# Module 1

## Introduction

The first integrated circuit was flip-flop with two transistors built by Jack Kilby at Texas Instruments in the year 1958. In the year 2008, Intel's Itanium microprocessor contained more than 2 billion transistors and a 16 Gb Flash memory contained more than 4 billion transistors. So in the range of over 50 years there is the growth rate is around 53%. This incredible growth has come from steady miniaturization of transistors and improvements in manufacturing processes. As transistors became smaller, they also became faster, dissipate less power, and are got cheaper to manufacture. The memory once needed for an entire company's accounting system is now carried by a teenager in her iPod. Improvements in integrated circuits have enabled space exploration, made automobiles safer and more fuel efficient, revolutionized the nature of warfare, brought much of mankind's knowledge to our Web browsers, and made the world a flatter place.

- During the first half of the twentieth century, electronic circuits used large, expensive, power-hungry, and unreliable vacuum tubes.
- In 1947, John Bardeen and Walter Brattain built the first functioning point contact transistor at Bell Laboratories, shown in Figure 1.1(a).
- Later it was introduced by the Bell Lab and named it as **Transistor, T-R-A-N-S-I-S-T-O-R**, because it is a resistor or semiconductor device which can amplify electrical signals as they are transferred through it from input to output terminals.
- Ten years later, Jack Kilby at Texas Instruments realized the potential for miniaturization if multiple transistors could be built on one piece of silicon. Figure 1.1(b) shows his first prototype of an integrated circuit, constructed from a germanium slice and gold wires.

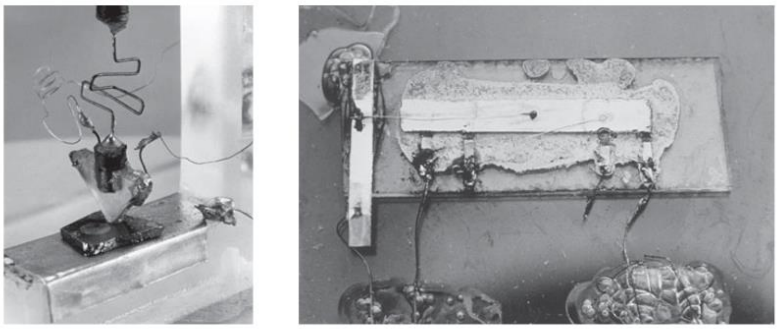



Fig. 1.1(a) First transistor (b) First Integrated Circuit

- Transistors are electrically controlled switches with a control terminal and two other terminals that are connected or disconnected depending on the voltage or current applied to the control.
- After the invention of point contact transistor, Bell Labs developed the bipolar junction transistor, which were more reliable, less noisy and more power-efficient.
- Early integrated circuits used mainly bipolar transistors, which required a small current into the control (base) terminal to switch much larger currents between the other two (emitter and collector) terminals.
- The problem seen with bipolar transistors were the power dissipated by the base current which limited the maximum number of transistors that can be integrated onto a single die.

- Then in 1960 came Metal Oxide Semiconductor Field Effect Transistors (MOSFETs). The advantages seen in MOSFETs were that they draw almost zero control current while idle. It was available in 2 forms as: nMOS and pMOS, using n-type and p-type silicon, respectively.
- In 1963, the first logic gates using MOSFETs was introduced at Fairchild. It included gates used both nMOS and pMOS transistors. This gave the name Complementary Metal Oxide Semiconductor, or CMOS. The circuits used discrete transistors but consumed only nanowatts of power, which was about six times lesser than bipolar transistors.
- MOS ICs became popular because of their low cost, each transistor occupied less area and the fabrication process was simpler. Early commercial processes used only pMOS transistors but it suffered from poor performance, yield, and reliability. Later on Processes using nMOS transistors became common in the 1970s.
- Even though nMOS process was less expensive compared to CMOS, nMOS logic gates consumed power while they were idle. Power consumption became a major issue in the 1980s as hundreds of thousands of transistors were integrated onto a single die. CMOS processes were widely adopted and have essentially replaced nMOS and bipolar processes for nearly all digital logic applications.
- In 1965, Gordon Moore observed that plotting the number of transistors that can be most economically manufactured on a chip gives a straight line on a semi-logarithmic scale. Also he found transistor count doubling every 18 months. This observation has been called **Moore's Law**.
    - Fig 1.2 shows that the number of transistors in Intel microprocessors has doubled every 26 months since the invention of the 4004.
    - Moore's Law is based on scaling down the size of transistors and to some extent building larger chips.

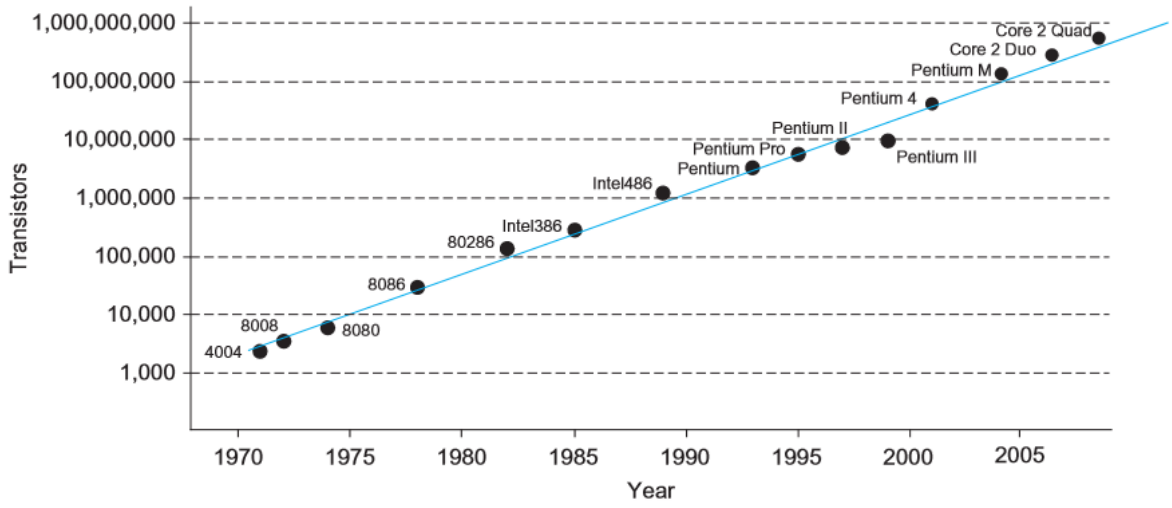



Fig 1.2 Transistors in Intel microprocessors

Level of Integration:

The process of integration can be classified as small, medium, large, very large.

1. Small-Scale Integration (SSI): The number of components is less than 10 in every package. Logic Gates like inverters, AND gate, OR gate and etc. are products of SSI.

2. Medium Scale Integration (MSI): MSI devices has a complexity of 10 to 100 electronic components in a single package. Ex: decoders, adders, counters, multiplexers, and demultiplexers.

3. Large Scale Integration (LSI): Products of LSI contain between 100 and 10,000 electronic components in a single package. Ex: memory modules, I/O controllers, and 4-bit microprocessor systems.

4. Very Large Scale Integration (VLSI): Devices that are results of VLSI contain between 10,000 and 300,000 electronic components. Ex: 8bit, 16-bit, and 32-bit microprocessor systems.

- The feature size of a CMOS manufacturing process refers to the minimum dimension of a transistor that can be reliably built. The 4004 had a feature size of 10μ m in 1971. The Core 2 Duo had a feature size of 45nm in 2008. Feature sizes specified in microns ($10^{-6}$m), while smaller feature sizes are expressed in nanometers ($10^{-9}$ m).

## MOS Transistor:

- Silicon (Si), a semiconductor, forms the basic starting material for most integrated circuits
- Silicon is a Group IV element in periodic table, it forms covalent bonds with four adjacent atoms, as shown in Figure 1.3(a). As the valence electrons of it are involved in chemical bonds, pure silicon is a poor conductor.
- However its conductivity can be increased by introducing small amounts of impurities, called dopants, into the silicon lattice.
- A dopant from Group V of the periodic table, such as arsenic, having five valence electrons. It replaces a silicon atom in the lattice and still bonds to four neighbors, so the fifth valence electron is loosely bound to the arsenic atom, as shown in Figure 1.3(b). Thermal vibration at room temperature is sufficient to free the electron. This results in As+ ion and a free electron. The free electron can carry current and this is an n-type semiconductor.

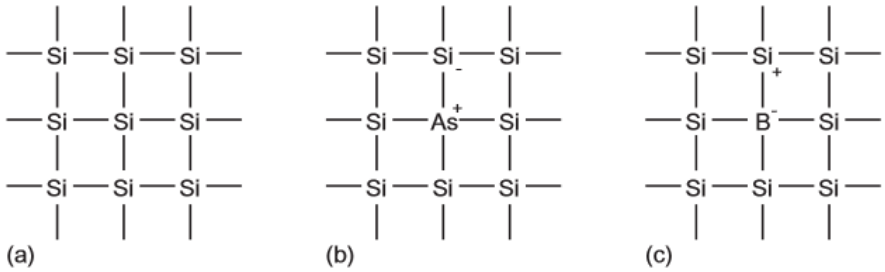



Fig 1.3 Silicon lattice and dopant atoms

- A Group III dopant, such as boron, having three valence electrons, as shown in Fig 1.3(c). The dopant atom can borrow an electron from a neighboring silicon atom, which in turn becomes short by one electron. That atom in turn can borrow an electron, and so forth, so the missing electron, or hole, can propagate about the lattice. The hole acts as a positive carrier so we call this a p-type semiconductor.
- A Metal-Oxide-Semiconductor (MOS) structure is created by superimposing several layers of conducting and insulating materials to form a sandwich-like structure.

- Transistors can be built on a single crystal of silicon, which are available as thin flat circular wafer of 15–30 cm in diameter. CMOS technology provides two types of transistors an n-type transistor (nMOS) and a p-type transistor (pMOS).
- Transistor operation is controlled by electric fields so the devices are also called Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) or simply FETs. Cross-sections and symbols of these transistors are shown in Figure 1.4. The n+ and p+ regions indicate heavily doped n- or p-type silicon.

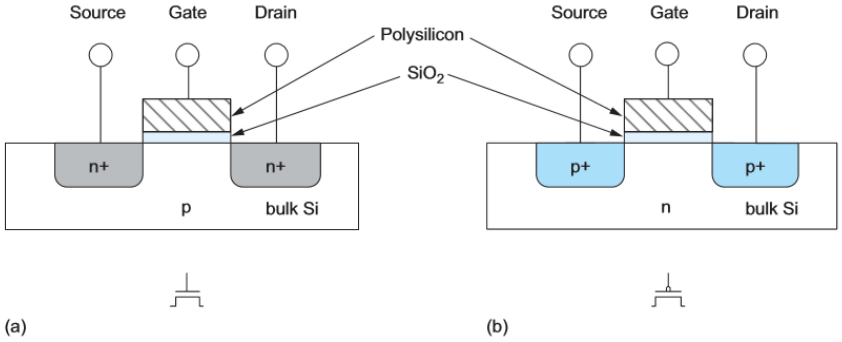



Fig 1.4 (a) nMOS transistor and (b) pMOS transistor

- Each transistor has conducting gate, an insulating layer of silicon dioxide (SiO2, also known as glass), and the silicon wafer, also called the substrate/body/bulk. Gates of early transistors were built from metal, so was called Metal-Oxide-Semiconductor, or MOS.
- Even though the gate has been formed from polycrystalline silicon (polysilicon), the name is still metal.
- An nMOS transistor is built with a p-type body and has regions of n-type semiconductor adjacent to the gate called the source and drain. They are physically equivalent and they can be interchangeable. The body is typically grounded.
- A pMOS transistor is just the opposite, consisting of p-type source and drain regions with an n-type body.
- In both the gate is the control input.
- nMOS Transistor:
    - It controls the flow of electrical current between the source and drain.
    - Considering an nMOS transistor, its body is generally grounded so the p–n junctions of the source and drain to body are reverse-biased. If the gate is also grounded, no current flows through the reverse-biased junctions and the transistor is OFF.
    - If the gate voltage is raised, it creates an electric field that starts to attract free electrons to the underside of the Si–SiO2 interface.
    - If the voltage is raised enough, the electrons outnumber the holes and a thin region under the gate called the channel is inverted to act as an n-type semiconductor.
    - Hence, a conducting path is formed from source to drain and current can flow. This is the condition for transistor is ON state.
    - Thus when the gate of an nMOS transistor is high, the transistor is ON and there is a conducting path from source to drain. When the gate is low, the nMOS transistor is OFF and almost zero current flows from source to drain.

