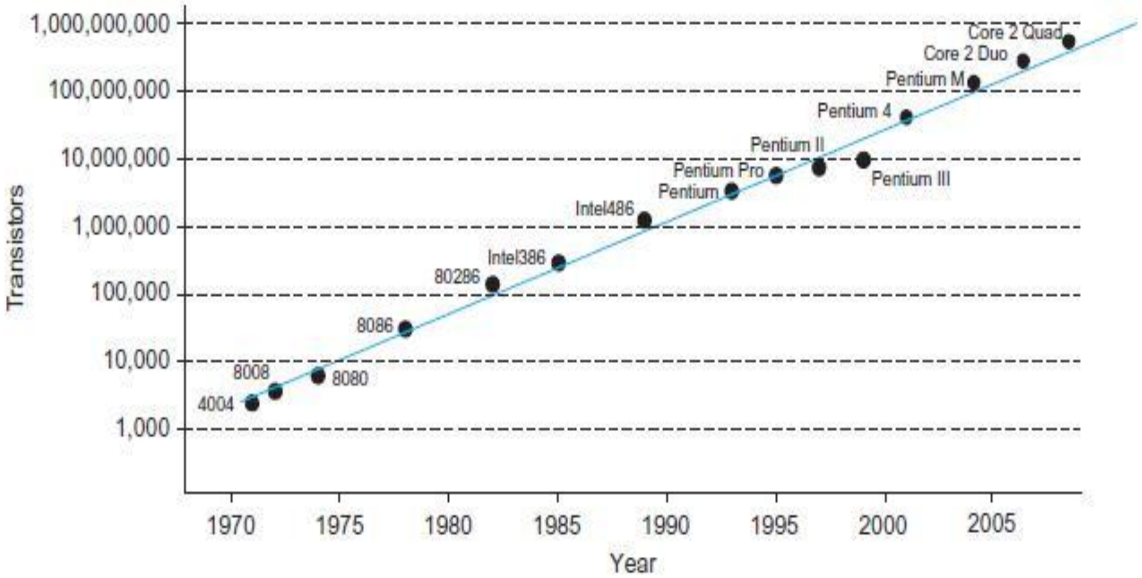
UNIT I - MOS TRANSISTOR PRINCIPLE

NMOS and PMOS transistors, Process parameters for MOS and CMOS, Electrical properties of CMOS circuits and device modeling, Scaling principles and fundamental limits, CMOS inverter scaling, propagation delays, Stick diagram, Layout diagrams.

1.1: INTRODUCTION: (VLSI)

- ✓ In 1958, Jack Kilby built the first integrated circuit flip-flop at Texas Instruments.
- ✓ Bell Labs developed the bipolar junction transistor. Bipolar transistors were more reliable, less noisy and more power-efficient.
- ✓ In 1960s, Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) began to enter in the production.
- ✓ MOSFETs offer the compelling advantage that, they draw almost zero control current while idle.
- ✓ They come in two flavors: nMOS and pMOS, using n-type and p-type silicon respectively.
- ✓ In 1963, Frank Wanlass at Fairchild described the first logic gates using MOSFETs. Fairchild's gates used both nMOS and pMOS transistors, naming as Complementary Metal Oxide Semiconductor (CMOS).
- ✓ 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 replaced nMOS and bipolar processes for 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.



○ ***Moore's Law* is defined as transistor count doubling every 18 months.**

The level of integration of chips is classified as

- Small Scale Integration (SSI)
- Medium Scale Integration (MSI)
- Large Scale Integration (LSI)

- Very Large Scale Integration (VLSI)
- Ultra Large Scale Integration (ULSI)

Small scale Integration:

✓ *Small-Scale Integration* (SSI) circuits have less than 10 gates. Example: 7404 inverter.

Medium scale Integration:

✓ *Medium-Scale Integration* (MSI) circuits have up to 1000 gates. Example: 74161 counter.

Large scale Integration:

✓ *Large-Scale Integration* (LSI) circuits have up to 10,000 gates. Example: 8-bit microprocessor (8085).

✓ Very large scale Integration (VLSI) with gates counting up to lakhs. Example: 16-bit microprocessor (8086).

✓ The feature size of a CMOS manufacturing process refers to the minimum dimension of a transistor that can be reliably built.

Ultra large scale Integration:

✓ Ultra Large-Scale Integration (ULSI) is the process of integrating millions of transistors on a single silicon semiconductor microchip.

****

1.2: nMOS and pMOS transistor

Explain the basic concept of nMOS and pMOS transistor with relevant symbol.

✓ A Metal-Oxide-Semiconductor (MOS) structure is created by superimposing layers of conducting and insulating materials.

✓ CMOS technology provides two types of transistors. They are n-type transistor (nMOS) and p-type transistor (pMOS).

✓ As transistor operation is controlled by electric fields, the devices are also called Metal Oxide Semiconductor Field Effect Transistors (MOSFETs).

✓ The transistor consists of a stack of the conducting gate, an insulating layer of silicon dioxide (SiO2) and the silicon wafer, also called as substrate, body or bulk.

✓ A pMOS transistor consists of p-type source and drain region with an n-type body.

✓ An nMOS transistor consists of n-type source and drain region with a p-type body.

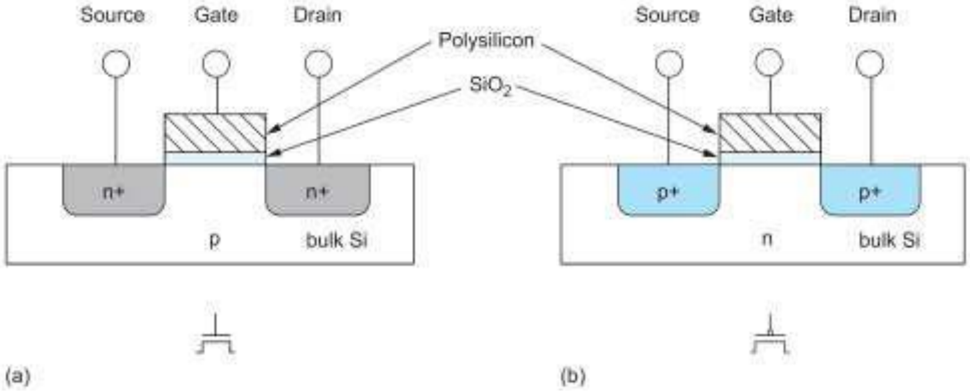


Figure 1: (a) n-MOS transistor (b) p-MOS transistor

nMOS Transistor:

✓ In an nMOS transistor, the body is grounded and the p–n junction of the source and drain to body are reverse-biased.

- ✓ As the gate is grounded, no current flows through junction. Hence, the transistor is OFF.
- ✓ If the gate voltage is raised, it creates an electric field, that start to attract free electrons to the underside of the Si–SiO2 interface.
- ✓ If the voltage is raised more, a thin region under the gate called the channel is inverted.
- ✓ Since a conducting path of electron carriers is formed from source to drain, current starts to flow. So, the transistor is said to be ON.

pMOS Transistor:

- ✓ For a pMOS transistor, the body is held at a positive voltage.
- ✓ When the gate terminal has a positive voltage, the source and drain junctions are reverse-biased and no current flows. So, the transistor is said to OFF.
- ✓ When the gate voltage is lowered, positive charges are attracted to the underside of the Si– SiO2 interface.
- ✓ When a sufficient low gate voltage is applied, the channel inverts and a conducting path of positive carriers is formed from source to drain, which makes the transistor ON.

NOTE:

- ✓ The symbol for the pMOS transistor has a bubble on the gate, indicating that the transistor behavior is opposite to nMOS.
- ✓ When the gate of an nMOS transistor is 1, the transistor is ON. When the gate is 0, the nMOS transistor is OFF.
- ✓ A pMOS transistor is ON when the gate is low(0) and OFF when the gate is high(1).

1.3: Modes of MOS TRANSISTOR

Explain the accumulation mode, depletion layer and inversion layer of MOS transistor with diagram.

- ✓ The MOS transistor is a majority-carrier device, in which the current in a conducting channel is controlled by gate voltage.
- ✓ In an nMOS transistor, the majority carriers are electrons.
- ✓ In a pMOS transistor, the majority carriers are holes.
- ✓ Figure 2 shows a simple MOS structure. The top layer of the structure is a good conductor called the gate.
- ✓ Transistor gate is polysilicon, i.e., silicon formed from many small crystals. The middle layer is a very thin insulating film of SiO_2, called the gate oxide. The bottom layer is the doped silicon body.
- ✓ The figure 2 shows a p-type body, in which the carriers are holes. The body is grounded and voltage is applied to the gate.
- ✓ The gate oxide is a good insulator, so almost zero current flows from the gate to the body.

Accumulation mode:

- ✓ In Figure 2(a), when a negative voltage is applied to the gate, negative charges are formed on the gate.
- ✓ The positively charged holes are attracted to the region under the gate. This is called the accumulation mode.

Depletion mode:

- ✓ In Figure 2(b), when a small positive voltage is applied to the gate, positive charges are formed on the gate.
- ✓ The holes in the body are repelled from the region directly under the gate, resulting in a depletion region forming below the gate.

Inversion layer:

- ✓ In Figure 2(c), when a higher positive potential greater than threshold voltage (V_t) is applied, more positive charges are attracted to the gate.
- ✓ The holes are repelled and some free electrons in the body are attracted to the region under the gate. This conductive layer of electrons in the p-type body is called the inversion layer.
- ✓ The threshold voltage depends on the number of dopants in the body and the thickness t_{ox} of the oxide.

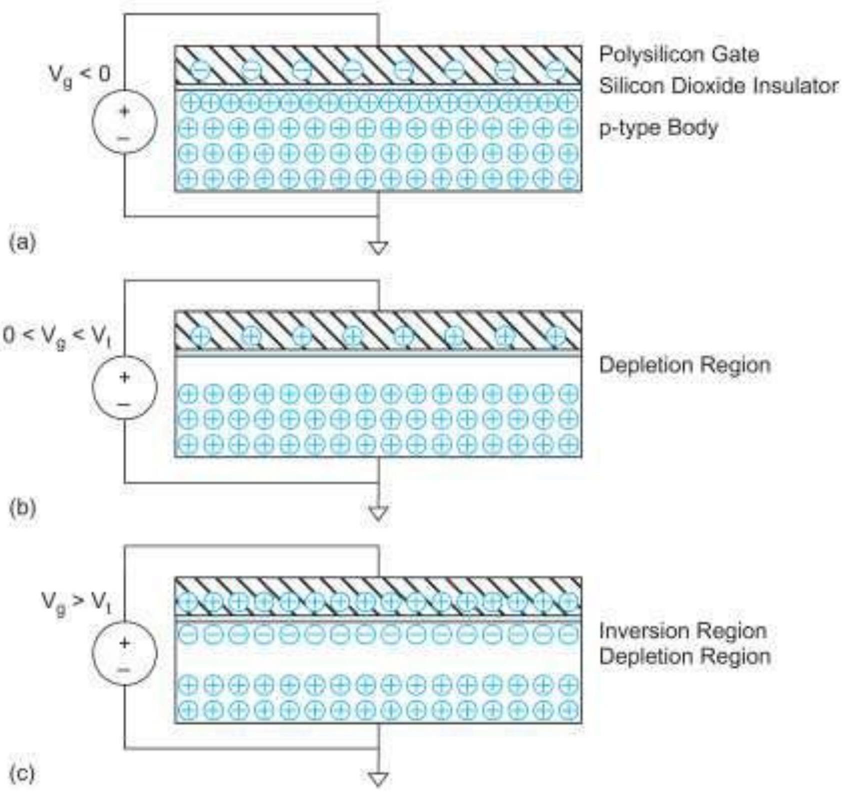


Figure 2: MOS structure demonstrating (a) accumulation, (b) depletion, and (c) inversion layer

1.4: Operating regions of MOS transistor:

Explain the three different types of modes of operation of nMOS transistor. (or)
Discuss the cutoff, linear and saturation region operation of MOS transistor. (Nov 2009)

- ✓ The MOS transistor operates in cutoff region, linear region and saturation region.

Cutoff region:

- In Figure 3(a), the gate-to-source voltage (V_{gs}) is less than the threshold voltage (V_t) and source is grounded.
- Junctions between the body and the source or drain are reverse biased, so no current flows. Thus, the transistor is said to be OFF and this mode of operation is called cutoff.
- If $V_{gs} < V_t$, the transistor is cutoff (OFF).

Linear Region:

- In Figure 3(b), the gate voltage is greater than the threshold voltage.
- An inversion region of electrons, called the channel connects the source and drain, creating a conductive path and making the transistor ON.
- If $V_{gs} > V_t$, the transistor turns ON. If V_{ds} is small, the transistor acts as a linear resistor, in which the current flow is proportional to V_{ds}.
- The number of carriers and the conductivity increases, with the gate voltage.

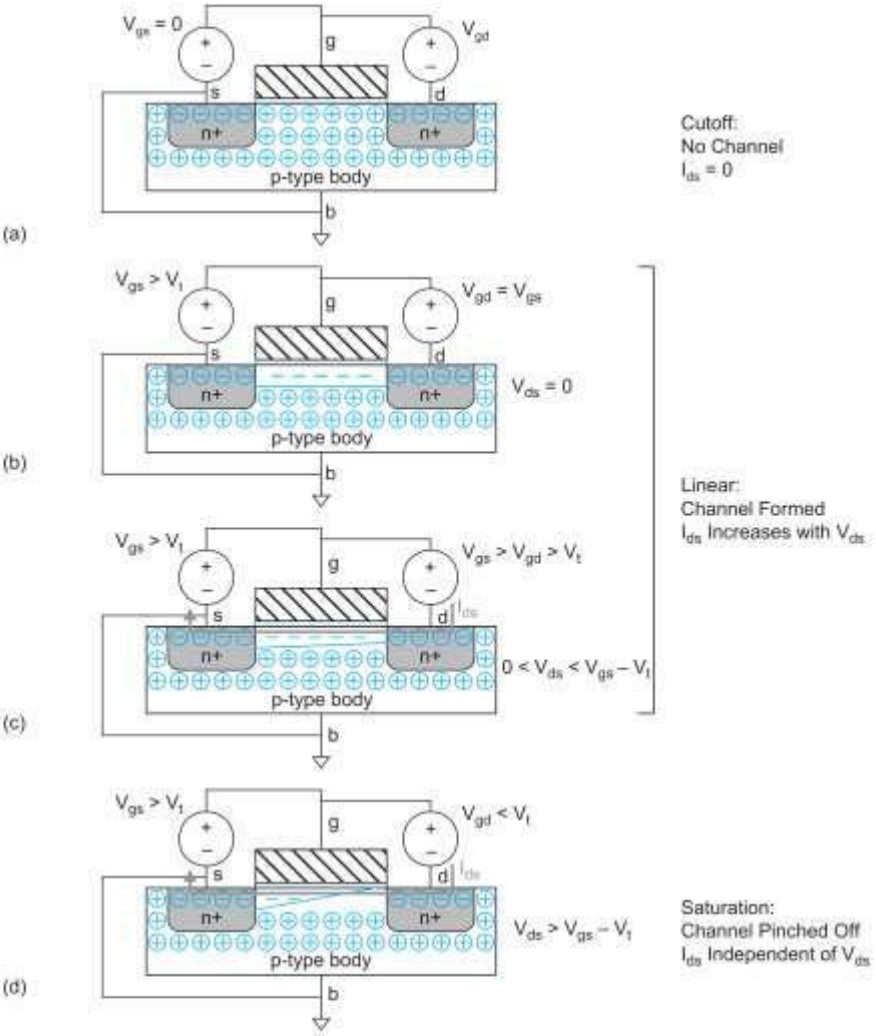


Figure 3: nMOS transistor demonstrating cutoff, linear, and saturation regions of operation

- The voltage between drain and source is $V_{ds} = V_{gs} - V_{gd}$. If $V_{ds} = 0$ (i.e., $V_{gs} = V_{gd}$), there is no electric field to push current from drain to source.
- When a small positive voltage V_{ds} is applied to the drain (Figure 3(c)), current I_{ds} flows through the channel from drain to source.
- This mode of operation is termed as linear, resistive, triode, nonsaturated, or unsaturated.

Saturation region:

- The current increases with increase in both the drain voltage and gate voltage.

- ✓ If V_{ds} becomes sufficiently large that $V_{gd} < V_t$, the channel is no longer inverted near the drain and becomes pinched off (Figure 3(d)).
- ✓ As electrons reach the end of the channel, they are injected into the depletion region near the drain and accelerated toward the drain.
- ✓ Above this drain voltage, current I_{ds} are controlled only by the gate voltage. This mode is called saturation.
- ✓ If $V_{gs} > V_t$ and V_{ds} is large, the transistor acts as a current source, in which the current flow becomes independent of V_{ds}.

Explain the three different types of modes of operation of pMOS transistor.

- ✓ The pMOS transistor in Figure 4 operates in just the opposite fashion. The n-type body is tied to high potential, junctions of p-type source and drains are normally reverse-biased.
- ✓ When the gate has high potential, no current flows between drain and source.
- ✓ When the gate voltage is lowered by a threshold V_t, holes are attracted to form a p-type channel beneath the gate, allowing current to flow between drain and source.

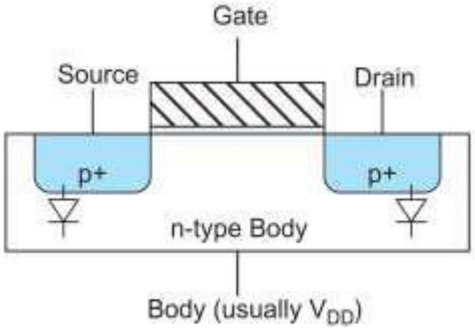


Figure 4: pMOS transistor

****

1.5: IDEAL I-V CHARACTERISTICS OF MOS TRANSISTOR

- ❖ **Explain in detail about the ideal I-V characteristics of a NMOS and PMOS device. (MAY 2013)**
- ❖ **Discuss in detail with necessary equations the operation of MOSFET and its current-voltage characteristics. (April/May 2011, May 2016).**
- ❖ **Derive drain current of MOS device in different operating regions. (Nov/Dec 2014)(May/June 2013) (Nov 2012, Nov 2016)**
- ❖ **Explain in detail about the ideal I-V characteristics and non-ideal I-V characteristics of a NMOS and PMOS device. (May/June 2013)**
- ❖ **Derive expressions for the drain-to-source current in the nonsaturated and saturated regions of operation of an nMOS transistor. (Nov 2007, Nov 2008)**

- ✓ MOS transistor has three regions of operation:
  - Cutoff or subthreshold region
  - Linear region
  - Saturation region
- ✓ The current through an OFF transistor is zero. When a transistor turns ON ($V_{gs} > V_t$), the gate attract electrons to form a channel.
- ✓ Current is measured from the amount of charge in the channel.
- ✓ The charge on each plate of a capacitor is $Q = CV$. Thus, the charge in the channel $Q_{channel}$ is

$$Q_{channel} = C_g (V_{gc} - V_t)$$

where C_g : Capacitance of the gate to the channel

$V_{gc} - V_t$: Amount of voltage attracting charge to the channel.

✓ If the source is at V_s and the drain is at V_d ,

✓ Average channel voltage is $V_c = (V_s + V_d)/2 = V_s + V_{ds}/2$.

✓ Gate and channel voltage V_{gc} is $V_g – V_c = V_{gs} – V_{ds}/2$,

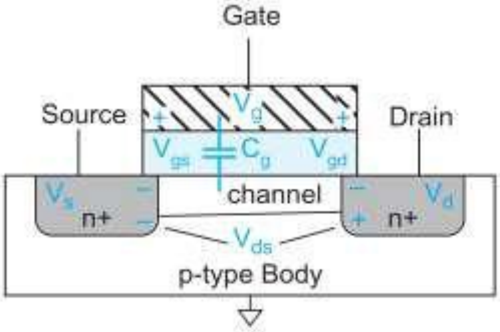


Figure 5: Average gate to channel voltage

✓ If the gate has length L and width W and the oxide thickness is t_{ox}, as shown in Figure 6, Then the capacitance C_g is

$$C_g = \kappa_{ox}\varepsilon_o \frac{WL}{t_{ox}} = \varepsilon_{ox}\frac{WL}{t_{ox}} = c_{ox}WL \quad \text{----------(1)}$$

Where, ε_o is the permittivity of free space, 8.85×10^{-14} F/cm,

Permittivity of SiO2 is $k_{ox} = 3.9$ times as great.

✓ The ε_{ox}/t_{ox} term is called as C_{ox}. Capacitance (C_{ox}) is a per unit area of the gate oxide.

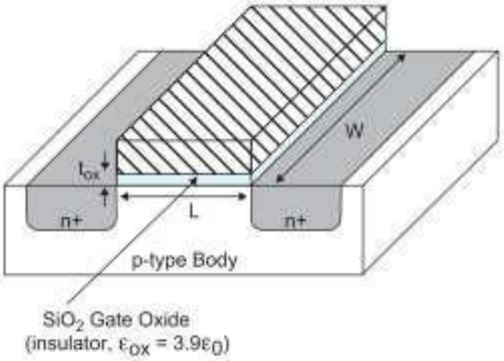


Figure 6: Transistor dimensions

✓ Average velocity (v) of carrier is proportional to the lateral electric field (field between source and drain). The constant of proportionality μ is called the mobility.

$$v = \mu E \text{ --------------(2)}$$

✓ The electric field E is the voltage difference between drain and source (V_{ds}) divided by the channel length (L).

$$E = \frac{V_{ds}}{L} \text{ -------------(3)}$$

✓ The time required for carriers to cross the channel is L divided by v.

✓ The current between source and drain is the total amount of charge in the channel divided by the time required to cross.

$$
\begin{aligned}
I_{ds} &= \frac{Q_{\text{channel}}}{L/v} \\
&= \mu C_{\text{ox}} \frac{W}{L}\left(V_{gs} - V_t - V_{ds}/2\right)V_{ds} \\
&= \beta\left(V_{GT} - V_{ds}/2\right)V_{ds}
\end{aligned}
$$

where

$$\beta = \mu C_{\text{ox}} \frac{W}{L};\ V_{GT} = V_{gs} - V_t \qquad \text{-----------}\ (4)$$

✓ Equation (4) is called linear or resistive, because when $V_{ds} << V_{GT}$, I_{ds} increases linearly with V_{ds}, like an ideal resistor.

✓ k' is the k prime, $k' = \mu C_{ox}$.

✓ If $V_{ds} > V_{dsat} = V_{GT}$, the channel is no longer inverted in the drain region. Channel is pinched off.

✓ Beyond this point (called the drain saturation voltage), increasing the drain voltage has no further effect on current.

✓ Substituting $V_{ds} = V_{dsat}$ in Eq (4), we can find an expression for the saturation current (I_{ds}) that is independent of V_{ds}.

$$I_{ds} = \frac{\beta}{2} V_{GT}^2 \text{--------------------} \qquad (5)$$

✓ This expression is valid for $V_{gs} > V_t$ and $V_{ds} > V_{dsat}$.

✓ Summarizes the current in the three regions:

$$
I_{ds} = \begin{cases}
0 & V_{gs} < V_t \quad \text{Cutoff} \\
\beta\left(V_{GT} - V_{ds}/2\right)V_{ds} & V_{ds} < V_{dsat} \quad \text{Linear} \\
\frac{\beta}{2} V_{GT}^2 & V_{ds} > V_{dsat} \quad \text{Saturation}
\end{cases}
$$

****

1.6: C – V CHARACTERISTICS OF MOS TRANSISTOR (AC characteristics)

❖ **Discuss the CV characteristics of the CMOS. (Nov 2012, May 2014, Nov 2015, Nov 2016)**

❖ **Explain the electrical properties CMOS. (Nov 2017)**

✓ Each terminal of an MOS transistor has capacitance to the other terminals. Capacitances are nonlinear and voltage dependent (C-V).

SIMPLE MOS CAPACITANCES MODEL:

✓ The gate of an MOS transistor is a good capacitor. Its capacitance is necessary to attract charge to invert the channel, so high gate capacitance is required to obtain high I_{ds}.

✓ The gate capacitor can be viewed as a parallel plate capacitor with the gate on top, channel on bottom and the thin oxide dielectric between.

✓ The capacitance is $C_g = C_{ox} WL$. ---------------(1)

$$C_g = C_{permicron}\ W \text{ ---------------}(2)$$

$$\text{Where } C_{permicron} = C_{ox}L = \frac{\varepsilon_{ox}}{t_{ox}}L$$

- ✓ In addition to the gate, the source and drain also have capacitances. These capacitances are called parasitic capacitors.
- ✓ The source and drain capacitances arise from the p–n junctions between the source or drain diffusion and the body. These capacitances are called diffusion capacitance C_{sb} and C_{db}.
- ✓ The depletion region acts as an insulator between the conducting p- and n-type regions, creating capacitance across the junction.
- ✓ The capacitance of junctions depends on the area and perimeter of the source and drain diffusion, the depth of the diffusion, the doping levels and the voltage.
- ✓ As diffusion has both high capacitance and high resistance, it is generally made as small as possible in the layout.
- ✓ Three types of diffusion regions are illustrated by the two series transistors in Figure 7.
- ✓ In Figure 7(a), each source and drain has its own isolated region of contacted diffusion.
- ✓ In Figure 7(b), the drain of the bottom transistor and source of the top transistor form a shared contacted diffusion region.
- ✓ In Figure 7(c), the source and drain are merged into an uncontacted region.

