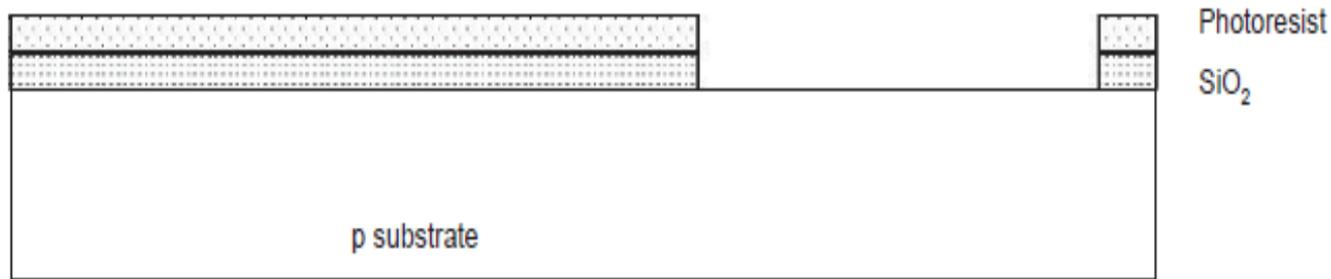
- The basic steps are mentioned below:
 - Expose photoresist through the n-well mask and strip off the exposed photoresist using organic solvents. The n-well mask used to define the n-well in this step is shown below.



CMOS Fabrication Steps-n-Well Process

Step 1

- The basic steps are mentioned below:
- Etch oxide with HF, which only attacks oxide where the resist has been exposed.



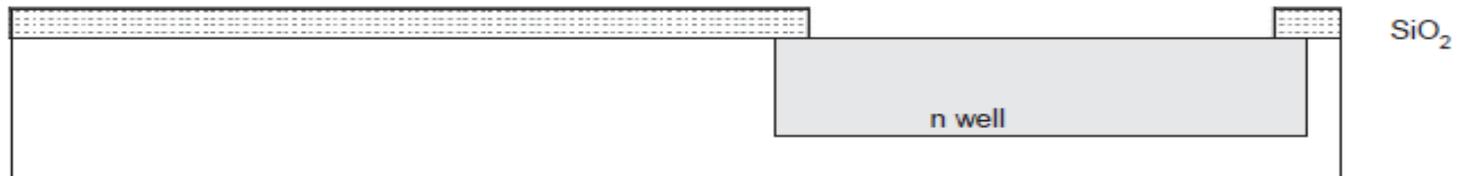
- Remove the photoresist, which exposes the wafer.



CMOS Fabrication Steps-**n-Well Process**

Step 1

- The basic steps are mentioned below:
- Implant or diffuse *n dopants into the exposed wafer using diffusion or ion implantation*. The ion implantation process allows shallower wells suitable for the fabrication of devices of smaller dimensions. The diffusion process occurs in all directions and deeper the diffusion more it spreads laterally. This affects how closely two separate structures can be fabricated.



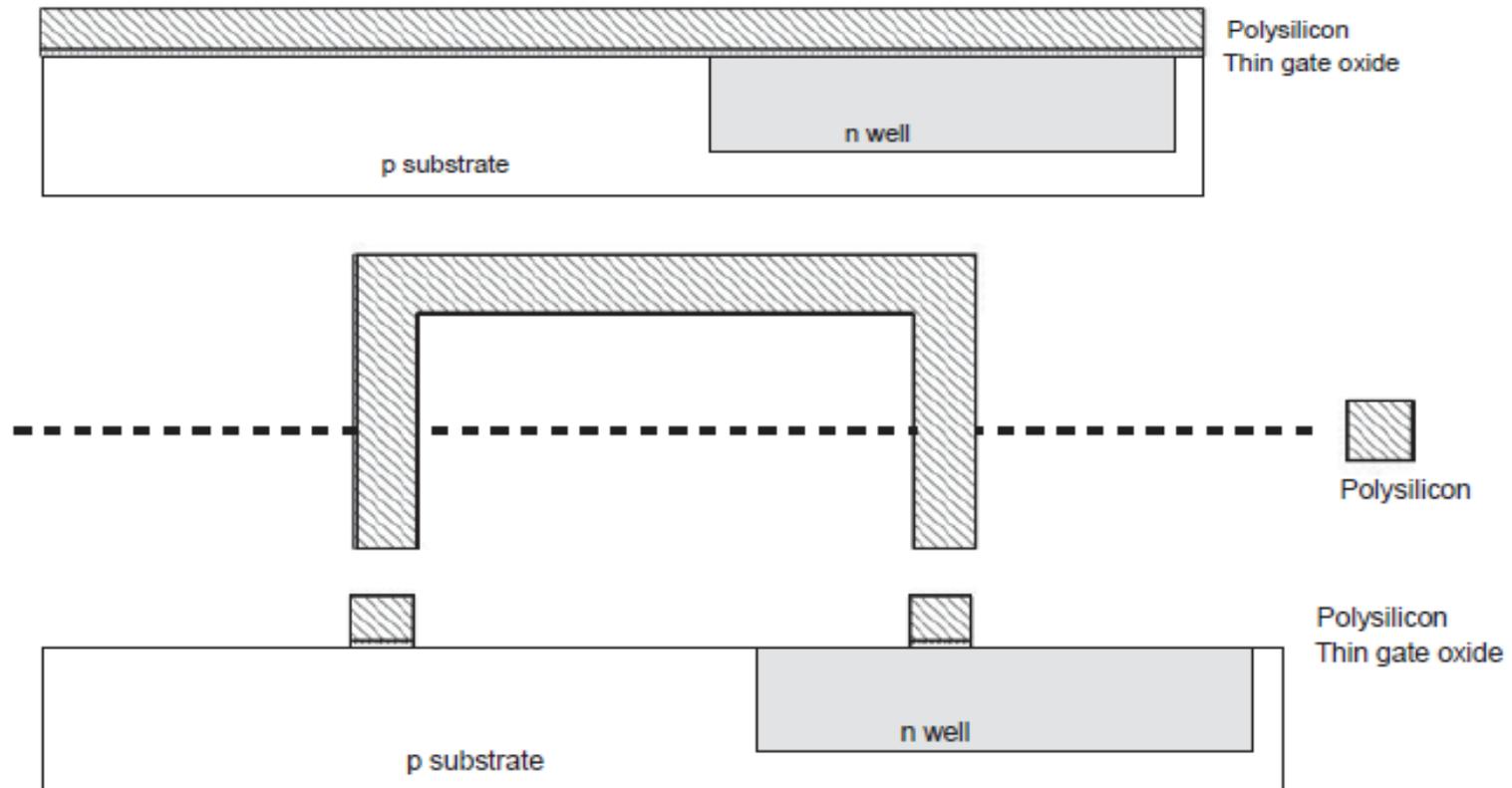
- Strip off SiO₂ leaving behind the p-substrate along with the n-well.



CMOS Fabrication Steps-**n-Well Process**

Step 2

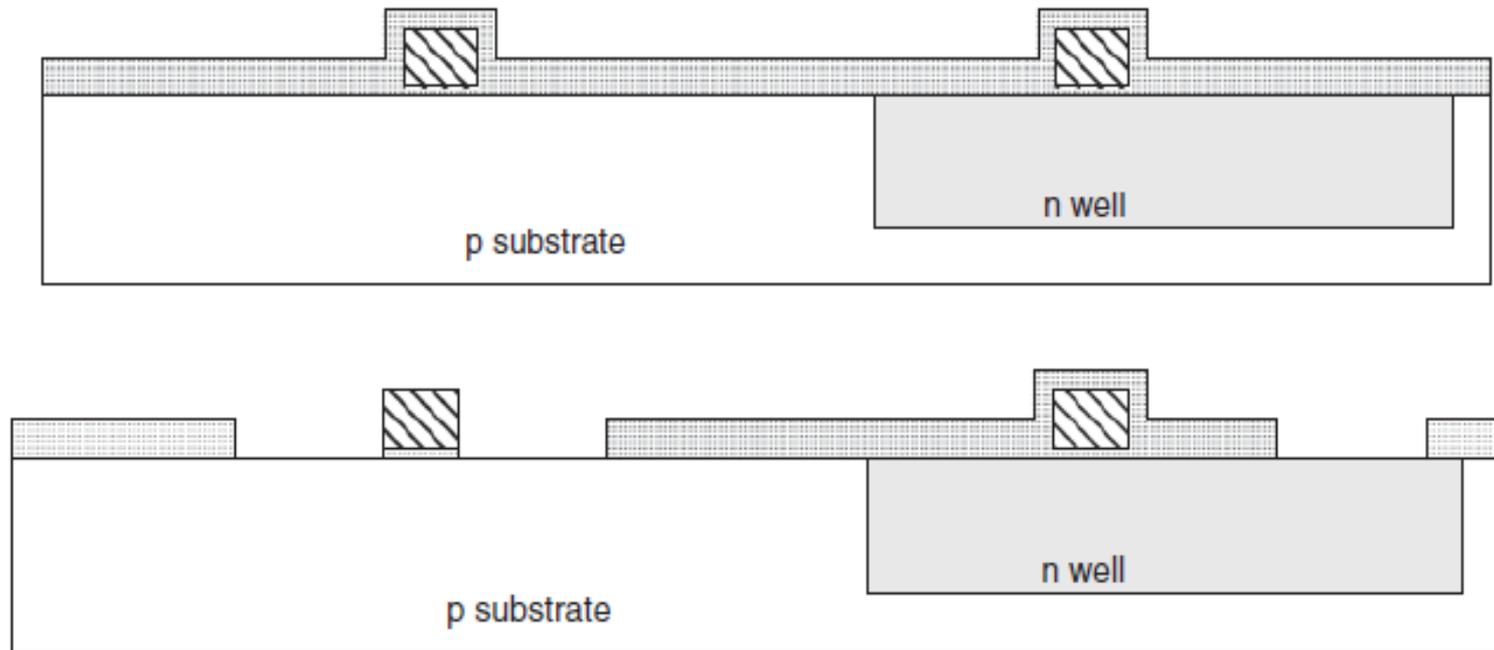
- The formation of thin oxide regions for the formation of p- and n-transistors requires MASK 2, which is also known as active mask because it defines the thin oxide regions where gates are formed.



CMOS Fabrication Steps-**n-Well Process**

Step 3

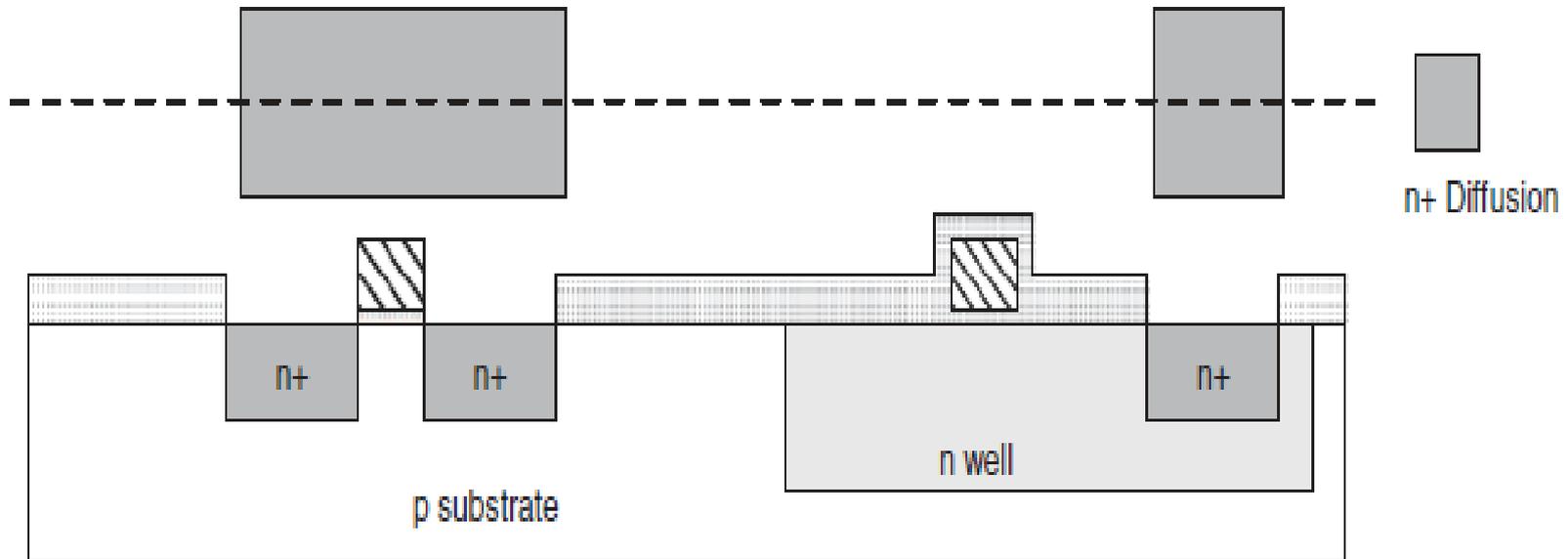
- The formation of patterned poly-silicon (nitride on the thin oxide) regions is done using MASK 3.
- Patterned poly-silicon is used for interconnecting different terminals.



CMOS Fabrication Steps-**n-Well Process**

Step 4

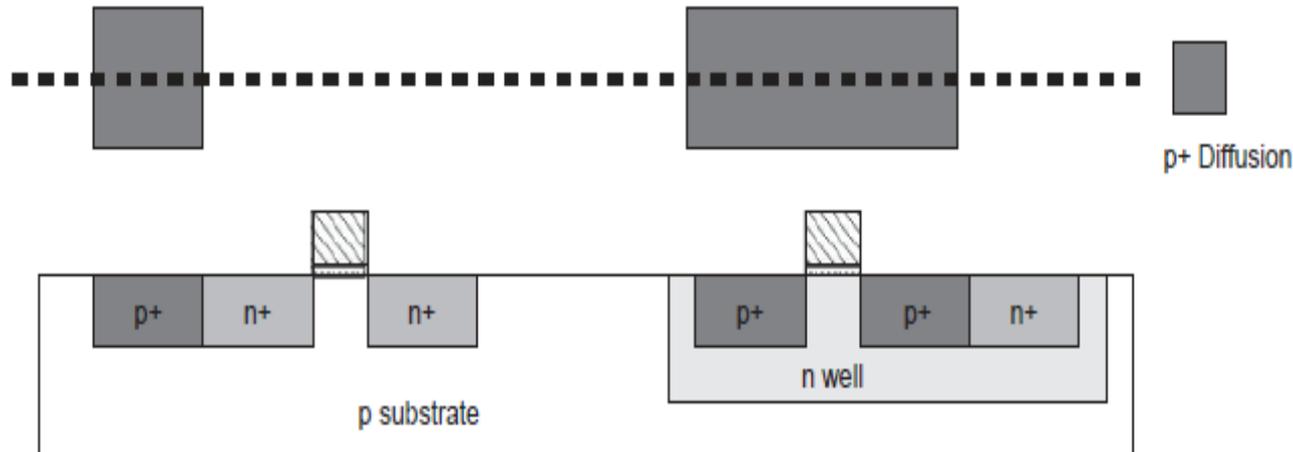
- The formation of n-diffusion is done with the help of the n+ mask, which is essentially MASK 4.



CMOS Fabrication Steps-n-Well Process

Step 5

- The formation of p-diffusion is done using the p+ mask, which is usually a negative form of the n+ mask. Similar sets of steps form p+ diffusion regions for the pMOS source and drain and substrate contact.

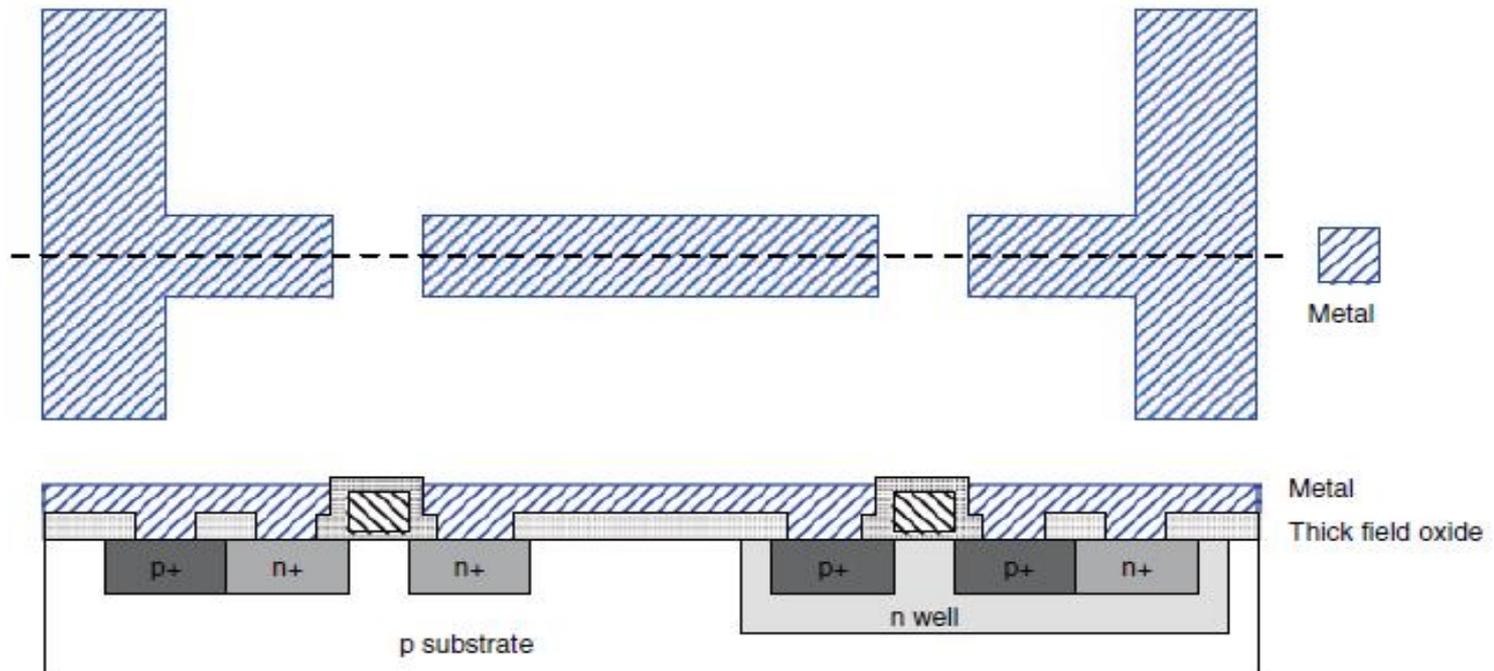


- Step 6 Thick SiO₂ is grown all over and then contact cut definition using another mask.

CMOS Fabrication Steps-**n-Well Process**

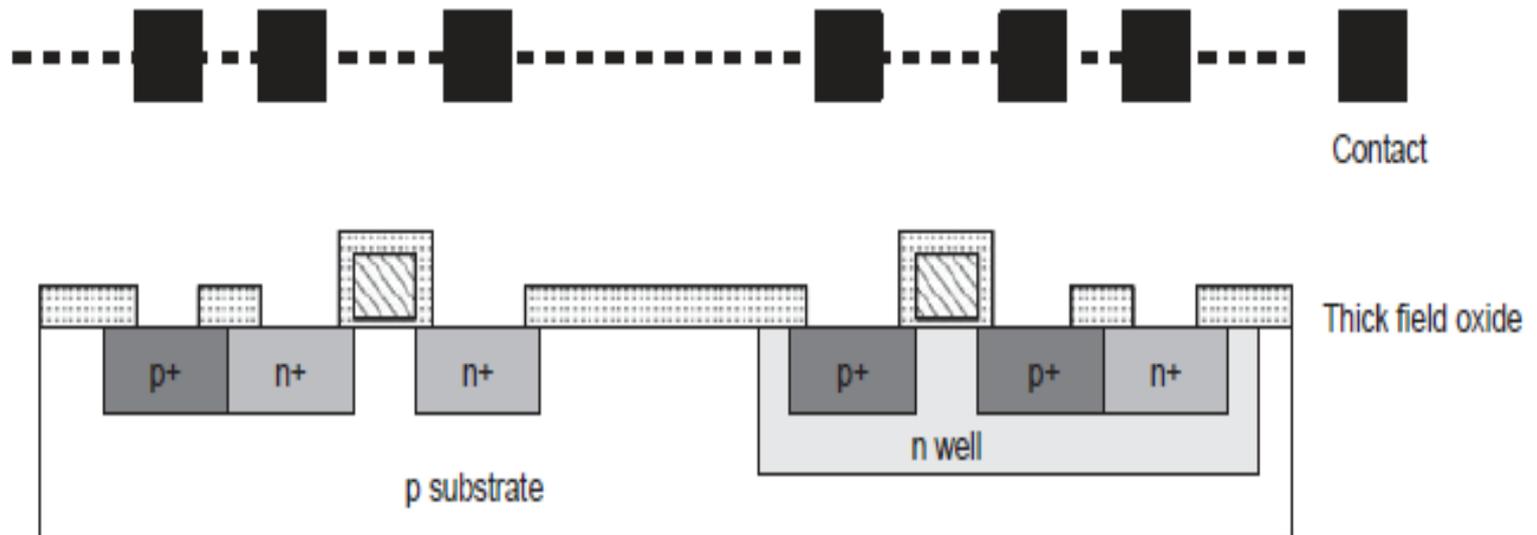
Step 7

- The whole chip then has metal deposited over its surface to a thickness of 1 μm .
- The metal layer is then patterned by the photolithographic process to form interconnection patterns using MASK 7.



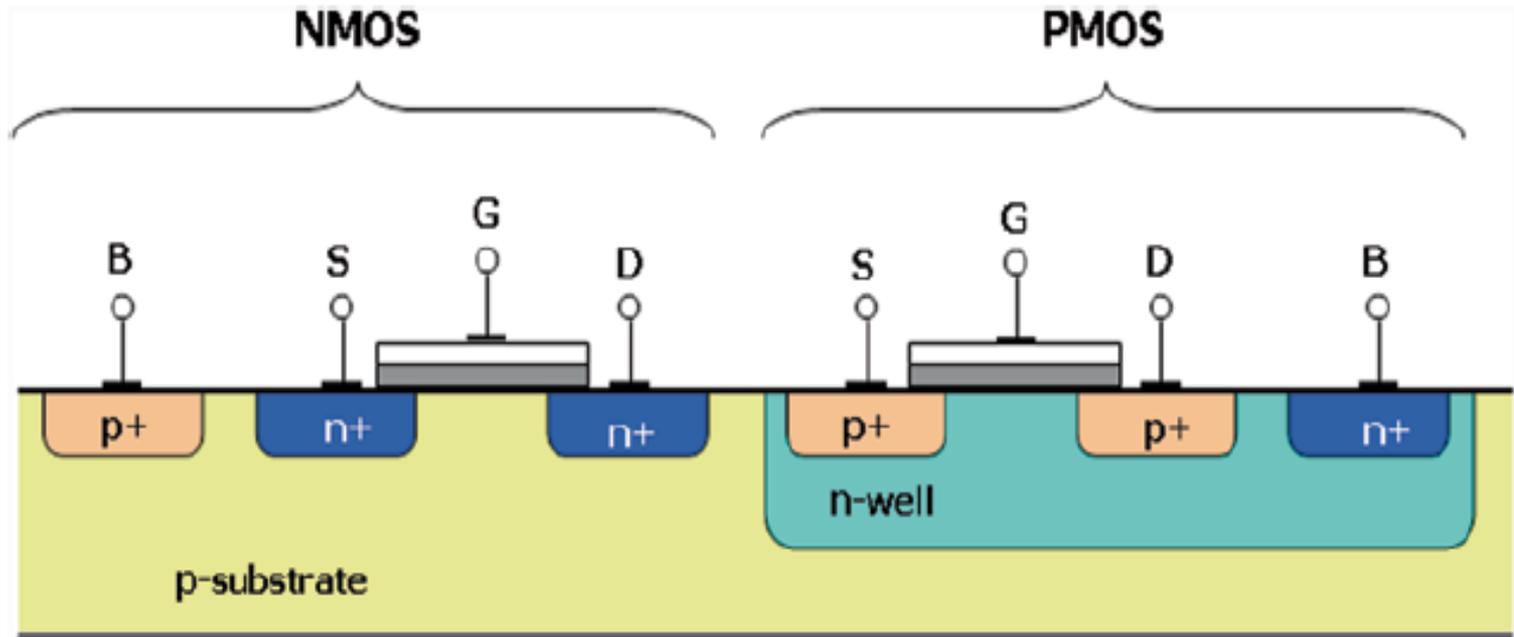
CMOS Fabrication Steps-**n-Well Process**

- **Step 8**
- Over-glassing is done by an overall passivation layer and a mask is required to define the openings for access to bonding pads (MASK 8).



CMOS Fabrication Steps-**n-Well Process**

- Two transistors, one pMOS and another nMOS, which can be used to realize a CMOS inverter are formed using the n-well process shown in Figure.



CMOS transistors realized using n-well process

CMOS Fabrication Steps-**p-Well Process**

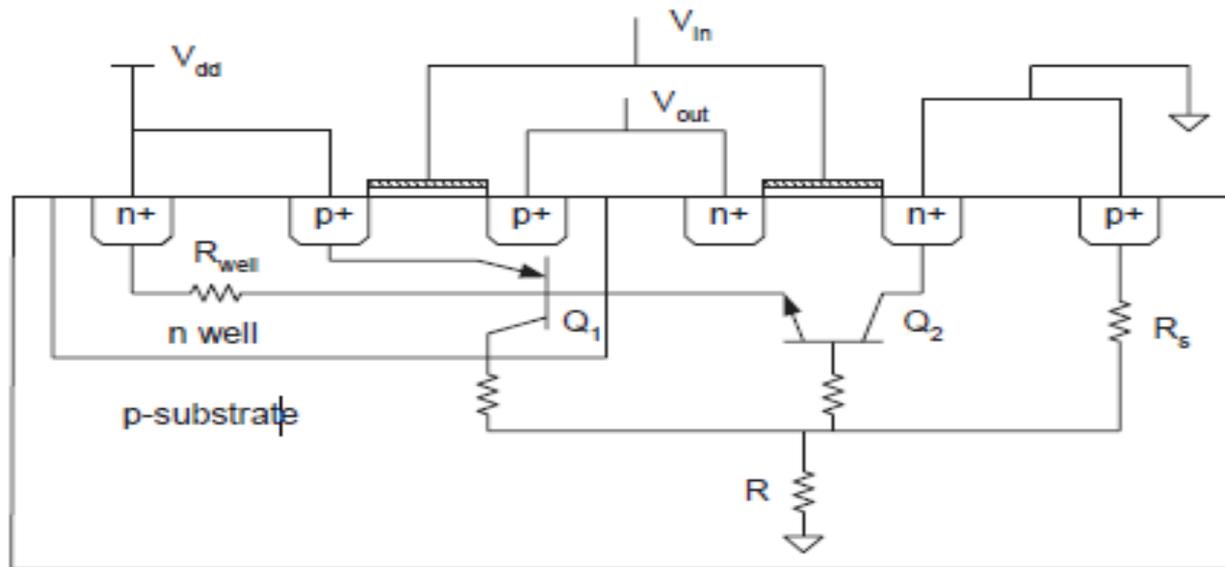
- Typical p-well fabrication steps are similar to an n-well process, except that a pwell is implanted to form n-transistors rather than an n-well.
- p-Well processes are preferred in circumstances where the characteristics of the n- and p-transistors are required to be more balanced than that achievable in an n-well process.
- Because the transistor that resides in the native substrate has been found to have better characteristics, the p-well process has better p-devices than an n-well process.

CMOS Fabrication Steps-Twin-Tub Process

- In the twin-tub process, the starting material is either an n+ or p+ substrate with a lightly doped epitaxial layer, which is used for protection against latch-up.
- The process is similar to the n-well process, involving the following steps:
 - Tub formation
 - Thin oxide construction
 - Source and drain implantations
 - Contact cut definition
 - Metallization

Latch-Up Problem and Its Prevention

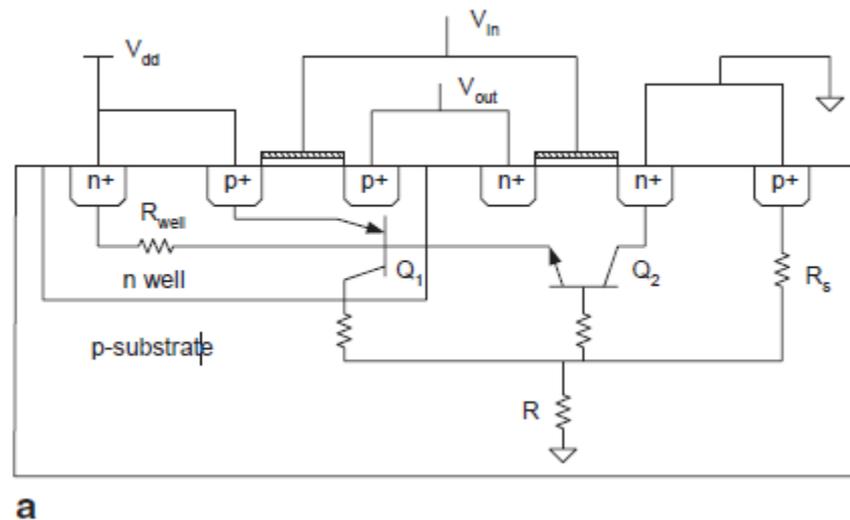
- The *latch-up* is an inherent problem in both *n-well-* and *p-well-based CMOS* circuits.
- The phenomenon is caused by the parasitic bipolar transistors formed in the bulk of silicon as shown in below figure for the *n-well* process.



a

Latch-Up Problem and Its Prevention

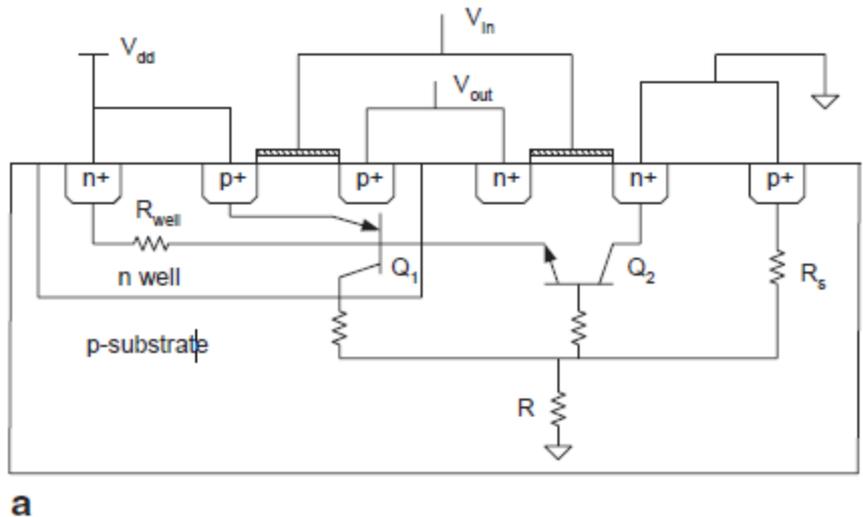
- Latch-up can be defined as the formation of a low-impedance path between the power supply and ground rails through the parasitic n-p-n and p-n-p bipolar transistors.



a
Cross section of a CMOS inverter

Latch-Up Problem and Its Prevention

- Two parasitic bipolar transistors, Q1 and Q2 are shown in the figure.
- The p-n-p transistor has its emitter formed by the p+source/drain implant used in the pMOS transistors.
- It may be noted that either the drain or the source may act as the emitter, although the source is the terminal that maintains the latch-up condition.

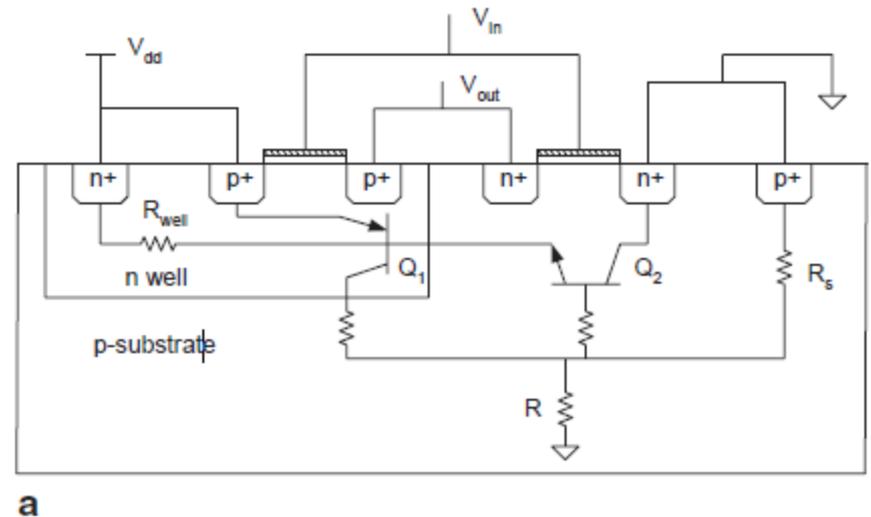


Cross section of a CMOS inverter

Latch-Up Problem and Its Prevention

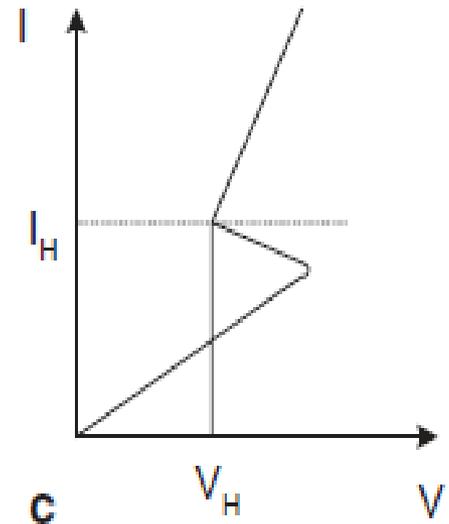
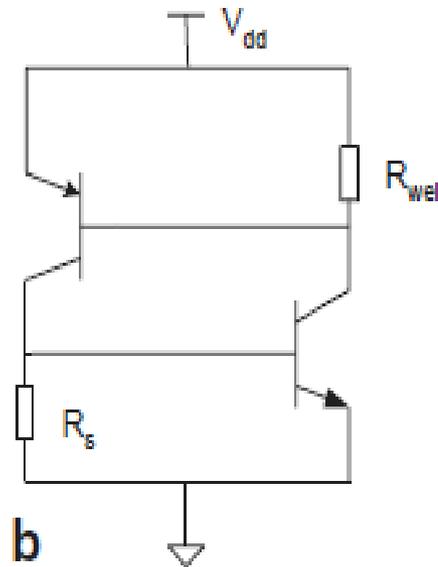
- The base is formed by the n-well, and the collector is formed by the p-substrate. The emitter of the n-p-n transistor is the n+ source/drain implant.
- The base is formed by the p-substrate and the collector is the n-well.
- The parasitic resistors R_{well} and R_s are formed because of the resistivity of the semiconductor material in the n-well and p-substrate, respectively.

Cross section of a CMOS inverter



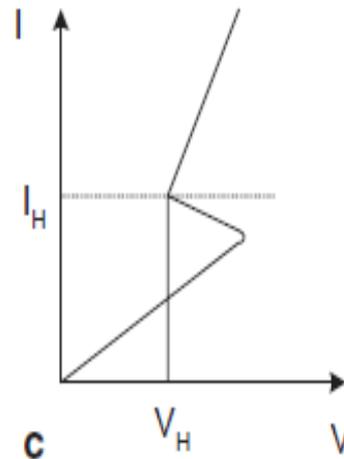
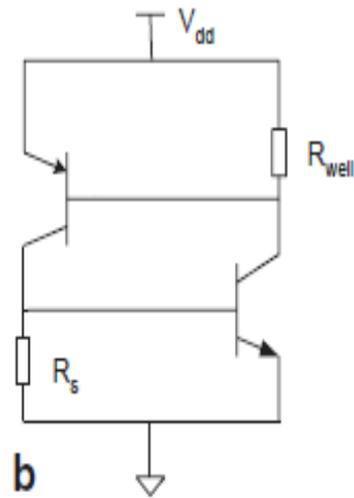
Latch-Up Problem and Its Prevention

- The bipolar junction transistors (BJTs) are cross-coupled to form the structure of a silicon-controlled rectifier (SCR) providing a short-circuit path between the power rail and the ground.
- Leakage current through the parasitic resistors can cause one transistor to turn on, which in turn turns on the other transistor due to positive feedback, leading to heavy current flow and device failure.



Latch-Up Problem and Its Prevention

- The latch-up condition is sustained as long as the current is greater than the *holding current* I_H ; the *holding current value depends* on the total parasitic resistance R_T in the current path.
- *There are several approaches* to reduce the tendency of latch-up. The slope of the I - V curve *depends on* the total parasitic resistance R_T in the current path.



Latch-Up Problem and Its Prevention

- The possibility of internal latchup can be reduced to a great extent by using the following rules:
 1. Every well must have an appropriate substrate contact.
 2. Every substrate contact should be directly connected to a supply pad by metal.
 3. Substrate contacts should be placed as close as possible to the source connection of transistors to the supply rails. This helps to reduce the value of both *R_s* and *R_{well}*.
 4. Alternatively, place a substrate contact for every 5–10 transistors.
 5. nMOS devices should be placed close to *V_{ss}* and pMOS devices close to *V_{dd}*.

Latch-Up Problem and Its Prevention

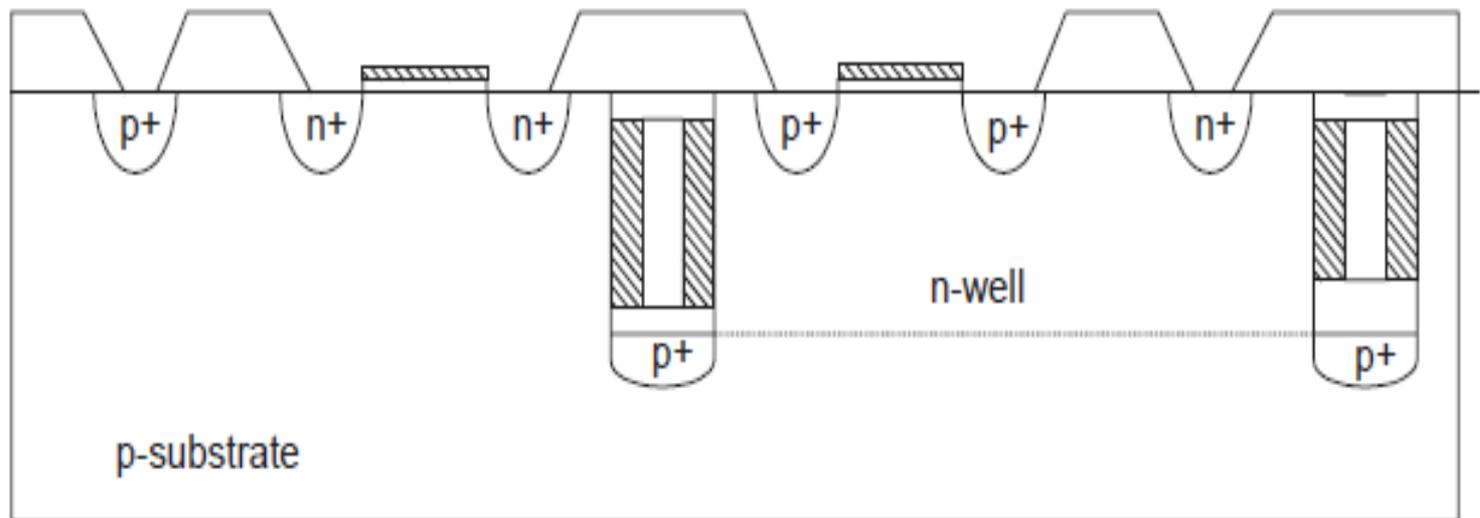
- **Guard Rings and Use of Trenches**
- The gain of the parasitic transistors can be reduced by using guard rings and making additional contacts to the ring as shown in Figure.



- This reduces parasitic resistance values and the contacts drain excess well or substrate leakage currents away from the active device such that trigger current which initiates latch-up is not attained.

Latch-Up Problem and Its Prevention

- **Guard Rings and Use of Trenches**
- Another approach to overcome the latch-up problem is to use trenches between the individual transistor devices of the CMOS structure, and highly doped field regions are formed in the bottom of the trenches.
- Each n- and p-well includes a retrograde impurity concentration profile and extends beneath adjacent trenches as shown in Figure.



Short-Channel Effects

- The channel length L is usually reduced to increase both the speed of operation and the number of components per chip.
- However, when the channel length is the same order of magnitude as the depletion-layer widths (x_{dD} , x_{dS}) of the source and drain junction, a metal–oxide–semiconductor field-effect transistor (MOSFET) behaves differently from other MOSFETs.
- This is known as short-channel effect (SCE).
- The SCEs are attributed to two physical phenomena:
 1. The limitation imposed on electron drift characteristics in the channel
 2. The modification of the threshold voltage due to the shortening of channel length
- Some of the important SCEs are mentioned below.

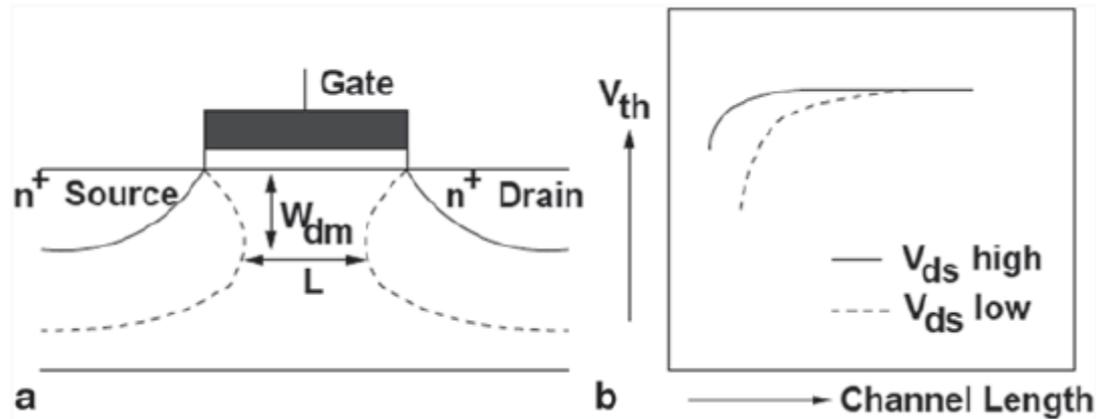
Channel Length Modulation Effect

Drain-Induced Barrier Lowering and Channel Punch Through

Short-Channel Effects

Channel Length Modulation Effect

- As the channel length is reduced, the threshold voltage of MOSFET decreases as shown in Figure.



Threshold voltage roll-off with channel length

- This reduction of channel length is known as *V_{th} roll-off*.
- The graph in Figure (b) shows the reduction of threshold voltage with reduction in channel length.
- This effect is caused by the proximity of the source and drain regions leading to a 2D field pattern rather than a 1D field pattern in short-channel devices as shown in Figure(a).

Emerging Technologies for Low Power

- Over the past two decades, industries have closely followed Moore's law by fabricating transistors with gate dielectric scaling using silicon dioxide (SiO₂).
- But, as transistor size shrinks, leakage current increases drastically.
- Managing that leakage is crucial for reliable high-speed operation.
- As a consequence, this is becoming an increasingly important factor in chip design.

Hi-K Gate Dielectric
Lightly Dopes Drain-Source
Silicon on Insulator(SOI)
FinFET

Emerging Technologies for Low Power

- High-K (Hi-K) materials are proposed to reduce the gate leakage current, a metal gate is used to suppress the poly-silicon gate depletion, and SOI technologies with single or multiple gate transistors offer opportunities for further scaling down of the transistor dimensions.
- Many other alternatives such as dual-gated SOI and substrate biasing have recently been proposed to address the conflicting requirement of high performance during active mode of operation and low leakage during sleep mode of operation.

Hi-K Gate Dielectric

Lightly Dopes Drain-Source

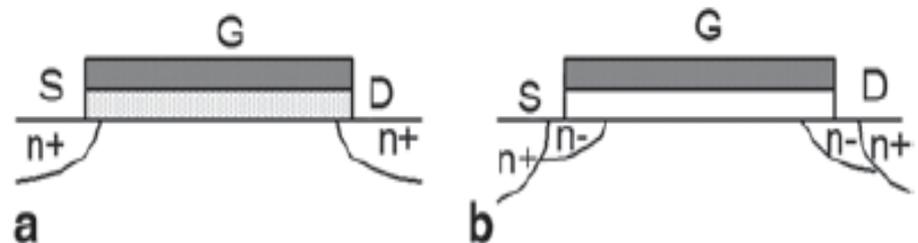
Silicon on Insulator(SOI)

FinFET

Emerging Technologies for Low Power

• *Lightly Doped Drain–Source*

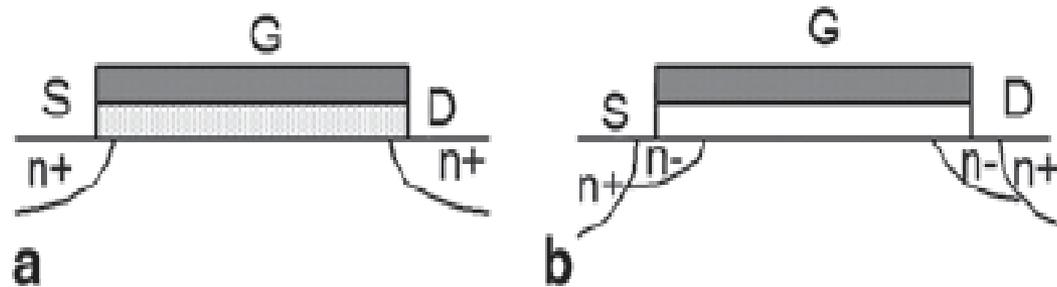
- A pattern of heavily doped n^+ regions is formed in the substrate adjacent to the dielectric spacer structure on the sidewalls of the structures and over the adjacent portions of the substrate which form LDD structures of an MOSFET device to form the said integrated circuit device as shown in Figure
- The n^+ regions provide smaller ohmic contacts required to avoid punch through. In the p-channel regions, the n-type LDD extensions are counter doped by the regular p^+ source/drain implant.



Emerging Technologies for Low Power

- ***Lightly Doped Drain–Source***

- This results in significant improvements in breakdown voltages, hot-electron effects, and short-channel threshold effects.
- A pattern of gate electrode structures is formed upon a semiconductor substrate whose structures each include a gate oxide and a poly-silicon layer as shown in Figure.

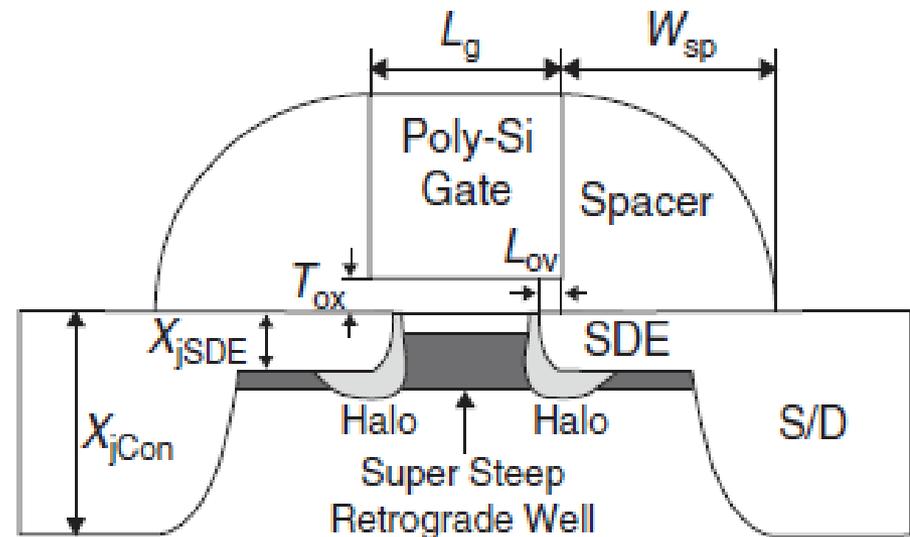


Emerging Technologies for Low Power

- ***Lightly Doped Drain–Source***

- The below figure shows a device with various channel-doping implants (source/drain extension, SDE; Gaussian halo; and vertical retrograde well) which have been developed to mitigate the SCEs and to improve the leakage characteristics.

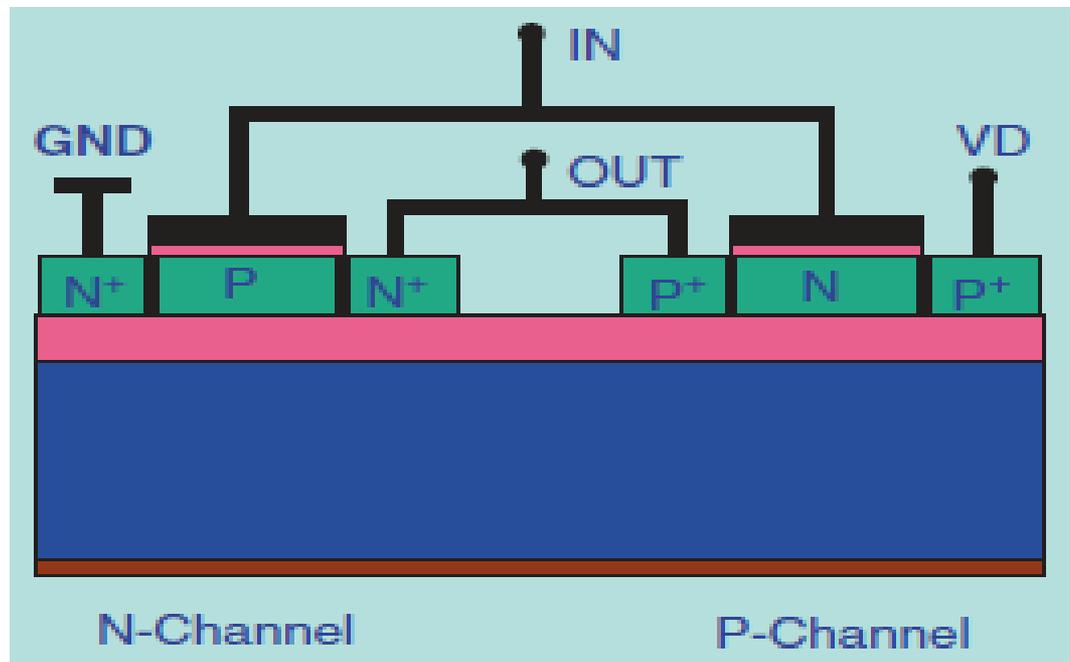
MOS transistor structure to overcome short channel effects



Emerging Technologies for Low Power

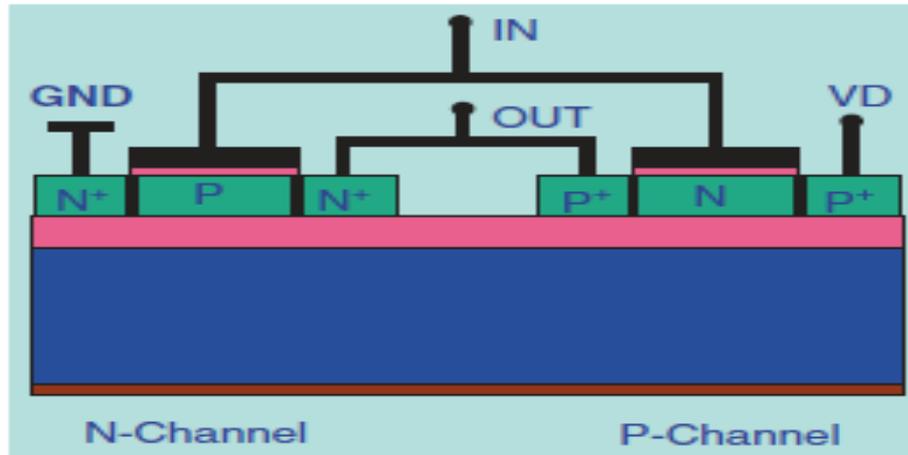
- ***Silicon on Insulator***

- Rather than using silicon as the substrate, technologies such as SOI have been developed that use an insulating substrate to improve process characteristics such as latch-up and speed.
- The below Figure shows a CMOS inverter fabricated using the SOI approach.



Emerging Technologies for Low Power

- *Silicon on Insulator(SOI)*



The steps used in a typical SOI CMOS process are as follows:

- A thin film (7–8 μm) of very lightly doped n-type Si is epitaxially grown over an insulator. Sapphire or SiO_2 is a commonly used insulator.
- An anisotropic etch is used to etch away the Si except where a diffusion area will be needed.
- Implantation of the p-island where an n-transistor is formed.
- Implantation of the n-island where a p-transistor is formed.
- Growing of a thin gate oxide (100–250 \AA).
- Depositing of phosphorus-doped poly-silicon film over the oxide.

Emerging Technologies for Low Power

- *Silicon on Insulator(SOI)*

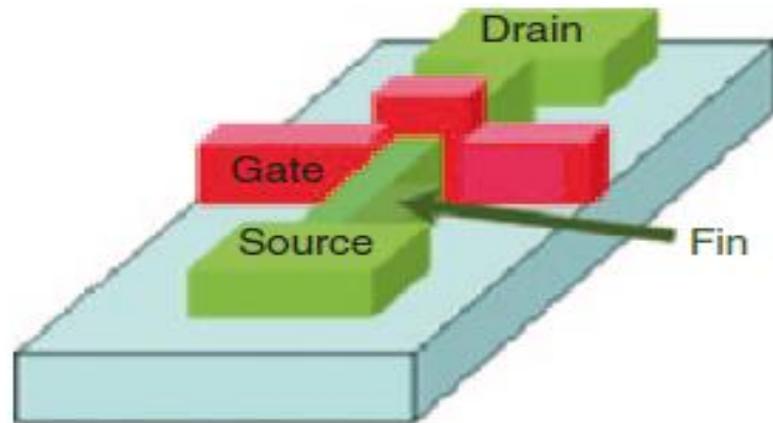
Advantages of SOI

- Due to the absence of wells, transistor structures denser than bulk silicon are feasible.
- Lower substrate capacitance.
- No field-inversion problems (the existence of a parasitic transistor between two normal transistors).
- No latch-up is possible because of the isolation of transistors by insulating substrate.

Emerging Technologies for Low Power

- *FinFET*

Simple FinFET structure



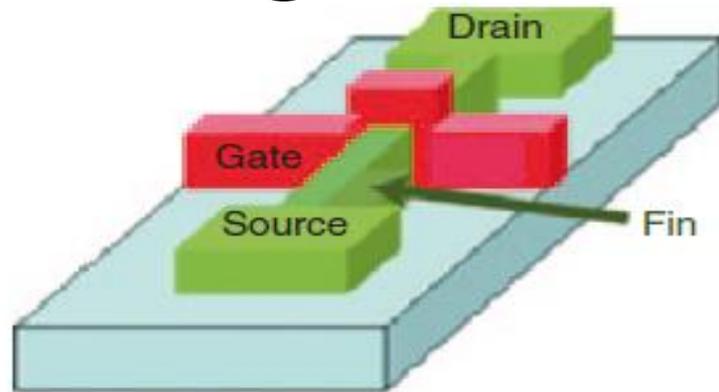
❖ The finFET is a transistor realization, first developed by Chenming Hu and his colleagues at the University of California at Berkeley, which attempts to overcome the worst types of SCE encountered by deep submicron transistors, such as DIBL.

❖ These effects make it difficult for the voltage on the gate electrode to deplete the channel underneath and stop the flow of carriers through the channel; in other words, to turn the transistor off.

Emerging Technologies for Low Power

- *FinFET*

Simple FinFET structure



- ❑ By raising the channel above the surface of the wafer instead of creating the channel just below the surface, it is possible to wrap the gate around up to three of its sides, providing much greater electrostatic control over the carriers within it.
- ❑ This led to the development of FinFET structure as shown in Figure.
- ❑ In current usage, the term FinFET has a less precise definition. Among microprocessor manufacturers, AMD, IBM, and Motorola describe their doublegate development efforts as FinFET development, whereas Intel avoids using the term to describe their closely related tri-gate architecture.

Unit-1:MOS Transistors

- Introduction
- The Structure of MOS Transistors
- The Fluid Model
 - The MOS Capacitor*
 - The MOS Transistor*
- Modes of Operation of MOS Transistors
- Electrical Characteristics of MOS Transistors
 - Threshold Voltage*
 - Transistor Transconductance g_m*
 - Figure of Merit*
 - Body Effect*
 - Channel-Length Modulation*
- MOS Transistors as a Switch
 - Transmission Gate*

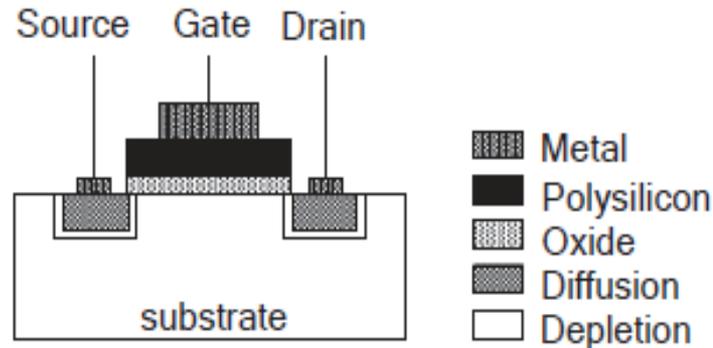
Introduction

- The base semiconductor material used for the fabrication of metal–oxide–semiconductor (MOS) integrated circuits is **silicon**.
- ***Metal, oxide, and semiconductor*** form the basic structure of MOS transistors.
- MOS transistors are realized on a single crystal of silicon by creating three types of conducting materials separated by intervening layers of an insulating material to form a sandwich-like structure.
- The three conducting materials are: ***metal, poly-silicon, and diffusion***.

Introduction

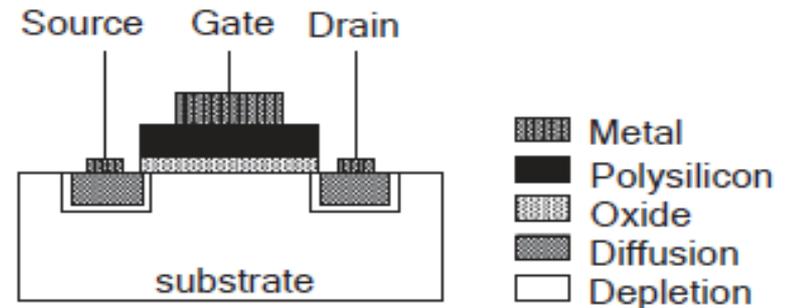
- *Aluminum* as metal and **polycrystalline silicon or polysilicon** are used for interconnecting different elements of a circuit.
- The insulating layer is made up of **silicon dioxide (SiO₂)**.
- Patterned layers of the conducting materials are created by a series of **photolithographic techniques 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.**

The Structure of MOS Transistors



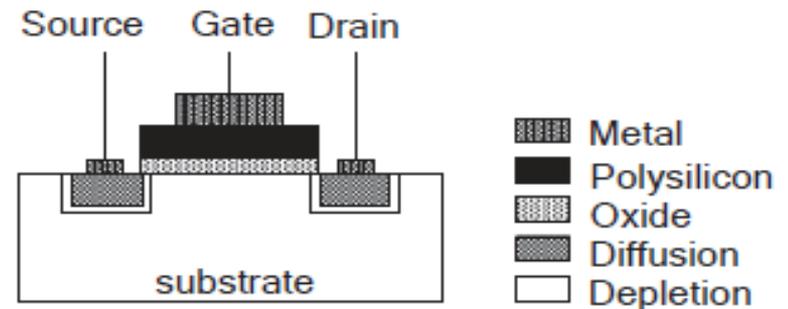
- The structure of an MOS transistor is shown in Figure.
- On a lightly doped substrate of silicon, two islands of diffusion regions of opposite polarity of that of the substrate are created.
- These two regions are called *source and drain*, which are connected via metal (or poly-silicon) to the other 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*.

The Structure of MOS Transistors



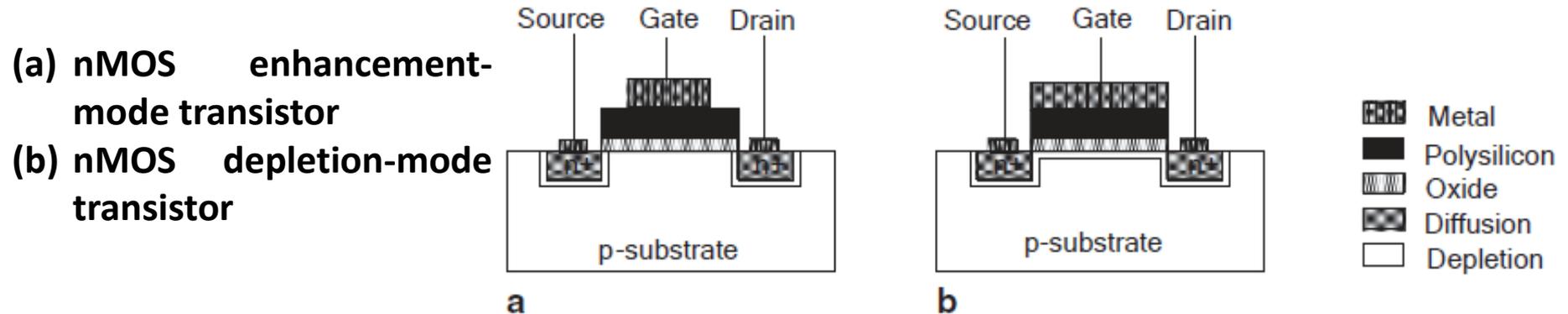
- *There* are two possible alternatives.
- The substrate can be lightly doped by either a p-type or an n-type material, leading to two different types of transistors.
- When the substrate is lightly doped by a p-type material, the two diffusion regions are strongly doped by an n-type material.
- In this 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.

The Structure of MOS Transistors



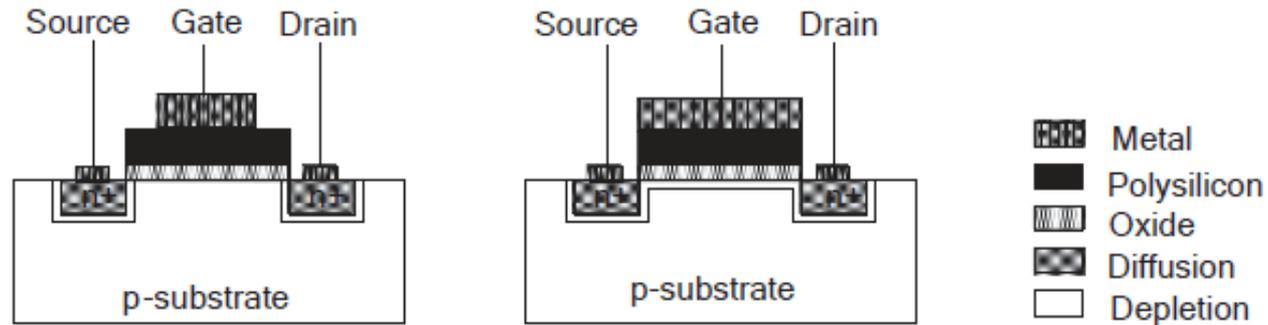
- The region between the two diffusion islands under the oxide layer is called the *channel region*.
- *The operation of an MOS transistor is based on the controlled flow of current between the source and drain through the channel region.*
- In order to make a useful device, there must be suitable means 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*.

The Structure of MOS Transistors



- After fabrication, the structure of an enhancement-mode nMOS transistor looks like Figure(a). 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.

The Structure of MOS Transistors



(a) nMOS enhancement-mode transistor (b) nMOS depletion-mode 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 Figure(b).*
- Here, the source and drain are normally connected by a conducting path, which can be removed by applying a suitable negative voltage to the gate.
- This is known as the *depletion mode of operation*

The Fluid Model

- The operation of an MOS transistor can be analyzed by using a suitable **analytical technique**, which will give **mathematical expressions** for **different device characteristics**.
- This, however, requires an in-depth knowledge of the physics of the device.
- Sometimes, it is possible to develop an intuitive understanding about the operation of a system by visualizing the physical behavior with the help of a simple but very effective model.

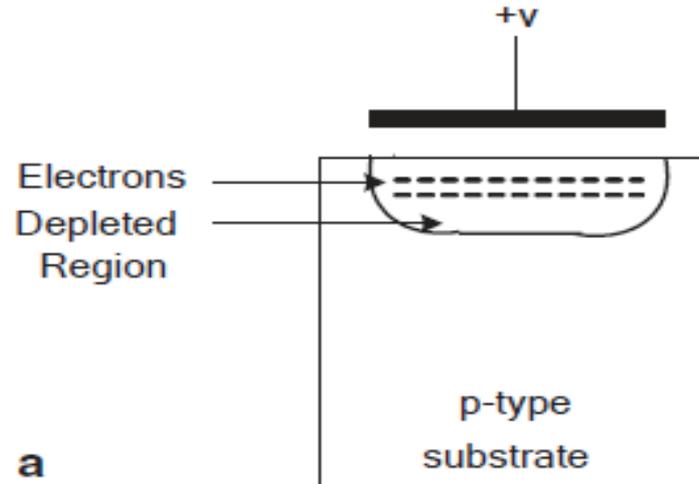
The Fluid Model

- 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).
- Using this model, even a novice can understand the operation of these devices.

The Fluid Model

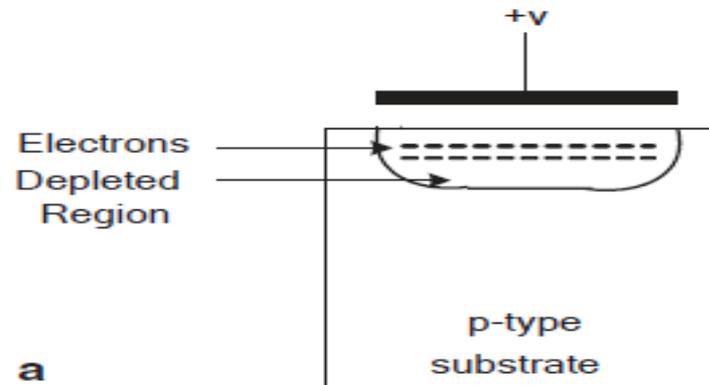
- The model is based on two simple ideas:
- (a) Electrical charge is considered as fluid, which can move from one place to another depending on the difference in their level, of one from the other, just like a fluid.
- (b) Electrical potentials can be mapped into the geometry of a container, in which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

The Fluid Model-The MOS Capacitor



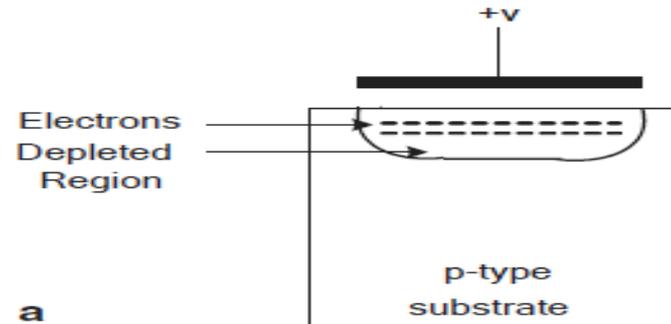
- From the knowledge of basic physics, we know that a simple parallel-plate capacitor can be formed with the help of two identical metal plates separated by an insulator.
- An MOS capacitor is realized by sandwiching a thin oxide layer between a metal or poly-silicon plate on a silicon substrate of suitable type as shown in Figure.

The Fluid Model-The MOS Capacitor



- 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 polysilicon 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.

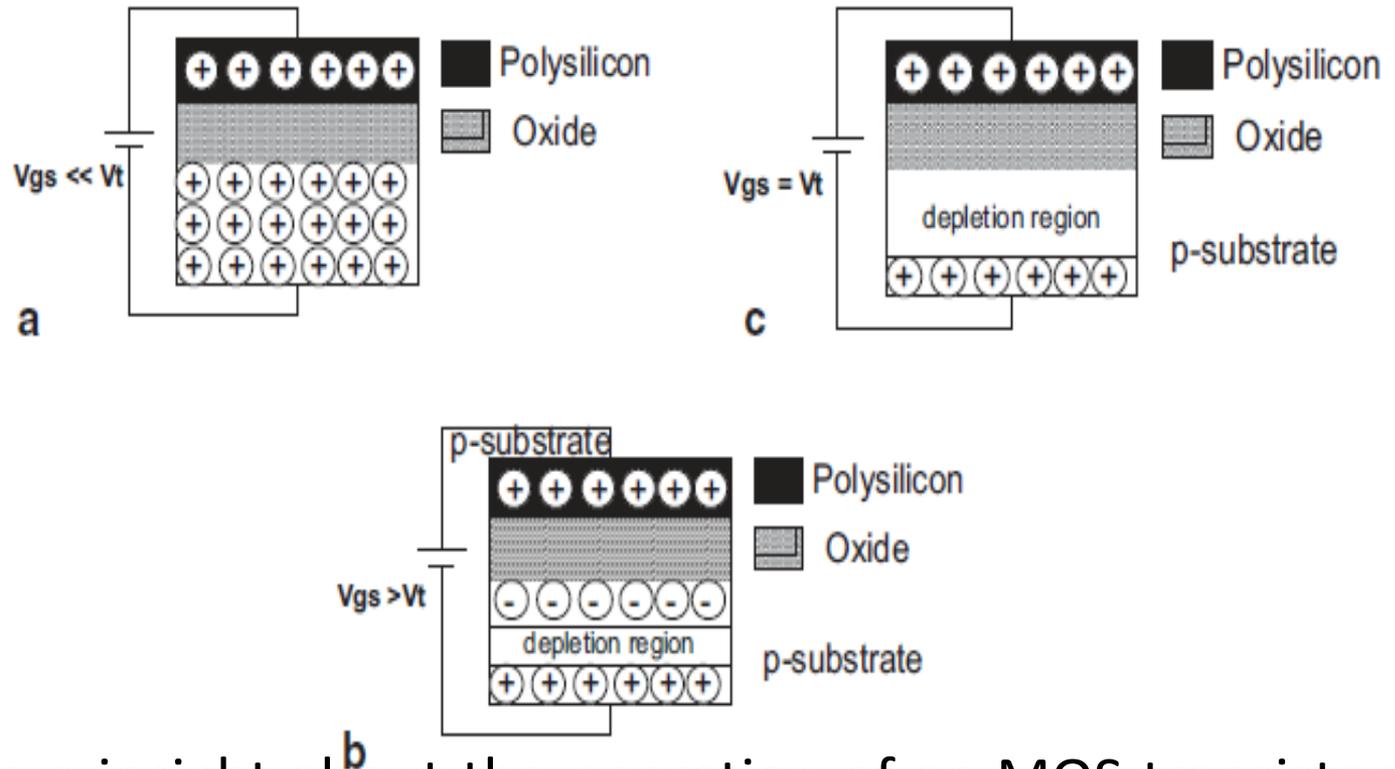
The Fluid Model-The MOS Capacitor



- 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.

Modes of Operation of MOS Transistors

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



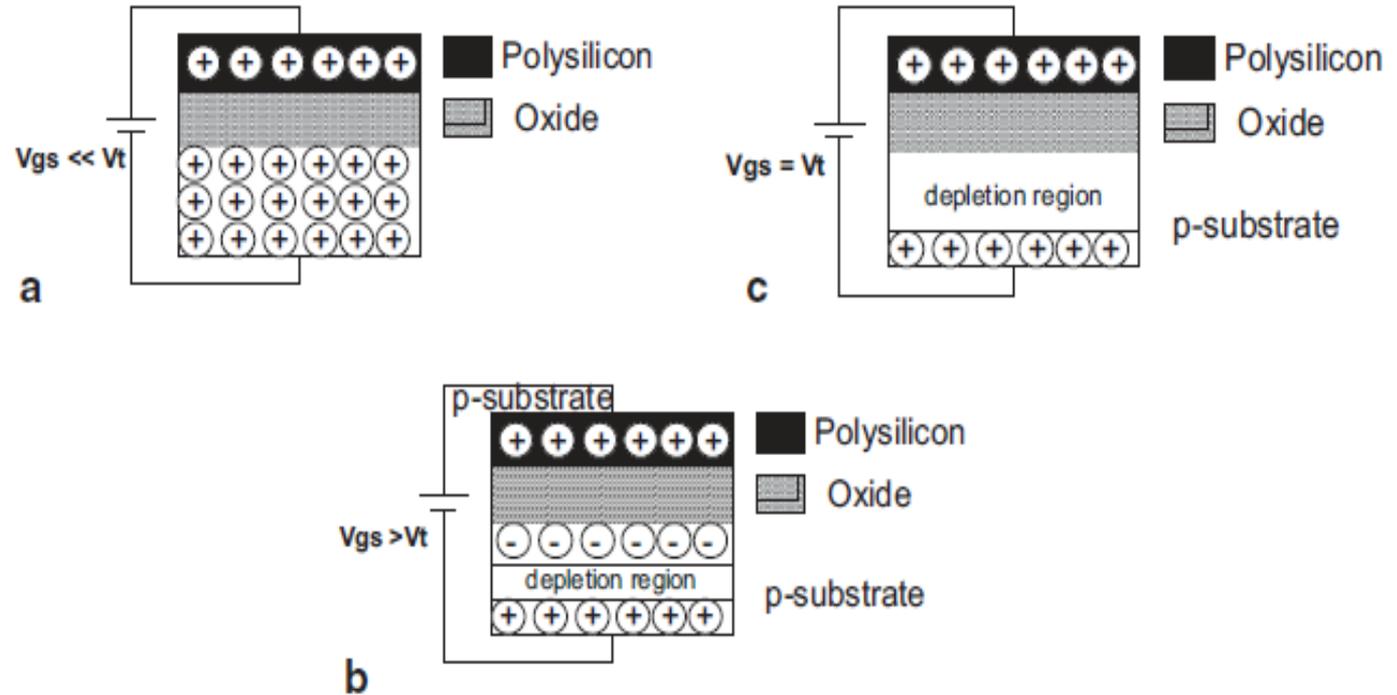
•After having some insight about the operation of an MOS transistor, let us now have a look at the charge distribution under the gate region under different operating conditions of the transistor.

•When the gate voltage is very small and much less than the threshold voltage, Fig. (a) 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*.