- pMOS Transistor:
    - The condition is reversed.
    - The body is held at a positive voltage and also when the gate is at a positive voltage, the source and drain junctions are reverse-biased and no current flows, the transistor is OFF.

- When the gate voltage is reduced, positive charges are attracted to the underside of the Si–SiO2 interface. A sufficiently low gate voltage inverts the channel and a conducting path of positive carriers is formed from source to drain, so the transistor is ON.
- The symbol for the pMOS transistor has a bubble on the gate, indicating that the transistor behavior is the opposite of the nMOS.
- A pMOS transistor is just the opposite of that of nMOS. It is ON when the gate is low and OFF when the gate is high

Transistor symbols and switch-level models is shown in Fig 1.5

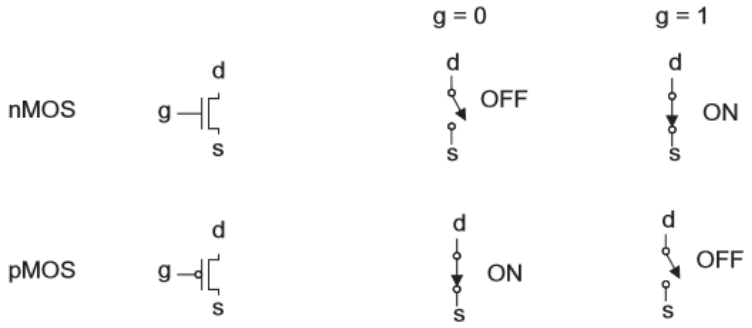



Fig 1.5 Transistor symbols and switch-level models

**MOS Transistor Theory:**

- MOS transistor is a majority-carrier device - current in channel between the source and drain is controlled by a voltage applied to the gate.
  - In nMOS transistor - majority carriers are electrons
  - In pMOS transistor - majority carriers are holes.
- To understand the behavior of MOS transistors, an isolated MOS structure with a gate and body but no source or drain is consider.
- It has top layer of good conducting gate layer. Middle layer is insulating oxide layer and bottom layer is the p-type substrate i.e doped silicon body. Since it is a p-type body carriers are holes

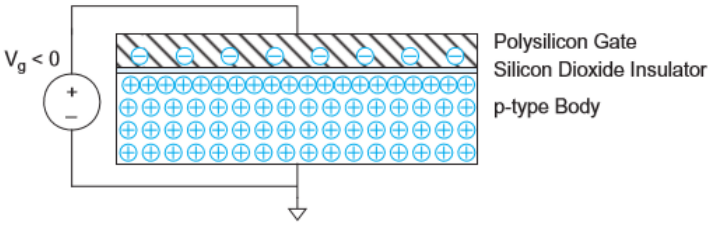



Fig 1.6 (a) Accumulation

- When a negative voltage is applied to the gate, the positively charged holes are attracted to the region beneath the gate. This is called the accumulation mode shown in Fig 1.6(a)
- When a small positive voltage is applied to the gate, the positive charge on the gate repels the holes resulting a depletion region beneath the gate as shown in Fig 1.6(b)

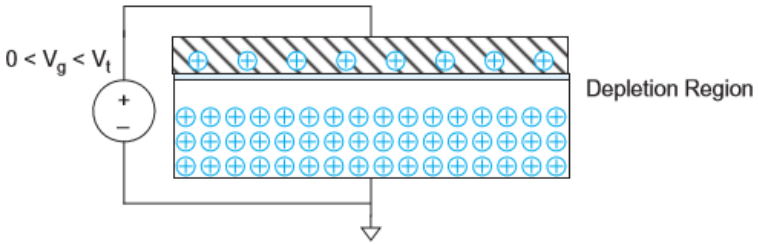



Fig 1.6(b) Depletion

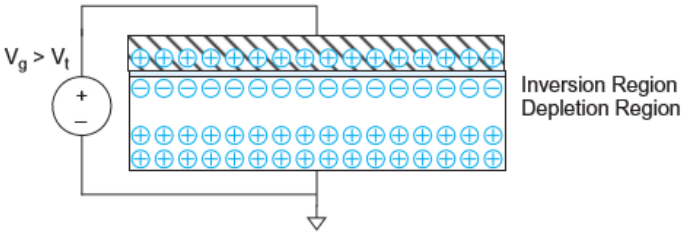



Fig 1.6(c) Inversion

- When a higher positive potential exceeding a critical threshold voltage Vt is applied, the holes are repelled further and some free electrons in the body are attracted to the region beneath the gate. This results a layer of electrons in the p-type body is called the inversion layer.

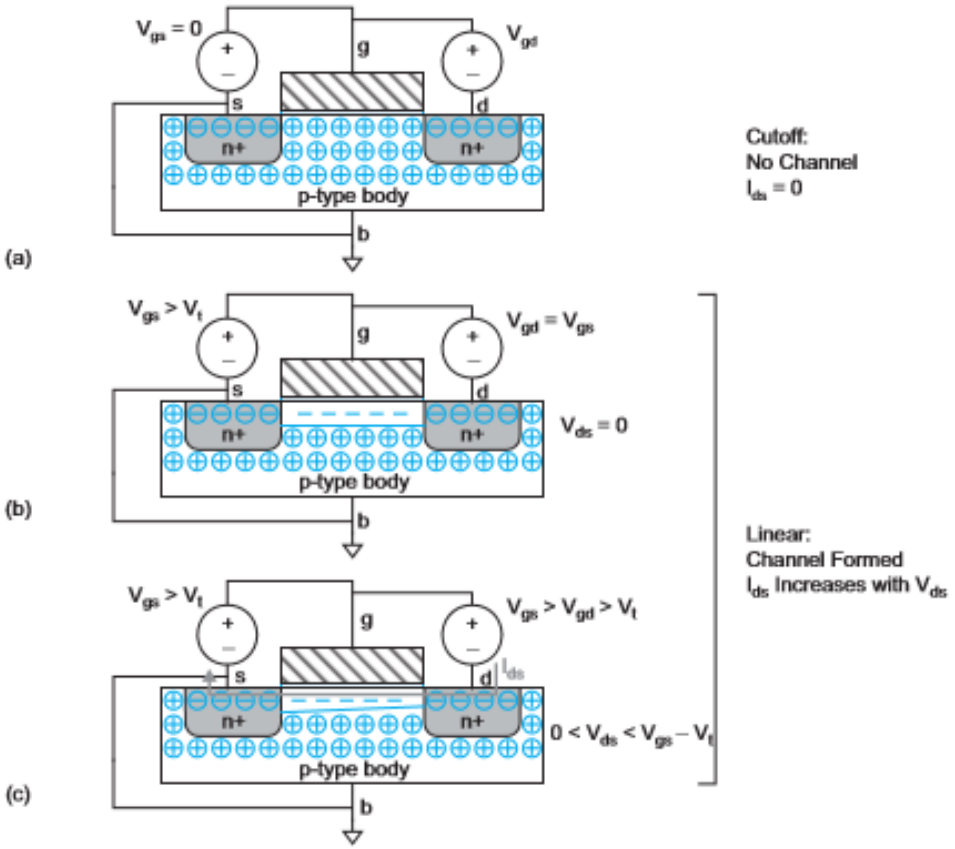



Fig 1.7 (a) nMOS demonstrating Cutoff and Linear operation

- Now considering transistor with MOS stack between two n-type regions called the source and drain the operation is considered.

- When gate-to-source voltage, Vgs is less than threshold voltage and if source is grounded, then the junctions between the body and the source or drain are zero-biased or reverse-biased and no current flows. We say the transistor is OFF, and this mode of operation is called **cutoff**. This is shown in above fig. 1.7(a)
- When the gate voltage is greater than the threshold voltage, an inversion region of electrons (majority carriers) called the channel connects the source and drain, creating a conductive path and turning the transistor ON Fig 1.7(b). The number of carriers and the conductivity increases with the gate voltage. The potential difference between drain and source is Vds= Vgs - Vgd. If Vds = 0 (i.e., Vgs =Vgd), there is no electric field tending to push current from drain to source. When a small positive potential Vds is applied to the drain, current Ids flows through the channel from drain to source. This mode of operation is termed **linear, resistive, triode, nonsaturated, or unsaturated** mode as shown in Fig 1.7 (c)
- If Vds becomes sufficiently large that Vgd < Vt, the channel is no longer inverted near the drain and becomes pinched off (Fig 1.7(d)). However, conduction is still brought about by the drift of electrons under the influence of the positive drain voltage. Above this drain voltage the current Ids is controlled only by the gate voltage and ceases to be influenced by the drain. This mode is called **saturation**.

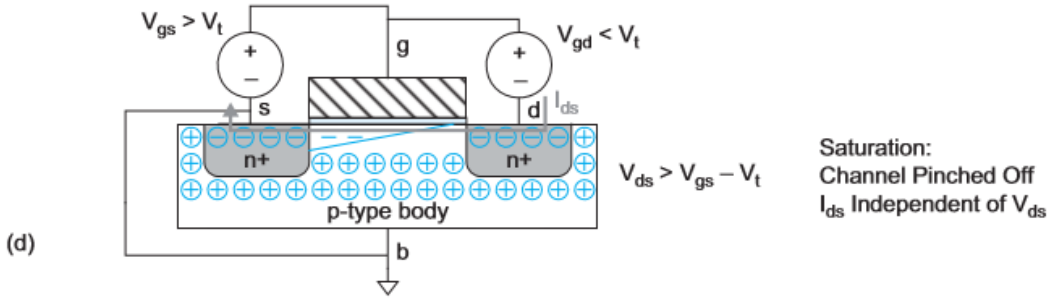



Fig 1.7 (d) Saturation

**pMOS Transistor**

- The pMOS transistor in Fig 1.8 operates in just the opposite fashion. The n-type body is tied to a high potential so the junctions with the p-type source and drain are normally reverse-biased. When the gate is also at a high potential, no current flows between drain and source. When the gate voltage is lowered by a threshold Vt, holes are attracted to form a p-type channel immediately beneath the gate, allowing current to flow between drain and source.

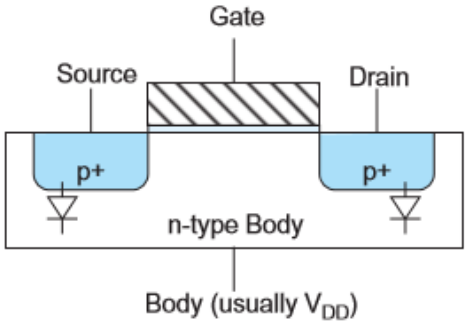



Fig 1.8 pMOS Transistor

**Ideal I-V Characteristics:**

- Considering Shockley model, which assumes the current through an OFF transistor is 0 i.e., when Vgs < Vt there is no channel and current from drain to source is 0.