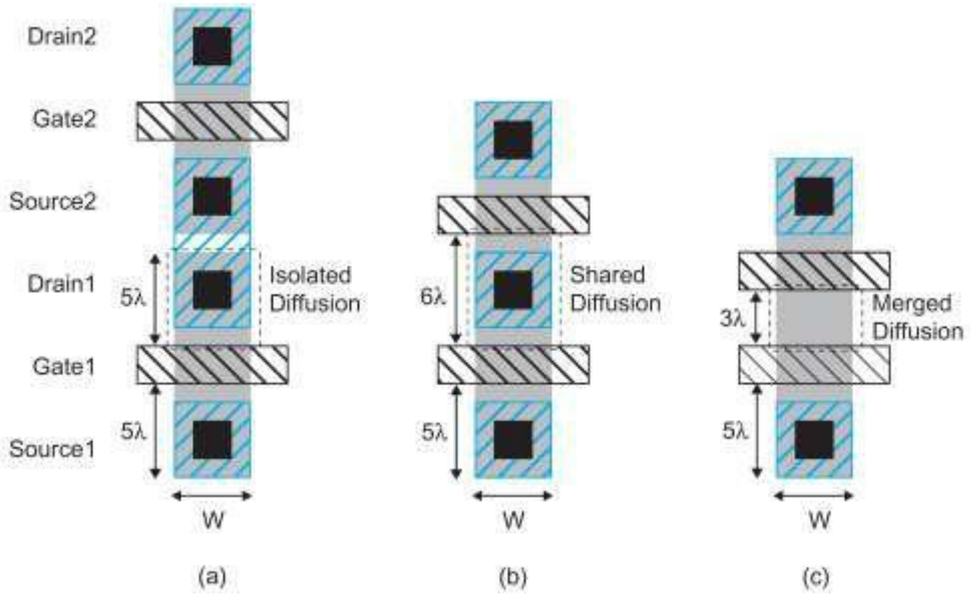


Figure7: Diffusion region geometries

DETAILED MOS GATE CAPACITANCE MODEL:

- ✓ MOS gate places above the channel and may partially overlap the source and drain diffusion areas.
- ✓ The gate capacitance has two components, (i) the intrinsic capacitance C_{gc} (over the channel) and (ii) the overlap capacitances C_{gol} (to the source and drain).
- ✓ The intrinsic capacitance was approximated as a simple parallel plate with capacitance
- ✓ $C_0 = WLC_{ox}$.
- ✓ The intrinsic capacitance has three components representing the different terminals connected to the bottom plate are C_{gb} (gate-to-body), C_{gs} (gate-to-source), and C_{gd} (gate-to-drain).
- ✓ The behavior in three regions (Cutoff, Linear and Saturation) can be approximated as shown in Table 1.

Parameter	Cutoff	Linear	Saturation
C_{gb}	$\leq C_0$	0	0
C_{gs}	0	$C_0/2$	$2/3\ C_0$
C_{gd}	0	$C_0/2$	0
$C_g = C_{gs} + C_{gd} + C_{gb}$	C_0	C_0	$2/3\ C_0$

Table1: Approximation for intrinsic MOS gate capacitance

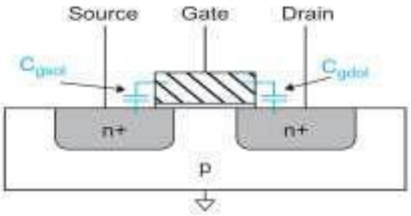


Figure 8: Overlap capacitances

DETAILED MOS DIFFUSION CAPACITANCE MODEL:

✓ The capacitance depends on both the area AS and sidewall perimeter PS of the source diffusion region. The area is AS = WD.

✓ The perimeter is PS = 2W +2D.

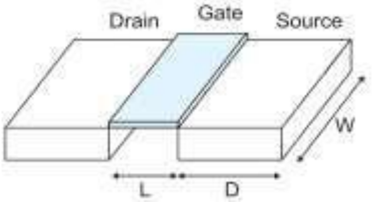


Figure 9: Diffusion region geometry

✓ The total source parasitic capacitance is $C_{sb} = AS * C_{jbs} + PS * C_{jbssw}$

Where, C_{jbs} - Capacitance of the junction between the body and the bottom of the source

C_{jbssw} - Capacitance of the junction between the body and the side walls of the source

✓ In summary, MOS transistor can be viewed as a four-terminal device with capacitances between each terminal pair, as shown in Figure 10.

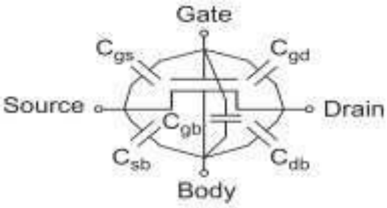


Figure 10: Capacitance of a MOS Transistor

✓ The gate capacitance includes an intrinsic component and overlap terms with the source and drain. The source and drain have parasitic diffusion capacitance to the body.

****

1.7: DC TRANSFER CHARACTERISTICS

- **Explain the DC transfer characteristic of CMOS inverter.[APRIL-2015, Nov 2015]**
- **Draw and explain the DC and transfer characteristics of a CMOS inverter with necessary conditions for the different regions of operation. (Nov/Dec 2011) (Nov/Dec 2012) (May/June 2013) (April/May 2012) (May/June 2014) (Nov/Dec 2013) (May 2016, May 2017, Nov 2008)**
- **Explain the CMOS inverter DC characteristics. (Nov 2007, Nov 2009)**

✓ The DC transfer characteristics of a circuit relate the output voltage to the input voltage.

(i) Static CMOS inverter DC Characteristics:

✓ The DC transfer function (V_{out} Vs. V_{in}) for the static CMOS inverter shown in Figure 11.

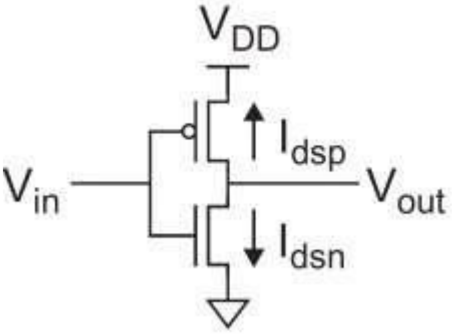


Figure 11: A Static CMOS inverter

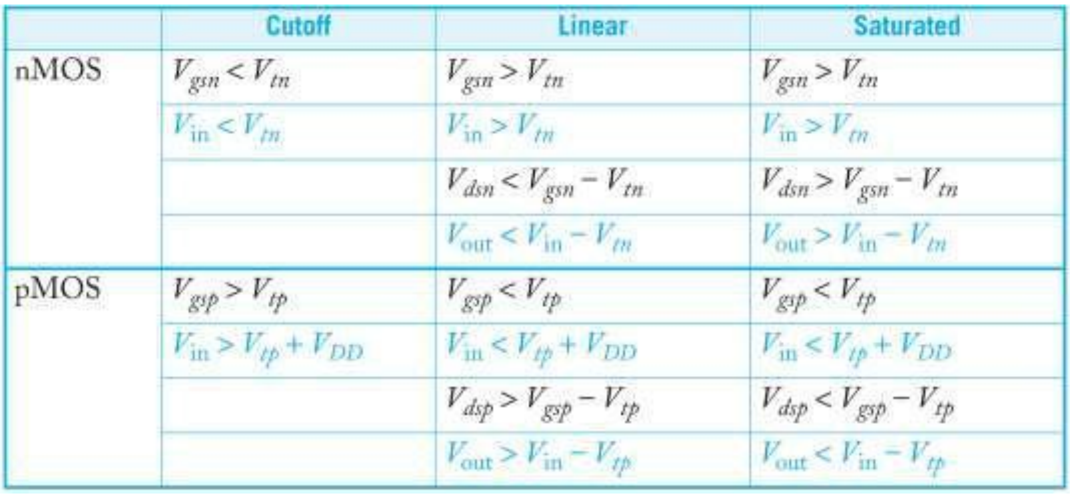
✓ Table 2, shows various regions of operation for the n and p transistors.

✓ In this table, V_{tn} is the threshold voltage of the n-channel device, and V_{tp} is the threshold voltage of the p-channel device. V_{tp} is negative.

✓ The equations are given both in terms of V_{gs} /V_{ds} and V_{in} /V_{out}.

✓ As the source of the nMOS transistor is grounded, $V_{gsn} = V_{in}$ and $V_{dsn} = V_{out}$.

✓ As the source of the pMOS transistor is tied to V_{DD}, $V_{gsp} = V_{in} - V_{DD}$ and $V_{dsp} = V_{out} - V_{DD}$.



	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	$V_{gsp} > V_{tp}$	$V_{gsp} < V_{tp}$	$V_{gsp} < V_{tp}$
	$V_{in} > V_{tp} + V_{DD}$	$V_{in} < V_{tp} + V_{DD}$	$V_{in} < V_{tp} + V_{DD}$
		$V_{dsp} > V_{gsp} - V_{tp}$	$V_{dsp} < V_{gsp} - V_{tp}$
		$V_{out} > V_{in} - V_{tp}$	$V_{out} < V_{in} - V_{tp}$

Table 2: Relationships between voltages for the three regions of operation of a CMOS inverter

✓ Figure 12(a), shows I_{dsn} and I_{dsp} in terms of V_{dsn} and V_{dsp} for various values of V_{gsn} and V_{gsp}.

✓ Figure 12(b), shows the same plot of I_{dsn} and $|I_{dsp}|$ in terms of V_{out} for various values of V_{in}.

✓ Operating points are plotted on V_{out} vs. V_{in} axes in Figure 12(c) to show the inverter DC transfer characteristics.

✓ The supply current $I_{DD} = I_{dsn} = |I_{dsp}|$ is plotted against V_{in} in Figure 13(d) showing that both transistors are momentarily ON as V_{in}.

✓ The operation of the CMOS inverter can be divided into five regions as indicated on figure 12(c).

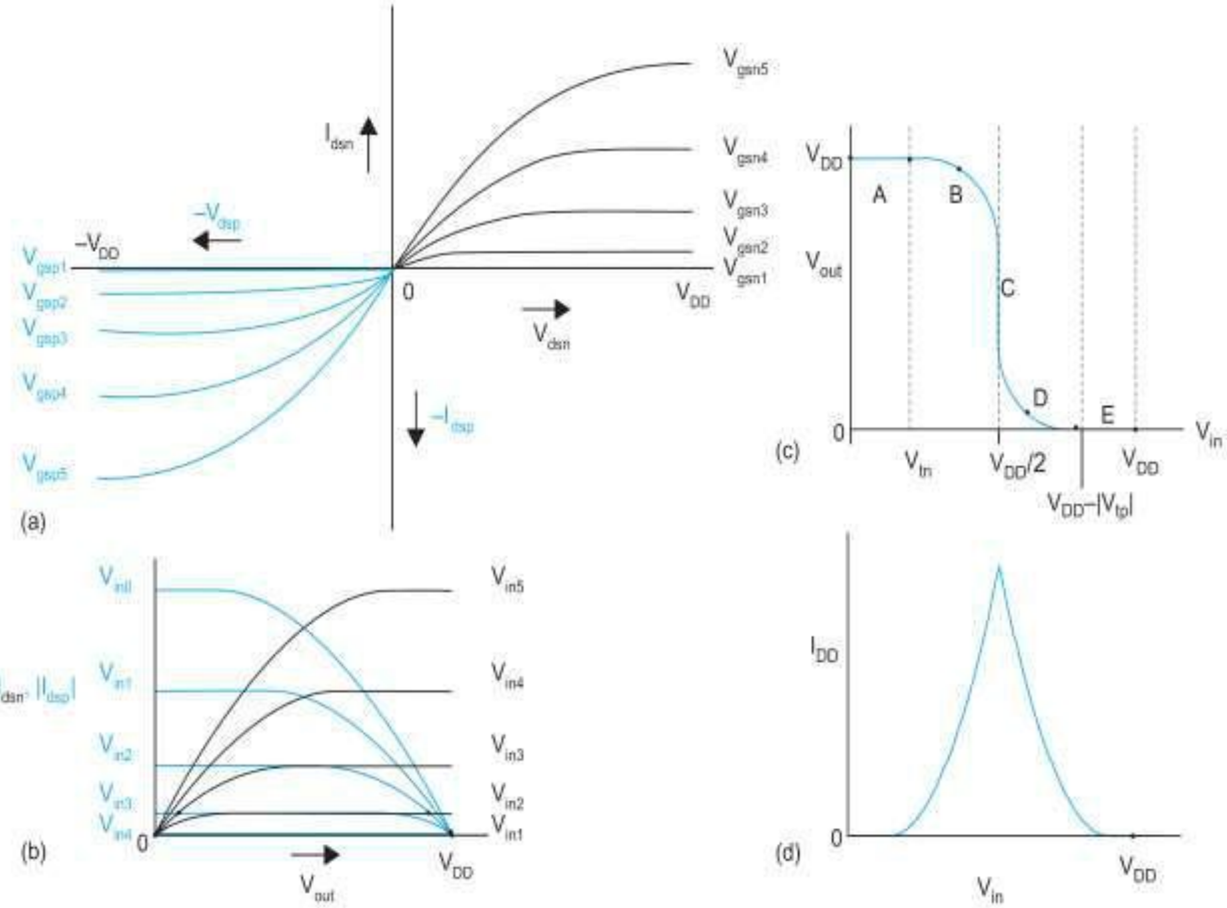
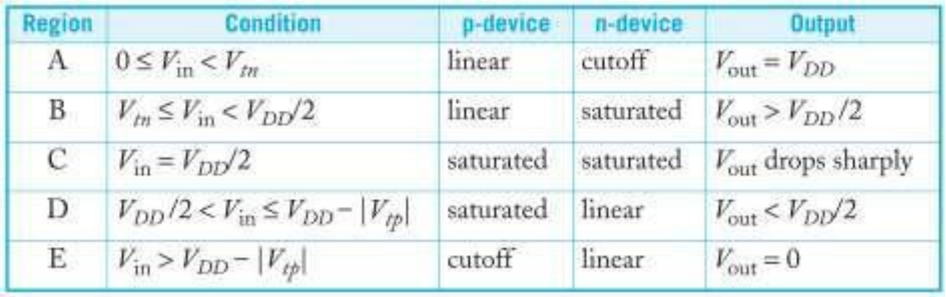


Figure 12: CMOS inverter DC characteristic

✓ The state of each transistor in each region is shown in Table 3.



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 3: Summary of CMOS inverter operation.

✓ In region *A*, the nMOS transistor is OFF and the pMOS transistor pulls the output to V_{DD}.

✓ In region *B*, the nMOS transistor starts to turn ON. It is 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, making the nMOS transistor to pull the output down to GND.

(ii) Beta ratio Effects:

✓ For $\beta_p = \beta_n$, the inverter threshold voltage V_{inv} is $V_{DD}/2$.

✓ It allows a capacitive load to charge and discharge in equal times by providing equal current source and equal sink capabilities.

✓ Inverter with different beta ratios $r = \beta_p / \beta_n$ is called skewed inverter.

✓ If r > 1, the inverter is HI-skewed. If r < 1, the inverter is LO-skewed. If r = 1, the inverter has normal skew or is unskewed.

✓ Figure 13, shows the impact of skewing the beta ratio on the DC transfer characteristics.

✓ As the beta ratio is changed, the switching threshold is varied.

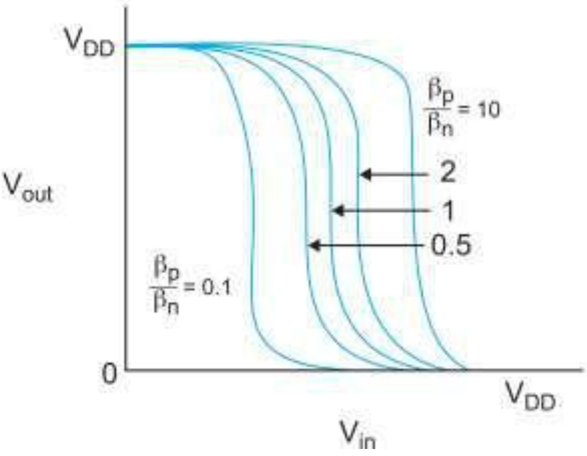


Figure 13: Transfer characteristics of skewed inverters

❖ **Derive the noise margins for a CMOS inverter. (May 2010, Nov 2016)**

(iii) Noise Margins:

✓ Noise margin (Noise immunity) is related to the DC voltage characteristics.

✓ Noise Margin allows determining the allowable noise voltage on the input of a gate, so that the output will not be corrupted.

✓ Two parameters of the noise margin are LOW noise margin (NM_L), and the HIGH noise margin (NM_H).

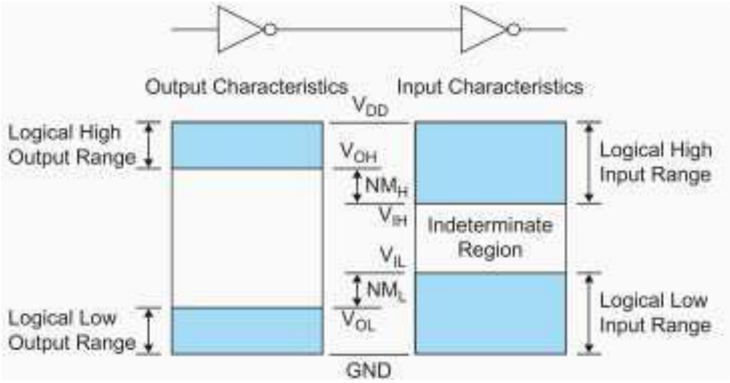


Figure 14: Noise Margin Definitions

✓ NM_L is defined as the difference in maximum LOW input voltage V_{IL} and the maximum LOW output voltage V_{OL}. $NM_L = V_{IL} - V_{OL}$

✓ The value of NM_H is the difference between the minimum HIGH output voltage V_{OH} and the minimum HIGH input voltage V_{IH}. i.e., $NM_H = V_{OH} - V_{IH}$

✓ Inputs between V_{IL} and V_{IH} are said to be in the indeterminate region or forbidden zone.

(iv) Pass Transistor DC Characteristics:

✓ The nMOS transistors pass 0's well but 1's poorly. Figure 15(a), shows an nMOS transistor with the gate and drain tied to V_{DD}.

✓ Initially at $V_s = 0$. $V_{gs} > V_{tn}$, so the transistor is ON and current flow.

✓ Therefore, nMOS transistors attempting to pass a 1 never pull the source above $V_{DD} - V_{tn}$. This loss is called a threshold drop.

✓ The pMOS transistors pass 1's well but 0's poorly.

✓ If the pMOS source drops below $|V_{tp}|$, the transistor cuts off.

✓ Hence, pMOS transistors only pull down to a threshold above GND, as shown in Figure 15(b).

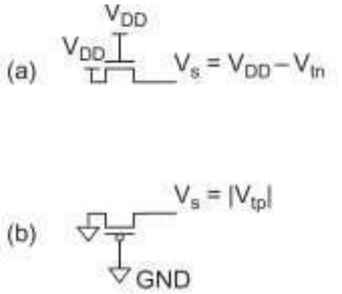


Figure 15: Pass Transistor threshold drops

****

1.8: NON IDEAL I-V EFFECTS

❖ **Explain in detail about the non ideal I-V characteristics of a CMOS device. (MAY 2013)**

❖ **Explain channel length modulation and body effect. (Nov 2009, May 2013)**

✓ MOS characteristics degrade with temperature. It is useful to have a qualitative understanding of non ideal effects to predict their impact on circuit behavior.

(i) Mobility Degradation and Velocity Saturation:

✓ Current is proportional to the lateral electric field $E_{lat} = V_{ds}/L$ between source and drain.

✓ A high voltage at the gate of the transistor attracts the carriers to the edge of the channel, causing carriers collision with the oxide interface that slows the carriers. This is called mobility degradation.

✓ Carriers approach a maximum velocity (v_{sat}) when high fields are applied. This phenomenon is called velocity saturation.

(ii) Channel Length Modulation:

✓ Current I_{ds} is an independent of V_{ds} for a transistor in saturation.

✓ The p–n junction between the drain and body forms a depletion region with a width L_d that increases with V_{db}, as shown in Figure 16.

✓ The depletion region effectively shortens the channel length to $L_{eff} = L - L_d$

✓ To avoid the body voltage into calculations, assume the source voltage is close to the body voltage i.e $V_{db} = V_{ds}$.

✓ Hence, increasing V_{ds} decreases the effective channel length.

✓ Shorter channel length results in higher current. Thus, I_{ds} increases with V_{ds} in saturation, as shown in Figure 16.

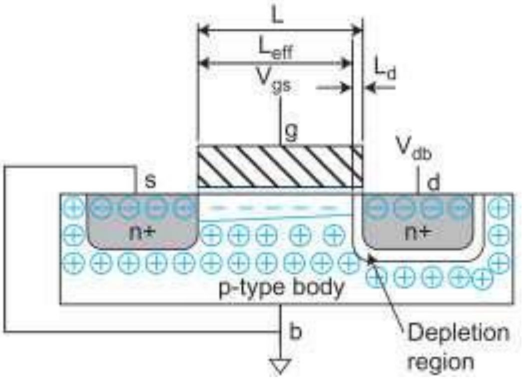


Figure 16: Depletion region shortens effective channel length

✓ In Saturation region, I_{ds} is

$$I_{ds} = \frac{\beta}{2} V_{GT}^2 \left(1 + \frac{V_{ds}}{V_A}\right)$$

✓ Hence, V_A is proportional to channel length. This channel length modulation model is a gross oversimplification of nonlinear behavior.

(iii) Threshold Voltage (V_t) Effects

❖ **Explain in detail about effect and its effect in MOS device. (May 2016)**

✓ Threshold voltage V_t increases with the source voltage, decreases with the body voltage, decreases with the drain voltage and increases with channel length.

Body Effect:

✓ When a voltage V_{sb} is applied between the source and body, it increases the amount of charge required to invert the channel. Hence, it increases the threshold voltage.

✓ The threshold voltage can be modeled as

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

where V_{t0} is the threshold voltage when the source is at the body potential, ϕ_s is the surface potential at threshold and γ is the body effect coefficient.

(iv) Leakage:

✓ Even when transistors are OFF, transistors leak small amounts of current.

✓ Leakage mechanisms include subthreshold conduction between source and drain, gate leakage from the gate to body and junction leakage from source to body and drain to body.

✓ Subthreshold conduction is caused by thermal emission of carriers over the potential barrier set by the threshold.

✓ Gate leakage is a quantum-mechanical effect caused by tunneling through the extremely thin gate dielectric.

✓ Junction leakage is caused by current through the p-n junction between the source/drain diffusions and the body.

****

1.9: Device models:

Explain the following: Device models and device characteristics. (MAY 2014)

✓ SPICE (Simulation Program with Integrated Circuit Emphasis) provides a wide variety of MOS transistor models with various trade-offs between complexity and accuracy.

✓ Level 1 and Level 3 models were important, but they are no longer adequate to accurately model very small modern transistors.

✓ BSIM models are more accurate and are presently the most widely used.

i.Level 1 model:

✓ The SPICE Level 1, or Shichman-Hodges Model is closely related to the Shockley model, enhanced with channel length modulation and the body effect.

✓ The basic current model is:

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t & \text{cutoff} \\ \text{KP}\dfrac{W_{\text{eff}}}{L_{\text{eff}}}\left(1+\text{LAMBDA}\times V_{ds}\right)\left(V_{gs}-V_t-\dfrac{V_{ds}}{2}\right)V_{ds} & V_{ds} < V_{gs}-V_t & \text{linear} \\ \dfrac{\text{KP}}{2}\dfrac{W_{\text{eff}}}{L_{\text{eff}}}\left(1+\text{LAMBDA}\times V_{ds}\right)\left(V_{gs}-V_t\right)^2 & V_{ds} > V_{gs}-V_t & \text{saturation} \end{cases}$$

✓ The parameters from the SPICE model are given in ALL CAPS.

✓ β is written instead as KP (W_{eff} /L_{eff}), where KP is a model parameter. W_{eff} and L_{eff} are the effective width and length.

✓ The LAMBDA term (LAMBDA =1/V_A) models channel length modulation.

✓ The threshold voltage is modulated by the source-to-body voltage V_{sb} through the body effect.

✓ For non negative V_{sb} , the threshold voltage, V_t is

$$V_t = \text{VTO} + \text{GAMMA}\left(\sqrt{\text{PHI}+V_{sb}} - \sqrt{\text{PHI}}\right)$$

Where, VTO is the “zero-bias” threshold voltage V_{t0},
GAMMA is the body effect coefficient , and PHI is the surface potential.