Modes of Operation of MOS Transistors

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



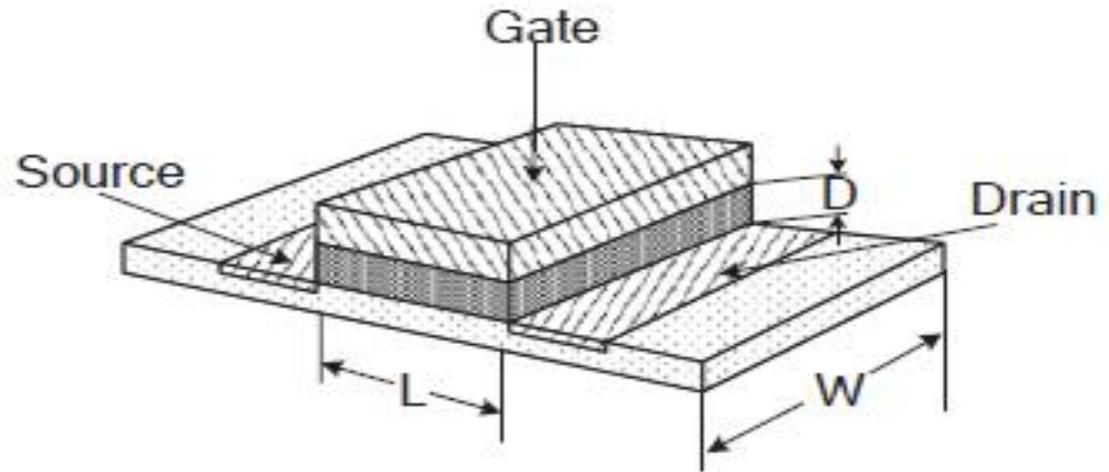
- As the gate voltage is increased, the holes are repelled from the SiO₂–substrate interface and a depletion region is created under the gate when the gate voltage is equal to the threshold voltage.
- In this condition, the device is said to be in *depletion mode* as shown in Fig. (b).
- As the gate 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.(c).
- The transistor is now said to be *in inversion mode*.

Electrical Characteristics of MOS Transistors

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

Electrical Characteristics of MOS Transistors

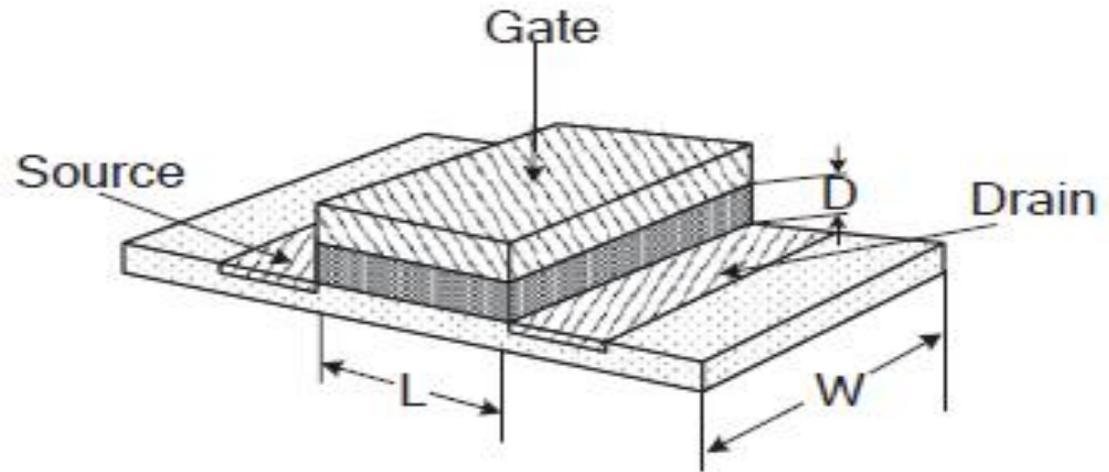
Structural view of an MOS transistor



- 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 above the Figure,
- The above figure shows the three important parameters of MOS transistors, the channel length L , the channel width W , and the dielectric thickness D .

Electrical Characteristics of MOS Transistors

Structural view of an MOS transistor



- The expression for the drain current is given by

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

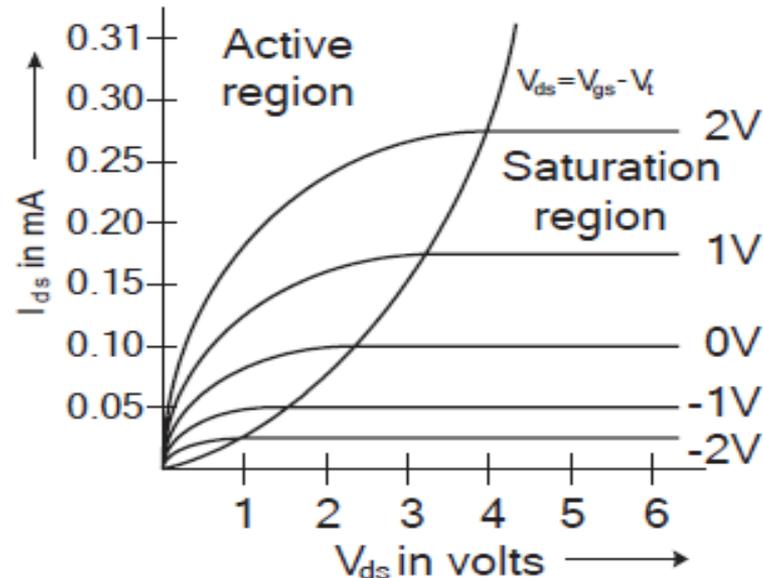
- Let us separately find out the expressions for Q_c and t_n .

Electrical Characteristics of MOS Transistors

- In the depletion-type nMOS transistor, a channel is created by implanting suitable impurities in the region between the source and drain during fabrication prior to depositing the gate insulation layer and the poly-silicon layer.
- As a result, channel exists even when the gate voltage is 0 V. Here, the channel current can also be controlled by the gate voltage.

Voltage–current characteristics of nMOS Depletion type transistor

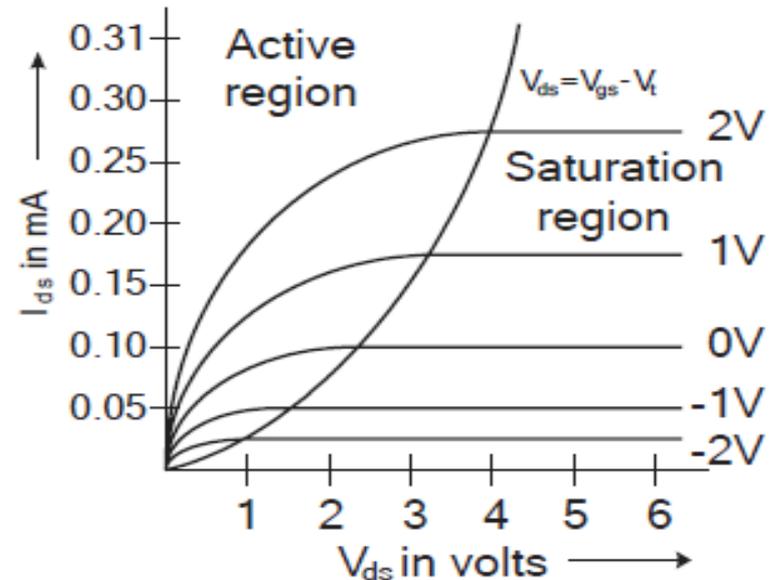
- A positive gate voltage increases the channel width resulting in an increase of drain current.
- A negative gate voltage decreases the channel width leading to a reduced drain current.



Electrical Characteristics of MOS Transistors

- A suitable negative gate voltage fully depletes the channel isolating the source and drain regions.
- The characteristic curve, as shown in Figure , is similar except the threshold voltage, which is a negative voltage in case of a depletion-mode nMOS transistor.
- In a similar manner, the expression for drain current can be derived and voltage–current characteristics can be drawn for pMOS enhancement-mode and pMOS depletion mode transistors.

Voltage–current characteristics of nMOS Depletion type transistor



Electrical Characteristics of MOS Transistors

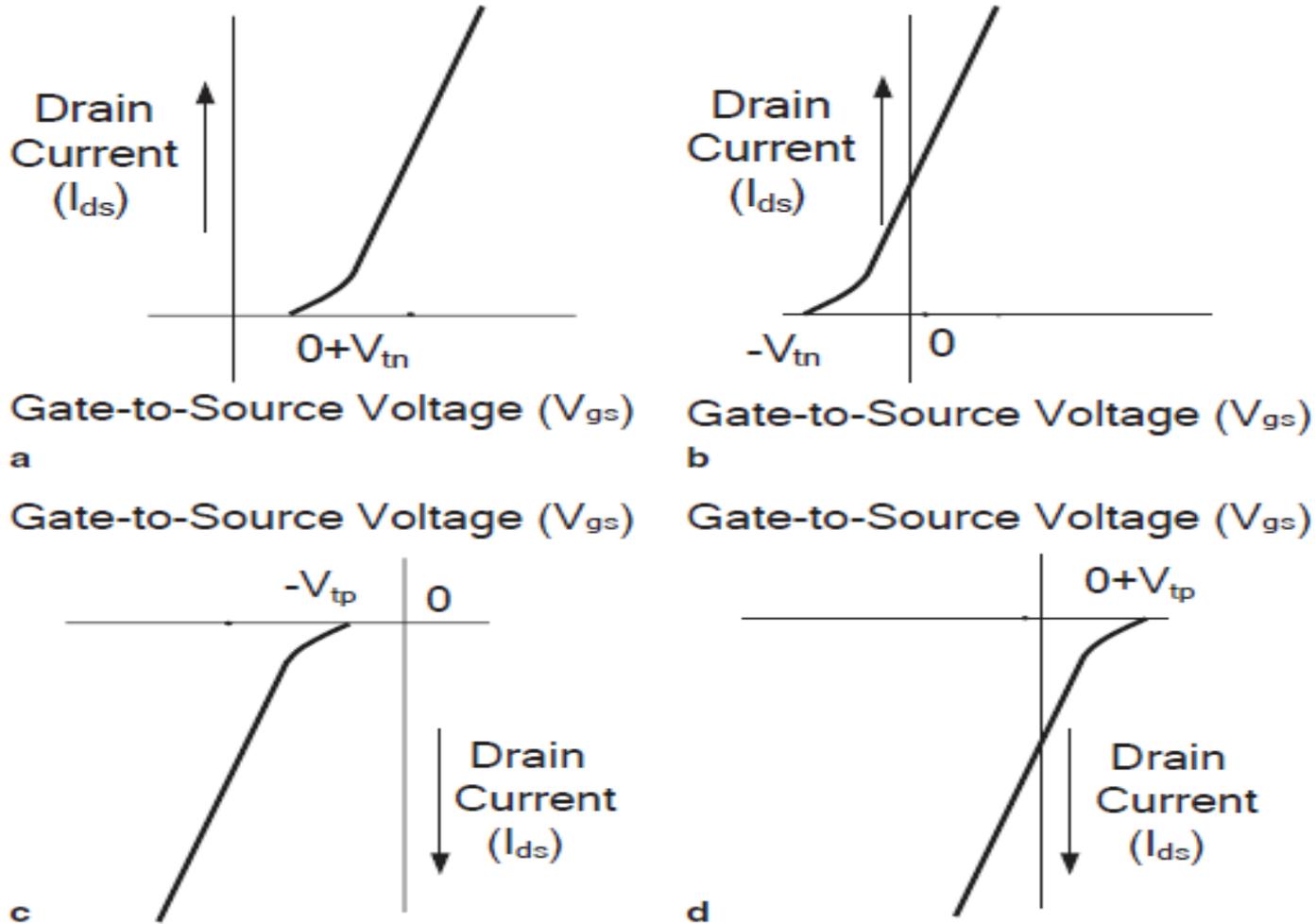
- *Threshold Voltage*
- *Transistor Transconductance g_m*
- *Figure of Merit*
- *Body Effect*
- *Channel-Length Modulation*

Electrical Characteristics of MOS Transistors

Threshold Voltage

- One of the parameters that characterize the switching behavior of an MOS transistor is its threshold voltage V_t .
- *As we know, this can be defined as the gate voltage at which an MOS transistor begins to conduct.*
- Typical value for threshold voltage for an nMOS enhancement-type transistor is $0.2 V_{dd}$, *i.e., for a supply voltage of 5 V, $V_{tn} = 1.0$ V.*
- *As we have seen, the drain current depends on both the gate voltage and the drain voltage with respect to the source.*

Electrical Characteristics of MOS Transistors-*Threshold Voltage*



For a fixed drain-to-source voltage, the variation of conduction of the channel region (represented by the drain current) for different gate voltages is shown in Fig. 3.11 for four different cases: nMOS depletion, nMOS enhancement, pMOS enhancement, and pMOS depletion transistors, as shown in Fig. 3.12a–d, respectively.

Electrical Characteristics of MOS *Transistor*

Transconductance (gm)

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

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

This can be derived from

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

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

Electrical Characteristics of MOS Transistor

Transconductance (g_m)

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

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

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

$$\text{or } g_m = \frac{\delta I_{ds}}{\delta V_{gs}} = \frac{C_g \mu_n V_{ds}}{L^2},$$

in saturation $V_{ds} = (V_{gs} - V_t),$

and substituting $C_g = \frac{\epsilon_{ins} \epsilon_0 WL}{D}.$

We get
$$g_m = \frac{\mu_n \epsilon_{ins} \epsilon_0}{D} \frac{W}{L} (V_{gs} - V_t).$$

Electrical Characteristics of MOS Transistor

Figure of Merit

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

$$W_0 = \frac{g_m}{C_g} = \frac{\mu_n}{L^2} (V_{gs} - V_t)$$
$$= \frac{1}{t_{sd}}$$

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

Electrical Characteristics of MOS *Transistor*

Body Effect

- ❑ All MOS transistors are usually fabricated on a common substrate and substrate (body) voltage of all devices is normally constant.
- ❑ When circuits are realized using a number of MOS devices, several devices are connected in series.
- ❑ This results in different source potentials for different devices.

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

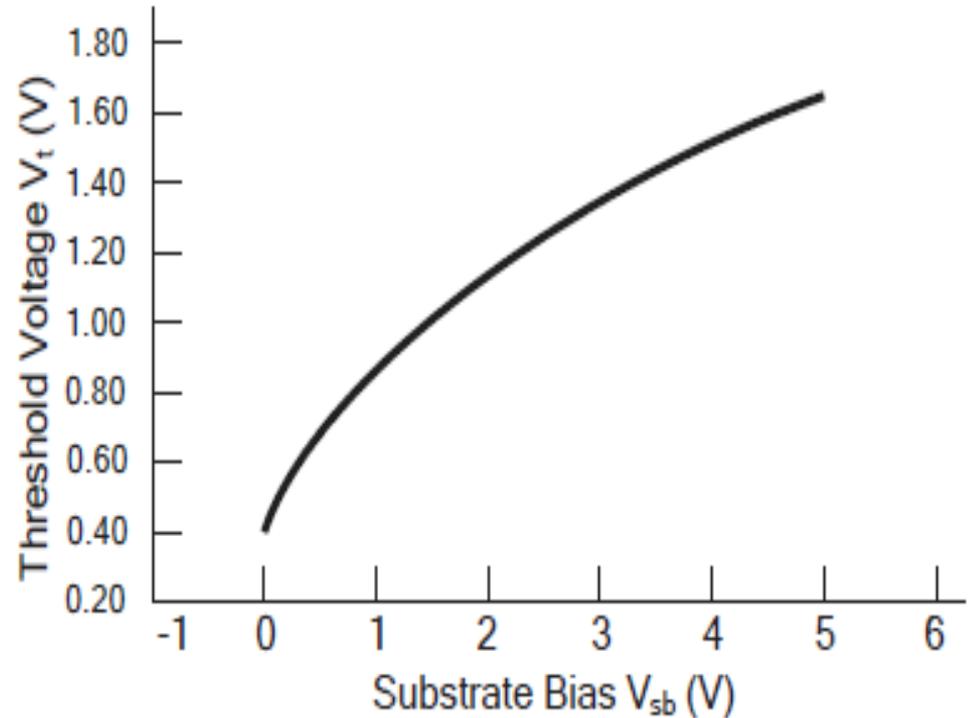
- ❖ It may be noted from above equation that the threshold voltage V_t is *not* constant with respect to the voltage difference between the substrate and the source of the MOS transistor.
- ❖ This is known as the substrate-bias effect or *body effect*.
- ❖ *Increasing the V_{sb} causes the channel to be depleted of charge carriers, and this leads to an increase in the threshold voltage.*

Electrical Characteristics of MOS *Transistor*

Body Effect

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

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



Electrical Characteristics of MOS Transistor

Channel Length Modulation

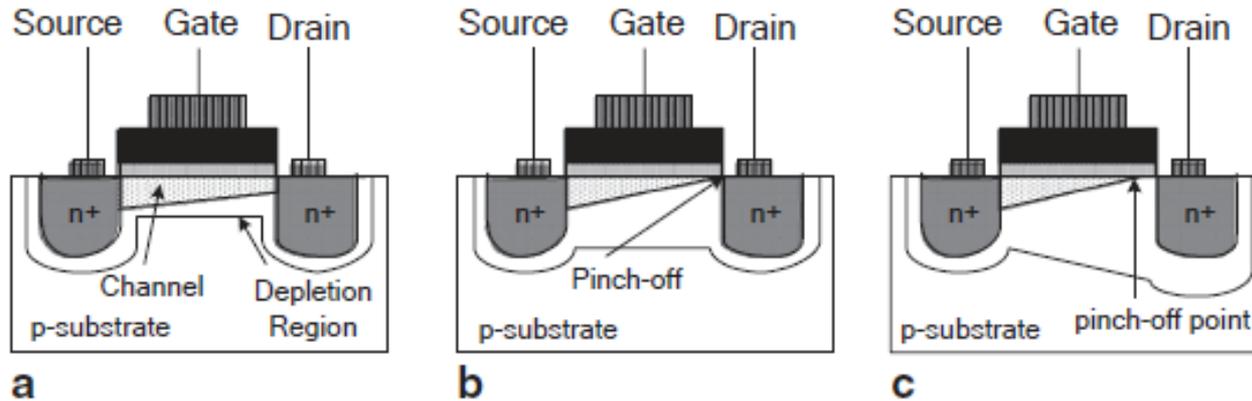


Fig. 3.14 a Nonsaturated region. b Onset of saturation. c Deep in saturation

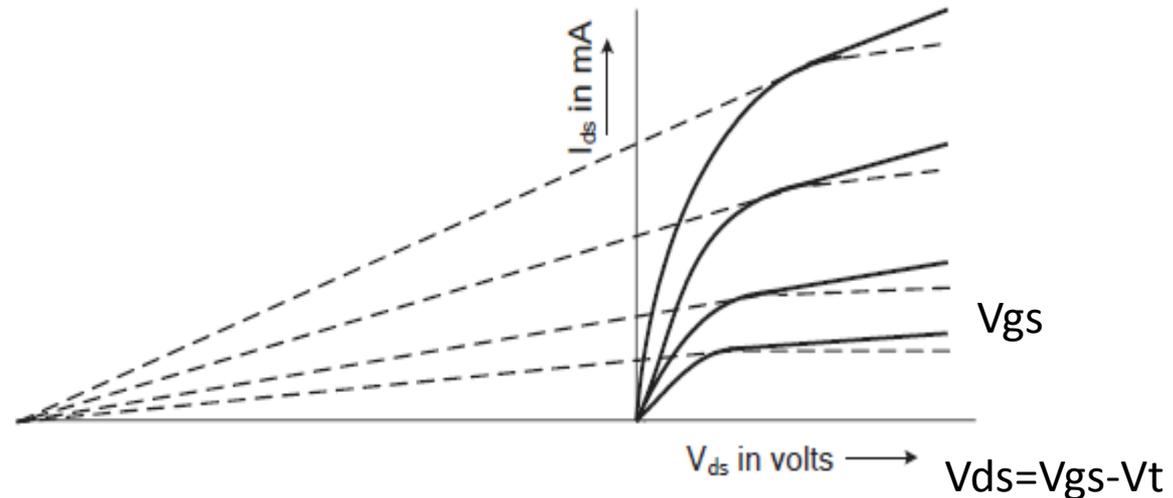


Fig. 3.15 Drain-current variations due to channel-length modulation

Electrical Characteristics of MOS Transistor

Channel Length Modulation

This effective channel length L_{eff} can be represented by

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

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

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

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

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

The channel-length modulation coefficient λ has the value in the range of 0.02–0.005 per volt.

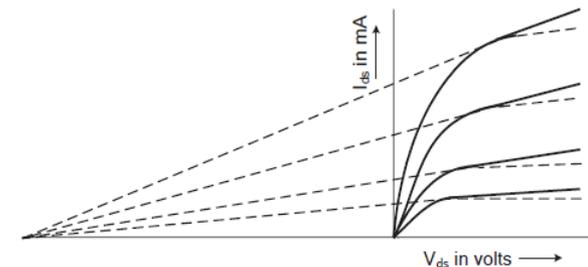


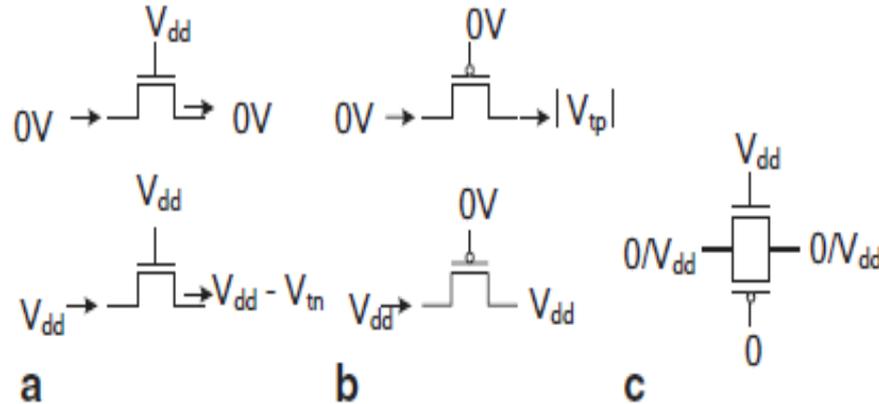
Fig. 3.15 Drain-current variations due to channel-length modulation

MOS Transistors as a Switch

a nMOS pass transistor.

b pMOS pass transistor.

c Transmission gate



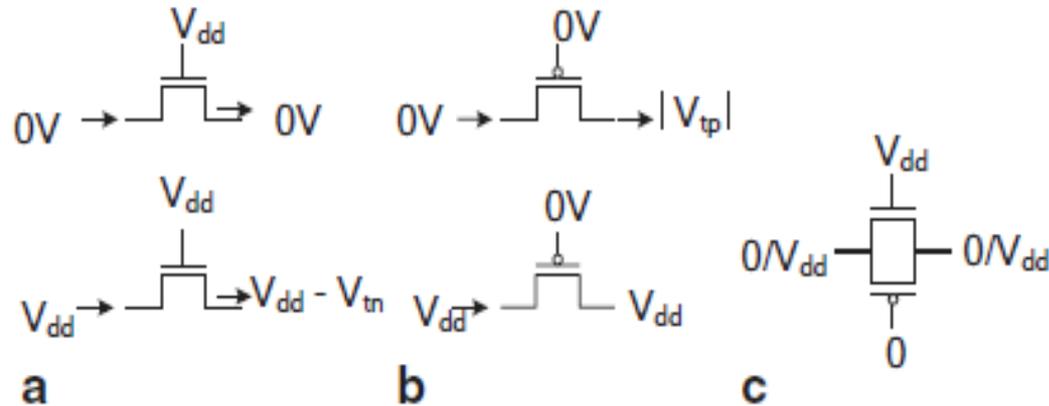
- We have seen that in the linear region (when the drain-to-source voltage is small) an MOS transistor acts as a variable resistance, which can be controlled by the gate voltage.
- An nMOS transistor can be switched from very high resistance when the gate voltage is less than the threshold voltage, to low resistance when V_{gs} exceeds the threshold voltage V_{tn} .
- *This has opened up the possibility of using an MOS transistor as a switch, just like a relay.*

MOS Transistors as a Switch

a nMOS pass transistor.

b pMOS pass transistor.

c Transmission gate

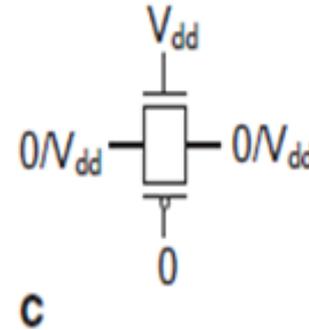


❑ For example, an nMOS transistor when used as a switch is **OFF** when $V_{gs} = 0\text{ V}$ and **ON** when $V_{gs} = V_{dd}$. However, its behavior as a switch is not ideal.

❑ When $V_{gs} = V_{dd}$, the switch turns on but the on resistance is not zero.

❑ As a result, there is some voltage drop across the switch, which can be neglected when it is in series with a large resistance.

MOS Transistors as a Switch

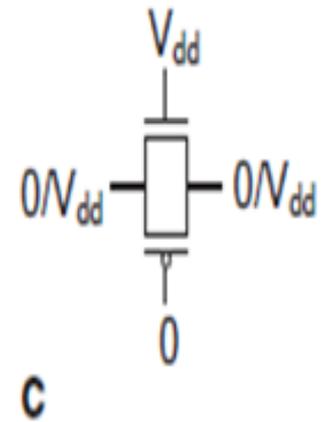


- ***Transmission Gate***

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

- ✓ The characteristics of a transmission gate, which is realized by using **one nMOS and one pMOS pass transistors** connected in parallel, can be constructed by combining the characteristics of both the devices.
- ✓ It may be noted that the operation of a transmission gate requires a dual-rail (both true and its complement) control signal.

MOS Transistors as a Switch



- ***Transmission Gate***

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

- Both the devices are **off** when “0” and “1” logic levels are applied to the gates of the nMOS and pMOS transistors, respectively.
- In this situation, no signal passes through the gate.
- Therefore, the output is in the high-impedance state, and the intrinsic load capacitance associated to the output node retains the high or low voltage levels, whatever it was having at the time of turning off the transistors.

MOS Transistors as a Switch

➤ To understand the operation of a transmission gate, let us consider two situations.

➤ In the first case, the transmission gate is connected to a relatively **large capacitive load**, and the output changes the state from low to high or high to low as shown in Fig. 3.17.

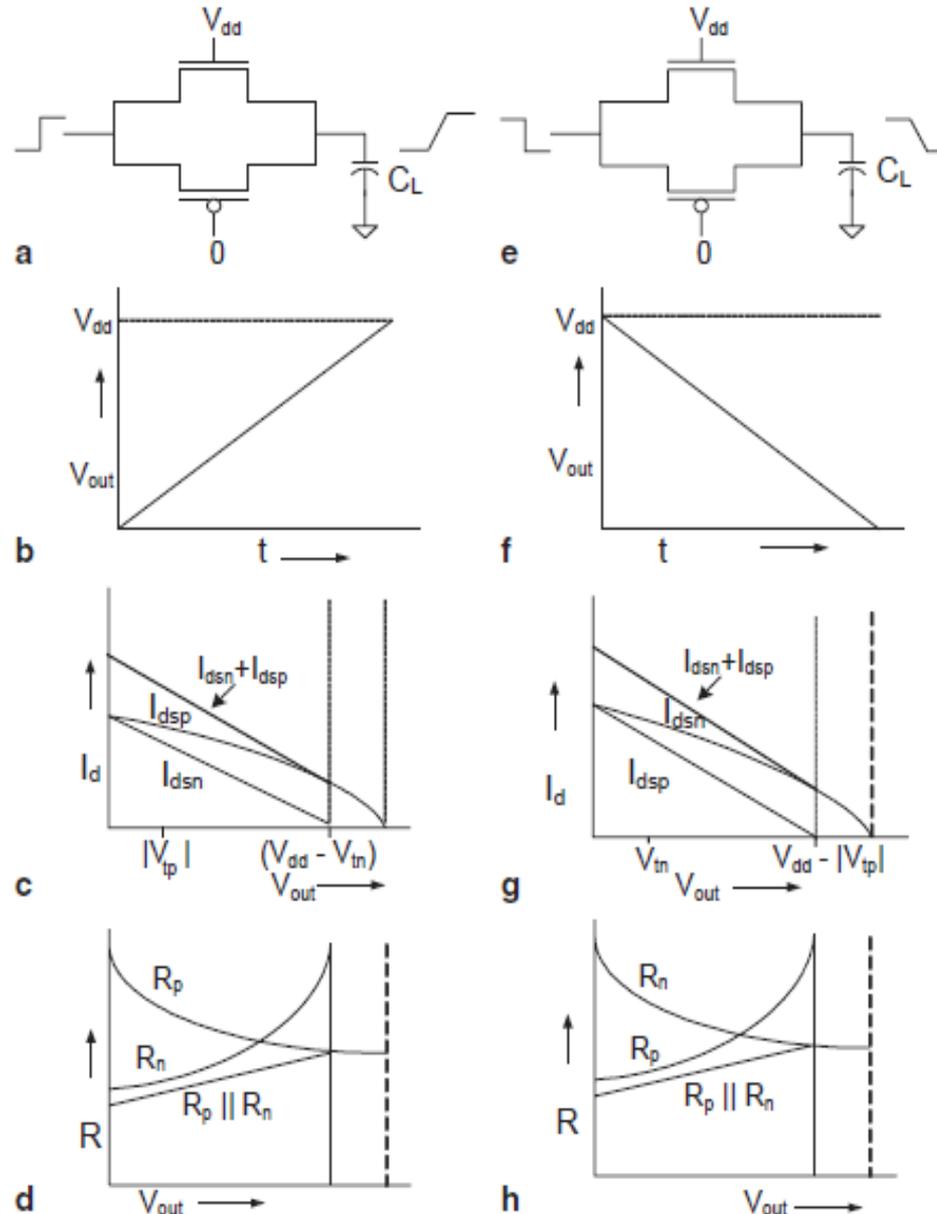


Fig. 3.17
a and e Output node charges from low-to-high level or high-to-low level.

b and f The output voltage changing with time for different transitions.

c and g The drain currents through the two transistors as a function of the output voltage.

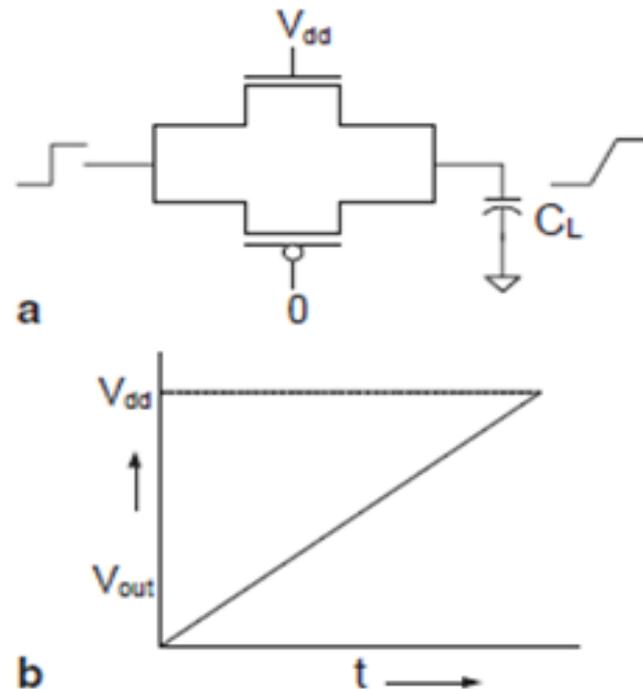
d and h The equivalent resistances as a function of the output voltage

MOS Transistors as a *Switch-Transmission Gate*

- **Case I: Large Capacitive Load**

❖ First, consider the case when the input has changed quickly to V_{dd} from 0 V and the output of the switch changes slowly from 0 V (V_{ss}) to V_{dd} to charge a load capacitance C_L .

❖ This can be modeled by using V_{dd} as an input and a ramp voltage generated at the output as the capacitor charges from V_{ss} to V_{dd} .



MOS Transistors as a Switch-Transmission Gate

- Region I: Here,

$$V_{dsn} = V_{dd} - V_{out}$$

$$V_{gsn} = V_{dd} - V_{out}$$

$$V_{dsp} = V_{out} - V_{dd}$$

$$V_{gsp} = -V_{dd}$$

- The current contributing to charge the load capacitor by the two transistors is

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

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

- for the nMOS and pMOS transistors, respectively.

MOS Transistors as a *Switch-Transmission Gate*

- Now, the equivalent resistances for the two transistors are

$$R_{\text{eqn}} = \frac{V_{\text{dd}} - V_{\text{out}}}{I_{\text{dsn}}} = \frac{2L_{\text{n}}}{K_{\text{n}}W_{\text{n}}} \cdot \frac{(V_{\text{dd}} - V_{\text{out}})}{(V_{\text{dd}} - V_{\text{out}} - V_{\text{tn}})^2}$$

- and

$$R_{\text{eqp}} = \frac{V_{\text{dd}} - V_{\text{out}}}{I_{\text{sdp}}} = \frac{2L_{\text{p}}}{K_{\text{p}}W_{\text{p}}} \cdot \frac{(V_{\text{dd}} - V_{\text{out}})}{(V_{\text{dd}} - |V_{\text{tp}}|)^2}$$

MOS Transistors as a *Switch-Transmission Gate*

- **Region II:** In this region, the nMOS transistor remains in saturation region, whereas the pMOS transistor operates in the linear region. Therefore, in this case

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

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

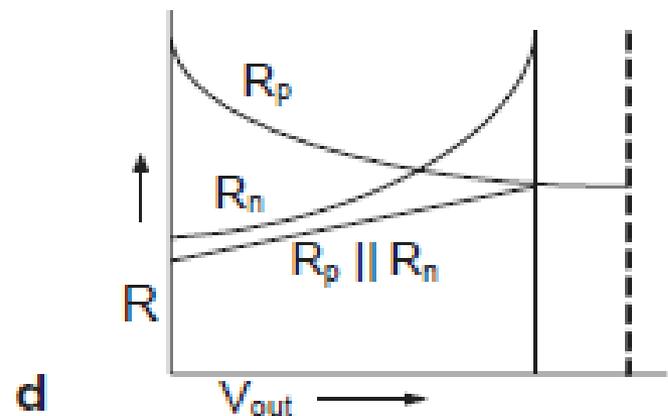
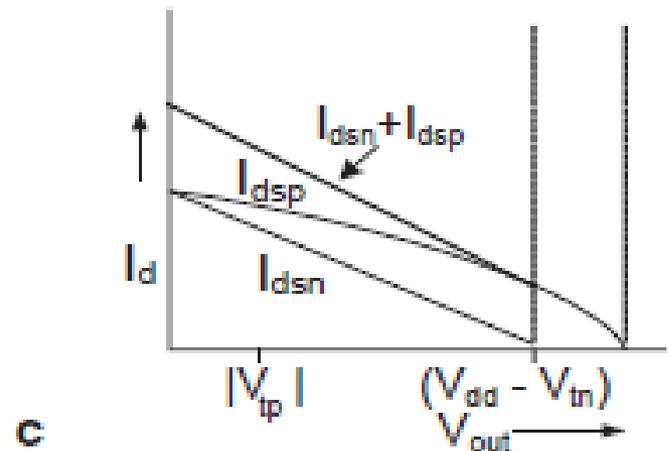
MOS Transistors as a Switch-Transmission Gate

- **Region III:** In this region, the nMOS transistor turns off and pMOS transistor continues to operate in the linear region.

➤ These individual nMOS and pMOS currents and the combined current are shown in Fig. 3.17c.

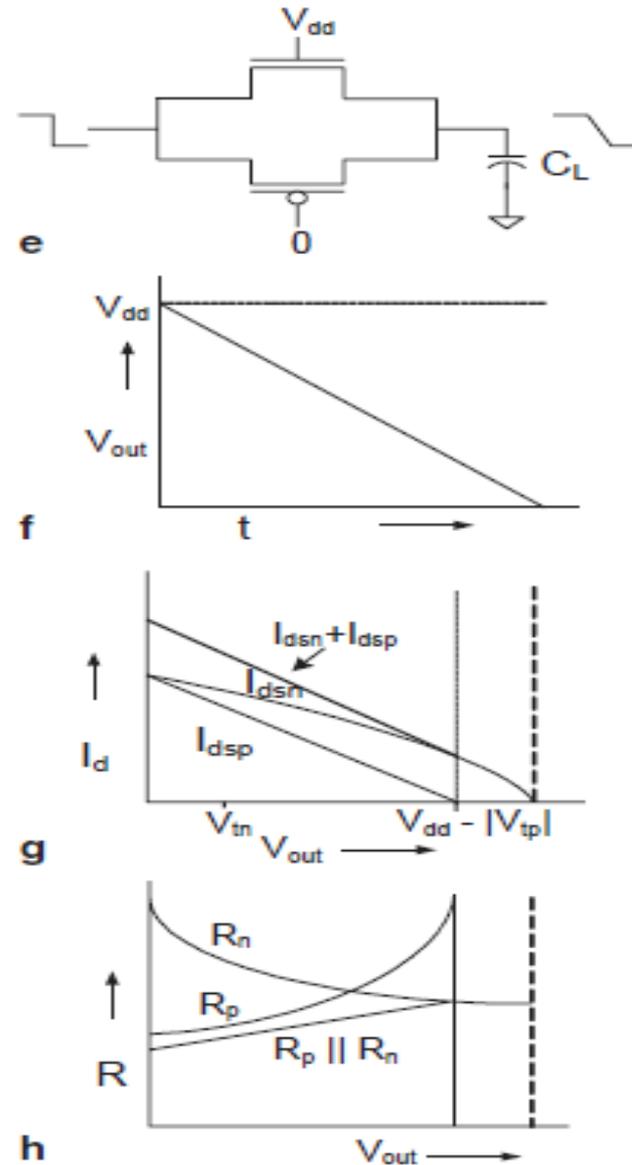
➤ It may be noted that the current decreases linearly as voltage builds up across the capacitor CL .

➤ *The equivalent resistances and their combined values are shown in Fig. 3.17d.*



MOS Transistors as a Switch-Transmission Gate

- Similarly, when the input voltage changes quickly from V_{dd} to 0 V and the load capacitance discharges through the switch, it can be visualized by Fig. 3.17e–h.





Low Power VLSI Circuits and Systems

Unit-2

Unit-2

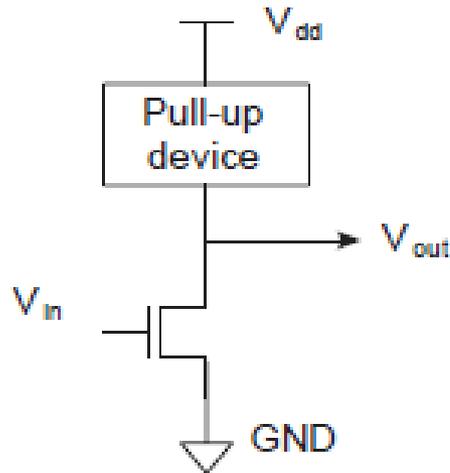
MOS Inverters

- ❖ Introduction
- ❖ Inverter and Its Characteristics
- ❖ MOS Inverter Configurations
- ❖ Inverter Ratio in Different Situations
- ❖ Switching Characteristics
- ❖ Delay Parameters
- ❖ Driving Large Capacitive Loads

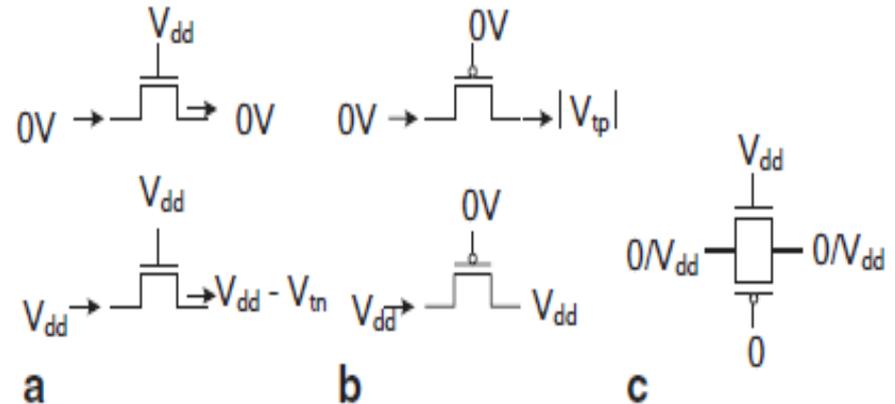
MOS Combinational Circuits

- ❖ Introduction
- ❖ Pass-Transistor Logic
- ❖ Gate Logic
- ❖ MOS Dynamic Circuits

MOS Inverters-Introduction

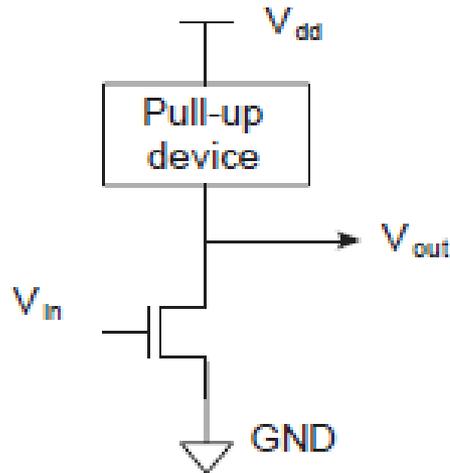


General structure of an nMOS inverter. nMOS n-type metal-oxide-semiconductor



- Metal–Oxide–Semiconductor (MOS) transistor can be considered as a **voltage-controlled resistor**.
- This basic property can be used to realize **digital circuits** using MOS transistors.
- We discuss the realization of various types of MOS inverters.

MOS Inverters-Introduction



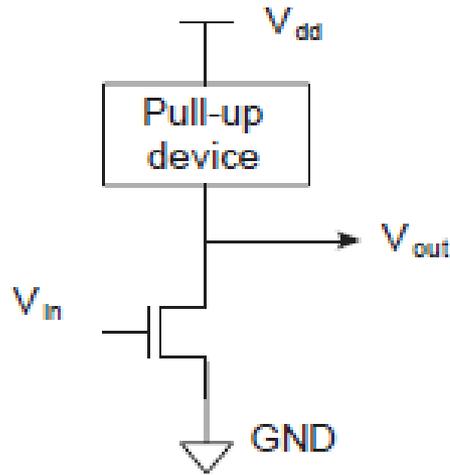
General structure of an nMOS inverter. nMOS n-type metal-oxide-semiconductor

- ❖ A pull-up device when energized will pull the output to supply (i.e. "1") and a pull-down will pull the output to ground (i.e. "0").
- ❖ Usually PMOS is used for pull-up since it can provide GOOD "1" (HIGH) i.e. V_{DD} and NMOS is pull-down since it can provide a GOOD "0" i.e. (LOW).

❖ The inverter forms the basic building block of **gate-based digital circuits**.

❖ An inverter can be realized with the source of an n-type metal-oxide-semiconductor (nMOS) enhancement transistor connected to the ground, and the drain connected to the positive supply rail V_{dd} through a **pull-up device**.

MOS Inverters-Introduction



General structure of an nMOS inverter. nMOS n-type metal-oxide-semiconductor



❖ The **pull-up device** can be realized in **several ways**. The **characteristics** of the inverter strongly depend on the **pull-up device** used to realize the inverter.

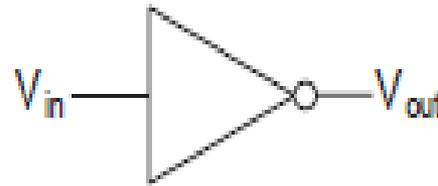
❖ Theoretically, a **passive resistor** of suitable value can be used. Although the use of a possible resistor may be possible in realizing an inverter using discrete components, this is not feasible in very-large-scale integration (VLSI) implementation

MOS Inverters

Inverter and Its Characteristics

Truth table

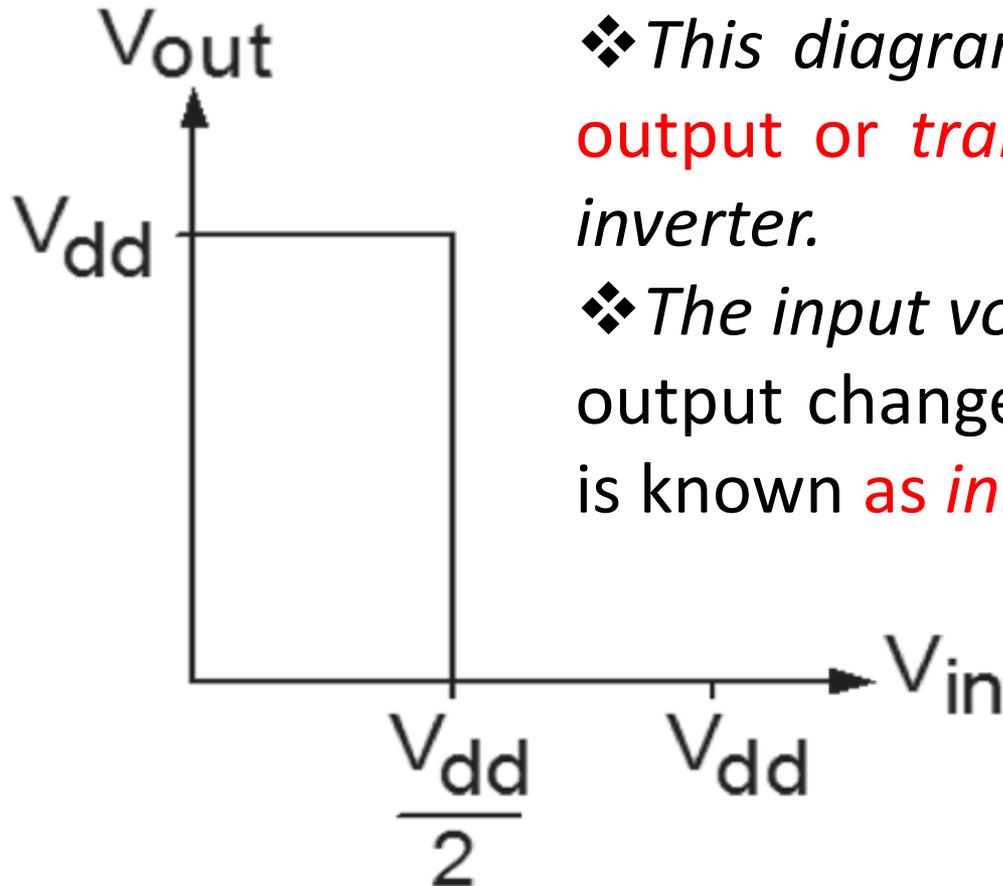
V_{in}	V_{out}
0	1
1	0



Truth table and
logic symbol of the inverter

MOS Inverters

Inverter and Its Characteristics



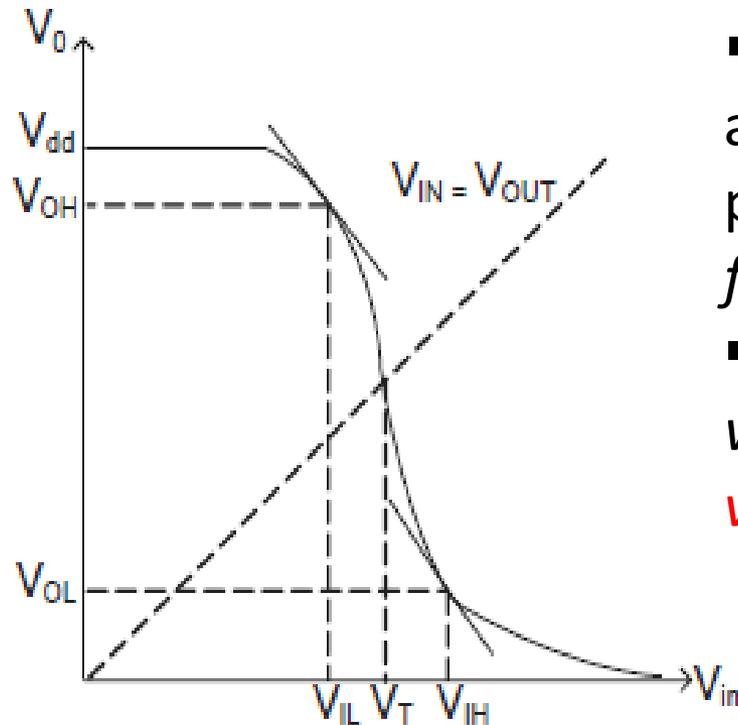
❖ This diagram is known as the **input-output or transfer characteristic** of the inverter.

❖ The input voltage, $V_{dd}/2$, at which the output changes from high '1' to low '0', is known **as inverter threshold voltage**.

Ideal transfer characteristics of an inverter

MOS Inverters

Inverter and Its Characteristics

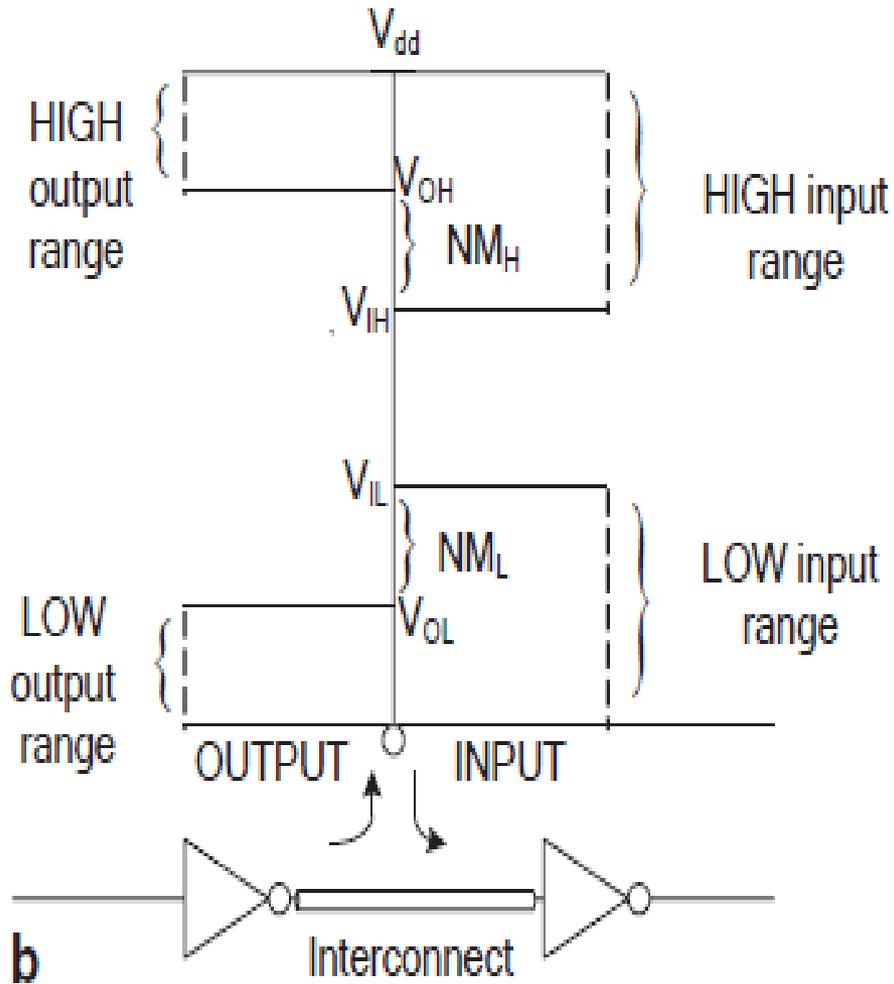


a

- Because of some voltage drop across the pull-up device, the output high voltage level is less than V_{dd} for the *low input voltage level*.
- This voltage is represented by **V_{OH}** , which is the *maximum output voltage level* for output level '1'.

Various voltage levels on the transfer characteristics

MOS Inverters-Inverter and Its Characteristics



b low- and high-level noise margins

✓ An important parameter called the *noise margin* is associated with the input-output voltage characteristics of a gate.

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

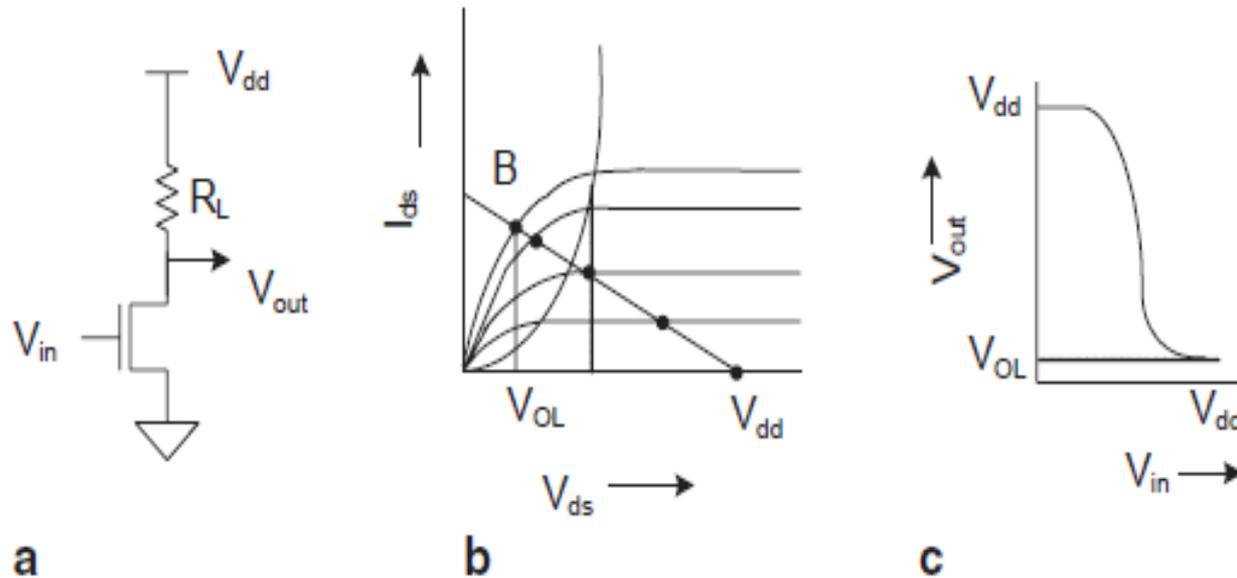
✓ The deviations in logic levels from the ideal values, which are restored as the signal propagates to the output, can be obtained from the DC characteristic curves.'1'.

MOS Inverters-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 Inverters-MOS Inverter Configurations

❖ *Passive Resistive as Pull-up Device*

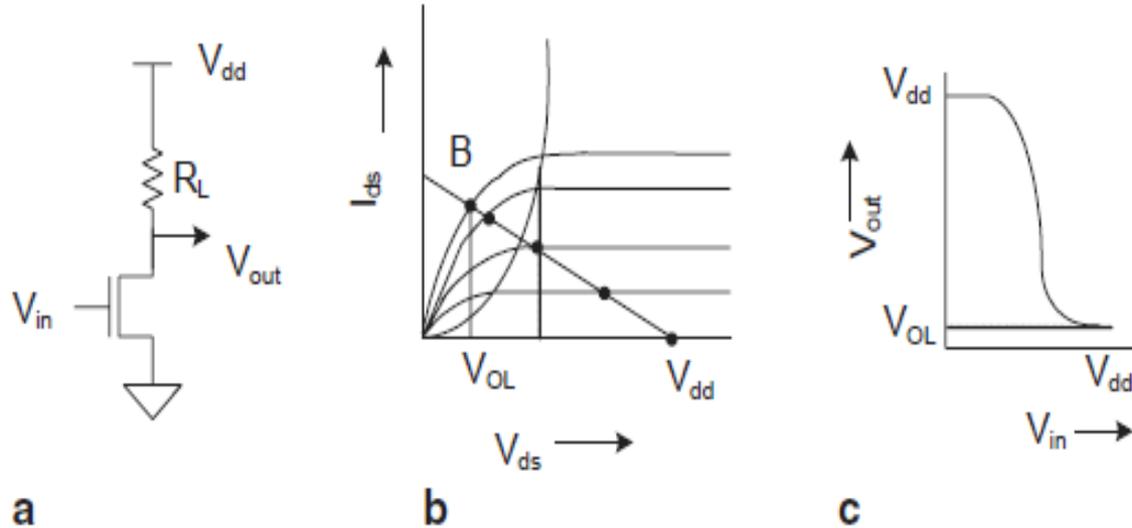


a An nMOS inverter with resistive load; b voltage–current characteristic; c transfer characteristic. nMOS n-type–metal–oxide semiconductor

- ❖ A passive resistor R_L can be used as the pull-up device as shown in Figure.
- ❖ The value of the resistor should be chosen such that the circuit functionally behaves like an inverter.

MOS Inverters-MOS Inverter Configurations

❖ *Passive Resistive as Pull-up Device*



a An nMOS inverter with resistive load; b voltage-current characteristic; c transfer characteristic. nMOS n-type-metal-oxide semiconductor

- ❑ When the input voltage V_{in} is less than V_{tn} , the transistor is OFF and the output capacitor charges to V_{dd} .
- ❑ Therefore, we get V_{dd} as the output for any input voltage less than V_{tn} .
- ❑ When V_{in} is greater than V_{tn} , the MOS transistor acts as a resistor R_c , where R_c is the channel resistance with $V_{gs} > V_{tn}$.

MOS Inverters-MOS Inverter Configurations

❖ *Passive Resistive as Pull-up Device*

This implementation of this inverter has a number of **disadvantages**:

- As the **charging** of the **output capacitor** takes place through the **load resistor R_L** and **discharge** through **R_c** and *their values must be different, there is asymmetry in the ON-to-OFF and OFF-to-ON switching times.*
- To have higher speeds of operation, the value of both **R_c** and **R_L** should be reduced.
- However, this **increases the power dissipation** of the circuit.
- Moreover, as we shall see later, to achieve a smaller value of **R_c** , *the area of the MOS inverter* needs to be **increased**.

MOS Inverters-MOS Inverter Configurations

❖ *Passive Resistive as Pull-up Device*

This implementation of this inverter has a number of disadvantages:

- The resistive load can be fabricated by two approaches—using a diffused resistor approach or using an undoped poly-silicon approach.

- ❑ In the first case, an n-type or a p-type isolated diffusion region can be fabricated to realize a resistor between the power supply line and the drain of the nMOS transistor.

- ❑ To realize a resistor of the order of **few $K \Omega$** , *as required for proper operation of the circuit*, the length to width must be large.

MOS Inverters-MOS Inverter Configurations

❖ *nMOS Depletion-Mode Transistor as Pull up*

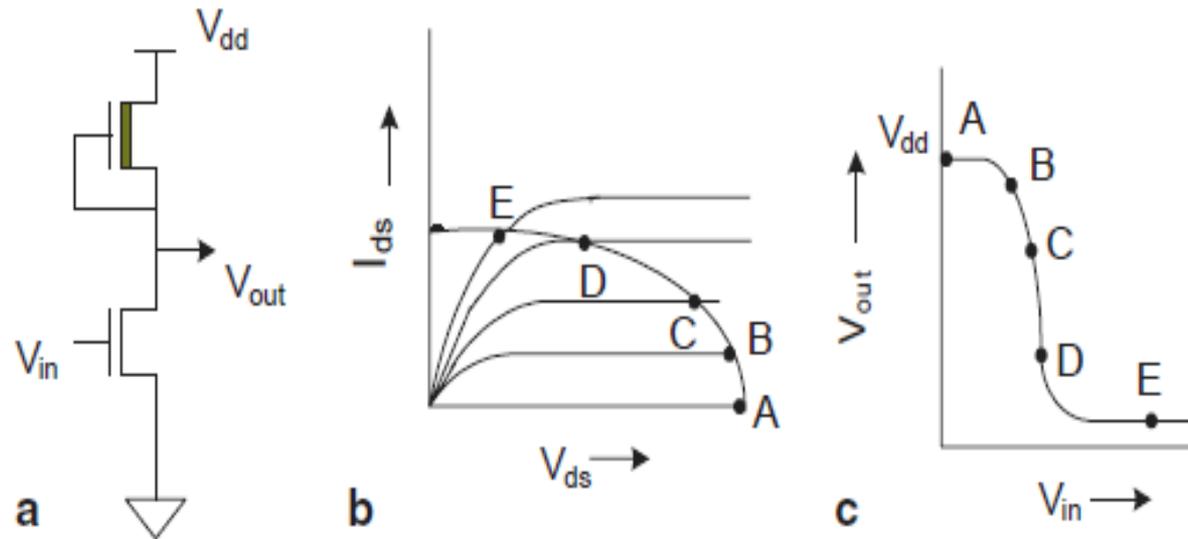


Fig. 4.7 a nMOS inverter with depletion-mode transistor as pull-up device; b voltage current characteristic; c transfer characteristic. nMOS n-type metal-oxide-semiconductor

- ❖ Any one of the transistors can be used as a pull-up device.
- ❖ First, we consider the use of an nMOS depletion-mode transistor as an active pull-up (pu) device as shown in Fig. 4.7a.
- ❖ As the output of an inverter is commonly connected to the gate of one or more MOS transistors in the next stage, there is no fan-out current, and the currents flowing through both the transistors must be equal.

MOS Inverters-MOS Inverter Configurations

❖ *nMOS Depletion-Mode Transistor as Pull up*

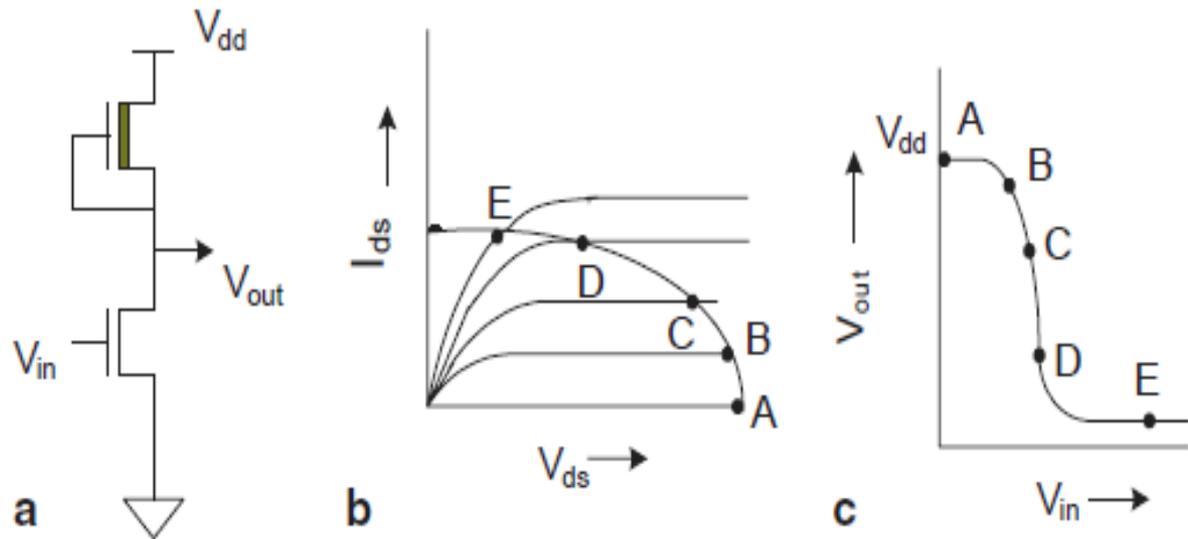


Fig. 4.7 a nMOS inverter with depletion-mode transistor as pull-up device; b voltage current characteristic; c transfer characteristic. nMOS n-type metal-oxide-semiconductor

❖ The input voltage is applied to the gate of the pull-down (pd) transistor, and the output is taken out from the drain of the pd device.

MOS Inverters-MOS Inverter Configurations

❖ *nMOS Enhancement-Mode Transistor as Pull up*

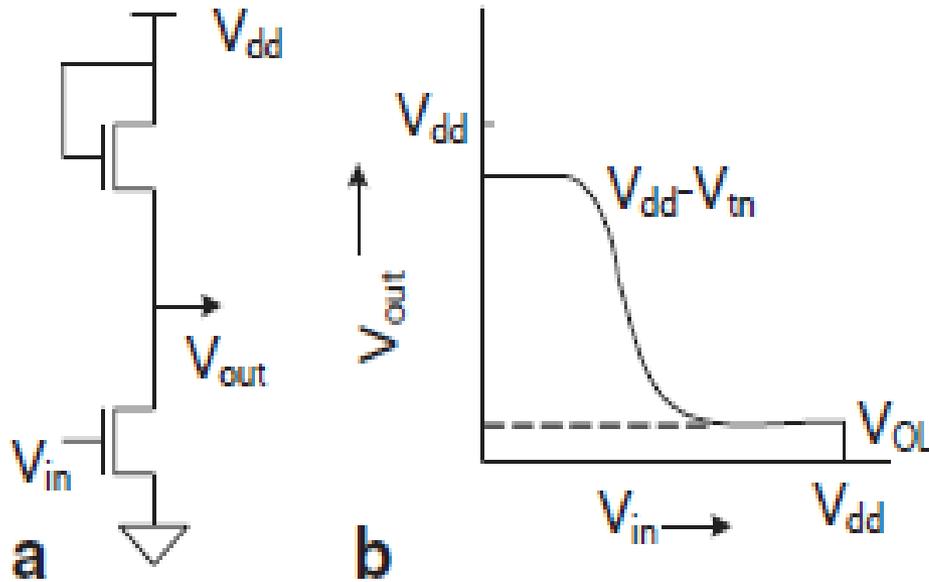


Fig. 4.8 a nMOS inverter with enhance-mode transistor as a pull-up device;
b transfer characteristic. nMOS n-type metal-oxide-semiconductor

❖ Let us consider the output voltage for two situations—when **$V_{in} = 0$ and $V_{in} = V_{dd}$** .

❖ *In* the first case, the desired output is V_{dd} .

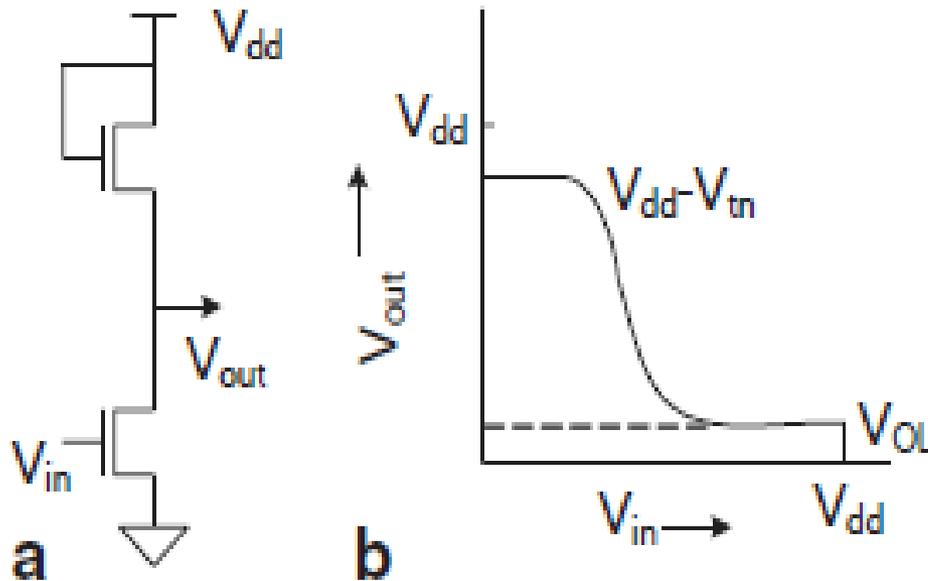
❖ *But as the output, V_{out} , approaches the voltage ($V_{dd} - V_{tn}$), the pull-up transistor turns off.*

❖ *Therefore, the output voltage cannot reach V_{dd} .*

❖ *The maximum output voltage that can be attained is ($V_{dd} - V_{tn}$), where V_{tn} is the threshold voltage of the enhancement-mode pull-up transistor.*

MOS Inverters-MOS Inverter Configurations

❖ *nMOS Enhancement-Mode Transistor as Pull up*



The output voltage for $V_{in} = V_{dd}$ is not 0 V, because in this case both the transistors are conducting and act as a voltage divider.

The transfer characteristic is shown in Fig. 4.8b.

Fig. 4.8 a nMOS inverter with enhance-mode transistor as a pull-up device;
b transfer characteristic. nMOS n-type metal-oxide-semiconductor

MOS Inverters-MOS Inverter Configurations

❖ *The pMOS Transistor as Pull Up*

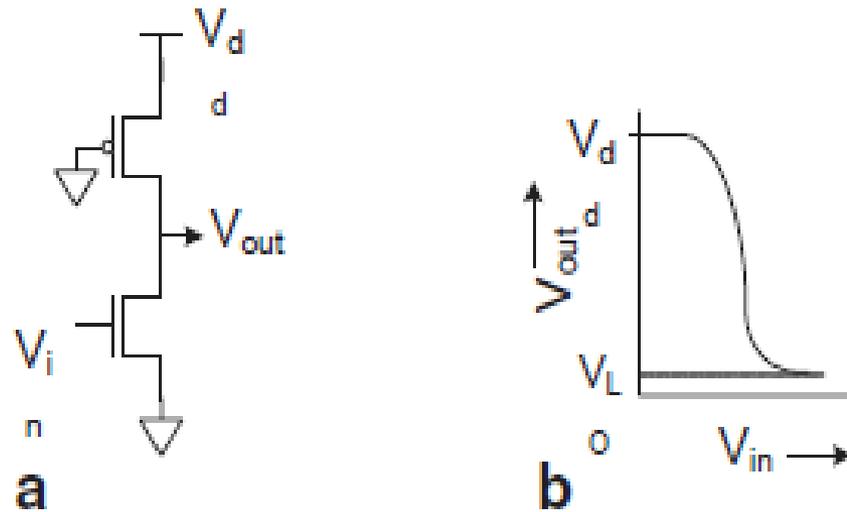


Fig. 4.9 a A pseudonMOS inverter;

b transfer characteristic.

PseudonMOS pseudo-n-type metal-oxide-semiconductor

❖ We can realize another type of inverter with a pMOS transistor as a pull-up device with its gate permanently connected to the ground as shown in Fig. 4.9a.

❖ As it is functionally similar to a depletion-type nMOS load, it is called a 'pseudo-nMOS' inverter.

MOS Inverters-MOS Inverter Configurations

❖ *The pMOS Transistor as Pull Up*

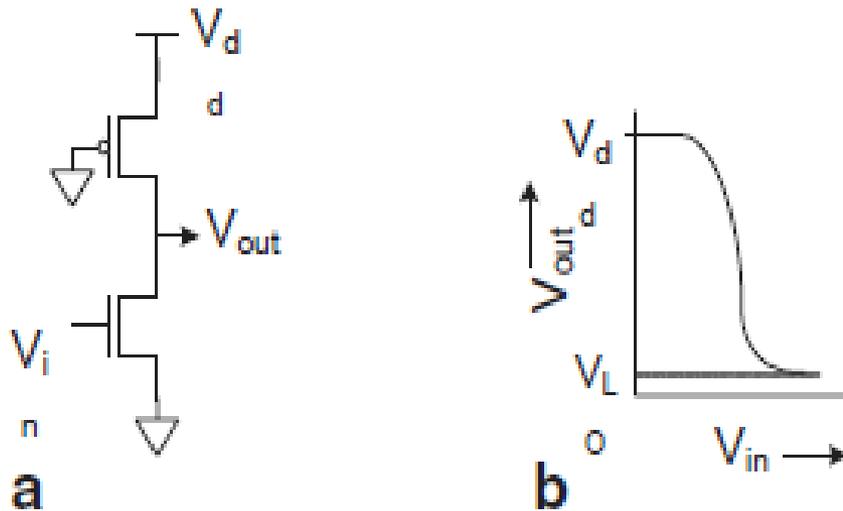


Fig. 4.9 a A pseudonMOS inverter;
b transfer characteristic.

PseudonMOS pseudo-n-type metal-oxide-
 semiconductor

- ❖ Unlike the CMOS inverter, discussed in Sect. 4.2.4, the pull-up transistor always remains ON, and there is DC current flow when the pull-down device is ON.
- ❖ The low-level output is also not zero and is dependent on the β_n / β_p ratio like depletion-type nMOS load.
- ❖ The voltage-transfer characteristic is shown in Fig. 4.9b.

MOS Inverters-MOS Inverter Configurations

❖ *pMOS Transistor as a Pull Up in Complementary Mode*

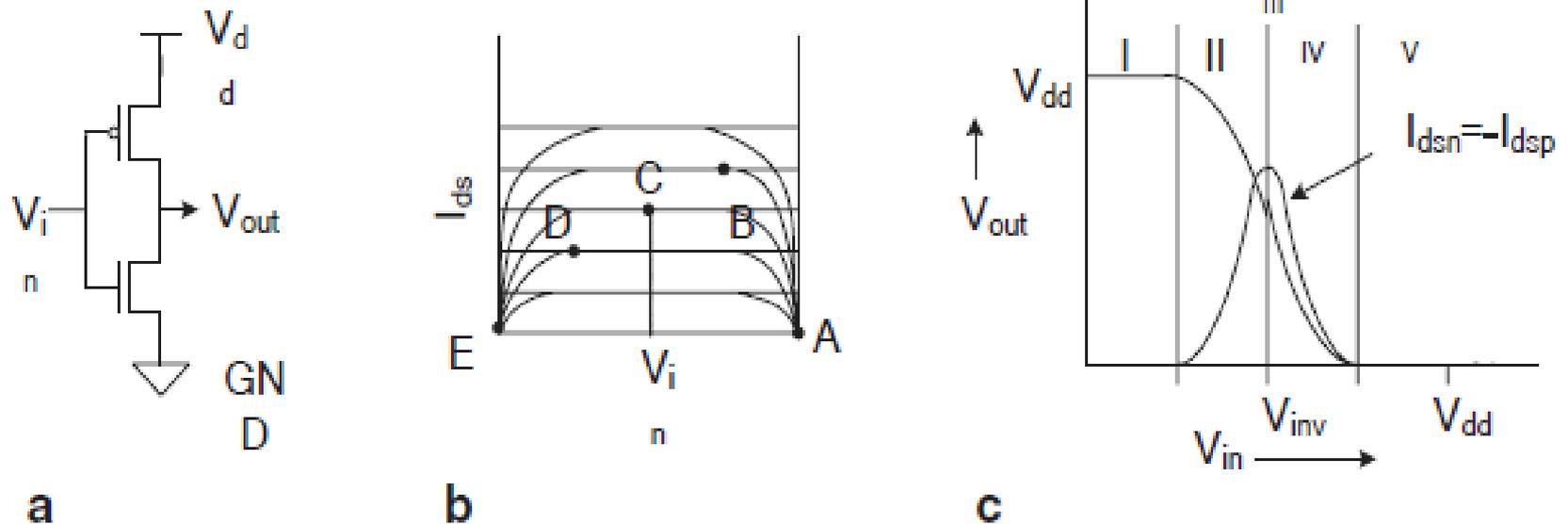


Fig. 4.10 a CMOS inverter; b voltage–current characteristic; and c transfer characteristic

- ❖ In this case, a **pMOS enhancement type transistor** is used as a pull-up device.
- ❖ However, here the **gates of both the pull-up and pull-down transistors are tied together and used as input** as shown in Fig. 4.10a.
- ❖ Output is taken from the drain of the pulldown device as usual.