- In other 2 regions (linear and saturation) channel is formed and electrons flow from source to drain at a rate proportional to electric field (field between source and drain)
- If the amount of charge in the channel and the rate at which it moves is known, we can determine the current.
- The charge on parallel plate of capacitor is given by, Q = C.V
- Here the charge in the channel is denoted by Qchannel and is given by

  Qchannel = Cg . Vc

  Where Cg – capacitance of gate to the channel

  Vc – amount of voltage attracting charge to the channel
- If we model the gate as a parallel plate capacitor, then capacitance is given by

  Area/Thickness

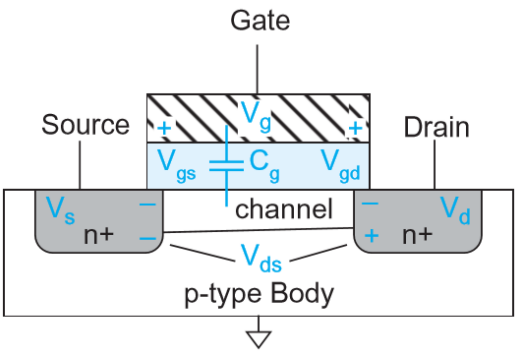



Fig a. Capacitance effect at the gate terminal

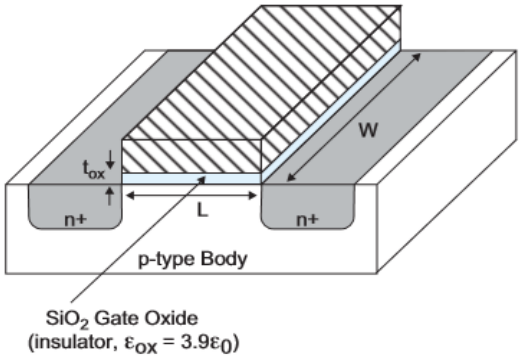



Fig b. Transistor dimensions

- If gate is having length L and width W and the oxide thickness is tox, as shown in Fig b, the capacitance is given by

$$Cg = \frac{\text{Ɛox } W\ L}{tox}$$

  Where Ɛox is the permittivity of oxide and it is 3.9 Ɛo.

  Ɛo is permittivity of free space, $8.85 \times 10^{-14}$ F/cm,
- Often, the Ɛox/tox term is called Cox, the capacitance per unit area of the gate oxide.
- Thus capacitance is now Cg = Cox W L
- Now the charges induced in channel due to gate voltage is determined by taking the average voltage between source and drain (Fig. a) and it is given by

  Vc = (Vs + Vd)/2

To form the channel and carriers to flow, the voltage condition at source and drain is as follows:

Vs = Vgs – Vt

Vd = (Vgs – Vt) – Vds

Thus average voltage is now

$$Vc = \frac{(Vgs-Vt)+(Vgs-Vt)-Vds}{2}$$

Upon simplification, Vc is now

Vc = (Vgs –Vt) – Vds/2

Thus Qchannel = $C_{ox}WL[(Vgs - Vt) - Vds/2]$

- The velocity of charge carrier in the channel is proportional to lateral electric field (field between source and drain) and it is given by,

  $$v = \mu E$$

  Where μ is the proportionality constant called 'mobility'
- The electric field E is the voltage difference between drain and source to the length of channel. Given by,

  $$E = \frac{Vds}{L}$$
- The current in the channel is given by the total amount of charge in channel and time taken by them to cross. The time taken is given by length to velocity.

  i.e., $$Ids = \frac{total\ charge}{time\ to\ cross\ channel} = \frac{Qchannel}{L/v}$$

  $$Ids = \frac{Cg.Vc}{L}\ v = \frac{Cg.Vc}{L}\ \mu E$$

  $$Ids = \frac{Cg.Vc}{L}\ \mu\left(\frac{Vds}{L}\right)$$

  $$Ids = \frac{Cox\ W\ L\ [(Vgs-Vt)-\frac{Vds}{2}]}{L}\ \mu(\frac{Vds}{L})$$

  Upon simplification, Ids is given by:

  $$Ids = \mu\ Cox\frac{W}{L}\left[(Vgs - Vt) - \frac{Vds}{2}\right]Vds$$

  $$Ids = \beta\ \left[(Vgs - Vt) - \frac{Vds}{2}\right]Vds$$

  Where $\beta = \mu\ Cox\frac{W}{L}$
- The above equation for current describes linear region operation for Vgs > Vt
- When Vds is increased to larger value i.e., Vds > Vsat = Vgs – Vt, the channel is no longer inverted and at the drain channel gets pinched off.
- Beyond this is the drain current is independent of Vds and depends only on the gate voltage called as saturation current.
- The expression for the saturation current is given by

  $$Ids = \mu\ Cox\frac{W}{L}\left[(Vgs - Vt) - \frac{Vds}{2}\right]Vds$$

  $$Ids = \mu\ Cox\frac{W}{L}\left[(Vgs - Vt) - \frac{(Vgs-Vt)}{2}\right](Vgs - Vt)$$

  $$Ids = \mu\ Cox\frac{W}{L}\left[\frac{(Vgs-Vt)}{2}\right](Vgs - Vt)$$

  $$Ids = \beta/2\ (Vgs - Vt)^2$$

  Where $\beta = \mu\ Cox\frac{W}{L}$

Summarizing the currents in all the 3 regions is

Ids = 0 for Vgs < Vt cutoff

$Ids = \beta \left[(Vgs - Vt) - \frac{Vds}{2}\right] Vds$ for Vds < (Vgs-Vt) linear region

$Ids = \beta \left[(Vgs - Vt) - \frac{Vds}{2}\right] Vds$ for Vds > (Vgs-Vt) saturation region

The plot of current and voltage i.e., I-V Characteristics is shown in the fig.

**pMOS Transistor:**

pMOS transistors behave in the same way, but with the signs of all voltages and currents reversed. The I-V characteristics are in the third quadrant, as shown in Fig.

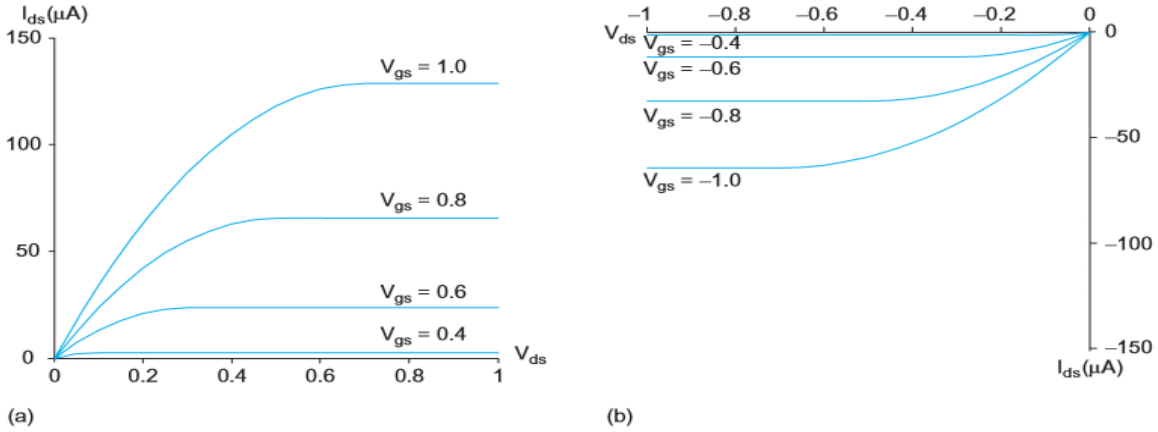



Fig. Plot of I-V characteristics of (a) nMOS and (b) pMOS

**Non ideal I-V Effects:**

- The ideal I-V model does not consider many effects that are important to modern devices. These effects are as follows:

**Velocity saturation:**

- Electron velocity is related to electric field through mobility by the equation
  v = μ E , where E is the lateral electric field or field between drain and source.
- It is assumed that μ is constant and independent parameter w.r.t, E
- At higher E, μ is no more constant and it varies and is due to velocity saturation effect
- When electric field reaches a critical value say $E_{sat}$, the velocity of charge carriers tend to saturate due to scattering effect at $E_{sat}$. This is shown in graph below.
- The impact of velocity saturation is modelled as follows:
  Before the velocity reaches critical value,

$$v = \frac{\mu\, Elat}{1 + \frac{Elat}{Esat}}$$

When the velocity reaches critical and greater it is given by,

$$v = Vsat$$

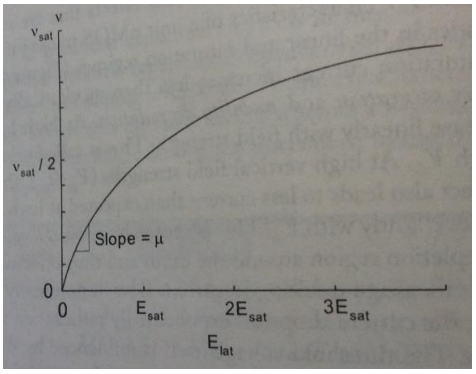



Fig. carrier velocity vs electric field

- When transistor is not velocity saturated, current Ids is given by

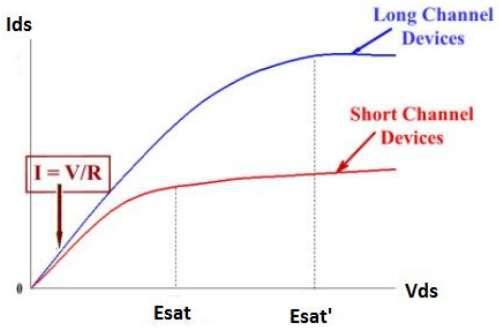
$$Ids = \mu Cox \frac{W}{L} \frac{(Vgs - Vt)^2}{2}$$

and with velocity saturation, current Ids is given as

$$Ids = Cox\ W\ (Vgs - Vt) Vsat$$

Observing both the expression we can say that
Ids depend quadratically on voltage without saturation and depends linearly when fully saturated

- As shown in graph for short channel devices it has extended saturation region (from Esat to Esat') due to velocity saturation.




- As channel length becomes shorter, lateral electrical field increases and transistor becomes more velocity saturated and this decreases drain current Ids.

**Mobility degradation:**

- Velocity of charge carriers depend on electric field and when these carriers travel along the length of channel, they get attracted to the surface (i.e., Gate) by the vertical electric field (field created by gate voltage)
- Hence they bounce against the surface during their travel
- This reduces surface mobility in comparison with the mobility along the channel.
- This is known as mobility degradation and has an impact on I-V characteristics.
- As mobility decreases the current also decreases.

**Channel length Modulation:**

- Ideally drain current Ids is independent on Vds in the saturation region making transistor a perfect current source.

- When Vds is increased further, near the drain barrier is build due to depletion region and reduces the length of the channel.
- This results in reducing the length of the channel by $L_d$. This is shown in Fig below. Thus in saturation the effective channel length is modelled as

$$L_{eff} = L - L_d$$

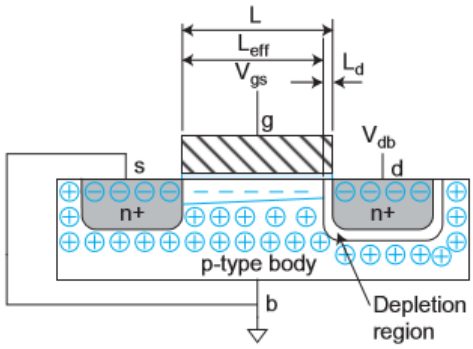



Fig. Channel length modulation in saturation mode

- To avoid introducing the body voltage into our calculations, assume the source voltage is close to the body voltage so Vdb ~ Vds. Hence, increasing Vds decreases the effective channel length. Shorter channel length results in higher current; thus, Ids increases with Vds in saturation
- This is modeled as

$$Ids = \frac{\beta}{2}(Vgs - Vt)^2(1 + \lambda Vds)$$

Where λ is empirical channel length modulation factor

- The equation can also be written as $Ids = \frac{\mu Cox}{2}\frac{W}{L}(Vgs - Vt)^2(1 + \lambda Vds)$

Thus as L decreases W/L ratio increases, this in turn increases Ids. Thus transistor in saturation is no more a constant current source.