✓ Level 1 model is easy to correlate with hand analysis, but it is too simplistic for modern design.

ii. Level 2 and 3 models

✓ The SPICE Level 2 and 3 models add effects of velocity saturation, mobility degradation, subthreshold conduction and drain-induced barrier lowering.

✓ The Level 2 model is based on the Grove-Frohman equations.

✓ Level 3 model is based on empirical equations that provide similar accuracy, faster simulation times and better convergence.

✓ These models are not efficient models for measuring I-V characteristics of modern transistors.

iii. BSIM models

✓ The Berkeley Short-Channel IGFET Model (BSIM) is a very elaborate model that is now widely used in circuit simulation.

✓ The models are derived from the fundamental device physics but uses many number of parameters to fit the behavior of modern transistors.

✓ BSIM versions 1, 2, 3v3, and 4 are implemented as SPICE levels 13, 39, 49, and 54, respectively.

✓ BSIM is quite good for digital circuit simulation.

✓ Features of the model are

- Continuous and differentiable I-V characteristics across subthreshold, linear and saturation regions for good convergence.
- Sensitivity of parameters such as V_t to transistor length and width.
- Detailed threshold voltage model including body effect and drain-induced barrier Lowering (DIBL).
- Velocity saturation, mobility degradation and other short-channel effects.
- Multiple gate capacitance models.
- Diffusion capacitance and resistance models.
- Gate leakage models.

****

1.10: SCALING

❖ **Discuss the scaling principles and its limits. (MAY 2013, Nov 2017)**

❖ **Discuss the principle of constant field and lateral scaling. Write the effects of the above scaling methods on the device characteristics. (Nov 2012, Dec 2011, Nov 2015, May 2016)**

❖ **Explain need of scaling, scaling principles and fundamental units of CMOS inverter. (May 2017)**

✓ In VLSI design, the transistor size has reduced by 30% every two to three years. Scaling is reducing feature size of transistor.

✓ Nowadays, transistors become smaller, switch faster, dissipate less power and cheaper.

✓ Designers need to predict the effect of feature size scaling on chip performance to plan future products and ensure existing products for cost reduction.

✓ Dennard's Scaling Law predicts that the basic operational characteristics of a MOS transistor can be preserved and the performance can be improved.

✓ Parameters of a device are scaled by a dimensionless factor S.

✓ These parameters include the following:

- All dimensions (in the x, y, and z directions)
- Device voltages
- Doping concentration densities

✓ In **constant field scaling,** electric fields remain the same as both voltage and distance shrink.

✓ 1/S scaling is applied to all dimensions, device voltages and concentration densities.

- I_{ds} per transistor are scaled by 1/S.
- No. of transistors per unit area is scaled by S^2.
- Current density is scaled by S and power density remains constant .
    - e.g., $\left(\frac{1}{S} * \frac{1}{S}\right) * S^2$

Lateral scaling (gate-shrink):

✓ Another approach is **lateral scaling**, in which only the gate length is scaled.

✓ This is commonly called as gate shrink, because it can be done easily to an existing mask database for a design.

  - I_{ds} per transistor are scaled by S.
  - No. of transistors per unit area is scaled by S.
  - Current density is scaled by S^2 and power density is scaled by S^2.

✓ The industry generally scales process generations with 30% shrink.

✓ It reduces the cost (area) of a transistor by a factor of two.

✓ A 5% gate shrink (S = 1.05) is commonly applied as a process, becomes mature to boost the speed of components in that process.

✓ **Constant voltage scaling (Fixed scaling)** offers quadratic delay improvement as well as cost reduction.

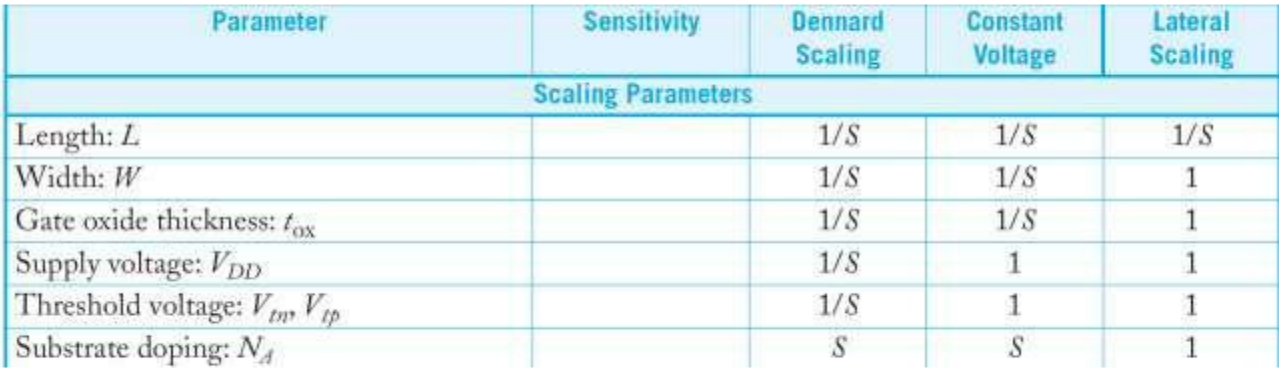
✓ It is also maintaining continuity in I/O voltage standards. Constant voltage scaling increases the electric fields in devices.

  - I_{ds} per transistor are scaled by S.
  - No. of transistors per unit area is scaled by S^2.
  - Current density is scaled by S^3 and power density is scaled by S^3.

✓ A 30% shrink with Dennard scaling improves clock frequency by 40% and cuts power consumption per gate by a factor of 2.

✓ Maintaining a constant field has the further benefit, that many nonlinear factors and wear out mechanisms are unaffected.

✓ From 90nm generation technology, voltage scaling is dramatically slowed down due to leakage. This may ultimately limit CMOS scaling.



Parameter	Sensitivity	Dennard Scaling	Constant Voltage	Lateral Scaling
Scaling Parameters				
Length: L		$1/S$	$1/S$	$1/S$
Width: W		$1/S$	$1/S$	1
Gate oxide thickness: t_{ox}		$1/S$	$1/S$	1
Supply voltage: V_{DD}		$1/S$	1	1
Threshold voltage: V_{tn}, V_{tp}		$1/S$	1	1
Substrate doping: N_A		S	S	1

Table: Influence of scaling on MOS device characteristics

Interconnecting Scaling:

✓ Wires to be scaled equally in width and thickness to maintain an aspect ratio close to 2.

✓ Wires can be classified as local, semiglobal and global.

✓ Local wires run within functional units and use the bottom layers of metal.

✓ Semiglobal wires run across larger blocks or cores, typically using middle layers of metal.

✓ Both local and semiglobal wires are scaling with feature size.

✓ Global wires run across the entire chip using upper levels of metal.

✓ Global wires do not scale with feature size. Indeed, they may get longer (by a factor of DC, on the order of 1.1) because, die size has been gradually increasing.

✓ When wire thickness is scaled, the capacitance per unit length remains constant.

****

1.11: Delay estimation

Derive an expression for the rise time, fall time, fall time and propagation delay of a CMOS inverter. (DEC 2013, APRIL-2015)

- ✓ Important definitions for delay estimation:

Propagation delay time (t_{pd}):

- ✓ Propagation delay time is defined as maximum time from the input crossing 50% to the output crossing 50%.

Contamination delay time (t_{cd}):

- ✓ Contamination delay time is defined as minimum time from the input crossing 50% to the output crossing 50%.

Rise time (t_r):

- ✓ Rise time is defined as time for a waveform to rise from 20% to 80% of its steady-state value

Fall time (t_f):

- ✓ Fall time is defined as time for a waveform to fall from 80% to 20% of its steady-state value
- ✓ Edge rate is average of rise and fall time, (t_{rf}) = (t_r + t_f)/2

Delay estimation response curve:

- ✓ When an input changes, the output will retain its old value for at least the contamination delay and take on its new value in, at most the propagation delay.
- ✓ Delays for the output rising is t_{pdr} /t_{cdr} and the output falling is t_{pdf} /t_{cdf} .
- ✓ Rise/fall times are also called as slopes or edge rates.
- ✓ Propagation and contamination delay times are also called as max-time and min-time respectively.

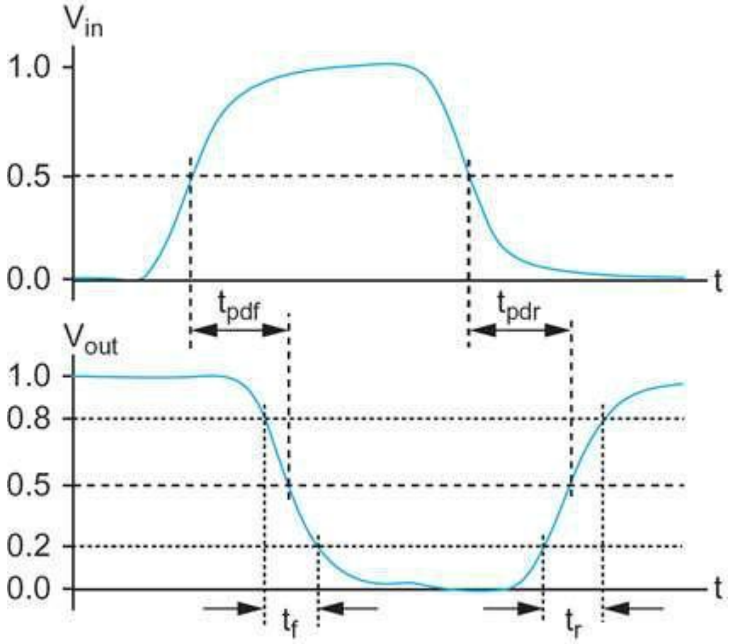


Figure: Delay estimation of CMOS inverter

- ✓ The gate that charges or discharges a node is called the driver. The gates and wires being driven, are called the load. Propagation delay is usually called as delay.
- ✓ Arrival times and propagation delays are defined separately for rising and falling transitions.
- ✓ The delay of a gate may be different from different inputs. Earliest arrival times can also be computed based on contamination delays.
- ✓ Expression of delay for rising output is $t_{PLH} = 0.69\ R_P.C_L$

Where, R_P – effective resistance of pMOS transistor
C_L - load capacitance of CMOS inverter.

- ✓ Expression of delay for falling output is $t_{PHL} = 0.69\ R_N.C_L$
Where, R_N – effective resistance of nMOS transistor
- ✓ Propagation delay of CMOS inverter is $t_P = (t_{PLH} + t_{PHL}) / 2$

1.11.1: RC Delay Model:

Discuss in detail about the resistive and capacitive delay estimation of a CMOS inverter circuit. (MAY 2013) (or)
Briefly explain about the RC delay model.

- ✓ RC delay model approximates the nonlinear transistor I-V and C-V characteristics with an average resistance and capacitance over the switching range of the gate.

Effective Resistance:

- ✓ The RC delay model treats a transistor as a switch in series with a resistor.
- ✓ The effective resistance is the ratio of V_{ds} to I_{ds}.
- ✓ A unit nMOS transistor is defined to have effective resistance R.
- ✓ An nMOS transistor of k times unit width has resistance R/k, because it delivers k times as much current.
- ✓ A unit pMOS transistor has greater resistance, generally in the range of 2R–3R, because of its lower mobility.
- ✓ According to the long-channel model, current decreases linearly with channel length (L) and hence resistance is proportional to L.

Gate and Diffusion Capacitance:

- ✓ Each transistor has gate and diffusion capacitance.
- ✓ C is the gate capacitance of a unit transistor. A transistor of k times unit width has capacitance kC.
- ✓ Diffusion capacitance depends on the size of the source/drain region.
- ✓ Wider transistors have proportionally greater diffusion capacitance. Increasing channel length, increases gate capacitance proportionally but does not affect diffusion capacitance.

Equivalent RC Circuits:

- ✓ Figure shows equivalent RC circuit models for nMOS and pMOS transistors of width k with contacted diffusion on both source and drain.
- ✓ The pMOS transistor has approximately twice the resistance of the nMOS transistor, because holes have lower mobility than electrons.

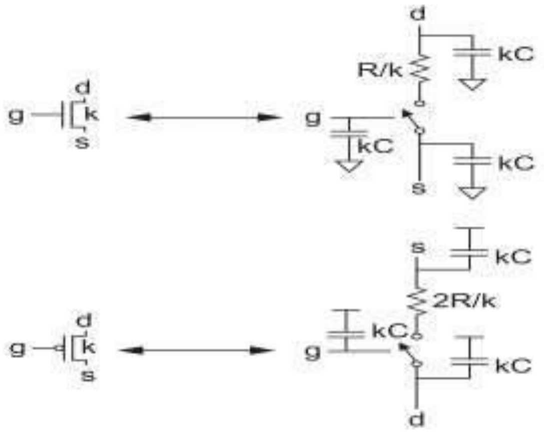


Figure: RC model of nMOS &pMOS transistors

****

1.12: Stick diagram

Explain about stick diagram in VLSI design. (April 2008)

- A **stick diagram** is a cartoon of a chip layout. A "stick diagram" is a paper and pencil tool that use to plan the layout of a cell.
- The stick diagram resembles the actual layout, but uses "sticks" or lines to represent the devices and conductors. Figure 17, shows a stick diagram for an inverter.
- The stick diagram represents the rectangles with lines, which represent wires and component symbols.
- The stick diagram does not represent all the details of a layout, but it makes some relationship much clearer and it is simple to draw.
- Layouts are constructed from rectangles, but stick diagrams are built from cartoon symbols for components and wires.

Stick diagram Rules:

- **Rule 1:** When two or more 'sticks' of the same type cross or touch each other, that represents electrical contact.
- **Rule 2:** When two or more 'sticks' of the different type cross or touch each other, there is no electrical contact. If electrical contact is needed, we have to show the connection explicitly.
- **Rule 3:** When a poly crosses diffusion, it represents a transistor. If a contact is shown, then it is not a transistor. A transistor exists where a polysilicon (red) stick crosses either an n-diffusion (green) stick or a p-diffusion (yellow) stick.
- **Rule 4:** In CMOS, a demarcation line is drawn to avoid touching of p-diff with n-diff. All pMOS must lie on one side of the line and all nMOS will have to be on the other side.

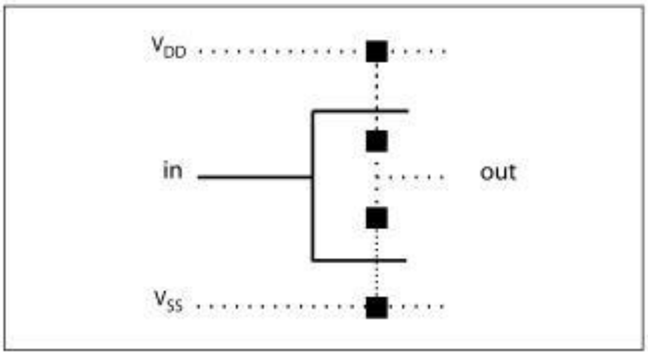


Figure 17: Stick diagram for an inverter

The symbols for wires used on various layers are shown in Figure 18.

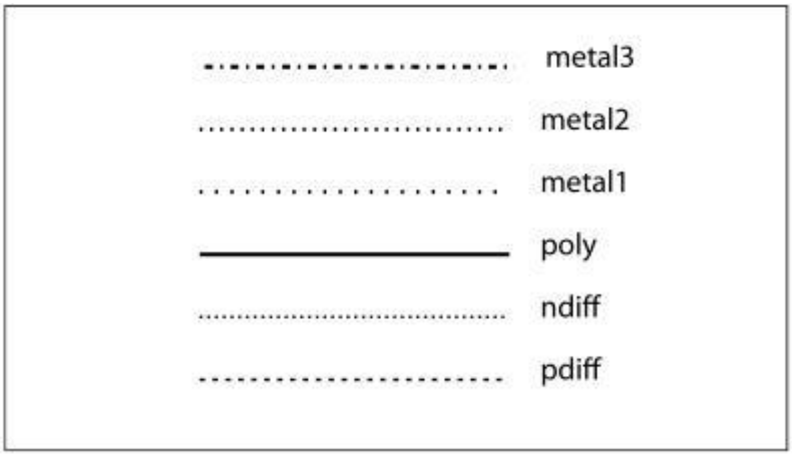


Figure 18: Symbols for wires used on various layers

- ✓ **Drawing stick diagrams in color**: Red for poly, green for n-diffusion, yellow for p-diffusion, and shades of blue for metal are typical colors.
- ✓ A few simple rules for constructing wires from straight-line segments ensure that, the stick diagram corresponds to a feasible layout.
- ✓ Wires cannot be drawn at arbitrary angles. Only horizontal and vertical wire segments are allowed.
- ✓ Two wire segments on the same layer, which cross are electrically connected.
- ✓ Vias to connect wires, which do not normally interact, are drawn as black dots.
- ✓ Figure 19, shows the stick figures for transistors.
- ✓ Each type of transistor is represented as poly and diffusion crossings, much as in the layout.

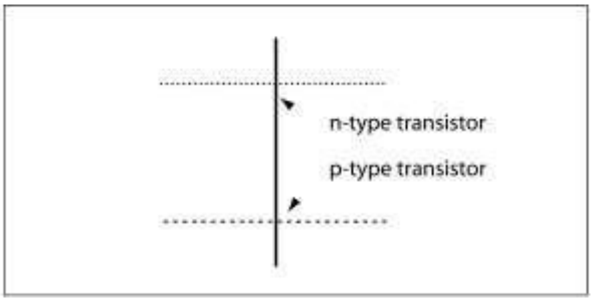
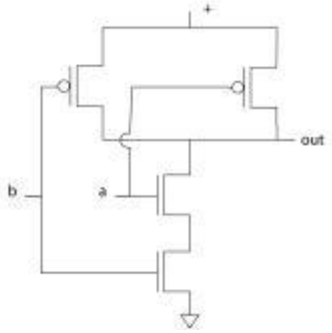


Figure 19: Stick figures for transistors

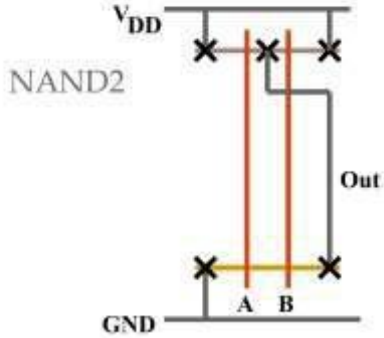
- ✓ Area and aspect ratio are also difficult to estimate from stick diagrams.
- ✓ Stick diagrams are especially important tools for layouts built from large cells and for testing the connections between cells.

Example:

Here is the transistor schematic for a two-input NAND gate:



And here is a stick diagram for the two-input NAND:



****

1.13: Layout Design Rules

- ❖ **Draw and explain briefly the n-well CMOS design rules. (NOV 2007, April 2008, MAY 2014)**
- ❖ **Discuss in detail with a neat layout, the design rules for a CMOS inverter.**
- ❖ **Write the layout design rules and draw diagram for four input NAND and NOR. (Nov 2016)**

- ✓ Layout rules also referred to as **design rules**.
- ✓ It can be considered as prescription for preparing the photomasks, which are used in the fabrication of integrated circuits.
- ✓ The rules are defined in terms of feature sizes (widths), separations and overlaps.
- ✓ The main **objective of the layout rules is to build reliable functional circuits in as small area as possible.**
- ✓ Layout design rules describe how small features can be and how closely they can be reliably packed in a particular manufacturing process.
- ✓ Design rules are a set of geometrical specifications that dictate the design of the layout masks.
- ✓ A design rule set provides numerical values for minimum dimensions and line spacing.
- ✓ Scalable design rules are based on a single parameter (λ), which characterizes the resolution of the process. λ is generally half of the minimum drawn transistor channel length.
- ✓ This length is the distance between the source and drain of a transistor and is set by the minimum width of a polysilicon wire.

Lambda based rule (Scalable design rule):

- ✓ Lambda-based rules are round up dimensions of scaling to an integer multiple of λ.

- ✓ Lambda rules make scaling layout small. The same layout can be moved to a new process, simply by specifying a new value of λ.
- ✓ The minimum feature size of a technology is characterized as 2λ.
- ✓ The MOSIS rules are expressed in terms of lambda.
- ✓ These rules allow some degree of scaling between processes.
- ✓ Only need to reduce the value of lambda and the designs will be valid in the next process down in size.
- ✓ These processes rarely shrink uniformly.
- ✓ Thus, industry usually uses the actual micron design rules for layouts.
- ✓ There are set of micron design rules for a hypothetical 65 nm process.
- ✓ We can observe that, these rules differ slightly but not immensely from lambda based rules with lambda = 0.035 micro meter.
- ✓ Upper level metal rules are highly variable depending on the metal thickness. Thicker wires require greater widths, spacing and bigger vias.
- ✓ Two metal layers in an n-well process has the following:
  - Metal and diffusion have minimum width and spacing of 4 λ.
  - Contacts are 2 λ × 2 λ and must be surrounded by 1 λ on the layers above and below.
  - Polysilicon uses a width of 2 λ.
  - Polysilicon overlaps diffusion by 2 λ where a transistor is desired and has a spacing of 1 λ away where no transistor is desired.
  - Polysilicon and contacts have spacing of 3 λ from other polysilicon or contacts.
  - N-well surrounds pMOS transistors by 6 λ and avoids nMOS transistors by 6 λ.

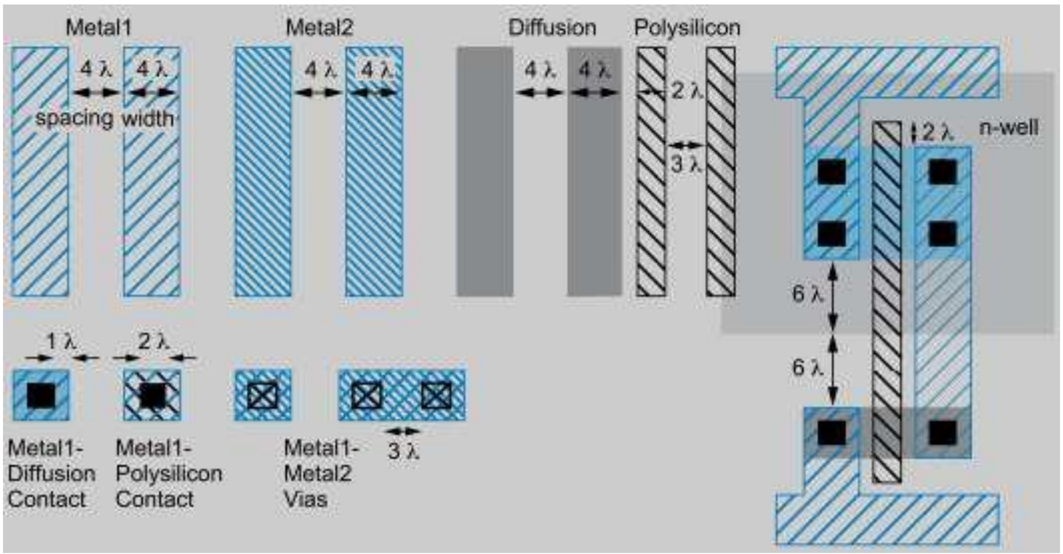


Figure: Simplified λ -based design rules with CMOS inverter layout diagram

Design Rule:

Well Rules:

- ✓ The n-well is usually a deeper implant than the transistor source/drain implants.
- ✓ Therefore, it is necessary to provide sufficient clearance between the n-well edges and the adjacent *n*+ diffusions.

Transistor Rules:

- ✓ CMOS transistors are generally defined by at least four physical masks.

✓ There are active (also called diffusion, diff, thinox, OD, or RX), n-select (also called n-implant, n-imp, or nplus), p-select (also called p-implant, pimp, or pplus) and polysilicon (also called poly, polyg, PO, or PC).

✓ The active mask defines all areas, where n- or p-type diffusion is to be placed *or* where the gates of transistor are to be placed.

Contact Rules:

✓ There are several generally available contacts:

- Metal to p-active (p-diffusion)
- Metal to n-active (n-diffusion)
- Metal to polysilicon
- Metal to well or substrate

Metal Rules:

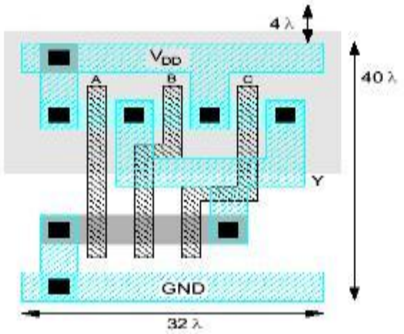
✓ Metal spacing may vary with the width of the metal line.

✓ Metal wire width of minimum spacing may be increased. This is due to etch characteristics versus large metal wires.

✓ Processes may allow vias to be placed over polysilicon and diffusion regions.