MOS Inverters-MOS Inverter Configurations

❖ *pMOS Transistor as a Pull Up in Complementary Mode*

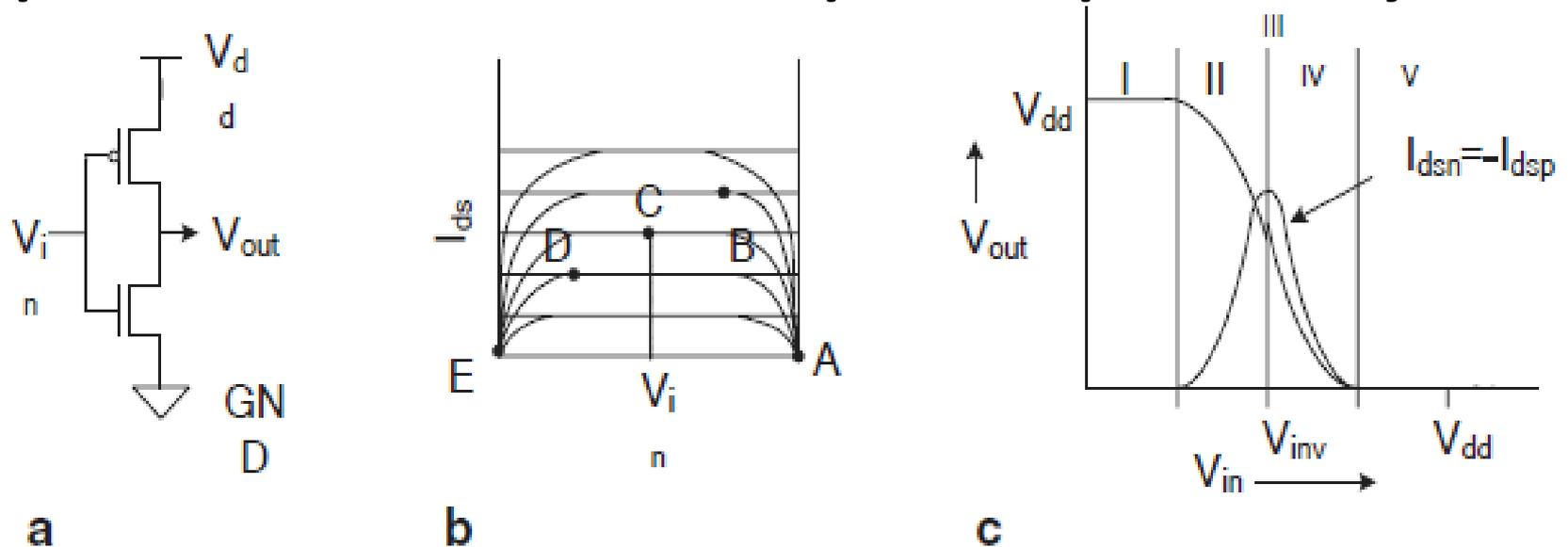


Fig. 4.10 a CMOS inverter; b voltage–current characteristic; and c transfer characteristic

❖ In this case, when the input voltage $V_{in} = 0\text{ V}$, the gate input of the pull-up transistor is below V_{dd} of its source voltage, i.e., $V_{gs} = -V_{dd}$, which makes the pull-up transistor ON, and the pull-down transistor OFF.

❖ So, there is no DC current flow between V_{dd} to ground.

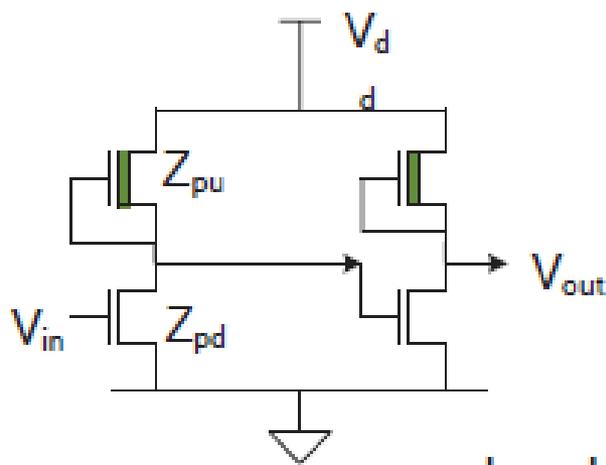
Comparison of the Inverters

Table 4.1 Comparison of the inverters

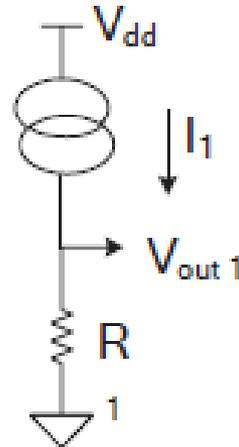
Inverters	V_{LO}	V_{HI}	Noise margin	Power
Resistor	Weak	Strong	Poor for low	High
nMOS depletion	Weak	Strong	Poor for low	High
nMOS enhancement	Weak	Weak	Poor for both low and high	High
Pseudo-nMOS	Weak	Strong	Poor for low	High
CMOS	Strong	Strong	Good	Low

nMOS n-type metal–oxide–semiconductor, *CMOS* complementary metal–oxide–semiconductor

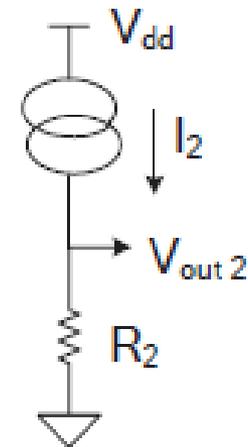
An nMOS Inverter Driven by Another Inverter



Inverter with $V_{in} = V_{dd}$



b



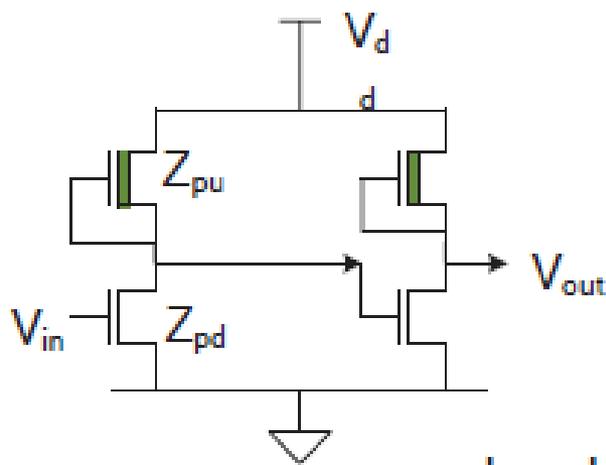
Inverter with $V_{in} = V_{dd} - V_t$

c

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

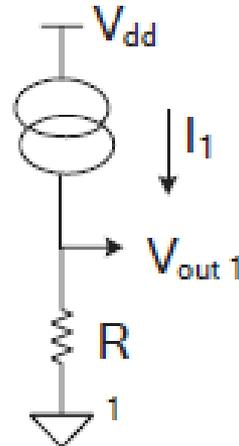
- In order to cascade two or more inverters without any degradation of voltage levels, we have to meet the condition $V_{in} = V_{out} = V_{inv}$; and for equal margins, let us set $V_{inv} = 0.5 V_{dd}$. This condition is satisfied when both the transistors are in saturation, and the drain current is given by

An nMOS Inverter Driven by Another Inverter



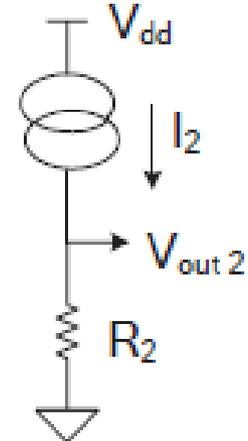
a

Inverter with $V_{in} = V_{dd}$



b

Inverter with $V_{in} = V_{dd} - V_t$



c

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

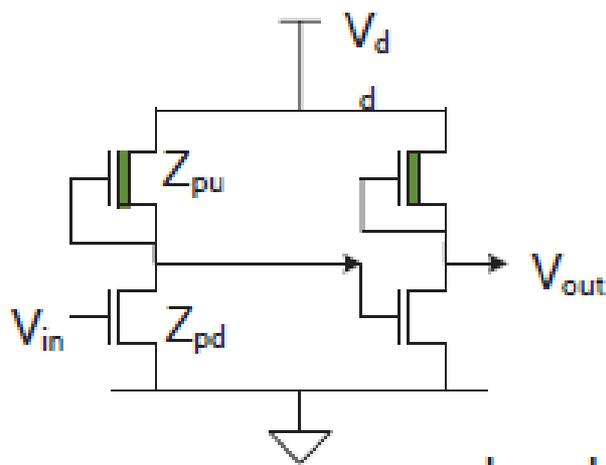
For the depletion-mode transistor,

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

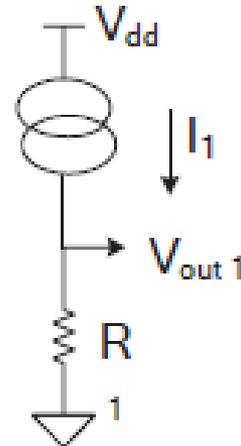
$$I_{ds}^{pu} = K \frac{W_{pu}}{L_{pu}} \frac{(-V_{tdp})^2}{2},$$

where V_{tdp} is the threshold voltage of the depletion-mode transistor and $V_{gs} = 0$.

An nMOS Inverter Driven by Another Inverter

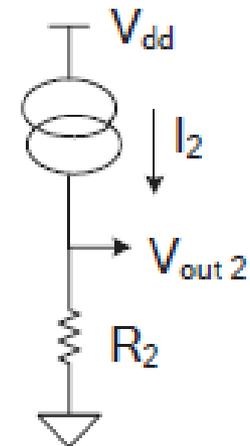


a



b

Inverter with $V_{in} = V_{dd}$



c

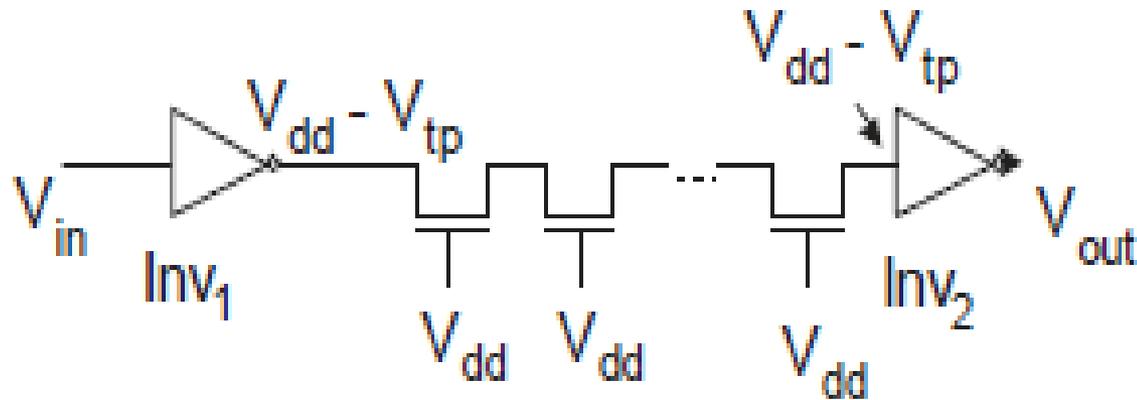
Inverter with $V_{in} = V_{dd} - V_t$

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

For the enhancement-mode transistor,

$$I_{ds} = K \frac{W_{pd} (V_{inv} - V_{tn})^2}{2L_{pd}}$$

An nMOS Inverter Driven Through Pass Transistors

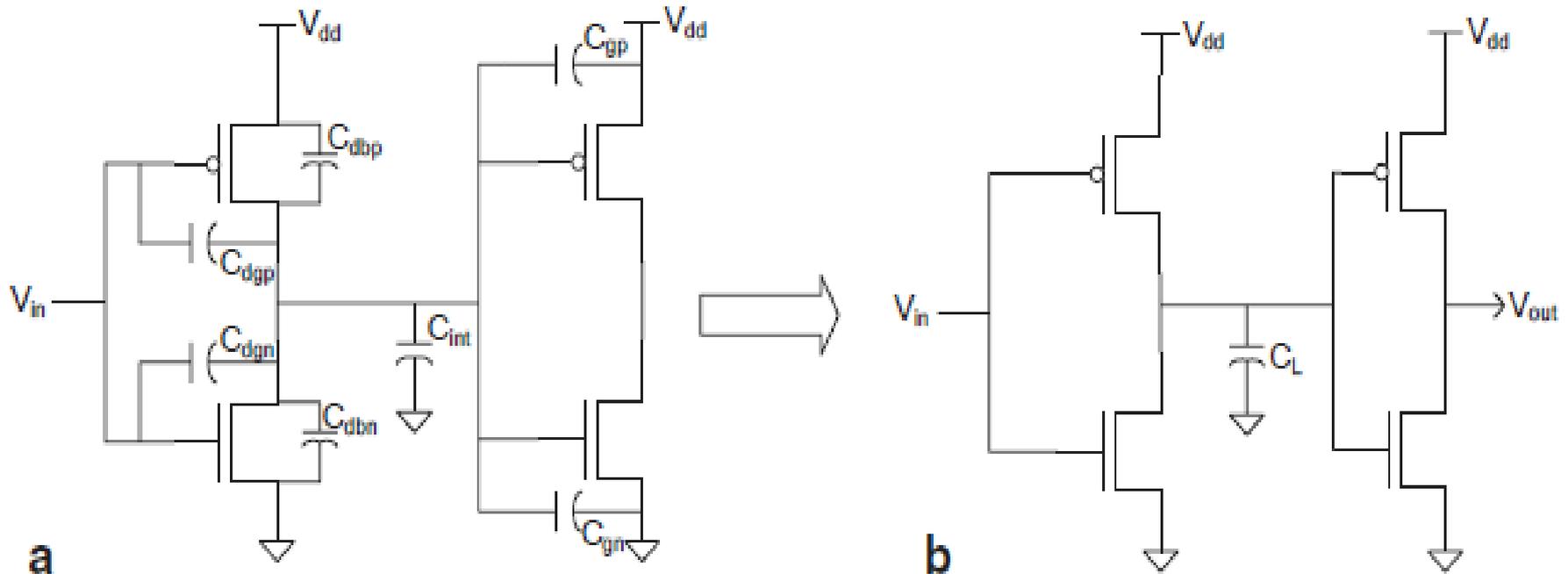


- ✓ Pass transistor passes a weak high level.
- ✓ If V_{dd} is applied to the input of a pass transistor, at the output, we get $(V_{dd} - V_{tp})$, where V_{tp} is the threshold voltage of the pass transistor.
- ✓ Therefore, instead of V_{dd} , a degraded high level $(V_{dd} - V_{tp})$ is applied to the second inverter.
- ✓ We have to ensure that the same voltage levels are produced at the outputs of the two Inverters in spite of different input voltage levels.

Switching Characteristics

- *Delay-Time Estimation*
- *Ring Oscillator*

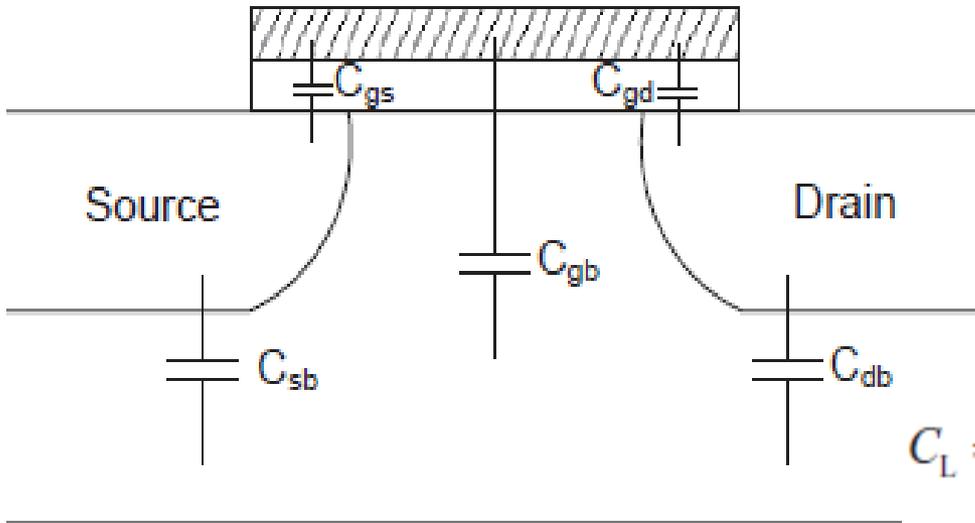
Switching Characteristics



a Parasitic capacitances of a CMOS inverter. b CMOS complementary metal–oxide– semiconductor

- The delay t_d is the time difference between the midpoint of the input swing and the midpoint of the swing of the output signal.
- The load capacitance shown at the output of the inverter represents the total of the input capacitance of driven gates, the parasitic capacitance at the output of the gate itself, and the wiring capacitance.

Switching Characteristics



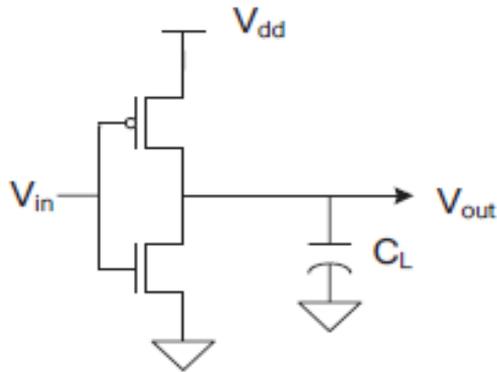
$$C_L = C_{dgn} + C_{dgp} + C_{dbn} + C_{dbp} + C_{int} + C_{gn} + C_{gp}.$$

Internal parasitic capacitances of an MOS transistor.

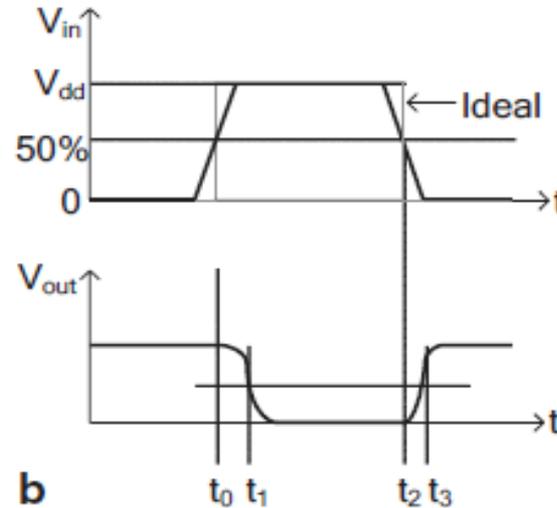
❑ The capacitances C_{gd} and C_{gs} are mainly due to gate overlap with the diffusion regions, whereas C_{db} and C_{gb} are voltage-dependent junction capacitances.

❑ The capacitance C_{out} is the lumped value of the distributed capacitances due to interconnection and C_{gn} and C_{gp} are due to the thin oxide capacitances over the gate area of the nMOS and pMOS transistors, respectively

Switching Characteristics-Delay-Time Estimation



a

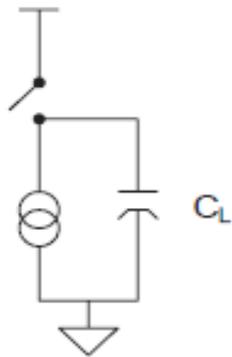


b

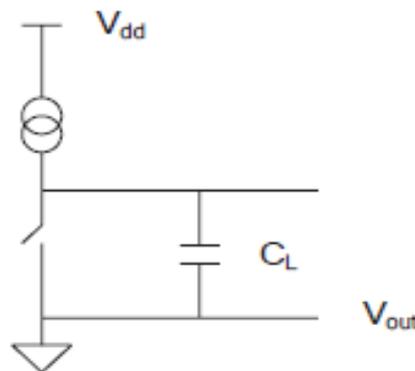
$$\tau_{phl} = t_1 - t_0$$

$$\tau_{plh} = t_3 - t_2$$

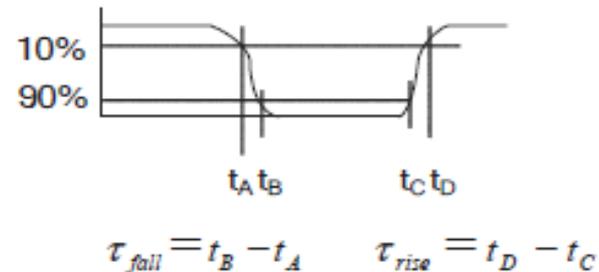
$$\tau_d = \frac{\tau_{phl} + \tau_{plh}}{2}$$



c



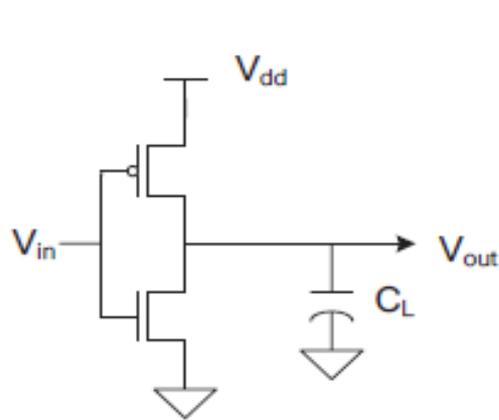
d



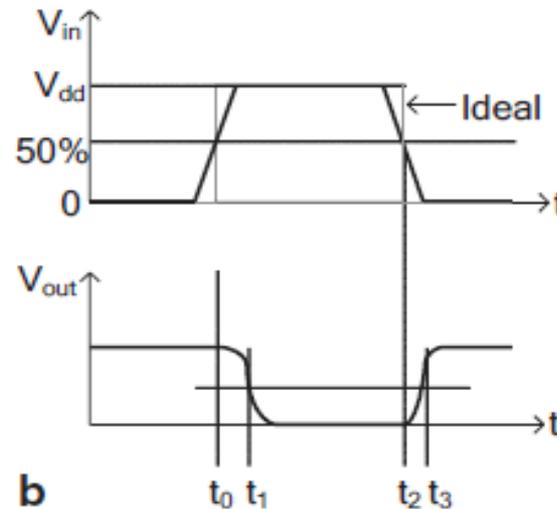
e

a CMOS inverter; **b** delay-time timings; **c** fall-time model; **d** rise-time model; **e** Rise time and fall times

Switching Characteristics-Delay-Time Estimation



a

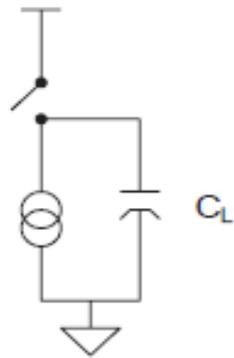


b

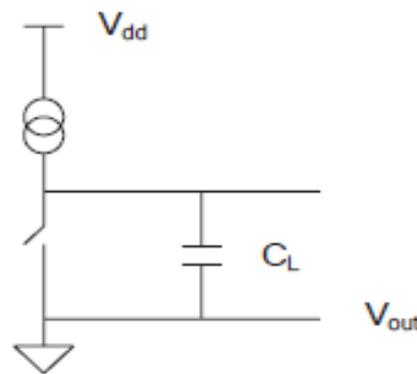
$$\tau_{pHL} = t_1 - t_0$$

$$\tau_{pLH} = t_3 - t_2$$

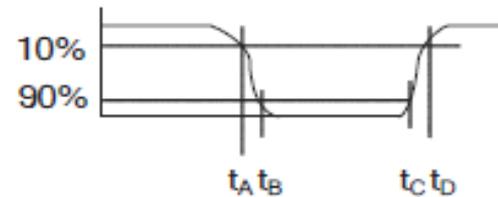
$$\tau_d = \frac{\tau_{pHL} + \tau_{pLH}}{2}$$



c



d



$$\tau_{fall} = t_B - t_A$$

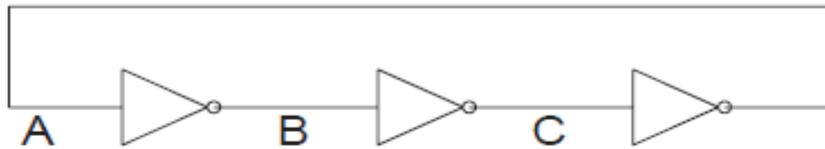
$$\tau_{rise} = t_D - t_C$$

e

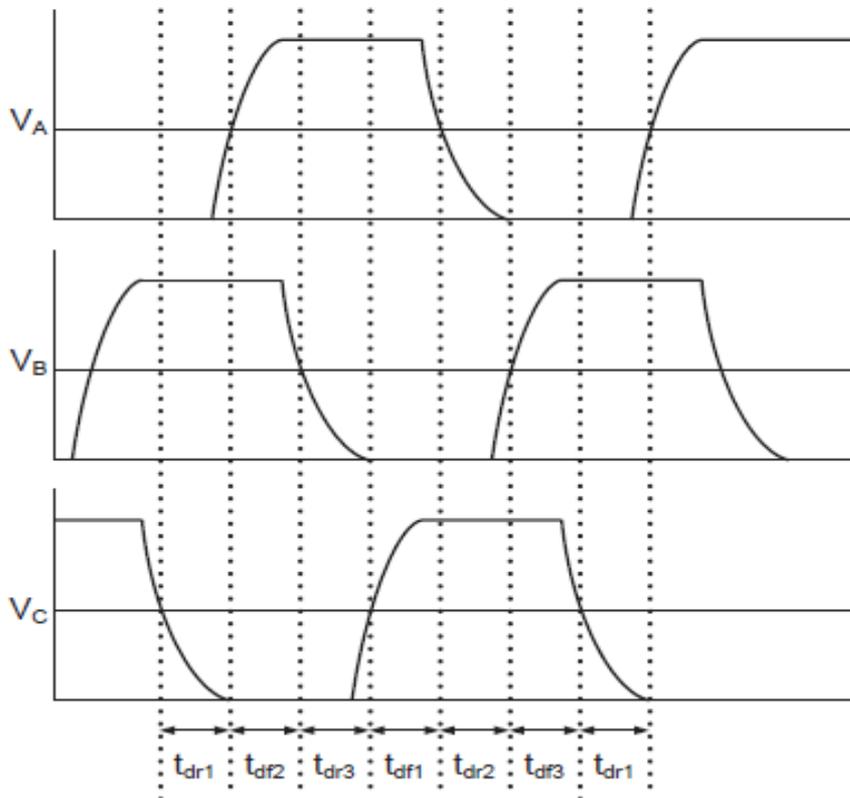
a CMOS inverter; b delay-time timings; c fall-time model; d rise-time model; e Rise time and fall times

The circuit for high-to-low propagation delay time t_{PHL} estimation

Switching Characteristics-*Ring Oscillator*



Output waveform of a three-stage ring oscillator



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

$$\begin{aligned}
 T &= t_{phl1} + t_{plh2} + t_{phl3} + t_{plh1} + t_{plh2} + t_{plh3} \\
 &= (t_{phl1} + t_{plh1}) + (t_{plh2} + t_{phl2}) + t_{phl3} + t_{plh3} \\
 &= 2t_d + 2t_d + 2t_d \\
 &= 6t_d \\
 &= 2.3t_d.
 \end{aligned}$$

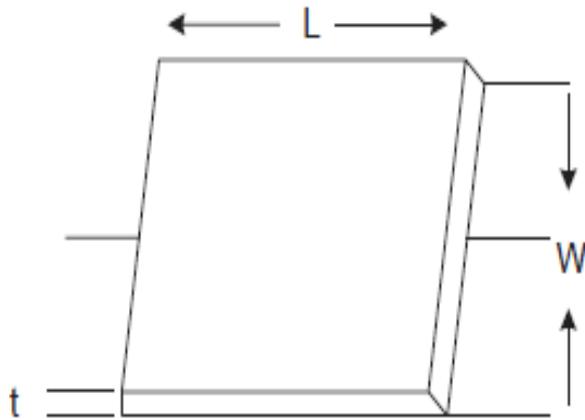
Switching Characteristics-*Ring Oscillator*

- ❖ For an n -stage (where n is an odd number) inverter, the time period $T = 2 \cdot n \cdot t_d$
- ❖ Therefore, the frequency of oscillation $f = 1/2nt_d$ or $t_d = 1/2nf$.
- ❖ *It may be noted* that the delay time can be obtained by measuring the frequency of the ring oscillator.
- ❖ For better accuracy of the measurement of frequency, a suitable value of n (say 151) can be chosen.
- ❖ This can be used for the characterization of a fabrication process or a particular design.
- ❖ The ring oscillator can also be used for on-chip clock generation.
- ❖ However, 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

- *Resistance Estimation*
- *Area Capacitance of Different Layers*
- *Standard Unit of Capacitance C_g*
- *The Delay Unit*

Delay Parameters-Resistance Estimation



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

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

One slab of conducting material

Let us consider a rectangular slab of conducting material of resistivity ρ , of width W , of thickness t and length L

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

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 a value between 1000 to 3000 Ω/sq .
- ❖ Since the mobility and the threshold voltage are functions of temperature, the channel resistance will vary with temperature.

Delay Parameters-*Resistance Estimation*

Typical sheet resistances for 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	-	$2.5 \times 10_4$	-

Delay Parameters-*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 in pF/ μm^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
Metal 2 to poly	0.3×10^{-4}	0.075

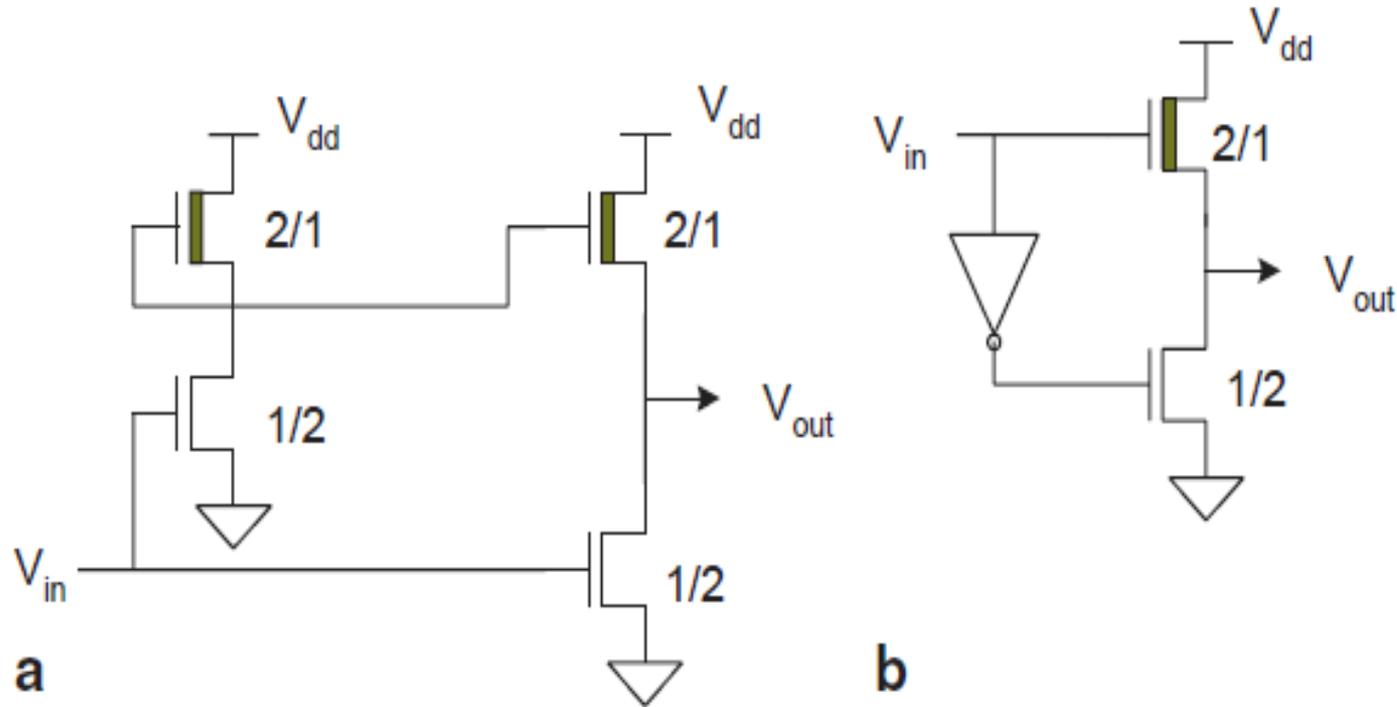
Delay Parameters-*Standard Unit of Capacitance C_g*

Capacitance	Value of pF/ μm^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

Driving Large Capacitive Loads

- There are situations when a large load capacitance such as, long buffers, off-chip capacitive load or I/O buffer are to be driven by a gate .
- In such cases, the delay can be very high if driven by a standard gate.
- A super buffer or a BiCMOS inverter is used or cascade of such gates of increasing can be used to achieve smaller delays
- ***Super Buffers***
- ***BiCMOS Inverters***
- ***Buffer Sizing***

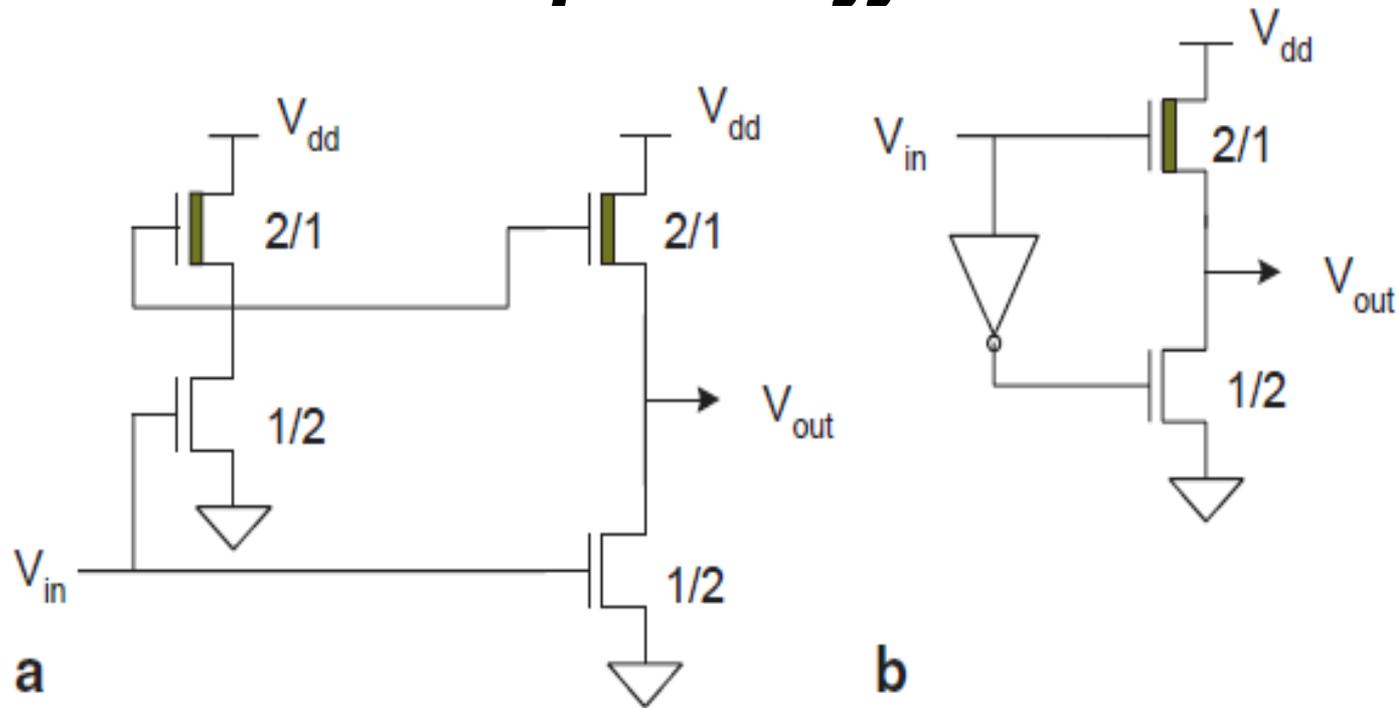
Super Buffers



❖ We have seen that one important drawback of the basic nMOS inverters (because of ratioed logic) in driving capacitive load is asymmetric drive capability of pull-up and pull-down devices.

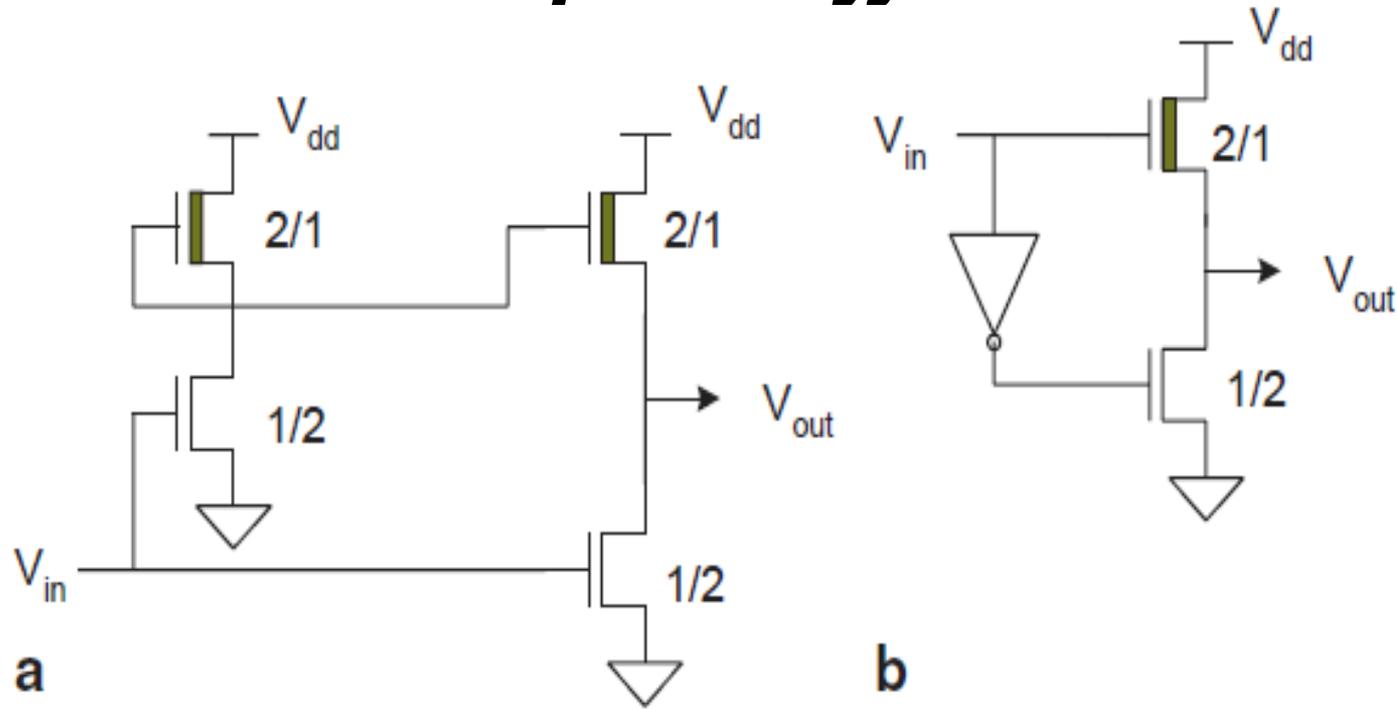
❖ This is because of longer channel length (four times) of the pull-up device.

Super Buffers



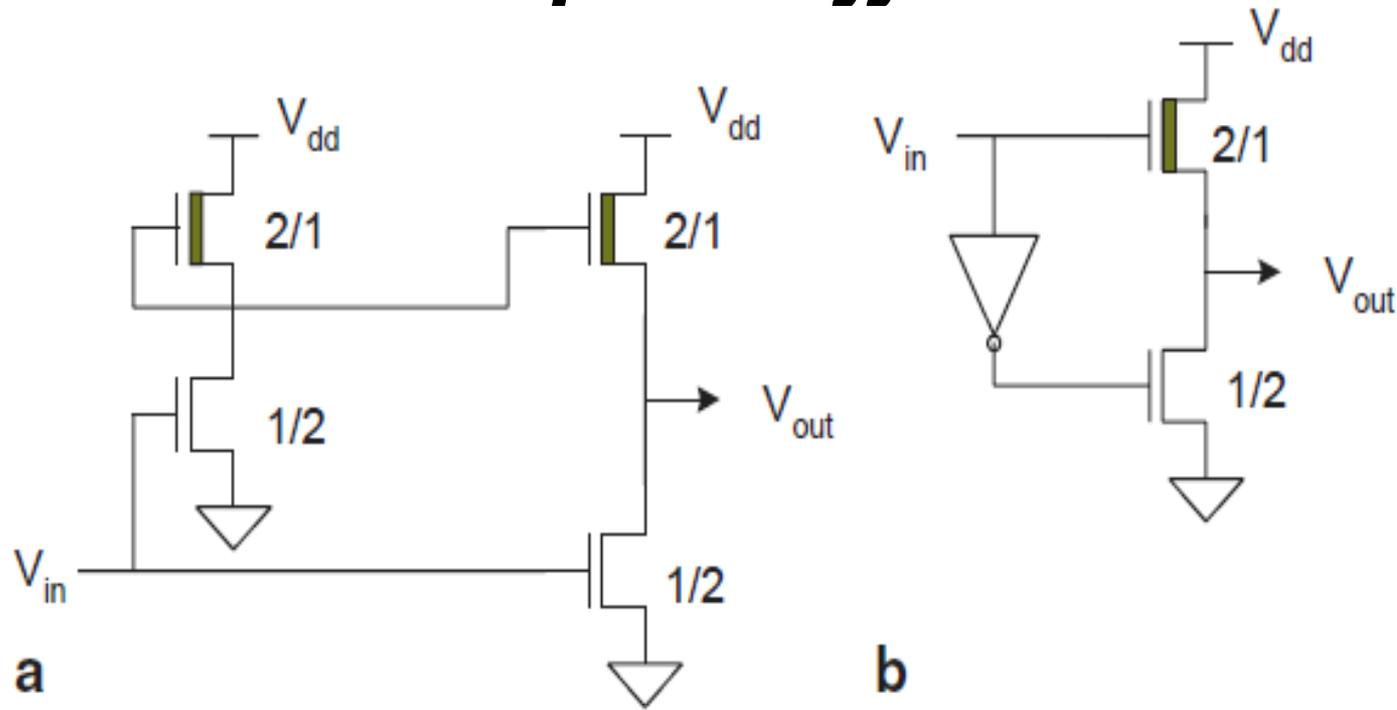
- ❖ Moreover, when the pull-down transistor is ON, the pull-up transistor also remains ON.
- ❖ As a consequence, a complete pull-down current is not used to discharge the load capacitance, but part of it is neutralized by the current passing through the pull-up transistors.
- ❖ This asymmetry can be overcome in especially designed circuits known as **super buffers**.

Super Buffers



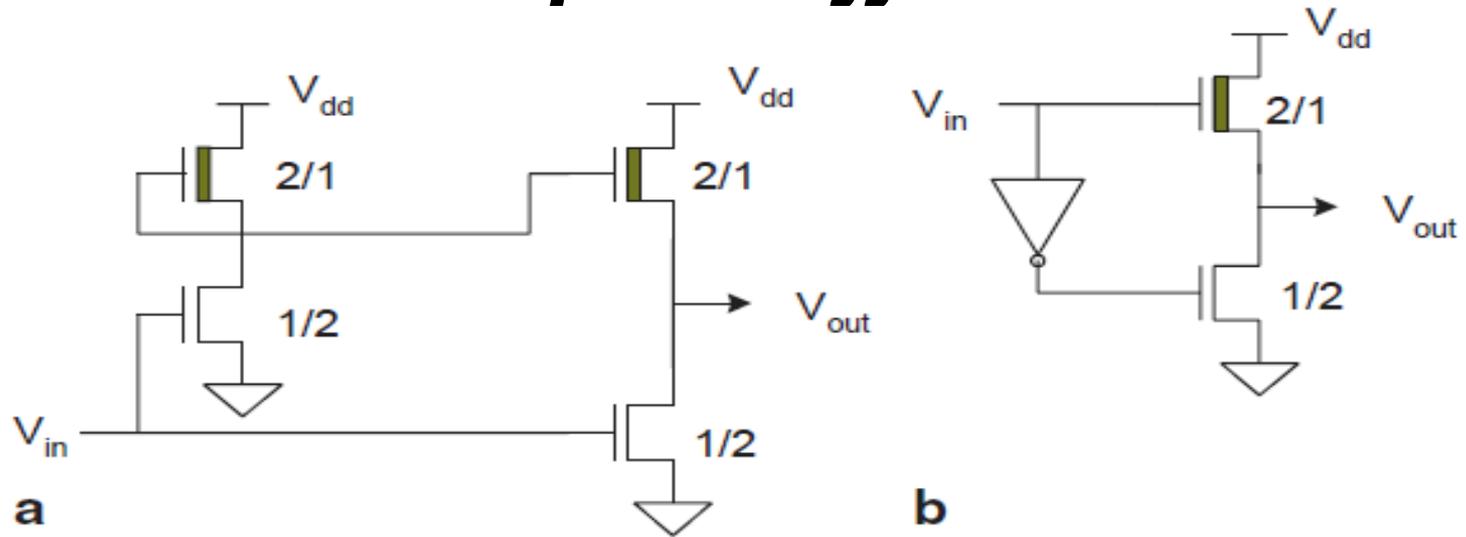
- ❖ In standard nMOS inverter, the gate of the depletion-mode pull-up device is tied to the source.
- ❖ Instead, the gate of the pull-up device of the super buffer is driven by another inverter with about twice the gate drive.

Super Buffers



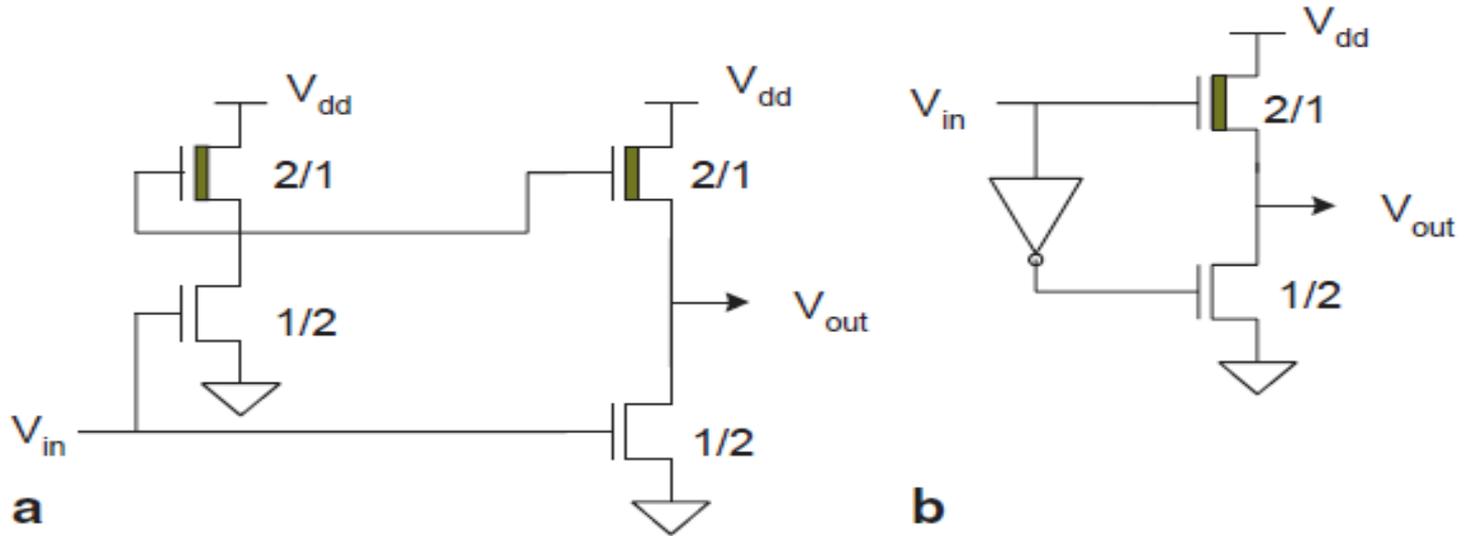
- Thus, the pull-up device is capable of sourcing about four times the current of the standard nMOS inverter.
- This is the key idea behind both the inverting and noninverting types of super buffers.
- This not only overcomes the asymmetry but also enhances the drive capability.

Super Buffers



- The output stage is a push–pull stage.
- The gate of the pull-up device is driven by a signal of opposite level of the pull-down device, generated using a standard inverter.
- For the inverting-type super buffer, when the input voltage is low, the gates of both the pull-down transistors are low, the gates of both the pull-up devices are high, and the output is high.

Super Buffers



- For a standard inverter, the drain current of the pull-up device in saturation ($0 < V_o < 2V$) and linear region are given as follows:

$$I_{ds}(\text{sat}) = \frac{\beta_{pu}}{2} (V_{gs} - V_{tdep})^2, \quad \text{where } V_{tdep} = -3V$$

$$= \frac{\beta_{pu}}{2} (0 + 3)^2 = 4.5 \beta_{pu} .$$

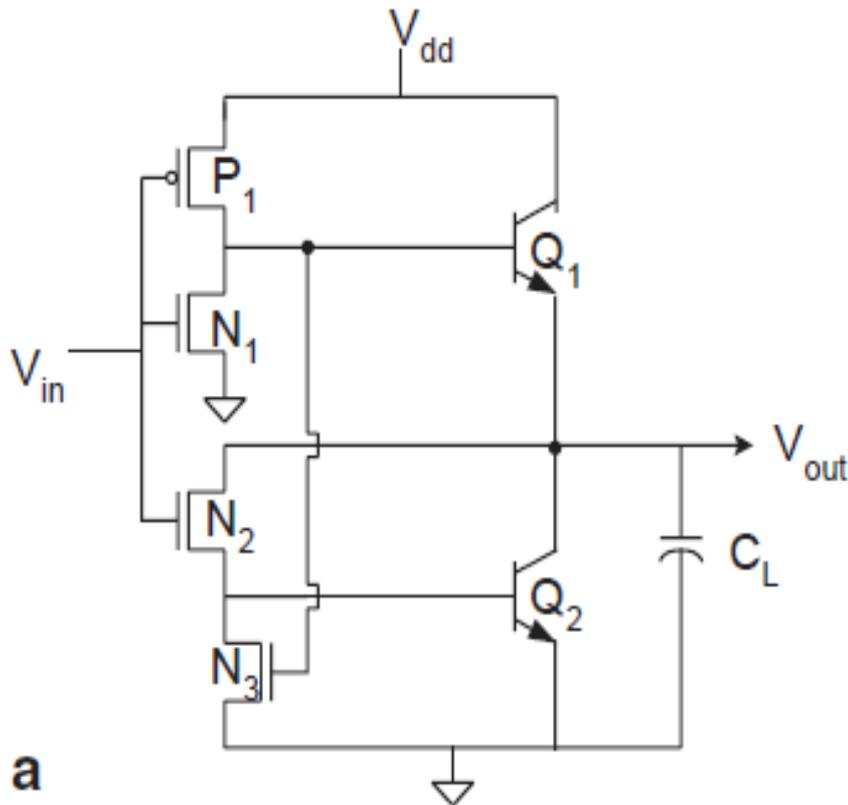
Super Buffers

The average source current $= \frac{(27.5+10.62)\beta_{pu}}{2} = 19.06\beta_{pu}$.

❖ Therefore, the current drive is $= 19.06 / 4.44 = 4.3$ times that of standard inverter for the super buffer using totem pole output configuration.

❖ This high drive also alleviates the asymmetry in driving capacitive loads and makes the nMOS inverter behave like a ratioless circuit.

BiCMOS Inverters



- ❖ Higher current drive capability of bipolar NPN transistors is used in realizing BiCMOS inverters.

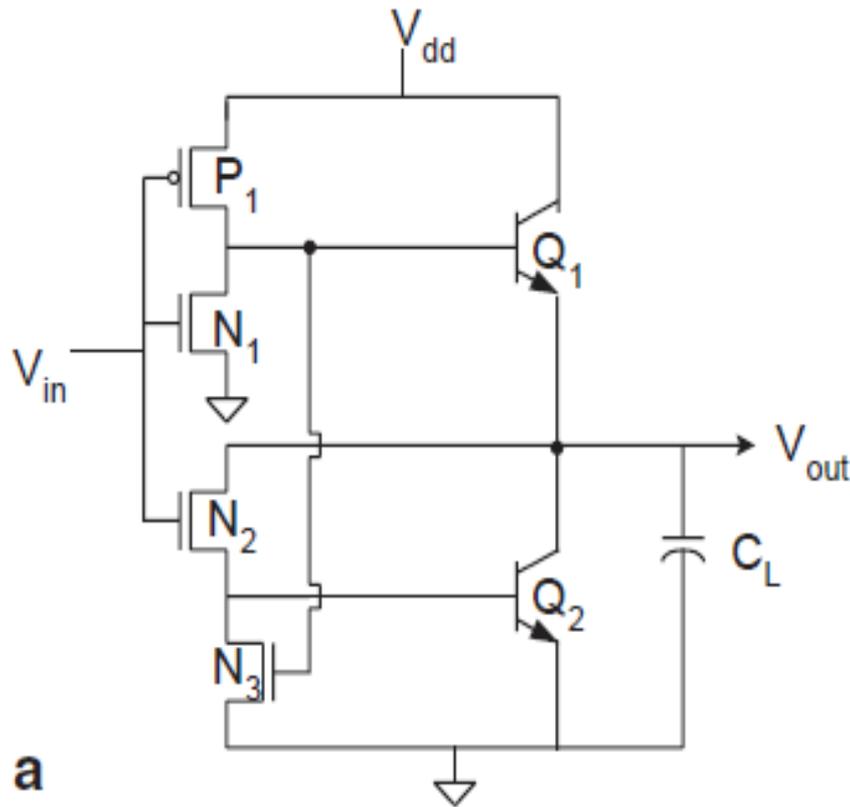
- ❖ The circuit requires four MOS transistors and two bipolar NPN transistors Q1 and Q2.

- When the input V_{in} is low, the pMOS transistor P1 is ON, which drives the base of the bipolar transistor Q1 to make it ON.

- The nMOS transistor N1 and N2 are OFF, but transistor N3 is ON, which shunts the base of Q2 to turn it OFF.

- The current through Q1 charges capacitor C_L and at the output V_{out} , we get a voltage ($V_{dd} - V_{be}$), where V_{be} is the base-emitter voltage drop of the transistor Q1.

BiCMOS Inverters

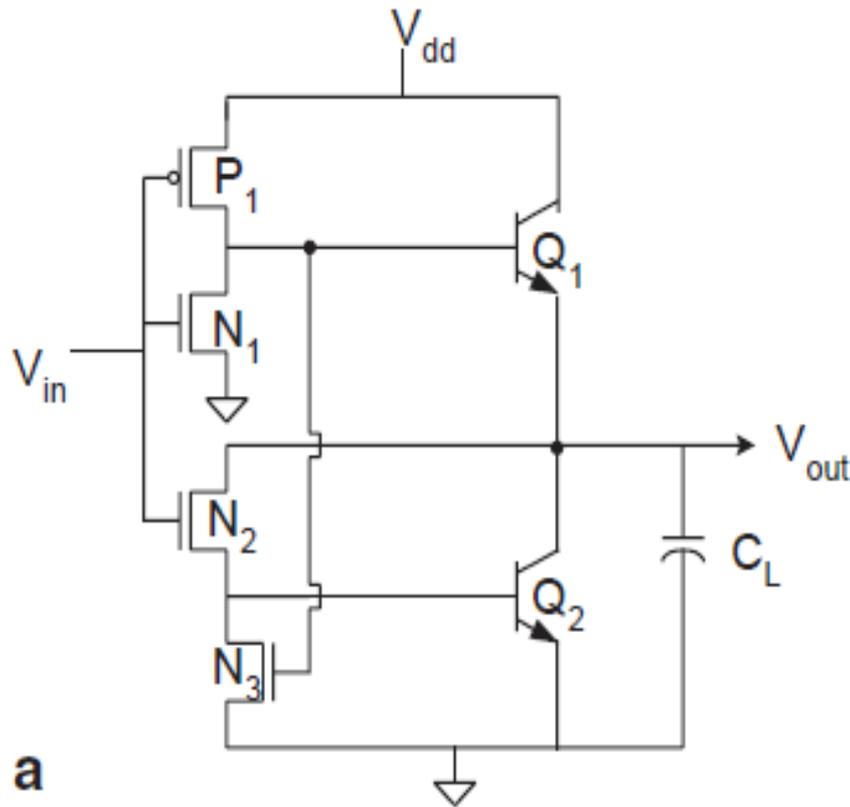


✓ If the input is high (V_{dd}), transistors $P1$ and $N3$ are OFF and $N1$ and $N2$ are ON.

✓ The drain current of $N2$ drives the base of the transistor $Q2$, which turns ON and the transistor $N1$ shunts the base of $Q1$ to turn it OFF

The capacitor C_L discharges through $Q2$. The conventional BiCMOS gate gives high-current-drive capability, zero static power dissipation and high input impedance.

BiCMOS Inverters

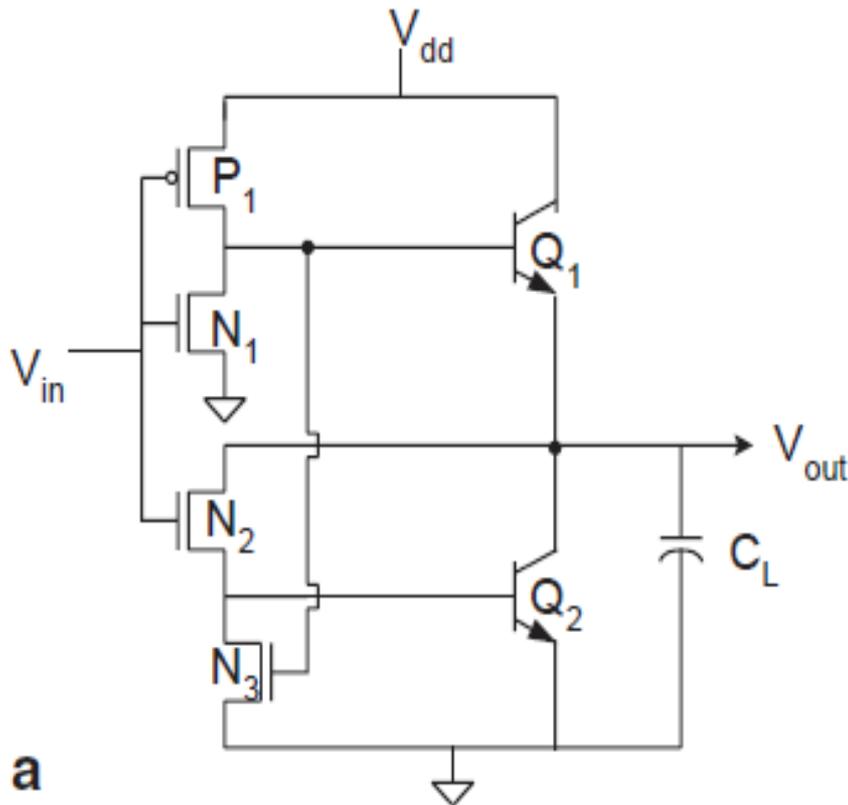


✓ If the input is high (V_{dd}), transistors $P1$ and $N3$ are OFF and $N1$ and $N2$ are ON.

✓ The drain current of $N2$ drives the base of the transistor $Q2$, which turns ON and the transistor $N1$ shunts the base of $Q1$ to turn it OFF

The capacitor C_L discharges through $Q2$. The conventional BiCMOS gate gives high-current-drive capability, zero static power dissipation and high input impedance.

BiCMOS Inverters



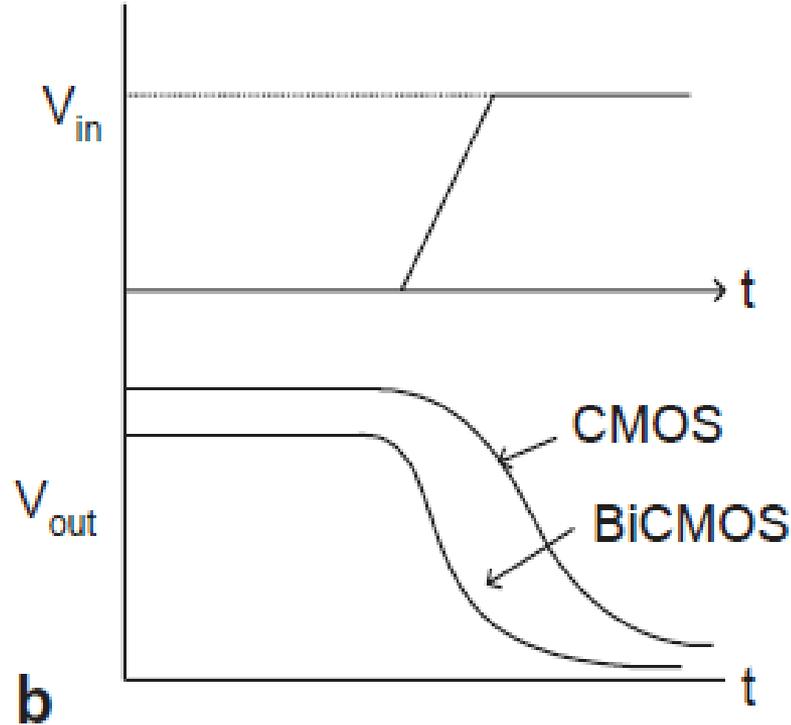
❖ For a zero input voltage, the pMOS transistor operates in the linear region.

❖ This drives the NPN transistor Q1, and the output we get is $V_{dd} - V_{be}$, where V_{be} is the base-emitter voltage drop of Q1.

❖ As input voltage increases, the subthreshold leakage current of the transistor N3 increases leading to a drop in the output voltage.

For $V_{in} = V_{inv} (V_{dd}/2)$, both P1 and N2 transistors operate in the saturation region driving both the NPN transistors (Q1 and Q2) ON.

BiCMOS Inverters



✓ In this region, the gain of the inverter is very high, which leads
✓ to a sharp fall in the output voltage, as shown in Fig. 4.22b. As the input voltage is further increased, the output voltage drops to zero.

✓ The output voltage characteristics of CMOS and BiCMOS are compared in Fig. 4.22b.

✓ It may be noted that the BiCMOS inverter does not provide strong high or strong low outputs.

✓ High output is $V_{dd} - V_{CE1}$, where V_{CE1} is the saturation voltage across $Q1$ and the low-level output is V_{CE2} , which is the saturation voltage across $Q2$.

Buffer Sizing

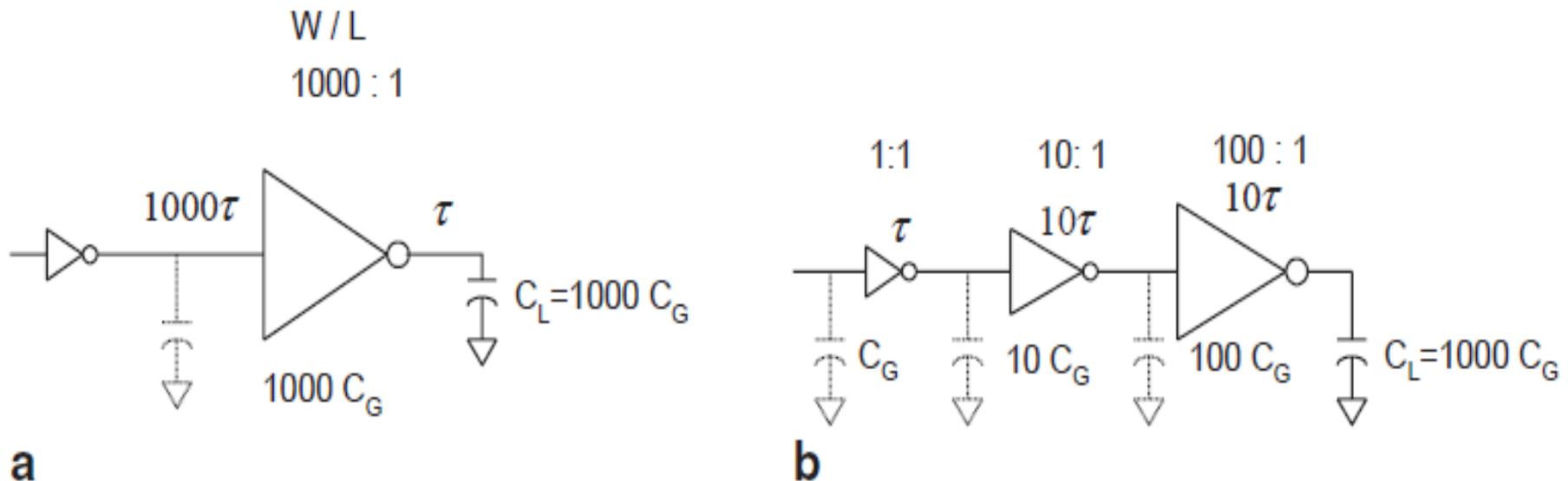
- It may be observed that an MOS transistor of unit length (2λ) has gate capacitance proportional to its width (W), *which may be multiple of λ .*
- *With the increase of the width, the current driving capability is increased.*
- But this, in turn, also increases the gate capacitance.
- As a consequence, the delay in driving a load capacitance C_L by a transistor of gate capacitance C_g is *given by the relationship (c_L / c_g) , τ where τ is the unit delay, or delay in driving an inverter by another of the same size.*

Buffer Sizing

- Let us now consider a situation in which a large capacitive load, such as an output pad, is to be driven by an MOS gate.
- The typical value of such load capacitance is about 100 pF, which is several orders of magnitude higher than C_g .
- *If such a load* is driven by an MOS gate of minimum dimension ($2\lambda \times 2\lambda$), then the delay will be $10^3\tau$.

Buffer Sizing

- To reduce this delay, if the driver transistor is made wider, say $10^3 \times 2\lambda$, the delay of this stage becomes τ , *but the delay in driving this driver stage is 1000τ , so, the total delay is 1001τ , which is more than the previous case.*
- *It has been observed that the overall delay can be minimized by using a cascaded stage of inverters of increasing size as shown in Fig. 4.24.*



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.***

MOS Combinational Circuits-Introduction

- 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.

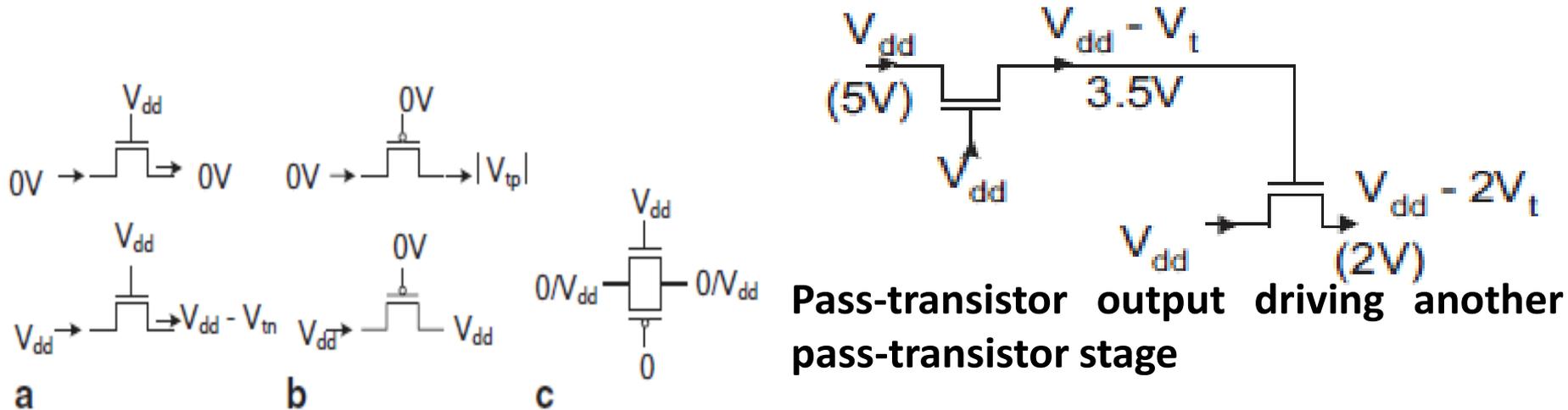
MOS Combinational Circuits

- **Pass-Transistor Logic**
- **Gate Logic**
- **MOS Dynamic Circuits**

Pass-Transistor Logic

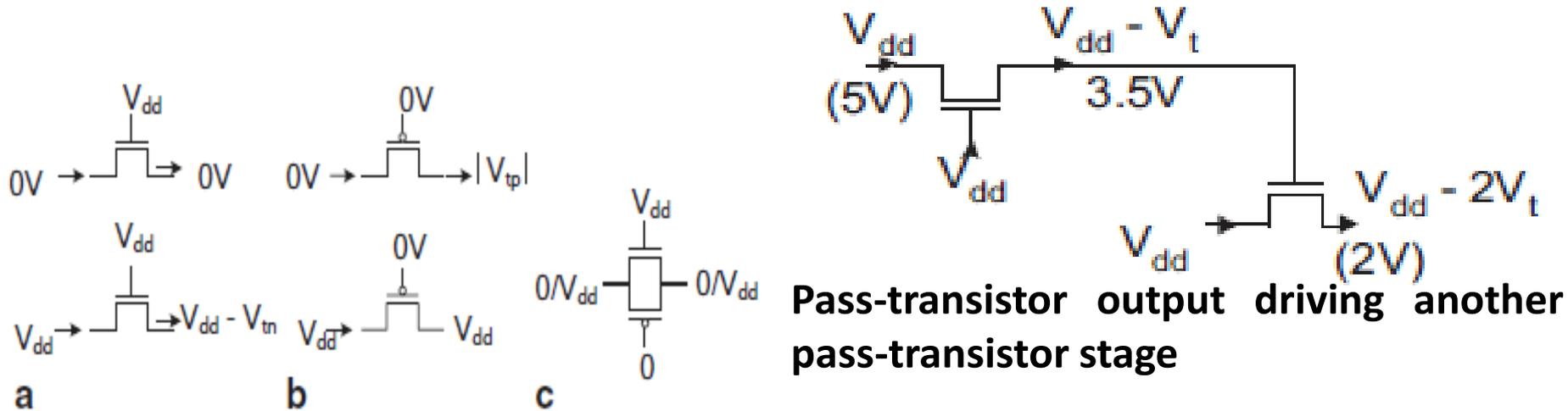
- As pass transistors functionally behave like **switches, logic functions** can be realized using pass transistors in a manner similar to **relay contact networks**.
- **Relay contact networks** have been presented in the classical text of Caldwell (1958).
- However, a closer look will reveal that there are some basic differences between **relay circuits and pass-transistor networks**.
- Some of the important differences are discussed in this section [1].
- In our discussion, we assume that the pass transistors are nMOS devices.

Pass-Transistor Logic



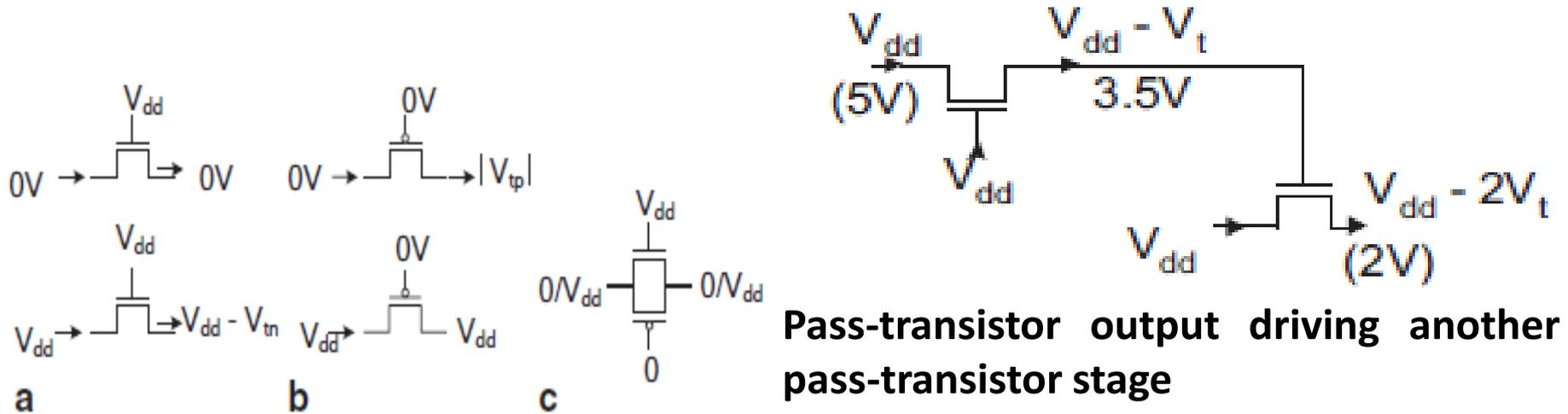
- ✓ A high-level signal gets degraded as it is steered through an nMOS pass transistor.
- ✓ For example, if both drain and gate are at some high voltage level, the source will rise to the lower of the two potentials: V_{dd} and $(V_{gs} - V_t)$.
- ✓ If both drain and gate are at V_{dd} , the source voltage cannot exceed $(V_{dd} - V_t)$.

Pass-Transistor Logic



➤ We have already seen that when a high-level signal is steered through a pass transistor and applied to an inverter gate, the inverter has to be designed with a high inverter ratio (8:1) such that the gate drive (3.5 V assuming the threshold voltage equal to 1.5 V) is sufficient enough to drive the inverter output to an acceptable low level.

Pass-Transistor Logic



□ This effect becomes more prominent when a pass-transistor output is allowed to drive the gate of another pass-transistor stage as shown in Fig. 5.1.

□ As the gate voltage is 3.5 V, the high-level output from the second pass transistor can never exceed 2.0 V, even when the drain voltage is 5 V as shown in Fig. 5.1.