Note: Channel length modulation is important in analog designs as it reduces the gain of the amplifier. But for digital circuits channel length modulations has no much importance.

**Body Effect:**

- MOSFETs have 4$^{th}$ implicit terminal called body/substrate along with gate, source and drain.
- The threshold voltage Vt which is assumed to remain constant is no more a constant value and varies as potential between source and body is varied. This effect is called body effect.
- The variation in the threshold voltage is modeled by the equation

$$V_t = V_{t0} + \gamma\left(\sqrt{\phi_s + V_{sb}} - \sqrt{\phi_s}\right)$$

Where Vto is the threshold voltage when source and body are at same potential
$\Phi_s$ is the surface potential
γ is the body effect coefficient and these two are given by
Vsb is the source to body potential

$$\phi_s = 2v_T \ln\frac{N_A}{n_i}$$

$$\gamma = \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2q\varepsilon_{si}N_A} = \frac{\sqrt{2q\varepsilon_{si}N_A}}{C_{ox}}$$

$\upsilon_T$ is voltage at room temperature ($\upsilon_T$ = KT/q at $30^0$ it is 26mV)
$N_A$ is the doping concentration level
$n_i$ is the intrinsic carrier concentration
q is charge (q = 1.6 $\times 10^{-19}$ C)
tox oxide thickness
Ɛox is the permitivity of oxide and is given by 3.9 Ɛo, where Ɛo is the permittivity of free space = $8.825\times10^{-14}$ F/cm
Ɛsi is the permittivity of silicon and given by 11.7 Ɛo and Ɛo=$8.825\times10^{-14}$ F/cm

- Body effect parameter γ depends on doping level concentration, thus by varying γ threshold voltage can be varied
- Also Vt depend on Vsb thus by proving appropriate potential threshold voltage can be varied.
- Thus a proper body bias can intentionally be applied to alter the threshold voltage, permitting trade-offs between performance and subthreshold leakage current

**Subthreshold Conduction:**

- The ideal I-V model assumes current flows from source to drain only when Vgs >Vt (when gate voltage is high). But in practical transistors, current does not abruptly cut off below threshold, but rather drops off exponentially.
- This regime of Vgs <Vt is called weak inversion/ subthreshold.
- This conduction of current is known as leakage and is undesired when the transistor is off
- The subthreshold conduction is modeled using equation given below

$$Ids = Idso\, e^{\frac{Vgs-Vt}{n\,VT}}\left[1 - e^{\frac{-Vds}{VT}}\right]$$

and $$I_{ds0} = \beta v_T^2 e^{1.8}$$

$I_{dso}$ is the current at saturation and is dependent on process and device geometry
Vt is the threshold voltage and $\upsilon_T$ voltage at room temperature.

- In the expression Ids is 0 if $V_{ds}$ is 0 and increases to full when $V_{ds}$ is few multiples of $\upsilon_T$
- Graph shows conduction in the subthreshold region

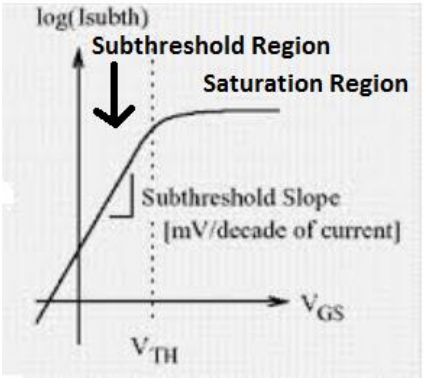



Fig. Subthreshold conduction

- Subthreshold conduction is useful for designing low power analog circuits and dynamic circuits as it reduces threshold voltage and results in low power consumptions.

**Drain Induced Barrier Lowering (DIBL):**

- As the drain voltage Vds is increased it creates an electric field that affects the threshold voltage.
- This effect is called drain-induced barrier lowering (DIBL) and this effect is especially pronounced in short-channel transistors.
- As the channel length decreases, the DIBL effect shows up and the variation caused in the threshold voltage can be modeled as

$$V_t = V_{t0} - \eta V_{ds}$$

η is the DIBL coefficient

**Junction Leakage:**

- The MOS structure is considered there exists p–n junctions between diffusion and the substrate. With CMOS structures p–n junctions between diffusion and the substrate or well, forming diodes, as shown in Fig. The well-to-substrate junction is another diode.

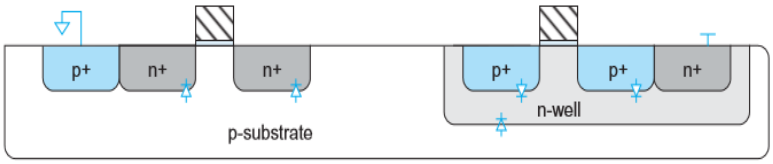



Fig. CMOS structure showing formation of p-n junctions between diffusion and substrate and also between well and substrate

- The substrate and well are tied to GND or $V_{DD}$ so that these diodes does not get into forward biased condition until voltage is applied in normal operation.
- But in reverse-biased conditions these diodes still conduct a small amount of current $I_D$. This leakage current is modeled using equation

$$I_D = I_S\left(e^{\frac{V_D}{\phi_T}} - 1\right)$$

Where, $I_D$ is the diode current
$I_S$ is the diode reverse bias saturation current
$V_D$ is the diode voltage (either Vsb or Vdb)

- $I_S$ depends on doping levels and on the area and perimeter of the diffusion region (geometry) and $V_D$
- Leakage current usually lies in the range of 0.1 – 0.01 fA/μm$^2$, which is negligible when compared to other leakage currents.

**Tunneling (Focoler Nordheium Tunneling):**

- According to quantum mechanics, for thinner gate oxides there is a nonzero probability that an electron in the gate will find itself on the other side of the oxide, (i.e., in the region below gate/ channel).
- This effect of carriers crossing a thin barrier is called tunneling, and results in leakage current through the gate called gate leakage current.
- Thus gate oxide cannot be considered as an ideal insulator. This effects the circuit functionality and increases power consumption due to static gate current.

- Fig shows plot of gate leakage current density $J_G$ against voltage for different oxide thickness. It can be observed that as oxide thickness decreases the leakage current density increases.

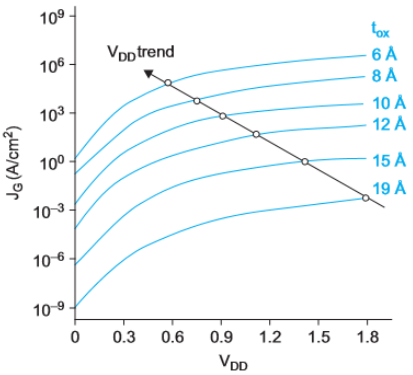



Fig. plot of gate leakage current density vs voltage for different $t_{ox}$

- Research is going on in finding an alternate to silicon dioxide and silicon nitrate is one contender for this.

Note: As mobility of electrons is more than holes in silicon, tunneling current magnitude for nMOS is more compared pMOS.

**Temperature Dependence:**

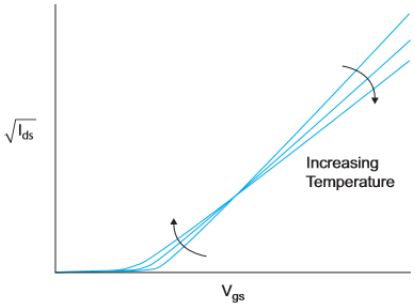
- Transistor characteristics are influenced by temperature
  - Carrier mobility – decreases with temperature and this is modeled using the relation

$$\mu(T) = \mu(T_r)\left(\frac{T}{T_r}\right)^{-k_\mu}$$

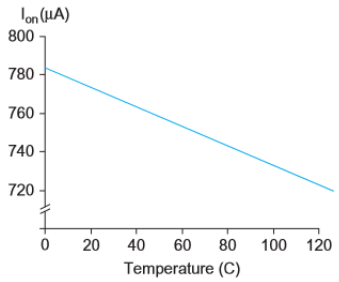
    where T is absolute temperature, Tr is room temperature, $k_\mu$ is fitting constant.
  - Threshold voltage – magnitude of threshold voltage decreases linearly with temperature and can be modeled as

$$V_t(T) = V_t(T_r) - k_{vt}(T - T_r)$$

    where $k_{vt}$ is typically about 1–2 mV/K.
  - Junction Leakage – increases with temperature because Is (diode reverse bias current) strongly depends on temperature
  - Velocity saturation – occurs sooner with temperature.
  - With increase in temperature drain current decreases with temperature when transistor is ON and when transistor is OFF, the junction leakage and subthreshold conduction contribute to leakage current and this increase. This condition is shown in the graph.




I–V characteristics of nMOS transistor in saturation at various temperatures




$I_{dsat}$ vs. temperature

- However, the circuit performance can be improved by providing cooling systems like heat sinks, water cooling, thin film refrigerator and liquid nitrogen.
- Advantages of using at lower temperatures are
  1. Leakages can be reduced
  2. With lower temperature, reducing threshold voltage it can be used in power saving
  3. Most wear out mechanisms are temperature dependent and if used at lower temp they are more reliable

**Geometry Dependence:**

- The layout designer would draws transistors with width and length $W_{drawn}$ and $L_{drawn}$.
- While mask preparation the actual gate dimensions may differ by $X_W$ and $X_L$.
- While diffusion process, the source and drain would tend to diffuse laterally under the gate by $L_D$, causing a smaller effective channel length that the carriers must traverse between source and drain. Similarly, $W_D$ accounts for smaller width while diffusion.
- Combing all these factors transistor, lengths and widths that should be used in place of L and W is given by

$$L_{\text{eff}} = L_{\text{drawn}} + X_L - 2L_D$$
$$W_{\text{eff}} = W_{\text{drawn}} + X_W - 2W_D$$

- If there is variations in the length and width of the transistor there will be variations in the performance. For example, if the currents have to be matched then length should not be varied.

**DC Transfer Characteristics**

- DC transfer characteristics of a circuit relate the output voltage to the input voltage, assuming the input changes slowly enough that capacitances have plenty of time to charge or discharge,
  CMOS Inverter Static Characteristics

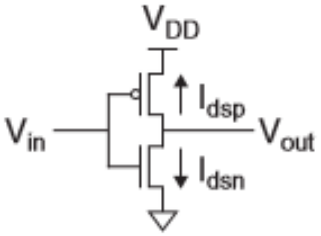



Fig 1.9 CMOS Inverter

CMOS inverter shown in Fig 1.9. Table below outlines various regions of operation for the n- and p-transistors. In this table, Vtn is the threshold voltage of the n-channel device, and Vtp is the threshold voltage of the p-channel device. Note that Vtp is negative. The equations are given both in terms of Vgs/Vds and Vin/Vout. As the source of the nMOS transistor is grounded, Vgsn = Vin and Vdsn = Vout. As the source of the pMOS transistor is tied to $V_{DD}$, Vgsp =Vin – $V_{DD}$ and Vdsp =Vout – $V_{DD}$.

| | Cutoff | Linear | Saturated |
|---|---|---|---|
| nMOS | $V_{gsn} < V_{tn}$ | $V_{gsn} > V_{tn}$ | $V_{gsn} > V_{tn}$ |
| | $V_{in} < V_{tn}$ | $V_{in} > V_{tn}$ | $V_{in} > V_{tn}$ |
| | | $V_{dsn} < V_{gsn} - V_{tn}$ | $V_{dsn} > V_{gsn} - V_{tn}$ |
| | | $V_{out} < V_{in} - V_{tn}$ | $V_{out} > V_{in} - V_{tn}$ |
| pMOS | $V_{gsp} > V_{tp}$ | $V_{gsp} < V_{tp}$ | $V_{gsp} < V_{tp}$ |
| | $V_{in} > V_{tp} + V_{DD}$ | $V_{in} < V_{tp} + V_{DD}$ | $V_{in} < V_{tp} + V_{DD}$ |
| | | $V_{dsp} > V_{gsp} - V_{tp}$ | $V_{dsp} < V_{gsp} - V_{tp}$ |
| | | $V_{out} > V_{in} - V_{tp}$ | $V_{out} < V_{in} - V_{tp}$ |