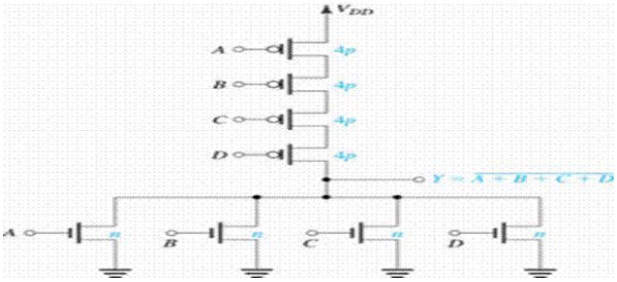
✓ Some processes allow vias to be placed within these areas, but do not allow the vias to the boundary of polysilicon or diffusion.

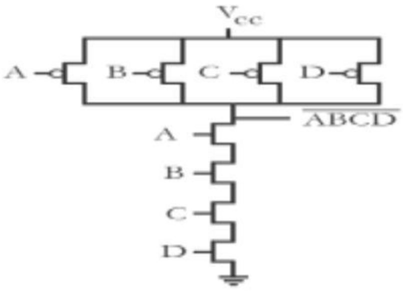
- Horizontal N-diffusion and p-diffusion strips
- Vertical polysilicon gates
- Metal1 VDD rail at top
- Metal1 GND rail at bottom



❖ **Draw diagram for four input NAND and NOR gate. (Nov 2017)**

4 input NOR gate **4 input NANA gate**





****

1.14: Latchup problem:

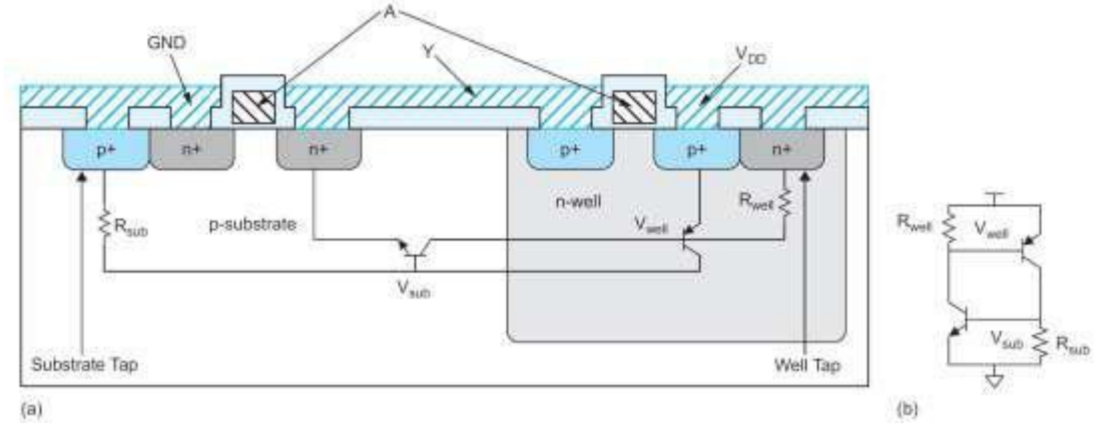
❖ ***Discuss the orgin of latch up problems in CMOS circuits with necessary diagrams. Explain the remedial measures. (Nov 2007, April 2008)***

✓ A CMOS process is slowed down by developing low-resistance paths between V_{DD} and GND, causing catastrophic meltdown. The phenomenon is called latchup.

✓ Latchup problem arises when parasitic bipolar transistors are formed by the substrate, well and diffusion.

✓ The cause of the latchup effect can be understood by examining the process cross-section of a CMOS inverter, as shown in Figure (a).

✓ The schematic shows, a circuit composed of an npn-transistor, a pnp-transistor, and two resistors connected between the power and ground rails (Figure (b)).



✓ The npn transistor is formed between the grounded n-diffusion source of the nMOS transistor, the p-type substrate and the n-well.

✓ The resistors are due to the resistance through the substrate or well to the nearest substrate and well taps.

✓ The cross-coupled transistors form a bistable silicon-controlled rectifier (SCR). Both parasitic bipolar transistors are OFF.

✓ Latchup can be triggered, when transient currents flow through the substrate during normal chip power-up.

✓ Latchup prevention is easily accomplished by

- Minimizing Rsub and Rwell.
- Use of guard rings

✓ SOI process avoids latchup entirely, because they have no parasitic bipolar structures.

****

1.15: Process parameters for MOS and CMOS:

CMOS TECHNOLOGIES:

The four main CMOS technologies are

- n-Well process
- p-Well process
- Twin-tub Process
- Silicon on Insulator

❖ **Explain the different steps involved in n-well CMOS fabrication process with neat diagrams. (Nov 2007, Nov 2009, Nov 2016)**

❖ **Describe with neat diagram the n-well and channel formation in CMOS process. (Nov/Dec 2014)(Nov/Dec 2011) (April/May 2011) (Nov/Dec 2012)**

n-WELL PROCESS:

Step 1: Start with blank wafer

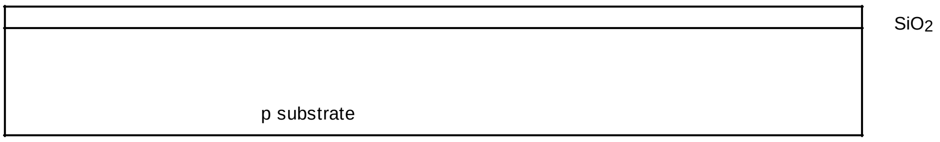
First step will be to form the n-well

– Cover wafer with protective layer of SiO_2 (oxide)

– Remove layer where n-well should be built.

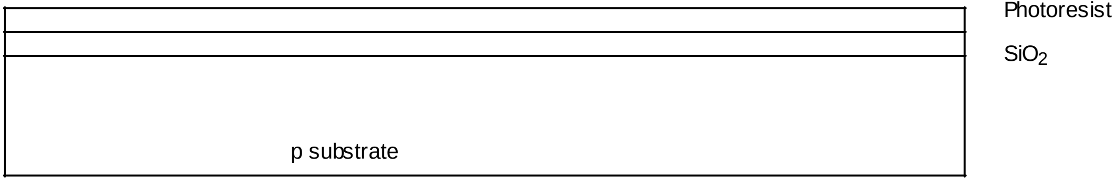
Step 2: Oxidation

Grow SiO_2 on top of Si wafer, at 900 – 1200^0 C with H_2O or O_2 in oxidation furnace.



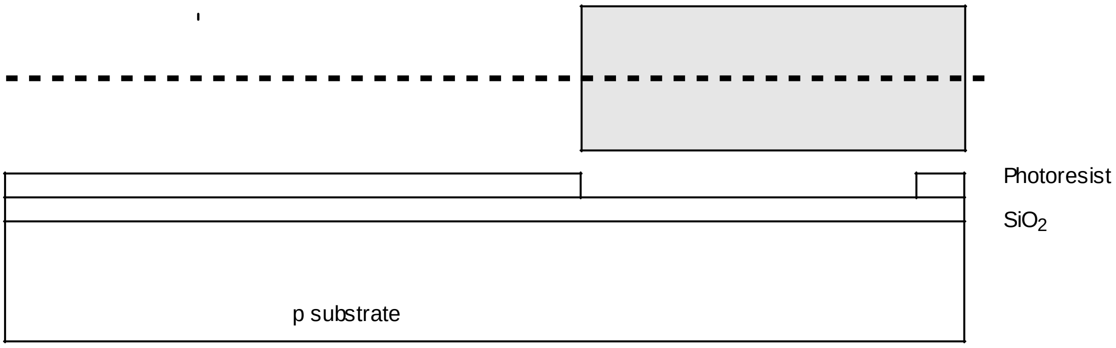
Step 3: Photoresist

- Spin on photoresist
  - Photoresist is a light-sensitive organic polymer.
  - Softens, where exposed to light.



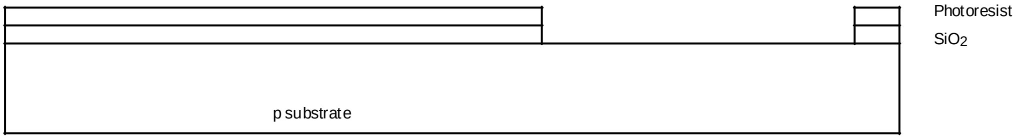
Step 4: Lithography

- Expose photoresist through n-well mask.
- Strip off exposed photoresist.



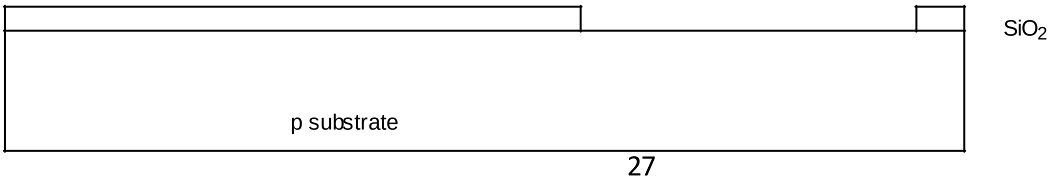
Step 5: Etch

- Etch oxide with hydrofluoric acid (HF).
- Only attracts oxide, where resist has been exposed.



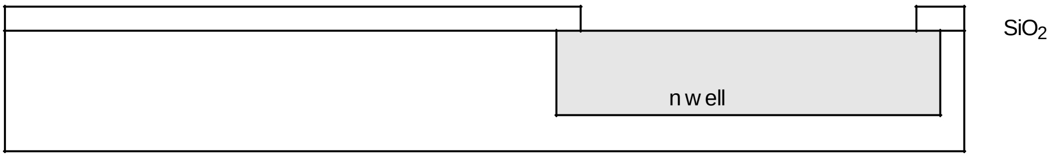
Step 6: Strip Photoresist

- Etch the remaining photoresist using a mixture of acids.



Step 7: n-well

n-well is formed with diffusion or ion implantation.

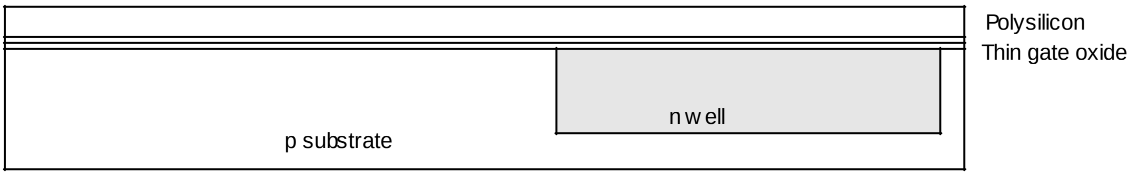


Step 8: Strip Oxide

- Strip off the remaining oxide using HF.
- Back to bare wafer with n-well.
- Subsequent steps involve similar series of steps.

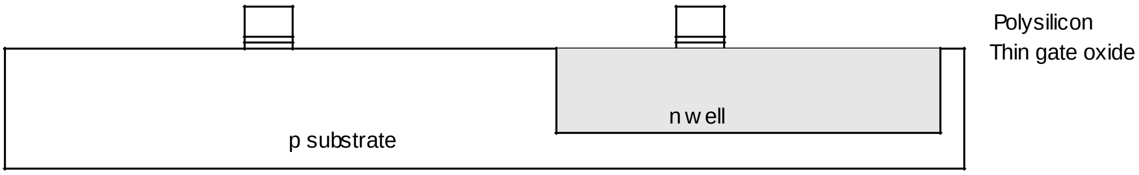
Step 9:Polysilicon

- Deposit thin layer of oxide. Use CVD to form poly and dope heavily to increase conductivity.



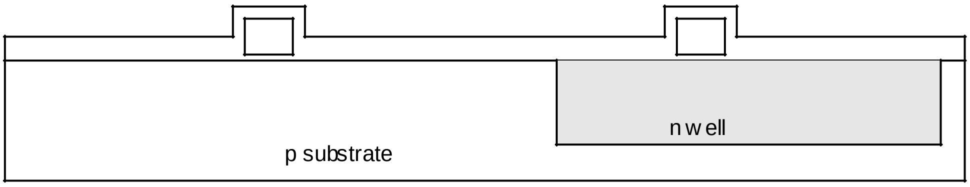
Step 10: Polysilicon Patterning

- Use same lithography process to pattern polysilicon.



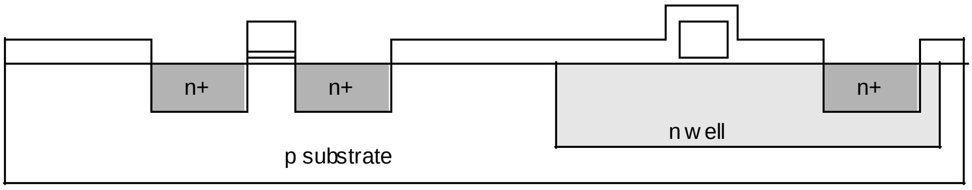
Step 11: Self-Aligned Process

Cover with oxide to define n diffusion regions.



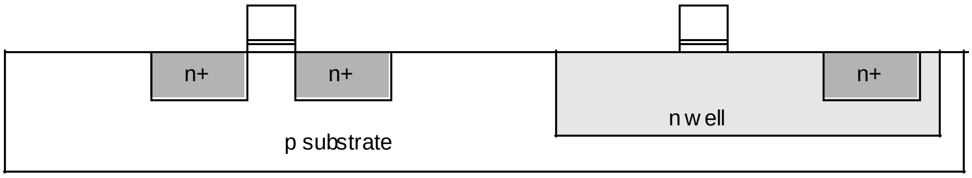
Step 12: N-diffusion

- Pattern oxide, using n+ active mask to define n diffusion regions.
- Diffusion or ion implantation is used to create n diffusion regions.



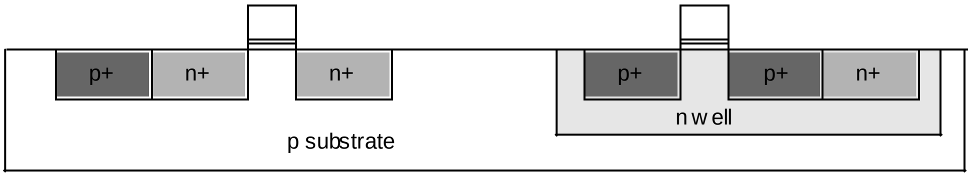
Step 13:

- Strip off oxide to complete patterning step.



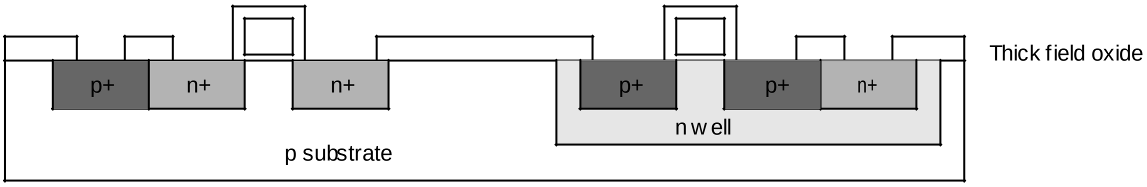
Step 14:P-Diffusion

- Similar set of steps are followed to form p+ diffusion regions for pMOS source and drain and substrate contact.



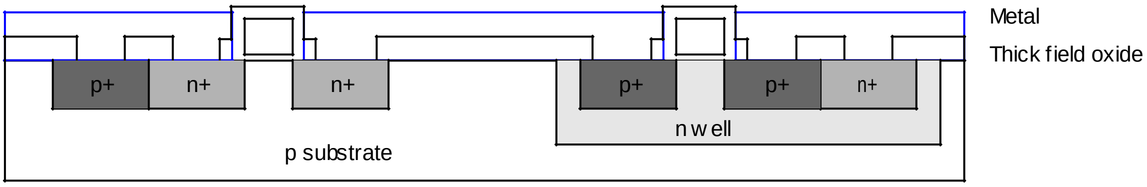
Step 15: Contacts

- Now, we need to wire together the devices.
- Cover chip with thick field oxide.
- Etch oxide, where contact cuts are needed.



Step 16: Metallization

- Sputter on aluminum over whole wafer.
- Pattern to remove excess metal, leaving wires.



****

P-WELL PROCESS:

- A common approach to p-well CMOS fabrication is to start with moderately doped n-type substrate (wafer), create the p-type well for the n-channel devices and build the p-channel transistor in the native n-substrate.

❖ **Explain the twin tub process with a neat diagram. (Nov 2007, April 2008) Twin-tub process:**

Step 1:

n- Substrate is taken initially, which is shown in figure.

Step 2:

Next step is epitaxial layer deposition. Lightly doped epitaxial layer is deposited above n-substrate.

Step 3:

The next step is tub formation. Two wells are formed namely n-well and p-well. Polysilicon layer is formed above overall substrate.

Step 4:

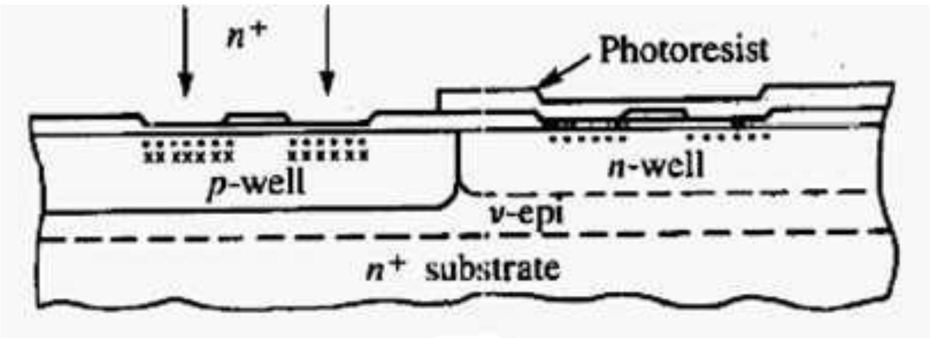
Polysilicon gates are formed for n-well and p-well by using photo-etching process.

Step 5:

n^+ diffusion is formed in n-well, P^+ diffusion is formed in p-well. These are used for V_{DD} contact and V_{SS} contact. These are known as substrate formation.

Step 6:

Then, contact cuts are defined as in n-well process. Then metallization is processed.



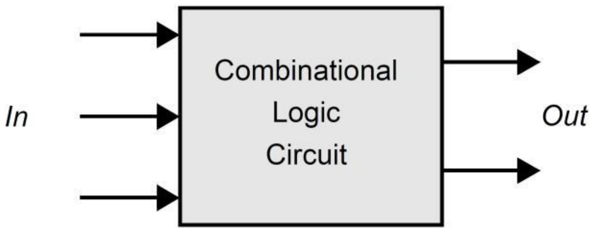
****
*

UNIT II - COMBINATIONAL LOGIC CIRCUITS

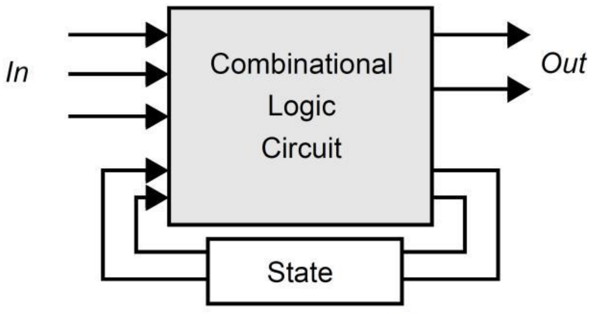
Examples of Combinational Logic Design, Elmore‟s constant, Pass transistor Logic, Transmission gates, static and dynamic CMOS design, Power dissipation – Low power design principles.

2.1:Introduction (Combinational Logic Circuit):

- Digital logics are divided into combinational and sequential circuits.
- Combinational circuits are circuits where outputs depend only on the present inputs.
- For sequential or regenerative circuit, the output is not only a function of the current input data, but also of previous values of the input signals.
- A sequential circuit includes a combinational logic portion and a memory module that holds the state. Example are registers, counters and memory.
- The building blocks for combinational circuits are logic gates, while the building blocks for sequential circuits are registers and latches.
- The delay of a logic gate depends on its output current I, load capacitance C and output voltage swing V.



(a) Combinational



(b) Sequential

- Alternative (ratioed circuits, dynamic circuits and pass transistor circuits) CMOS logic configurations are called circuit families.
- nMOS transistors provide more current than pMOS for the same size and capacitance, so nMOS networks are preferred.

Examples of combinational circuits

(i) CMOS inverter:

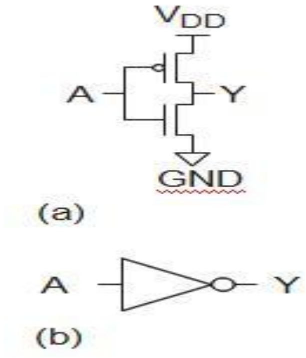


Figure: Inverter (a) schematic (b) symbol $Y = \overline{A}$

(ii) **Two input NAND gate:**

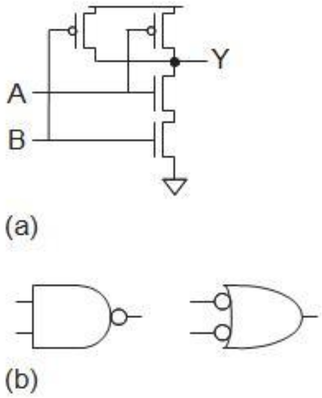


Figure: 2 input NAND gate (a) schematic (b) symbol

(iii) **Three input NAND gate:**

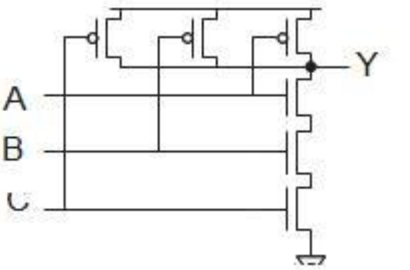


Figure: 3 –input NAND gate Schematic Y=A.B.C

(iv) **Two input NOR gate:**

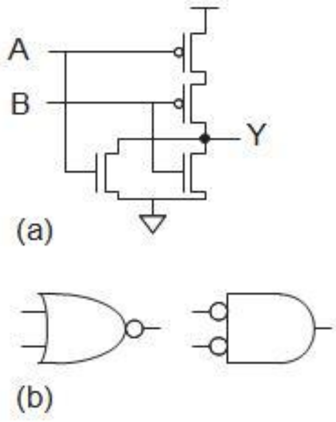
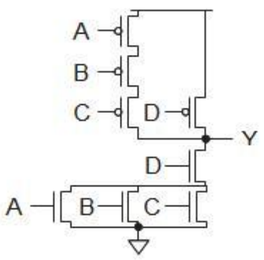


Figure: 2-input NOR gate (a) schematic (b) Symbol $Y = \overline{A + B}$

Example:

Sketch a static CMOS gate computing $Y = \overline{(A + B + C) \cdot D}$.



* * *

2.2: Circuit Families

Briefly discuss about the classification of circuit families and comparison of the circuit families. (May 2014, APRIL-2015)

Draw and explain the function of static CMOS.

2.2.1: Static CMOS

- Static CMOS circuits with complementary nMOS pulldown and pMOS pullup networks are used for the majority of logic gates in integrated circuits.

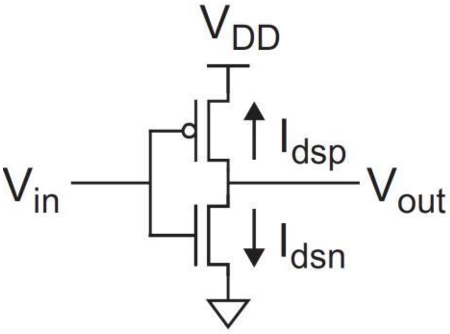


Figure: Static CMOS inverter

Advantages of static CMOS:

- Static CMOS circuits have good noise margins
- Static CMOS circuits are fast, low power, easy to design.
- Static CMOS circuits are widely supported by CAD tools.
- Static CMOS circuits are available in standard cell libraries.

Drawback of static CMOS

- It requires both nMOS and pMOS transistors for each input.
- It has a relatively large logical effort.
- Gate delay is

increased. **a. Bubble pushing**

- CMOS stages are inherently inverting, so AND and OR functions must be built from NAND and NOR gates.
- DeMorgan‟s law helps with this conversion:

$$\overline{A.B} = \overline{A} + \overline{B}$$

$$\overline{A + B} = \overline{A}.\overline{B}$$

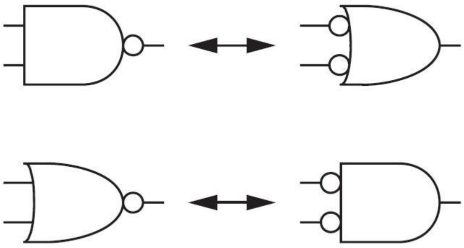


Figure: Bubble pushing with DeMorgan‟s law

- A NAND gate is equivalent to an OR of inverted inputs.
- A NOR gate is equivalent to an AND of inverted inputs.
- The same relationship applies to gates with more inputs.
- Switching between these representations is easy and is often called bubble pushing.

b. Compound Gates

- Static CMOS also efficiently handles compound gates computing various inverting combinations of AND/OR functions in a single stage.
- The function F = AB +CD can be computed with an AND-OR INVERT- 22 (AOI22) gate and an inverter, as shown in Figure.