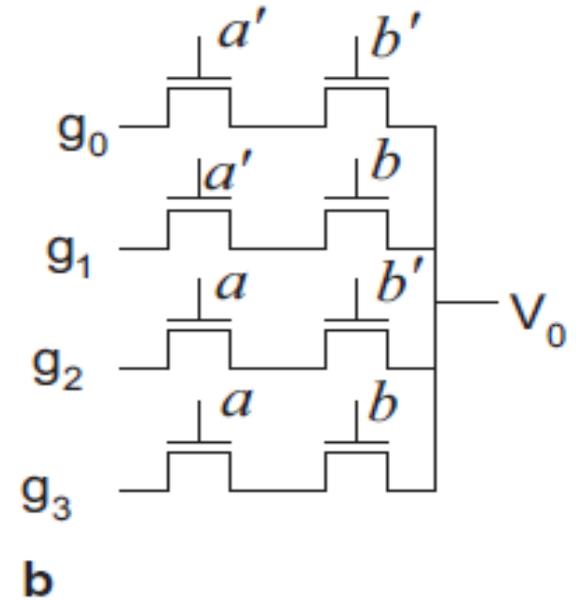
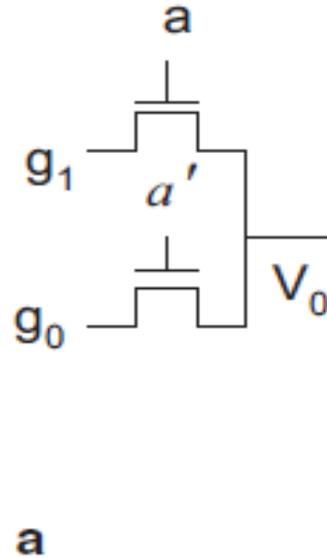
Pass-Transistor Logic

- Finally, care should be taken such that advertently or inadvertently an **output point is not driven by signals of opposite polarity.**
- In such a situation, each transistor acts as a **resistor** and a **voltage about half of the high level is generated at the output.**
- Presence of this type of path is known as ***sneak path.***
- *In such a situation, an undefined voltage level is generated at a particular node.*
- This should be avoided.

Realizing Pass-Transistor Logic

➤ Pass transistors can be used to realize multiplexers of different sizes.

- (a) 2-to-1 multiplexer.
- (b) 4-to-1 multiplexer circuit using pass-transistor network

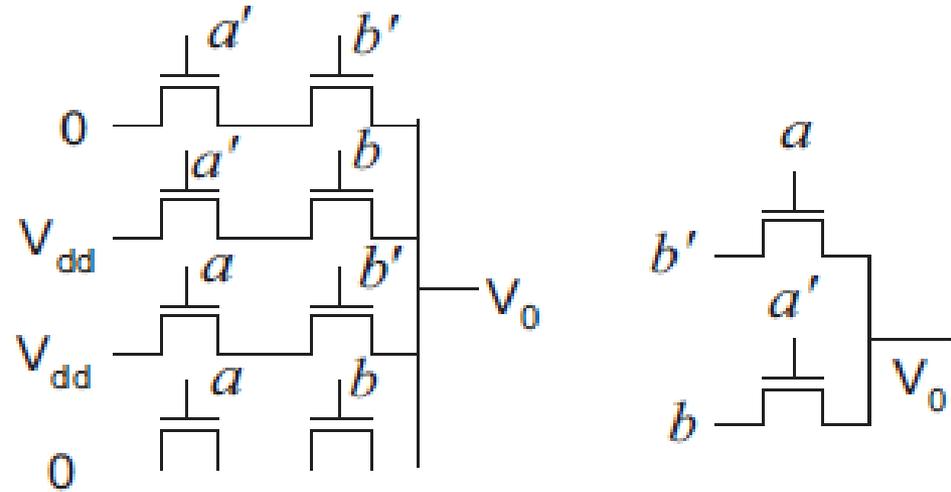


$$f(a,b) = g_0 a'b' + g_1 a'b + g_2 ab' + g_3 ab.$$

Realizing Pass-Transistor Logic

(a) Multiplexer realization of $f = a'b + ab'$.

(b) Minimum transistor pass-transistor realization of $f = a'b + ab'$



- ✓ This approach may be termed as universal logic module (ULM)-based approach because of the ability of multiplexers to realize any function up to a certain number of input variables.
- ✓ However, the circuit realized in the above manner may not be optimal in terms of the number of pass transistors.

Realizing Pass-Transistor Logic- *Advantages*

❑ **A. Ratioless:**

❑ For a Reliable operation of an **inverter (or gate logic)**, the **width/length (W/L) ratio** of the **pull-up device** is **four times (or more)** that of the **pull-down device**.

❑ As a result, the geometrical dimension of the transistors is not minimum (i.e., 2×2).

❑ The pass transistors, on the other hand, can be of minimum dimension.

❑ This makes pass-transistor circuit realization very area Efficient.

Realizing Pass-Transistor Logic- *Advantages*

b. Powerless:

- ❑ In a pass-transistor circuit, there is no direct current (DC) path from supply to ground (GND).
- ❑ So, it does not require any standby power, and power dissipation is very small.
- ❑ Moreover, each additional input requires only a minimum geometry transistor and adds no power dissipation to the circuit.

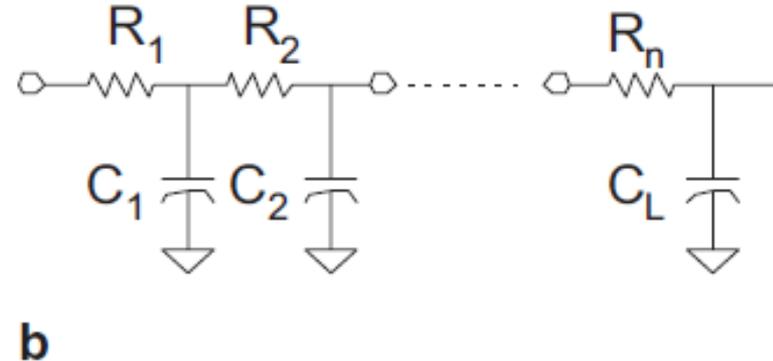
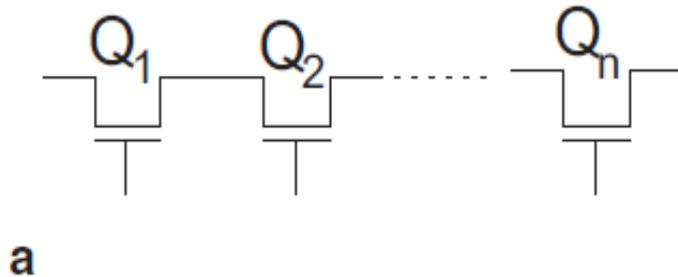
c. Lower area:

- ✓ Any one type of pass-transistor networks, nMOS or p-type MOS (pMOS), is sufficient for the logic function realization.
- ✓ This results in a smaller number of transistors and smaller input loads.
- ✓ This, in turn, leads to smaller chip area, lower delay, and smaller power consumption.

Realizing Pass-Transistor Logic- *Disadvantages*

1. When a signal is steered through several stages of pass transistors, the delay can be considerable.

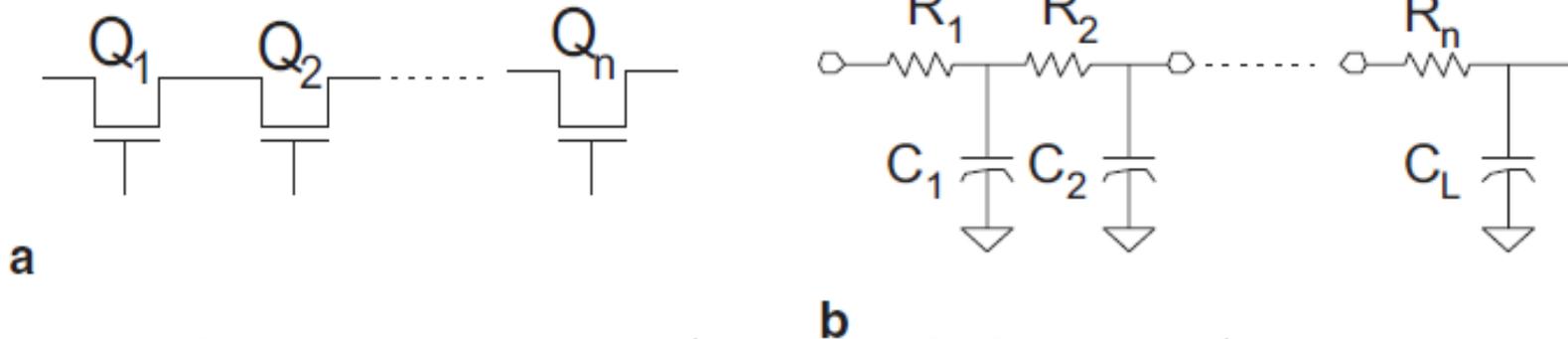
(a) Pass-transistor network. (b) RC model for the pass-transistor network.



1. The ON resistance of each pass transistor is R_{pass} and capacitance C_L .
2. The value of R_{pass} and C_L depends on the type of switch used.
3. The **time constant** $R_{pass} C_L$ approximately gives the time constant corresponding to the time for C_L to charge to 63 % of its final value.

Realizing Pass-Transistor Logic- *Disadvantages*

(a) Pass-transistor network. (b) RC model for the pass-transistor network.



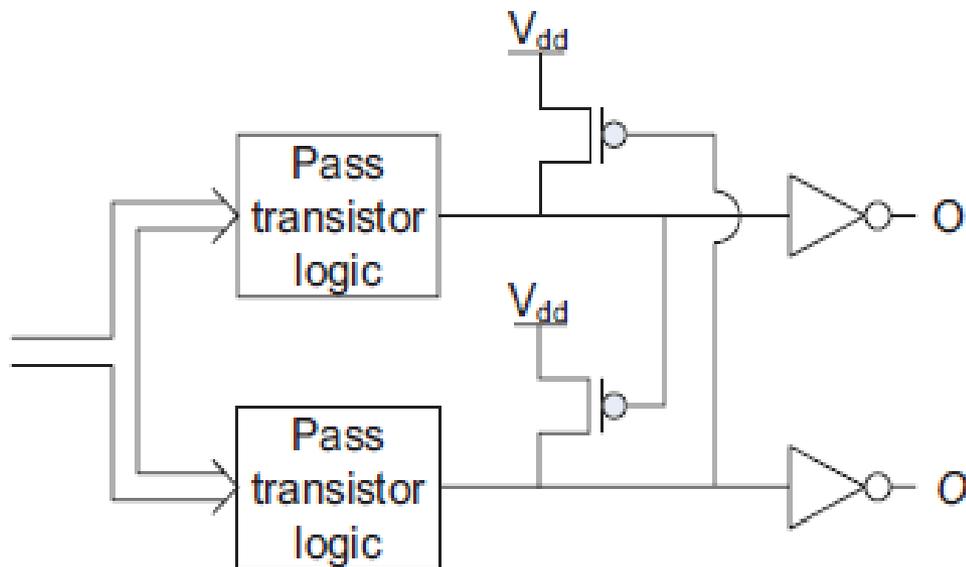
1. To calculate the delay of a cascaded stage of n transistors, we can simplify the equivalent circuit by assuming that all resistances and capacitances are equal and can be lumped together into one resistor of value $n \times R_{pass}$ and one capacitor of value $n \times C_L$.
2. The equivalent time constant is $n^2 R_{pass} C_L$.

Pass-Transistor Logic Families

- ❖ Conventional nMOS Pass-transistor logic or Complementary Pass-transistor Logic (CPL)
- ❖ Dual Pass-transistor Logic (DPL)
- ❖ Swing-Restored Pass-Transistor Logic (SRPL)

Complementary Pass-transistor Logic (CPL)

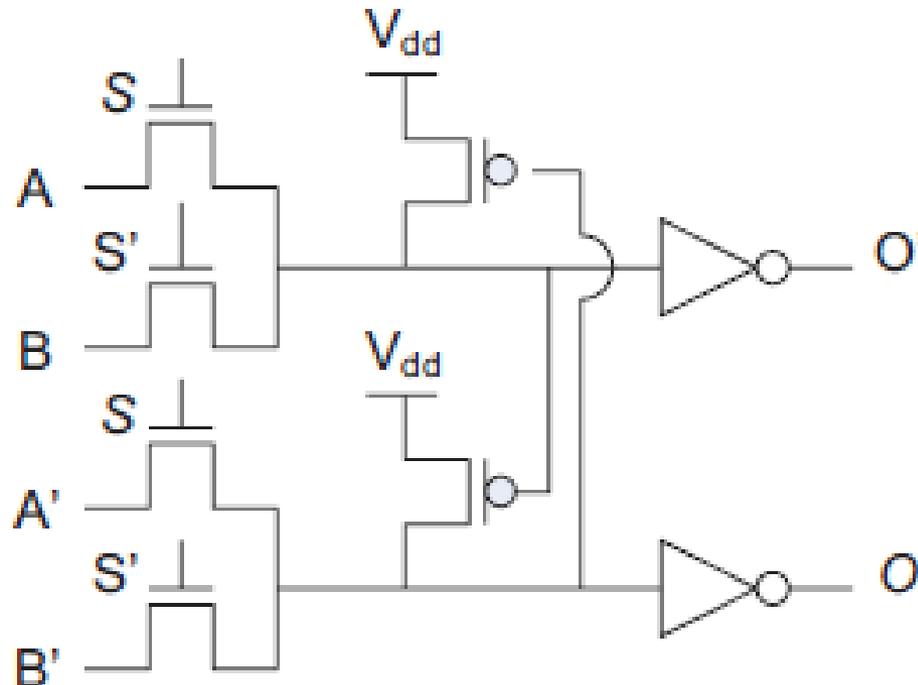
- ✓ **Two nMOS pass-transistor logic network** (one for each rail)
- ✓ **Two small pull-up pMOS transistors** for **swing restoration**
- ✓ **Two output inverters** for the **complementary output signals**.



Basic complementary pass-transistor logic (CPL) structure

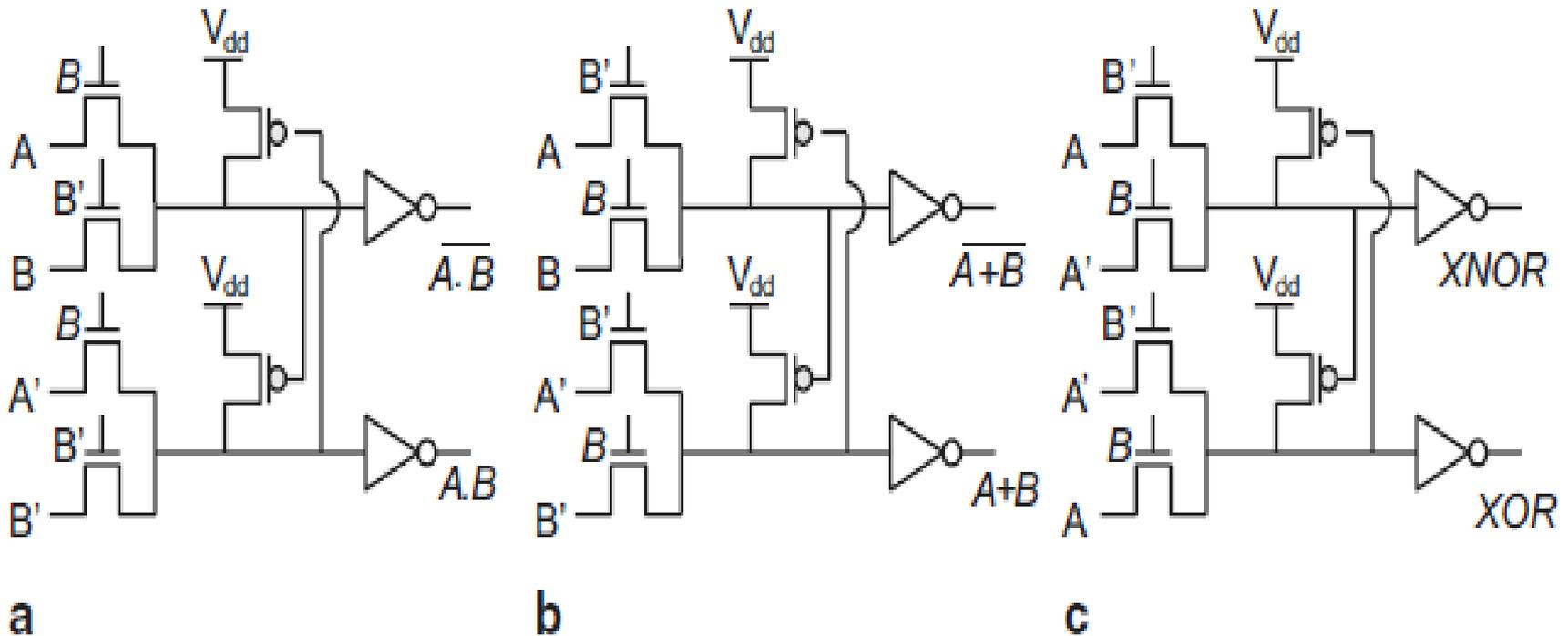
Complementary Pass-transistor Logic (CPL)

- ❖ 2-to-1 Multiplexer Realization using CPL Logic
- ❖ This is the basic and minimal gate structure with ten transistors.
- ❖ All two-input functions can be implemented by this basic gate structure.



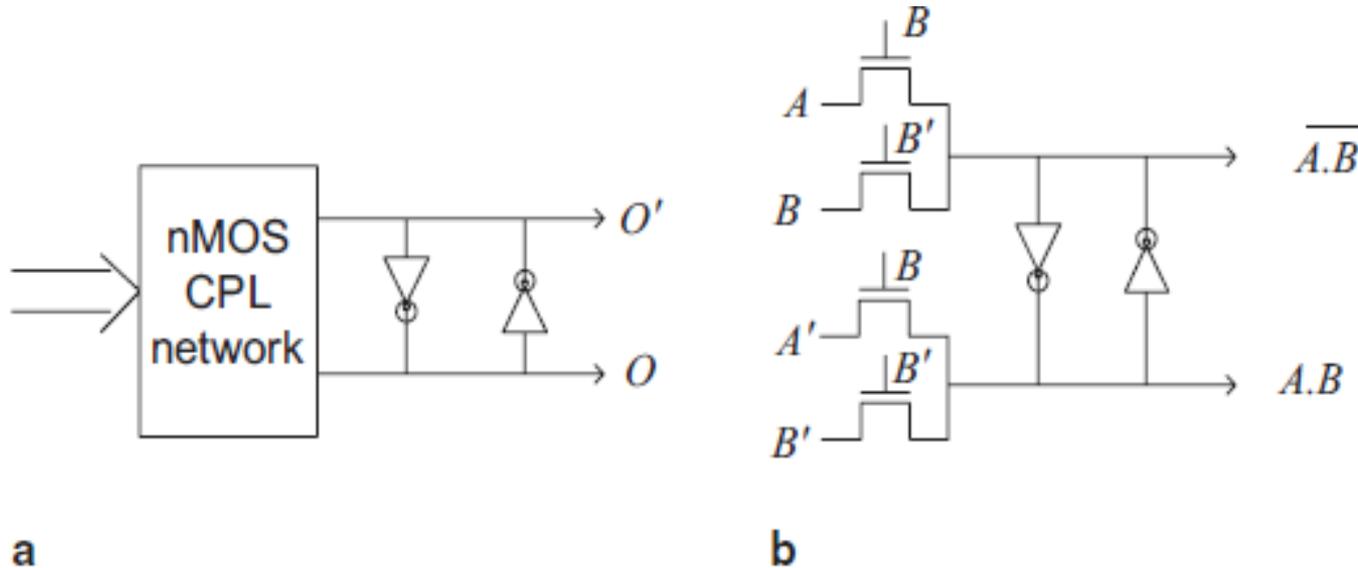
Complementary Pass-transistor Logic (CPL)

Complementary pass-transistor logic (CPL) logic circuit for (a) 2-input AND/NAND, (b) 2-input OR/NOR, and (c) 2-input EX-OR gate structure.



Swing-Restored Pass-Transistor Logic

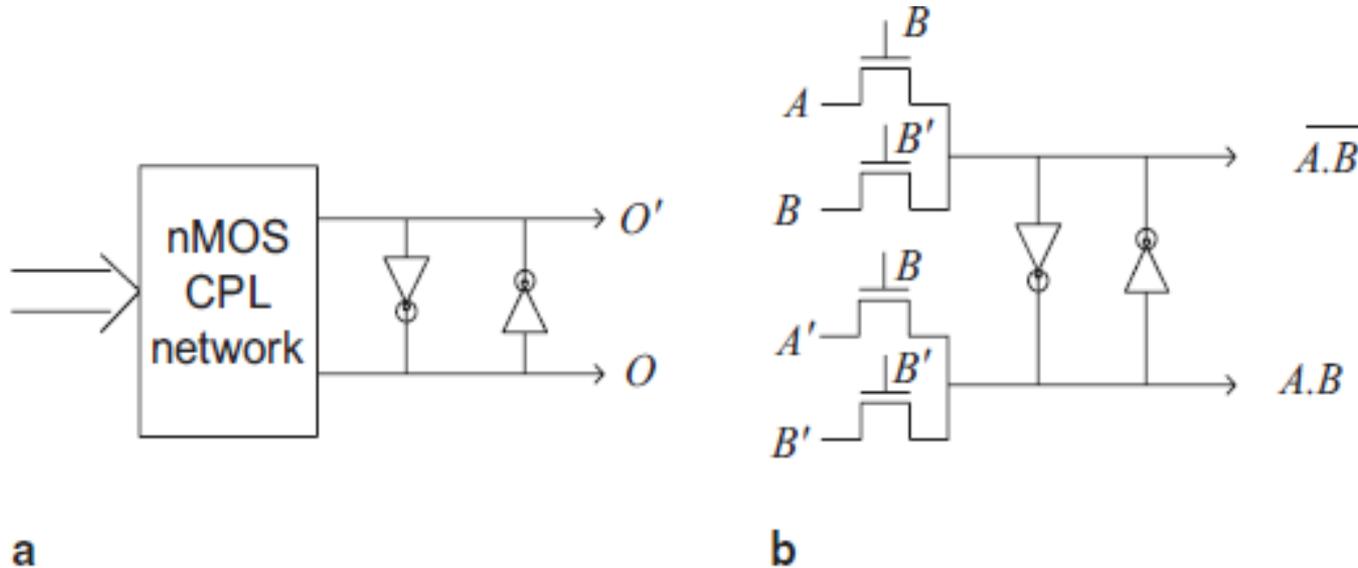
(a) Basic swing-restored pass-transistor logic (SRPL) configuration; and (b) SRPL realization of 2-input NAND gate



- ❑ The SRPL logic style is an **extension of CPL** to make it suitable for **low-power low-voltage applications**.
- ❑ Output inverters are cross-coupled with a latch structure, which performs both **swing restoration and output buffering**.

Swing-Restored Pass-Transistor Logic

(a) Basic swing-restored pass-transistor logic (SRPL) configuration; and (b) SRPL realization of 2-input NAND gate



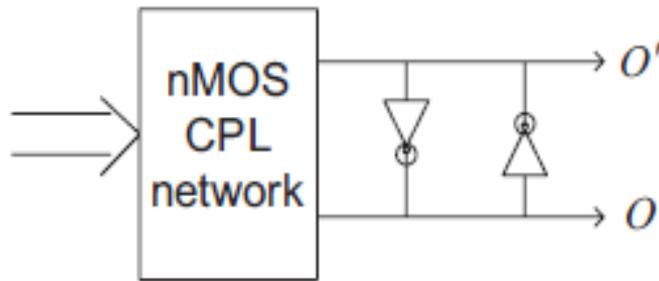
❑ The SRPL logic style is an **extension of CPL** to make it suitable for **low-power low-voltage applications**.

❑ Output inverters are cross-coupled with a latch structure, which performs both **swing restoration and output buffering**.

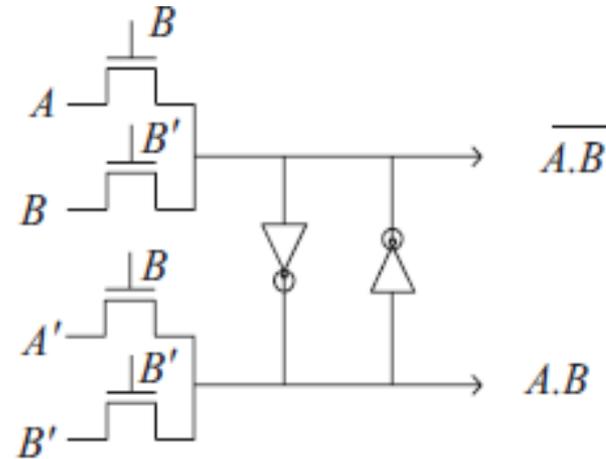
❑ The **pull-up pMOS transistors** are not required anymore and that the output nodes of the nMOS networks are the gate outputs.

Swing-Restored Pass-Transistor Logic

(a) Basic swing-restored pass-transistor logic (SRPL) configuration; and (b) SRPL realization of 2-input NAND gate



a

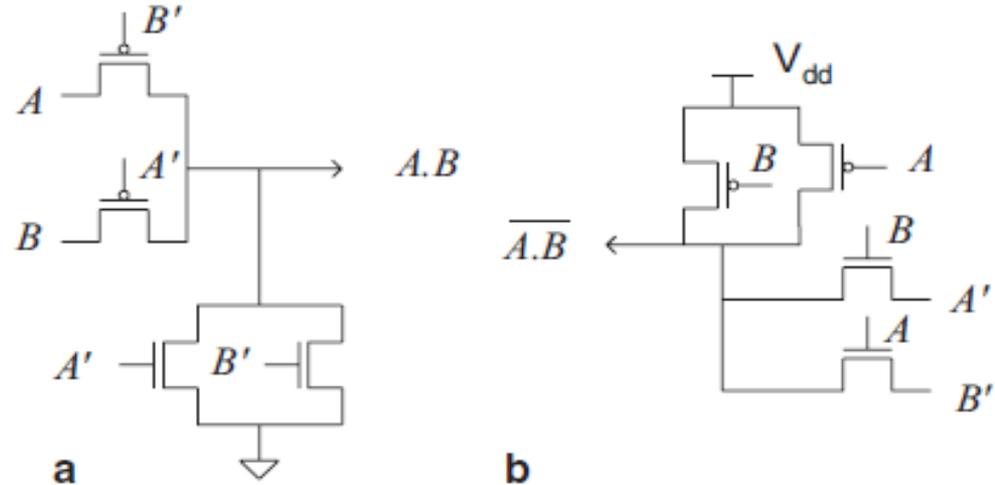


b

- ❑ The SRPL logic style is an **extension of CPL** to make it suitable for **low-power low-voltage applications**.
- ❑ Output inverters are cross-coupled with a latch structure, which performs both **swing restoration and output buffering**.
- ❑ The **pull-up pMOS transistors** are not required anymore and that the output nodes of the nMOS networks are the gate outputs.

Double Pass-Transistor Logic

Double pass-transistor logic (DPL) realization of 2-input AND/NAND function

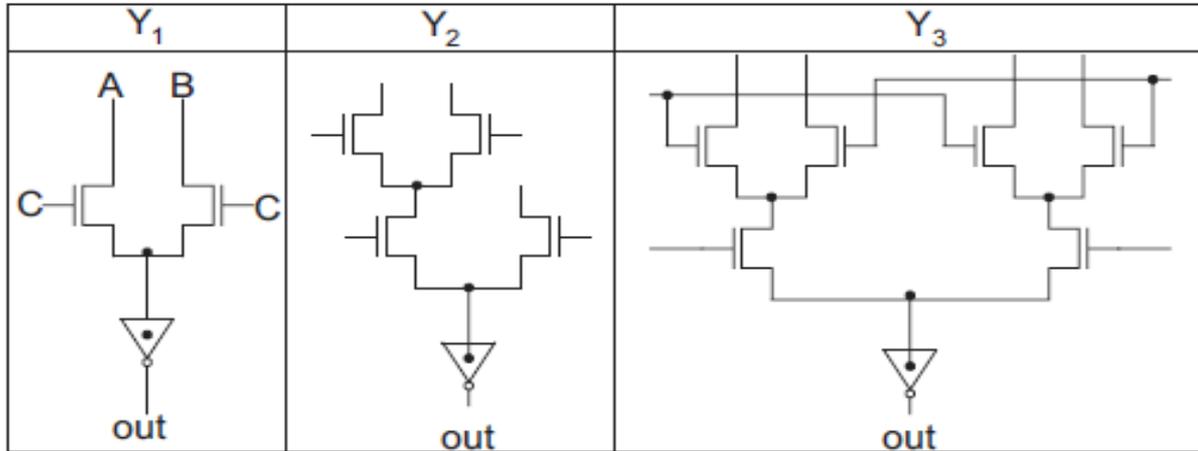


❖ The DPL has a **balanced input capacitance**, and the delay is independent of the input delay contrary to the CPL and conventional CMOS pass-transistor logic, where the input capacitance for different signal inputs is the same.

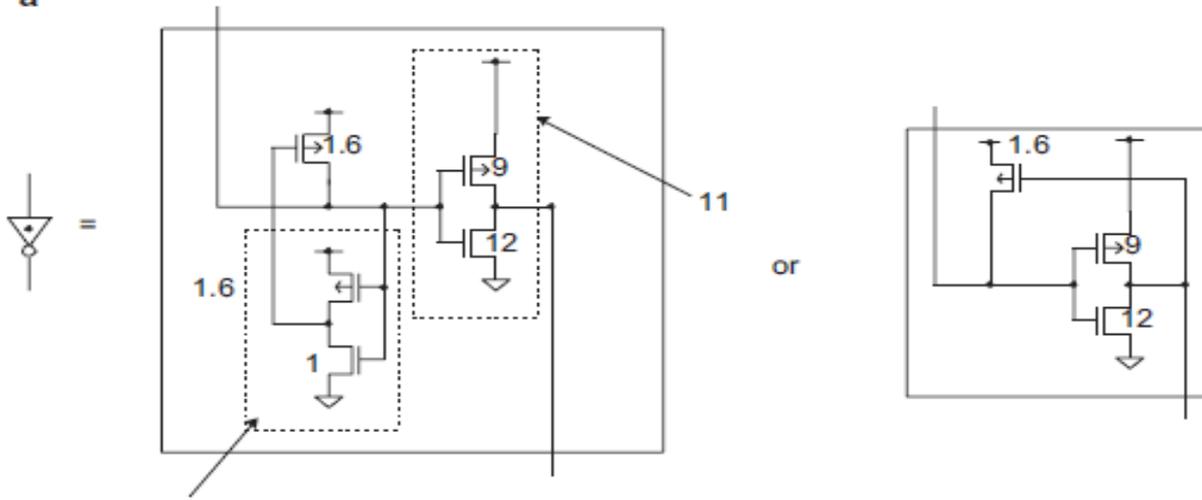
❖ As **two current paths** always drive the output in DPL, the **reduction in speed** due to the additional transistor is compensated.

Single-Rail Pass-Transistor Logic

Single-rail pass-transistor logic (LEAP) cells



a

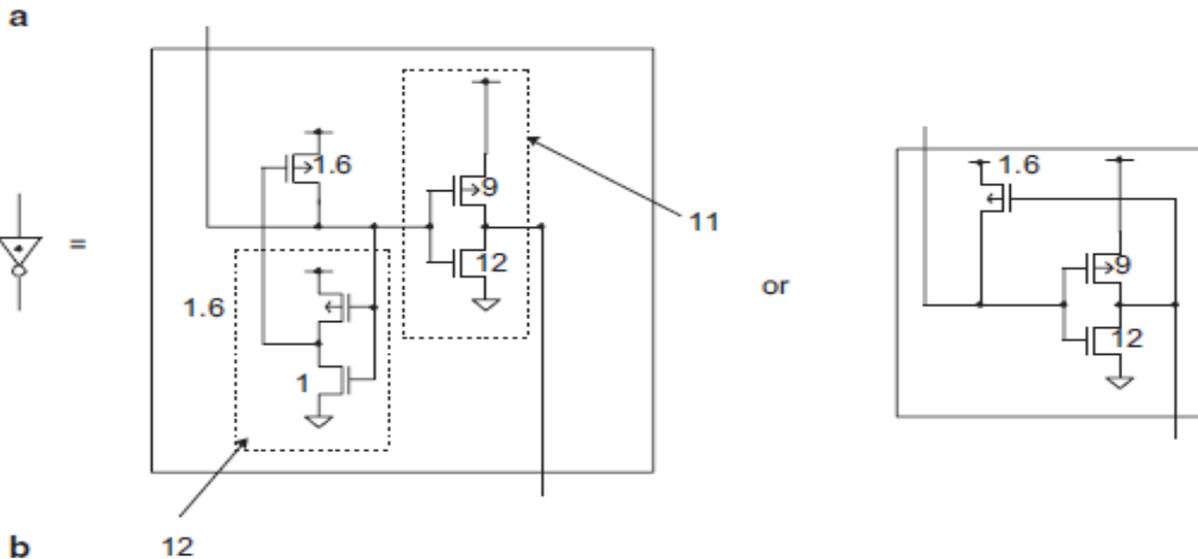
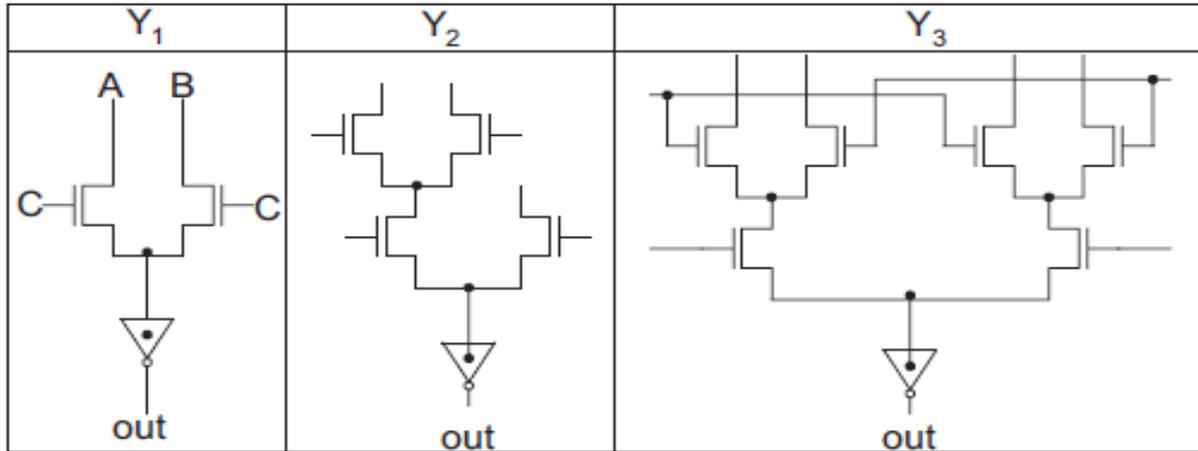


b

A single-rail pass-transistor logic style has been adopted in the single-rail pass-transistor logic **(LEAP; LEAn integration with pass transistors)** design which exploits the full functionality of the **multiplexer structure scheme.**

Single-Rail Pass-Transistor Logic

Single-rail pass-transistor logic (LEAP) cells



- Swing restoration is done by a feedback pull-up pMOS transistor.
- This is slower than the cross-coupled pMOS transistors of CPL working in differential mode.

Comparisons of the Logic Styles

Table 5.1 Qualitative comparisons of the logic styles

Logic style	#MOS networks	Output driving	I/O decoupling	Swing restoration	# Rails	Robustness
CMOS	$2n$	Med/good	Yes	No	Single	High
CPL	$2n+6$	Good	Yes	Yes	Dual	Medium
SRPL	$2n+4$	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

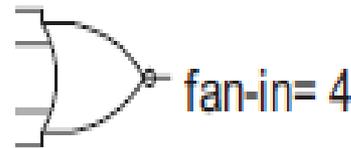
MOS metal–oxide–semiconductor, *I/O* input/output, *CMOS* complementary metal–oxide–semiconductor, *CPL* complementary pass-transistor logic, *SRPL* swing-restored pass-transistor logic, *DPL* double pass-transistor logic, *LEAP* single-rail pass-transistor logic, *DCVSPG* differential cascode voltage switch pass gate

Gate Logic

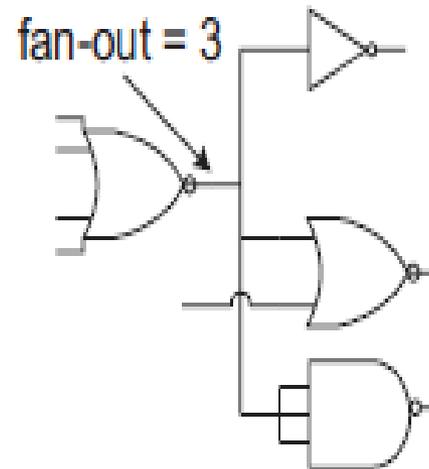
- *Fan-In and Fan-Out*
- *nMOS NAND and NOR Gates*
- *CMOS Realization*
 - CMOS NAND Gates
 - CMOS NOR Gates
- *Switching Characteristics*
- *CMOS NOR Gate*
- *CMOS Complex Logic Gates*

Gate *Logic-Fan-In and Fan-Out*

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



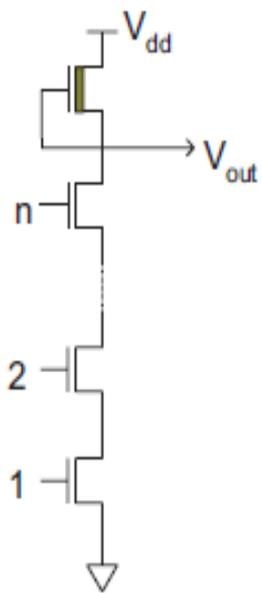
a



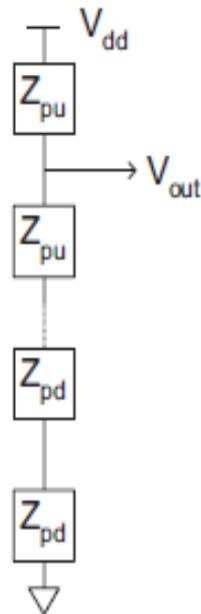
b

Gate Logic- *nMOS NAND and NOR Gates*

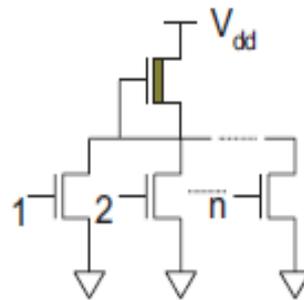
- (a) *n*-input *nMOS* NAND gate; (b) equivalent circuits; and (c) *n*-input *nMOS* NOR gate.



a



b



c

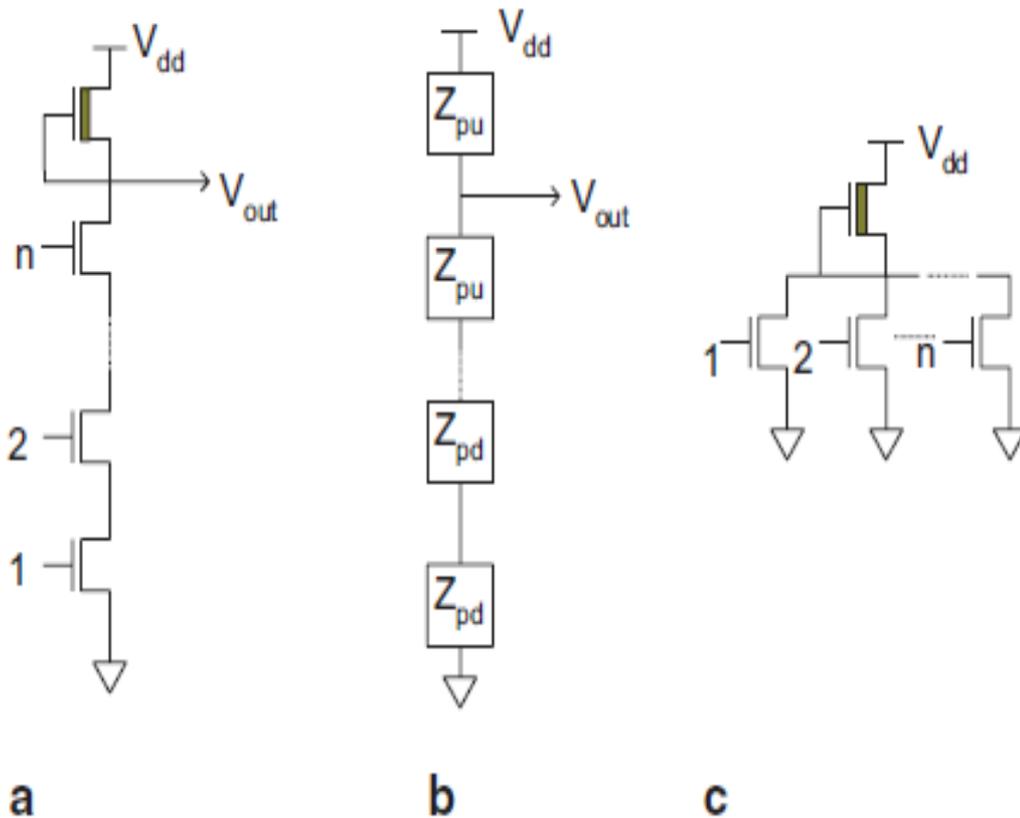
➤ Realizations of *nMOS* NAND gate with two or more inputs.

➤ Let us consider the generalized realization with *n* inputs with a depletion-type *nMOS* transistor as a pull-up device and *n* enhancement-type *nMOS* transistors as pull-down devices.

➤ In this kind of realization, the length/width (L/W) ratio of the pull-up and pull-down transistors should be carefully chosen such that the desired logic levels are maintained.

Gate Logic- *nMOS NAND and NOR Gates*

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



❑ The critical factor here is the low-level output voltage, which should be sufficiently low such that it turns off the transistors of the following stages.

❑ To satisfy this, the output voltage should be less than the threshold voltage, i.e., $V_{out} \leq V_t = 0.2 V_{dd}$.

❑ To get a low-level output, all the pull-down transistors must be ON to provide the GND path.

Gate Logic- *nMOS NAND and NOR Gates*

- It may be noted that, not only one pull-down transistor is required per input of the NAND gate stage but also the size of the pull-up transistor has to be adjusted to maintain the required overall ratio.
- This requires a considerably larger area than those of the corresponding nMOS inverter.
- Moreover, the delay increases in direct proportion to the number of inputs.
- If each pull-down transistor is kept of minimum size, then each will represent one gate capacitance at its input and resistance of all the pull-down transistors will be in series. Therefore, for an *n-input NAND gates*, we have a delay of *n-times that of an inverter*, i.e.,

Gate Logic- *nMOS NAND and NOR Gates*

- Therefore, for an *n-input NAND gates*, we have a **delay of *n-times* that of an inverter**, i.e.,

$$\tau_{\text{NAND}} = n\tau_{\text{inv}}$$

- As a consequence, nMOS NAND gates are **only used when absolutely necessary and the number of inputs is restricted** so that **area and delay** remain within **some limit**.

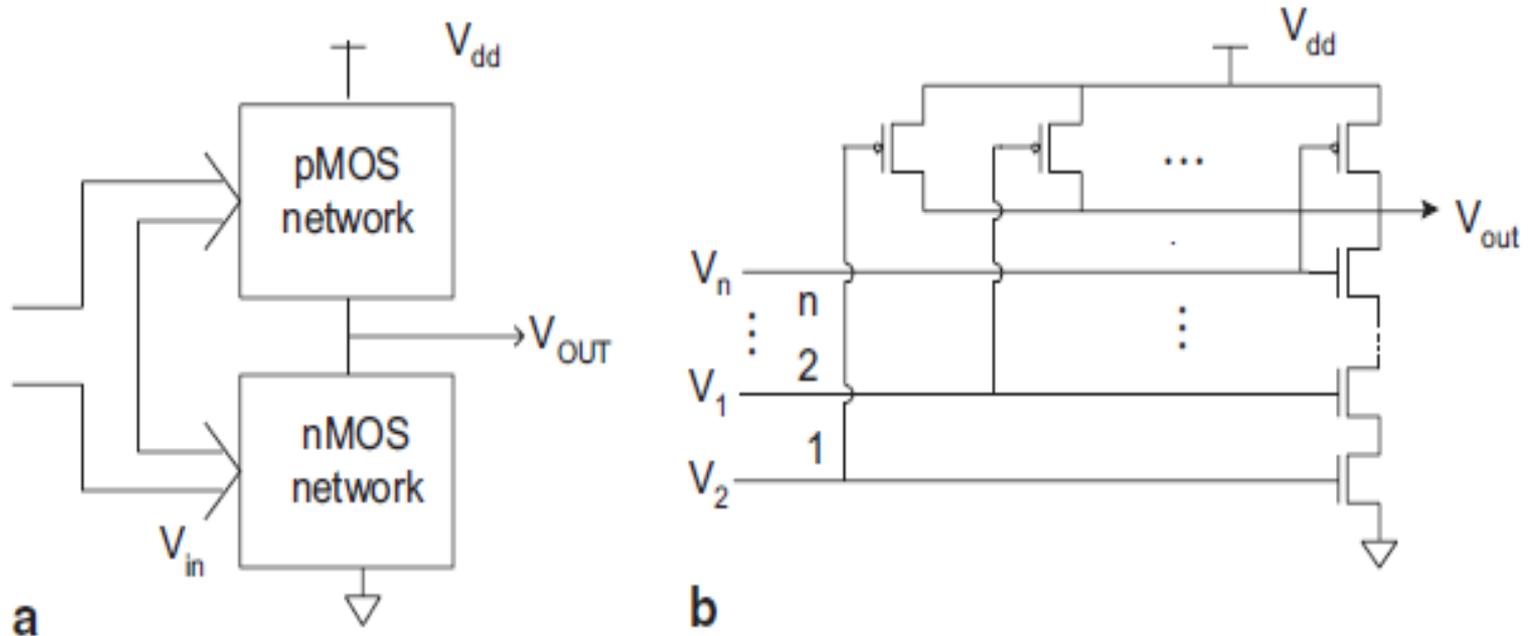
Gate Logic- *nMOS NAND and NOR Gates*

- ***n-input NOR***
- *Unlike the NAND gate, here the **output** is **low** when any one of the **pull-down transistors** is **ON**, as it happens in the case of an inverter.*
- *As a consequence, the **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.*

Gate Logic- *nMOS NAND and NOR Gates*

- The **area occupied** by the nMOS NOR gate is **reasonable**, because the pull-up transistor geometry is not affected by the number of inputs.
- The **worst-case delay** of an **NOR gate** is also **comparable** to the corresponding inverter.
- As a consequence, the use of **NOR gate** in circuit realization is **preferred** compared to that of **NAND gate**, when there is a choice.

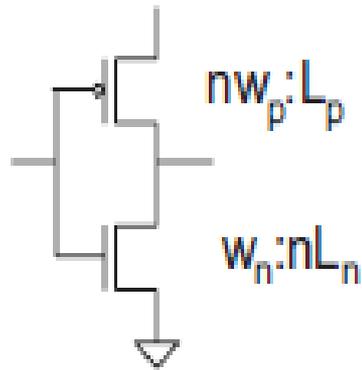
Gate Logic- *CMOS Realization*



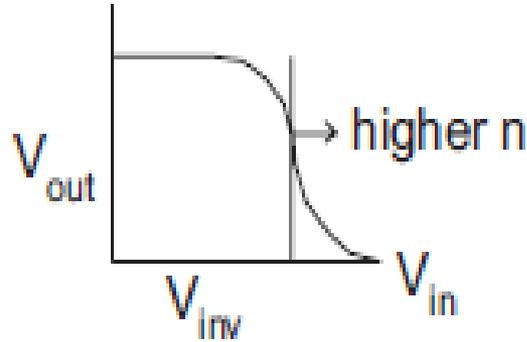
(a) General CMOS network; and (b) n -input CMOS NAND gate.

- It has fan-in of n having n number of nMOS transistor in series in the pull-down path and n number of pMOS transistors in parallel in the pull-up network.
- We assume that all the transistors are of the same size having a width $W = W_n = W_p$ and length $L = L_n = L_p$.
- If all inputs are tied together, it behaves like an inverter.

Gate Logic- *CMOS Realization*



a



b

(a) Equivalent circuit of n-input complementary MOS (CMOS) NAND gate; and

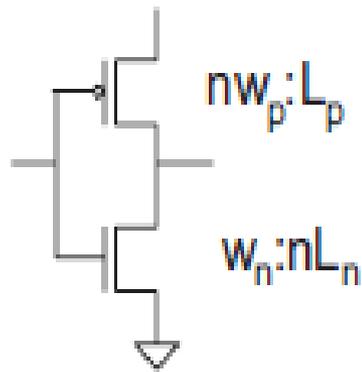
(b) Transfer characteristics of n-input CMOS NAND gate

✓ To determine the inversion point, pMOS transistors in parallel may be equated to a single transistor with the width *n times that of a single transistor*.

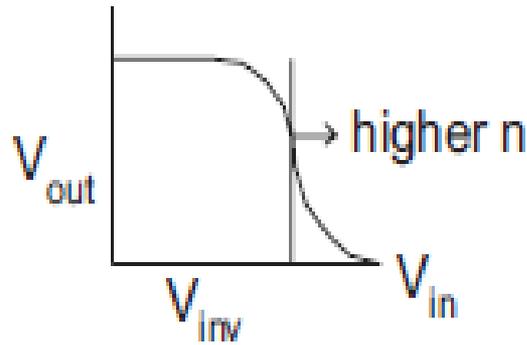
✓ And nMOS transistors in series may be considered to be equivalent to have a length equal to *n times that of a single transistor*

✓ This makes the trans-conductance ratio $= \beta_n / \beta_p n^2$.

Gate Logic- *CMOS Realization*



a



b

(a) Equivalent circuit of n-input complementary MOS (CMOS) NAND gate; and

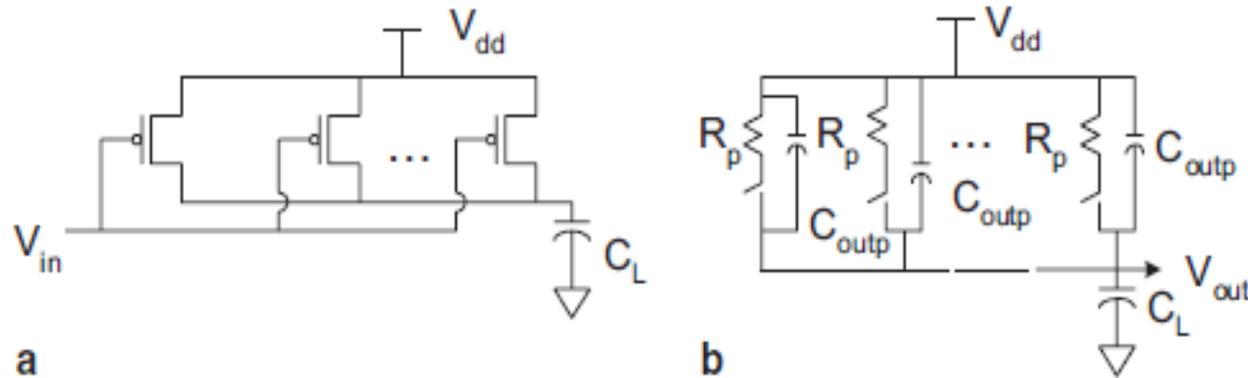
(b) Transfer characteristics of n-input CMOS NAND gate

✓ This makes the trans-conductance ratio = $\beta_n / \beta_p n^2$

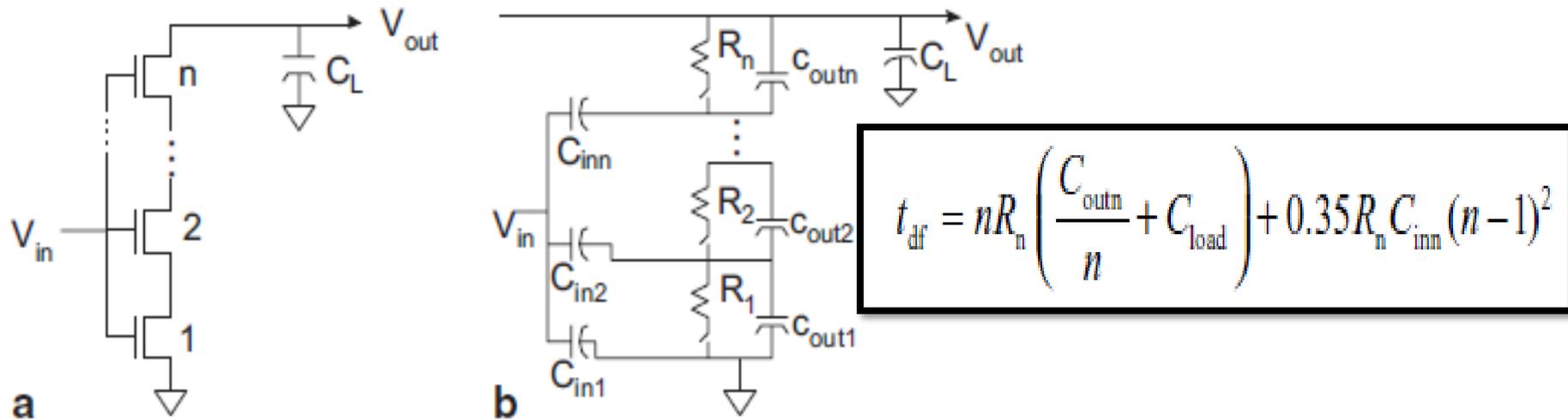
$$V_{inv} = \frac{V_{dd} + V_{tp} + V_{in} \sqrt{\frac{\beta_n}{n^2 \beta_p}}}{1 + \sqrt{\frac{\beta_n}{n^2 \beta_p}}} \quad V_{inv} = \frac{V_{dd} + V_{tp} + \frac{V_{th}}{n} \sqrt{\frac{\beta_n}{\beta_p}}}{1 + \frac{1}{n} \sqrt{\frac{\beta_n}{\beta_p}}}$$

1. As the fan-in number n increases, the β_n / β_p ratio (trans-conductance ratio) decreases leading to increase in V_{inv} .
2. In other words, with the increase in fan-in, the switching point (inversion point) moves towards the right.
3. It may be noted that in our analysis, we have ignored the body effect.

Gate Logic- *Switching Characteristics*



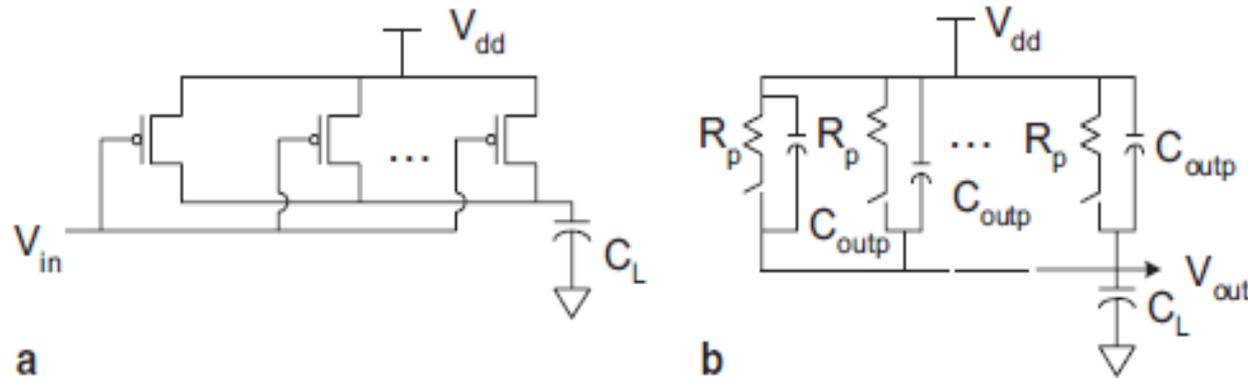
a Pull-up transistor tied together with a load capacitance; and b equivalent circuit



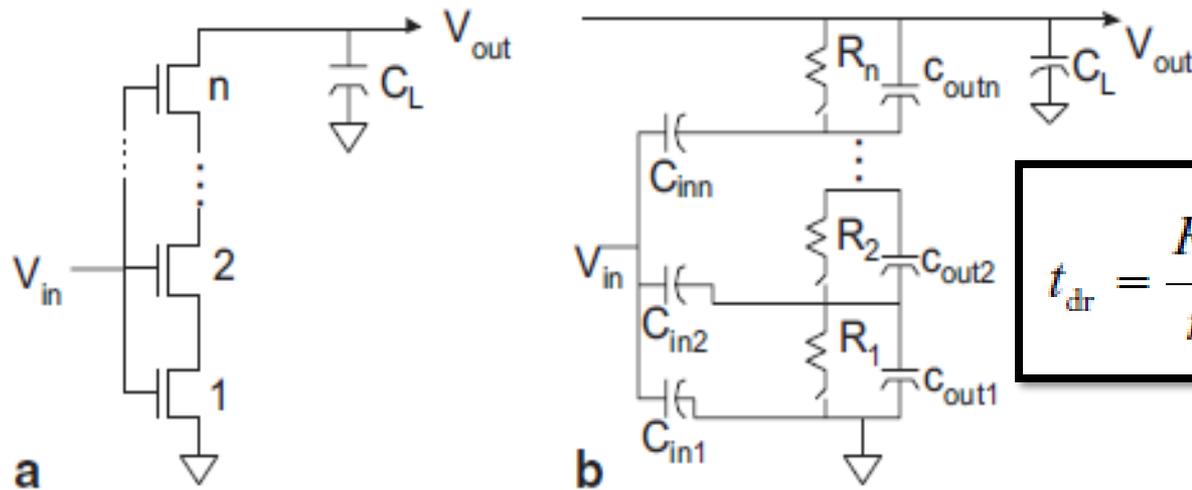
$$t_{df} = nR_n \left(\frac{C_{outn}}{n} + C_{load} \right) + 0.35R_n C_{inn} (n-1)^2$$

a Pull-down transistors along with load capacitance C_L , and b equivalent circuit

Gate Logic- *Switching Characteristics*



a Pull-up transistor tied together with a load capacitance; and b equivalent circuit



a Pull-down transistors along with load capacitance C_L , and b equivalent circuit

$$t_{dr} = \frac{R_p}{n} \left(nC_{outp} + \frac{C_{outn}}{n} + C_{load} \right)$$

Gate Logic- CMOS NOR Gate

In a similar manner, the rise and fall times of an n -input NOR gate can be obtained as

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

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

It may be noted that in case of NAND gate, the discharge path is through the series-connected-nMOS transistors.

As a consequence, the high-to-low delay increases with the number of fan-in.

If the output load capacitance is considerably larger than other capacitances then the fall time reduces to $t_{df} = nR_n C_L$ and $t_{dr} = R_p C_L$

Gate Logic- CMOS NOR Gate

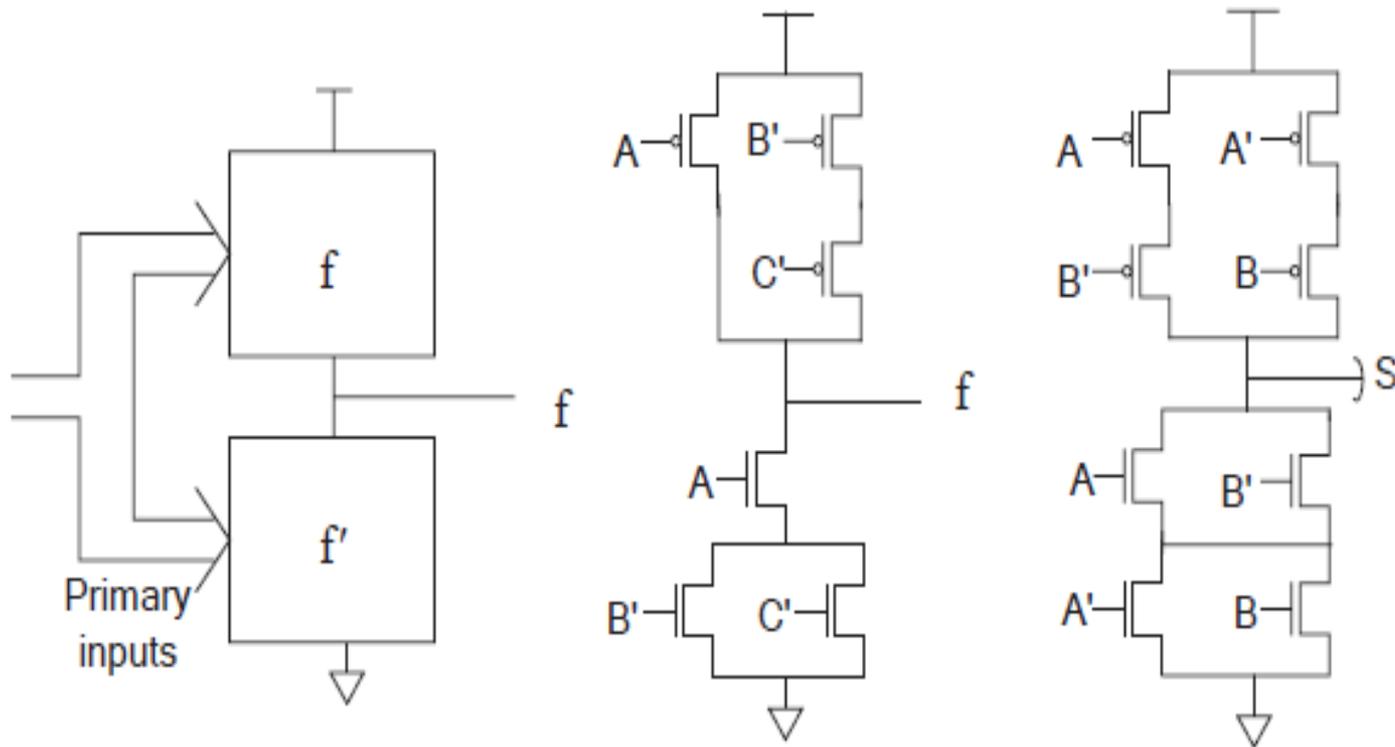
In a similar manner, the rise and fall times of an n -input NOR gate can be obtained as

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

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

1. On the other hand, in case of NOR gate the charging of the load capacitance take place through the series connected pMOS transistors giving rise time $t_{dr} = nR_p C_L$ and $t_{df} = R_n C_L$.
2. As $R_p > R_n$, the rise-time delay for NOR gates is more than the fall-time delay of NAND gates with the same fan-in.
3. This is the reason why NAND gates are generally a better choice than NOR gates in complementary CMOS logic.
4. NOR gates may be used for limited fan-out.

Gate Logic- CMOS Complex Logic Gates



a Realization of a function f by complementary MOS (CMOS) gate; b realization of $f = A' + BC$; and c realization of $S = A \oplus B$

$$f = A' + BC \Rightarrow f' = (A' + BC)' = A(B' + C').$$

$$S = A \oplus B = A'B + AB' \Rightarrow S' = (A'B + AB')' = (A + B')(A' + B).$$

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.

MOS Dynamic Circuits

- 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).

MOS Dynamic Circuits

- In the static combinational circuits, capacitance is regarded as a parameter responsible for **poor performance** of the circuit, and therefore considered as an undesirable circuit element.
- But, a capacitance has the important property of holding charge.
- Information can be stored in the form of charge.
- This information storage capability can be utilized to realize digital circuits.
- 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.

MOS Dynamic Circuits

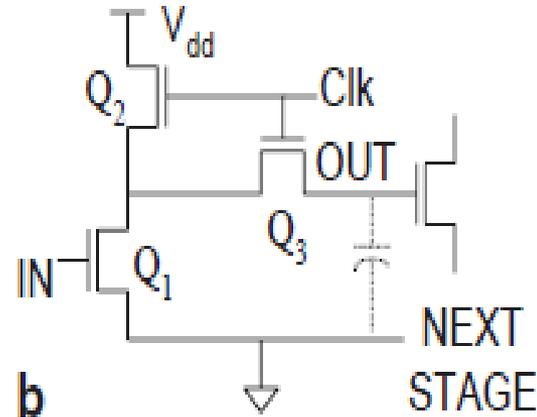
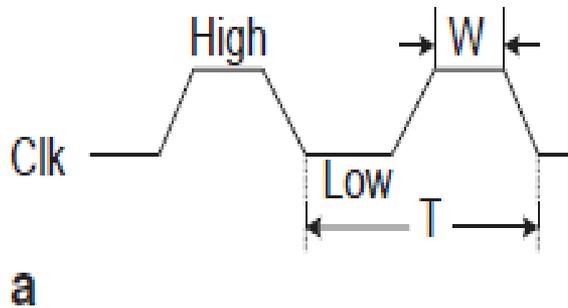
- The advantage of **low power** of full-complementary CMOS circuits and **smaller chip area** of nMOS circuits are combined in dynamic circuits leading to circuits of **smaller area and lower power dissipation**.
- MOS dynamic circuits are **also faster** in speed.
- However, these are not free from disadvantages.

MOS Dynamic Circuits

- *Single-Phase Dynamic Circuits*
- *Two-Phase Dynamic Circuits*
- *CMOS Dynamic Circuits*
- *Advantages and Disadvantages*
 - Charge Leakage Problem
 - Charge Sharing Problem
 - Clock Skew Problem
- *Domino CMOS Circuits*
- *NORA Logic*

MOS Dynamic Circuits

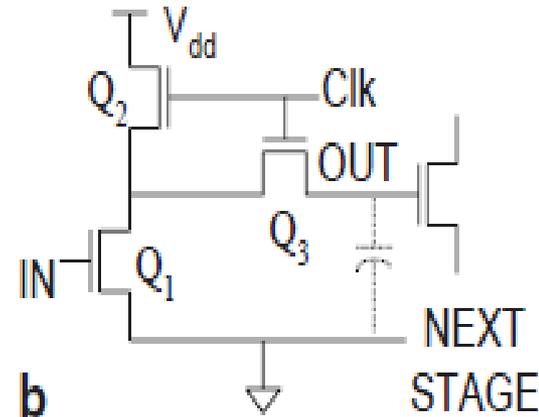
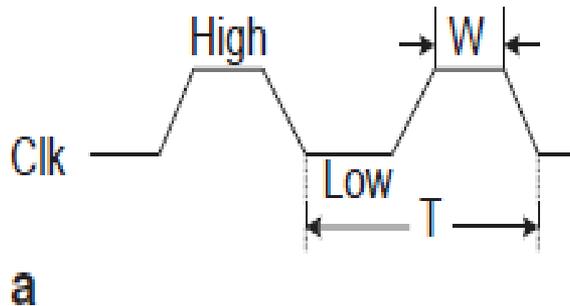
- *Single-Phase Dynamic Circuits*
- Two types of clocks are commonly used; *single-phase clock and nonoverlapping two-phase clock*



- The single-phase clock consists of a sequence of pulses having high (logic 1) and low (logic 0) levels with width W and time period T .
- A *single phase* clock has two states (low and high) and two edges per period.

MOS Dynamic Circuits

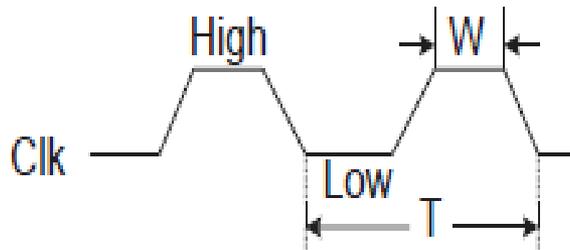
- *Single-Phase Dynamic Circuits*



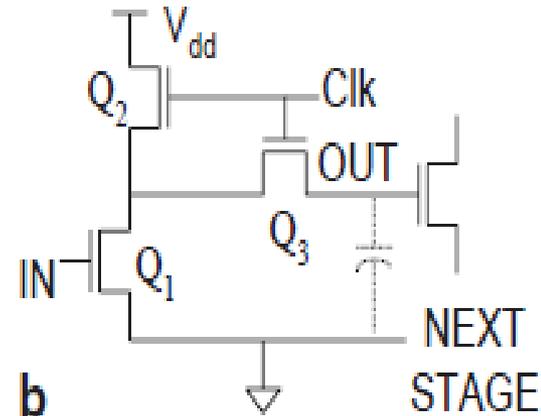
- ❖ When the clock is in the high state, both transistors Q_2 and Q_3 are ON.
- ❖ Depending on the input, Q_1 is either ON or OFF.
- ❖ If the input voltage is low, Q_1 is OFF and the output capacitor (gate capacitor of the next stage) charges to V_{dd} through Q_2 and Q_3 .

MOS Dynamic Circuits

- *Single-Phase Dynamic Circuits*



a



b

❖ When the input voltage is high, Q_1 is ON and the output is discharged through it to a low level.

MOS Dynamic Circuits

- *Single-Phase Dynamic Circuits*

- ❖ The circuits realized using the single-phase clocking scheme has the **disadvantage** that the **output voltage level** is dependent on the **inverter ratio** and the **number of transistors** in the current path to GND.

- ❖ In other words, single-phase dynamic circuits are **ratioed logic**.

- ❖ Moreover, as we have mentioned above, the circuit dissipates power when the output is low and the clock is high.

- ❖ Another problem arising out of single-phase clocked logic is known as *clock skew problem*.

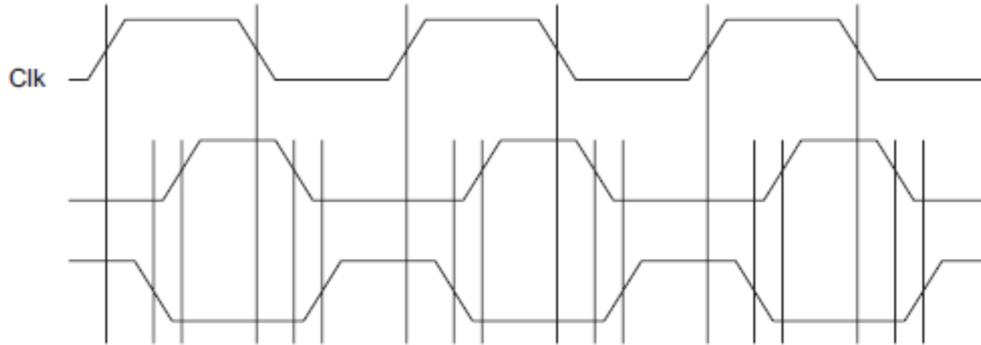
- ❖ *This is due to a delay in a clock signal during its journey through a number of circuit stages.*

- ❖ This results in undesired signals like glitch, hazards, etc.

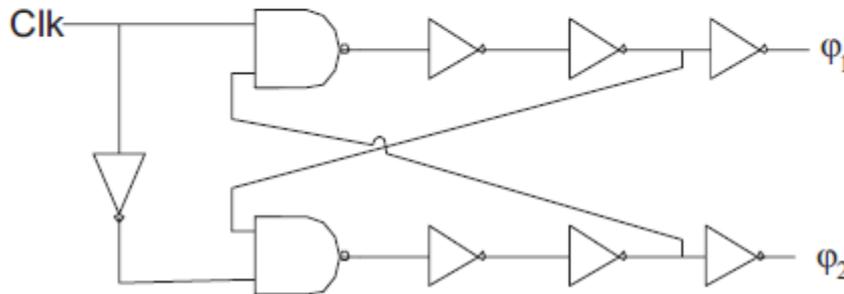
- ❖ Some of the problems can be overcome using two-phase clocking scheme as discussed in the following subsection.

MOS Dynamic Circuits

- *Two-Phase Dynamic Circuits*



a



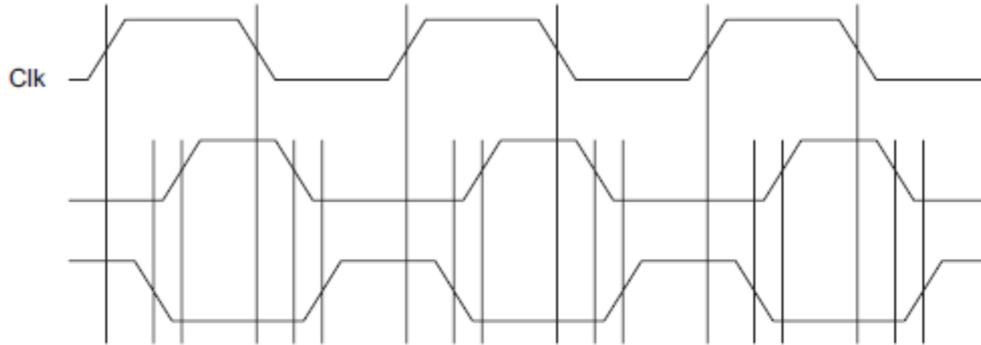
b

a Two-phase clock; and b a two-phase clock generator circuit

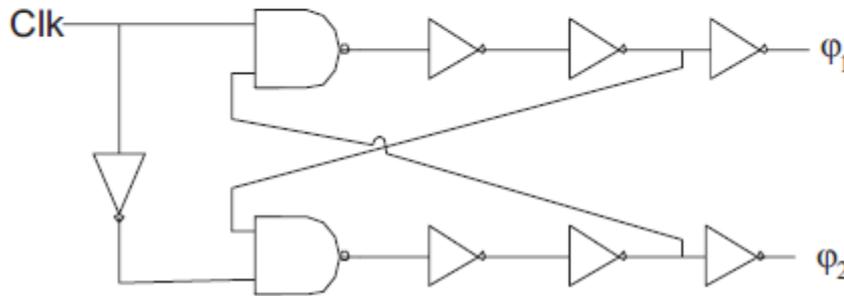
A two-phase nonoverlapping clock. As the two phases (ϕ_1 and ϕ_2) are never high simultaneously, the clock has three states and four edges and satisfies the property $\phi_1 \cdot \phi_2' = 0$.

MOS Dynamic Circuits

- ***Two-Phase Dynamic Circuits***



a



b

a Two-phase clock; and b a two-phase clock generator circuit

➤ The circuit takes a single-phase clock as an input and generates two-phase clock as the output.

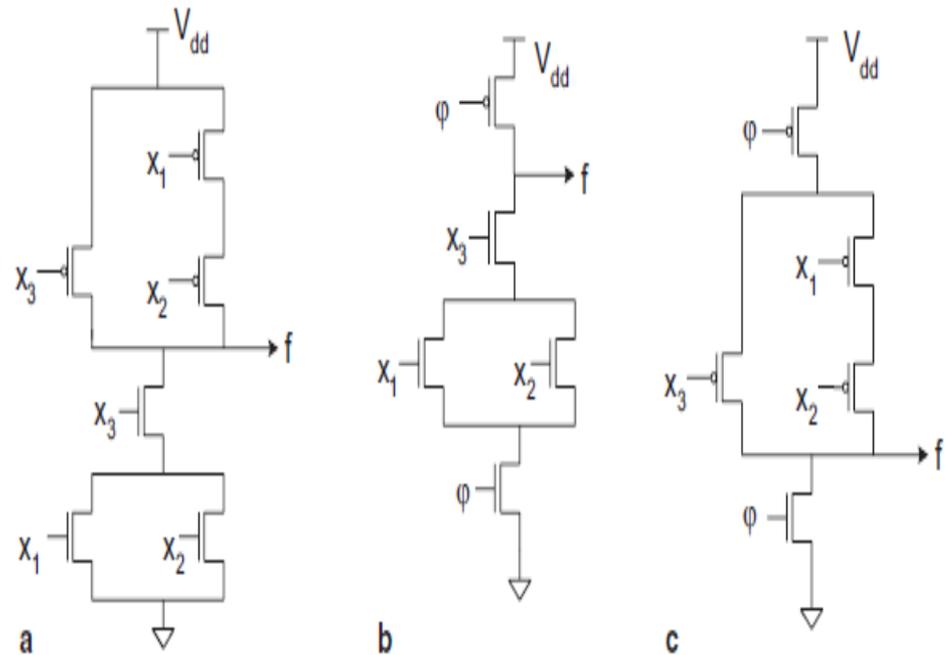
➤ The dead time, Δ , is decided by the delay through the NAND gates and the two inverters.