- The objective is to find the variation in output voltage (Vout) as a function of the input voltage (Vin). This may be done graphically, for simplicity, we assume Vtp = –Vtn and that the pMOS transistor is 2–3 times as wide as the nMOS transistor so βn = βp.
- The plot shows Idsn and Idsp in terms of Vdsn and Vdsp for various values of Vgsn and Vgsp using drain current equation.
- Fig 1.10(b) shows the same plot of Idsn and |Idsp| now in terms of Vout for various values of Vin. The possible operating points of the inverter, marked with dots, are the values of Vout where Idsn = |Idsp| for same Vin.
- These operating points are plotted on Vout vs. Vin axes in Fig. (c) to show the inverter DC transfer characteristics.
- The supply current $I_{DD}$ = Idsn = |Idsp| is also plotted against Vin in Fig (d) showing that both transistors are momentarily ON as Vin passes through voltages between GND and $V_{DD}$, resulting in a pulse of current drawn from the power supply.
- The operation of the CMOS inverter can be divided into five regions indicated on Fig 1.10(c). The state of each transistor in each region and state of output is shown in Table 2.
    - In region A, the nMOS transistor is OFF so the pMOS transistor pulls the output to $V_{DD}$.
    - In region B, the nMOS transistor starts to turn ON, pulling the output down.
    - In region C, both transistors are in saturation.
    - In region D, the pMOS transistor is partially ON
    - In region E, pMOS is completely OFF, leaving the nMOS transistor to pull the output down to GND.

| Region | Condition | p-device | n-device | Output |
|---|---|---|---|---|
| A | $0 \le V_{in} < V_{tn}$ | linear | cutoff | $V_{out} = V_{DD}$ |
| B | $V_{tn} \le V_{in} < V_{DD}/2$ | linear | saturated | $V_{out} > V_{DD}/2$ |
| C | $V_{in} = V_{DD}/2$ | saturated | saturated | $V_{out}$ drops sharply |
| D | $V_{DD}/2 < V_{in} \le V_{DD} - \lvert V_{tp} \rvert$ | saturated | linear | $V_{out} < V_{DD}/2$ |
| E | $V_{in} > V_{DD} - \lvert V_{tp} \rvert$ | cutoff | linear | $V_{out} = 0$ |

Table 2. Summary of CMOS Inverter Operation

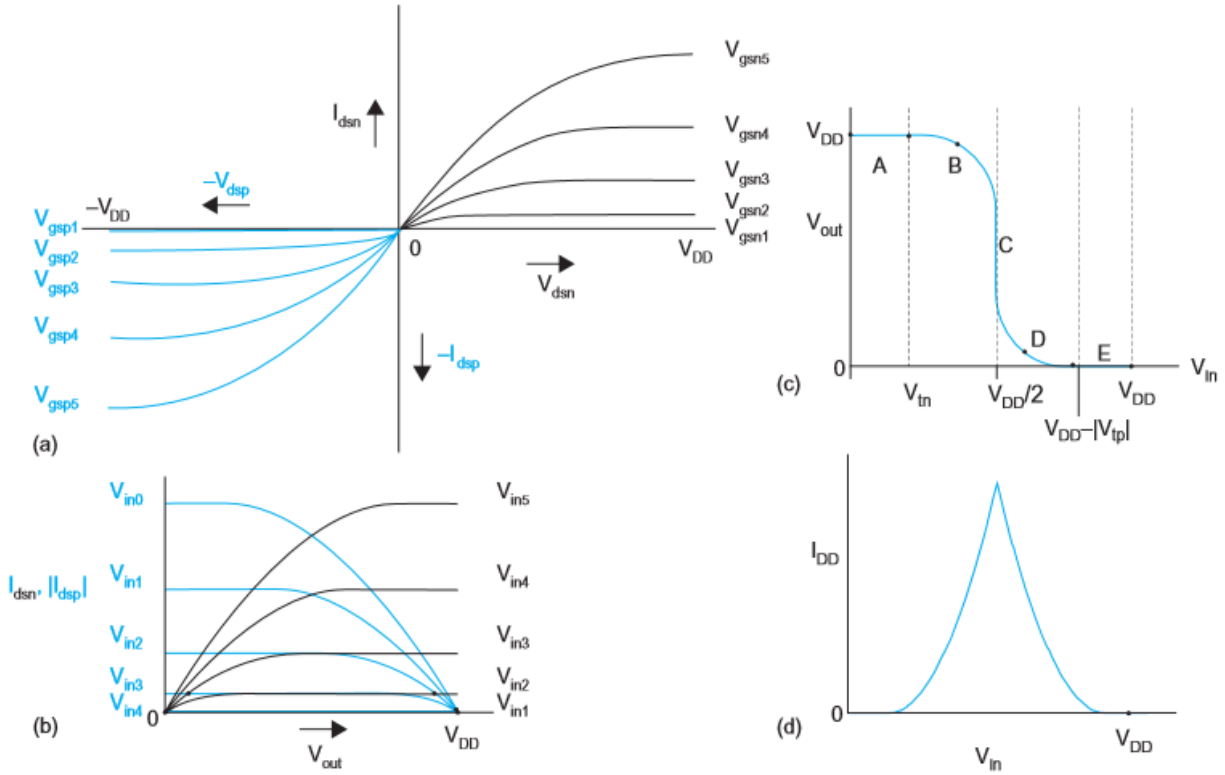
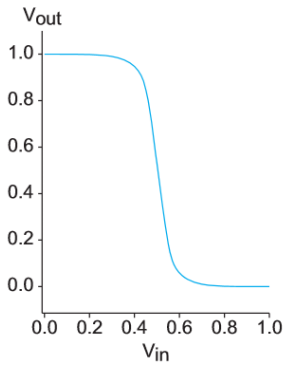



Fig 1.10 Graphical Derivation of CMOS Inverter DC Characteristics




In the fig. the crossover point where Vin = Vout, is called the **'input threshold'**

Fig. CMOS inverter Transfer Characteristics

**Beta Ratio Effects:**

- We have seen that for βn = βp the inverter threshold voltage Vinv is $V_{DD}/2$. This may be desirable because it maximizes noise margins.
- Inverters with different beta ratios βp/βn are called skewed inverters. If βp/βn > 1, the inverter is HI-skewed. If βp/βn < 1, the inverter is LO-skewed. If βp/βn = 1, the inverter has normal skew or is unskewed.
- A HI-skew inverter has a stronger pMOS transistor. Therefore, if the input is $V_{DD}/2$, we would expect the output will be greater than $V_{DD}/2$.
- LO-skew inverter has a weaker pMOS transistor and thus a lower switching threshold.
- Figure explores the impact of skewing the beta ratio on the DC transfer characteristics. As the beta ratio is changed, the switching threshold moves. However, the output voltage transition remains sharp. Gates are usually skewed by adjusting the widths of transistors while maintaining minimum length for speed.

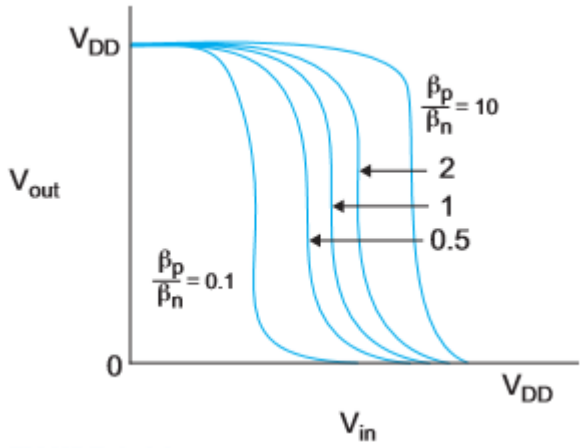



Fig. Transfer Characteristics of Skewed Inverters

**Noise Margin:**

- Noise margin is closely related to the DC voltage characteristics. This parameter allows you to determine the allowable noise voltage on the input of a gate so that the output will not be corrupted.
- The specification most commonly used to describe noise margin (or noise immunity) uses two parameters: the LOW noise margin, $NM_L$, and the HIGH noise margin, $NM_H$.
- With reference to Fig1.12, $NM_L$ is defined as the difference in maximum LOW input voltage recognized by the receiving gate and the maximum LOW output voltage produced by the driving gate.

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

- Similarly $NM_H$ is the difference between the minimum HIGH output voltage of the driving gate and the minimum HIGH input voltage recognized by the receiving gate.

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

Where $V_{IH}$ = minimum HIGH input voltage
$V_{IL}$ = maximum LOW input voltage
$V_{OH}$= minimum HIGH output voltage
$V_{OL}$= maximum LOW output voltage

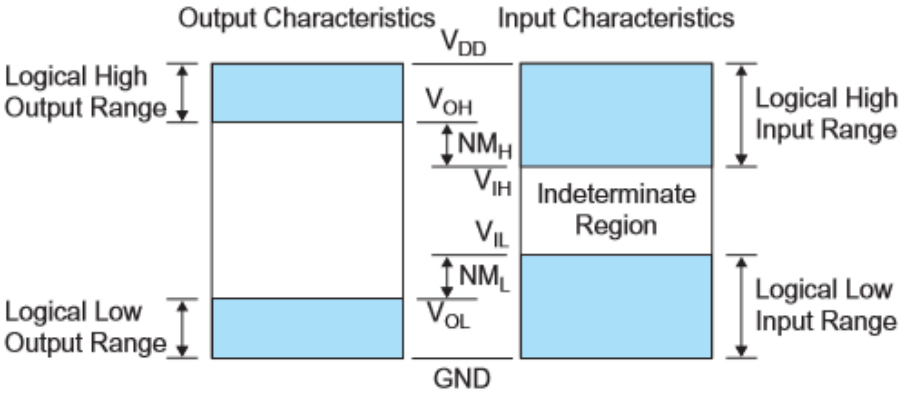



Fig. Noise Margin Definitions

- Inputs between $V_{IL}$ and $V_{IH}$ are said to be in the indeterminate region or forbidden zone and do not represent any legal digital logic levels. Therefore, it is generally desirable to have $V_{IH}$ as close as possible to $V_{IL}$ and for this value to be midway in the "logic swing,"

$V_{OL}$ to $V_{OH}$. This implies that the transfer characteristic should switch abruptly; that is, there should be high gain in the transition region.

- DC analysis gives us the static noise margins specifying the level of noise that a gate may see for an indefinite duration.

**Pass Transistor DC characteristics:**

- nMOS transistors pass '0's well but 1s poorly. Figure (a) shows an nMOS transistor with the gate and drain tied to $V_{DD}$. Imagine that the source is initially at Vs = 0. Vgs > Vtn, so the transistor is ON and current flows. If the voltage on the source rises to Vs = $V_{DD}$ – Vtn, Vgs falls to Vtn and the transistor cuts itself OFF.
- Therefore, nMOS transistors attempting to pass a 1 never pull the source above $V_{DD}$ – Vtn. This loss is sometimes called a threshold drop.
- Similarly, pMOS transistors pass 1s well but 0s poorly. If the pMOS source drops below |Vtp|, the transistor cuts off. Hence, pMOS transistors only pull down to within a threshold above GND, as shown in Fig (b).
- As the source can rise to within a threshold voltage of the gate, the output of several transistors in series is no more degraded than that of a single transistor Fig (c ).
- However, if a degraded output drives the gate of another transistor, the second transistor can produce an even further degraded output Fig(d).