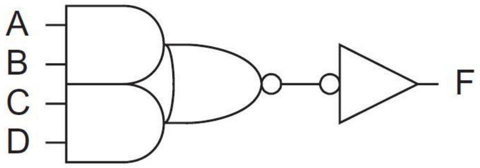


Figure: Logic using AOI22 gate

- Logical effort of compound gates can be different for different inputs.
- Figure shows, how logical efforts can be estimated for the AOI21, AOI22 and a more complex compound AOI gate.

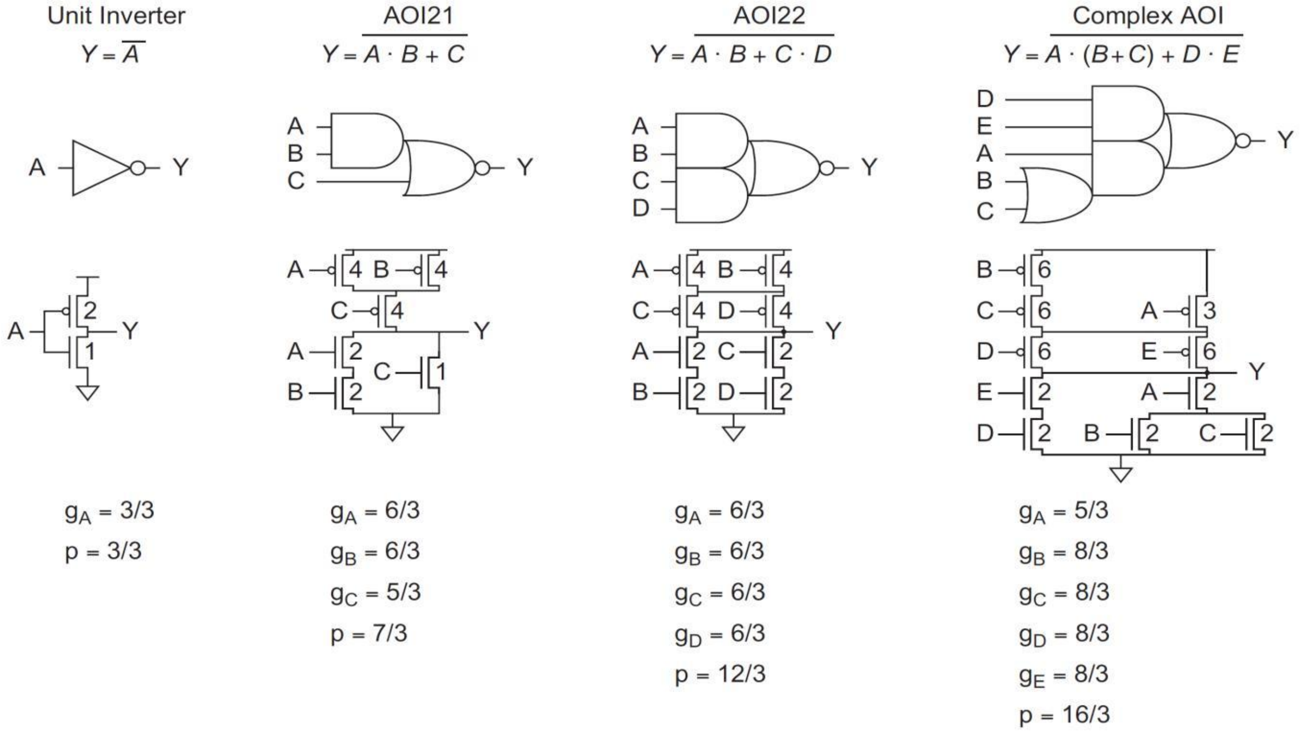


Figure: Logical efforts and parasitic delays of AOI gates

c. Input ordering delay effect

- The logical effort and parasitic delay of different gate inputs are different.
- Consider the falling output transition occurring, when one input hold a stable 1 value and the other rises from 0 to 1.

- If input B rises last, node x will initially be at $V_{DD} – V_t = V_{DD}$, because it was pulled up through the nMOS transistor on input A.
- The Elmore delay is (R/2)(2C) + R(6C) =7RC=2.33 τ
- If input A raises last, node x will initially be at 0 V, because it was discharged through the nMOS transistor on input B.
- No charge must be delivered to node x, so the Elmore delay is simply R(6C) =6RC =2τ.

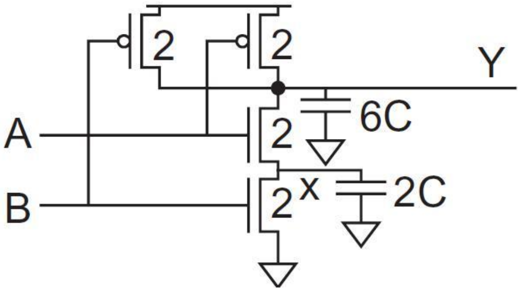
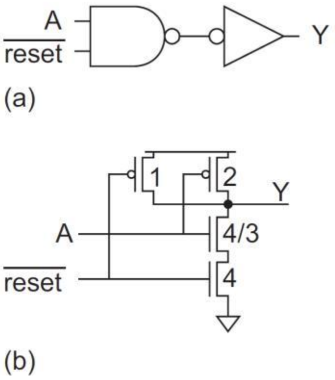


Figure: 2 –input NAND gate Schematic Y=A.B

- We define the outer input to be the input closer to the supply rail (e.g., B) and the inner input to be the input closer to the output (e.g., A).
- Therefore, if one signal is known to arrive later than the others, the gate is faster when that signal is connected to the inner input.

d. Asymmetric gates

- When one input is far less critical than another, even symmetric gates can be made asymmetric to favor the late input at the expense of the early one.
- In a series network, this involves connecting the early input to the outer transistor and making the transistor wider, so that, it offers less series resistance when the critical input arrives.
- In a parallel network, the early input is connected to a narrower transistor to reduce the parasitic capacitance.
- Consider the path in Figure (a). Under ordinary conditions, the path acts as a buffer between A and Y.
- When reset is asserted, the path forces the output low.
- If reset only occurs under exceptional circumstances and take place slowly, the circuit should be optimized for input-to-output delay at the expense of reset.
- This can be done with the asymmetric NAND gate in Figure (b).



(a)

(b)

Figure: Resettable buffer optimized for data input

e. Skewed gates

What is meant by skewed gate and give functions of skewed gate with schematic diagrams?

- One input transition is more important than the other. HI-skew gates to favor the rising output transition. LO-skew gates to favor the falling output transition.
- This favoring can be done by decreasing the size of the noncritical transistor.
- The logical efforts for the rising (up) and falling (down) transitions are called g_u and g_d, respectively.
- Figure (a) shows, how a HI-skew inverter is constructed by downsizing the nMOS transistor.
- This maintains the same effective resistance for the critical transition, while reducing the input capacitance relative to the unskewed inverter of Figure (b).
- Thus reducing the logical effort on that critical transition to $g_u = 2.5/3 = 5/6$.
- The logical effort for the falling transition is estimated by comparing the inverter to a smaller unskewed inverter with equal pulldown current, shown in Figure (c), giving a logical effort of $g_d = 2.5/1.5 = 5/3$.

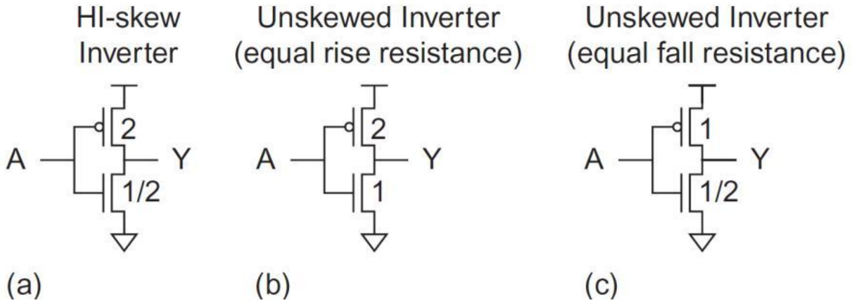


Figure: Logical effort calculation for HI-skew inverter

- Figure shows, HI skew and LO-skew gates with a skew factor of two. Skewed gates are sometimes denoted with an H or an L on their symbol in a schematic.

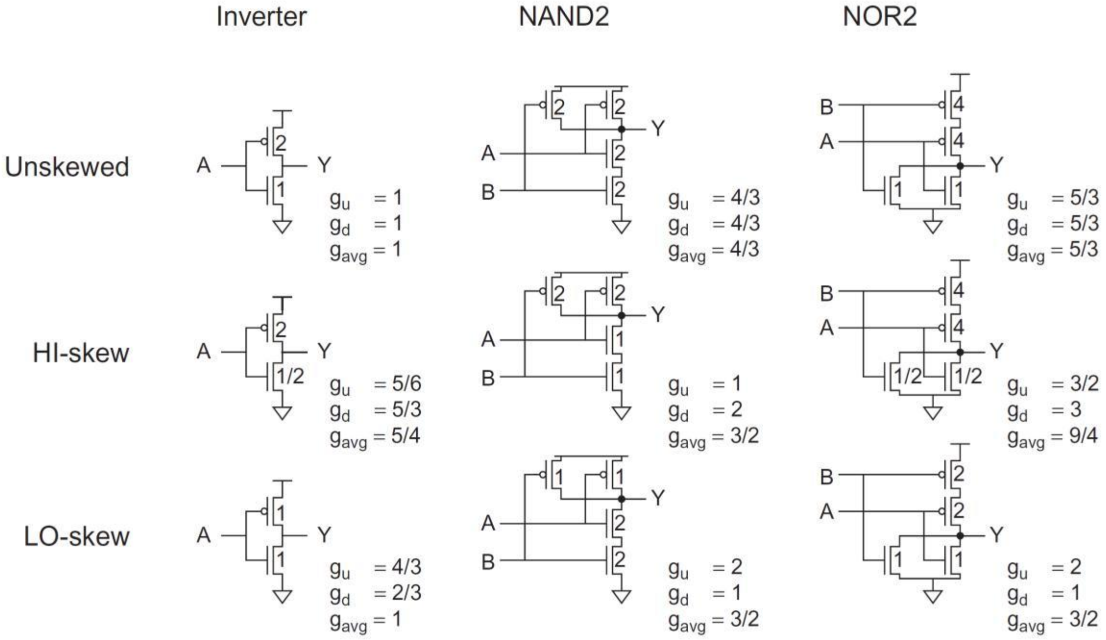


Figure: List of skewed gates

f. P/N ratios

- By accepting a slower rise delay, the pMOS transistors can be downsized to reduce input capacitance and average delay significantly.
- P/N ratio is defined as the ratio of pMOS to nMOS transistor width. For processes, a mobility ratio of µn/µp = 2.

g. Multiple threshold voltages

- Some CMOS processes offer two or more threshold voltages.
- Transistors with lower threshold voltages produce more ON current, but also leak exponentially more OFF current.
- Libraries can provide both high- and low-threshold versions of gates.
- The low-threshold gates can be used carefully to reduce the delay of critical paths.
- Skewed gates can use low-threshold devices on, only the critical network of transistors.

****

2.3: Elmore's Delay

What is meant by Elmore's delay and give expression for Elmore's delay?

- The Elmore delay model estimates the delay from a source, switching to one of the leaf nodes. Delay is the sum over each node i of the capacitance C_i on the node multiplied by the effective resistance R.

Propagation delay time :

$$t_{pd} \approx \sum_{\text{nodes } i} R_{i-to-source} C_i$$

$$= R_1C_1 + (R_1 + R_2)C_2 + ... + (R_1 + R_2 + ... + R_N)C_N$$

- Delay of an ideal fanout-of-1 inverter with no parasitic capacitance is $\tau = 3RC$.
- The normalized delay *d* relative to this inverter delay:

$$d = \frac{t_{pd}}{\tau}$$

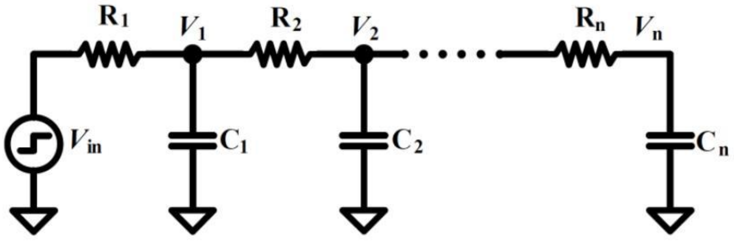
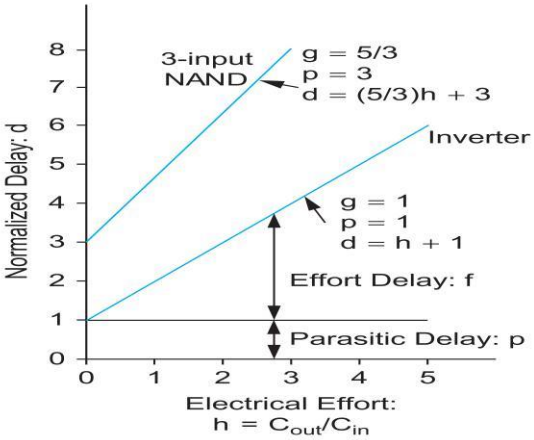


Figure: RC delay equivalent for series of transistors

Linear delay model

- The RC delay model is one, where delay is a linear function of the fanout of a gate.
- The normalized delay of a gate can be expressed in units of Y as d = f + p.
  Where p is the parasitic delay inherent to the gate when no load is attached.
  f is the effort delay or stage effort that depends on the complexity.
- Effort delay of the gate is f = gh.
  Where g is the logical effort (An inverter has a logical effort of 1).
- Logical effort is defined as the ratio of the input capacitance of a gate to the input capacitance of an inverter delivering the same output current.
- h is the fanout or electrical effort. Electrical effort is defined as ratio of the output capacitance to input capacitance.
- More complex gates have greater logical efforts, indicating that they take longer time to drive a given fanout.
- For example, the logical effort of the 3-input NAND gate is 5/3.
- The electrical effort can be computed as $h = \frac{C_{out}}{C_{in}}$

  Where C_{out} is the capacitance of the external load being driven and C_{in} is the capacitance of the gate.
- Normalized delay vs electrical effort for an idealized inverter and 3-input NAND gate shown in diagram.
- The y-intercepts indicate the parasitic delay. The slope of the lines is the logical effort.
- The inverter has a slope of 1. The NAND gate has a slope of 5/3.



****

2.4: Pass Transistor circuits:

Explain Pass transistor logic with neat sketches. (April 2008)

- In pass-transistor circuits, inputs are applied to the source/drain diffusion terminals.
- These circuits build switches using either nMOS pass transistors or parallel pairs of nMOS and pMOS transistors called as transmission gates.

- The nMOS transistors pass „0‟s well but 1‟s poorly. Figure (a) shows an nMOS transistor with the gate and drain tied to V_{DD}.
- Initially at $Vs = 0$. $V_{gs} > V_{tn}$, so the transistor is ON and current flows.
- Therefore, nMOS transistors attempting to pass a 1 never pull the source above $V_{DD} – V_{tn}$. This loss is called a threshold drop.
- The pMOS transistors pass 1‟s well but 0‟s poorly.
- If the pMOS source drops below $|V_{tp}|$, the transistor cuts off.
- Hence, pMOS transistors only pull down to a threshold above GND, as shown in Figure (b).

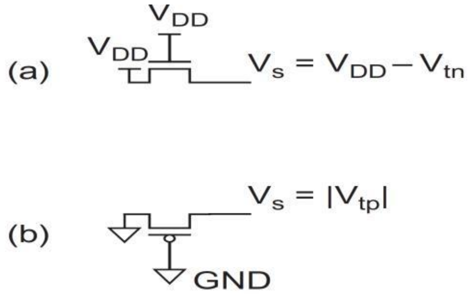
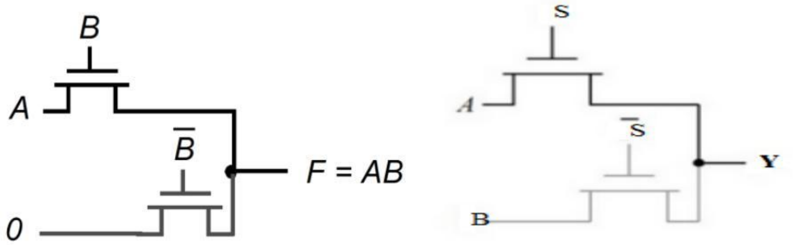


Figure : Pass Transistor threshold drops

- Figures show an implementation of the AND function and 2x1 multiplexer using only NMOS transistors.



AND Logic 2x1 multiplexer

- In AND gate, if the B input is high, the top transistor is turned ON and copies the input A to the output F.
- When B is low, the bottom pass transistor is turned ON and passes a 0.
- In 2x1 multiplexer, if the S selection input is high, the top transistor is turned ON and allows input A to the output Y.
- When S is low, the bottom pass transistor is turned ON and passes the B input.
- An NMOS device is effective at passing a 0 but is poor at pulling a node to V_{DD}. When the pass transistor pulls a node high, the output only charges up to V_{DD} -V_{tn}.

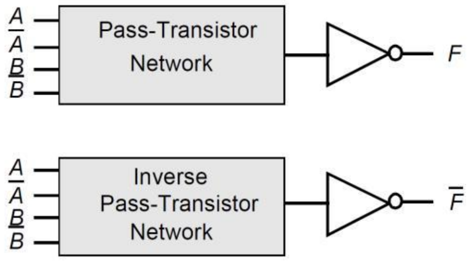
Application:

- Pass transistors are essential to the design of efficient 6-transistor static RAM cells used in modern systems.

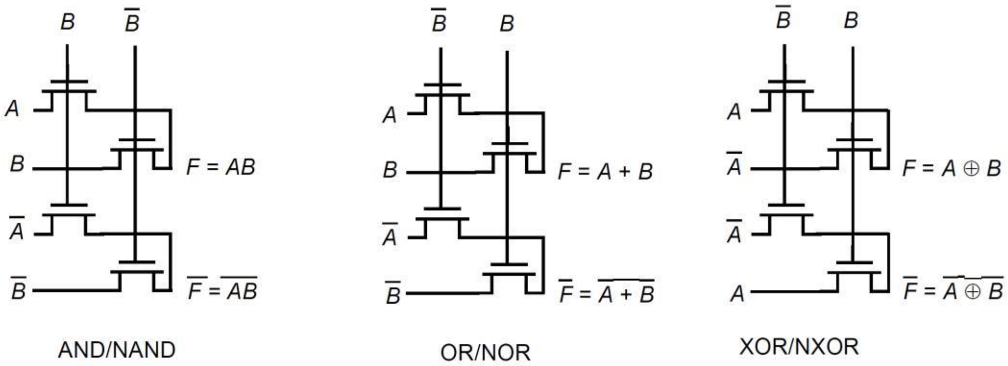
2.4.1: Differential Pass Transistor Logic

- For high performance design, a differential pass-transistor logic family, called CPL or DPL, is commonly used.
- The basic idea is to accept true and complementary inputs and produce true and complementary outputs.
- A number of CPL gates (AND/NAND, OR/NOR, and XOR/NXOR) are shown in Figure.

- Since the circuits are *differential*, complementary data inputs and outputs are always available.
- Both polarities of every signal eliminate the need for extra inverters, as is often the case in static CMOS or pseudo-NMOS.
- CPL belongs to the class of *static* gates, because the output-defining nodes are always connected to either V_{DD} or GND through a low resistance path.
- This is advantage for the noise flexibility.



(a) Basic concept



(b) Example pass-transistor networks

Figure: Complementary pass-transistor logic (CPL).

2.4.2: CMOS with transmission gates

❖ **Discuss in detail the characteristics of CMOS Transmission gates.(May 2016, May 2017, Nov 2017)**

❖ **Explain Transmission gates with neat sketches. (April 2008)**

- A transmission gate in conjunction with simple static CMOS logic is called CMOS with transmission gate.
- A transmission gate is parallel pairs of nMOS and pMOS transistor.
- A single nMOS or pMOS pass transistor suffers from a threshold drop.
- Transmission gates solve the threshold drop but require two transistors in parallel.
- The resistance of a unit-sized transmission gate can be estimated as R for the purpose of delay estimation.
- Current flow the parallel combination of the nMOS and pMOS transistors. One of the transistors is passing the value well and the other is passing it poorly.
- A logic-1 is passed well through the pMOS but poorly through the nMOS.
- Estimate the effective resistance of a unit transistor passing a value in its poor direction as twice the usual value: 2R for nMOS and 4R for pMOS.

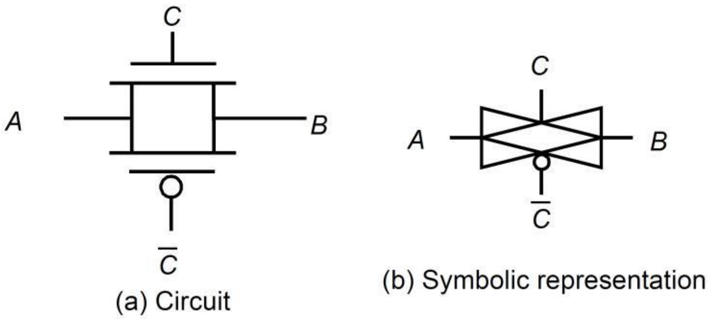


Figure: CMOS Transmission gate

- The given below figure shows the parallel combination of resistances. When passing a 0, the resistance is R || 4R = (4/5)R.
- The effective resistance passing a 1 is 2R || 2R = R.
- Hence, a transmission gate made from unit transistors is approximately R in either direction.
- Transmission gates are built using equal-sized nMOS and pMOS transistors.
- Boosting the size of the pMOS transistor only slightly improves the effective resistance while significantly increasing the capacitance.

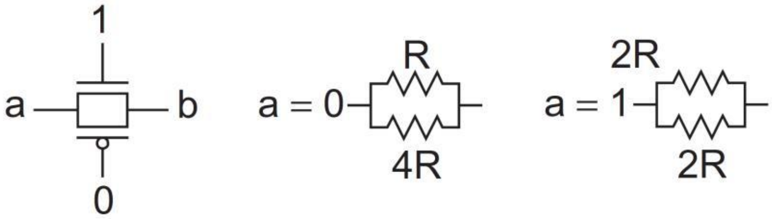


Figure: Effective resistance of a unit transmission gate

- Figure (a) redraws the multiplexer to include the Inverters that drive the diffusion inputs but to exclude the output inverter. Figure (b) shows this multiplexer drawn at the transistor level.

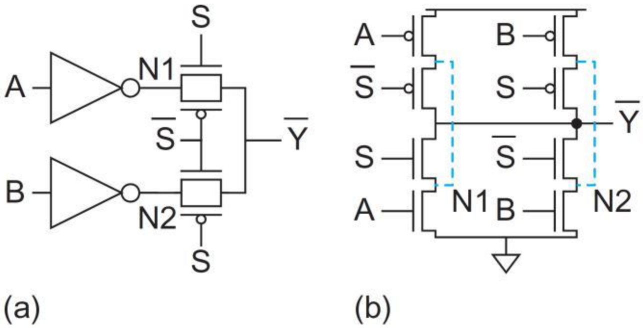


Figure: CMOSTG in a 2-input inverting multiplexer

****

2.5: Static CMOS design:

2.5.1:Ratioed Circuits:

Write short notes on ratioed circuits. (Nov 2016)

- The ratioed gate consists of an nMOS pulldown network and pullup device called the static load.
- When the pulldown network is OFF, the static load pulls the output to 1.
- When the pulldown network turns ON, it fights the static load.
- The static load must be weak enough that, the output pulls down to an acceptable 0. Hence, there is a ratio constraint between the static load and pulldown network.

Advantage: Stronger static loads produce faster rising outputs.

Disadvantages:

- Degrade the noise margin and burn more static power when the output is 0.
- A resistor is a simple static load, but large resistors consume a large layout area in typical MOS processes.

- Another technique is to use an nMOS transistor with the gate tied to V_{GG} (Shown in fig.(b)). If $V_{GG} = V_{DD}$, the nMOS transistor will only pull up to $V_{DD} - V_t$.
- Figure (c) shows depletion load ratioed circuit.

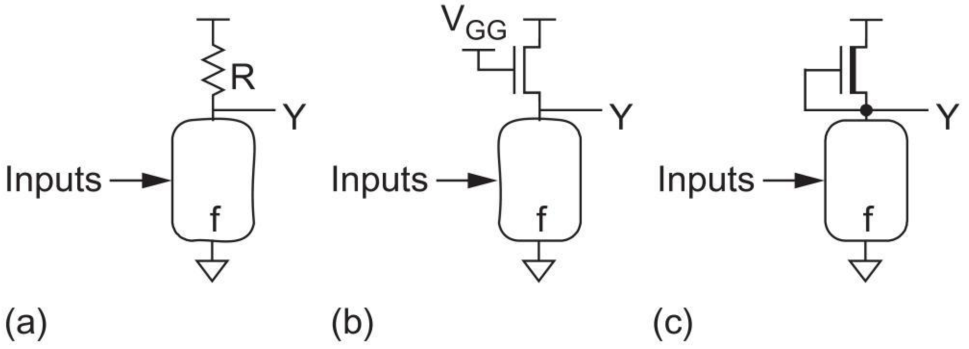


Figure: nMOS ratioed gates

2.5.2: pseudo nMOS

Explain the detail about pseudo-nMOS gates with neat circuit diagram. (April/May 2011) (Nov/Dec 2013)

- Figure (a) shows a pseudo-nMOS inverter.
- The static load is built from a single pMOS transistor that has its gate grounded, so it is always ON.
- The beta ratio affects the shape of the transfer characteristics and the V_{OL} of the inverter.
- Larger relative pMOS transistor size offer faster rise time, but less sharp transfer characteristics.
- **Drawback:** Pseudo-nMOS gates will not operate correctly if $V_{OL} > V_{IL}$ of the receiving gate.