➤ As the clock (clk) signal goes low, ϕ_1 also goes low after four gate delays.

➤ ϕ_2 can go high after seven gate delays.

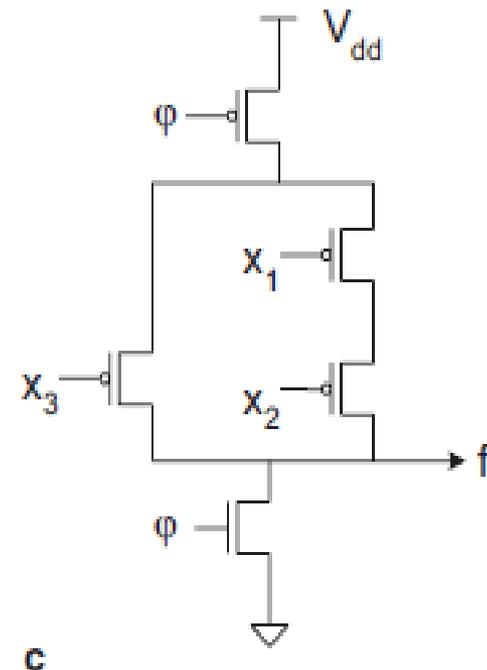
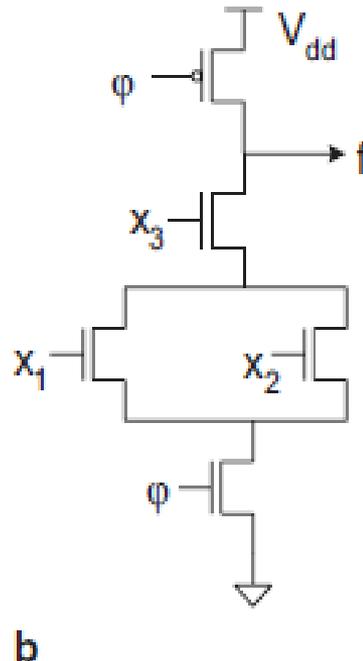
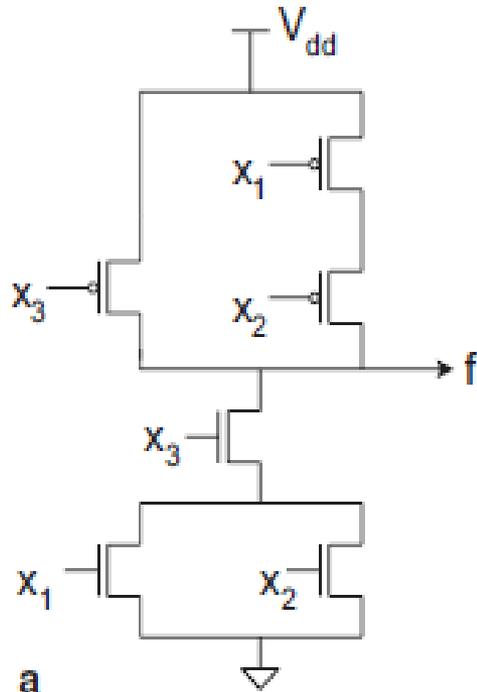
MOS Dynamic Circuits

- **CMOS Dynamic Circuits**
- Dynamic CMOS circuits avoid area penalty of static CMOS and at the same time retains the low-power dissipation of static CMOS, and they can be considered as extension of pseudo-nMOS circuits.
- The function of these circuits can be better explained using the idea of *pre-charged logic*.



MOS Dynamic Circuits

- *CMOS Dynamic Circuits*



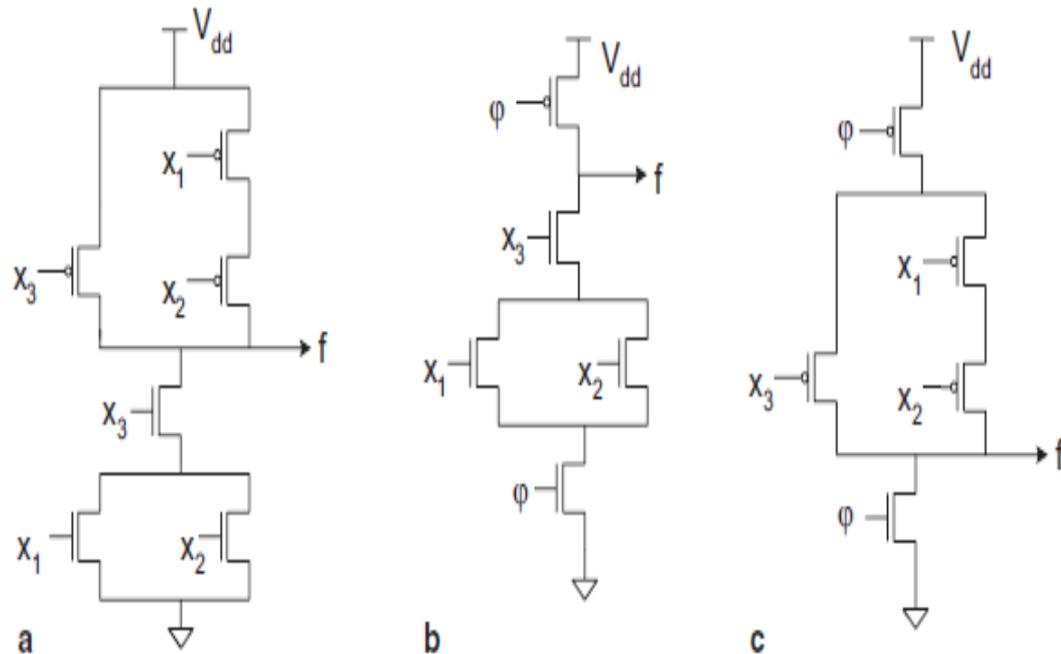
Realization of function $f = x_3(x_1 + x_2)$ using
a static complementary MOS (CMOS)

b dynamic CMOS with n-block

c dynamic CMOS with p-block

MOS Dynamic Circuits

- **CMOS Dynamic Circuits**
- Irrespective of several alternatives available in dynamic circuits, in all cases the output node is pre-charged to a particular level, while the current path to the other level is turned OFF.
- The charging of inputs to the gate must take place during this phase.



MOS Dynamic Circuits

- ***Advantages***

- ❖ The dynamic CMOS circuits have a number of advantages. The number of transistors required for a circuit with fan-in N is $(N + 2)$, in contrast to $2N$ in case of static 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.

MOS Dynamic Circuits

- *Advantages*

- 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

MOS Dynamic Circuits

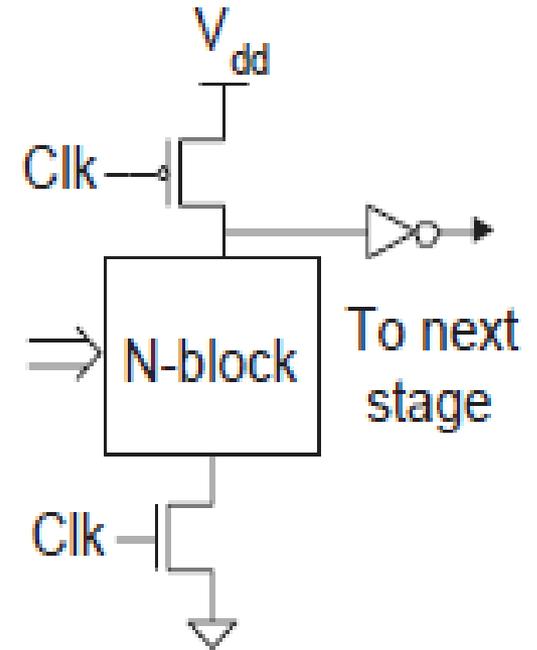
- *Disadvantages*
 - a. Charge Leakage Problem
 - b. Charge Sharing Problem
 - c. Clock Skew Problem

MOS Dynamic Circuits

- **Domino CMOS Circuits**

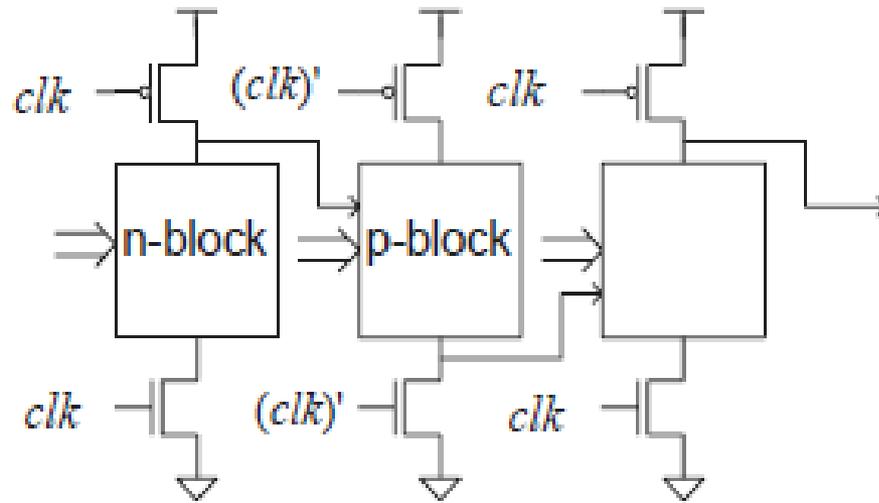
Domino CMOS circuits have the following advantages:

- 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.



MOS Dynamic Circuits

- NORA Logic





Low Power VLSI Circuits and Systems

Unit-3

Unit-3

Sources of Power Dissipations

- ❖ Introduction
 - ❖ Short-Circuit
Dissipation
 - ❖ Switching
Dissipation
 - ❖ Glitching
Dissipation
 - ❖ Leakage
Dissipation
- Power
- Power
- Power

Supply Voltage Scaling for Low Power

- ❖ Device Feature Size Scaling
- ❖ Architectural-Level
Approaches
- ❖ Voltage Scaling Using High-
Level Transformations
- ❖ Multilevel Voltage Scaling
- ❖ Challenges in MVS
- ❖ Dynamic Voltage and
Frequency Scaling
- ❖ Adaptive Voltage Scaling
- ❖ Subthreshold Logic Circuits

Introduction

- In order to develop techniques for minimizing power dissipation, it is essential to identify **various sources of power dissipation** and **different parameters** involved in each of them.
- Power dissipation may be specified in two ways.
- One is maximum power dissipation, which is represented by **“peak instantaneous power dissipation.”**
- Peak instantaneous power dissipation occurs when a circuit draws maximum power, which leads to a supply voltage spike due to resistances on the power line.
- Glitches may be generated due to this heavy flow of current and the circuit may malfunction, if proper care is not taken to suppress power-line glitches.

Introduction

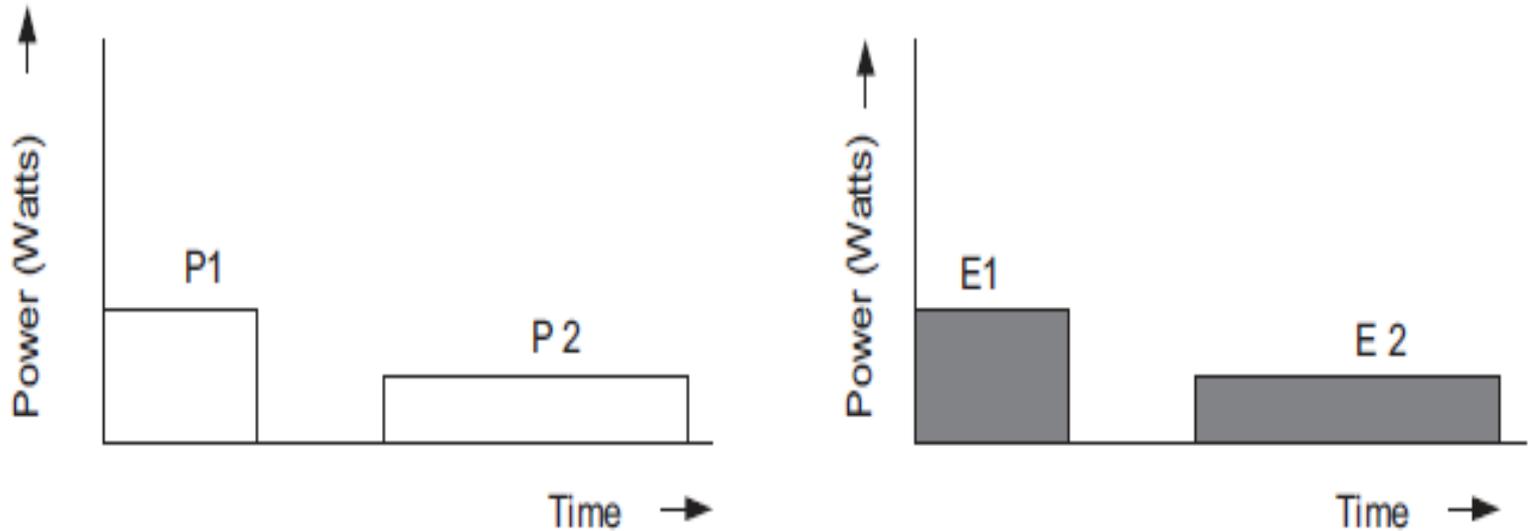
- The second one is the **“average power dissipation,”** which is important in the context of battery-operated portable devices.
- The average power dissipation will decide the battery lifetime.
- Here, we will be concerned mainly with the average power dissipation, although the techniques used for reducing the average power dissipation will also lead to the reduction of peak power dissipation and improve reliability by reducing the possibility of power-related failures.

Introduction

- Power dissipation can be divided into two broad categories—*static and dynamic*.
- **Static power dissipation** takes place continuously even if the inputs and outputs do not change.
- For some logic families, such as nMOS and pseudo-nMOS, both pull-up and pull-down devices are simultaneously ON for low output level causing direct current (DC) flow.
- This leads to static power dissipation..

Introduction

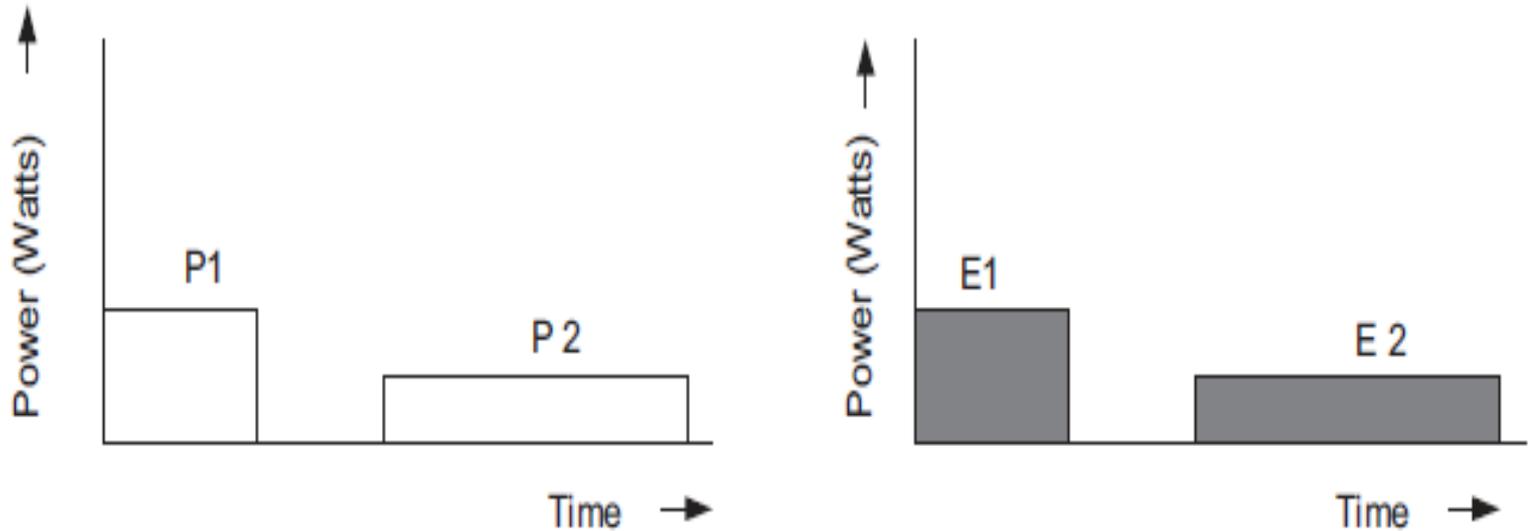
- Power and Energy



- Power dissipation is essentially the rate at which energy is drawn from the power supply, which is proportional to the average power dissipation.
- Power dissipation is important from the viewpoint of cooling and packaging of the integrated circuit (IC) chips.

Introduction

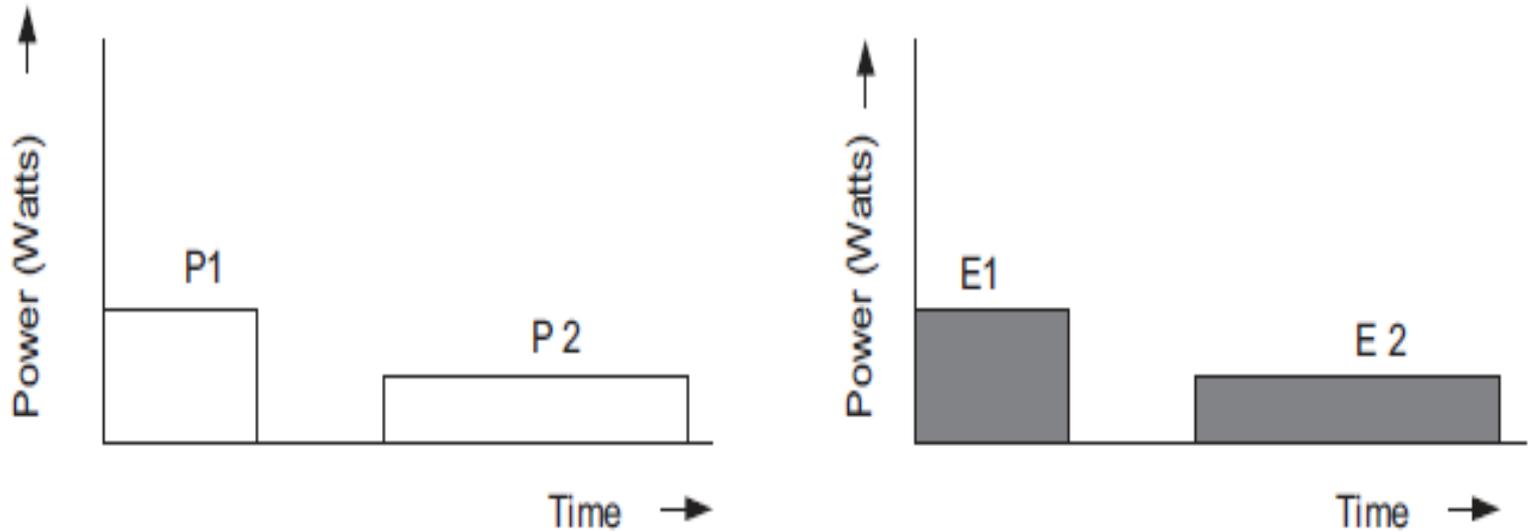
- Power and Energy



- On the other hand, energy consumed is important for battery-driven portable systems.
- As power dissipation is proportional to the clock frequency, to reduce power dissipation by half we may reduce the clock frequency by half.

Introduction

- Power and Energy



- Although this will reduce the power dissipation and keep the chip cooler, the time required to perform a computation will double that of the previous case.
- In this case, the energy is drawn from the battery at half the rate of the previous case.
- But, the total energy drawn from the battery for performing a particular computation remains the same.

Introduction

- **Dynamic power dissipation**
- In CMOS circuits, power dissipation can be divided into two broad categories: *dynamic and static*.
- *Dynamic power dissipation in CMOS circuits occur when the circuits are in working condition or active mode, that is, there are changes in input and output conditions with time.*
- In this section, we introduce the following three basic mechanisms involved in dynamic power dissipation:

Introduction

- **Dynamic power dissipation**
- *Short-circuit power: 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.
- *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*

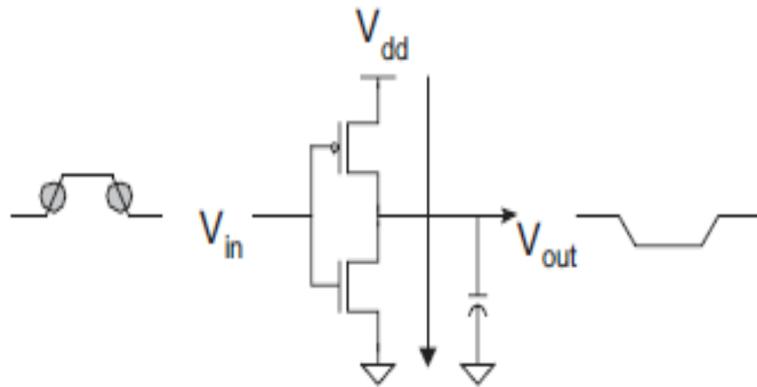
Introduction

- Static power dissipation
- 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 band of the p region to the conduction band of the n region, known as band-to-band-tunneling current
- Sub-threshold leakage current between source and drain when the gate voltage is less than the threshold voltage V_t

Introduction

- **Static power dissipation**
- 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

Short-Circuit Power Dissipation



Short-circuit power dissipation during input transition

- When there are finite rise and fall times at the input of CMOS logic gates, both pMOS and nMOS transistors are simultaneously ON for a certain duration, shorting the power supply line to ground.
- This leads to current flow from supply to ground.
- Short-circuit power dissipation takes place for input voltage in the range $V_{tn} < V_{in} < V_{dd} - |V_{tp}|$, when both pMOS and nMOS transistors turn ON creating a conducting path between V_{dd} and ground (GND).

Short-Circuit Power Dissipation

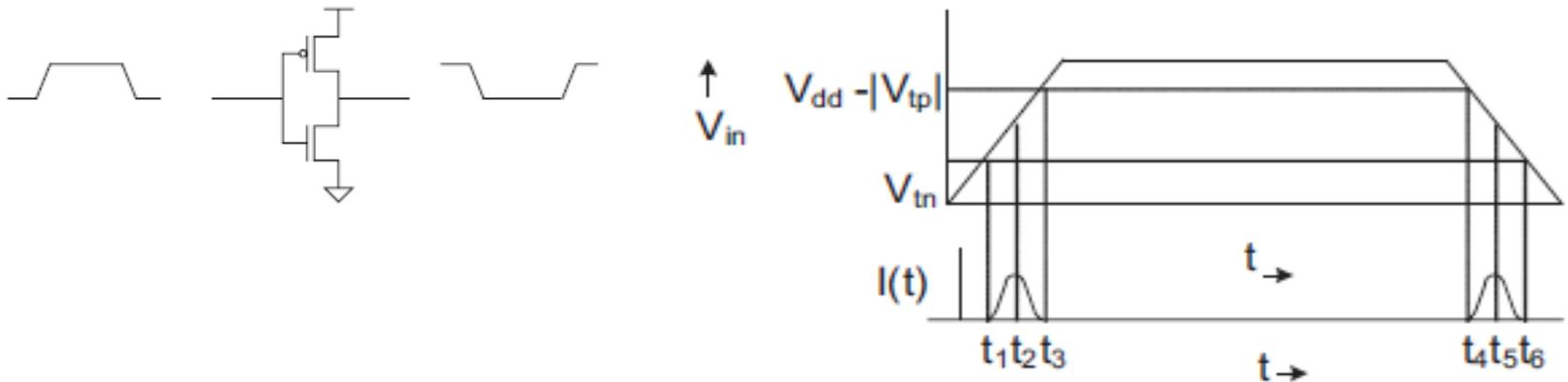
$$I_{\text{mean}} = 2 \times \frac{1}{T} \left[\int_{t_1}^{t_2} i(t) dt + \int_{t_2}^{t_3} i(t) dt \right].$$

$$I_{\text{mean}} = \frac{4}{T} \left[\int_{t_1}^{t_2} i(t) dt \right].$$

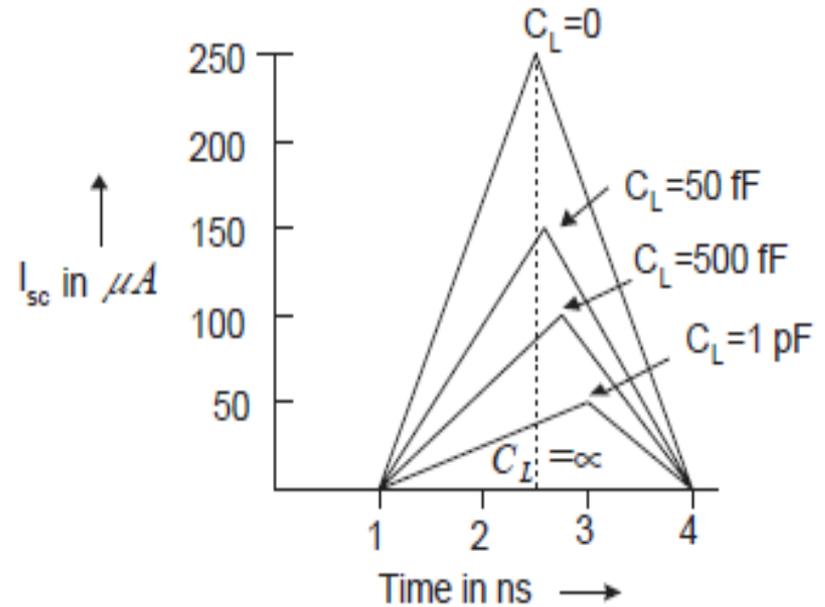
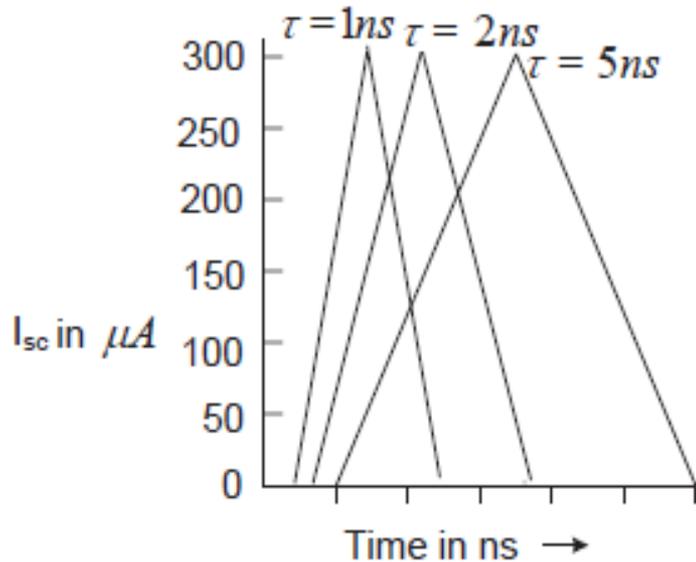
Because of the symmetry, we may write

As the nMOS transistor is operating in the saturation region,

$$i(t) = \frac{\beta}{2} (V_{\text{in}}(t) - V_t)^2 \quad I_{\text{mean}} = \frac{4}{T} \left[\int_{t_1}^{t_2} \frac{\beta}{2} (V_{\text{in}}(t) - V_t)^2 dt \right].$$



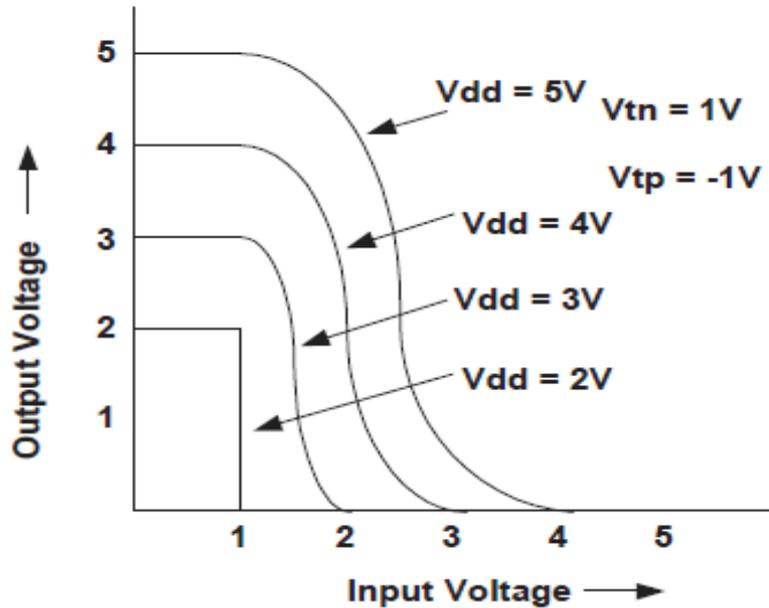
Short-Circuit Power Dissipation



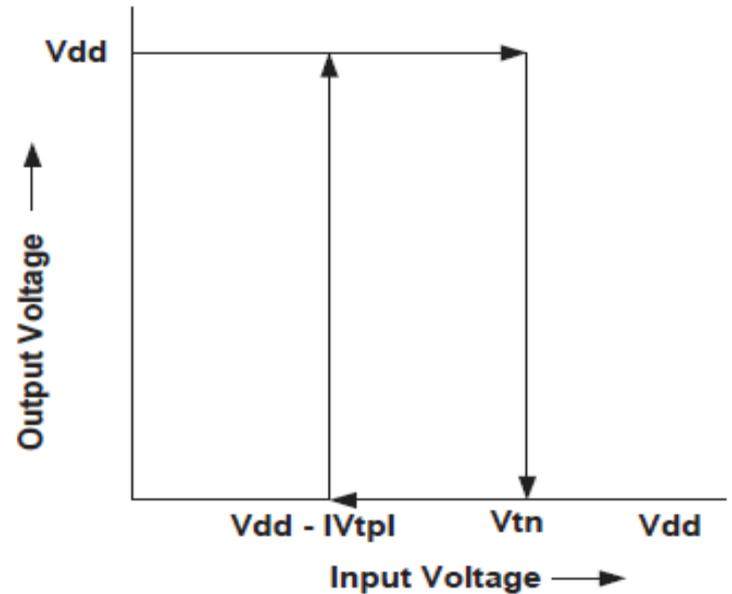
Variation of short-circuit current with load capacitance

Short-circuit current as a function of input rise/fall time

Short-Circuit Power Dissipation



Voltage transfer characteristics for $V_{dd} \geq (V_{tn} + |V_{tp}|)$



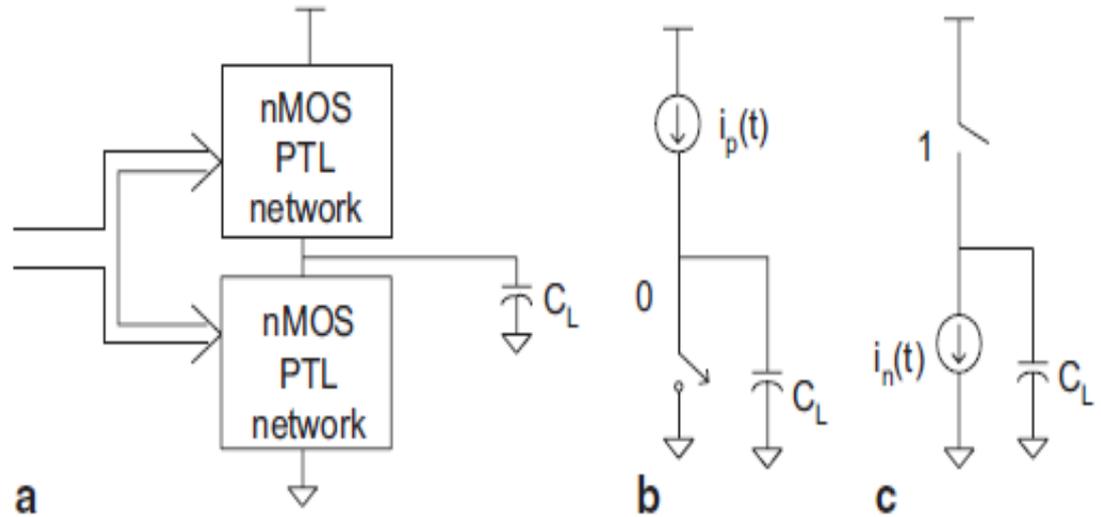
Transfer characteristics for $V_{dd} < (V_{tn} + |V_{tp}|)$

Switching Power Dissipation

- There exists capacitive load at the output of each gate.
- The exact value of capacitance depends on the fan-out of the gate, output capacitance, and wiring capacitances and all these parameters depend on the technology generation in use.
- As the output changes from a low to high level and high to low level, the load capacitor charges and discharges causing power dissipation.
- This component of power dissipation is known as switching power dissipation.

Switching Power Dissipation

Dynamic Power Dissipation Model

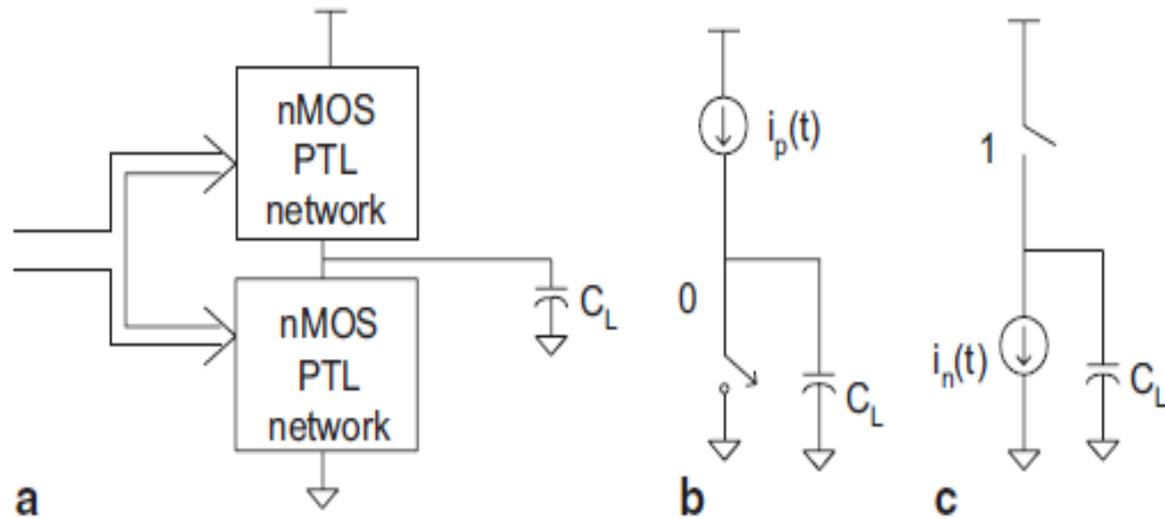


- ❖ For some input combinations, the pMOS network is ON and nMOS network is OFF.
- ❖ In this state, the capacitor is charged to V_{dd} by drawing power from the supply.
- ❖ For some other input combinations, the nMOS network is ON and pMOS network is OFF,
- ❖ In this state, the capacitor discharges through the nMOS network.
- ❖ For simplicity, let us assume that the CMOS gate is an inverter.

Switching Power Dissipation

Dynamic Power Dissipation Model

During the transition of the output from 0 to V_{dd} , the energy drawn from the power supply is given by



$$E_{0 \rightarrow 1} = \int_0^{V_{dd}} p(t) dt = \int V_{dd} \cdot i(t) dt,$$

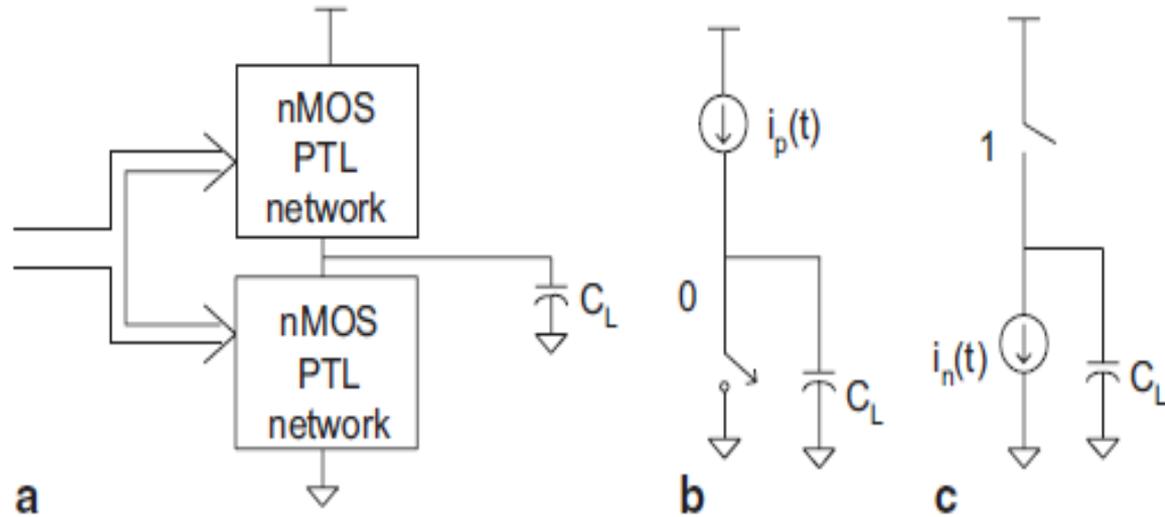
where $i(t)$ is an instantaneous current drawn from the supply voltage V_{dd} , and it can be expressed as

$$i(t) = C_L \frac{dV_0}{dt},$$

$$E_{0 \rightarrow 1} = V_{dd} \int_0^{V_{dd}} C_L dV_0 = C_L V_{dd}^2.$$

Switching Power Dissipation

Dynamic Power Dissipation Model



$$E_{0 \rightarrow 1} = V_{dd} \int_0^{V_{dd}} C_L dV_0 = C_L V_{dd}^2.$$

Regardless of the waveform and time taken to charge the capacitor, $C_L V_{dd}^2$ is the energy drawn from the power supply for 0 to V_{dd} transition at the load capacitance.

Switching Power Dissipation

$$P_d = \frac{1}{T} C_L V_{dd}^2 = C_L V_{dd}^2 f.$$

- ❖ The switching power is proportional to the **switching frequency** and **independent of device parameters**.
- ❖ As the switching power is proportional to the **square of the supply voltage**, there is a strong dependence of switching power on the supply voltage.
- ❖ Switching power reduces by **56 %**, if the supply voltage is reduced from **5 to 3.3 V**, and if the supply voltage is lowered to **1 V**, the switching power is reduced by **96 %** compared to that of **5V**.
- ❖ This is the reason why **voltage scaling** is considered to be the most dominant approach to reduce switching power.

Switching Power Dissipation

- *Dynamic Power for a Complex Gate*
- *Reduced Voltage Swing*
- *Internal Node Power*
- *Switching Activity*
- *Switching Activity of Static CMOS Gates*
- *Inputs Not Equiprobable*
- *Mutually Dependent Inputs*
- *Transition Probability in Dynamic Gates*
- *Power Dissipation due to Charge Sharing*

Switching Power Dissipation

- *Dynamic Power for a Complex Gate*

- ❖ For an inverter having a load capacitance C_L , the dynamic power expression is $C_L V_{dd}^2 f$.

- ❖ Here, it is assumed that the output switches from rail to rail and input switching occurs for every clock.

- ❖ This simple assumption does not hold good for complex gates because of several reasons.

- ❖ **First**, apart from the output load capacitance, there exist capacitances at other nodes of the gate.

Switching Power Dissipation

- *Dynamic Power for a Complex Gate*

□ As these internal nodes also charge and discharge, dynamic power dissipation will take place on the internal nodes.

□ This leads to two components of **dynamic power dissipation-load power and internal node power**.

□ **Second**, at different nodes of a gate, the voltage swing may not be from rail to rail.

□ This **reduced voltage swing** has to be taken into consideration for estimating the dynamic power.

Switching Power Dissipation

- *Dynamic Power for a Complex Gate*

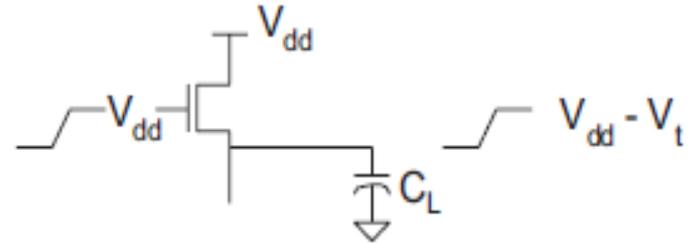
❖ **Finally**, to take into account the condition when the **capacitive node of a gate** might not switch when the clock is switching, a concept **known as *switching activity is introduced***.

❖ *Switching activity determines how often switching occurs on a capacitive node.*

❖ These three issues are considered in the following subsections.

Switching Power Dissipation

- ***Reduced Voltage Swing***

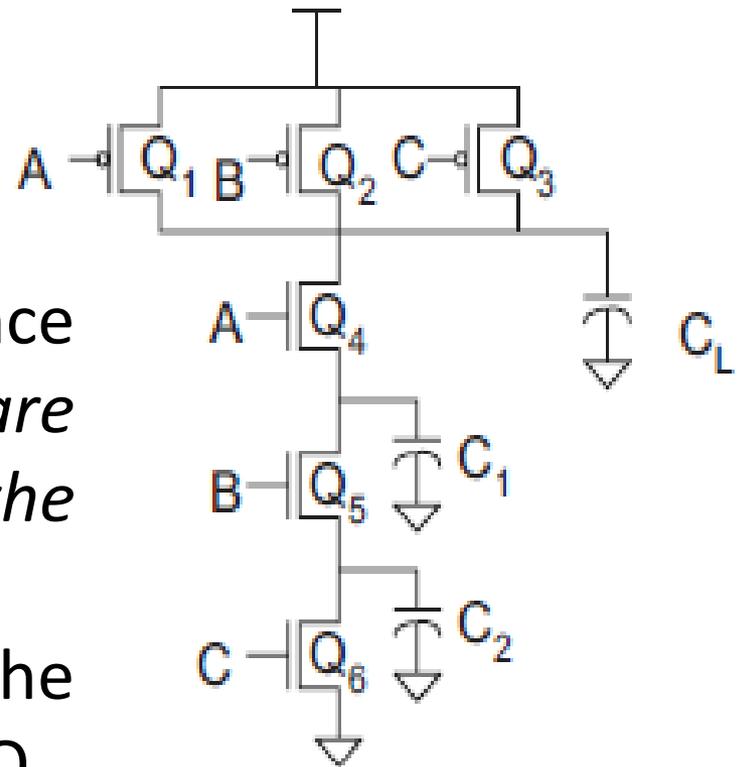


- There are situations where a rail-to-rail swing does not take place on a capacitive node.
- This situation arises in pass transistor logic and when the pull-up device is an enhancement-type nMOS transistor in nMOS logic gates
- In such cases, the output can only rise to $V_{dd} - V_t$.
- This situation also happens in internal nodes of CMOS gates.

Switching Power Dissipation

- *Internal Node Power*

Switching nodes of a three-input NAND gate



➤ Apart from the output capacitance C_L , two capacitances C_1 and C_2 are shown in two internal nodes of the gate.

➤ For input combination 110, the output is “1” and transistors Q_3 , Q_4 , and Q_5 are ON.

Switching Power Dissipation

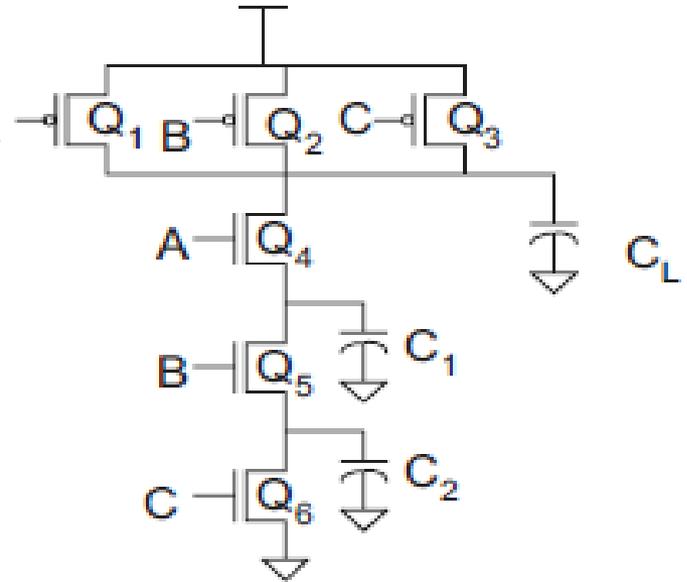
- **Internal Node Power**

Switching nodes of a three-input NAND gate

- All the capacitors will draw energy from the supply.
- Capacitor C_L will charge to V_{dd} through Q_3 , capacitor C_1 will charge to $(V_{dd} - V_t)$ through Q_3 and Q_4 .
- Capacitor C_2 will also charge to $(V_{dd} - V_t)$ through Q_3 , Q_4 , and Q_5 .
- For each 0-to- V_{dd} transition at an internal node, the energy drawn is given by

$$E_{0 \rightarrow 1} = C_i V_i V_{dd}$$

Where C_i is the internal node capacitance and V_i is internal voltage swing at node i .



Switching Power Dissipation

- ***Switching Activity***

➤ For a complex logic gate, the **switching activity** depends on two factors—the **topology of the gate** and the **statistical timing behavior of the circuit**.

➤ To handle the transition rate variation statistically, let **$n(N)$** be the number of 0-to- V_{dd} output transitions in the time interval $[0, N]$.

➤ **Total energy E_N** drawn from the power supply for this interval is given by

$$E_N = C_L V_{dd}^2 \times n(N).$$

Switching Power Dissipation

- **Switching Activity**

$$E_N = C_L V_{dd}^2 \times n(N).$$

The average power dissipation during an extended interval is $P_{avg} = \lim_{N \rightarrow \infty} \frac{E_N}{N} \times f$, where f is the clock frequency.

$$P_{avg} = \left(\lim_{N \rightarrow \infty} \frac{n(N)}{N} \right) C_L V_{dd}^2 f.$$

The term $\lim_{N \rightarrow \infty} (n(N)/N)$ gives us the expected (average) value of the number of transitions per clock cycle, which is defined as the switching activity.

$$\alpha_{0 \rightarrow 1} = \lim_{N \rightarrow \infty} \frac{n(N)}{N}, \quad P_d = \alpha_0 C_L V_{dd}^2 f + \sum_{i=1}^k \alpha_i C_i V_i V_{dd} f,$$

➤ Where α_0 is the switching activity at the output node, α_i is the switching activity on the i_{th} internal node, and f is the clock frequency.

➤ Here, it is assumed that there are k internal nodes.

Switching Power Dissipation

- ***Switching Activity of Static CMOS Gates***
- The switching activity at the output of a static CMOS gate depends strongly on the function it realizes.
- It is assumed that the inputs are independent to each other and the probability of occurrence of “0” and “1” is same, let P_0 be the probability that the output will be “0” and P_1 is the probability that the output will be “1.” Then, $P_{0 \rightarrow 1} = P_0 P_1 = P_0(1 - P_0)$.

Switching Power Dissipation

- ***Switching Activity of Static CMOS Gates***
- For an n -input gate, the total number of input combinations is 2^n .
- Out of 2^n combinations in the truth table, let n_0 be the total number of combinations for which the output is 0 and n_1 is the total number of combinations for which the output is “1,” then

A	B	F_{NAND}
0	0	1
0	1	1
1	0	1
1	1	0

Switching Power Dissipation

- ***Switching Activity of Static CMOS Gates***

- For an n -input gate, the total number of input combinations is 2^n .
- Out of 2^n combinations in the truth table, let n_0 be the total number of combinations for which the output is 0 and n_1 is the total number of combinations for which the output is “1,” then

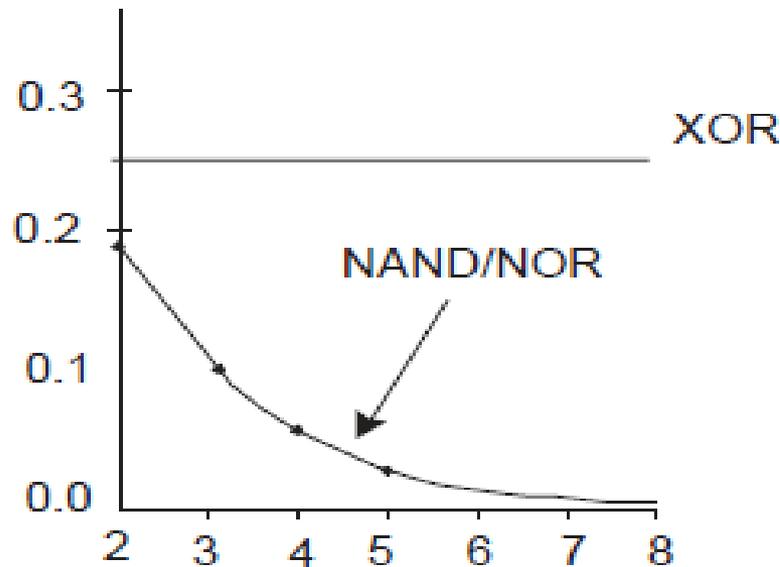
$$P_0 = \frac{n_0}{2^n} \text{ and } P_1 = \frac{n_1}{2^n}.$$

$$P_{0 \rightarrow 1} = \frac{n_0}{2^n} \times \frac{n_1}{2^n} = \frac{n_0(2^n - n_0)}{2^{2n}}.$$

Switching Power Dissipation

- ***Switching Activity of Static CMOS Gates***
- The variation of the switching activity at the output of NAND, NOR, and EX-OR gates with the increase in the number of inputs is shown in Fig. 6.11.

Variation of switching activity with increase in the number of inputs



Switching Power Dissipation

- *Inputs Not Equiprobable*
- we have assumed that the inputs are independent of each other and equiprobable.
- But, inputs might not be equiprobable.
- In such cases, the probability of transitions at the output depends on the probability of transitions at the primary inputs.

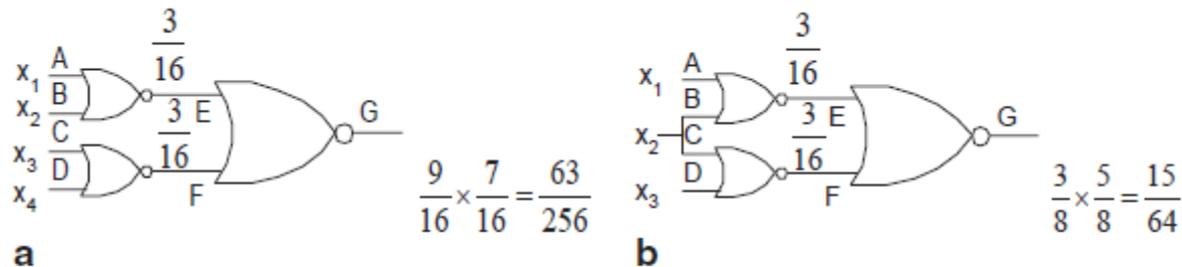
Switching Power Dissipation

- ***Inputs Not Equiprobable***
- Let us consider a two-input NAND gate with two inputs A and B having mutually independent signal probabilities of P_A and P_B .
- *The probability of logical “0” at the output is $P_0 = P_A P_B$ and the probability of “1” at the output is $P_1 = (1 - P_0) = (1 - P_A P_B)$.*

$$P_{0 \rightarrow 1} = P_0 P_1 = (1 - P_A P_B) P_A P_B.$$

Switching Power Dissipation

- ***Mutually Dependent Inputs***
- When a multilevel network of complex gates are considered, the inputs to a gate at a particular stage might not be independent.
- This can happen because the signal from one input may propagate through multiple paths and again recombine at a particular stage.
- This problem of re-convergent fan-out is shown with the help of a simple example given in Fig. 6.12.



a Circuit without re-convergent fan-out. b Circuit with re-convergent fan-out

Switching Power Dissipation

- ***Mutually Dependent Inputs***

Characteristics of the standard cells

Gate type	Area (cell unit)	Input capacitance (fF)	Output capacitance (fF)	Average delay (ns)
INV	2	85	48	$0.22 + 1.00 C_0$
NAND2	3	105	48	$0.30 + 1.24 C_0$
NAND3	4	132	48	$0.37 + 1.50 C_0$
NAND6	7	200	48	$0.65 + 2.30 C_0$
NOR2	3	101	48	$0.27 + 1.50 C_0$
NOR3	4	117	48	$0.31 + 2.00 C_0$

➤ Provides details of various gates used for the realization of the three implementations.

➤ The table provides the area in terms of unit area called cell grid, output and input capacitances, and delay in terms of output capacitance C_0 in picofarad.

➤ For all transistors, $Wp = 2Wn = 10 \mu\text{m}$.

Switching Power Dissipation

- *Transition Probability in Dynamic Gates*
- The logic style has a strong influence on the node transition probability.
- For static CMOS, we calculated the transition probability as $P_0 P_1 = P_0(1 - P_0)$.
- *In the case of dynamic gates, the situation is different.*
- As we have seen, in the case of **domino or NORA logic styles**, the output nodes are **pre-charged** to a high level in the **Pre-charge phase** and then are discharged in the **evaluation phase**, depending on the outcome of evaluation.

Switching Power Dissipation

- *Transition Probability in Dynamic Gates*
- In other words, the output transition probability will depend on the signal probability P_0 , i.e.,

$$P_{0 \rightarrow 1} = P_0,$$

- where P_0 is the probability for the output is in the zero state. For n independent inputs to a gate

$$P_{0 \rightarrow 1} = \frac{N_0}{2^N},$$

- where N_0 is the number of zero entries in the truth table

Switching Power Dissipation

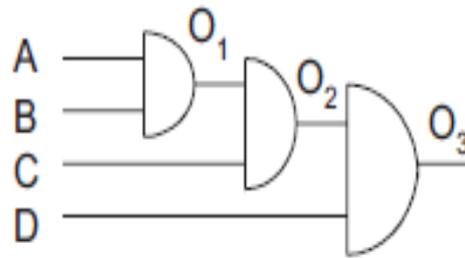
- ***Power Dissipation due to Charge Sharing***
- Moreover, in the case of dynamic gates, power dissipation takes place due to the phenomenon of charge sharing even when the output is not 0 at the time evaluation, i.e., the output load capacitance is not discharged, but part of the charge of the load capacitance might get redistributed leading to a reduction in the output voltage level.
- In the next pre-charge period, the output is again pre-charged back to V_{dd} .

Glitching Power Dissipation

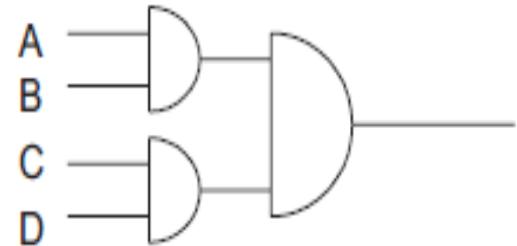
- In the power calculations so far, we have assumed that the **gates have zero delay**.
- In practice, the gates will have finite delay and this delay will lead to spurious undesirable transitions at the output.
- These spurious signals are known as *glitches*.
- In the case of a static CMOS circuit, the output node or internal nodes can make undesirable transitions before attaining a stable value.

Glitching Power Dissipation

- If the inputs ABC change value from 101 to 000, ideally for zero gate delay the output should remain at the 0 logic level.
- However, considering unit gate delay of the first gate stage, output O_1 is delayed compared to the C input



a

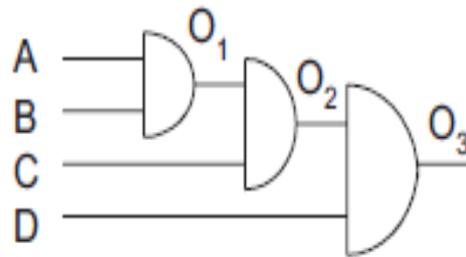


b

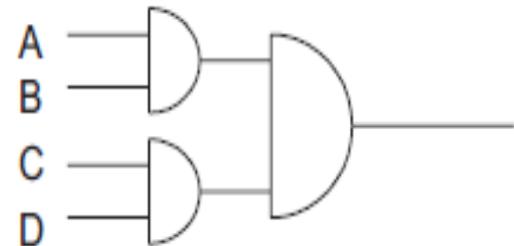
Realization of A, B, C, and D, a in cascaded form, b balanced realization

Glitching Power Dissipation

- As a consequence, the output switches to 1 logic level for one gate delay duration.
- This transition increases the dynamic power dissipation and this component of dynamic power is known as *glitching power*.



a

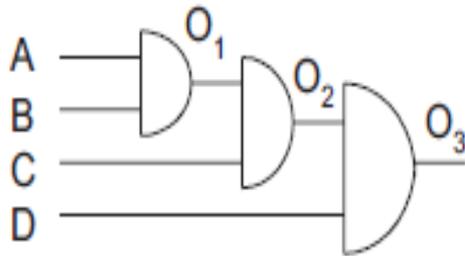


b

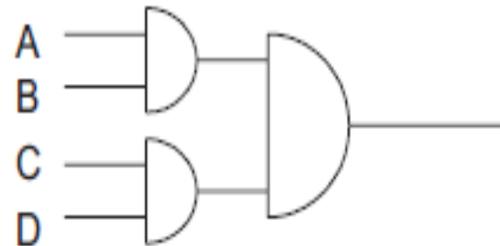
Realization of A, B, C, and D, a in cascaded form, b balanced realization

Glitching Power Dissipation

- *Glitching power may constitute a significant portion of dynamic power, if circuits are not properly designed.*



a

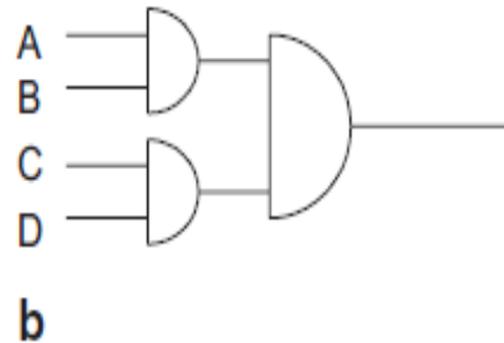
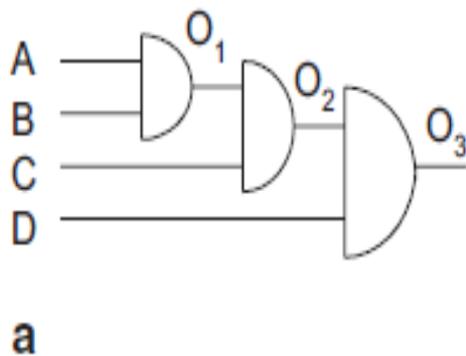


b

Realization of A, B, C, and D, a in cascaded form, b balanced realization

Glitching Power Dissipation

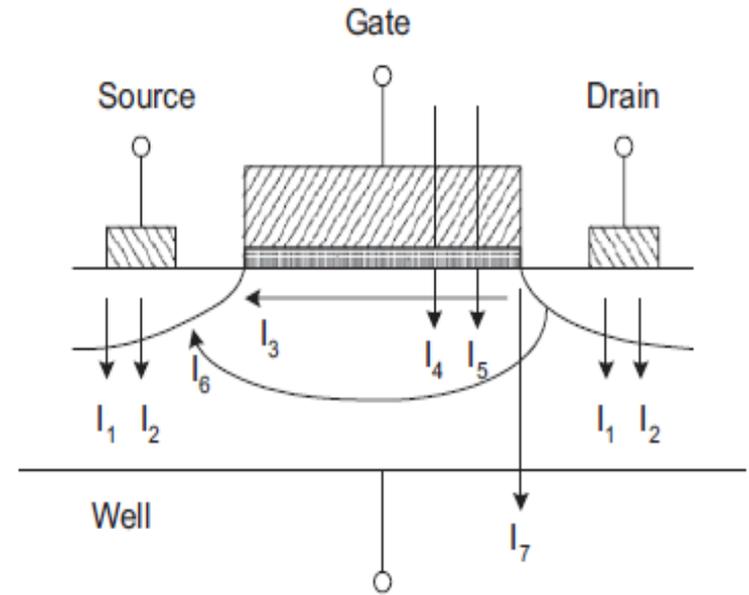
- Usually, cascaded circuits as shown in Fig. 6.16a exhibit high glitching power.
- The glitching power can be minimized by realizing a circuit by balancing delays, as shown in Fig. 6.16b.
- On highly loaded nodes, buffers can be inserted to balance delays and cascaded implementation can be avoided, if possible, to minimize glitching power.



Realization of A, B, C, and D, a in cascaded form, b balanced realization

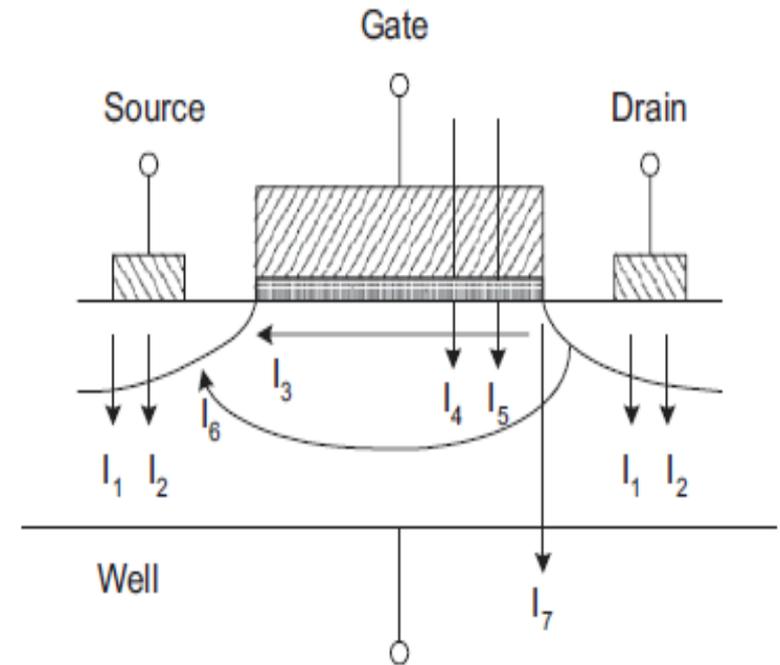
Leakage Power Dissipation

- When the circuit is not in an active mode of operation, there is **static power dissipation** due to various **leakage mechanisms**.
- In deep-submicron devices, these leakage currents are becoming a significant contributor to power dissipation of CMOS circuits.



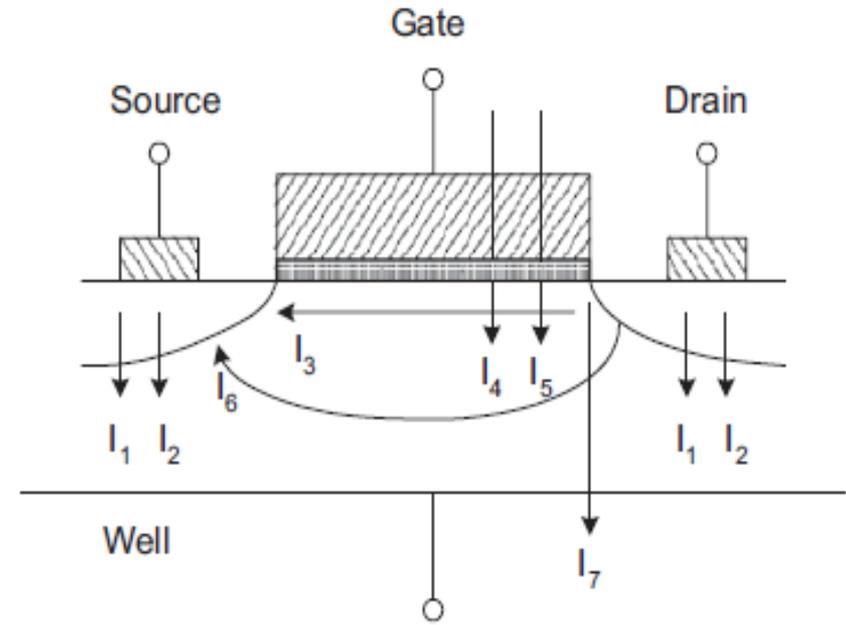
Leakage Power Dissipation

- 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 sub-threshold leakage current between the source and the drain when the gate voltage is less than the threshold voltage V_t ;

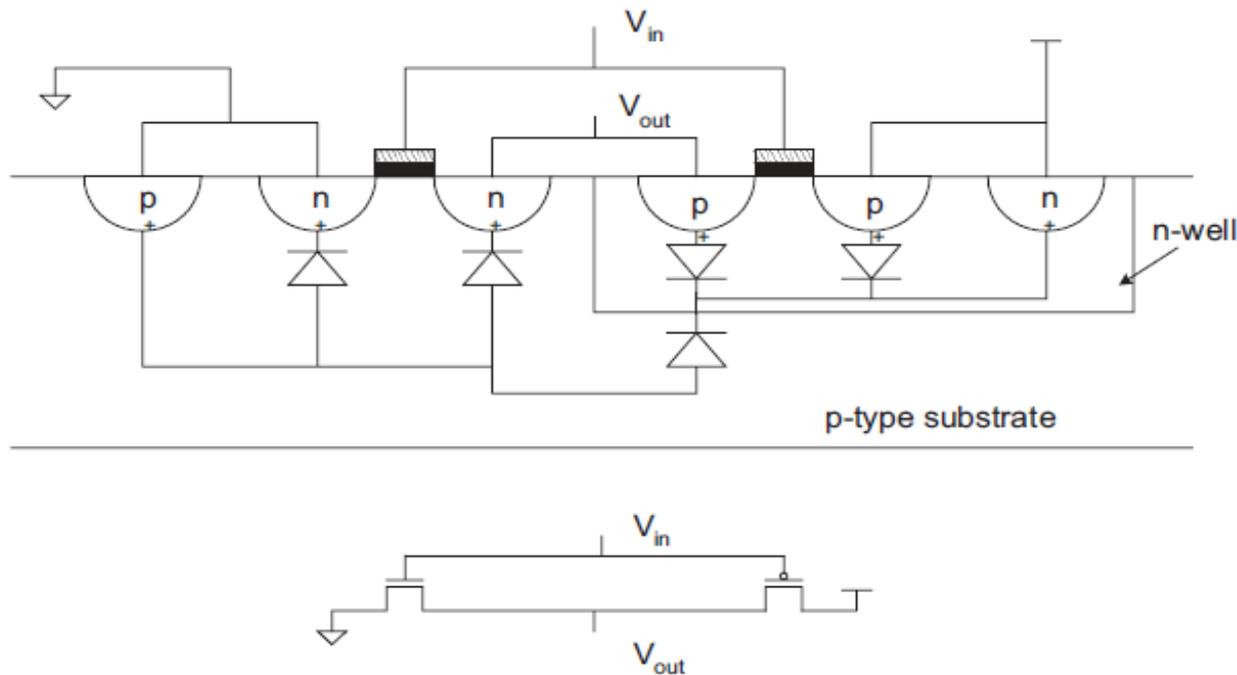


Leakage Power Dissipation

- 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; and I_7 is the **channel punch-through current** due to the close proximity of the drain and the source in short-channel devices.



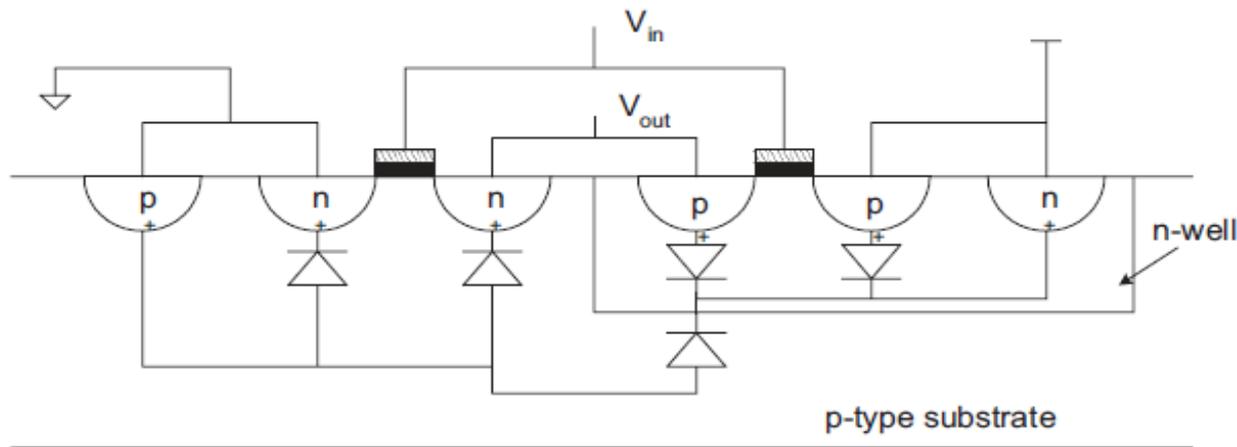
p–n Junction Reverse-Biased Current



nMOS Inverter and
its Physical
Structure

- ❖ Let us consider the physical structure of a CMOS inverter shown in Fig. 6.18.
- ❖ As shown in the figure, source–drain diffusions and n-well diffusions form parasitic diodes in the bulk of silicon substrate.
- ❖ As parasitic diodes are reverse-biased, their leakage currents contribute to static power dissipation.

p–n Junction Reverse-Biased Current



nMOS Inverter and its Physical Structure

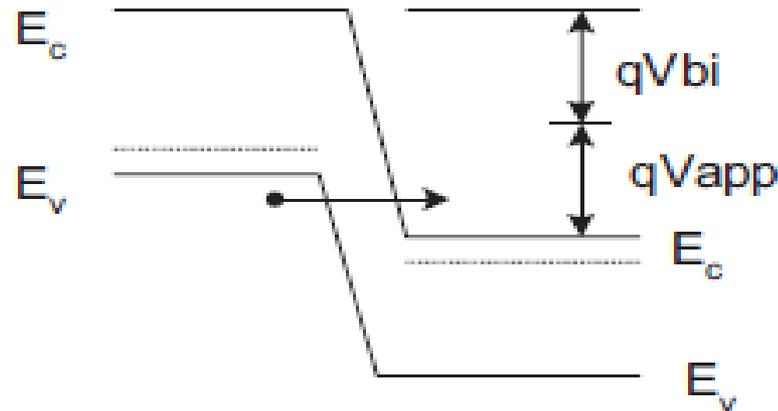


The current for one diode is given by

$$I_{rdlc} = AJ_s \left[e^{-\frac{qV_d}{nKT}} - 1 \right],$$

where J_s is the reverse saturation current density (this increases with temperature), I_s is the AJ_s , V_d is the diode voltage, n is the emission coefficient of the diode (sometimes equal to 1), q is the charge of an electron (1.602×10^{-19}), K is the Boltzmann constant (1.38×10^{-23} j/k), T is the temperature in K, $V_T = KT/q$ is known as the thermal voltage. At room temperature, $V_T = 26$ mV

Band-to-Band Tunneling Current

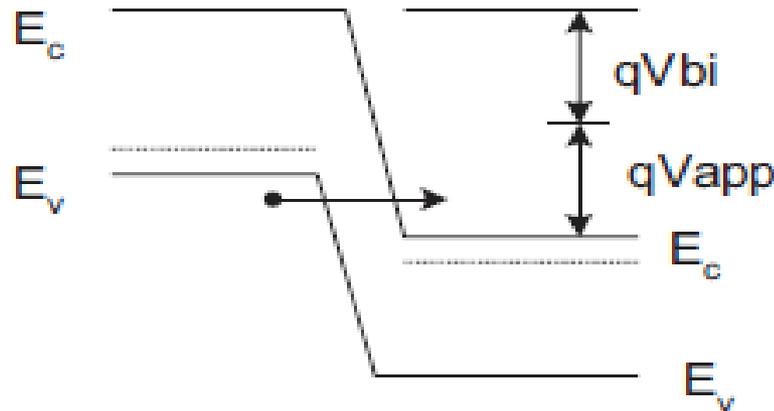


BTBT in reverse biased
p-n junction

✓ When both n regions and p regions are heavily doped, a high electric field across a reverse biased p-n junction causes a significant current to flow through the junction due to tunneling of electrons from the valence band of the p region to the conduction band of n region.

✓ This is illustrated in Fig. 6.19. It is evident from this diagram that for the voltage drop across that junction should be more than the band gap.

Band-to-Band Tunneling Current



BTBT in reverse biased
p-n junction

The tunneling current density is given by

$$J_{b-b} = A \frac{EV_{app}}{E_g^{1/2}} \exp\left(-B \frac{E_g^{3/2}}{E}\right),$$

$$A = \frac{\sqrt{2m^* q^3}}{4\pi^3 \hbar^2} \text{ and } B = \frac{4\sqrt{2m^*}}{3q\hbar},$$

where m^* is the effective mass of electron, E is the electric field at the junction, q is the electronic charge, and h is the reduced Planck's constant ($1/2\pi$ times).

Subthreshold Leakage Current

- The **subthreshold leakage current** in CMOS circuits is due to carrier diffusion between the source and the drain regions of the transistor in weak inversion, when the gate voltage is below V_t .

Subthreshold Leakage Current

- *The behavior of an MOS transistor in the subthreshold operating region is similar to a bipolar device, and the subthreshold current exhibits an exponential dependence on the gate voltage.*
- The amount of the subthreshold current may become significant when the gate-to-source voltage is smaller than, but very close to, the threshold voltage of the device.

Sub-threshold Leakage Current

- The sub-threshold current expression as given by the BSIM3 model is stated below:

$$I_{\text{stlc}} = A e^{\frac{q}{N'KT}(V_{\text{gs}} - V_{\text{th}})} \left(1 - e^{\frac{-qV_{\text{ds}}}{kT}} \right)$$

$$= \mu_0 c_{\text{ox}} \frac{W}{L} (m-1) (V_{\text{T}})^2 \times e^{(V_{\text{gs}} - V_{\text{th}})/mV_{\text{T}}} \times \left(1 - e^{-V_{\text{ds}}/V_{\text{T}}} \right)$$

$$m = 1 + \frac{c_{\text{dm}}}{c_{\text{ox}}} = 1 + \frac{\epsilon_{\text{si}}/w_{\text{dm}}}{\epsilon_{\text{ox}}/t_{\text{ox}}} = 1 + \frac{3t_{\text{ox}}}{w_{\text{dm}}}$$

The typical value of this current for a single transistor is **1–10 nA**.