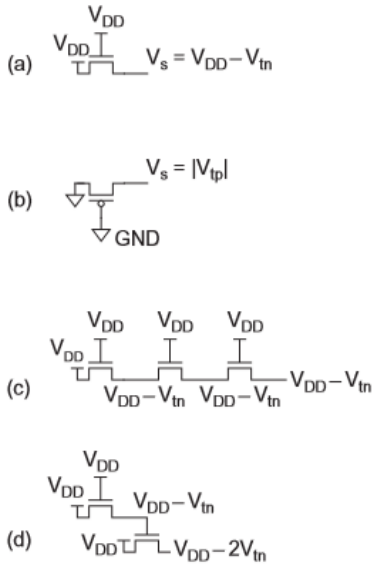



Fig. Pass Transistor Threshold drop

- The problem seen with nMOS and pMOS of not passing strong 1's and strong 0's respectively can be overcome by using Transmission gate.
- It has an nMOS and pMOS connected in parallel as shown in fig below.

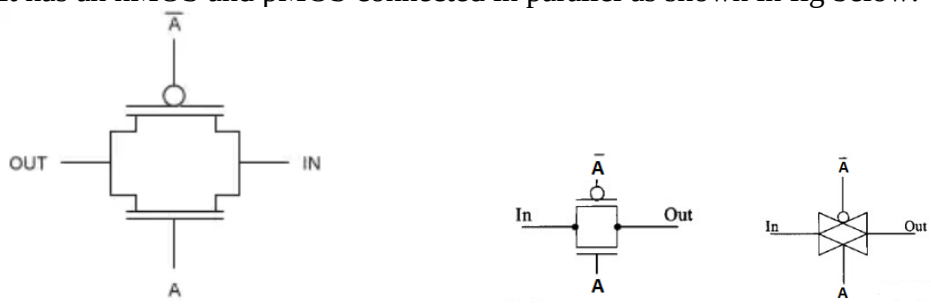



Fig. Schematic and symbol of Transmission gate (TG)

- When A is logic high both transistors are ON and TG is said to be ON. When input is provided as nMOS is not able to transmit strong 1, pMOS will do the function. Similarly when pMOS is not able to transmit strong 0, nMOS will do this function.
- Thus transmission gate is able to send both strong 0 and strong 1 without any signal degradation.
- Transmission gate can be used as
  - Multiplexing element
  - Analog switch
  - Latch element

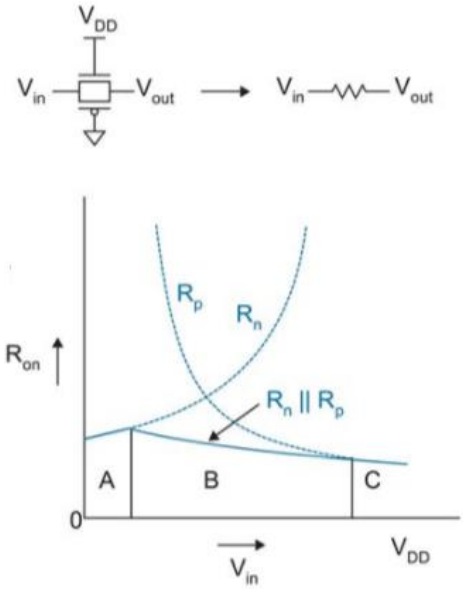
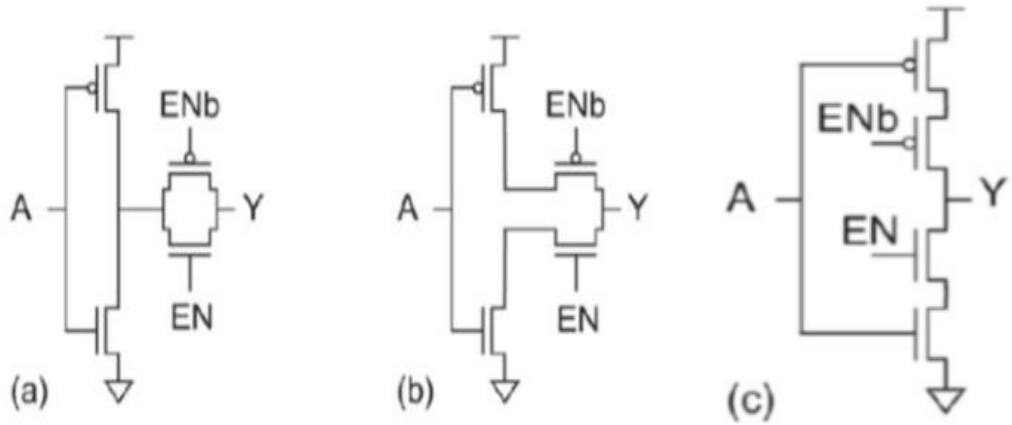



Fig. Resistance of Transmission gate as a function of input voltage

**Tristate inverters:**




- By cascading a transmission gate and an inverter forms a tristate inverter as shown in Fig (a)
- When EN = 1, EN'= 0, thus transmission gate is ON and transmits the output Y as the compliment of inverter input A.
- When EN = 0 and EN' =1, transmission gate is OFF and the output Y is in tristate or high impedance state.
- Fig (b) and (c) shows other configurations of tristate inverters

**Ratioed Inverters Transfer Characteristics**

- Other than CMOS inverter there are also other forms of inverters. One such is shown in the fig. below which has an nMOS with load as resistor.
- This is an nMOS inverter circuit. When Vin = 0, nMOS is OFF and output goes to Vdd through the Rload.
- When Vin = 1, nMOS is ON and pulls the output to gnd.
- When we consider the transfer characteristics and I-V characteristics, we see that as load is increased $V_{OL}$ decreases also the current decreases. Thus choosing load resistor compromises between current and $V_{OL}$.

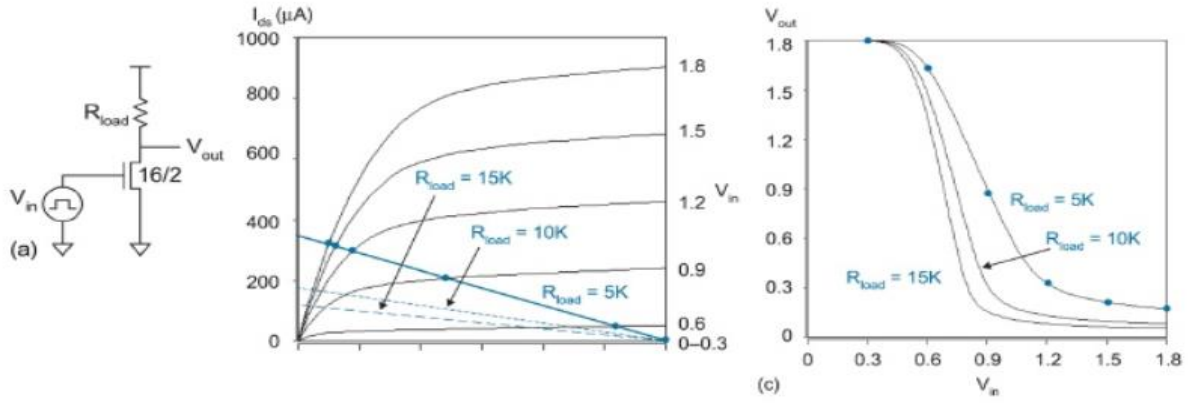



Fig. nMOS inverter with resistive load, I-V characteristics and transfer characteristics

- An alternate to this is using a more practical circuit called pseudo-nMOS inverter circuit, which uses a pMOS transistor as a load with its gate terminal always grounded.
- Here pMOS will be in ON state. When Vin = 0, nMOS is OFF and as pMOS is ON the output rises to Vdd. When Vin = 1, nMOS will be ON and pulls the output to gnd.
- When the transfer characteristics is observed as the W/L ratio is varied for pMOS in the pseudo-nMOS inverter circuit, the shape of the transfer characteristics varies.
- As parameter P (i.e., as W is decreased sharper characteristics is obtained) is varied characteristics varies with higher value of P less sharper characteristics is seen.
- In the circuit P/2 represents the W/L ratio.

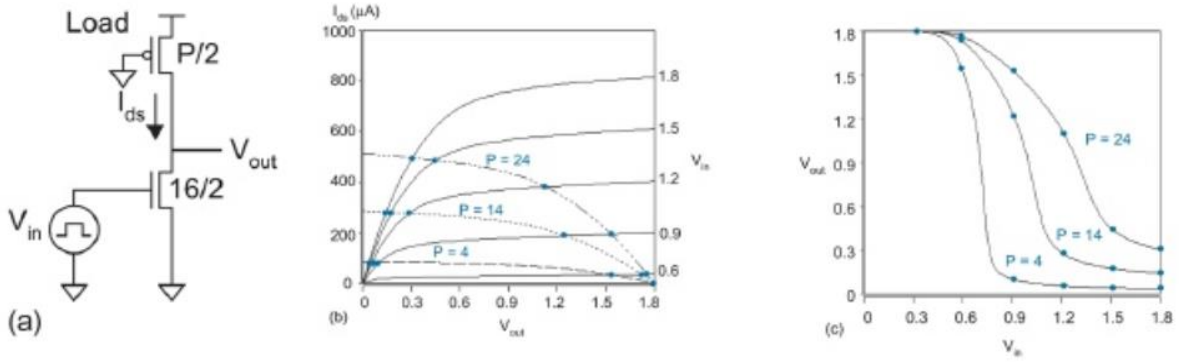



Fig. pseudo-nMOS inverter with I-V characteristics and transfer characteristics

- These types of gates are called as ratioed circuits as transfer function depends on the strength of pull down (pMOS) to pull up (nMOS) devices.
- In these types of circuits ratios must be chosen properly so that circuit operates properly.
- Disadvantage seen with these ratioed circuits are
    - Constant power dissipation
    - Poor noise margin
- However these circuits are used under limited circumstances such as reduced input capacitance and smaller area.

# Fabrication

## nMOS Fabrication:

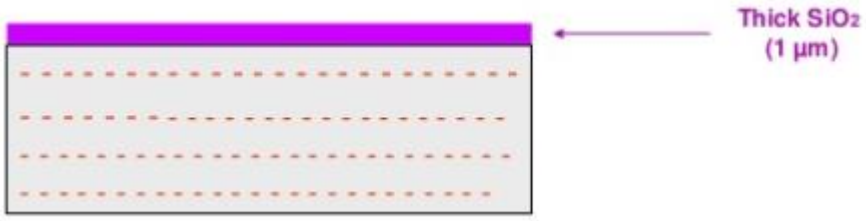
Semiconductor device fabrication is the process of creating integrated circuits in multiple-step sequence of photolithographic and chemical processing during which electronic circuits are gradually created on a wafer made of pure semiconducting material.

The following steps gives general aspect of nMOS fabrication process.

1. Processing is carried on thin wafer cut from single silicon crystal of high purity to which p-type impurities are introduced as crystal is grown. Wafers are around 75 to 150 mm in diameter and 0.4 mm thick. They are doped with boron (p-type) impurity concentration of $10^{15}/cm^3$ to $10^{16}$ $/cm^3$.
2. On this a thick layer if silicon dioxide (SiO2) of 1μm. This protects the surface, act as barrier to dopants and also act as an insulating layer on which other layers can be deposited and patterned.
3. The surface is now covered with photoresist and it is spun to achieve even distribution of required thickness.
4. A mask is used and the photoresist layer on the wafer is exposed to UV light. Mask defines those regions into which diffusion will take place and these regions remain unaffected after exposing to UV light and other region gets hardened.
5. The UV exposed regions are etched away along with the silicon dioxide layer so that the wafer surface is exposed in the window defined by the mask.
6. The remaining photoresist is removed and a thin layer of SiO2 is grown over entire surface and then polysilicon is deposited on top of this to form gate structure.
7. The thin oxide is removed to expose areas into which n-type impurities are diffused to form source and drain.
8. Thick oxide is grown all over again and then masked with photoresist and etched to expose selected area of polysilicon gate and drain and the source areas where connections are to be made.
9. The whole chip is then has metal (Al) deposited over its surface to a thickness of 1μm. This metal layer is then masked and etched to form the required interconnection pattern.