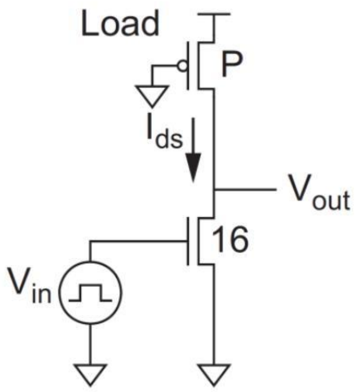


Figure (a): pseudo nMOS inverter

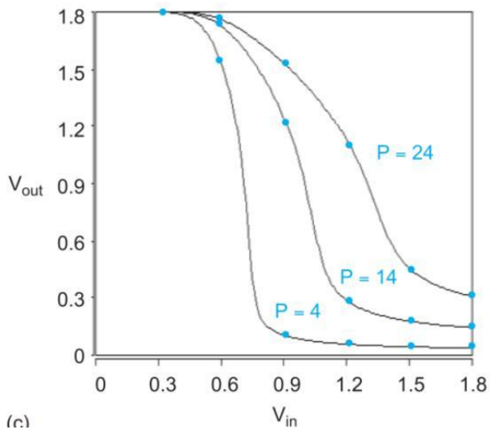


Figure: D.C Charateristics

- Figure shows several pseudo-nMOS logic gates.

Implement NAND gate using pseudo- nMOS logic. (Nov 2013)

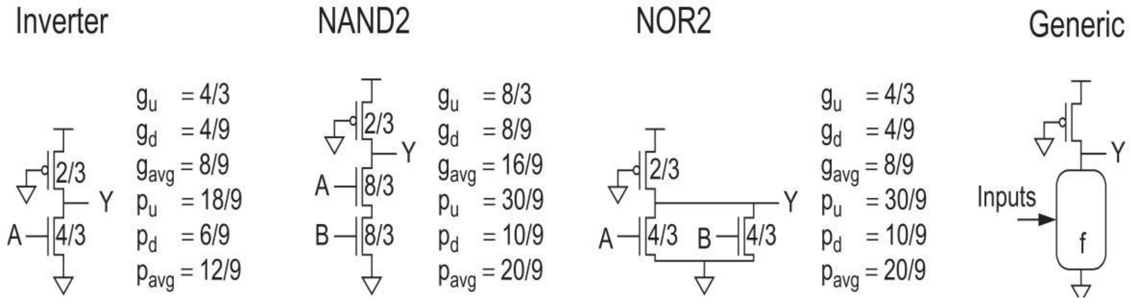


Figure: Pseudo-nMOS logic gates

2.5.3: Ganged capacitor:

- Figure shows pairs of CMOS inverters ganged together.
- The truth table is given in Table, showing that the pair compute the NOR function. Such a circuit is sometimes called a symmetric ^{2} NOR, or ganged CMOS.

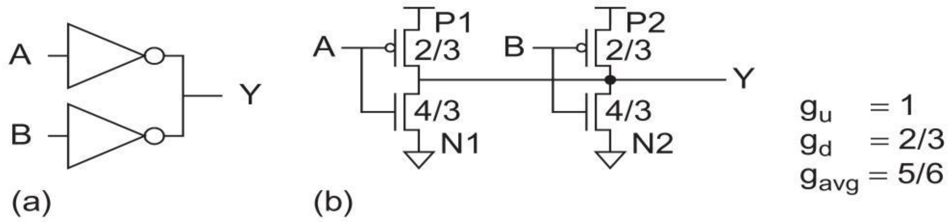


Figure: symmetric ^{2} NOR gate.

A	*B*	*N1*	*P1*	*N2*	*P2*	*Y*
0	0	OFF	ON	OFF	ON	1
0	1	OFF	ON	ON	OFF	~ 0
1	0	ON	OFF	OFF	ON	~ 0
1	1	ON	OFF	ON	OFF	0

Table: Operation of symmetric NOR

- When one input is 0 and the other 1, the gate can be viewed as a pseudo-nMOS circuit with appropriate ratio constraints.
- When both inputs are 0, both pMOS transistors turn on in parallel, pulling the output high faster than they would, in an ordinary pseudo nMOS gate.
- When both inputs are 1, both pMOS transistors turn OFF, saving static power dissipation.

2.5.4: Cascode voltage switch logic

Explain about DCVSL logic with suitable example. (May 2017)

- Cascode Voltage Switch Logic (CVSL) seeks the benefits of ratioed circuits without the static power consumption.
- It uses both true and complementary input signals and computes both true and complementary outputs using a pair of nMOS pulldown networks, as shown in Figure (a).
- The pulldown network f implements the logic function as in a static CMOS gate, while $\overline{f}$ uses inverted inputs feeding transistors arranged in the conduction complement.
- For any given input pattern, one of the pulldown networks will be ON and the other OFF.
- The pulldown network that is ON will pull that output low.
- This low output turns ON the pMOS transistor to pull the opposite output high.
- When the opposite output rises, the other pMOS transistor turns OFF, so no static power dissipation occurs.
- Figure (b) shows a CVSL AND/NAND gate.

Advantage:

- CVSL has a potential speed advantage because all of the logic is performed with nMOS transistors, thus reducing the input capacitance.

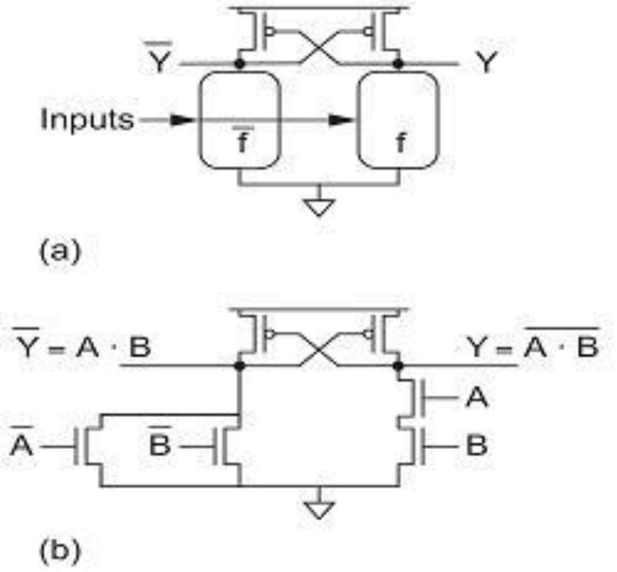


Figure: CVSL gates

****

2.6:Dynamic CMOS design:

Describe the basic principle of operation of dynamic CMOS, domino and NP domino logic with neat diagrams. (NOV 2011)

Dynamic Circuits:

Write short notes on Dynamic CMOS circuits. (Nov 2016)

- Ratioed circuits reduce the input capacitance by replacing the pMOS transistors connected to the inputs with a single resistive pullup.
- The drawbacks of ratioed circuits include
    - Slow rising transitions,
    - Contention on the falling transitions,
    - Static power dissipation and a nonzero V_{OL}.
- Dynamic circuits avoid these drawbacks by using a clocked pullup transistor rather than a pMOS that is always ON.
- Figure compares (a) static CMOS, (b) pseudo-nMOS, and (c) dynamic inverters.

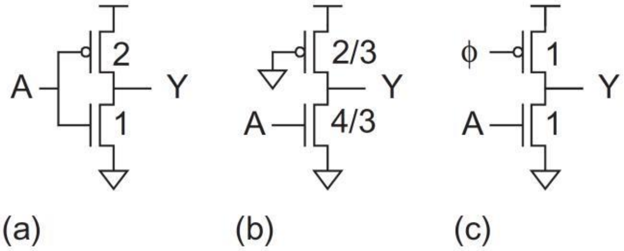


Figure: Comparison of (a) static CMOS, (b) pseudo-nMOS, and (c) dynamic inverters

- Dynamic circuit operation is divided into two modes, as shown in Figure.

  (i) During precharge, the clock ϕ is 0, so the clocked pMOS is ON and initializes the output Y high.

  (ii) During evaluation, the clock is 1 and the clocked pMOS turns OFF. The output may remain high or may be discharged low through the pulldown network.

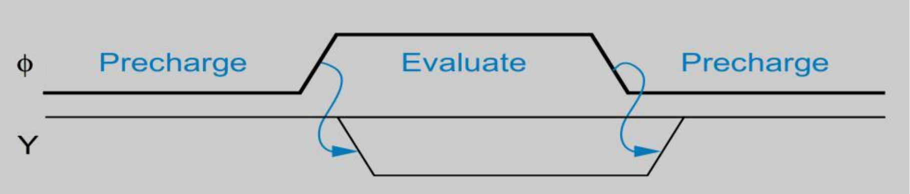


Figure: Precharge and evaluation of dynamic gates

Advantages:

- Dynamic circuits are the fastest used circuit family because they have lower input capacitance and no contention during switching.
- Zero static power dissipation.

Disadvantags:

- They require careful clocking, consume significant dynamic power and are sensitive to noise during evaluation mode.

Foot transistor:

- In Figure (c), if the input A is 1 during precharge, contention will take place because both the pMOS and nMOS transistors will be ON.
- When the input cannot be guaranteed to be 0 during precharge, an extra clocked evaluation transistor can be added to the bottom of the nMOS stack.
- To avoid contention as shown in the below figure, extra transistor is sometimes called as foot is added.

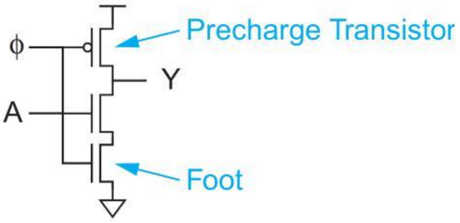


Figure: Footed dynamic inverter

- The given below figure shows generic footed and unfooted gates.

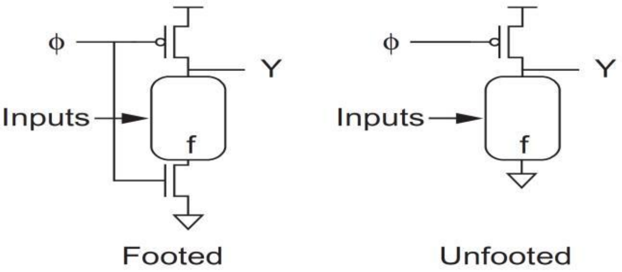


Figure: Generalized footed and unfooted dynamic gates

- The given below figure estimates the falling logical effort of both footed and unfooted dynamic gates.

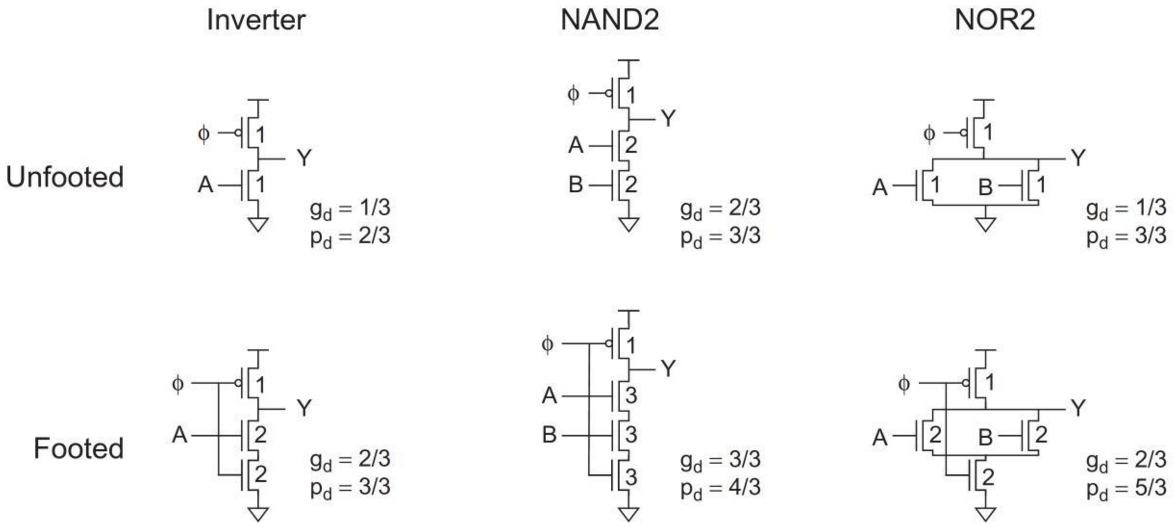


Figure: List of dynamic gates

- The pull down transistor‟s width is chosen to give unit resistance. Precharge occurs while the gate is idle and takes place more slowly.
- Therefore, the precharge transistor width is chosen for twice unit resistance.
- This reduces the capacitive load on the clock and the parasitic capacitance at the expense of greater rising delays.

- Footed gates have higher logical effort than their unfooted concept but are still an improvement over static logic.
- The parasitic delay does increase with the number of inputs, because there is more diffusion capacitance on the output node.
- A fundamental difficulty with dynamic circuits is the monotonicity requirement. While a dynamic gate is in evaluation, the inputs must be monotonically rising.
- That is, the input can start LOW and remain LOW, start LOW and rise HIGH, start HIGH and remain HIGH, but not start HIGH and fall LOW.
- Figure shows waveforms for a footed dynamic inverter in which the input violates monotonicity.

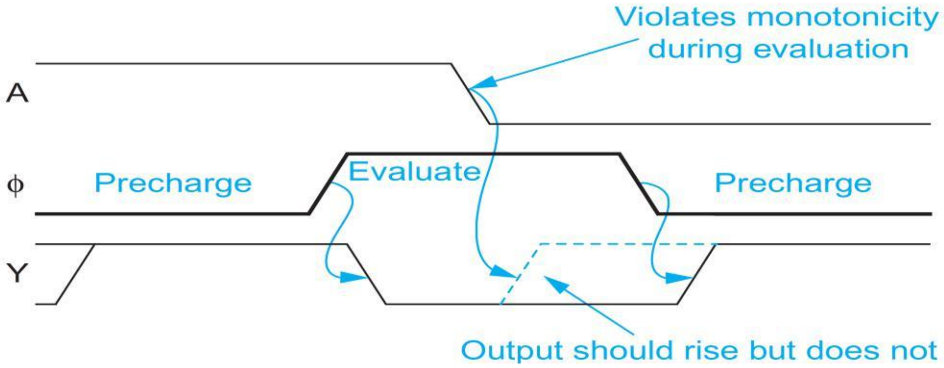


Figure: Monotonicity problem

- During precharge, the output is pulled HIGH.
- When the clock rises, the input is HIGH, so the output is discharged LOW through the pulldown network.
- The input later falls LOW, turning off the pulldown network. However, the precharge transistor is also OFF, so the output floats, staying LOW rather than rising.
- The output will remain low until the next precharge step.

- The inputs must be monotonically rising for the dynamic gate to compute the correct function.
- Unfortunately, the output of a dynamic gate begins HIGH and monotonically falls LOW during evaluation.
- This monotonically falling output X is not a suitable input to a second dynamic gate expecting monotonically rising signals, as shown in the below figure.
- Dynamic gates sharing the same clock cannot be directly connected.
- This problem is often overcome with domino logic.

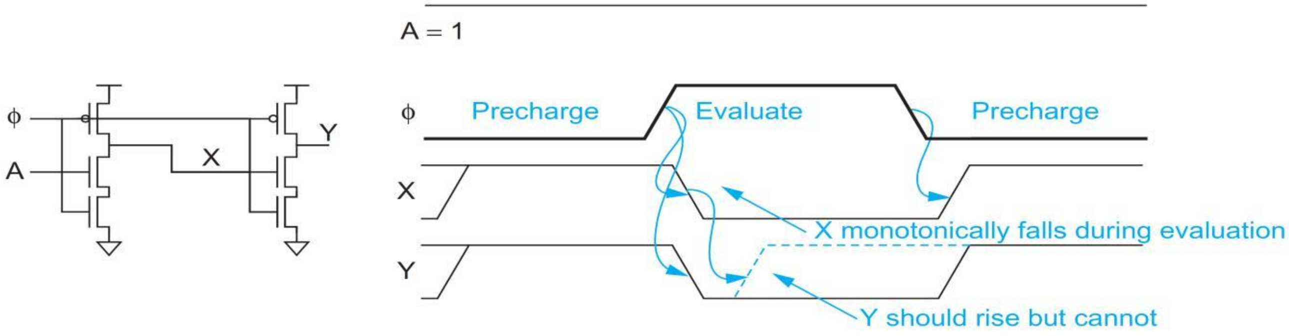


Figure: Incorrect connection of dynamic gates

2.6.1: Domino logic

Explain the domino logic families with neat diagrams. (NOV 2012, APRIL-2015, Nov 2017)

- The dynamic-static pair together is called a domino gate.
- The monotonicity problem can be solved by placing a static CMOS inverter between dynamic gates, as shown in figure (a).
- This converts the monotonically falling output into a monotonically rising signal suitable for the next gate, as shown in figure (b).
- A single clock can be used to precharge and evaluate all the logic gates within the chain.
- The dynamic output is monotonically falling during evaluation, so the static inverter output is monotonically rising.
- Therefore, the static inverter is usually a HI-skew gate to favor this rising output. Observe that precharge occurs in parallel, but evaluation occurs sequentially.

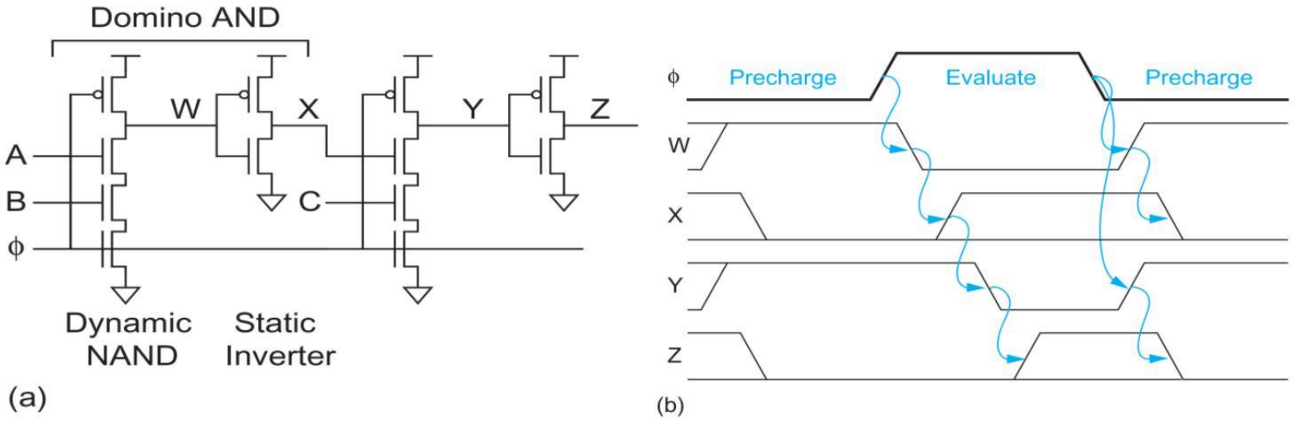


Figure: Domino gates

2.6.2: Dual Rail Domino Logic:

- Dual-rail domino gates encode each signal with a pair of wires. The input and output signal pairs are denoted with _h and _l, respectively.
- Table summarizes the encoding. The _h wire is asserted to indicate that the output of the gate is “high” or 1. The _l wire is asserted to indicate that the output of the gate is “low” or 0.
- When the gate is precharged, neither _h nor _l is asserted. The pair of lines should never be both asserted simultaneously during correct operation.

sig_h	*sig_l*	**Meaning**
0	0	Precharged
0	1	‘0’
1	0	‘1’
1	1	Invalid

Table: Dual-rail domino signal encoding

- Dual-rail domino gates accept both true and complementary inputs and compute both true and complementary outputs, as shown in Figure (a).
- This is identical to static CVSL circuits except that the cross-coupled pMOS transistors are instead connected to the precharge clock.
- Therefore, dual-rail domino can be viewed as a dynamic form of CVSL, sometimes called DCVS.
- Figure (b) shows a dual-rail AND/NAND gate and Figure (c) shows a dual-rail XOR/XNOR gate. The gates are shown with clocked evaluation transistors, but can also be unfooted.

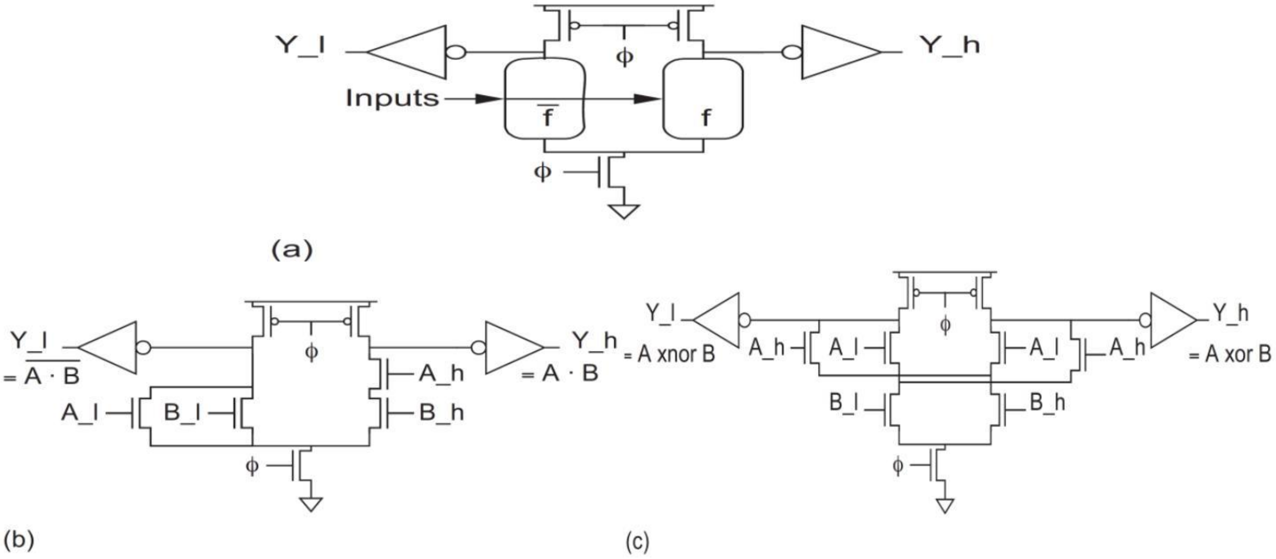


Figure: Dual-rail domino gates

Disadvantages:

- It requires more area, wiring and power.
- Dual-rail structures lose the efficiency of wide dynamic NOR gates.

Application:

- It is useful for asynchronous circuits.

2.6.3: Keepers

Explain the keeper logic family with neat diagrams.

- Dynamic circuits also suffer from charge leakage on the dynamic node.
- If a dynamic node is precharged high and then left floating, the voltage on the dynamic node will drift over time due to subthreshold, gate and junction leakage.
- Dynamic circuits have poor input noise margins.
- If the input rises above $V_{t,}$, while the gate is in evaluation, the input transistors will turn ON weakly and can incorrectly discharge the output.
- Both leakage and noise margin problems can be addressed by adding a keeper circuit.
- Figure shows a conventional keeper on a domino buffer. The keeper is a weak transistor that holds, or staticizes, the output at the correct level when it would otherwise float.
- When the dynamic node X is high, the output Y is low and the keeper is ON to prevent X from floating.

- When X falls, the keeper initially opposes the transition, so it must be much weaker than the pulldown network.
- Eventually Y rises, turning the keeper OFF and avoiding static power dissipation.

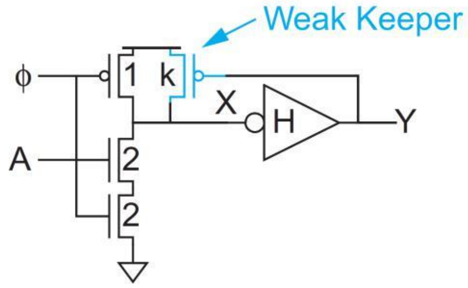


Figure: Conventional keeper

- The keeper must be strong enough to compensate for any leakage current drawn when the output is floating and the pulldown stack is OFF.
- Strong keepers also improve the noise margin, because when the inputs are slightly above V_t, the keeper can supply enough current to hold the output high.