Where m is the **sub-threshold swing coefficient**,

$V_{\text{T}} = kT/q$ is the **thermal voltage**,

μ_0 is the **zero bias mobility**,

c_{ox} is the **gate oxide capacitance per unit area**,

w_{dm} is the **maximum depletion layer width**, and t_{ox} is the **gate oxide thickness**.

Sub-threshold Leakage Current

- Various mechanisms which affect the sub-threshold leakage current are:
- Drain-induced barrier lowering (DIBL)
- Body effect
- Narrow-width effect
- Effect of channel length and V_{th} *roll-off*
- Effect of temperature

Drain-induced barrier lowering (DIBL)

- For long-channel devices, the sources and drain region are separated far apart and the depletion regions around the drain and source have little effect on the potential distribution in the channel region.
- So, the threshold voltage is independent of the channel length and drain bias for such devices.
- However, for short-channel devices, the source and drain depletion width in the vertical direction and the source drain potential have a strong effect on a significant portion of the device leading to variation of the subthreshold leakage current with the drain bias. This is known as the DIBL effect.

Drain-induced barrier lowering (DIBL)

- Because of the DIBL effect, the barrier height of a short-channel device reduces with an increase in the subthreshold current due to a lower threshold voltage.
- DIBL occurs when the depletion regions of the drain and the source interact with each other near the channel surface to lower the source potential barrier.

Body Effect

- As a negative voltage is applied to the substrate with respect to the source, the well-to-source junction, the device is reverse-biased and bulk depletion region is widened.
- This leads to an increase in the threshold voltage. This effect is known as the body effect.

Body Effect

- The threshold voltage equation given below gives the relationship of the threshold voltage with the body bias

$$V_{th} = V_{fb} + 2\psi_B + \frac{\sqrt{2\epsilon_{si} q N_a (2\tau_B + V_{bs})}}{C_{ox}},$$

Where V_{fb} is the flat-band voltage, N_a is the doping density in the substrate,



is the difference between the Fermi potential and the intrinsic potential in the substrate.

Body Effect

- The variation of the threshold voltage with respect to the substrate bias dV_{th}/dV_{bs} is referred to as the substrate sensitivity:

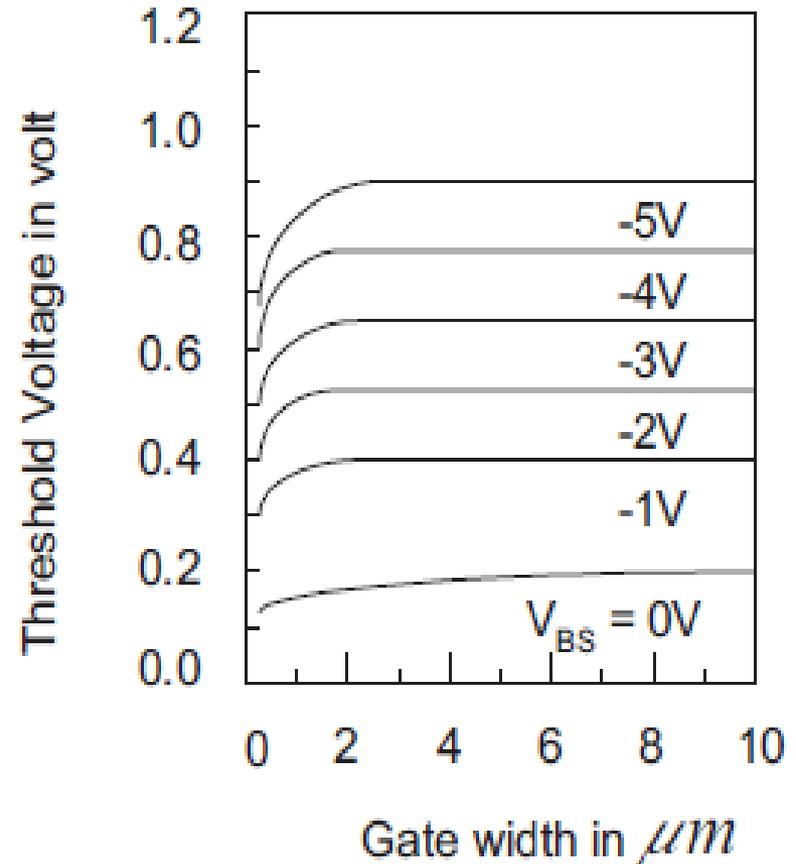
$$\frac{dV_{th}}{dV_{bs}} = \frac{\sqrt{\frac{\epsilon_{si} q N_a}{2(2\psi_B + V_{bs})}}}{C_{ox}}$$

Substrate sensitivity is higher for higher bulk doping concentration, and it decreases as the substrate reverse bias increases.

At $V_{bs} = 0$, it is equal to C_{dm}/C_{ox} or $m - 1$ where m is also called the body effect coefficient.

Narrow-Width Effect

- The width of a gate, particularly when it becomes narrow, affects the threshold voltage as shown in Fig. 6.25.
- The reduction in threshold voltage also leads to an increase in the subthreshold leakage current (Fig. 6.24).

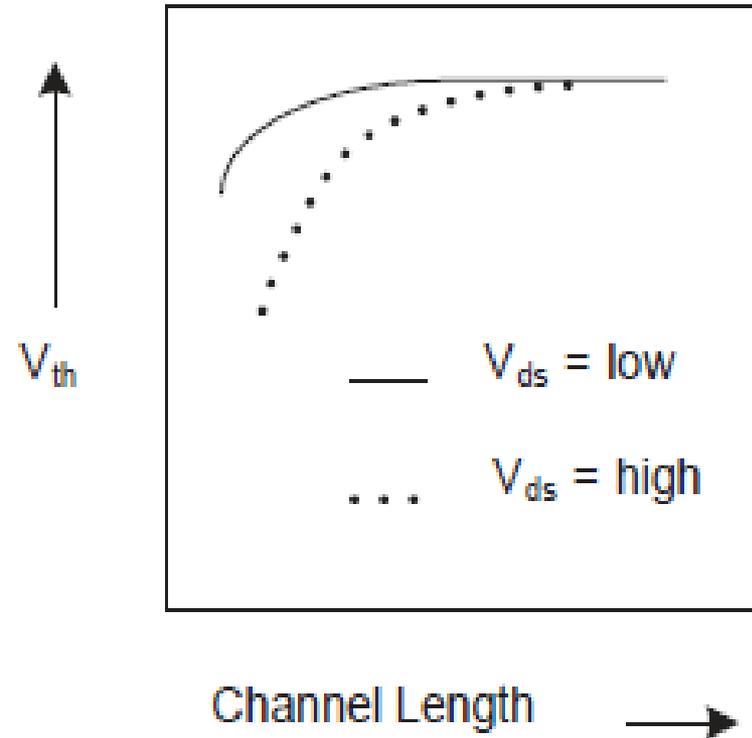


V_{th} Roll-Off

- As the channel length is reduced, the threshold voltage of metal–oxide–semiconductor field-effect transistor (MOSFET) decreases.
- This reduction of channel length is known as *V_{th} roll-off*.

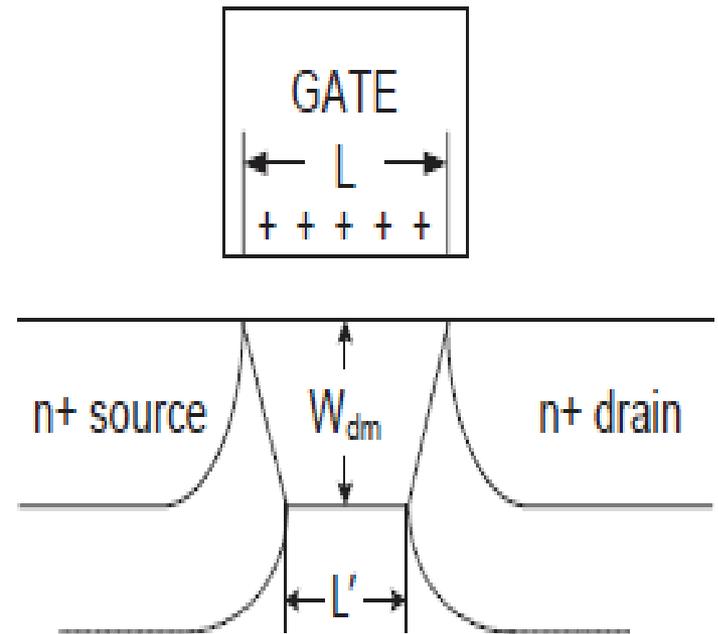
V_{th} Roll-Off

- Figure 6.26 shows the reduction of threshold voltage with a reduction in channel length.
- This effect is caused by the proximity of the source and drain regions leading to a 2D field pattern rather than a 1D field pattern in short-channel devices.



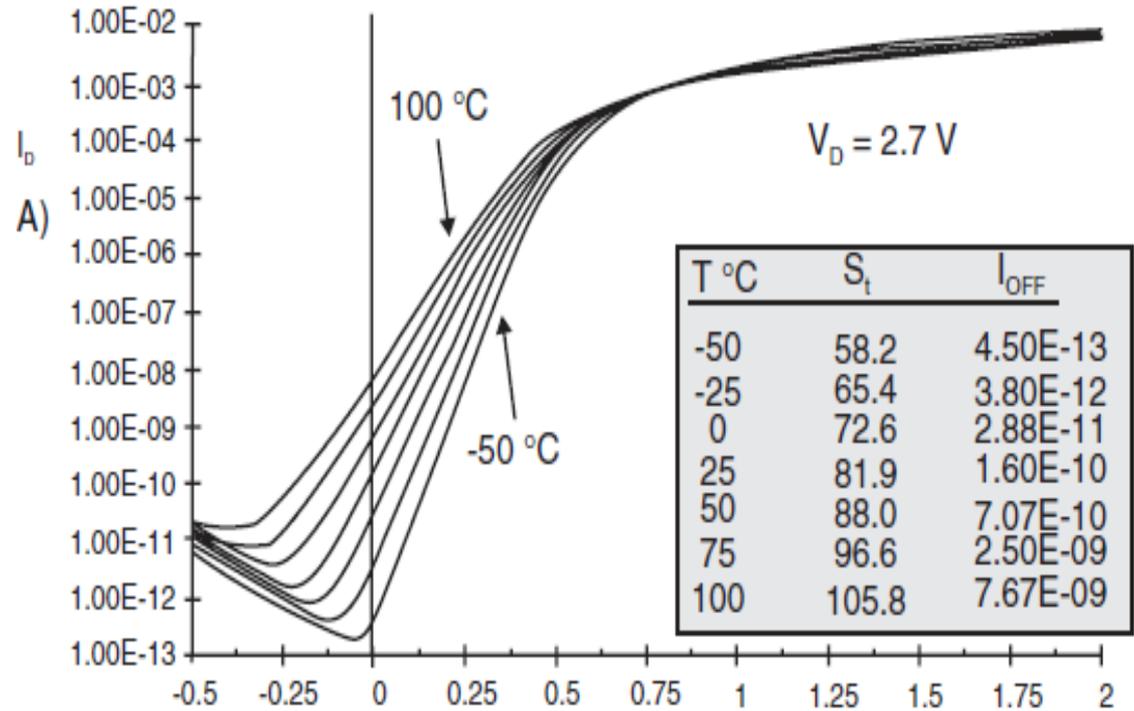
V_{th} Roll-Off

- The bulk charge that needs to be inverted by the application of gate voltage is proportional to the area under the trapezoidal region, shown in Fig. 6.27, given by $Q_B' W_{dm} (L + L') / 2$.



Temperature Effect

- For a 0.3- μm technology, S_t varies from 58.2 to 81.9 mV/decade.
- As the temperature varies from -50 to $+25$ $^{\circ}\text{C}$ in the 0.3- μm technology as shown in Fig. 6.28, the I_{OFF} increases from 0.45 to 100 pA, an increase by a factor of 356, for a 20- μm -wide device (23 fA/ μm to 8 pA/ μm)..

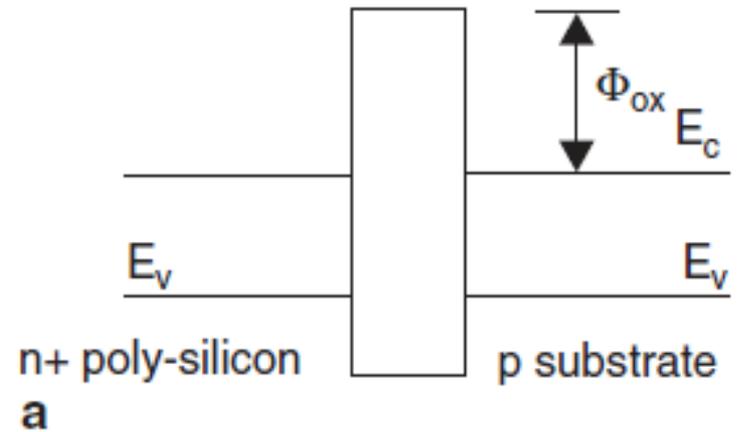


Tunneling Through Gate Oxide

- Device size scaling leads to a reduction in oxide thickness, which in turn results in an increase in the field across the oxide.
- The high electric field along with low oxide thickness leads to the tunneling of electrons from the substrate to the gate and vice versa, through the gate.
- The basic phenomenon of tunneling is explained with the help of a heavily doped n + type polysilicon gate and a p-type substrate.

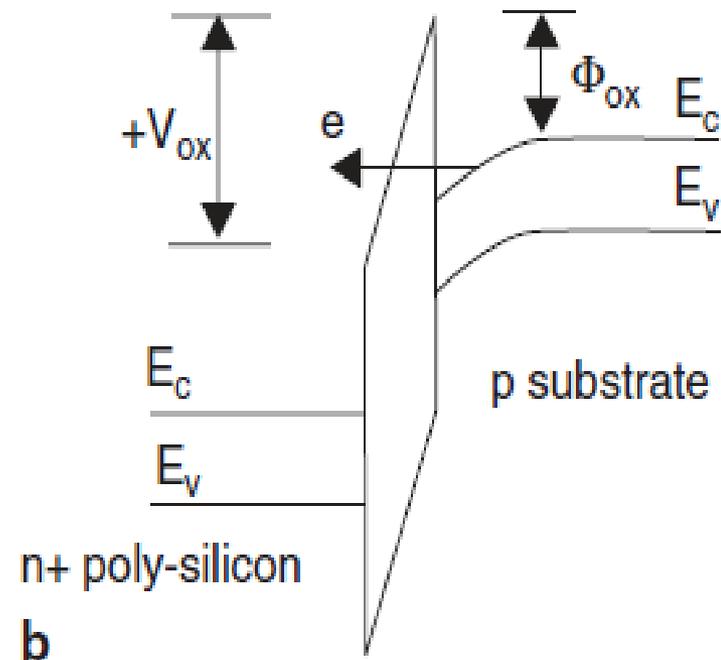
Tunneling Through Gate Oxide

- Because of higher mobility, primarily electrons take part in the tunneling process.
- An energy band diagram in flat-band condition is shown in Fig. 6.29a, where Φ is the Si–SiO₂ interface barrier height for electrons.



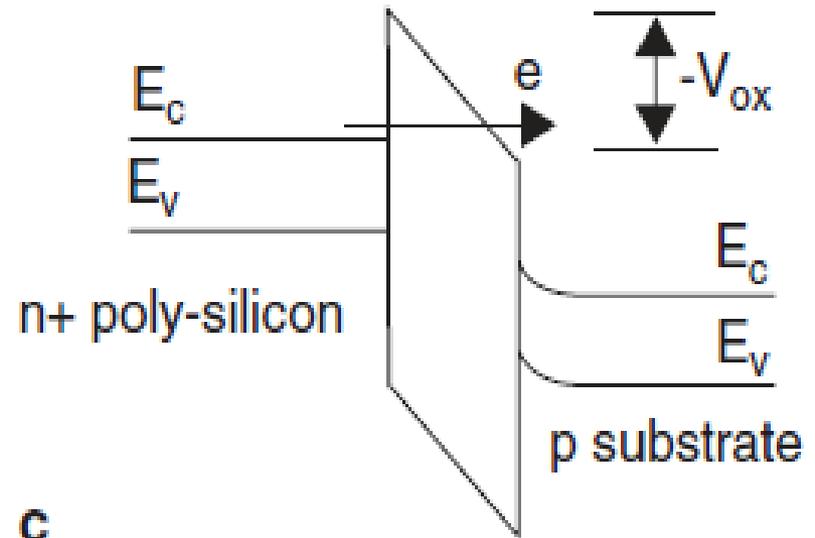
Tunneling Through Gate Oxide

- The energy band diagram changes to that of Fig. 6.29b, when a positive bias is applied to the gate.
- The electrons at the strongly inverted surface can tunnel through the SiO₂ because of the small width of the potential barrier arising out of small oxide thickness.



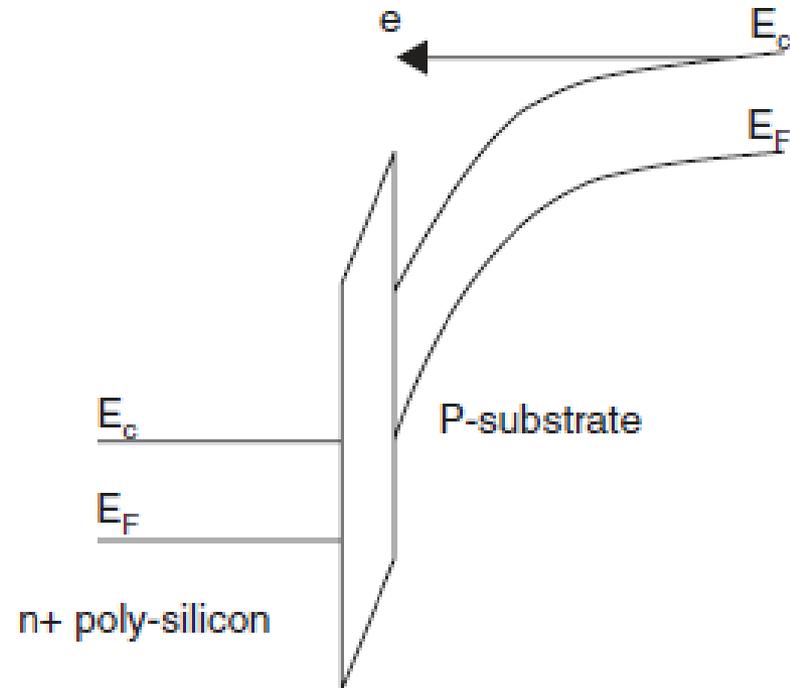
Tunneling Through Gate Oxide

- Similarly, if a negative gate bias is applied, the electrons from the n + poly-silicon can tunnel through the oxide layer as shown in Fig. 6.29c.
- This results in gate oxide-tunneling current, which violates the classical infinite input impedance assumption of MOS transistors and thus affects the circuit performance significantly.



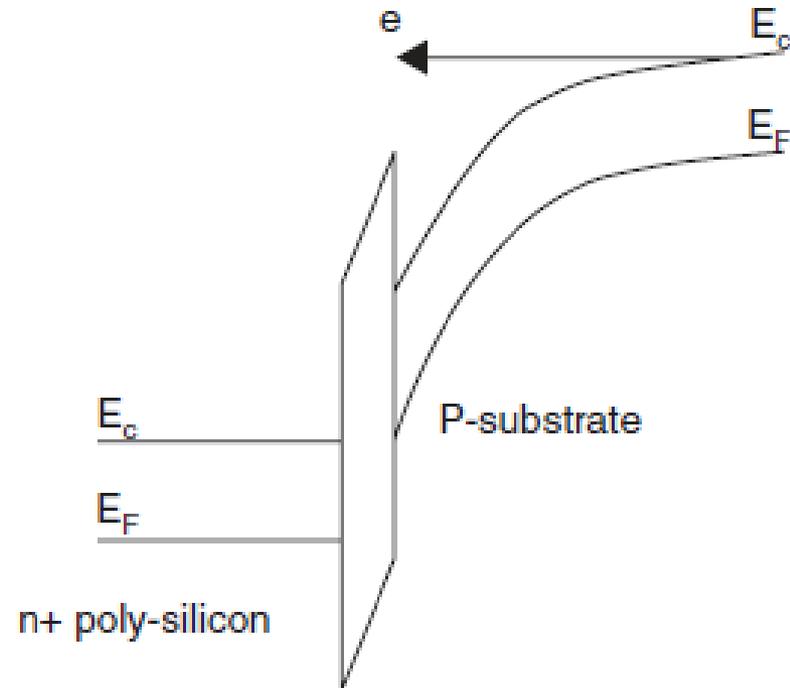
Hot-Carrier Injection

- Electrons and holes near the Si–SiO₂ interface can gain sufficient energy due to the high electric field in short-channel MOSFETs and cross the interface potential barrier and enter into the oxide layer as shown in Fig. 6.30.



Hot-Carrier Injection

- This phenomenon is known as the hot-carrier injection. Because of the lower effective mass and smaller barrier height (3.1 eV) for electrons compared to holes (4.5 eV), the possibility of an injection due to electrons is more than the holes.

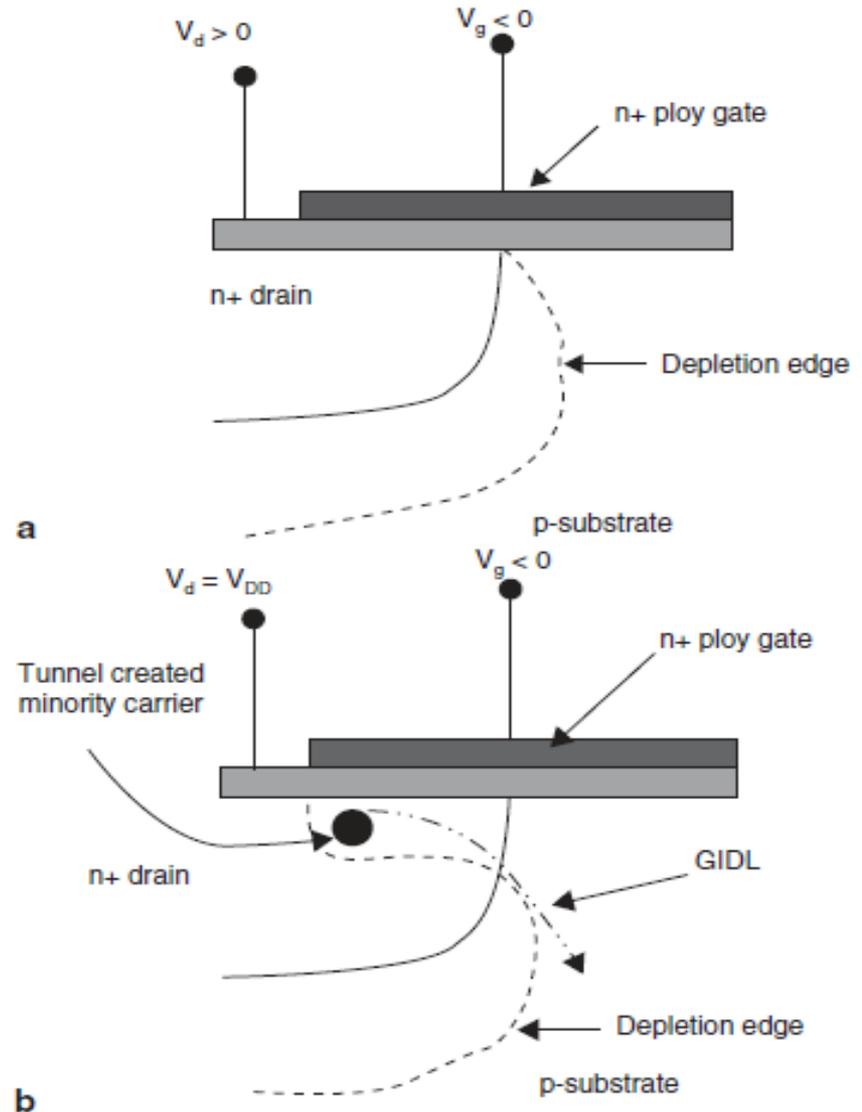


Gate-Induced Drain Leakage

- Due to a **high electric field** (V_{dg}) *across the oxide, a deep depletion region under the drain overlap region is created, which generates electrons and holes by band-to band tunneling.*
- The resulting drain to body current is called GIDL current.

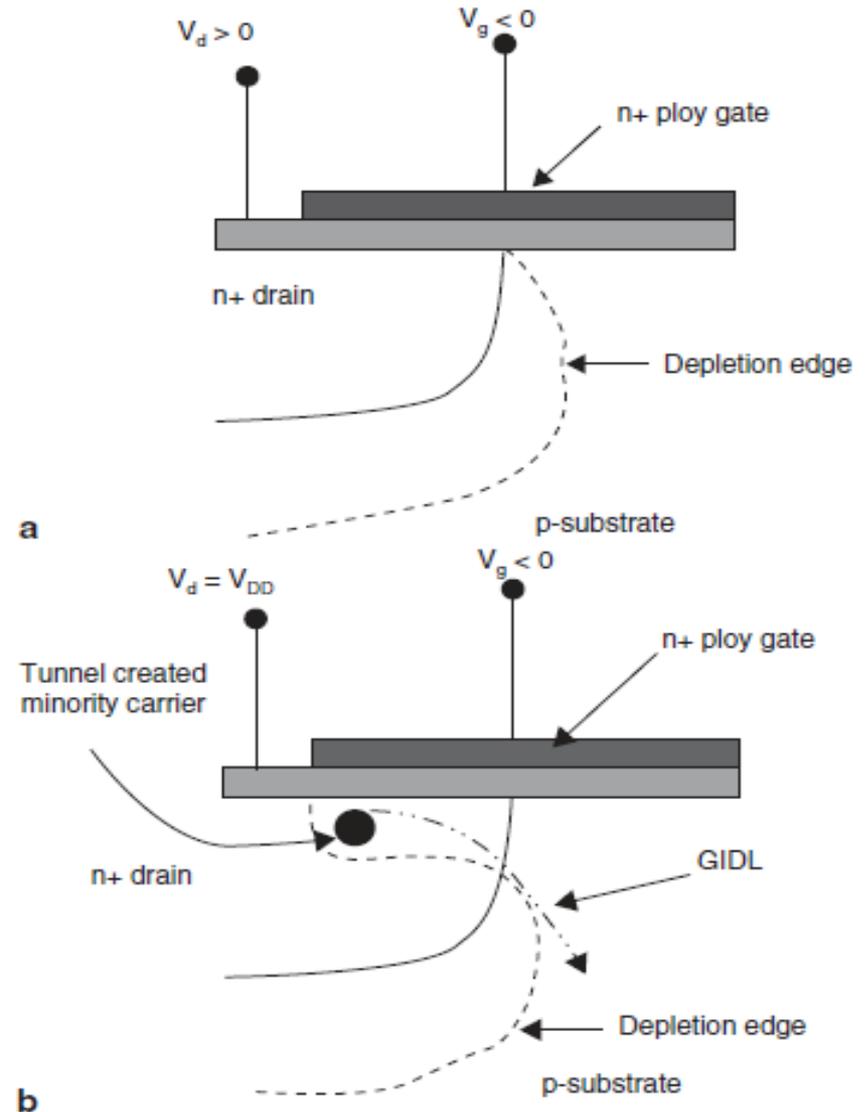
Gate-Induced Drain Leakage

- When the negative gate bias is large (v , gate at zero or negative voltages with respect to drain at V_{dd}), the n^+ region under the gate can be depleted and even inverted as shown in Fig. 6.31.



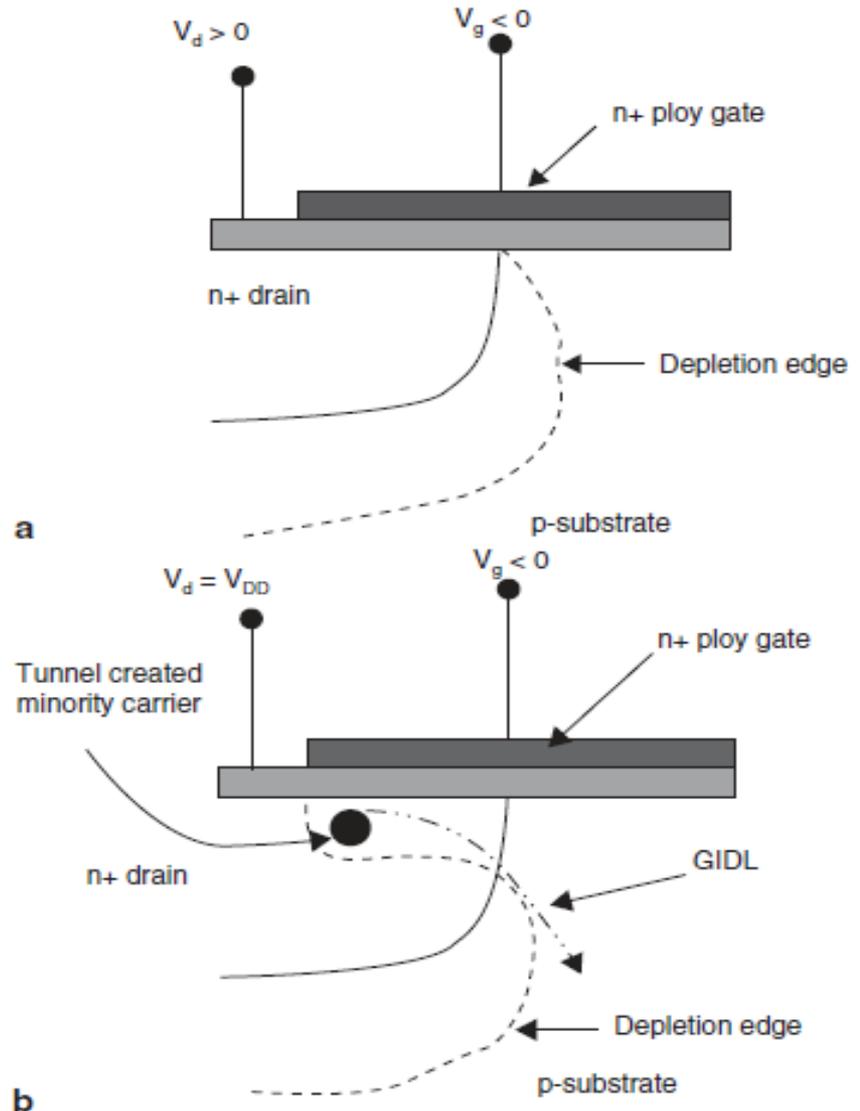
Gate-Induced Drain Leakage

- As a consequence, minority carriers are emitted in the drain region underneath the gate.
- As the substrate is at a lower potential, the minority carriers at the drain depletion region are swept away laterally to the substrate, creating a GIDL current.



Gate-Induced Drain Leakage

- At **low drain doping**, the electric field is not high enough to cause tunneling.
- At **very high drain doping**, the depletion width and hence the tunneling is limited.
- The GIDL is worse for moderated doping, when the depletion width and electric field are both considerable.



Punch-Through

- Due to the proximity of the drain and the source in short-channel devices, the depletion regions at the source–substrate and drain–substrate junctions extend into the channel.
- If the doping is kept constant while the channel length is reduced, the separation between the depletion region boundaries decreases.
- Increased reverse bias (higher V_{ds}) *across the junction further decreases the separation.*

Punch-Through

- When the depletion region merge, a majority of the carriers in the source enter into the substrate and get collected by the drain.
- This situation is known as punch-through condition.
- The net effect of punch through is an increase in the subthreshold current.
- Moreover, punch through degrades the subthreshold slope.

Punch-Through

- The punch-through voltage V_{PT} estimates the value of V_{ds} for which punch through occurs at $V_{gs} = 0$:

$$V_{PT} \propto N_B (L - W_j)^3,$$

where N_B is the doping concentration at the bulk, L is the channel length, and W_j is the junction width.

Supply Voltage Scaling for Low Power

- The **total power dissipation** can be represented by the simplified equation:

$$P_{\text{total}} = P_{\text{dynamic}} + P_{\text{static}}$$

- Although the **dynamic power has three components**, the **switching power** is the most dominant component.

Supply Voltage Scaling for Low Power

- The switching power $P_{\text{switching}} = \alpha_0 C_L V_{\text{dd}}^2 f$ caused by the **charging and discharging of capacitances** at different nodes of a circuit can be optimized by reducing each of the components such as the **clock frequency f** , **the total switched capacitance $\sum \alpha_i C_i$** , and **the supply voltage V_{dd}** .

Supply Voltage Scaling for Low Power

- Another dynamic power component, the **glitching power** is often neglected.
- But, it can account for up to **15 %** of the dynamic power.
- The third component, the **short-circuit power**, captures the power dissipation as a result of a short-circuit current, which flows between the supply voltage and ground (GND), when the CMOS logic gates switches from **0 to 1 or from 1 to 0**.
- This can be minimized by **minimizing the rise and fall times**.

Supply Voltage Scaling for Low Power

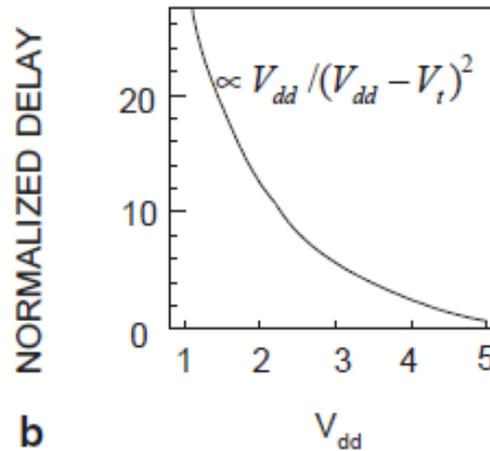
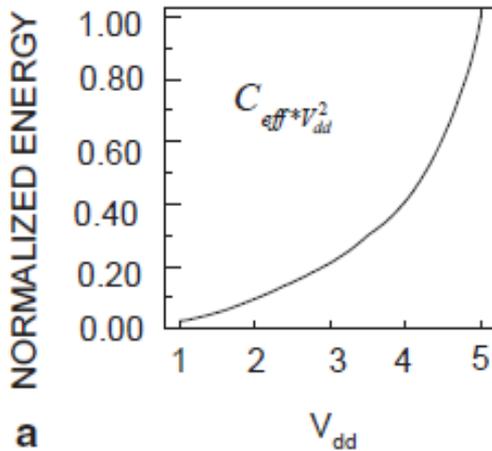
- The **static power dissipation** has also three dominant components.
- The most significant among them is the **sub-threshold leakage power** due to the flow of current between the drain and source.
- The second important component is the **gate leakage power** due to the tunneling of electrons from the bulk silicon through the gate oxide potential barrier into the gate.
- In **sub-50-nanometer devices**, the source–substrate and drain–substrate reversed p–n junction **band-to-band tunneling current**, the third component, is also large.

Supply Voltage Scaling for Low Power

- Because of the quadratic dependence of dynamic power on the supply voltage, supply voltage scaling was initially developed to reduce dynamic power.
- But, the supply voltage scaling also helps to reduce the static power because the subthreshold leakage power decreases due to the reduction of the drain induced barrier lowering (DIBL), the gate-induced drain leakage (GIDL), and the gate tunneling current as well.
- It has been demonstrated that the supply voltage scaling leads to the reduction of the subthreshold leakage and gate leakage currents of the order of V^3 and V^4 , respectively.

Supply Voltage Scaling for Low Power

- There is a performance penalty for the reduction in the supply voltage.
- If the threshold voltage is not scaled along with the supply voltage to avoid an increase in leakage current, a plot of the variation of the normalized delay with the supply voltage variation is shown in Fig. 7.1b.



$$\text{Delay} \propto \frac{V_{dd}}{(V_{dd} - V_t)^2} = \frac{1}{V_{dd} \left(1 - \frac{V_t}{V_{dd}}\right)^2}$$

a Variation of normalized energy with respect to supply voltage; **b** variation of delay with respect to supply voltage

Static Voltage Scaling

- **Device Feature Size Scaling (Physical Level Approach)**
 - *Constant Field Scaling*
 - *Constant Voltage Scaling*
- **Architectural-Level Approaches**
 - *Parallelism for Low Power*
 - *Multi-Core for Low Power*
 - *Pipelining for Low Power*
 - *Combining Parallelism with Pipelining*
 - *Voltage Scaling Using High-Level Transformations*

Device Feature Size Scaling

- In the **first physical-level-based approach**, the device feature size is scaled to overcome the loss in performance.
- Continuous improvements in process technology and photolithographic techniques have made the fabrication of MOS transistors of smaller and smaller dimensions to provide a higher packaging density.
- As a reduction in feature size reduces the gate capacitance, this leads to an improvement in performance.

Device Feature Size Scaling

- This has opened up the possibility of **scaling device feature sizes** to compensate for the loss in performance due to voltage scaling.
- The reduction of the size, i.e., the dimensions of MOSFETs, is commonly referred to as *scaling*.
- *To characterize the process of scaling, a parameter S , known as **scaling factor**, is commonly used.*
- *All **horizontal and vertical dimensions** are divided by this scaling factor, $S > 1$, to get the dimensions of the devices of the new generation technology.*

Device Feature Size Scaling

- Recent history of device size scaling

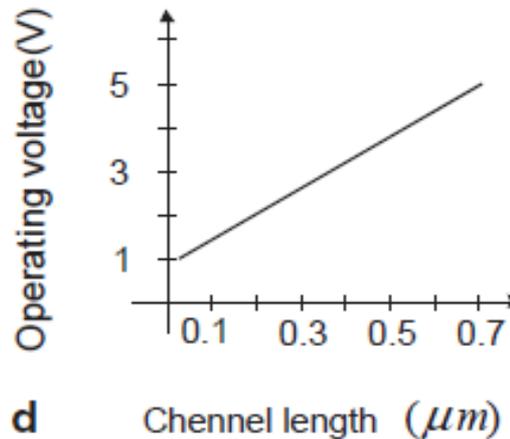
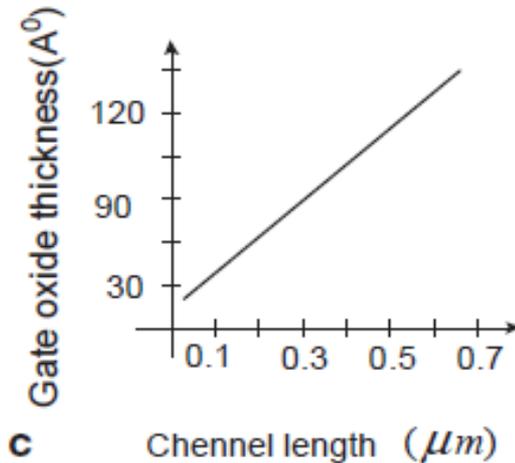
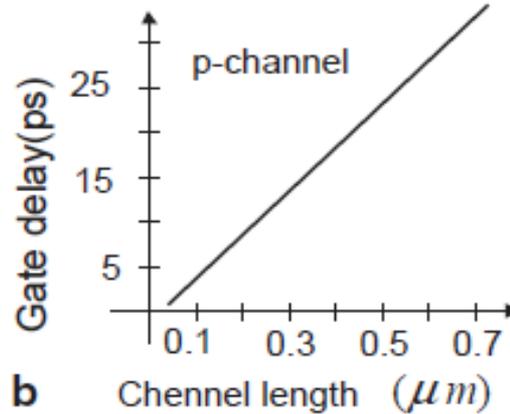
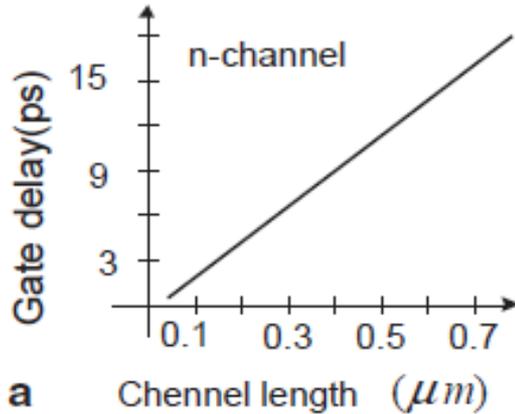
Table 7.1 Recent history of device size scaling for CMOS circuits

Year	1985	1987	1989	1991	1993	1995	1997	1999	2003	2005	2007	2009
Feature	2.5	1.7	1.2	1.0	0.8	0.5	0.35	0.25	0.18	0.090	0.065	0.045

CMOS complementary metal–oxide–semiconductor

Device Feature Size Scaling

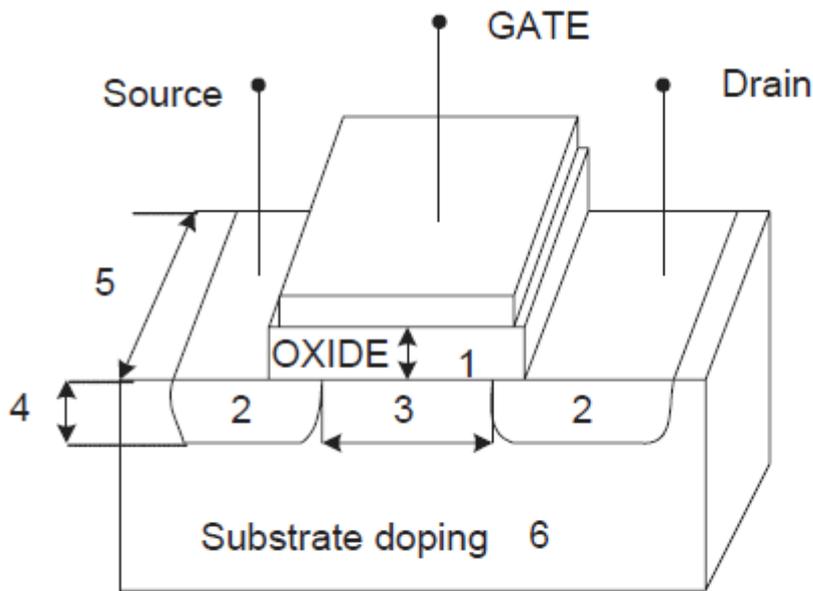
- Trends in metal–oxide–semiconductor (MOS) device scaling



It may be noted that the slope of all the curves in this figure is equal to the scaling parameter S .

Device Feature Size Scaling

- Basic geometry of an MOSFET and the various parameters scaled by a scaling factor S .



- $t_{ox}' = t_{ox} / S$
- $N_D' = N_D \times S$
- $L' = L / S$
- $X_j' = X_j / S$
- $W' = W / S$
- $N_A' = N_A / S$

• It may be noted that all the three dimensions are proportionally reduced along with a corresponding increase in doping densities.

• There are two basic approaches of device size scaling—*constant-field scaling* and *constant voltage scaling*.

Constant-Field Scaling

- In this approach, the magnitudes of all the internal electric fields within the device are preserved, while the dimensions are scaled down by a factor of S .
- *This requires* that all potentials must be scaled down by the same factor.
- Accordingly, supply and threshold voltages are scaled down proportionately.
- This also dictates that the doping densities are to be increased by a factor of S *to preserve the field conditions.*

Constant-Field Scaling

Table 7.2 Constant-field scaling of the device dimensions, voltages, and doping densities

Quantity	Before scaling	After scaling
Channel length	L	$L' = L / S$
Channel width	W	$W' = W / S$
Gate oxide thickness	t_{ox}	$t'_{\text{ox}} = t_{\text{ox}} / S$
Junction depth	x_j	$x'_j = x_j / S$
Power supply voltage	V_{dd}	$V'_{\text{dd}} = V_{\text{dd}} / S$
Threshold voltage	V_{T0}	$V'_{\text{T0}} = V_{\text{T0}} / S$
Doping densities	N_{A}	$N'_{\text{A}} = N_{\text{A}} \cdot S$
	N_{D}	$N'_{\text{D}} = N_{\text{D}} \cdot S$

- As a consequence of scaling, various electrical parameters are affected.
- For example, the gate oxide capacitance per unit area increases by a factor of S as given by the following relationship:

$$C'_{\text{ox}} = \frac{\epsilon_{\text{ox}}}{t'_{\text{ox}}} = S \frac{\epsilon_{\text{ox}}}{t_{\text{ox}}} = S \cdot C_{\text{ox}}.$$

Constant-Field Scaling

- As both length and width parameters are scaled down by the same factor, the W/L remains unchanged.
- So, the transconductance parameter K_n is also scaled by a factor S .
- Both linear-mode and saturation-mode drain currents are reduced by a factor of S , as given below:

$$\begin{aligned} I'_{ds}(\text{lin}) &= \frac{K'_n}{2} \left(2(V'_{gs} - V'_T) \cdot V'_{ds} - V'^2_{ds} \right) \\ &= \frac{SK_n}{2} \cdot \frac{1}{S^2} \left(2(V_{gs} - V) V_{ds} - V^2_{ds} \right) \\ &= \frac{I_D(\text{lin})}{S}. \end{aligned}$$

Constant-Field Scaling

- Effects of constant-field scaling on the key device parameters

Table 7.3 Effects of constant-field scaling on the key device parameters

Quality	Before scaling	After scaling
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	$P'/\text{area}' = (P/\text{area})$
Delay	t_d	$t'_d = t_d / S$
Energy	$E = P \cdot t_d$	$E' = \frac{P}{S^2} \times \frac{t_d}{S} = \frac{P \cdot t_d}{S^3} = \frac{1}{S^3} E$

Constant-Field Scaling

- Important benefits of constant-field scaling are:
 - (i) Smaller device sizes leading to a reduced chip size, higher yield, and more number of integrated circuits (ICs) per wafer,
 - (ii) Higher speed of operation due to smaller delay
 - (iii) Reduced power consumption because of the smaller supply voltage and device currents.

Constant-Voltage Scaling

- It may be necessary to use **multiple supply voltages** and complicated-level translators to resolve this problem.
- In such situations, constant-voltage scaling may be preferred.
- In a **constant-voltage scaling approach**, **power supply voltage** and the **threshold voltage** of the device remain unchanged.
- To preserve the charge–field relations, however, the doping densities have to be scaled by a factor of S^2 .

Short-Channel Effects

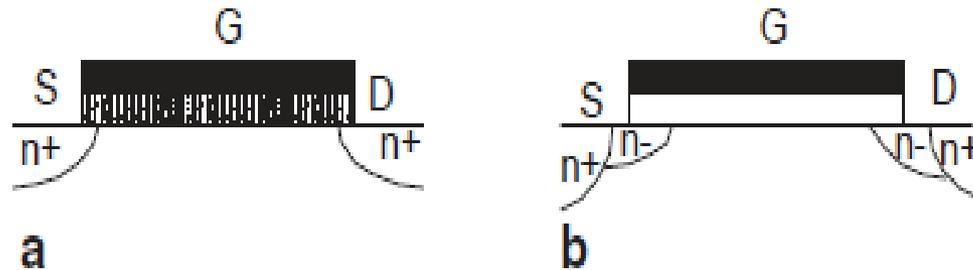
- Short-channel effects arise when channel length is of the same order of magnitude as depletion region thickness of the source and drain junctions or when the length is approximately equal to the source and drain junction depths.
- As the channel length is reduced below $0.6\ \mu\text{m}$, the short-channel effect starts manifesting.
- This leads to an increase in subthreshold leakage current, reduction in threshold voltage with V_{gs} , *and a linear increase in the saturation current* instead of square of the gate-to-source voltage.

Short-Channel Effects

- Moreover, if channel length is scaled down without scaling the supply voltage (constant-voltage scaling), the electric field across a gate oxide device continues to increase, creating a hot carrier.
- Hot carriers can cause an avalanche breakdown of the gate oxide.
- It is necessary to restrict the maximum electric field across the gate oxide to 7 MV/cm, which translates into 0.7 V/10 Å of gate oxide thickness.
- For gate oxide thickness of 70 Å, the applied gate voltage should be limited to 4.9 V for long-term reliable operation.

Short-Channel Effects

- One technique is to use *lightly doped drain structure* shown in Fig. 7.4.
- The physical device structure is modified so that the carriers do not gain energy from the field to become hot carriers.
- Of course, the performance of the device is traded to obtain long-term reliability.



a Conventional structure; **b** lightly doped drain structure

Architectural-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

- Parallel processing is traditionally used for the **improvement of performance** at the expense of a **larger chip area and higher power dissipation**.
- Basic idea is to use multiple copies of hardware resources, such as arithmetic logic units (ALUs) and processors, to operate in parallel to provide a higher performance.
- Instead of using parallel processing for improving performance, it can also be used to reduce power.

Parallelism for Low Power

- We know that supply voltage scaling is the most effective way to reduce power consumption.
- Unfortunately, the savings in power come at the expense of performances or, more precisely, maximum operating frequency.
- This follows from the equation:

$$f_{\max} \propto (V_{\text{dd}} - V_t)^2 / V_{\text{dd}} = V_{\text{dd}} \left(1 - \frac{V_t}{V_{\text{dd}}} \right)^2 .$$

Parallelism for Low Power

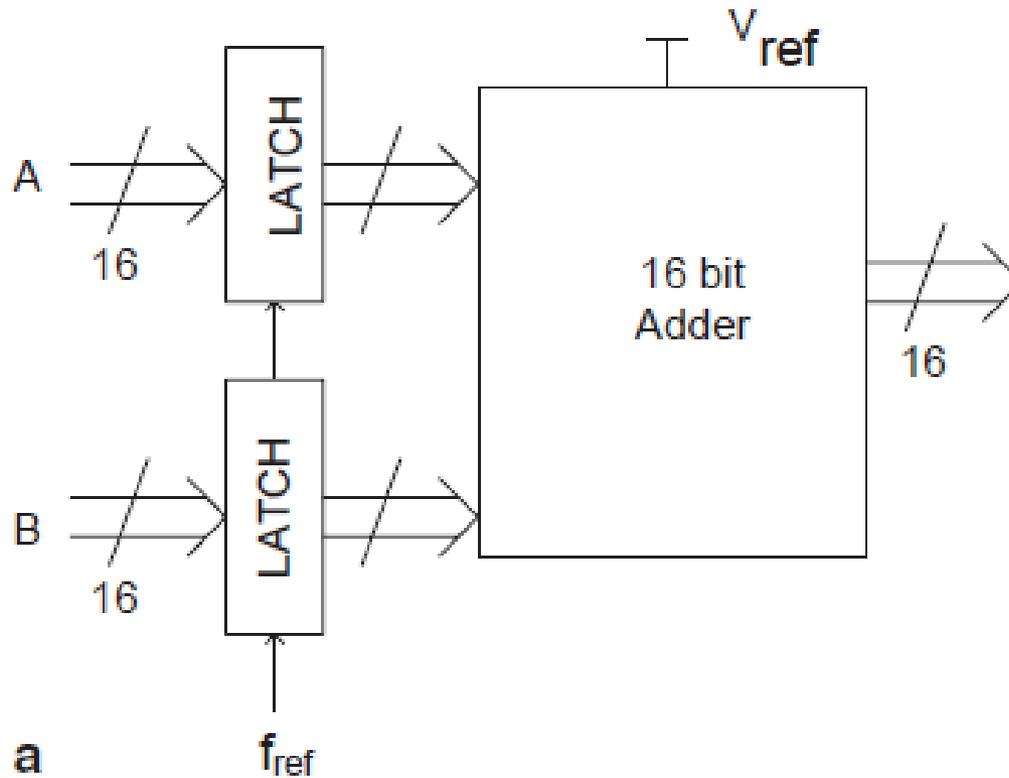
- If the threshold voltage is scaled by the same factor as the supply voltage, the maximum frequency of operation is roughly linearly dependent on the power supply voltage.
- Reducing the supply voltage forces the circuit to operate at a lower frequency.
- In simple terms, if the supply voltage is reduced by half, the power is reduced by one fourth and performance is lowered by half.

Parallelism for Low Power

- The loss in performance can be compensated by parallel processing.
- This involves splitting the computation into two independent tasks running in parallel.
- This has the potential to reduce the power by half without reduction in the performance.
- Here, the basic approach is to trade the area for power while maintaining the same throughput.

Parallelism for Low Power

- 16-bit adder



Parallelism for Low Power

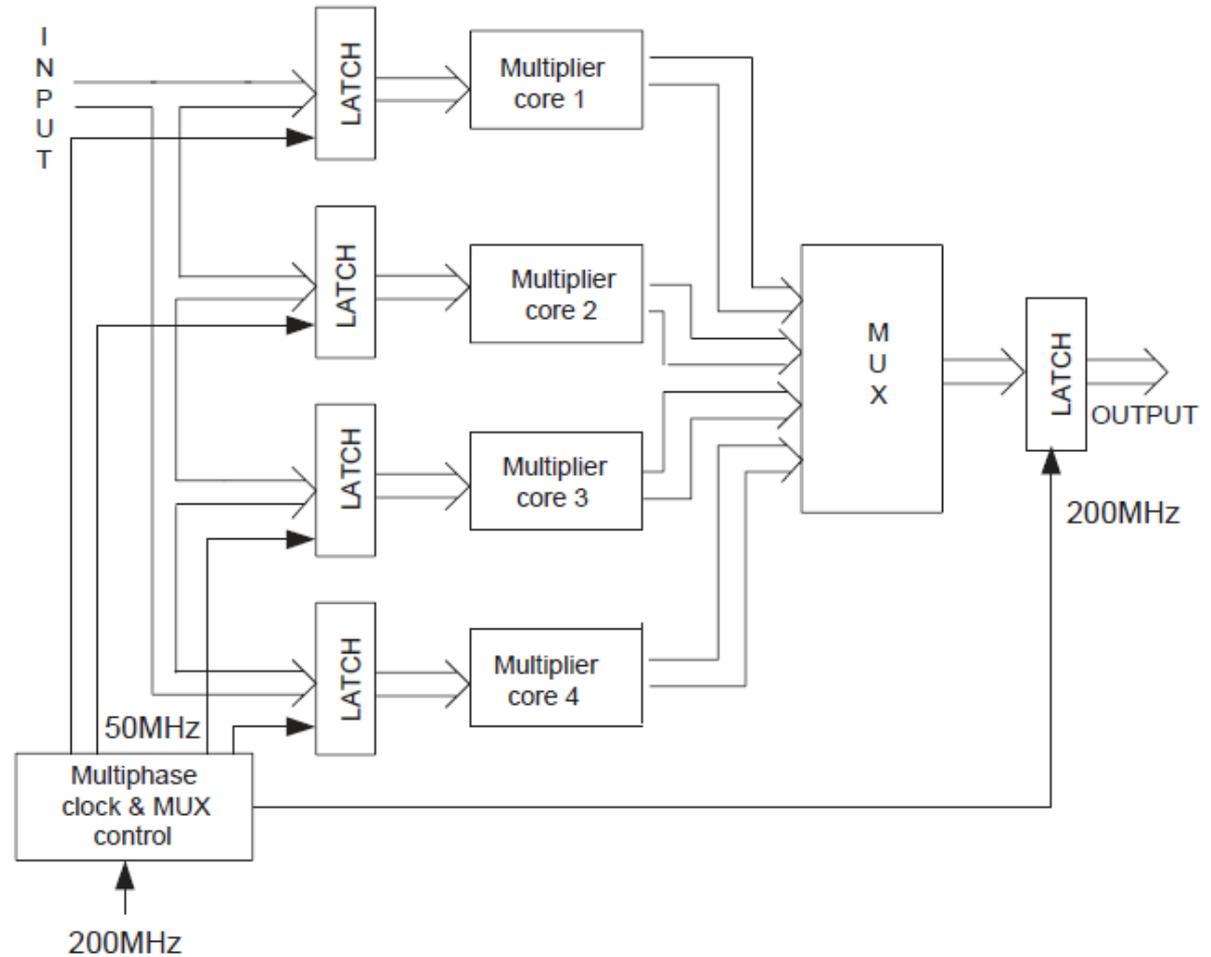
- Impact of parallelism on area, power, and throughput

Table 7.6 Impact of parallelism on area, power, and throughput

Parameter	Without V_{dd} scaling	With V_{dd} scaling
Area	2.2X	2.2X
Power	2.2X	0.227X
Throughput	2X	1X

Multi-Core for Low Power

- Four-core Multiplier Architecture.
MUX
Multiplexer



Multi-Core for Low Power

Table 7.7 Power in multi-core architecture

Number of cores	Clock in MHz	Core supply voltage	Total power
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

- Pipelined realization 16-bit adder

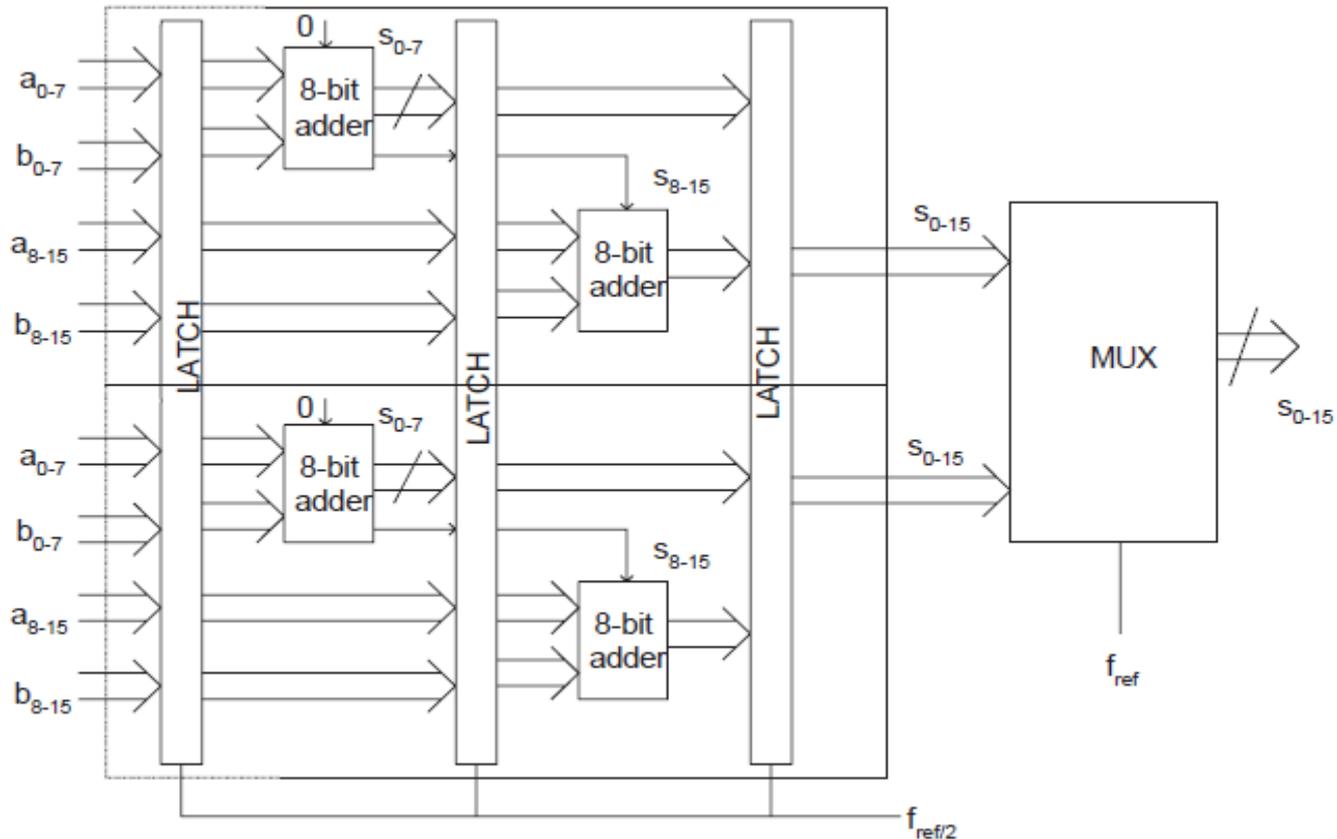
Table 7.8 Impact of pipelining on area, power, and throughput

Parameter	Without V_{dd} scaling	With V_{dd} scaling
Area	1.15X	1.15X
Power	2.30X	0.28X
Throughput	2X	1X

$$P_{\text{pipe}} = C_{\text{pipe}} \cdot V_{\text{pipe}}^2 \cdot f_{\text{pipe}} = (1.15C_{\text{ref}}) \cdot \left(\frac{V_{\text{ref}}}{2}\right)^2 \cdot f = 0.28P_{\text{ref}}.$$

Combining Parallelism with Pipelining

- Parallel-pipelined realization of 16-bit adder.
MUX multiplexer

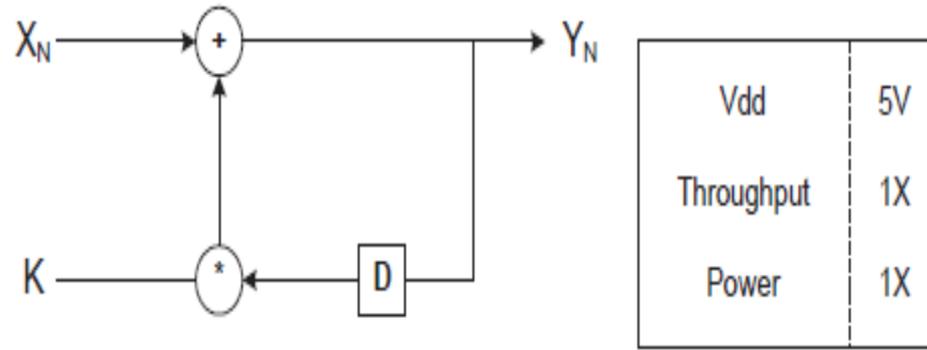


Voltage Scaling Using High-Level Transformations

- 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**.
- loop unrolling can be used to minimize power by voltage scaling

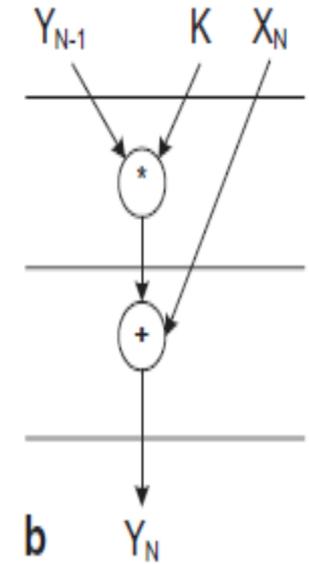
Voltage Scaling Using High-Level Transformations

- a A first-order infinite impulse response (IIR) filter;
- b directed acyclic graph (DAG) corresponding to the IIR filter



a

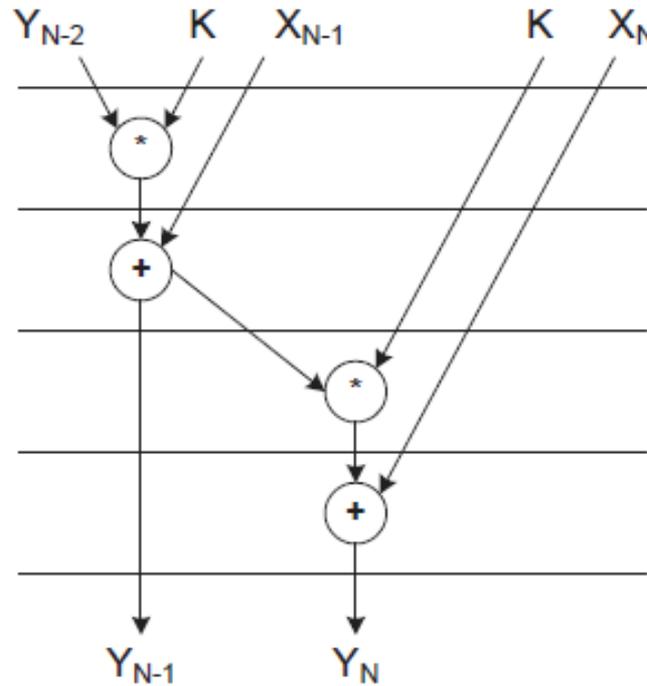
$$Y_N = X_N + K \times Y_{N-1}$$



Voltage Scaling Using High-Level Transformations

- Directed acyclic graph (DAG) after unrolling

Vdd	5V
Throughput	1X
Power	1X

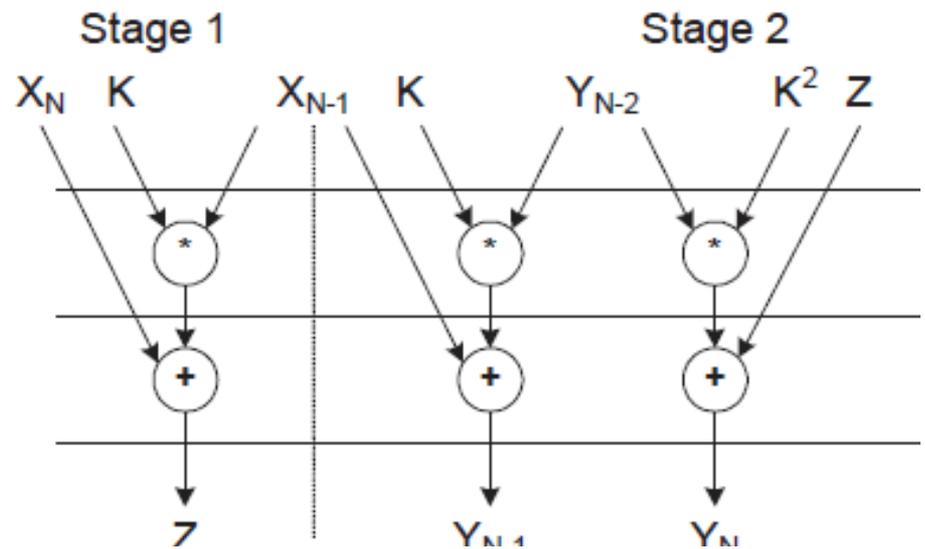


Voltage Scaling Using High-Level Transformations

- Directed acyclic graph (DAG) after unrolling and pipelining

Vdd	5V	2.9V
Throughput	2X	1X
Power	1.5X	0.5X

$$Z = X_N + K \times X_{N-1}$$

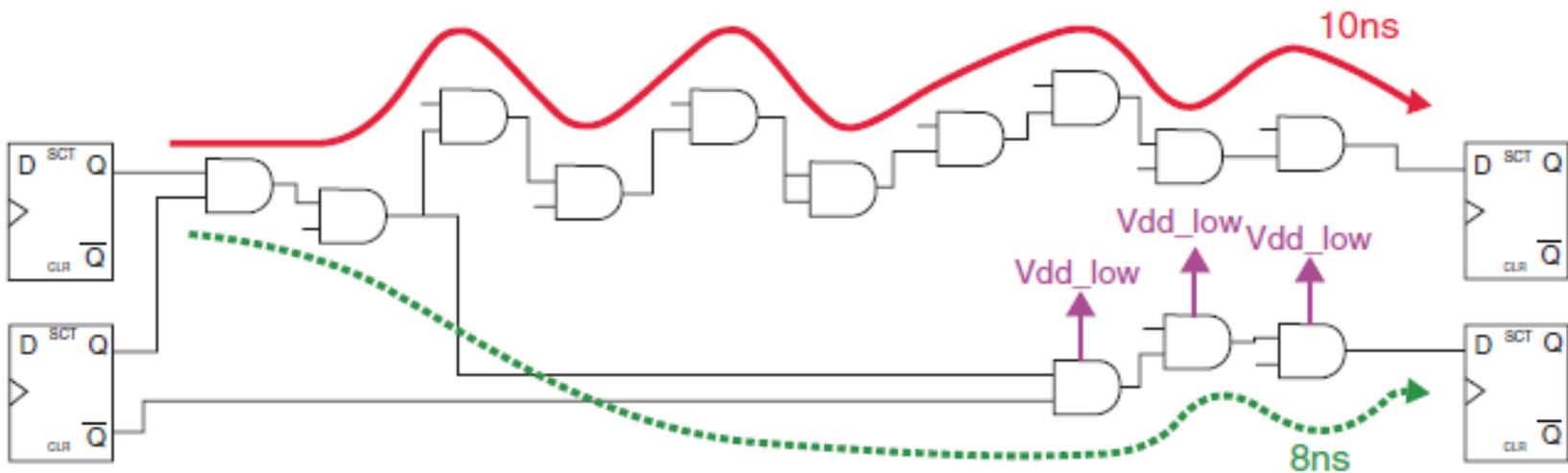


$$Y_{N-1} = X_{N-1} + K \times Y_{N-2}$$

$$Y_N = Z + K^2 \times Y_{N-2}$$

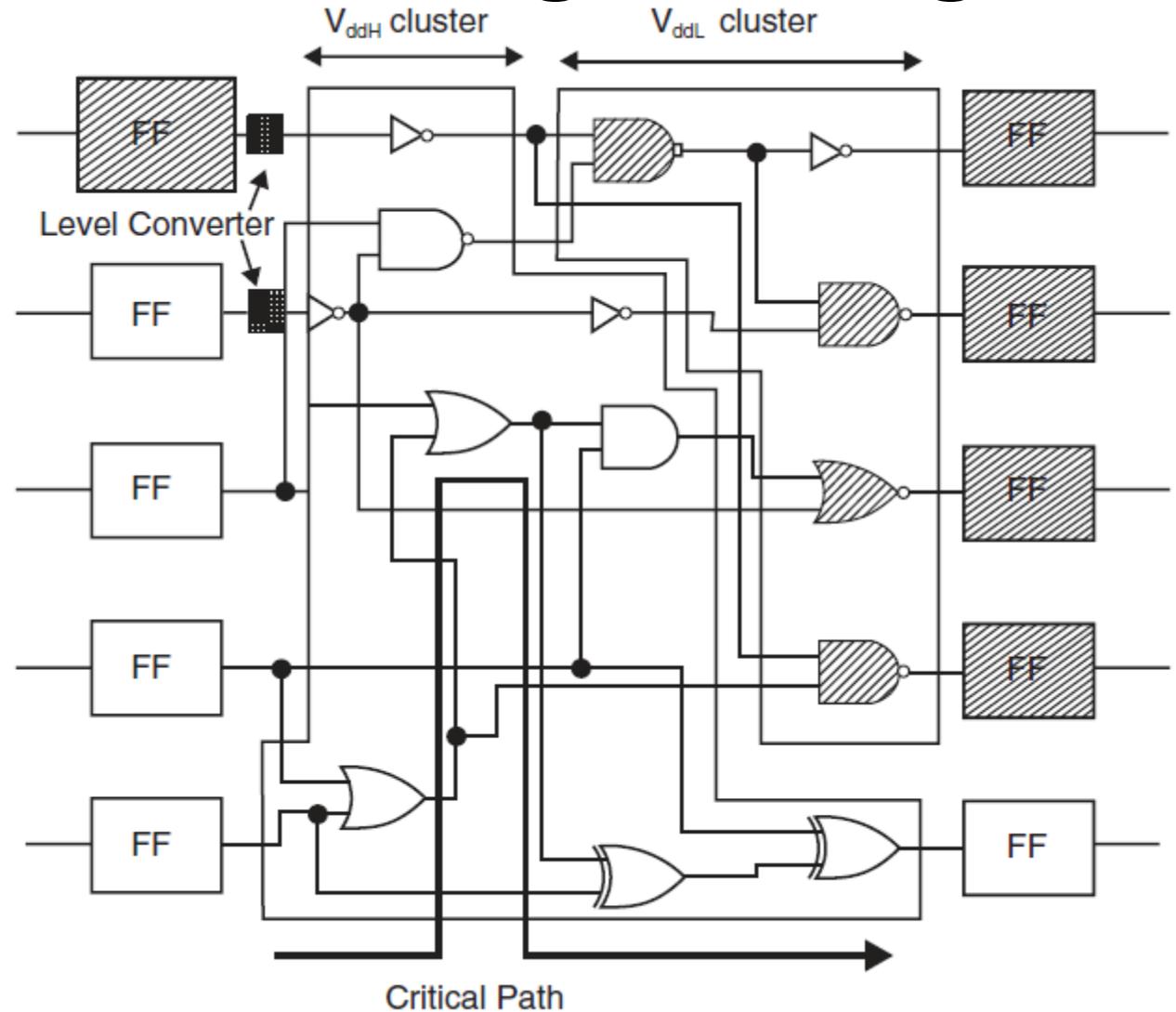
Multilevel Voltage Scaling

- Assignment of multiple supply voltages based on delay on the critical path



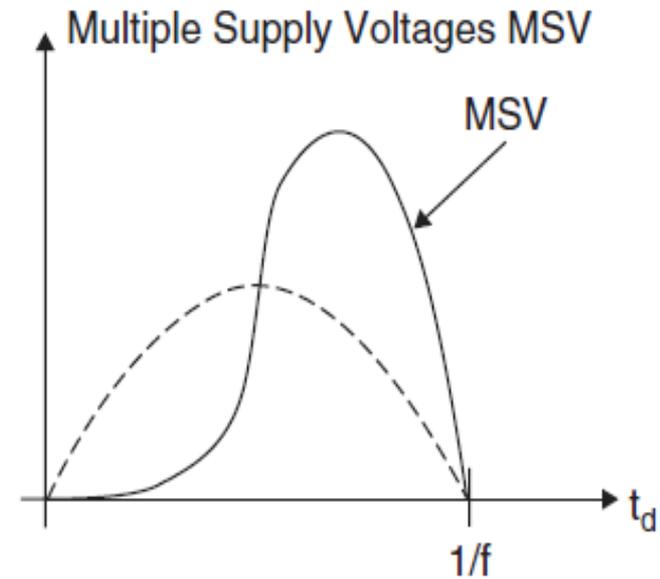
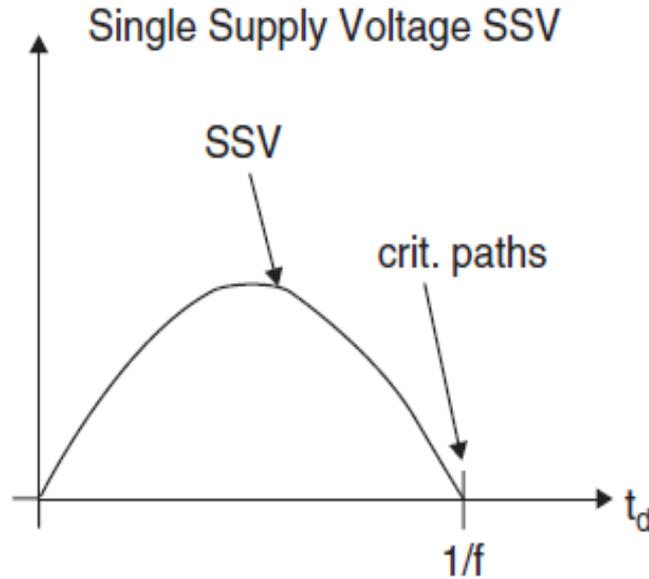
Multilevel Voltage Scaling

Clustered
voltage
scaling.
FF flip-
flop

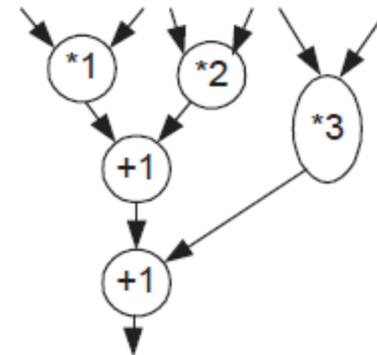


Multilevel Voltage Scaling

- Distribution of path delays under single supply voltage (SSV) and multiple supply voltage (MSV)



Macro-based voltage island approach to achieve low power



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 V_{dd} saturates quickly.