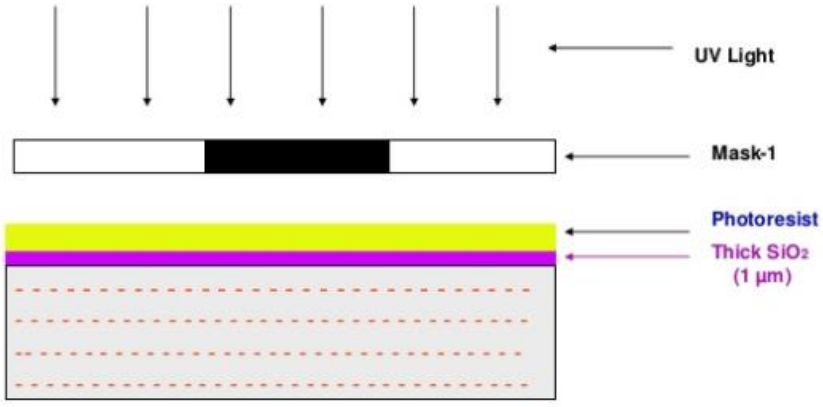
Fig below depicts the nMOS fabrication steps:


(1) Si with p-type impurities

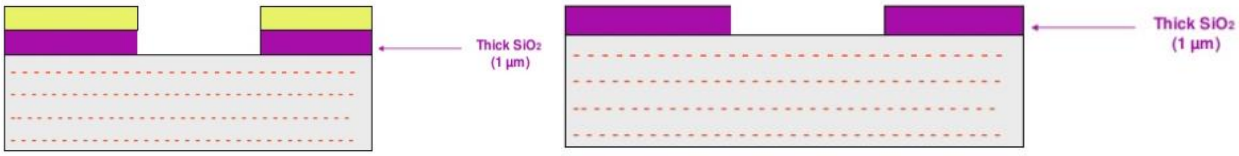



(2) Thin layer of SiO2 on substrate

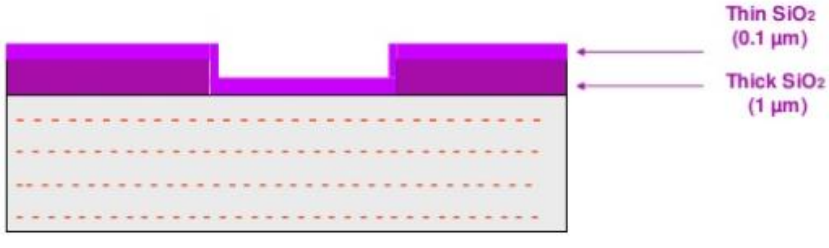

(3) Photoresist on the layer of SiO2



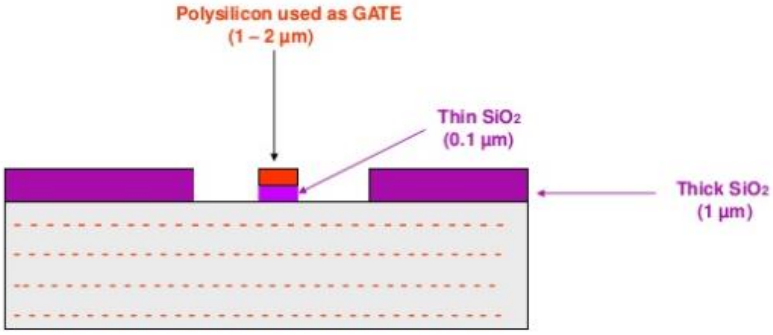

(4) Photoresist layer exposed to UV light through mask




(5) UV exposed regions are etched away

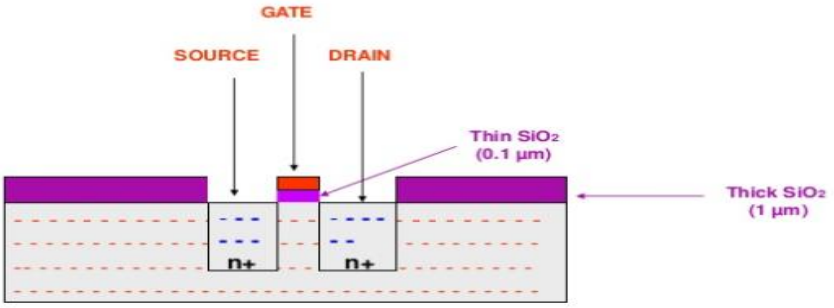



(a)

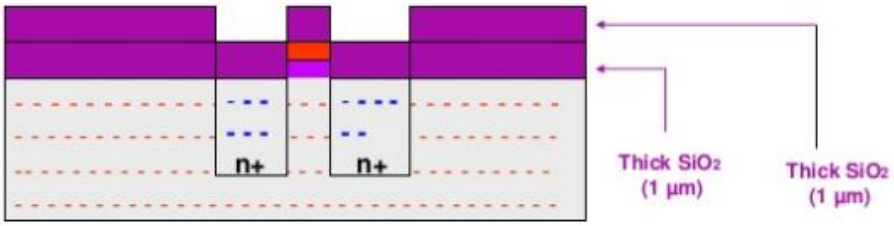



(b)

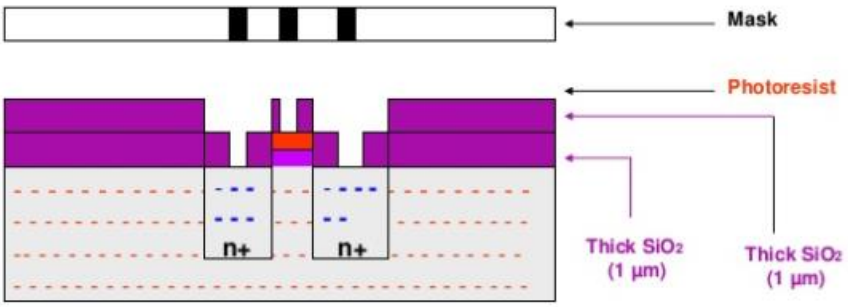
(6 a & b) thin SiO2 layer formation and deposition of polysilicon for gate terminal




(7) n+ diffusion for source and drain formation

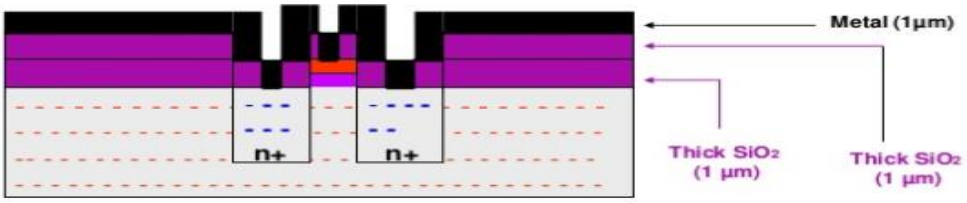



(a)

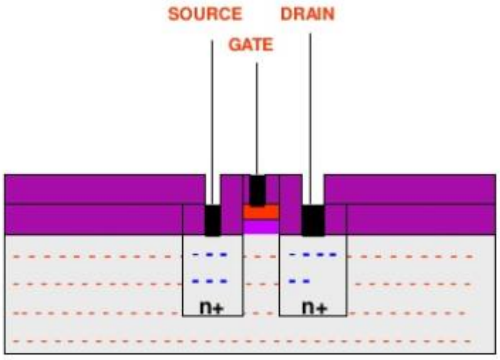



(b)

8(a & b) thick layer of SiO2 grown and masked with photoresist S and D contact cuts




(a)



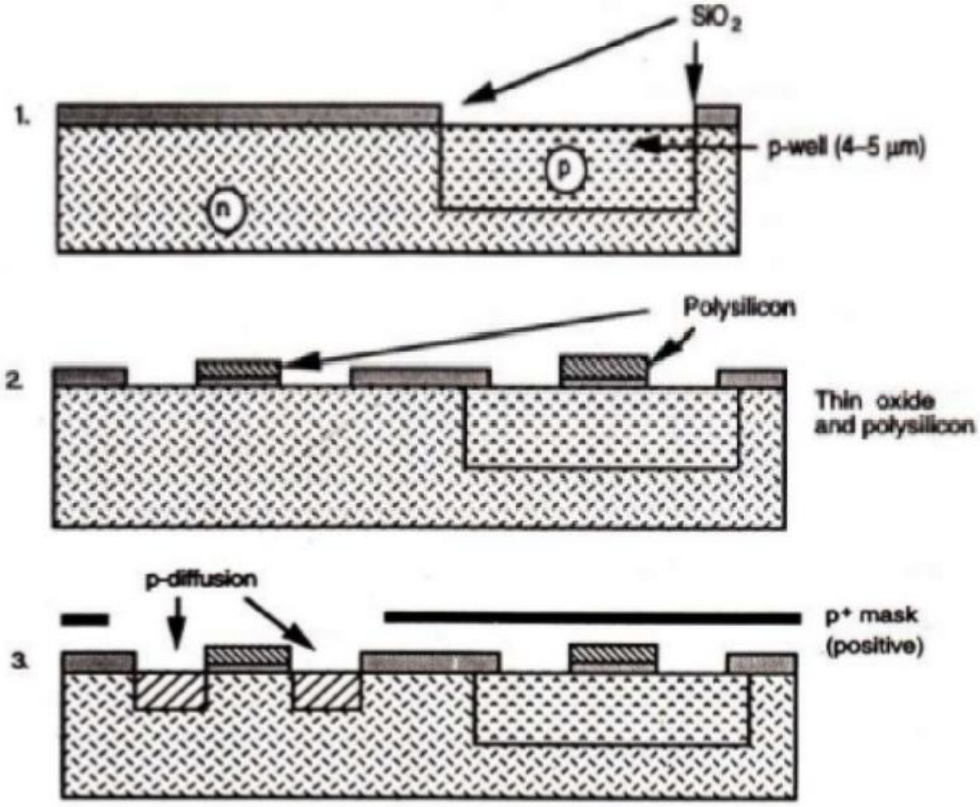

(b)

9(a & b) metal layer deposition and metal layer is masked and etched to form final nMOS transistor

## CMOS Fabrication

- There are a number of methods for CMOS fabrication, which includes p-well, n-well, twin tub and silicon-on-insulator (SOI) processes.
- The p-well process is widely used and the n-well process as it is an retrofit to existing nMOS technology.