2.6.3.1: *Differential keeper:*

- Figure shows a differential keeper for a dual-rail domino buffer.
- When the gate is precharged, both keeper transistors are OFF and the dynamic outputs float. As one of the rails evaluates low, the opposite keeper turns ON.
- The differential keeper is fast, because it does not oppose the falling rail.
- As long as one of the rails is guaranteed to fall promptly, the keeper on the other rail will turn on before excessive leakage or noise causes failure.

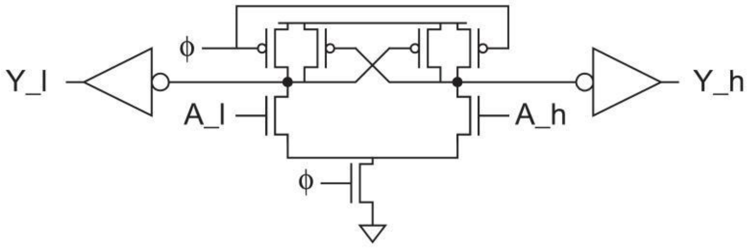


Figure: Differential keeper

2.6.4: *Secondary precharge devices*

- Dynamic gates are subject to problems with charge sharing.
- For example, consider the 2-input dynamic NAND gate in Figure (a). Suppose the output Y is precharged to V_{DD} and inputs A and B are low.

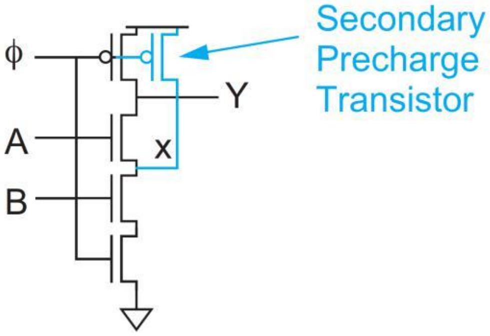


Figure: Secondary prechagre transistor

- Also suppose that the intermediate node x had a low value from a previous cycle.
- During evaluation, input A rises, but input B remains low, so the output Y should remain high.
- However, charge is shared between C_X and C_Y, shown in Figure (b). This behaves as a capacitive voltage divider and the voltages equalize at

$$V_x = V_Y = \frac{C_Y}{C_x + C_Y} V_{DD}$$

2.6.5: Charge sharing:

- Charge sharing is serious when the output is lightly loaded (small C_Y) and the internal capacitance is large.
- If the charge-sharing noise is small, the keeper will eventually restore the dynamic output to V_{DD}.
- If the charge-sharing noise is large, the output may flip and turn off the keeper, leading to incorrect results.
- Charge sharing can be overcome by precharging some or all of the internal nodes with secondary precharge transistors.
- These transistors should be small, because they only charge the small internal capacitances and their diffusion capacitance slows the evaluation.
- It is sufficient to precharge every other node in a tall stack.

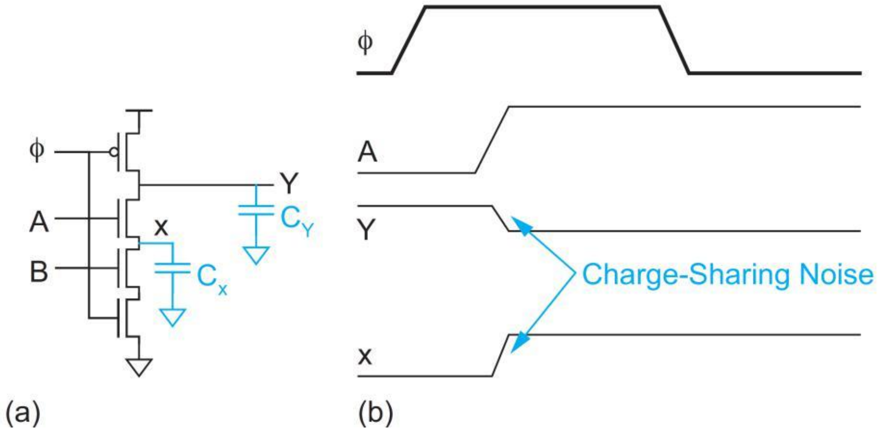


Figure: Charge-sharing noise

2.6.6: NP and Zipper Domino

Describe the basic principle of operation of NP domino logic. (NOV 2011)

- The HI-skew inverting static gates are replaced with predischarged dynamic gates using pMOS logic.
- A footed dynamic p-logic NAND gate is shown in Figure (b). When ϕ is 0, the first and third stages precharge high while the second stage predischarges low.
- When ϕ rises, all the stages evaluate. Domino connections are possible, as shown in Figure (c).

- The design style is called NP Domino or NORA Domino (NO RAce).

- NORA has two major drawbacks.
  - (i) The logical effort of footed p-logic gates is worse than that of HI-skew gates.
  - (ii) NORA is extremely susceptible to noise.

- In an ordinary dynamic gate, the input has a low noise margin (about V_t), but is strongly driven by a static CMOS gate.
- The floating dynamic output is more prone to noise from coupling and charge sharing, but drives another static CMOS gate with a larger noise margin.
- In NORA, however, the sensitive dynamic inputs are driven by noise prone dynamic outputs.
- Besides drawback and the extra clock phase requirement, there is little reason to use NORA.
- Zipper domino is a closely related technique, that leaves the precharge transistors slightly ON during evaluation by using precharge clocks. This swing between 0 and $V_{DD} - |V_{tp}|$ for the pMOS precharge and V_{tn} and V_{DD} for the nMOS precharge.

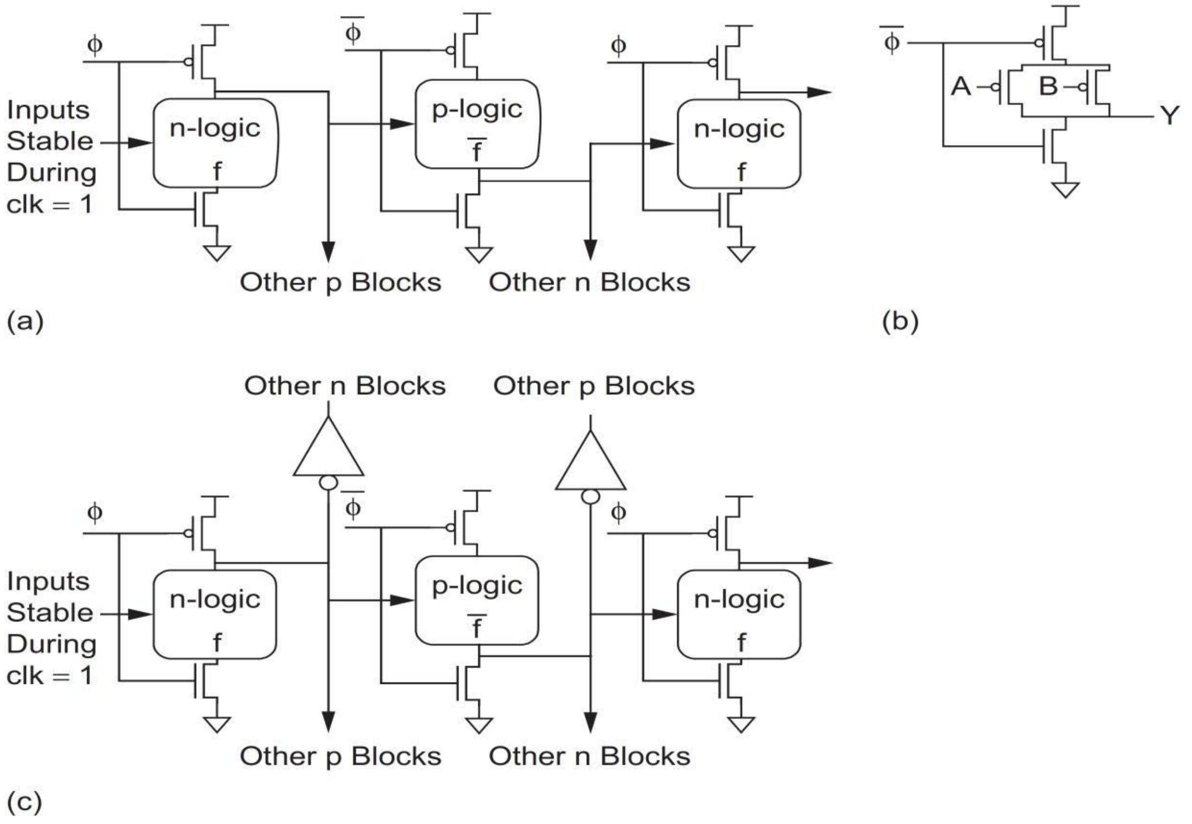


Figure : NP Domino

****

2.7: Power dissipation:

- ❖ **Explain the static and dynamic power dissipation in CMOS circuits with necessary diagrams and expressions. (DEC 2011, Nov 2015, NOV 2016, May 2017, May 2010)**
- ❖ **What are the sources of power dissipation in CMOS and discuss various design techniques to reduce power dissipation in CMOS? (Nov 2012, May 2013, Nov 2014, May 2016)**

- The instantaneous power P (t) consumed by a circuit element is the product of the current and the voltage of the element

$$P(t) = I(t)V(t)$$

- The energy consumed over time interval T is the integral of the instantaneous power

$$E = \int_0^T P(t)\,dt$$

- The average power is $P_{avg} = \frac{E}{T} = \frac{1}{T}\int_0^T P(t)\,dt$

  Power is expressed in units of Watts (W). Energy is usually expressed in Joules (J)

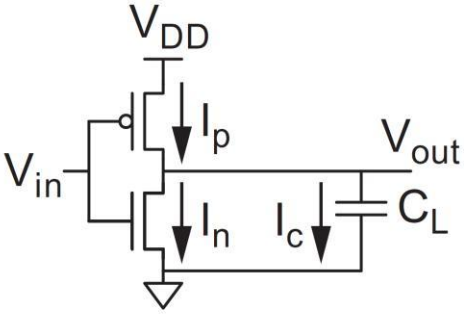
  - By Ohm‟s Law, V = IR, so the instantaneous power dissipated in the resistor is

$$P_R(t) = \frac{V_R^2(t)}{R} = I_R^2(t)\,R$$

- This power is converted from electricity to heat. V_{DD} supplies power proportional to its current $P_{VDD}(t) = I_{DD}(t)\,V_{DD}$
- When the capacitor is charged from 0 to V_C, it stores energy E_C

$$E_C = \int_0^\infty I(t)V(t)\,dt = \int_0^\infty C\frac{dV}{dt}V(t)\,dt = C\int_0^{V_C} V(t)\,dV = \tfrac{1}{2}CV_C^2$$

- Figure shows a CMOS inverter driving a load capacitance.



- When the input switches from 1 to 0, the pMOS transistor turns ON and charges the load to V_{DD}.
- According to E_C equation the energy stored in the capacitor is

$$E_C = \tfrac{1}{2}C_L V_{DD}^2$$

- The energy delivered from the power supply is

$$E_C = \int_0^\infty I(t)V_{DD}\,dt = \int_0^\infty C\frac{dV}{dt}V_{DD}\,dt = CV_{DD}\int_0^{V_{DD}} dV = CV_{DD}^2$$

- Gate switches at some average frequency f_{sw}.

- Over some interval T, the load will be charged and discharged Tf_{sw} times.
- Then, the average power dissipation is

$$P_{\text{switching}} = \frac{E}{T} = \frac{Tf_{\text{sw}}CV_{DD}^2}{T} = CV_{DD}^2 f_{\text{sw}}$$

- This is called the dynamic power because it arises from the switching of the load.
- Because most gates do not switch every clock cycle, it is often more convenient to express switching frequency f_{sw} as an activity factor α times the clock frequency f.
- The dynamic power dissipation may be rewritten as

$$P_{\text{switching}} = \alpha C V_{DD}^2 f$$

- The activity factor is the probability that the circuit node transitions from 0 to 1, because that is the only time the circuit consumes power.
- A clock has an activity factor of α = 1 because it rises and falls every cycle.
- The total power of a circuit is calculated as,

$$P_{\text{dynamic}} = P_{\text{switching}} + P_{\text{short circuit}}$$

$$P_{\text{static}} = \left(I_{\text{sub}} + I_{\text{gate}} + I_{\text{junct}} + I_{\text{contention}}\right) V_{DD}$$

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

2.7.1: Dynamic power:

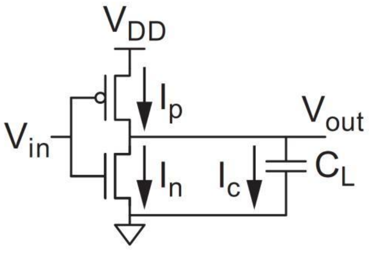
- Dynamic power consists mostly of the switching power.

$$P_{\text{switching}} = \alpha C V_{DD}^2 f$$

- The supply voltage V_{DD} and frequency f are known by the designer.
- To estimate dynamic power, one can consider each node of the circuit.
- The capacitance of the node is the sum of the gate, diffusion, and wire capacitances on the node.
- The activity factor can be estimated using switching probability or measured from logic simulations.
- The effective capacitance of the node is, its true capacitance multiplied by the activity factor.
- The switching power depends on the sum of the effective capacitances of all the nodes.

2.7.1.1: Sources of dynamic power dissipation:

- Dynamic dissipation due to
    - Charging and discharging load capacitances as gates switchs.
    - "Short-circuit" current while both pMOS and nMOS stacks are partially ON



2.7.1.2: Reducing dynamic power dissipation:

❖ **Explain various ways to minimize the static and dynamic power dissipation. (Nov 2013, May 2015)**

❖ **Discuss the low power design principles in detail. (Nov 2017)**

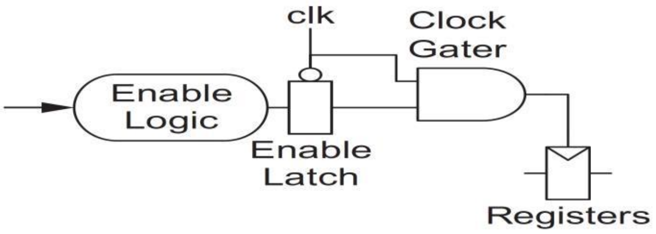
- Low power design involves considering and reducing each of the terms in switching power.
  i. As V_{DD} is a quadratic term, it is good to select the minimum V_{DD}.
  ii. Choose the lowest frequency.
  iii. The activity factor is reduced by putting unused blocks to sleep.
  iv. Finally, the circuit may be optimized to reduce the overall load capacitance.
- Switching power is consumed by delivering energy to charge a load capacitance, then dumping this energy to GND.

Activity factor:

- If a circuit can be turned OFF entirely, the activity factor and dynamic power go to zero.
- Blocks are typically turned OFF, by stopping the clock called as clock gating.
- The activity factor of a logic gate can be estimated by calculating the switching probability.

(a)Clock gating:

- Clock gating, AND‟s a clock signal with an enable to turn OFF the clock to idle blocks.
- The clock enable must be stable, while the clock is active.
- Figure shows how an enable latch can be used to ensure the enable does not change before the clock falls.



Capacitance:

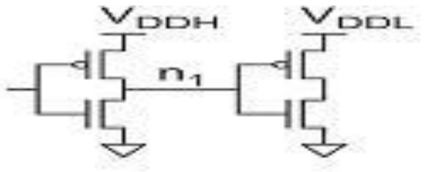
- Switching capacitance comes from the wires and transistors in a circuit.
- Wire capacitance is minimized through good floor planning and placement.
- Device-switching capacitance is reduced by choosing smaller transistors.

Voltage:

- Voltage has a quadratic effect on dynamic power.
- Therefore, choosing a lower power supply significantly reduces power consumption.
- The chip may be divided into multiple voltage domains, where each domain is optimized for the needs of certain circuits.

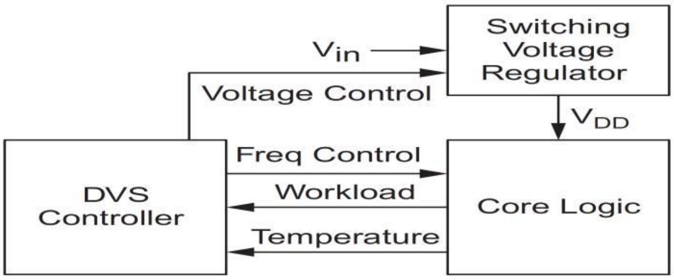
a. Voltage domains:

- Selecting, which circuits belong in which domain and routing power supplies to multiple domains.
- Figure **(Voltage domain crossing)** shows direct connection of inverters in two domains using high and low supplies, V_{DDH} and V_{DDL}, respectively.



b. Dynamic voltage scaling (DVS):

- Systems can save large amounts of energy by reducing the clock frequency, then reducing the supply voltage.
- This is called dynamic voltage scaling (DVS) or dynamic voltage/frequency scaling (DVFS).



- Figure shows a block diagram for a basic DVS system.
- It determines the supply voltage and clock frequency sufficient to complete the workload on schedule or to maximize performance without overheating.

- Dynamic power is directly proportional to frequency, so a chip should not run faster than necessary.
- Reducing the frequency allows downsizing transistors or using a lower supply voltage.

2.7.2: Static power:

- Static power is consumed even when a chip is not switching.
- Static CMOS gates have no contention current.

2.7.2.1: Sources of static power dissipation:

- Static dissipation due to
  - Subthreshold leakage through OFF transistors.
  - Gate leakage through gate dielectric.
  - Junction leakage from source/drain diffusions.
  - Contention current in ratioed circuits.

$$P_{static} = (I_{sub} + I_{gate} + I_{junc} + I_{contention})V_{DD}$$

1. **Subthreshold leakage current:**
   - Subthreshold leakage current flows when a transistor is OFF.
   - Subthreshold leakage current equation is

$$I_{\text{sub}} = I_{\text{off}} 10^{\frac{V_{gs}+\eta(V_{ds}-V_{DD})-k_{\gamma}V_{sb}}{S}}$$

where I_{off} is the subthreshold current at $V_{gs} = 0$ and $V_{ds} = V_{DD}$, and S is the subthreshold slope.

2. **Gate leakage:**

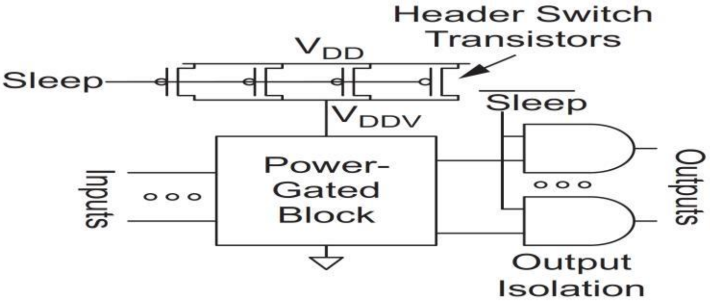
- Gate leakage occurs when carriers tunnel through a thin gate dielectric, when a voltage is applied across the gate (e.g., when the gate is ON).
- Gate leakage is a strong function of the dielectric thickness.

3. **Junction leakage:**
   - Junction leakage occurs when a source or drain diffusion region is at a different potential from the substrate.
   - Leakage of reverse-biased diodes is usually negligible.
4. **Contention current:**
   - Static CMOS circuits have no contention current. However, certain alternative circuits inherently draw current even while quiescent.

2.7.2.2: Methods of reducing static power:

Power gating:

- To reduce static current during sleep mode is, to turn OFF the power supply to the sleeping blocks. This technique is called power gating.



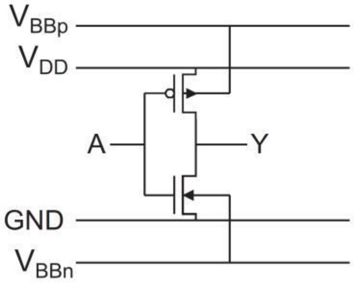
- The logic block receives its power from a virtual V_{DD} rail, V_{DDV}.
- When the block is active, the header switch transistors are ON, connecting V_{DDV} to V_{DD}.
- When the block goes to sleep, the header switch turns OFF, allowing V_{DDV} to float and gradually sink toward 0.

Multiple threshold voltage and oxide thickness:

- Selective application of multiple threshold voltages can maintain performance on critical paths with low-V_t transistors, while reducing leakage on other paths with high-V_t transistors.

Variable threshold voltage:

- Method to achieve high I_{on} in active mode and low I_{off} in sleep mode is, by adjusting the threshold voltage of the transistor by applying a body bias.
- This technique is sometimes called variable threshold CMOS (VTCMOS).
- Figure shows a schematic of an inverter using body bias.



****

UNIT – III

SEQUENTIAL LOGIC CIRCUITS

Static and Dynamic Latches and Registers, Timing issues, pipelines, clock strategies, Memory architecture and memory control circuits, Low power memory circuits, Synchronous and Asynchronous design.

3.1 Static Latches and Registers:

- ❖ **Discuss in detail various static latches and registers. (Nov 2016)**
- ❖ **Explain the methodology of sequential circuit design of Latches. (May 2014)**
- ❖ **Discuss the operation of a CMOS latch. (Nov 2007)**

3.1.1 The Bi-stability Principle

- Static memories use positive feedback to create a bistable circuit. A bistable circuit has two stable states that represent 0 and 1.
- The basic idea is shown in Figure 3.1a, which shows two inverters connected in cascade along with a voltage-transfer characteristic (VTC).
- The output of the second inverter Vo2 is connected to the input of the first V i1, as shown by the dotted lines in Figure 3.1a.
- The resulting circuit has only three possible operation points (A, B, and C).
- A and B are stable operation points, and C is a metastable operation point.

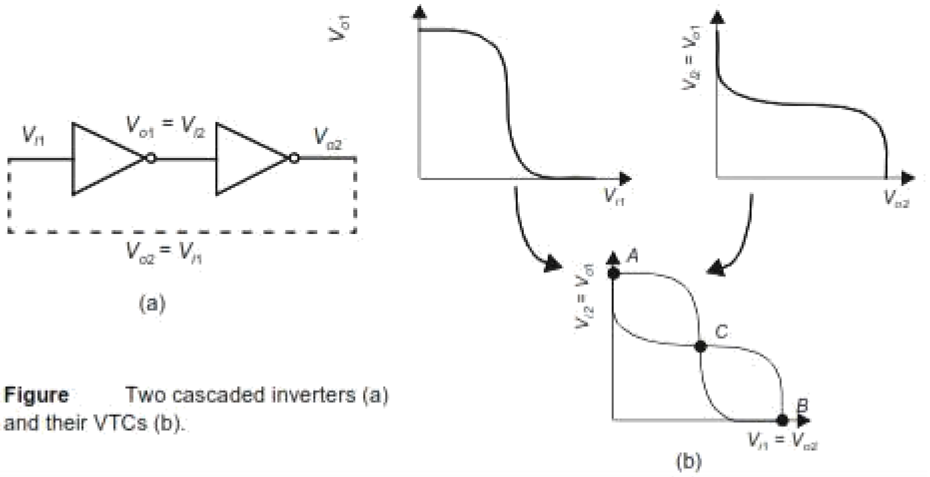
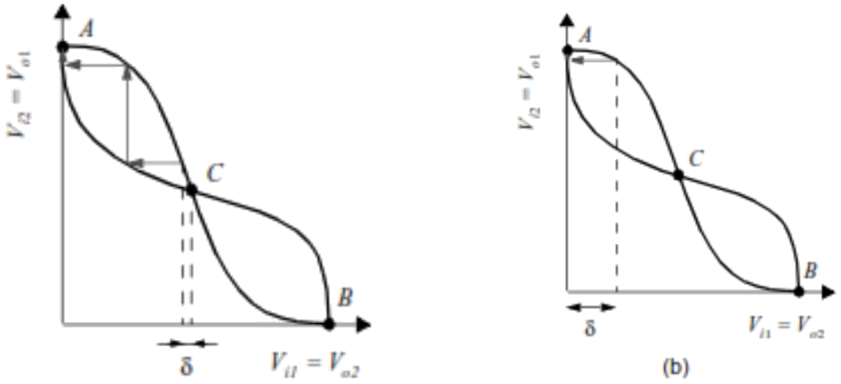


Figure Two cascaded inverters (a) and their VTCs (b).

Figure 3.1: a. Two inverters connected in cascade b. VTCs

- Cross-coupled inverter pair is biased at point C. It is amplified and regenerated around the circuit loop.
- The bias point moves away from C until one of the operation points A or B is reached.
- C is an unstable operation point. Every deviation causes the operation point to run away from its original bias. Operation points with this property are termed as metastable.

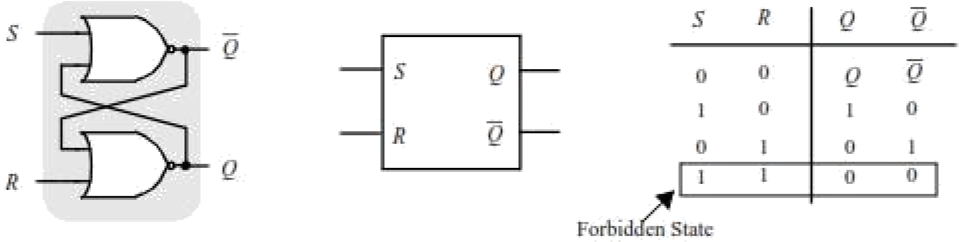


(a) Figure 3.2 Metastability.