Multilevel Voltage Scaling

- Challenges in MVS
- 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)

- 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.

Dynamic Voltage and Frequency Scaling(DVFS)

- 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.

Dynamic Voltage and Frequency Scaling

- Basic Approach
- Dynamic Frequency Scaling
- Dynamic Voltage and Frequency Scaling
- DVFS with Varying Work Load

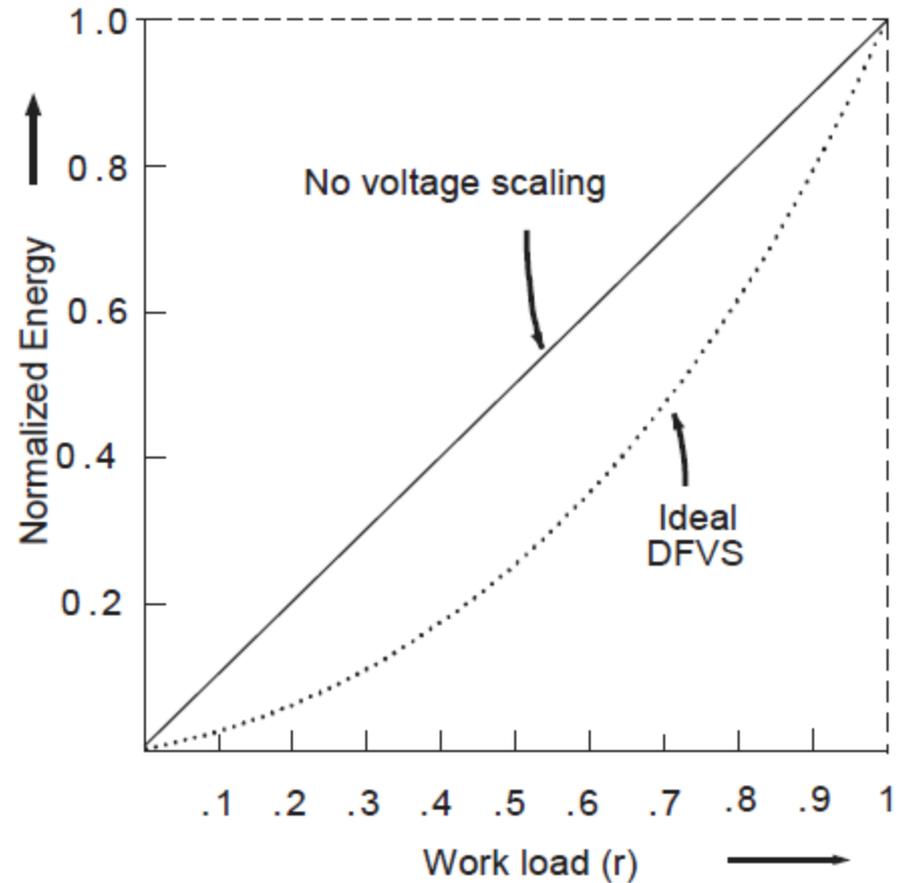
Dynamic Voltage and Frequency Scaling

- **Basic Approach**
- The energy drawn from the power supply can be reduced by using the following two approaches:
- **Dynamic Frequency Scaling**
- **Dynamic Voltage and Frequency Scaling**

Dynamic Voltage and Frequency Scaling

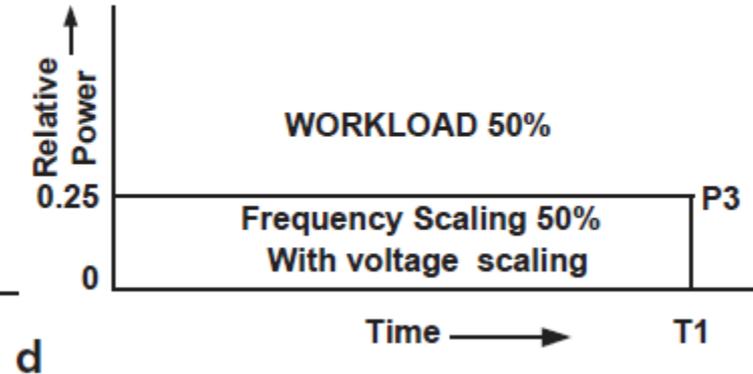
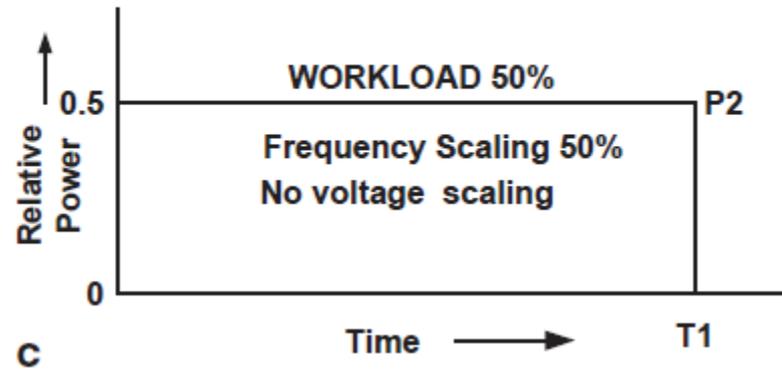
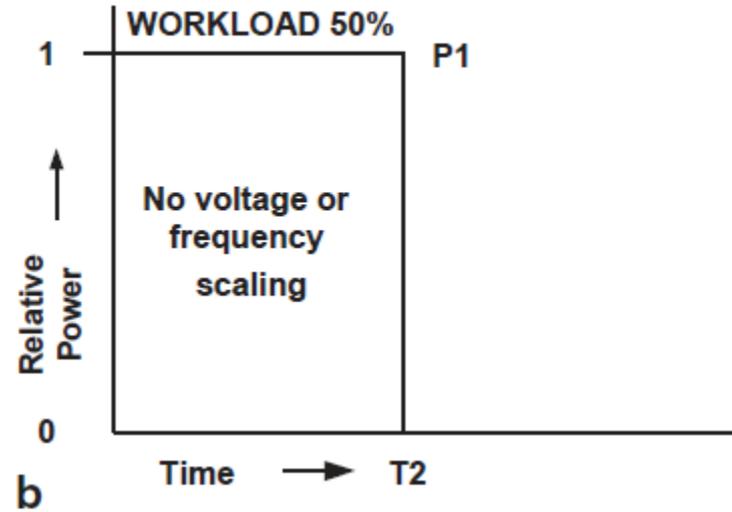
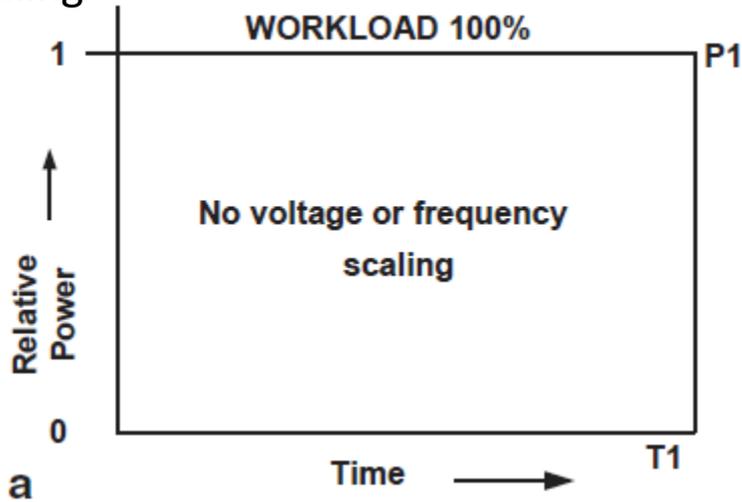
- Dynamic Frequency and voltage Scaling

$$\text{Delay}(D) \propto \frac{V_{dd}}{(V_{dd} - V_t)^2} = \frac{1}{V_{dd} \left(1 - \frac{V_t}{V_{dd}}\right)^2}$$



DVFS

Four different cases with two different workloads and with voltage and frequency scaling



DVFS with Varying Work Load

- Variable voltage processor $\mu(r)$
- The need of a processor which can operate over a frequency range with a corresponding lower supply voltage range can be manufactured using the present-day process technology and several such processors are commercially available.

DVFS with Varying Work Load

- Variable voltage processor $\mu(r)$
- Transmeta's TM 5400 or "Crusoe" processor and Strong ARM processor are examples of such variable voltage processors.
- Transmeta's Crusoe processor can operate over the voltage range of 1.65–1.1 V, with the corresponding frequency range of 700– 200 MHz.

DVFS with Varying Work Load

Table 7.10 Relationship between voltage, frequency, and power

Frequency (f) MHz	Voltage V_{dd}	Relative power
700	1.65	100
600	1.60	80.59
500	1.50	59.03
400	1.40	41.14
300	1.25	24.60
200	1.10	12.70

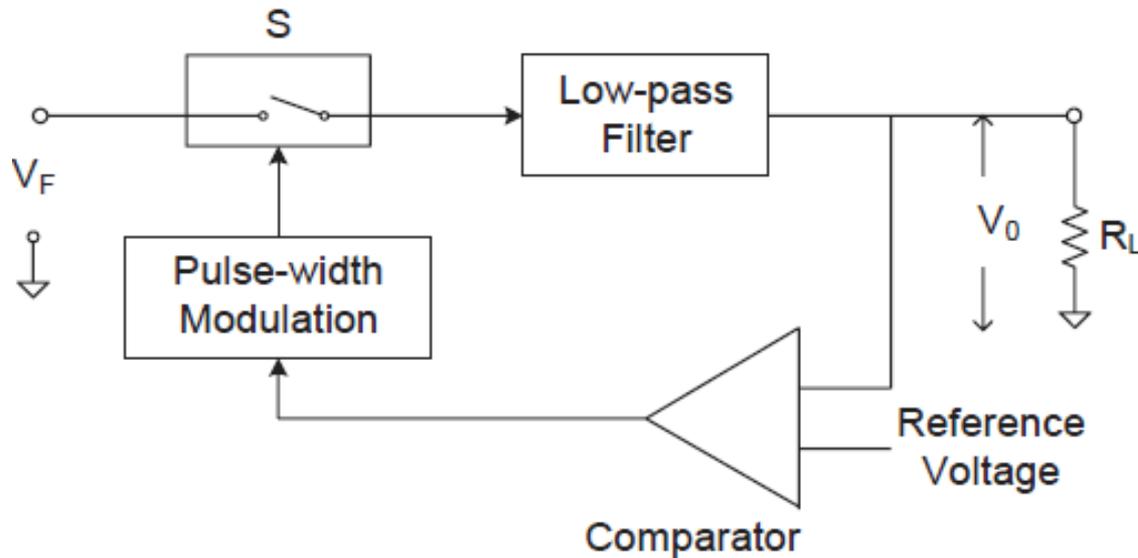
- Variable voltage processor $\mu(r)$

It allows the following adjustments:

- Frequency change in steps of 33 MHz
- Voltage change in steps of 25 mV
- Up to 200 frequency/voltage change per second

DVFS with Varying Work Load

- Variable voltage generator $V(r)$



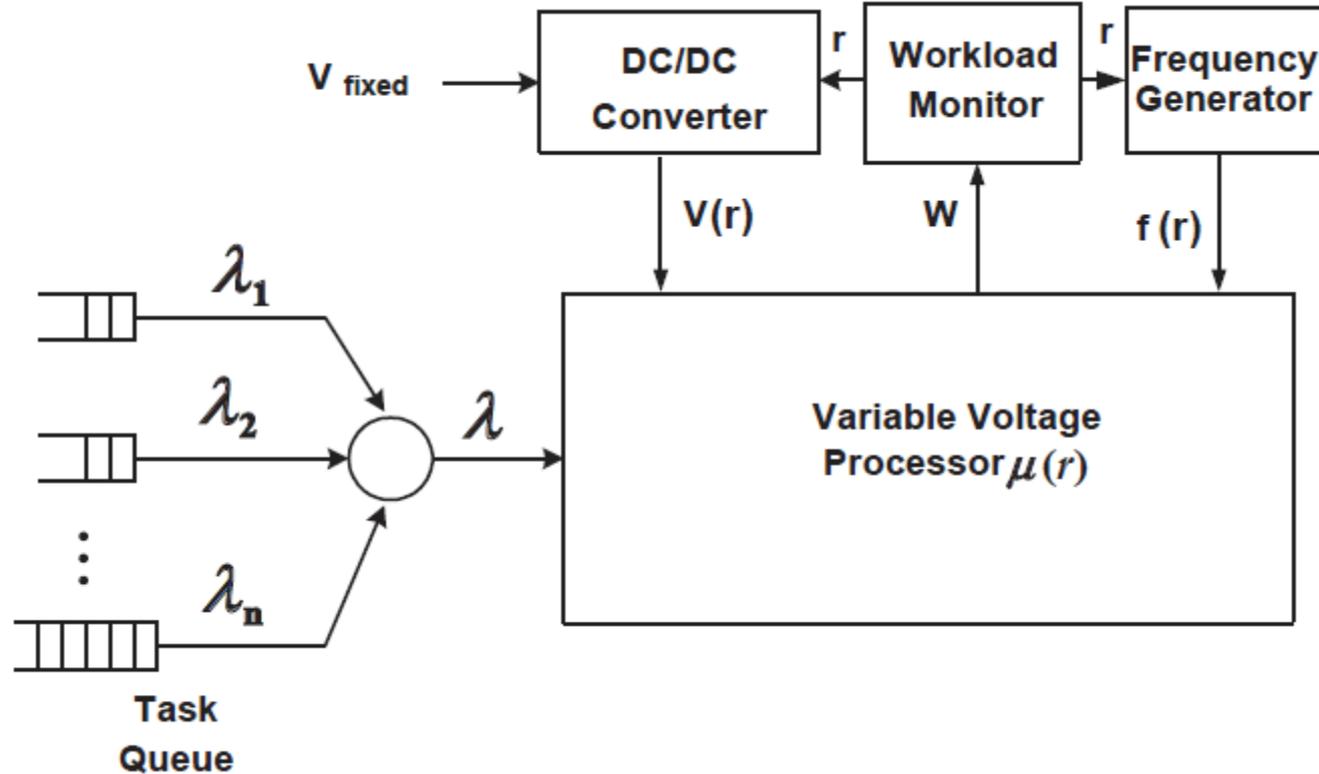
Block diagram of a direct current (DC)-to-DC converter

DVFS with Varying Work Load

- Variable frequency generator $f(r)$
- The variable frequency is generated with the help of a phase lock loop (PLL) system.
- The heart of the device is the high-performance PLL-core, consisting of a phase frequency detector (PFD), programmable on-chip filter, and voltage-controlled oscillator (VCO).
- The PLL generates a high-speed clock which drives a frequency divider.
- The divider generates the variable frequency $f(r)$.
- *The PLL and the divider together generate the independent frequencies related to the PLL operating frequency.*

DVFS

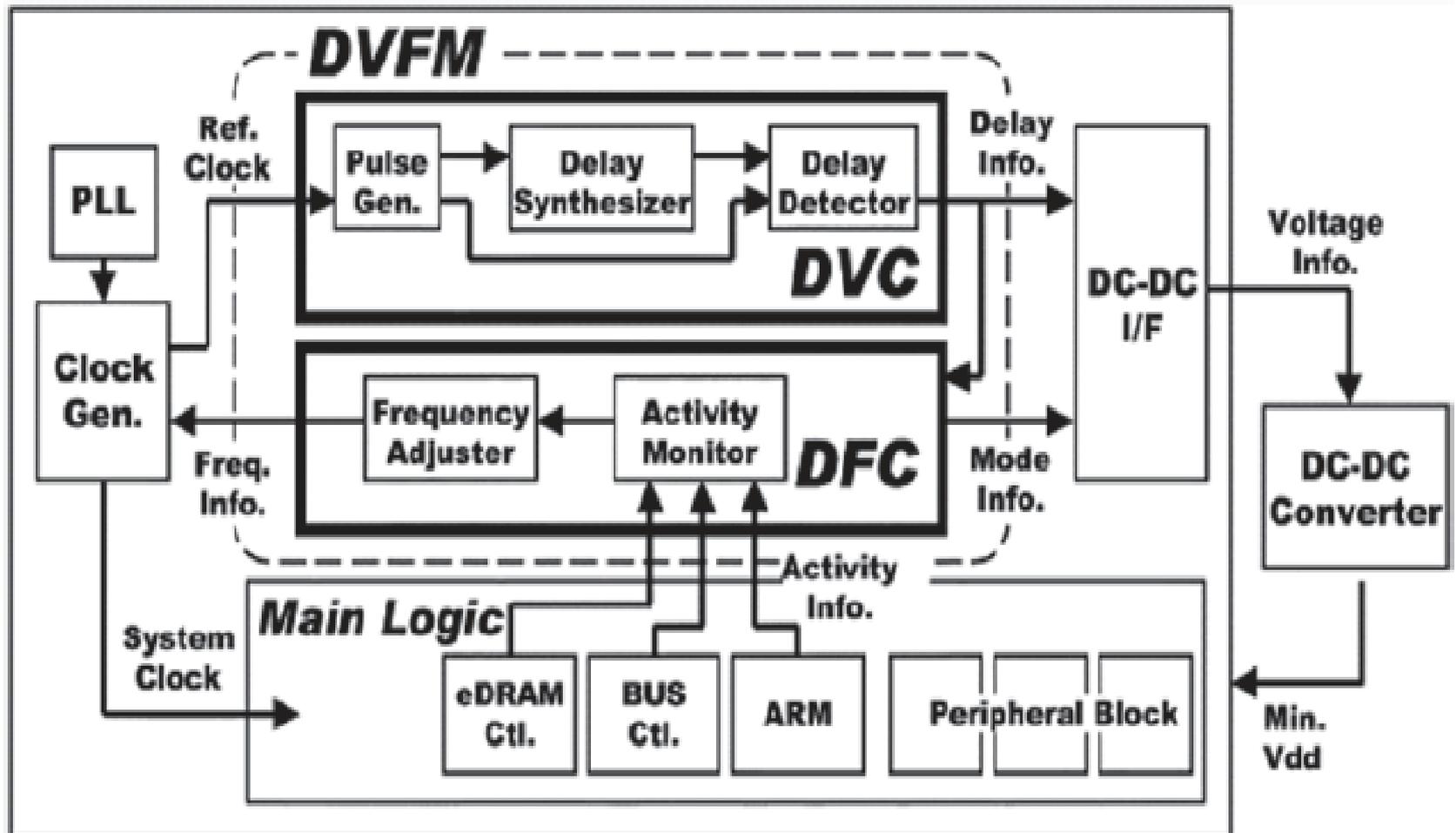
- Model for dynamic voltage scaling



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.

Adaptive Voltage Scaling



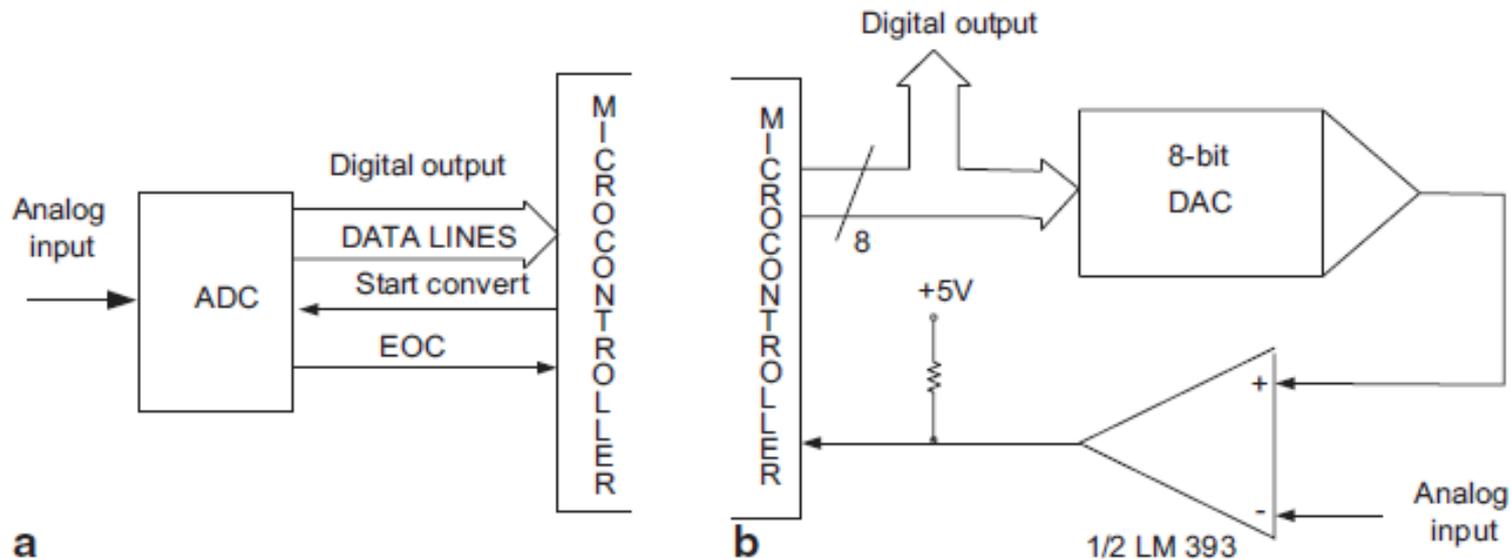
Unit-4

Switched Capacitance Minimization

- **System-Level Approach: Hardware–Software Co-design**
- **Transmeta's Crusoe Processor**
- **Bus Encoding**
- **Clock gating**
- **Gated-clock FSMs, FSM State Encoding**
- **FSM Partitioning**
- **Operand isolation**
- **Pre-computation**
- **Logic Styles for Low Power**

System-Level Approach: Hardware–Software Co-design

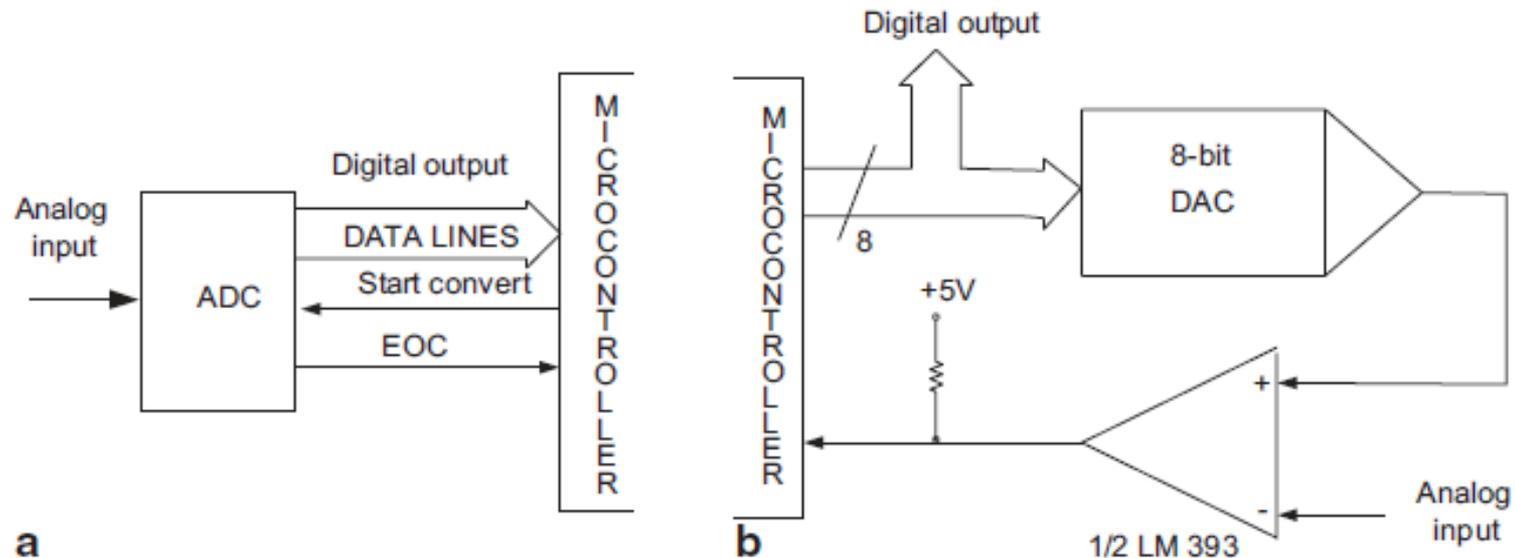
- It is well known that a particular functionality can be realized purely by hardware, or by a combination of both hardware and software



a Analog-to-digital converter (ADC) implemented by hardware and **b** ADC implemented by hardware–software mix

System-Level Approach: Hardware–Software Co-design

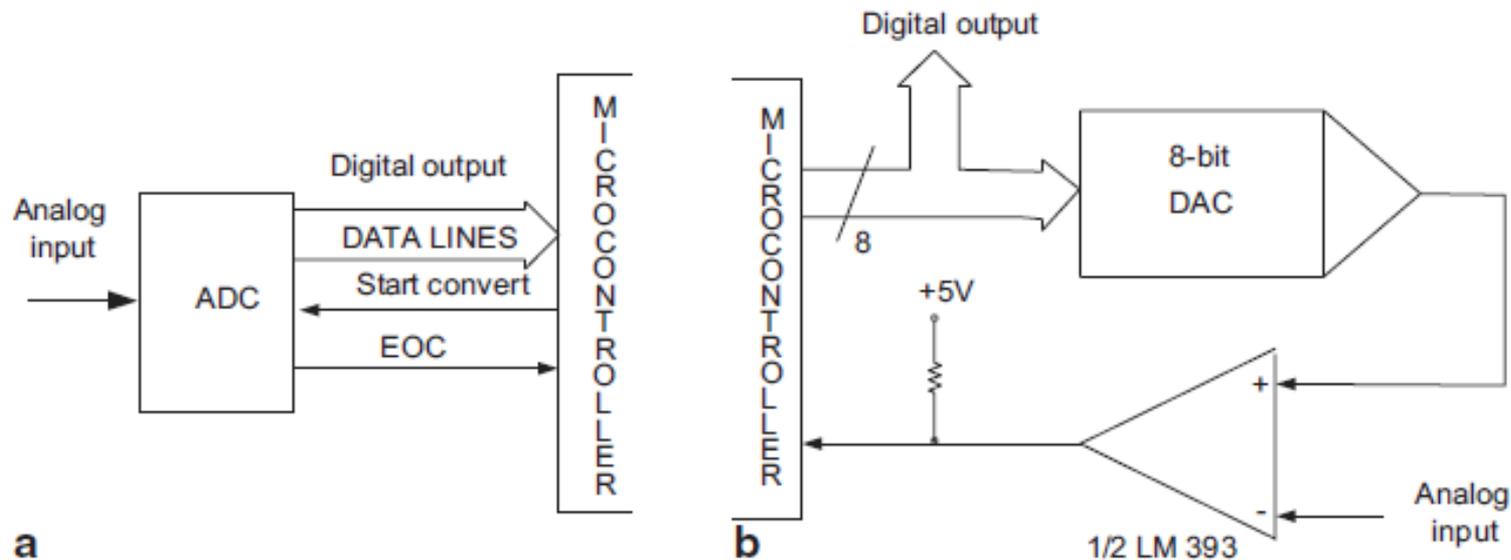
- **Approach-I:** This approach involves the use of a costly ADC chip along with a few lines of program code to read the ADC data.



a Analog-to-digital converter (ADC) *implemented by hardware* and **b** ADC *implemented* by hardware–software mix

System-Level Approach: Hardware–Software Co-design

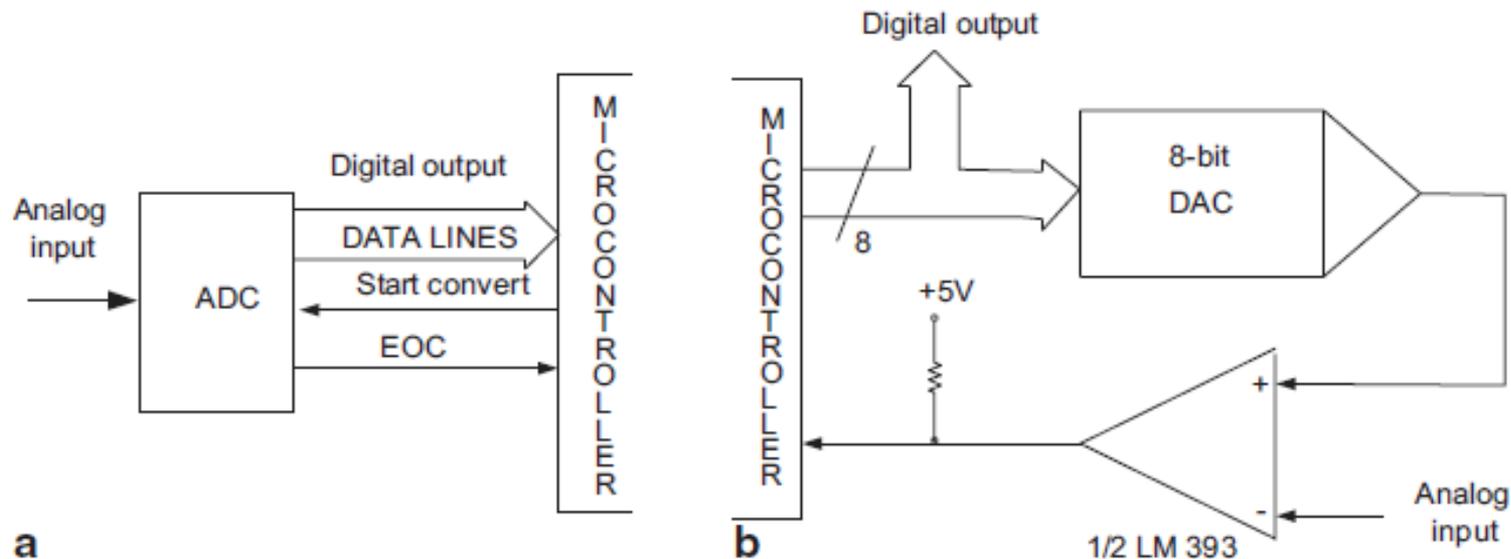
- **Approach-I:**
- The ADC chip can be selected based on the sampling rate and the precision of the digital data.
- The software overhead is very small.



a Analog-to-digital converter (ADC) implemented by hardware and **b** ADC implemented by hardware–software mix

System-Level Approach: Hardware–Software Co-design

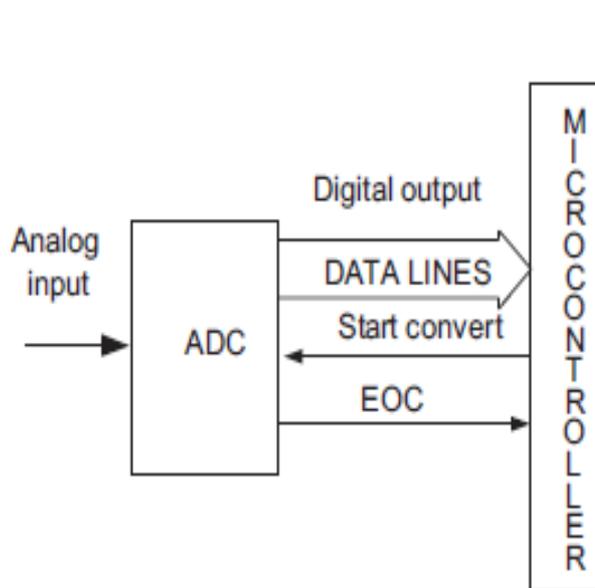
- **Approach-I:**
- This approach can provide higher performance in terms of conversion time and sampling rate.
- However, it involves **higher cost and higher power dissipation.**



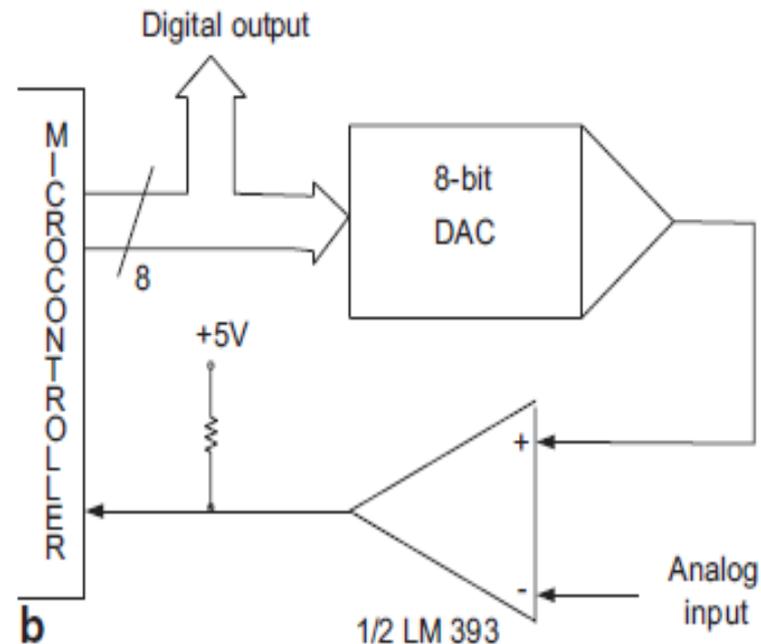
a Analog-to-digital converter (ADC) implemented by hardware and **b** ADC implemented by hardware–software mix

System-Level Approach: Hardware– Software Co-design

- **Approach-II:** ADC functionality can be implemented by software with few inexpensive external components such as a digital-to-analog converter (DAC) and a comparator



a



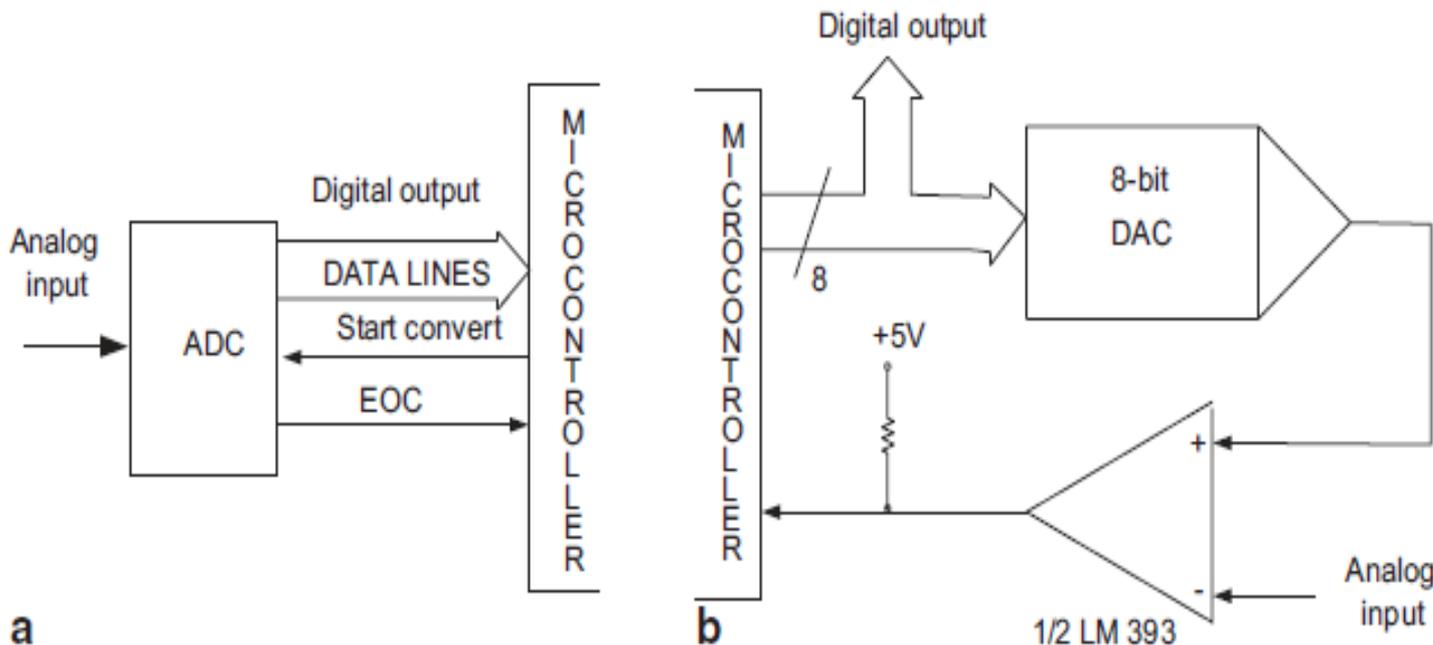
b

a Analog-to-digital converter (ADC) implemented by hardware and b ADC implemented by hardware–software mix

System-Level Approach: Hardware–Software Co-design

- **Approach-II:**

- Any of the A/D conversion algorithms such as a **successive approximation** can be implemented by the software utilizing the DAC and the comparator.
- As a consequence, it will have **higher software overhead**, but **lesser hardware cost of implementation**.



a Analog-to-digital converter (ADC) implemented by hardware and b ADC implemented by hardware–software mix

System-Level Approach: Hardware– Software Co-design

- The **first approach** provides a **fast conversion time** at the **cost of higher cost or larger chip area**.
- In the **second alternative**, the **hardware cost and chip area are lower**, but **conversion time is longer**.
- So, for a given application, there is a **trade-off** between how much is implemented by hardware and by software.

System-Level Approach: Hardware– Software Co-design

- This has led to the concept of hardware–software co-design, which involves partitioning of the system to be realized into two distinct parts: **hardware and software**.
- Choosing which functions to implement in **hardware** and which in **software** is a major engineering challenge that involves consideration of issues such as **cost, complexity, performance, and power consumption**.
- From the behavioral description, it is necessary to perform **hardware/ software partitioning**.

Transmeta's Crusoe Processor

- 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.

Transmeta's Crusoe Processor

- By virtualizing the x86 CPU with a hardware–software combine, the Transmeta engineers have drawn a line between hardware and software such that the **hardware part is relatively simple** and **high-speed**, but **much less power-hungry** **very long instruction word (VLIW) engine**.

Transmeta's Crusoe Processor

- Complex task of translating the x86 instructions into the instructions of the VLIW is performed by a piece of software known as the **code morphing software (CMS)**.

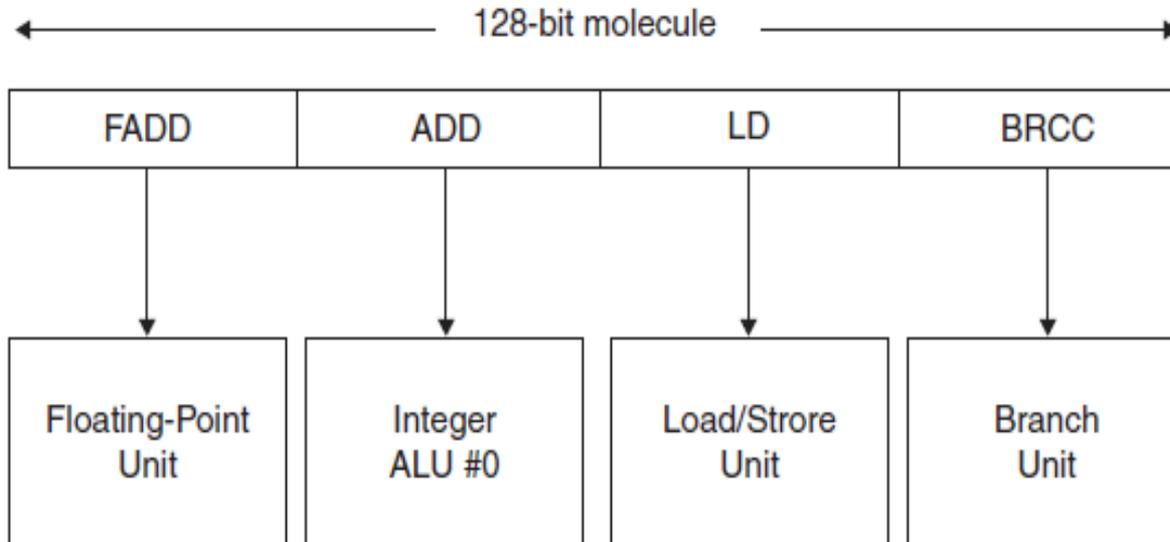
Transmeta's Crusoe Processor

- Hardware
- The Crusoe processor is a **very simple, high-performance VLIW processor** with **two integer units, a floating-point unit, a memory (load/store) unit, and a branch unit.**
- The long instruction word, called a **molecule**, can be 64 bits or 128 bits long, containing up to four **reduced instruction set computing (RISC)**-like instructions called **atoms**.

Transmeta's Crusoe Processor

- **Hardware**

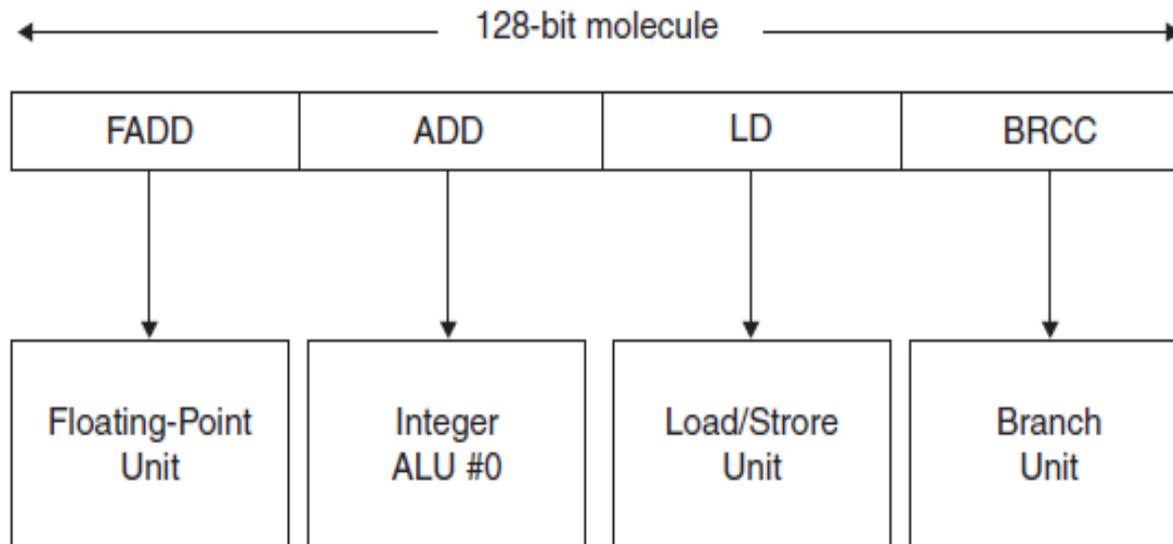
- All atoms within a molecule are executed in parallel and the format of the molecule directly determines how atoms get routed to the functional units.



Transmeta's Crusoe Processor

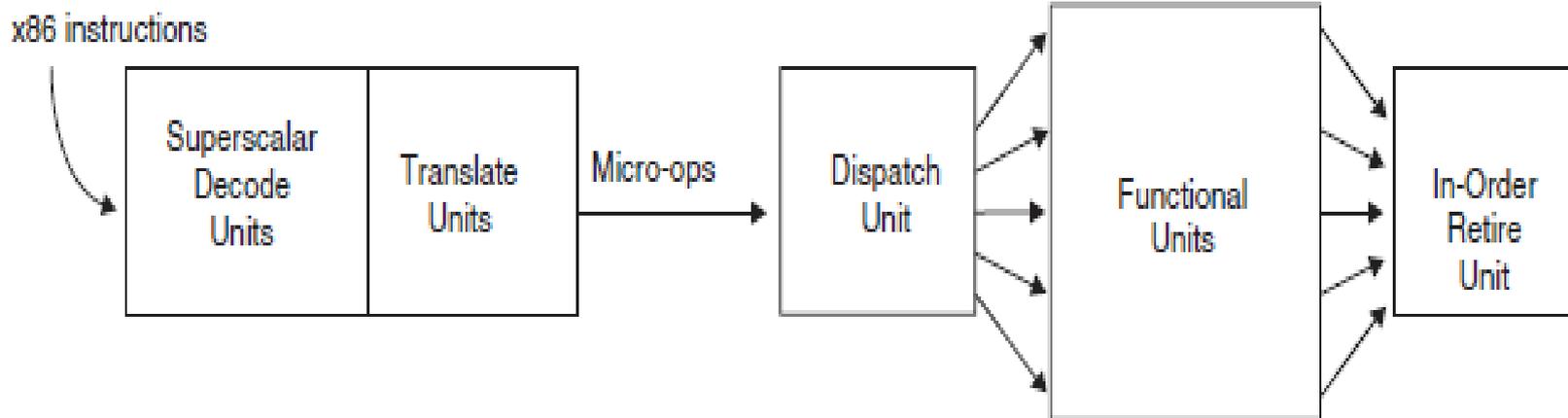
- **Hardware**

- A molecule can contain up to four atoms, which are executed in parallel. *FADD floating point addition, ADD addition, LD load, BRCC branch if carry cleared, ALU arithmetic logic unit*



Transmeta's Crusoe Processor

- Hardware
- Superscalar out-of-order architecture



Transmeta's Crusoe Processor

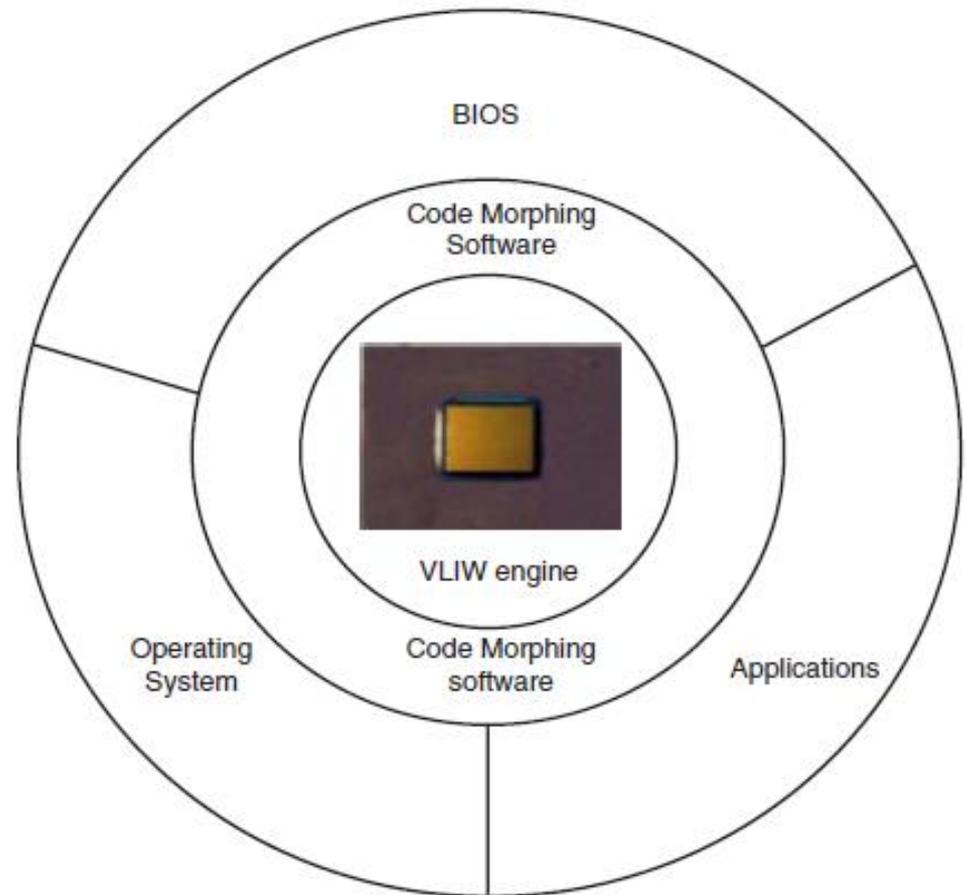
- Hardware
- Comparison of the die sizes

	Mobile P1	Mobile P11	Mobile P111	TM3120	TM5400
Process	0.25 m	0.25 m shrink	0.18 m	0.22 m	0.18 m
On-chip L1 cache (KB)	32	32	32	96	128
On-chip L2 cache (KB)	0	256	256	0	256
Die size (mm ²)	130	180	106	77	73

Transmeta's Crusoe Processor

- ***Software***

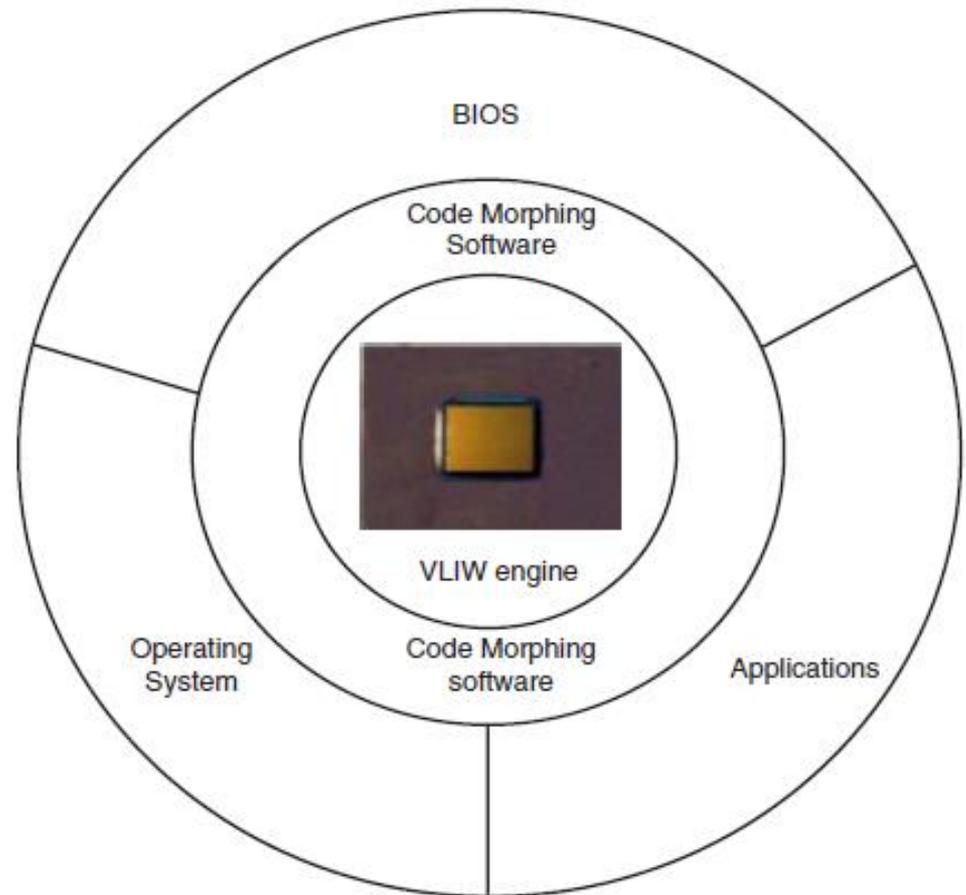
The success of the Crusoe processor lies in the use of an innovative software known as **CMS** that surrounds the VLIW hardware processor.



Transmeta's Crusoe Processor

- ***Software***

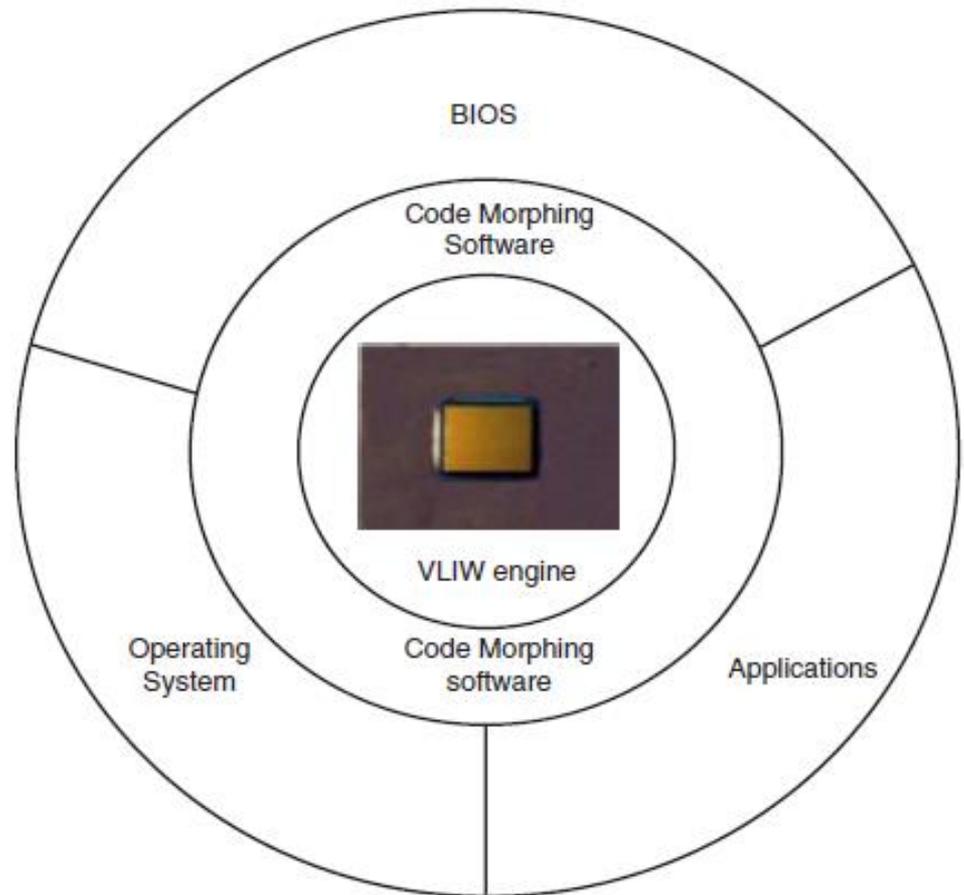
This software layer decouples the x86 instruction set architecture (ISA) from the underlying processor hardware, which is very different from that of the conventional x86 processor architectures.



Transmeta's Crusoe Processor

- ***Software***

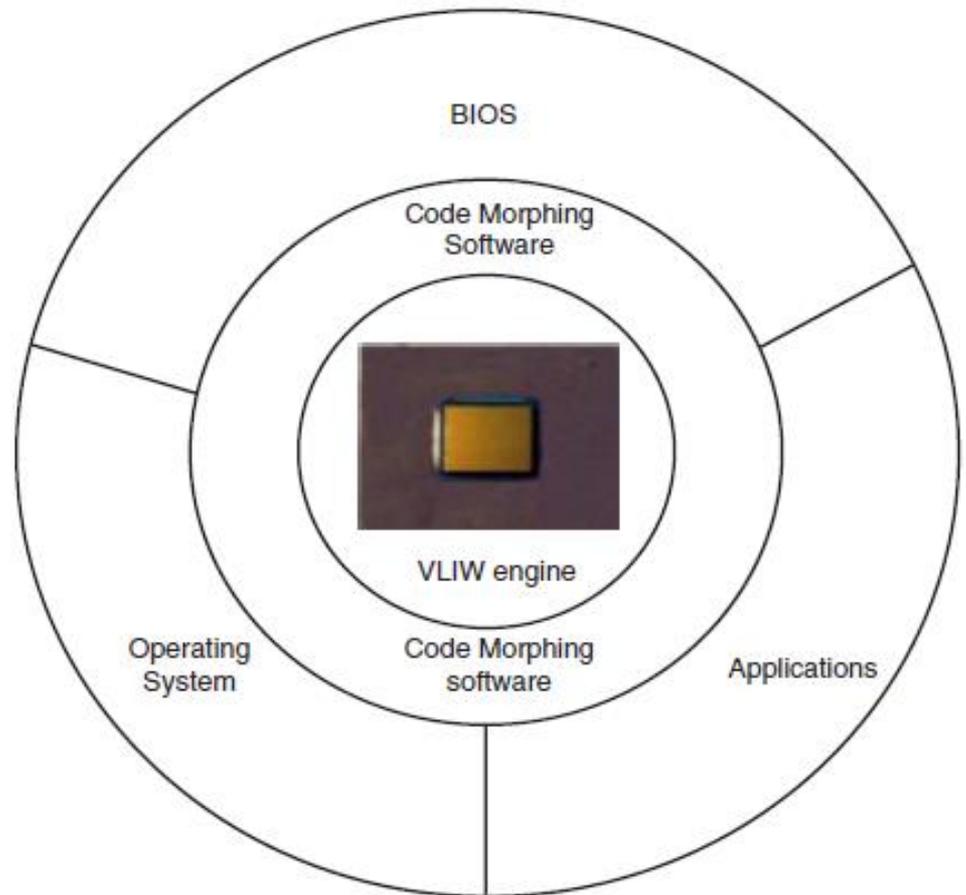
The CMS is essentially a dynamic translation system of a program that compiles the instruction of the ISA into the instruction of the VLIW processors.



Transmeta's Crusoe Processor

- ***Software***

The CMS is essentially a dynamic translation system of a program that compiles the instruction of the ISA into the instruction of the VLIW processors.

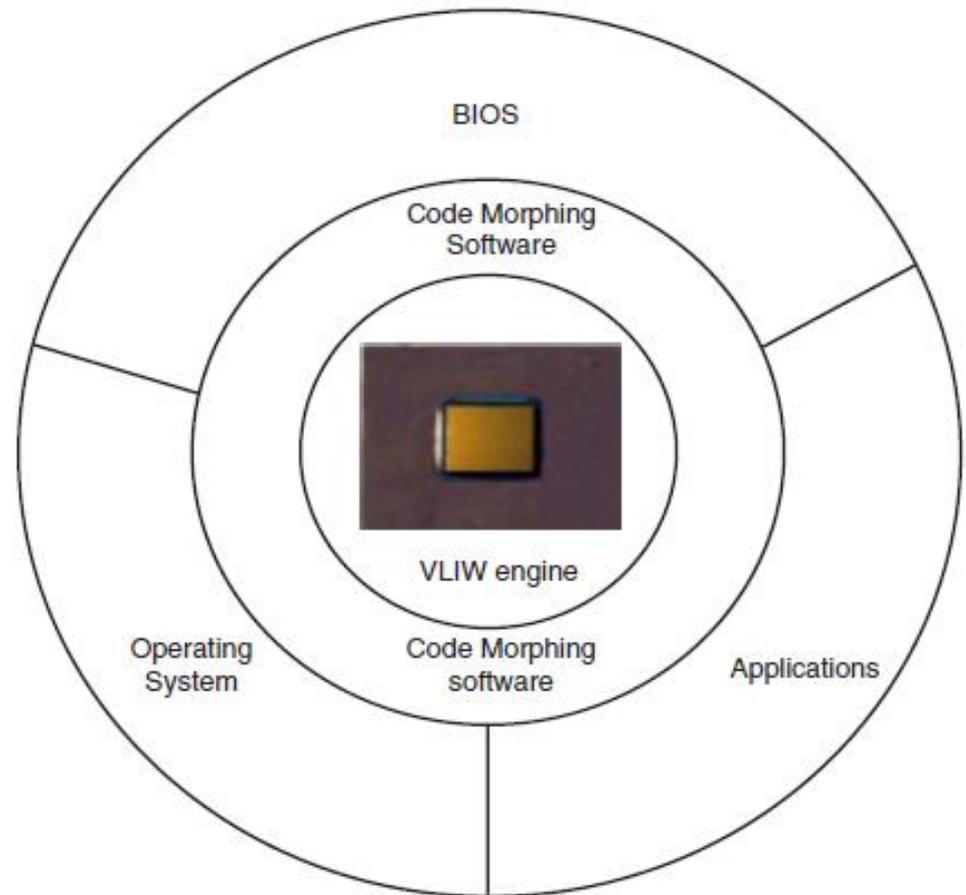


Transmeta's Crusoe Processor

- **Software**

- ❖ The CMS is the only program that is written directly for the VLIW processor as shown in Fig.

- ❖ All the other software, such as basic input/output system (BIOS), operating system (OS), and other application programs, are written using x86 ISA.



Transmeta's Crusoe Processor

- Decoding and Scheduling
- In a conventional superscalar processor such as **x86**, instructions are fetched from the memory and decoded into micro-operations, which are then recorded by an out-of-order dispatch hardware and fed to the functional units for parallel execution.
- In contrast to this, CMS translates an entire group of **x86** instructions at once and saves the resulting translation in a translation cache and in subsequent executions.

Transmeta's Crusoe Processor

- **Decoding and Scheduling**
- The system skips the translation step and directly executes the existing optimized translation.
- In a superscalar architecture, the out of- order dispatch unit has to translate and schedule instructions every time these are executed, and it must be done very quickly.
- In other words, power dissipation occurs each time when instructions are executed.

Transmeta's Crusoe Processor

- Decoding and Scheduling
- The cost of translation is amortized in the code morphing approach over many executions.
- This also allows much more sophisticated translation and scheduling algorithms.
- The generated code is better optimized, resulting in lesser number of instructions in the generated code.
- This not only speeds up execution but also reduces power dissipation.

Transmeta's Crusoe Processor

- **Caching**
- **A separate memory space is used to store the translation cache and the CMS.**
- The memory space is not accessible to x86 code, and the size of this memory can be set at boot time.
- One primary advantage of caching is to reuse the translated code by making use of the locality of reference property.

Transmeta's Crusoe Processor

- **Caching**
- In real-life applications, it is very common to execute a block of code many times over and over after it has been translated once.
- Because of the high repeat rates of many applications, code morphing has the opportunity to optimize execution and amortize the initial translation overhead.
- As a result, the approach of caching translations provides excellent opportunity of reuse in many real-life applications.

Transmeta's Crusoe Processor

- **Filtering**
- **It is of common knowledge that a small percentage (may be less than 10 %) of the code accounts for 95 % of the execution time.**
- **This opens up the possibility of applying different amounts of translation efforts to different portions of a program.**

Transmeta's Crusoe Processor

- **Filtering**
- The code morphing has a built-in feature of a wide choice of execution modes of x86 code, starting from interpretation, which has no translation overhead at all, but the execution is slower, to highly optimized code, which has a large overhead for code generation, but that runs the fastest once it is translated.
- A set of sophisticated heuristics helps to choose from several execution modes based on the dynamic feedback information gathered during the actual execution of the code.

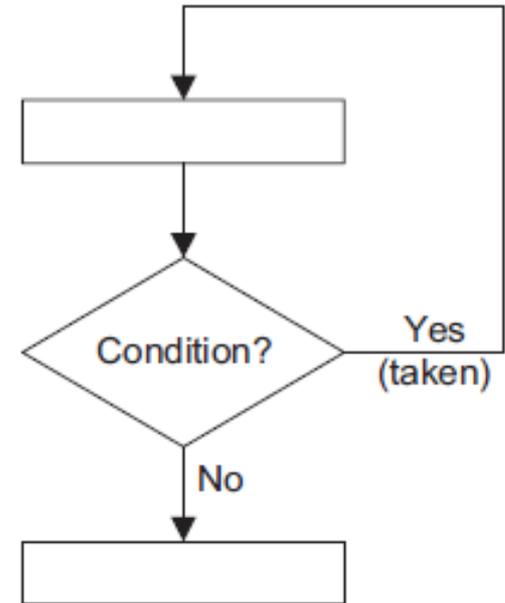
Transmeta's Crusoe Processor

- **Prediction and Path Selection**
- **The CMS can gather feedback information about**
- the x86 program with the help of an additional code present in the translator whose sole purpose is to collect information about block execution frequencies or branch history.
- Decision can be made about how much effort is required to optimize a code based on how often a piece of x86 code is executed

Transmeta's Crusoe Processor

- **Prediction and Path Selection**

- For example, whether a conditional branch instruction, as shown in Fig. 8.5, is balanced (50 % probability of taken) or biased in a particular direction, decision can be made about how much effort to put to optimize that code.
- In a conventional hardware-only x86 implementation, it would be extremely difficult to make similar kinds of decisions.



Flowchart of a program with a branch

Transmeta's Crusoe Processor

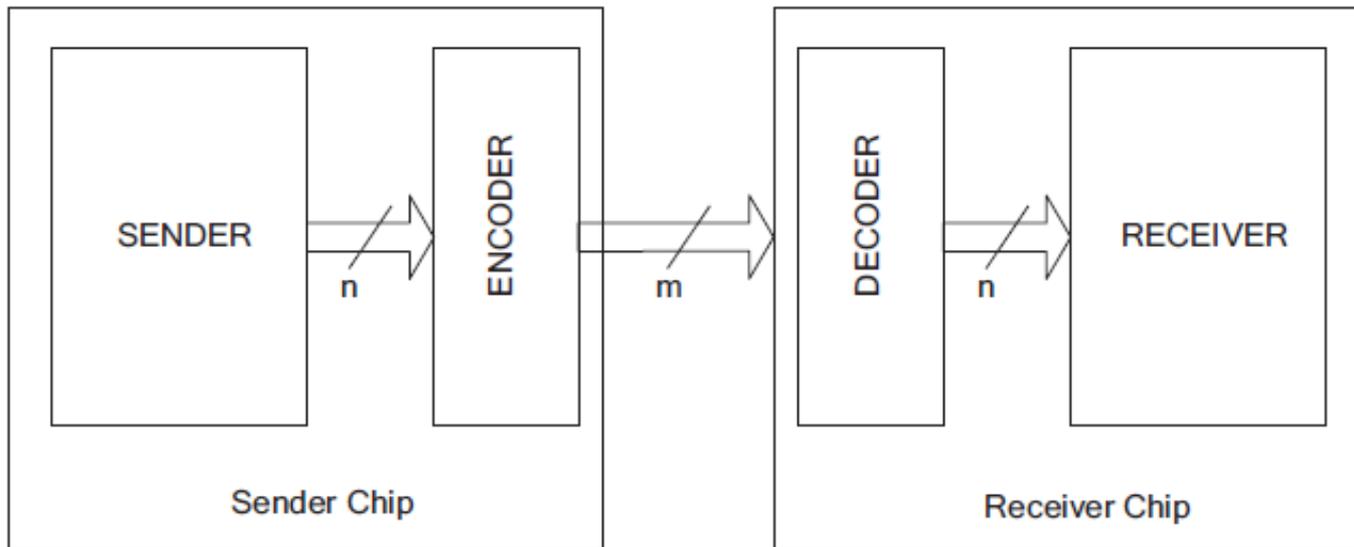
- The above VLIW instructions are executed **in-order by the hardware** and the molecules explicitly encode the instruction-level parallelism.
- As a consequence, the VLIW engine is **very simple.**

Bus Encoding

- 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 input/output (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.

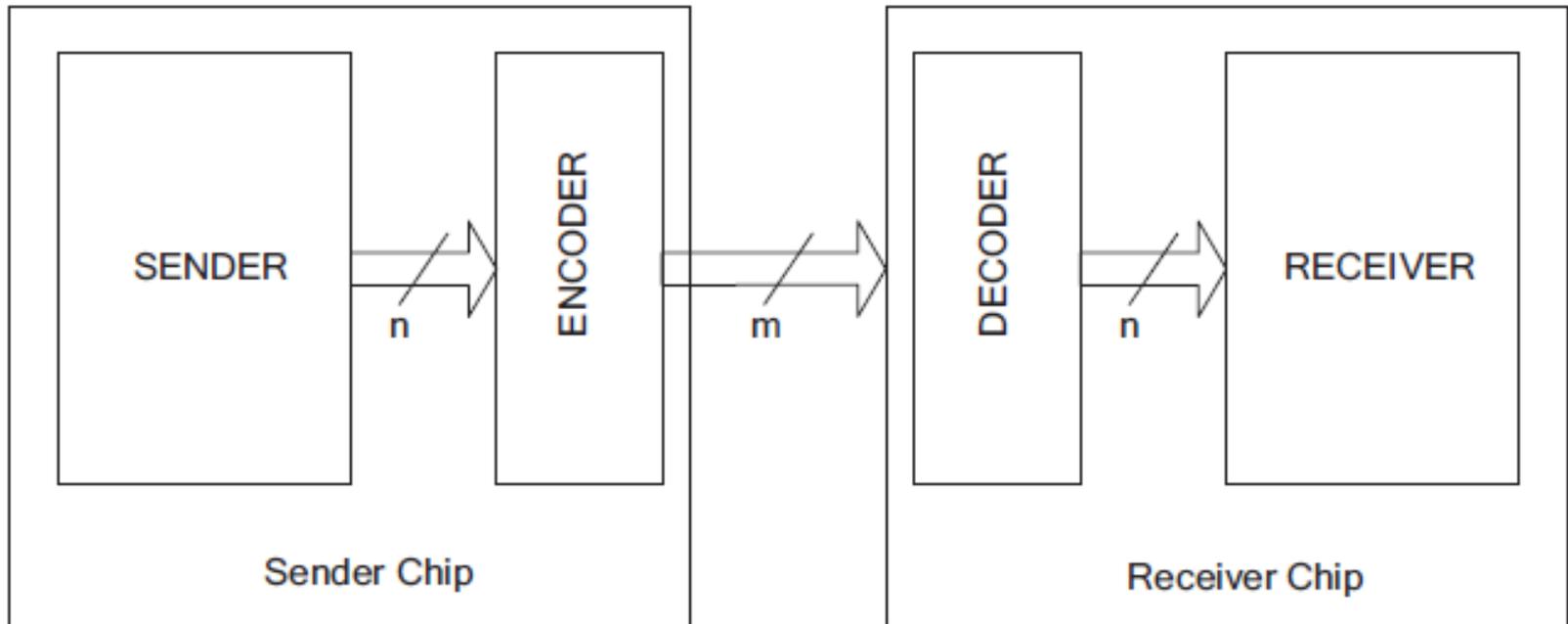
Bus Encoding

- One possible approach for **reducing the switching activity** is to suitably **encode** the data before sending over the I/O interface.



Bus Encoding

- A **decoder** is used to get back the original data at the receiving end



Bus Encoding

- **Encoding** can be used either to remove **undesired correlation among the data bits**, as it is done while doing **encryption**, or to **introduce controlled correlation**.
- Coding to **reduce the switching activity** falls under the second category.
- To reduce the switching activity, it is necessary to **increase the correlation** among the signal values transmitted over the I/O pins.

Bus Encoding

- Coding scheme can be broadly divided into two broad **categories—non-redundant and redundant.**
- In case of **non-redundant coding**, an *n-bit code* is translated into another *n-bit code* ($m = n$) and the 2^n code elements of *n-bit* are mapped among themselves.
- Although there is no additional lines for sending off chip data, the non-redundant coding is done based on the statistics of the sequence of words such that the switching activity is reduced.

Bus Encoding

- In case of **redundant coding technique**, additional lines are used for sending data, i.e., $m > n$.
- *Here, 2^n different n -bit words are mapped to a larger set of m -bit 2^m data words.*
- Here, there are two alternatives—the encoder/decoder may be **memory-less or may use memory elements** for the purpose of encoding and/or decoding.

Bus Encoding

- If an n -bit word has a unique m -bit code word, then *neither encoder nor decoder requires memory.*
- Another possibility is to use a static **one-to-many mapping** of 2^n unique words to 2^m code words.
- *The current state of the encoder selects a particular m -bit code word for transmission out of several alternatives for a particular n -bit data word.*
- In this scheme, an **encoder requires memory**, but the **decoder can be implemented without memory.**

Bus Encoding

- *Gray Coding*
- *One-Hot Coding*
- *Bus-Inversion Coding*
- *T0 Coding*

Bus Encoding

- ***Gray Coding***
- One popular example of the nonredundant encoding scheme is the Gray coding.
- Gray coding produces a code word sequence in which adjacent code words differ only by 1 bit, i.e., Hamming distance of 1 as shown in Table.

Bus Encoding

- ***Gray Coding***
- Gray coding produces a code word sequence in which adjacent code words differ only by 1 bit, i.e., Hamming distance of 1 as shown in 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

Bus Encoding

- ***Gray Coding***
- The number of transitions for binary representation is 30.
- On the other hand, the number of transitions for Gray code will always have 16.

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

Bus Encoding

- ***Gray Coding***
- As a consequence, the transition activity, and hence the power dissipation, reduces by about 50 %, and it is very useful when the data to be transmitted is sequential and highly correlated.

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

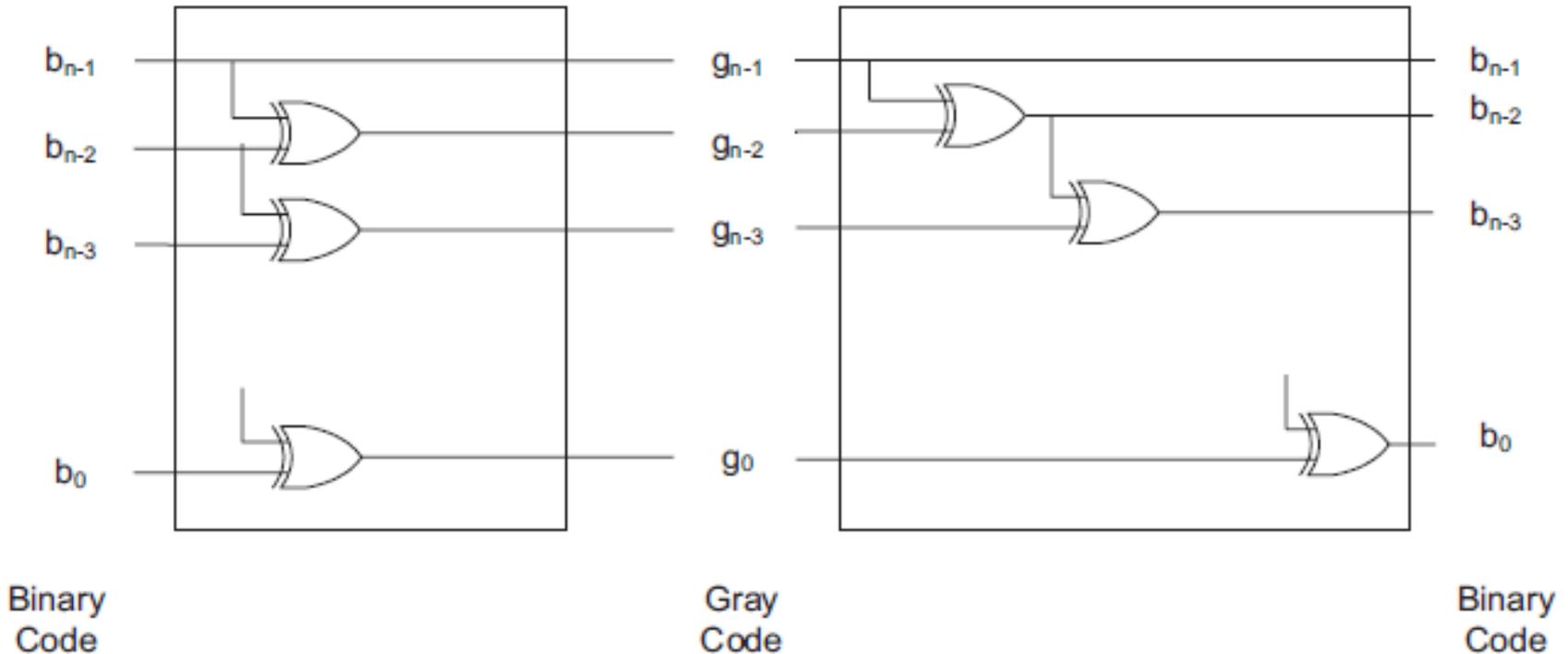
Bus Encoding

- Gray Coding**

Encoder and decoder for Gray code

Encoder

Decoder



Each encoder and decoder requires $(n - 1)$ two-input exclusive OR (EX-OR) gates.