**The p-well Process**



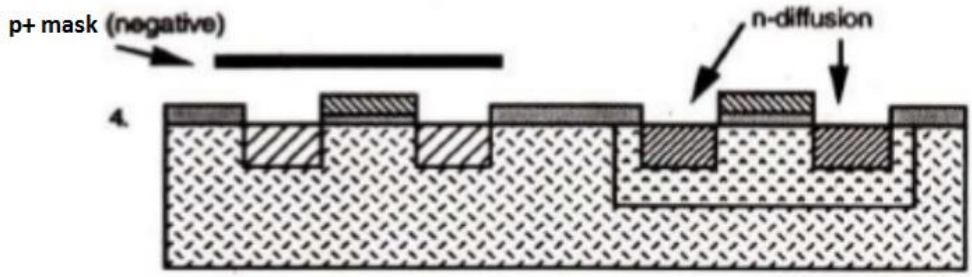



Fig. CMOS p-well process steps

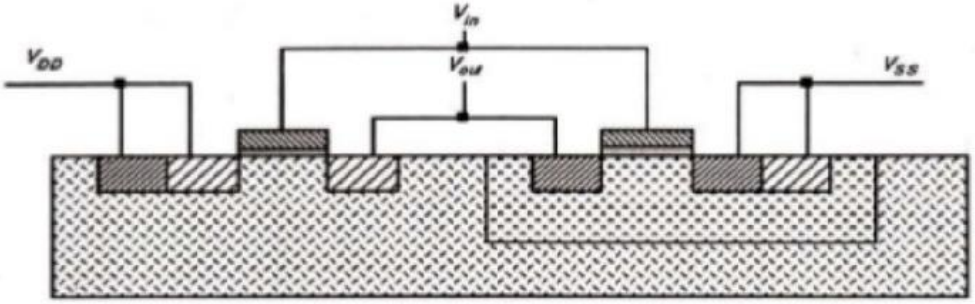



Fig. CMOS p-well inverter showing $V_{DD}$ and $V_{SS}$ substrate connections

- The p-well structure has an n-type substrate in which p-type devices can be formed with the help of masking and diffusion. In order to accommodate n-type devices, deep p-well is diffused into the n-type substrate. This is shown in Fig 1.
  Masking, patterning and diffusion process is same as that of nMOS fabrication. The summary of processing steps are:
  - Mask: defines the areas in which the deep p-well diffusion has to take place.
  - Mask 2: defines the thin oxide region (where the thick oxide is to be removed or stripped and thin oxide grown)
  - Mask 3: patterning the polysilicon layer which is deposited after thin oxide.
  - Mask 4: A p+ mask is used (to be in effect "AND" with mask 2) to define areas where p-diffusion is to take place.
  - Mask 5: –ve form of mask 4 (p+ mask) is used which defines areas where n-diffusion is to take place.
  - Mask 6: Contact cuts are defined using this mask.
  - Mask 7: The metal layer pattern is defined by this mask.
  - Mask 8: An overall passivation (over glass) is now applied and it also defines openings for accessing pads.
- In the process, the diffusion should be carried out with special care as p-well concentration and depth will affect the threshold voltage and also the breakdown voltage of the n-transistor.
- To achieve low threshold voltage either deep-well diffusion or high-well resistivity is required.
- But deep well require larger spacing between n- and p-type transistors and wires due to lateral diffusion and therefore needs larger chip area.
- The p-well acts as substrate for n-devices within parent n-substrate and two areas are electrically isolated

**The n-well Process**

- The p-well processes have been one of the most commonly available forms of CMOS. However, an advantage of the n-well process is that it can be fabricated on the same process line as conventional n MOS.

- n –well CMOS circuits are also superior to p-well because of the lower substrate bias effects on transistor threshold voltage and inherently lower parasitic capacitances associated with source and drain regions.
- Typically n-well fabrication steps are similar to a p-well process, except that an n-well is used which is illustrated in flow diagram
- The first masking step defines the n-well regions.
- The well depth is optimized to ensure against p-substrate to p+ diffusion breakdown without compromising the n-well to n+ mask separation.
- The next steps are to define the devices and diffusion paths, grow field oxide, deposit and pattern the polysilicon, carry out the diffusions, make contact cuts and metallization.
- An n-well mask is used to define n-well regions, as opposed to a p-well mask in a p-well process.
- Fig. Depicts inverted circuit fabricated by n-well process.

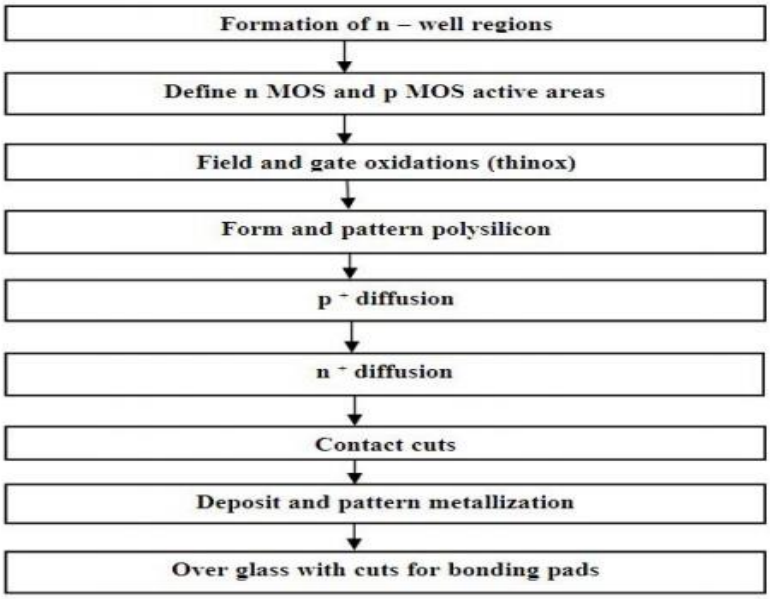



Fig. Main steps in typical n-well process

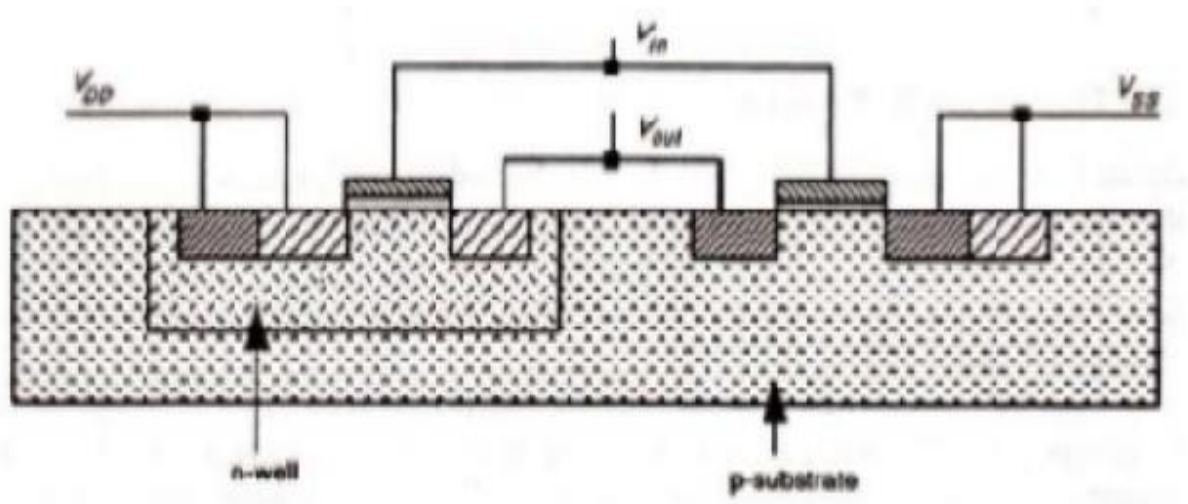
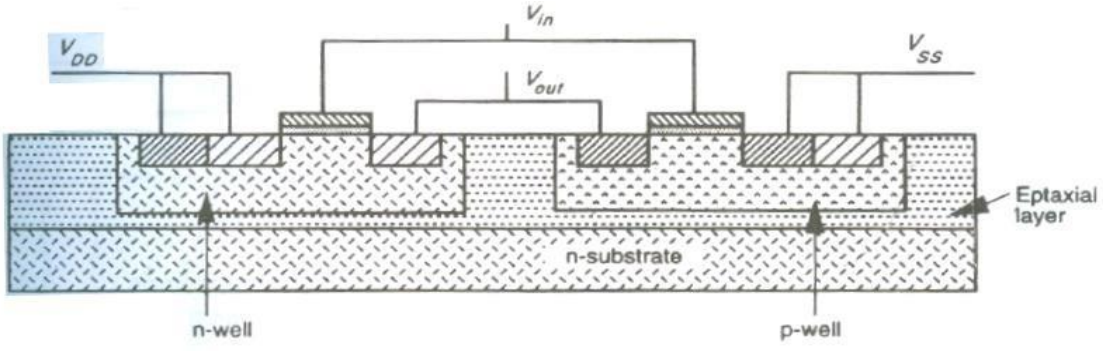



Fig. (A) Cross-sectional view of n-well CMOS Inverter

**The Twin-Tub process**

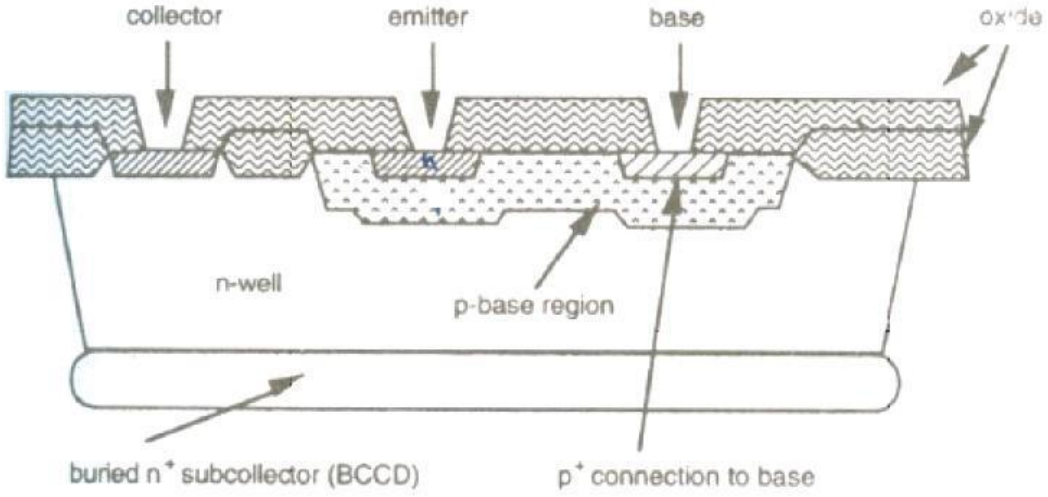
- Twin-tub CMOS technology provides the basis for separate optimization of the p-type and n-type transistors, thus making it possible for threshold voltage, body effect, and the gain associated with n- and p-devices to be independently optimized.
- Generally the starting material is either an n+ or p+ substrate with a lightly doped epitaxial or epi layer, which is used for protection against latch-up.
- The aim of epitaxial is to grow high purity silicon layers of controlled thickness with accurately determined dopant concentrations distributed homogeneously throughout the layer. The electrical properties for this layer are determined by the dopant and its concentration in the silicon.



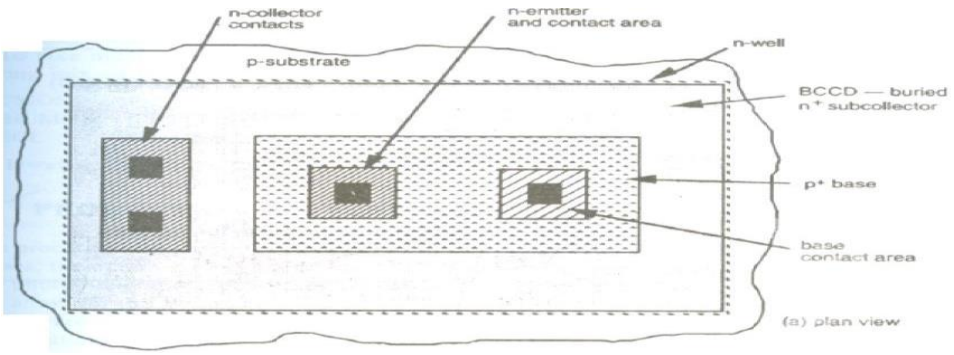

Twin-tub structure.

**BiCMOS Technology**

- The load driving capabilities of MOS transistors is less because of limited current sourcing and sinking abilities of both p and n transistors
- Bipolar transistors provide high gain, better noise and high frequency characteristics than MOS transistors.
- Thus Bipolar can be combined with CMOS technology to build high speed devices called as BiCMOS devices.




Cross section of BiCMOS process




Layout view of BiCMOS process.

| | CMOS technology | BiCMOS technology |
|---|---|---|
| 1. | It has bidirectional capability (source and drain are interchangeable) | Essentially unidirectional |
| 2. | Low static power dissipation | High power dissipation |
| 3. | It has high input impedance | It has Low input impedance |
| 4. | High Packing Density | Low Packing Density |
| 5. | It has Low gain | It has High gain |
| 6. | High delay sensitivity to load | Low delay sensitivity to load |