Agenda



- **Bus Encoding Techniques:**
 - **Bus Encoding Basics**
 - **Gray Coding**
 - **One-hot Encoding**
 - **Bus-Invert Encoding**
 - **T0 Encoding**



NPTEL

Ajit Pal

IIT Kharagpur

NPTEL

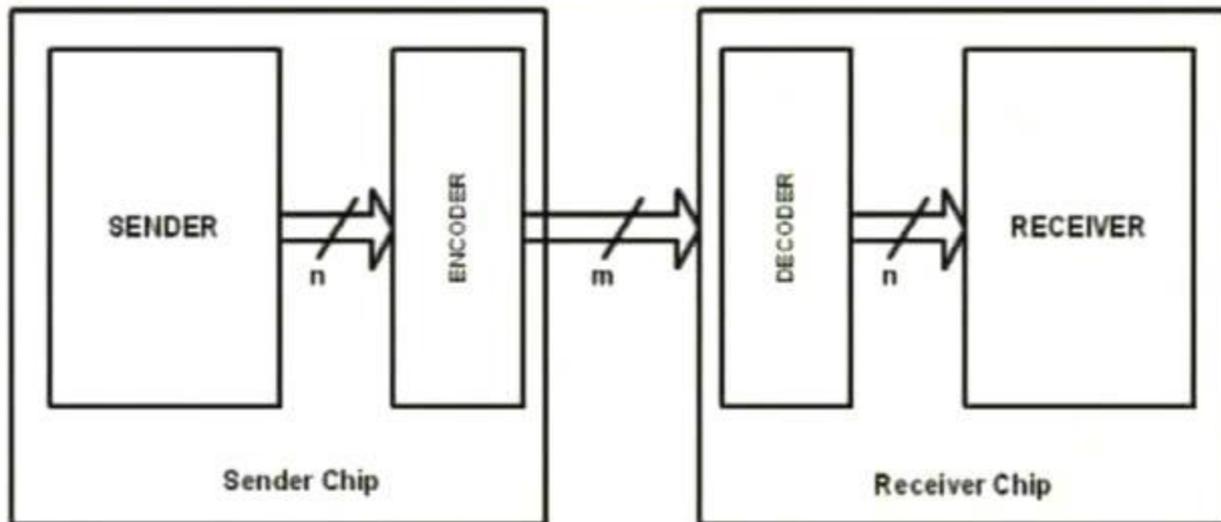
Bus Encoding



- **Goals of coding:**
 - Remove undesired correlation among information bits (Encryption)
 - Introduce controlled correlation (Spectrum shaping, timing recovery, error detection/correction)
- **Coding for reduced switching activity falls under the second category**
 - Introducing sample to sample correlation such that total number of bit transitions is reduced
- **Communicating data bits in an appropriately coded form can reduce the switching activity**



Bus Encoding



➤ Categories of encoding techniques:

- Redundant: $m > n$
- Non-redundant: $m = n$
- One-to-one: (memoryless)

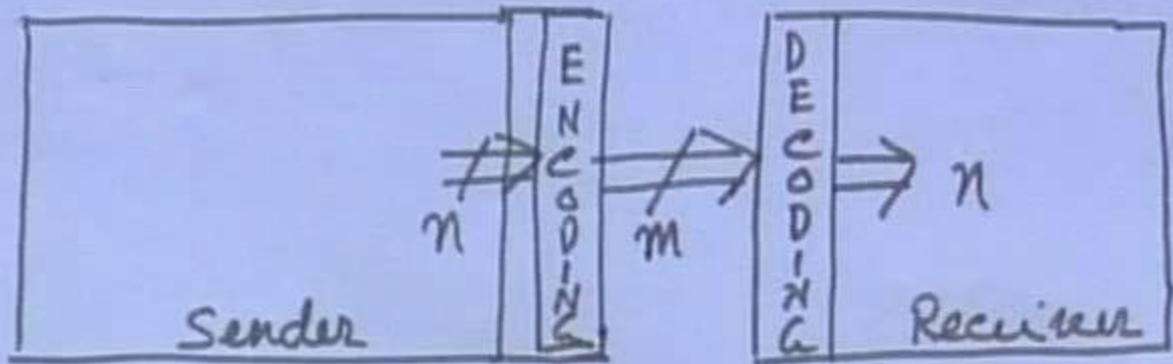


One-to-many (static): Encoder with memory, decoder without memory

One-to-many (dynamic): Encoder with memory, decoder with memory

BUS Encoding

© CET
I.I.T. KGP



- Redundant $m > n$
- Non redundant $m = n$
- one-to-one

OFF-CHIP capacitance

αC_L



Encoding Techniques

Non redundant

Redundant

$m=n$

one-to-one

one-to-many

Memory less.

Static

Dynamic

Encoder with M
Decoder Memoryless

Encoder
&
Decoder

with memory



NPTEL

Gray Coding Vs Binary Coding



- A gray code sequence is a set of numbers in which adjacent numbers have only one bit difference

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



Gray Coding Vs Binary Coding



- It is useful when the data is sequential and highly correlated, like Instruction addresses
- No of bit transitions is limited to one for sequential data
- For random data, the no of transitions for binary and gray code are approximately equal



NPTEL

Ajit Pal

IIT Kharagpur

NPTEL

Comparison of the temporal activity



Benchmark Program	Instruction Address		Data Address	
	Binary Coded	Gray Coded	Binary Coded	Gray Coded
Fastqueens	2.46	1.03	0.85	0.91
Qsort	2.64	1.33	1.32	1.25
Reducer	2.57	1.71	1.47	1.40
Circuit	2.33	1.47	1.33	1.18
Semigroup	2.68	1.99	1.38	1.34
Nand	2.42	1.57	1.25	1.16
Boyer	2.76	2.09	1.76	1.72
Browse	2.51	1.64	1.32	1.40
Chat	2.43	1.54	1.32	1.20

Bit transitions per instruction executed

Ref: Chandrakasan & Broderon, Low Power Digital CMOS Design, KAP



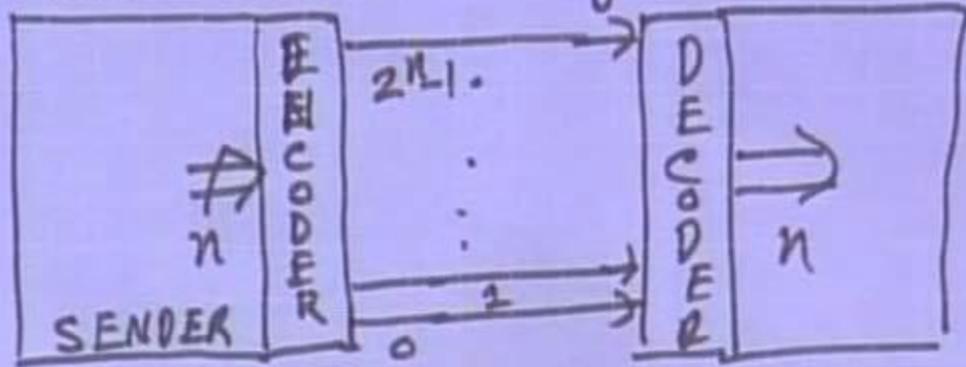
One Hot Coding



- Two chips are connected using $m=2^n$ wires, and a n -bit data word is encoded by placing '1' on the i^{th} wire, where $0 \leq i \leq 2^n - 1$ is the binary value corresponding to the bit pattern and '0' on the remaining $m-1$ wires.
- Guarantees precisely one $0 \rightarrow 1$ and one $1 \rightarrow 0$ bit transition when a different data word is sent
- Number of transitions is always two
- Number of wires required increases exponentially with the word size of the data
- For $n=8$, reduction in switching activity is 75%, but requires 256 lines



one-hot-encoding

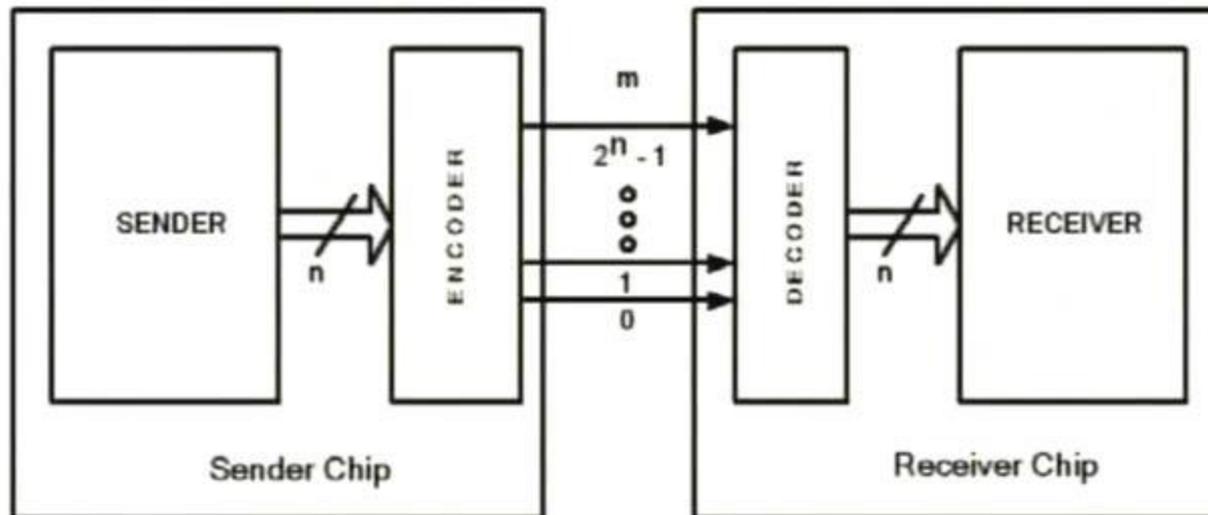


Binary code
000 0
000 1
00 1 0
⋮
⋮
⋮
1 1 1

2^n
one-hot code
0000 0000 0000 0001
0000 0000 0000 0010
⋮
1000 0000 0000 0000



One Hot Coding



- One-hot coding is a redundant ($m > n$),
- one-to-one (memoryless) coding technique



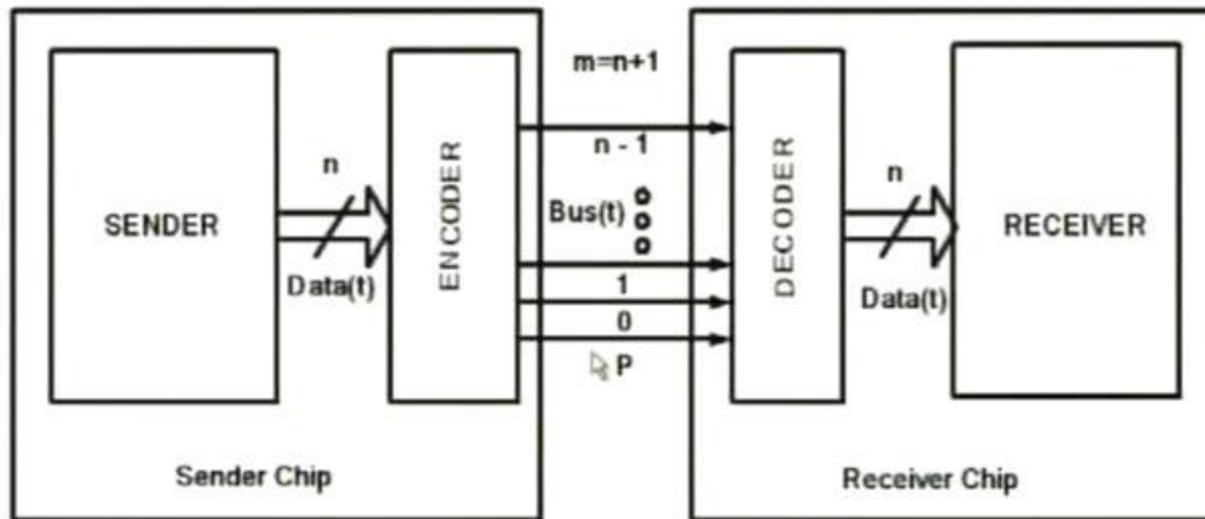
Bus Inversion Coding



- It is a redundant coding scheme where $m=n+1$
- If the i^{th} data word is S_i , then either S_i or $\sim S_i$ is transmitted depending on which would result in fewer no of bit transitions
- An extra bit **P** encodes the polarity of the data word

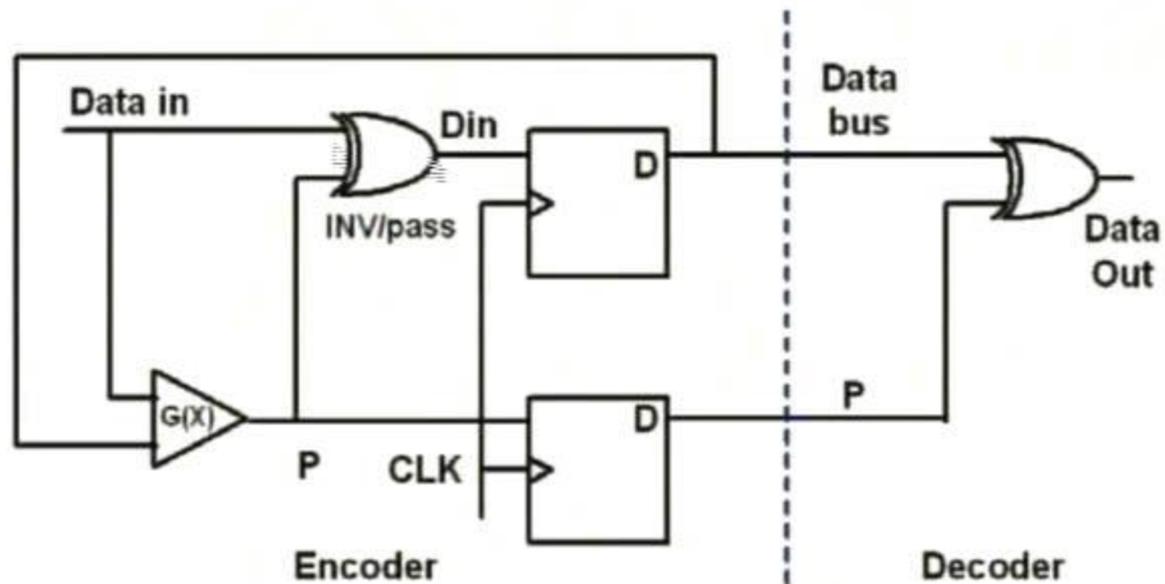


Bus Inversion Coding

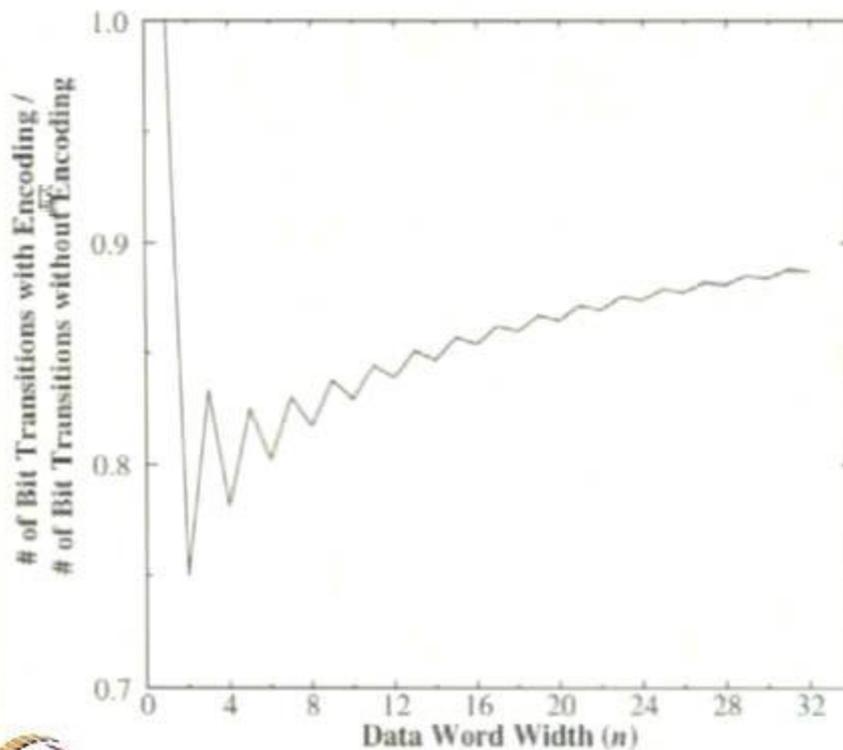


Bus Inversion coding is a redundant ($m=n+1$),
Decoder is memoryless, but encoder requires memory

Bus Inversion Coding



Bus Inversion Coding



➤ The coding technique works better for smaller values of n

➤ For $n=2$, switching activity reduction is 25%

➤ For $n=32$, switching activity reduction is 11%



Predicted reduction in Switching Activity

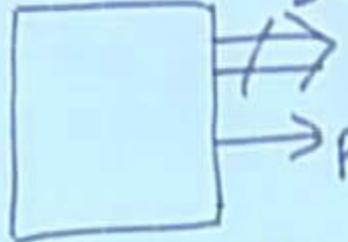
NPTEL

Ajit Pal

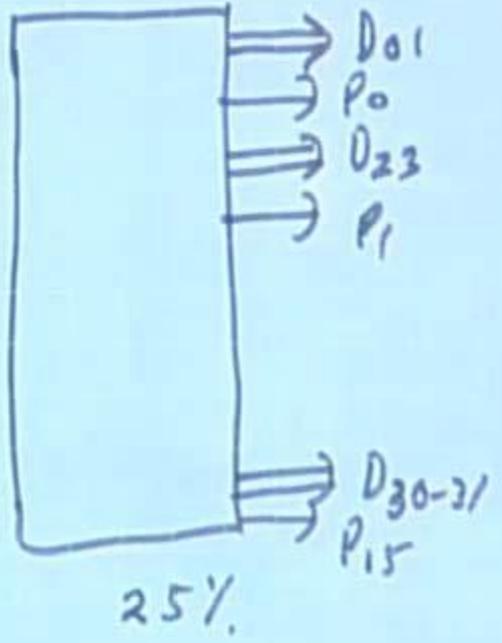
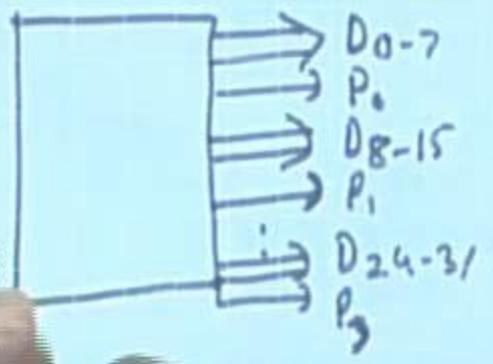
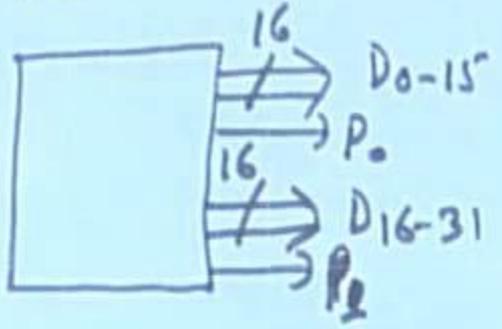
IIT Kharagpur

NPTEL

Case-I 32



11% 11%



T0 Encoding



- The Gray coding provides an asymptotic best performance of a single transitions for each address generated when infinite streams of consecutive addresses are considered.
- However, the code is optimum only in the class of irredundant codes, i.e. codes that employ exactly n -bit patterns to encode a maximum of 2^n words.
- By adding some redundancy to the code, better performance can be achieved by adapting the T0 encoding scheme, which requires a redundant line INC.
- The T0 code provides, zero transition property for infinite streams of consecutive addresses.

▪ **Encoding** $(B(t), INC(t)) = \begin{cases} B(t-1), 1, & \text{if } t > 0, b(t) = b(t-1) + S \\ b(t), 0, & \text{otherwise} \end{cases}$



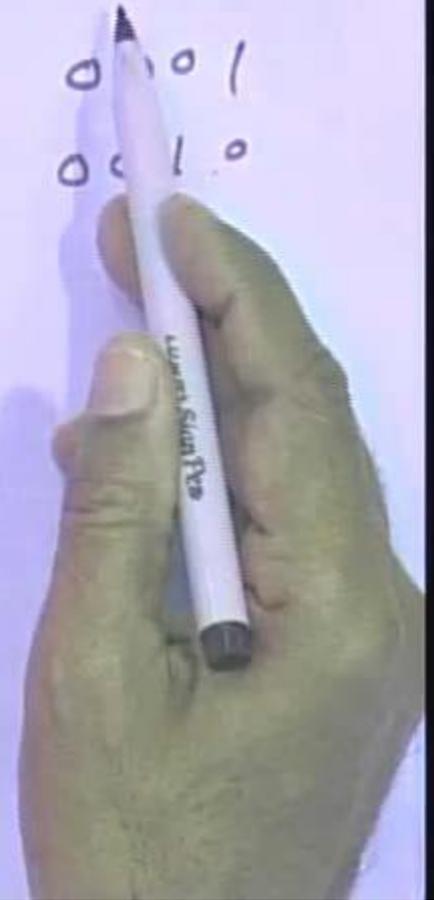
Decoding

$$b(t) = \begin{cases} b(t-1) + S & \text{if } INC = 1 \text{ and } t > 0 \\ B(t) & \text{if } INC = 0 \end{cases}$$

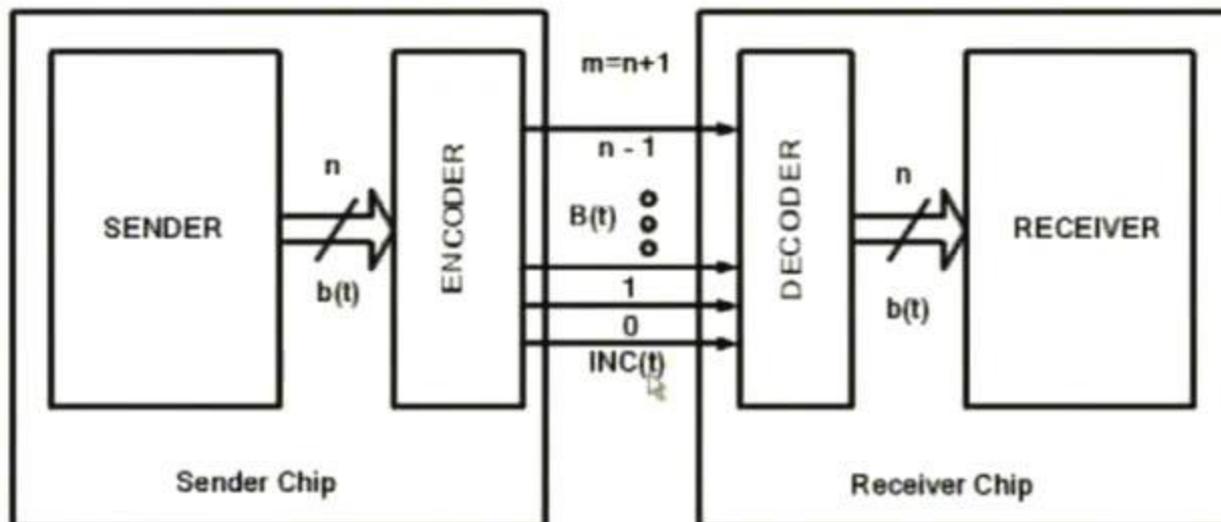
✓0000 . . .
0000
0000

0000 ✓
0001
0010

32-bit
2
—



T0 Encoding



- T0 coding is a redundant ($m=n+1$),
- Both decoder and encoder require memory



Conclusion



- **Bus Encoding Techniques:**
 - **Gray Coding:** Nonredundant, one-to-one, memoryless, suitable for address bus
 - **One-hot Encoding:** Redundant, one-to-one, memoryless, suitable for small bus size
 - **Bus-Invert Encoding:** Redundant, decoder memoryless, encoder with memory, suitable for both address and data bus
 - **T0 Encoding:** Redundant, both encoder and decoder requires memory, suitable for address bus
- **Have the potential to use in SoC and Multi-core Architecture**



NPTEL

Ajit Pal

IIT Kharagpur

NPTEL

Unit-5

Leakage Power Minimization

Agenda

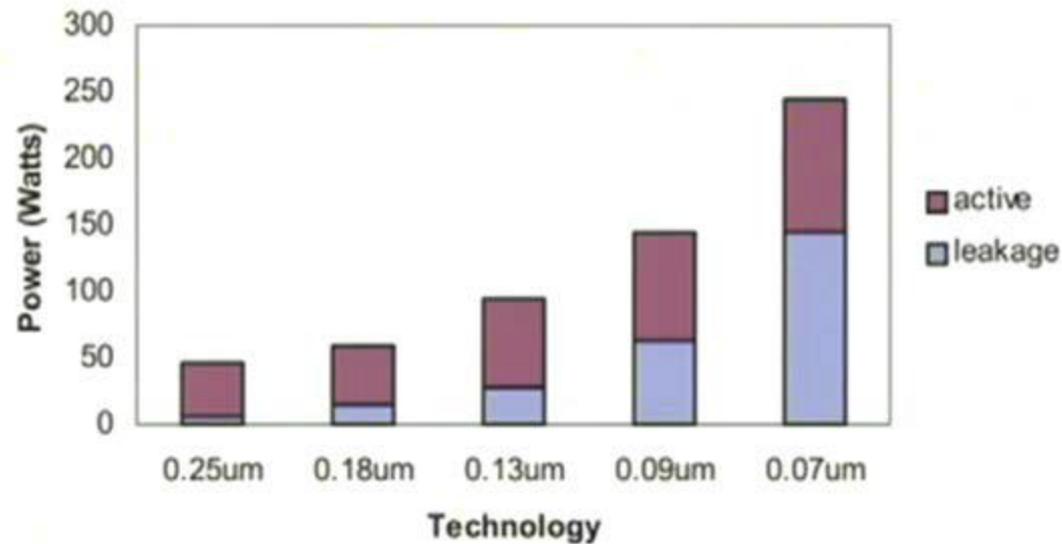


- ❑ Why Threshold Voltage Scaling?
- ❑ Fabrication of multiple threshold voltages
- ❑ Leakage power reduction techniques
 - Standby leakage reduction
 - Transistor stacking
 - Variable-threshold-voltage CMOS (VTCMOS)
 - Multi Threshold CMOS (MTCMOS)
 - Power gating
 - Combining power gating with DVFS
 - Run-time leakage reduction
 - Multilevel Vdd scaling
 - Dual-Vth assignment approach
 - Dynamic Vth scaling (Vt Hopping)



NPTEL

Why Leakage Power is an Issue?

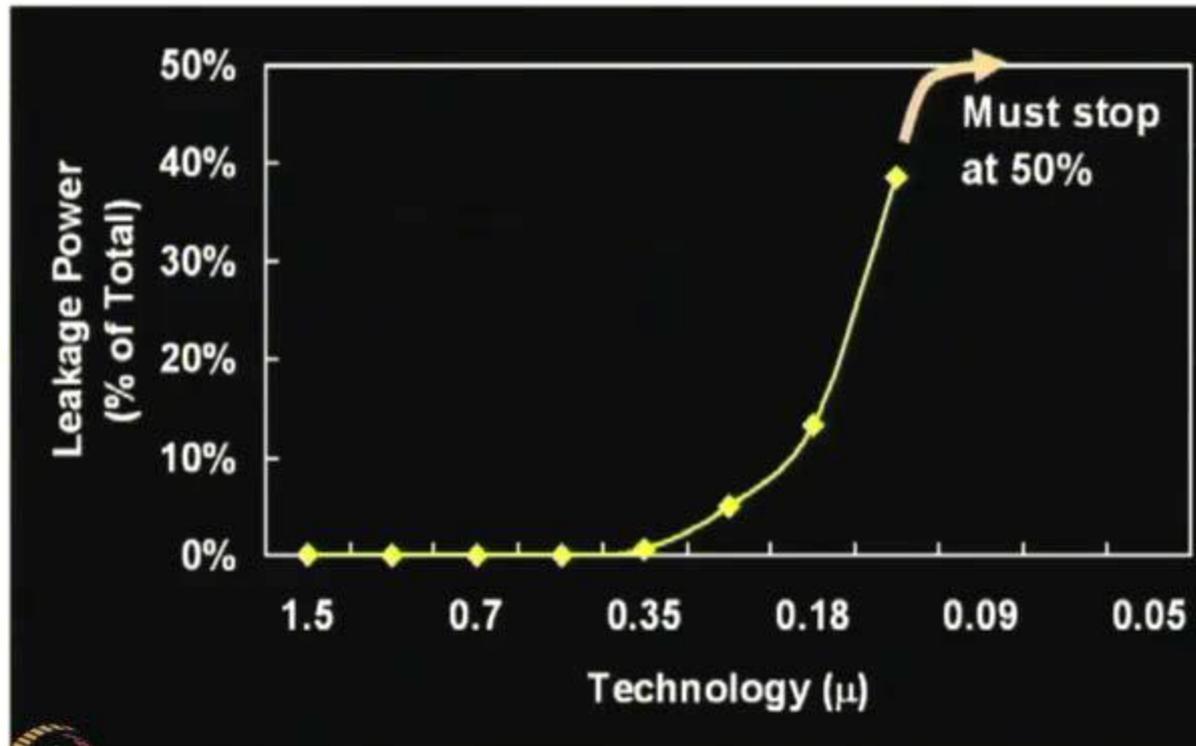


Source: Microprocessor power consumption, Intel

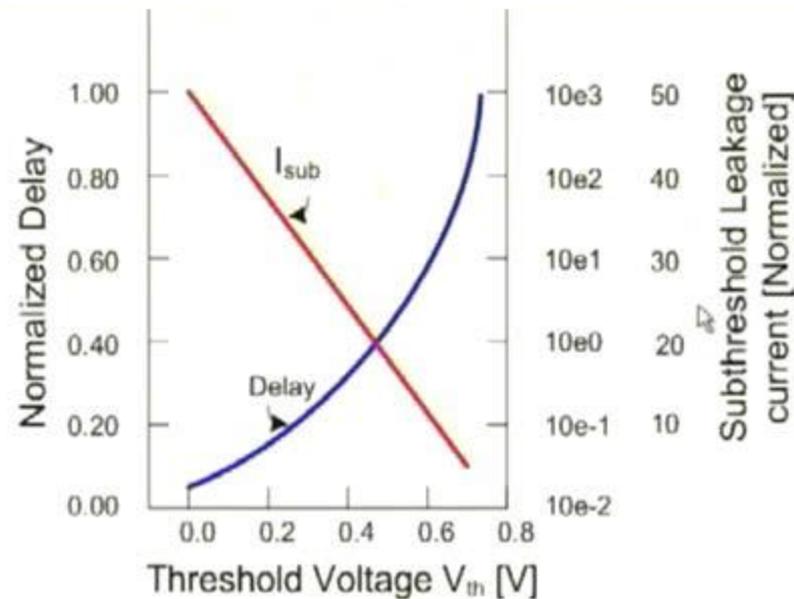
- Leakage power is becoming a large component of total power dissipation
- Reduction of runtime leakage power is important



Leakage Power Limits Vt Scaling



The Key Parameter: V_t



- As supply voltage is scaled down to reduce power dissipation, the threshold voltage is also scaled down to maintain performance



Threshold Voltage (V_{th}) Scaling



$$V_T \downarrow = \text{Delay} \downarrow + I_{\text{leakage}} \uparrow$$

Low $-V_T$: Provides high performance

$$V_T \uparrow = \text{Delay} \uparrow + I_{\text{leakage}} \downarrow$$

High $-V_T$: Reduces subthreshold leakage

$$0.2V_{DD} \leq V_T \leq 0.5V_{DD}$$

- Scale down the threshold voltage for low voltage low power circuits to increase performance
- Scale up threshold voltage without affecting performance



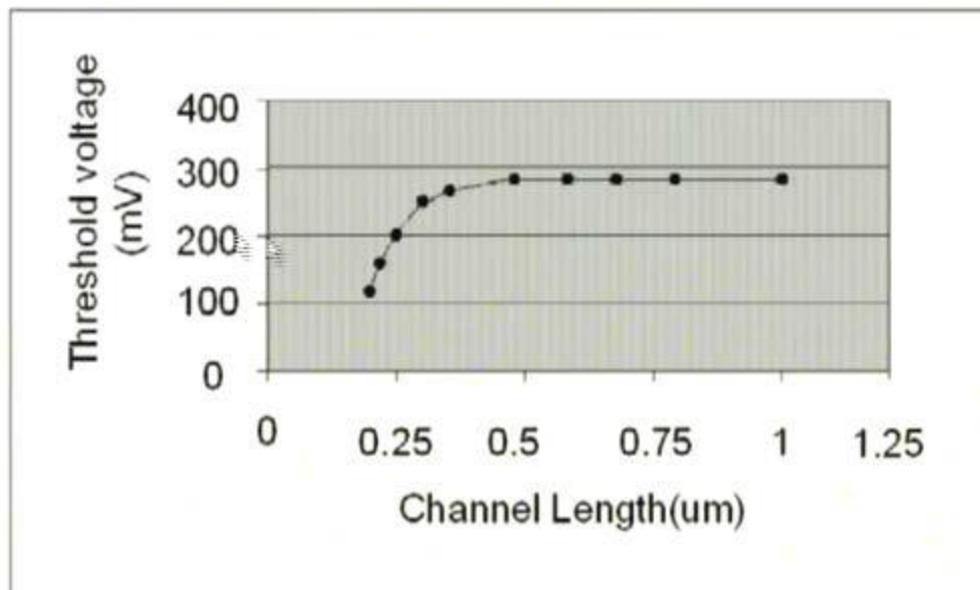
Fabrication of Multiple V_t



- Multiple channel doping
- Multiple Oxide thickness
- Multiple channel length
- Multiple body bias



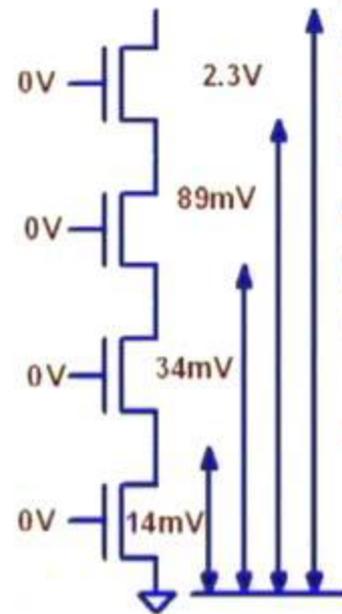
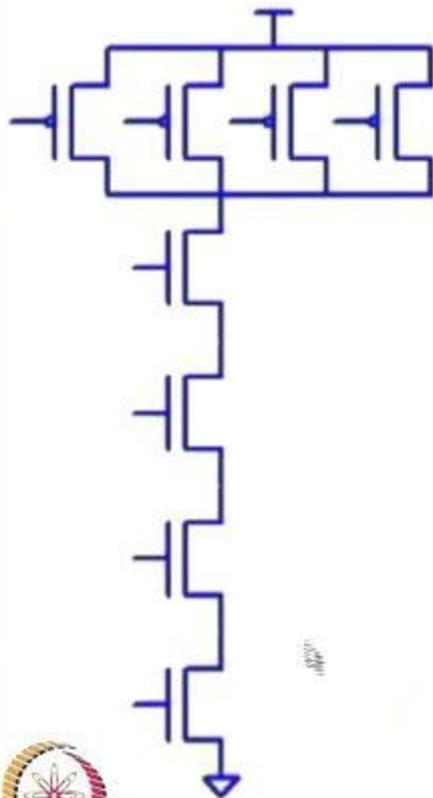
Multiple channel length



➤ The threshold voltage decreases as the channel length is reduced, which is known as V_{th} roll-off



Transistor Stacking



- When more than one transistor is in series in a CMOS circuit, the leakage current has strong dependence on the number of turned off transistor
- This is known as *stack effect*



Power Gating



➤ Who Controls?

- By control software to schedule power gating
- Part of the device driver or OS idle tasks
- Initiated by hardware timers
- Power Management Unit (PMU)

➤ Tradeoff:

- Amount of leakage power saving
- The entry and exit time
- Energy dissipated in entering and leaving such leakage saving mode
- The activity profile

➤ The SLEEP event initiates entry

➤ The WAKE event initiates return to active mode



Impact of Power Gating



➤ Multiprocessor Systems:

- It is assumed that a processor is powered down only when it has completed a task and waiting for another task to be assigned
- Power gating of individual CPUs gives very good leakage reduction
- Cache is empty when CPU is waken clean and reset-ready
- The number of cores to be power-gated and active can be optimized for varying workload conditions



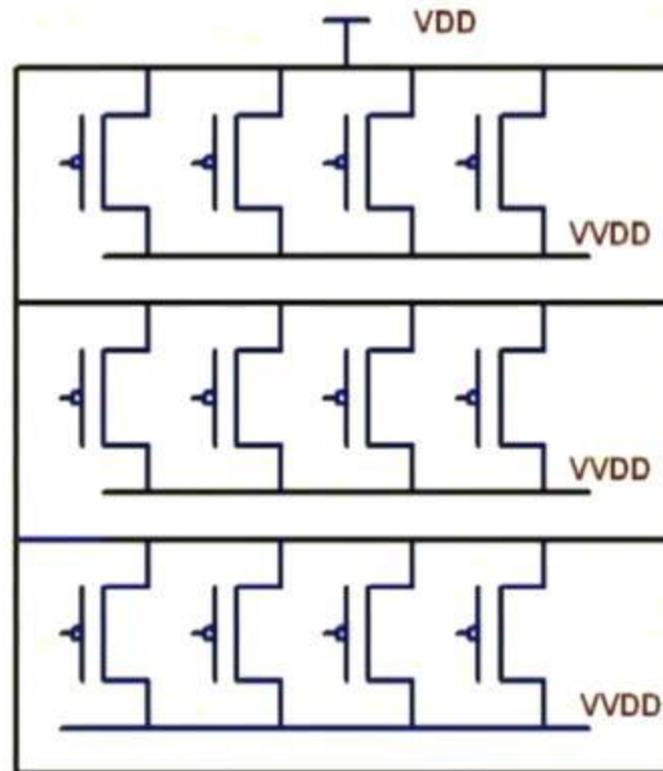
NPTEL

Grid Style



➤ The switches are distributed throughout the power gated region

- Hybrid style:
- Grid style at the top level
 - Ring style to certain power gated macros

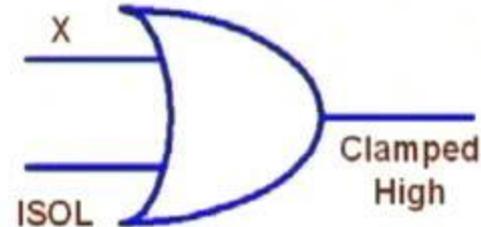
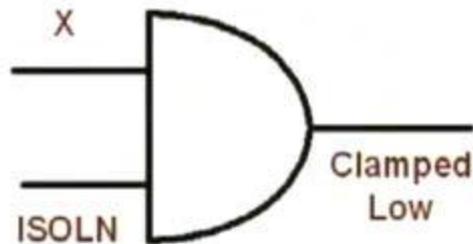


NPTEL

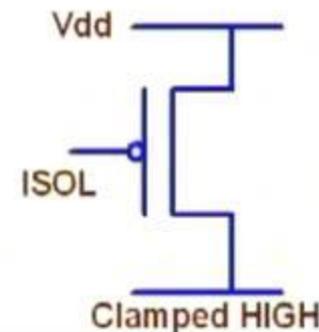
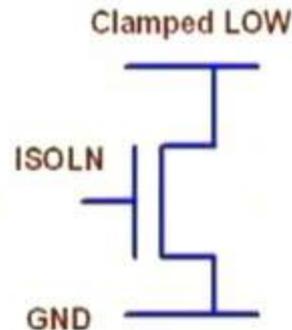
Clamping Circuits



- It is necessary to clamp the outputs to inactive states
- An AND-gate function can be used to clamp the output to '0'
- An OR-gate function can be used to clamp the output to '1'



- An alternative approach is to use pull-up or pull-down transistors to avoid full gate delay



Software based Approach

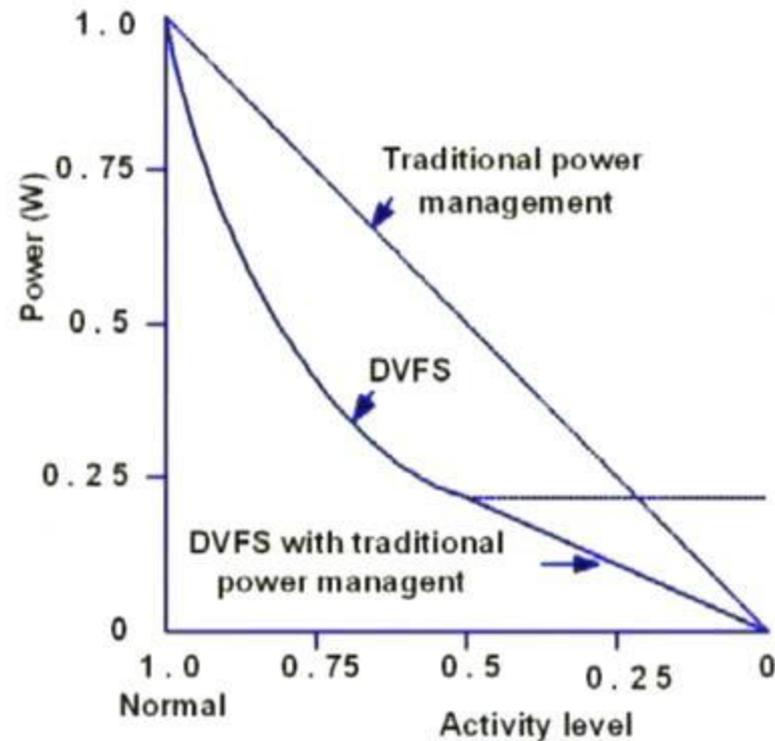


- In the software approach, the always-ON CPU reads the registers of the power-gated blocks and stores in the processor's memory
- During power-up sequence, the CPU writes back the registers from the memory
- Bus traffic slows down the power down and power up sequence
- Bus conflicts may make powering down unviable
- Software must be written and integrated into the system's software for handling power down and power up

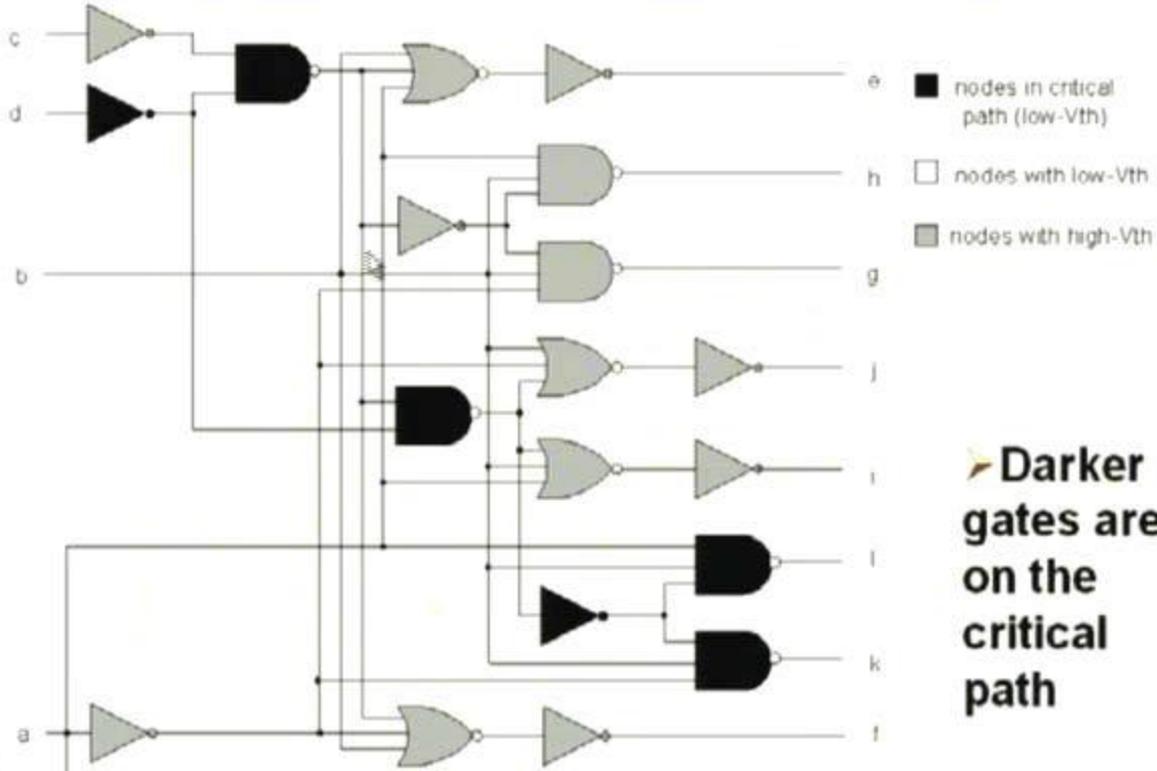
Combining DVFS with Power Management



- The DVFS ranges hit the limits of physics and the curve flattens out, when the minimum frequency-voltage point is reached
- At this point it is more efficient to switch over to traditional power management

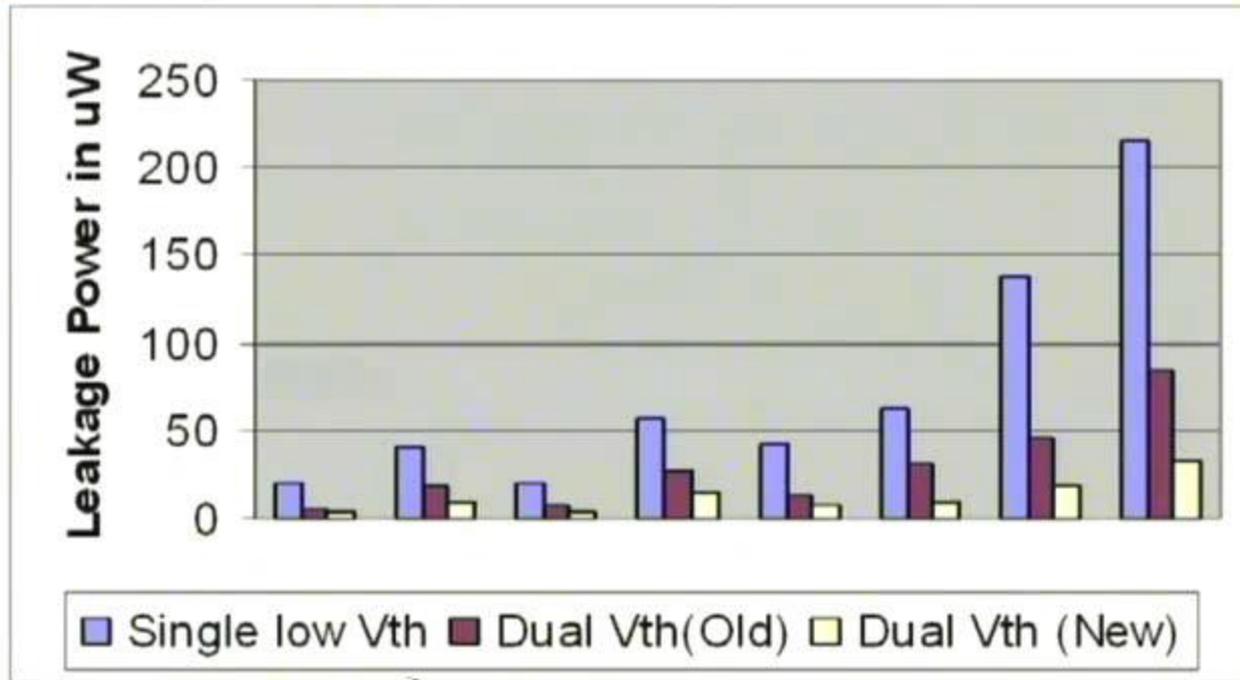


Dual Threshold CMOS Technology



HighVt = 0.25 assigned to all gates in the off-critical path

Comparison

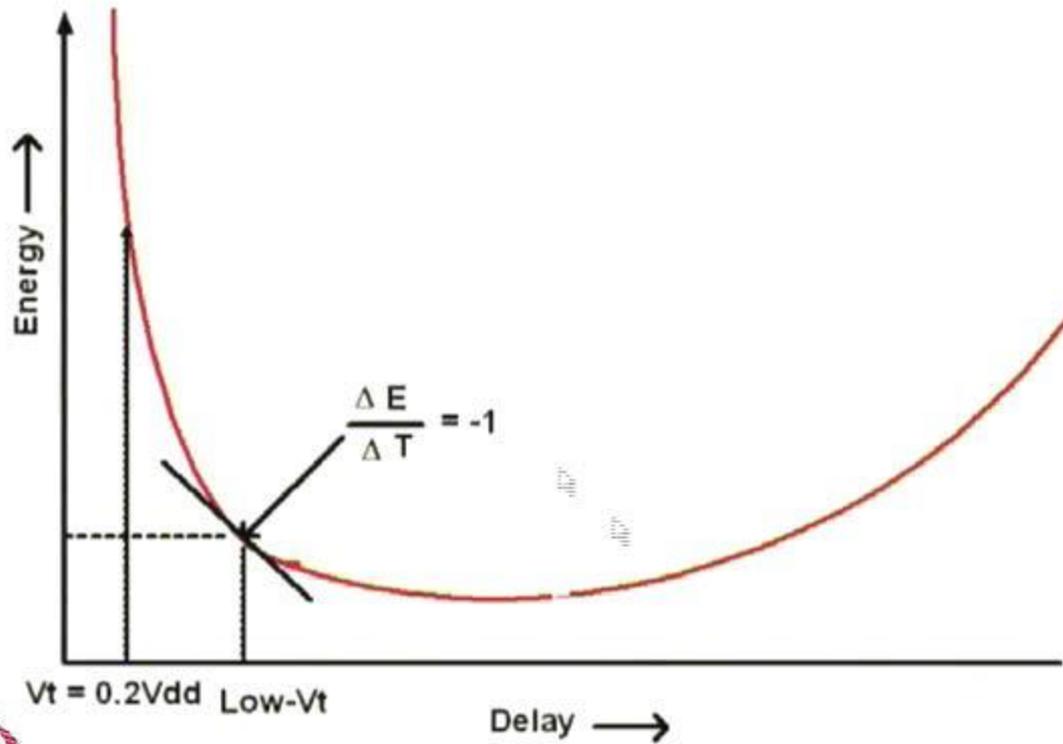


More reduction in leakage power

(Average 25% more reduction in leakage power)



Energy-Constrained Dual- V_T Assignment



Dynamic Vth scaling



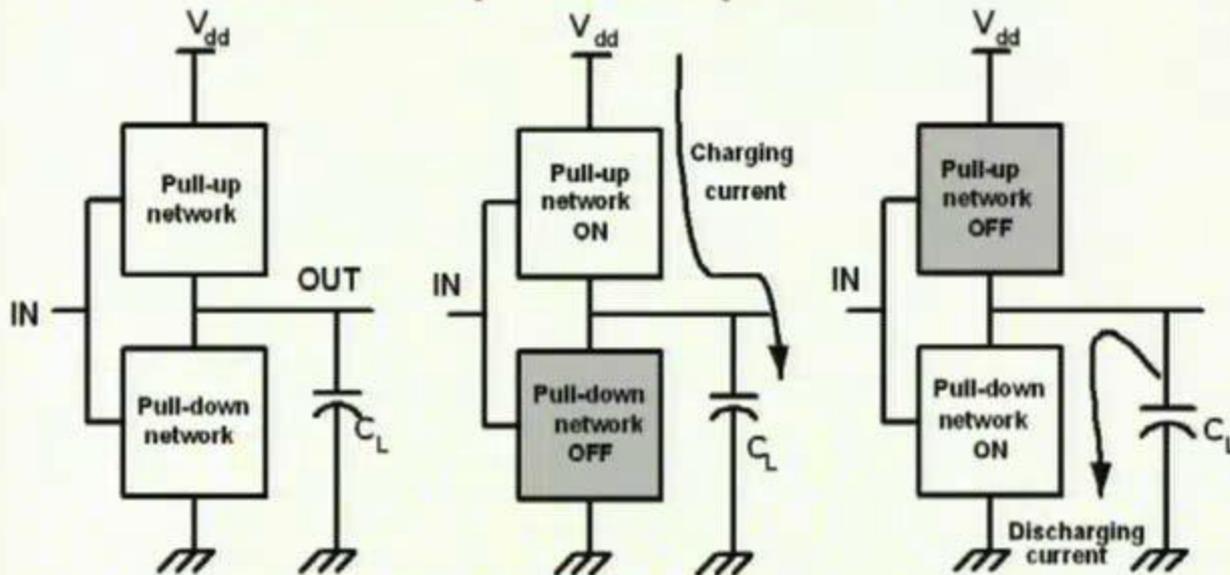
- Just like dynamic the Vdd scaling scheme, a **dynamic Vth scheme (DVTS)** can be used to reduce runtime leakage power in **sub-100-nm** generations, where leakage power is significant portion of the total power at runtime
- When the workload is less than the maximum, the processor is operated at lower clock frequency. Instead of reducing the supply voltage, the DVTS hardware raises the threshold voltage using **reverse body biasing** to reduce runtime leakage power
- Just enough throughput is delivered for the current workload by dynamically adjusting the Vth in an optimal manner to maximize leakage power reduction



Switching Power Dissipations



- Due to the charging and discharging of load and parasitic capacitors



$$P_d = \alpha_0 C_L V_{dd}^2 f + \sum_{i=1}^k \alpha_i C_i V_i V_{dd} f,$$



Adiabatic charging



The average current from 0 to t: $I(t) = \frac{C.V_c(t)}{t}$

The energy dissipation in R from 0 to t = T is given by:

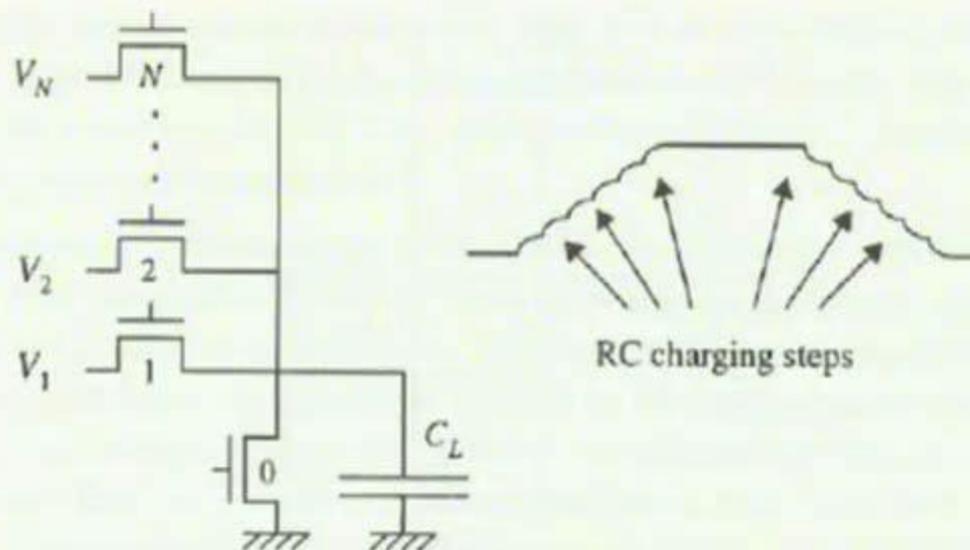
$$E_{diss} = R \int_0^T I^2 dt = RI^2(T)T = \frac{RC}{T} CV_c^2(T)$$

Observations:

- For $T > 2RC$, the dissipated energy is smaller than the conventional case
- The dissipation can be made arbitrarily small by further extending the charging time T
- The dissipated energy is proportional to R; a smaller R results in a lower dissipation unlike conventional case



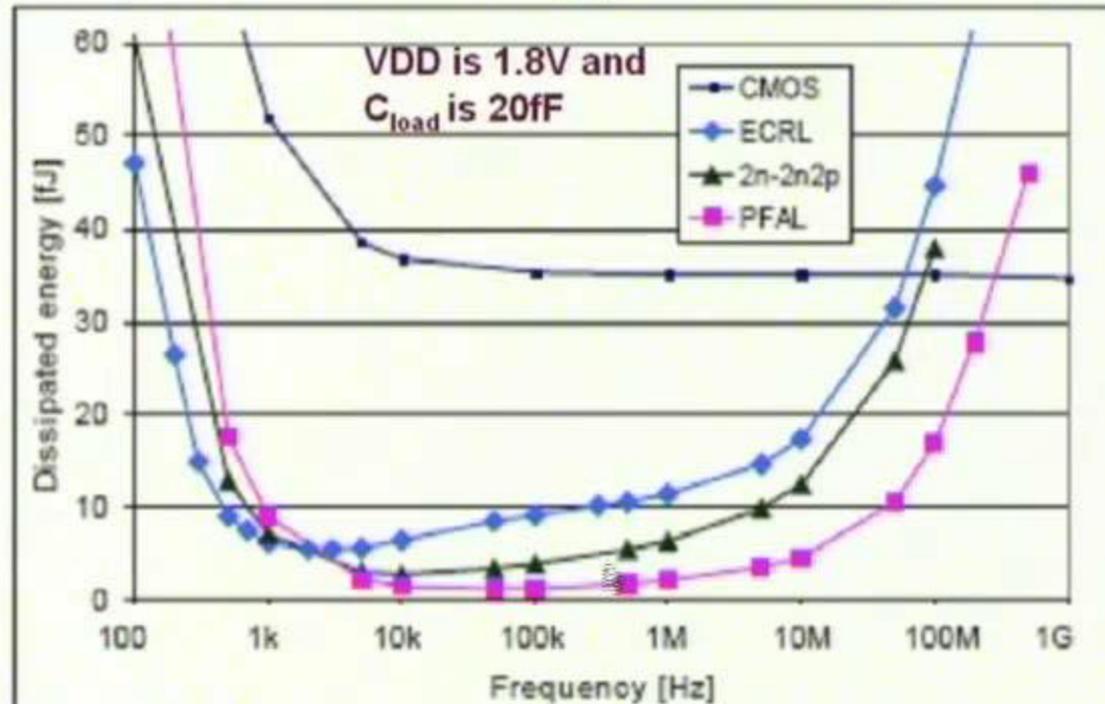
Stepwise Charging Circuits



- The total energy dissipation E_{stepwise}
- $= N \cdot E_{\text{step}} = N \cdot (C_L V^2 / 2N^2) = C_L V^2 / 2N$
- Multiple supply voltages are required
- Overhead of routing many supply lines



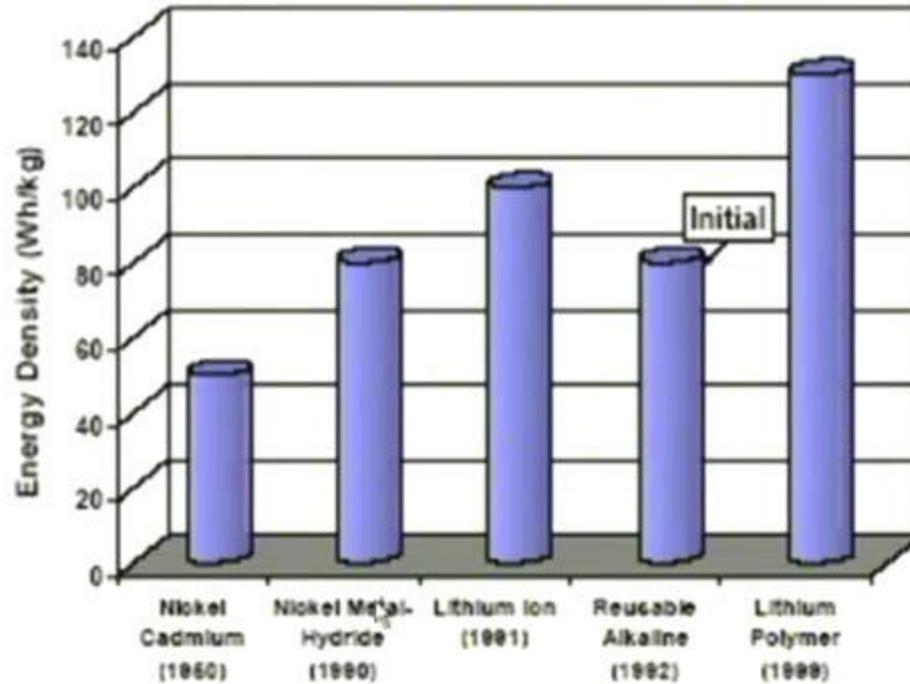
Comparison of Energy Consumption



➤ Energy consumption per switching operation versus frequency for a CMOS inverter, an ECRL inverter, a PFAL inverter and a 2N-2N2P inverter in case of nominal threshold voltage.

Battery-Driven System Design

Energy Density of Different Batteries



Battery-Driven System Design

Impact of Discharge characteristics on Battery Capacity



➤ **Rate capacity effects:** 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.

➤ **Recovery effects:** Depends of the concentration of positively charged ions near the cathode (rate of diffusion is affected). 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.



Battery-Driven System Design

Battery Modeling



- **Stochastic models:** Battery is represented by a finite number of charge units and the discharge behaviour is modeled using a discrete-time transient stochastic process. Can take into account variable loads, **rate capacity and recovery effects**, but not thermal effects. Computation request is modest.
- **Electrochemical Models:** Models electro-chemical, thermodynamic processes, physical construction, etc. Can analyze many discharge effects under variable loads, including **rate-capacity effects, thermal effects and recovery effects**. Most accurate but most computationally intensive



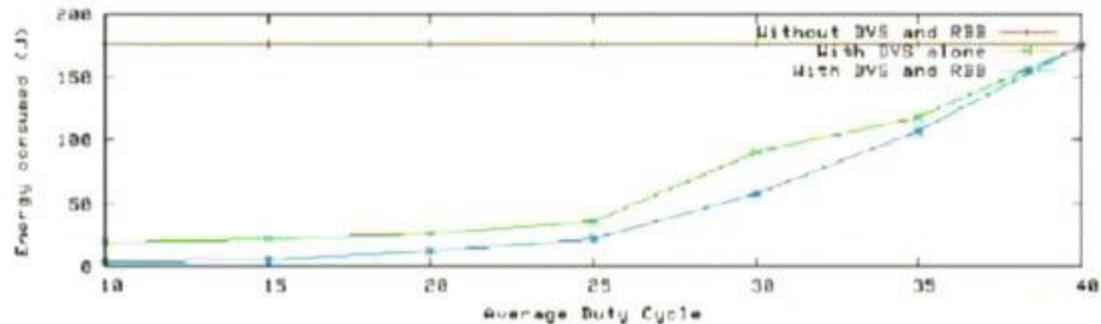
Battery-Driven System Design

Results of the simulation

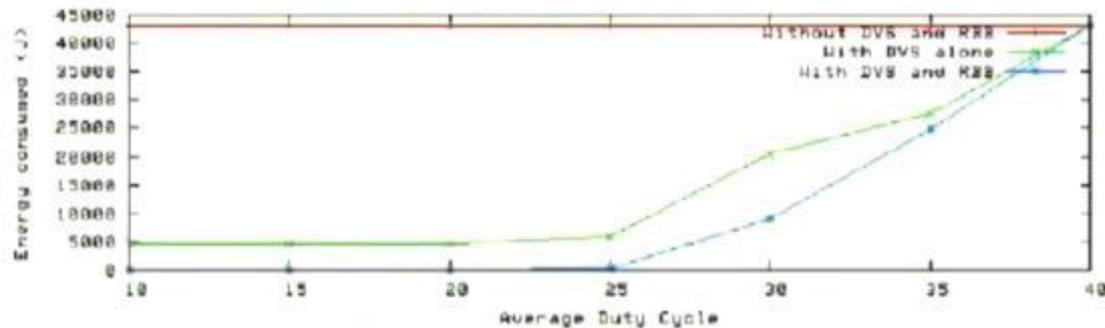


By varying the average duty cycle of the task profile

0.18 μ



0.07 μ





- ❑ In RTL coding there is no provision to use Multi-Vt , Multi-Vdd, Body biasing and power gating in RTL synthesis, 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



Benefits and Impact of Low Power Techniques



Technique	Dynamic Power Benefit	Static Power Benefit	Design Impact	Verification Impact	Implementation Impact
Clock Gating	Large	Small	Small	Small	Small
Multi-Vdd	Large	Small	Little	Low	Medium
DVFS	Large	Small	Medium	Large	Medium
Multi-Vt	Small	Large	Medium	Small	Medium
Power Gating	Small	Very Large	Medium	Large	Medium

- Above techniques can dramatically reduce power consumption in deep submicron chips
- However, these techniques traditionally require ad-hoc, time-consuming, risk-prone, and manual verification and implementation approaches, **unless automated using CAD tools**



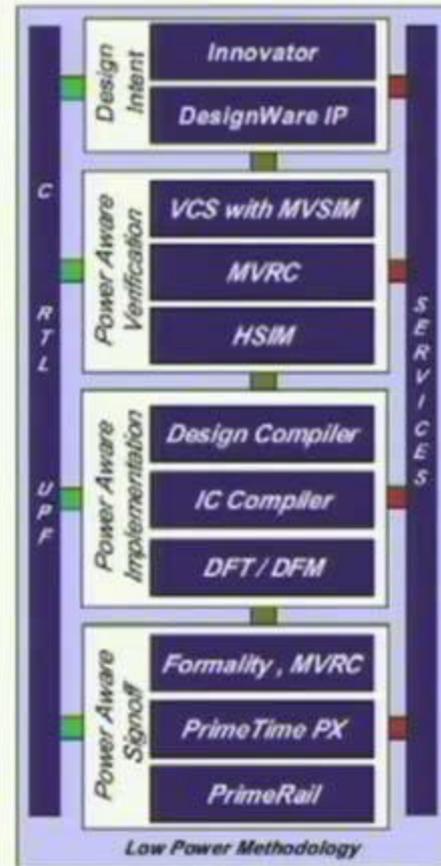
Synopsys' Eclipse Low Power Solution



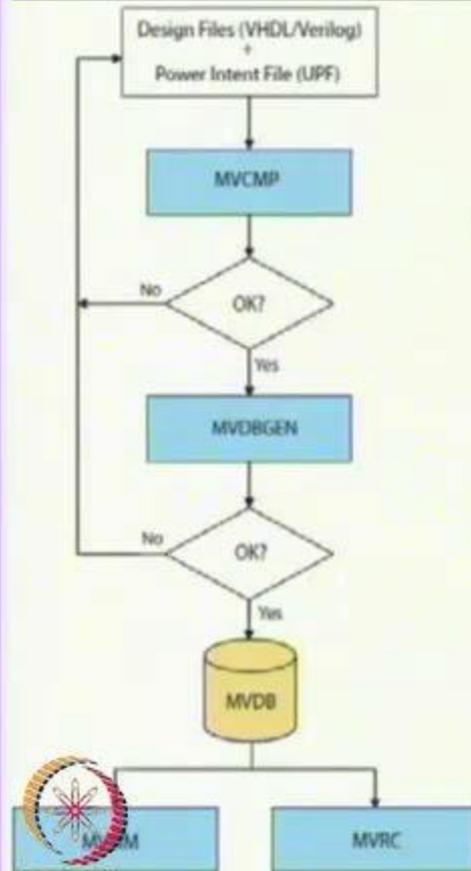
❑ Comprehensive approach – power-aware tools at all levels of design hierarchy starting from early architectural and system-level analysis to verification, RTL synthesis, test, physical implementation and sign-off

❑ Supports the Accellera Unified Power format – an open industry standard to specify power intent.

❑ Backed by the popular “Low Power Methodology Manual” (LPMM)



Multi-Voltage (MV) Verification Flow



❑ Multi-Voltage Compiler (MVCMP)

- Compiles Verilog RTL files and UPF
- Reports syntax & semantics errors

❑ Multi-Voltage Database Generator (MVDBGEN)

- Reads binary design view created by MVCMP, elaborates design & creates Multi-Voltage Database (MVDB)
- Database used by MVSIM and MVRC

❑ Multi-Voltage Rule Checker (MVRC)

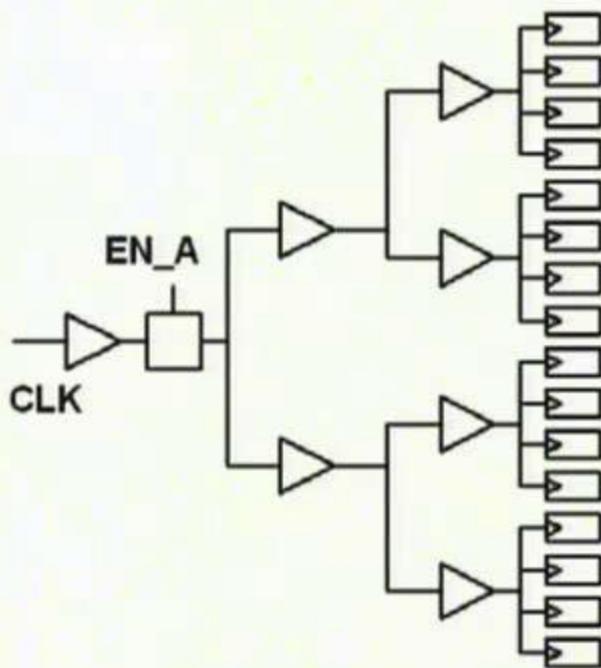
- Structural and power architectural checks on designs without vectors

❑ Multi-Voltage Simulator (MVSIM)

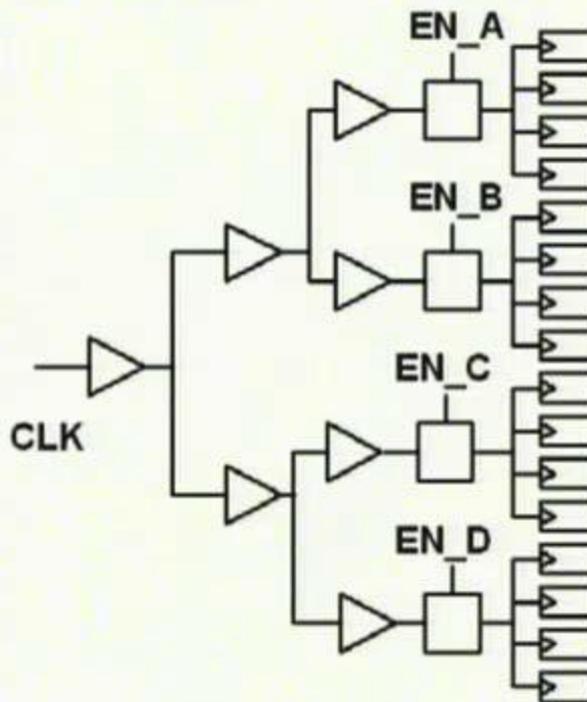
- Co-simulates with VCS to do voltage-aware functional simulation



Clock gating in the Clock Tree



Clock gating close to the root:
optimized for low power



Clock gating close to the registers:
optimized for clock skew

Design Compiler + IC Compiler

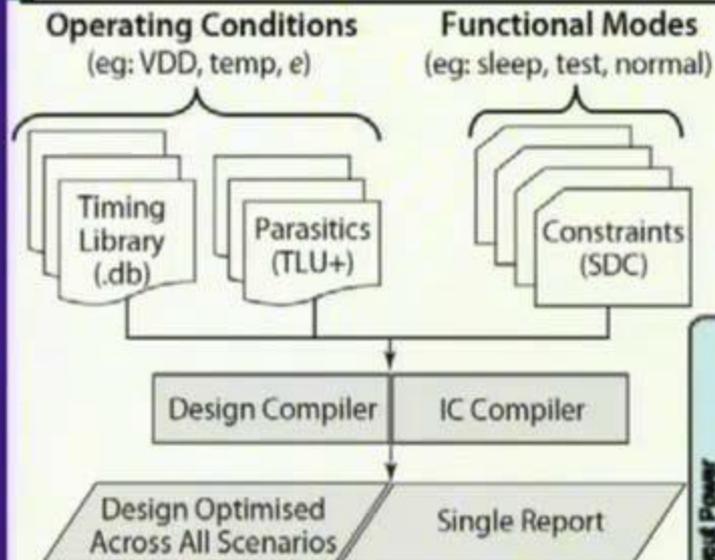


Figure: Multi-Corner Multi-Mode (MCMM) Optimization

➤ Concurrent MCMM-aware synthesis, placement, routing and optimization transforms reduce turn-around time of complex designs

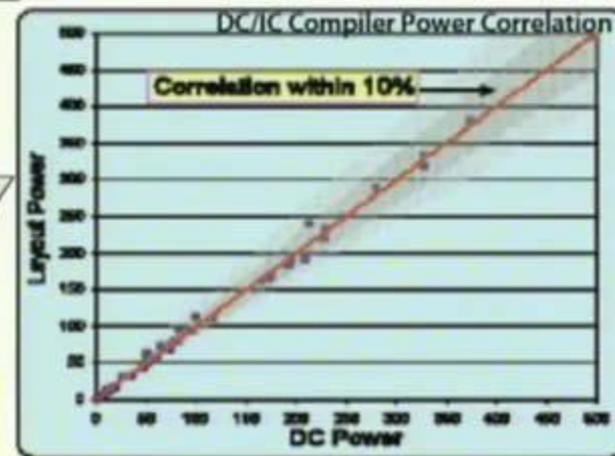


Figure: Power Correlation between Synthesis and Layout