- A bistable circuit has two stable states. In absence of any triggering, the circuit remains in a single state.
- A trigger pulse must be applied to change the state of the circuit.
- Common name for a bistable circuit is flip-flop.

3.1.2 SR Flip-Flops

- The SR or set-reset flip-flop implementation is shown in Figure (a) below.
- This circuit is similar to the cross-coupled inverter pair with NOR gates replacing the inverters.
- The second input of the NOR gates is connected to the trigger inputs (S and R), that make it possible to force the outputs Q and Q_{bar}.
- These outputs are complimentary (except for the SR = 11 state).
- When both S and R are 0, the flip-flop is in a quiescent state and both outputs retain their value.
- If a positive (or 1) pulse is applied to the S input, the Q output is forced into the 1 state (with Qbar going to 0).
- Vice versa, a 1 pulse on R resets the flip-flop and the Q output goes to 0.



S	R	Q	$\overline{Q}$
0	0	Q	$\overline{Q}$
1	0	1	0
0	1	0	1
1	1	0	0

(a) Schematic diagram (b) Logic symbol (c) Characteristic table

Figure 3.3

- When both S and R are high, both Q and Q_{bar} are forced to zero. This input mode is considered to be forbidden.
- An SR flip-flop can be implemented using a cross-coupled NAND structure as shown in Figure 3.4

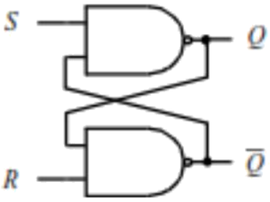


Figure 3.4 NAND based SR flip-flop

Clocked SR flip-flop:

- Clocked SR flip-flop (a level-sensitive positive latch) is shown in Figure 3.5.
- It consists of a cross-coupled inverter pair, plus 4 extra transistors to drive the flip-flop from one state to another and to provide clocked operation.
- Consider the case where Q is high and R pulse is applied.
- The combination of transistors M_4, M_7, and M_8 forms a ratioed inverter.
- In order to make the latch switch, we must succeed in bringing Q below the switching threshold of the inverter M_1-M_2.
- Once this is achieved, the positive feedback causes the flip-flop to invert states. This requirement forces to increase the sizes of transistors M_5, M_6, M_7, and M_8.

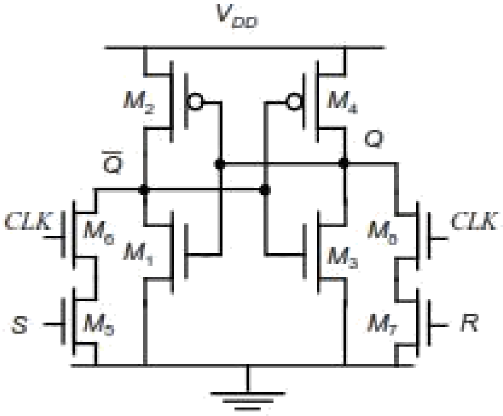


Figure 3.5 CMOS clocked SR flip-flop

- The clocked SR flip-flop does not consume any static power.

3.1.3 Multiplexer Based Latches:

- Multiplexer based latches can provide similar functionality to the SR latch.
- But sizing of devices only affects performance and is not critical to the functionality.
- Figure 3.6 shows an implementation of static positive and negative latches based on multiplexers.
- For a negative latch, when the clock signal is low, the input 0 of the multiplexer is selected, and the D input is passed to the output.
- When the clock signal is high, the input 1 of the multiplexer connected to the output of the latch.
- The feedback holds the output stable while the clock signal is high.
- Similarly in the positive latch, the D input is selected when clock is high and the output is held (using feedback) when clock is low.

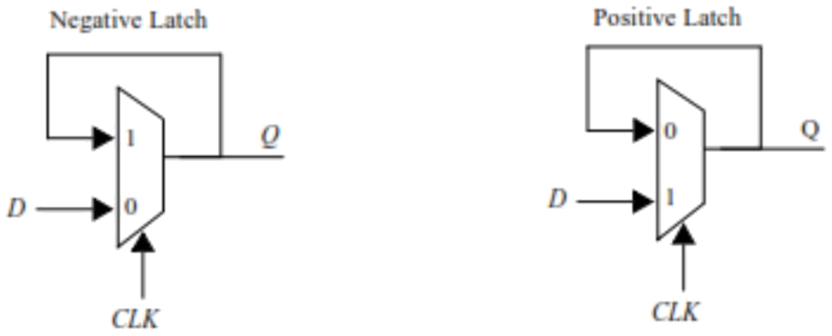


Figure 3.6 Negative and positive latches based on multiplexers.

Design a d-latch using transmission gate. (May 2015)

- A transistor level implementation of a positive latch is shown in Figure 3.7.
- When CLK is high, the bottom transmission gate is ON and the latch is transparent – i.e, the D input is copied to the Q output.
- During this phase, the feedback loop is open due to the top transmission gate is OFF.

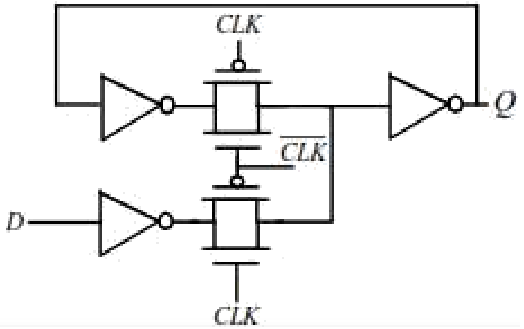
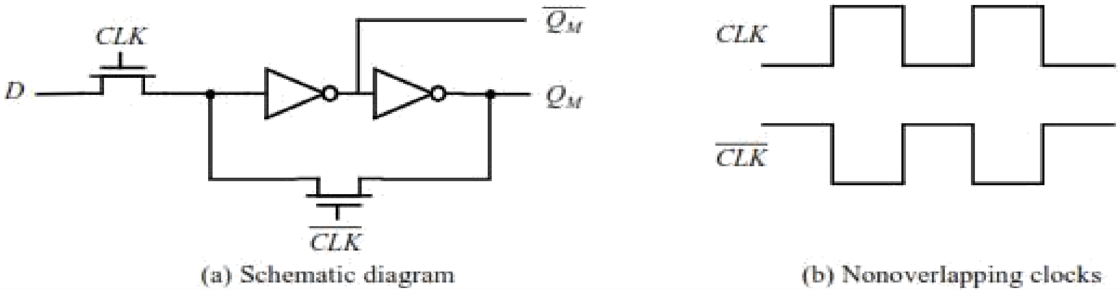


Figure 3.7 Transistor level implementation of a positive latch built using transmission gates.

- To reduce the clock load, implement a multiplexer based NMOS latch using two pass transistors as shown in Figure 3.8.
- The advantage of this approach is the reduced clock load of only two NMOS devices.
- When CLK is high, the latch samples the D input, while a low clock-signal enables the feedback-loop and puts the latch in the hold mode.



(a) Schematic diagram

(b) Nonoverlapping clocks

Figure 3.8 Multiplexer based NMOS latch using NMOS only pass transistors for multiplexers.

3.1.4 Master-Slave Based Edge Triggered Register:

- **Explain the operation of master-slave based edge triggered register. (May 2016)**
- **Draw and explain the operation of conventional CMOS, pulsed and resettable latches. (Nov 2012)**

- An edge-triggered register is to use a master-slave configuration as shown in Figure 3.9.
- The register consists of cascading a negative latch (master stage) with a positive latch (slave stage).
- A multiplexer based latch is used to realize the master and slave stages.
- On the low phase of the clock, the master stage is transparent and the D input is passed to the master stage output, Q_M.
- During this period, the slave stage is in the hold mode, keeping its previous value.
- On the rising edge of the clock, the master slave stops sampling the input and the slave stage starts sampling.
- During the high phase of the clock, the slave stage samples the output of the master stage (Q_M), while the master stage remains in a hold mode.
- A negative edge-triggered register can be constructed using the same principle by simply switching the order of the positive and negative latch (i.e., placing the positive latch first).

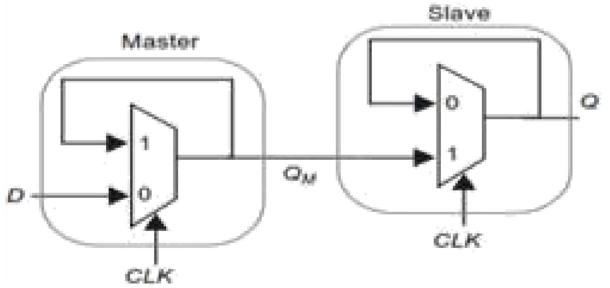


Figure 3.9: Positive edge-triggered register based on a master-slave configuration.

- A complete transistor level implementation of the master-slave positive edge-triggered register is shown in Figure 3.10.
- When clock is low (CLK bar = 1), T_1 is ON and T_2 is OFF and the D input is sampled onto node Q_M.
- During this period, T_3 is OFF and T_4 is ON and the cross-coupled inverters (I_5 & I_6) hold the state of the slave latch.

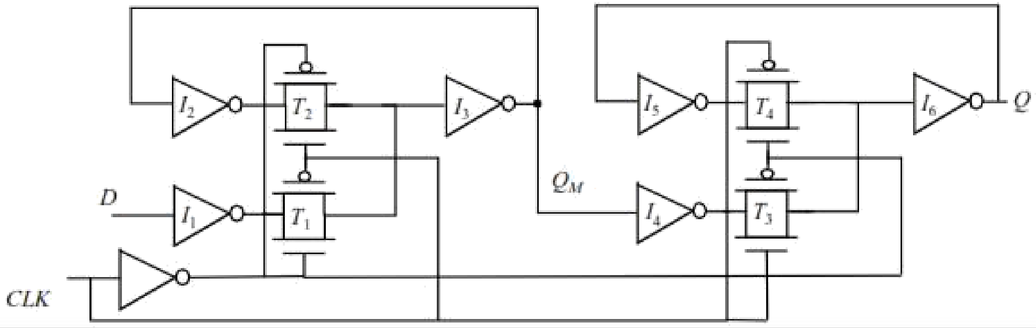
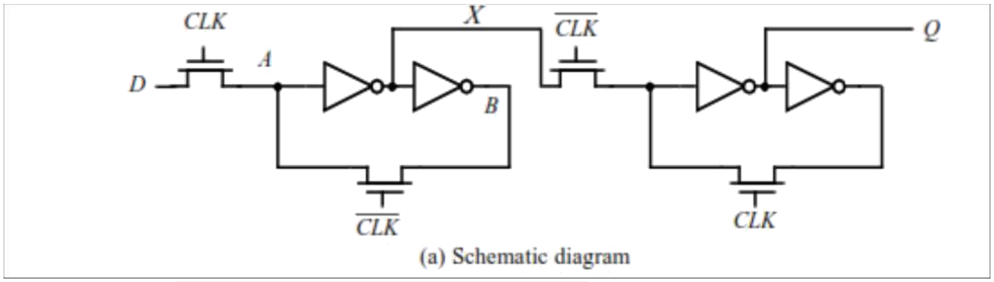


Figure 3.10 Transistor-level implementation of a master-slave positive edge-triggered register using multiplexers.

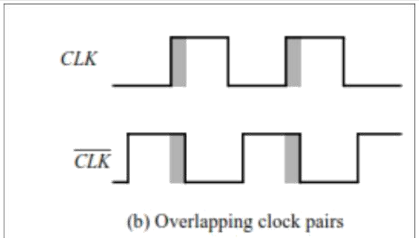
- When the clock goes high, the master stage stops sampling the input and goes into a hold mode.
- T_1 is OFF and T_2 is ON and the cross coupled inverters I_3 and I4 hold the state of Q_M. Also T_3 is ON and T_4 is OFF and Q_M is copied to output Q.

3.1.5 Non-ideal clock signals:

- We have assumed that $\overline{CLK}$ is a perfect inversion of CLK.
- Even if this was possible, this would still not be a good assumption.
- Variations can exist in the wires. It is used to route the two clock signals or the load capacitances can vary based on data stored in the connecting latches.
- This effect, known as clock skew is a major problem and causes the two clock signals to overlap as is shown in Figure 3.11 b.
- Clock-overlap can cause two types of failures, as illustrated for the NMOS-only negative master-slave register of Figure 3.11 a.



(a) Schematic diagram



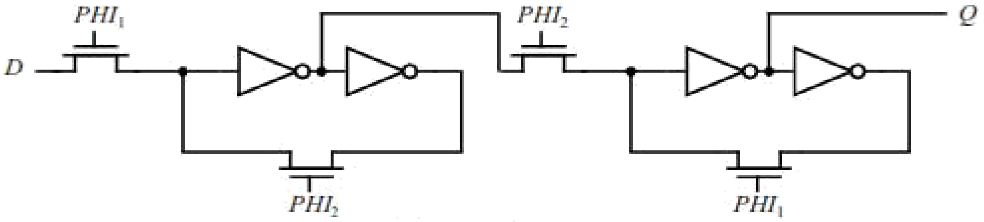
(b) Overlapping clock pairs

Figure 3.11 Master-slave register based on NMOS-only pass transistors.

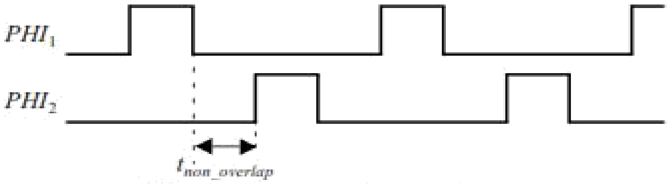
- When the clock goes high, the slave stage should stop sampling the master stage output and go into a hold mode.
- However, since CLK and $\overline{CLK}$ are both high for a short period of time (the overlap period).

- Both sampling pass transistors conduct and there is a direct path from the D input to the Q output.
- As a result, data at the output can change on the rising edge of the clock, which is undesired for a negative edge triggered register.
- The is known as a race condition in which the value of the output Q is a function of whether the input D arrives at node X before or after the falling edge of $\overline{CLK}$.
- If node X is sampled in the metastable state, the output will switch to a value determined by noise in the system.
- Those problems can be avoided by using two non-overlapping clocks PHI and PH2.
- By keeping the non-overlap time $t_{non\_overlap}$ between the clocks large enough, such that no overlap occurs even in the presence of clock-routing delays.

- During the non-overlap time, the FF is in the high-impedance state.
- Leakage will destroy the state, if this condition holds for too long a time.
- Hence the name ***pseudostatic***: the register employs a combination of static and dynamic storage approaches depending upon the state of the clock.



(a) Schematic diagram



(b) Two-phase nonoverlapping clocks

Figure 3.12 Pseudostatic two-phase D register.

3.1.6 Low-Voltage Static Latches:

- The scaling of supply voltages is critical for low power operation.
- At very low power supply voltages, the input to the inverter cannot be raised above the switching threshold, resulting in incorrect evaluation.
- Scaling to low supply voltages, hence requires the use of reduced threshold devices.
- Multiple Threshold devices as shown in Figure 3.13.
- The shaded inverters and transmission gates are implemented in low-threshold devices.
- The low threshold inverters are gated using high threshold devices to eliminate leakage.
- During normal mode of operation, the sleep devices are tuned on.
- During idle mode, the high threshold devices in series with the low threshold inverter are turned OFF (the SLEEP signal is high), eliminating leakage.

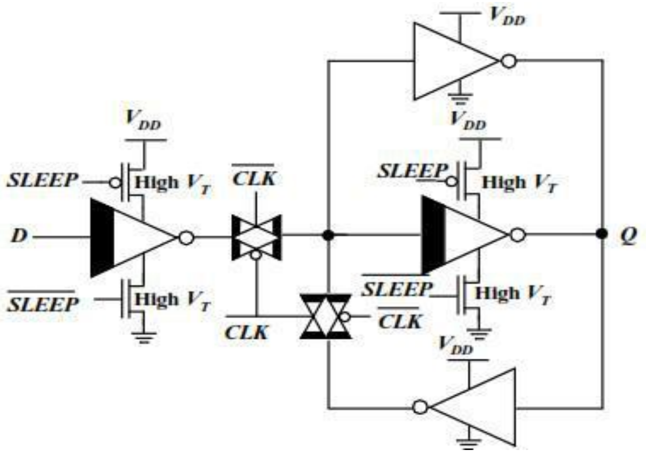


Figure 3.13: One solution for the leakage problem in low-voltage operation using MTCMOS.

****

3.2 Dynamic Latches and Registers:

❖ **Discuss about the design of sequential dynamic circuits. (Nov 2012, Nov 2017)**
Explain the methodology of sequential circuit design of flip-flop. (May 2014)

- A stored value remains valid as long as the supply voltage is applied to the circuit, hence the name static.
- The major disadvantage of the static gate is, its complexity.
- Registers are used in computational structures that are constantly clocked, such as pipelined data path.
- The requirement that the memory should hold state for extended periods of time.
- This results in circuits, based on temporary storage of charge on parasitic capacitors.
- The principle is identical to the dynamic logic. In dynamic logic, logic signal is a charge, stored on a capacitor.
- The absence of charge denotes as logic 0 and presence of charge denotes as logic 1.
- A stored value can be kept for a limited amount of time (range of milliseconds).
- A periodic refresh of its value is necessary.

3.2.1 Dynamic Transmission-Gate Based Edge-triggered Registers:

❖ **Design a d-flipflop using transmission gate. (Nov 2016)**

- A dynamic positive edge-triggered register based on the master-slave concept is shown in Figure 3.14.
- When CLK = 0, the input data is sampled on storage node 1. It has an equivalent capacitance of C1 consisting of the gate capacitance of I_1, the junction capacitance of T_1, and the overlap gate capacitance of T_1.
- During this period, the slave stage is in a hold mode with node 2 in a high-impedance state.
- On the rising edge of clock, the transmission gate T_2 turns on. The value is sampled on node1 before the rising edge propagates to the output Q.
- Node 2 stores the inverted version of node 1.
- The reduced transistor provides high-performance and low-power systems.

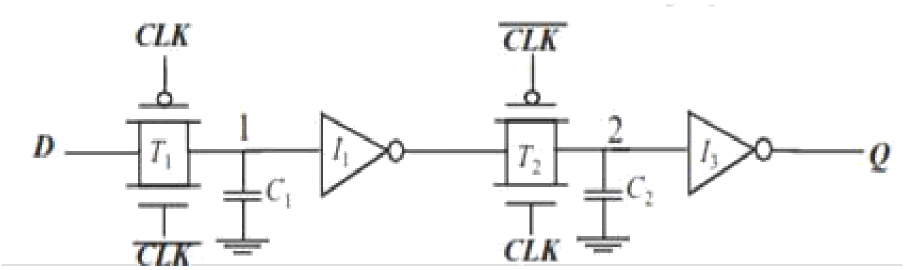


Figure 3.14 Dynamic edge-triggered register.

- The set-up time of this circuit is the delay of the transmission gate and time takes node 1 to sample the D input.
- The hold time is approximately zero, since the transmission gate is turned OFF.

- The propagation delay (t_{c-q}) is equal to two inverter delays plus the delay of the transmission gate T_2.
- Dynamic register has storage nodes (i.e., the state). A node has to be refreshed at periodic intervals to prevent a loss. Loss due to charge leakage and diode leakage.
- Clock overlap is an important for register. Consider the clock waveforms shown in Figure 3.15.
- During the 0-0 overlap period, the NMOS of T_1 and the PMOS of T_2 are simultaneously on.
- It is creating a direct path for data to flow from the D input of the register to the Q output. This is known as a race condition.
- The output Q can change on the falling edge, if the overlap period is large.
- Overlap period constraint is given as: $t_{overlab\,0-0} < t_{T1} + t_{I1} + t_{T2}$
- Similarly, the constraint for the 1-1 overlap is given as: $t_{hold} > t_{overlab1-1}$

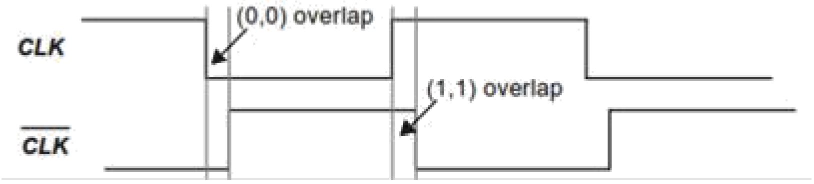


Figure 3.15 Impact of non-overlapping clocks

3.2.2 C^2MOS Dynamic Register:

The C^2MOS Register

- Figure 3.16 shows positive edge-triggered register based on the master-slave concept, which is insensitive to clock overlap. This circuit is called the C^2MOS (Clocked CMOS) register.

1. CLK = 0 ($\overline{CLK}$ = 1):
   - The first tri-state driver is turned ON. The master stage acts as an inverter, sampling the inverted of D on the internal node X.
   - The master stage is, in evaluation mode. The slave section is, in hold mode.
   - Both transistors M_7 and M_8 are OFF, decoupling the output from the input. The output Q retains its previous value stored on the output capacitor C_{L2}.

2. CLK = 1 ($\overline{CLK}$ = 0):
   - The master stage section is in hold mode (M_3-M_4 off), while the second section evaluates (M_7-M_8 on).
   - The value stored on C_{L1} propagates to the output node through the slave stage, which acts as an inverter.
   - The overall circuit operates as a positive edge-triggered master-slave register.

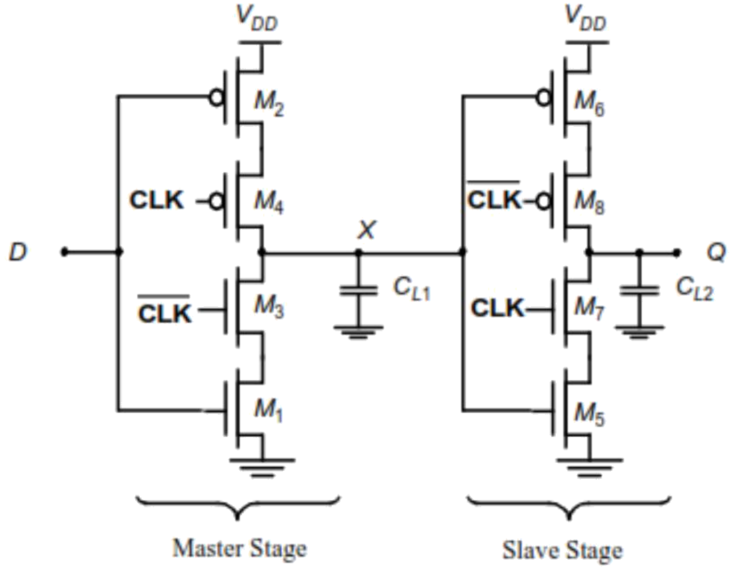


Figure 3.16 C^2MOS master-slave positive edge-triggered register.

- It is similar to the transmission-gate based register, presented earlier. However, there is an important difference.
- *A C^2MOS register with CLK-CLK clocking is insensitive to overlap, as long as the rise and fall times of the clock edges are small.*

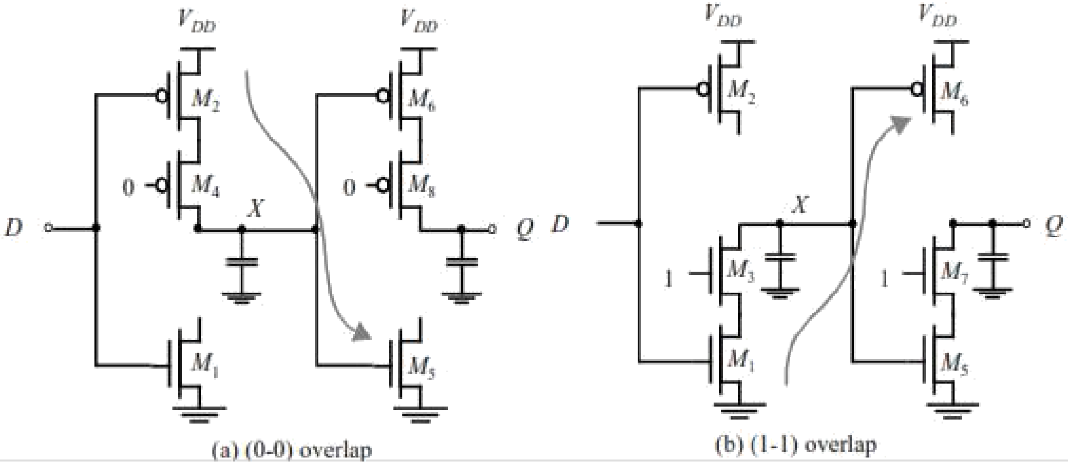


Figure 3.17 C^2MOS D FF during overlap periods.

3.2.3 True Single-Phase Clocked Register (TSPCR):

Explain the operation of True Single Phase Clocked Register. (Nov 2016, April 2017)

- In the two-phase clocking schemes, care must be taken in routing the two clock signals to ensure that overlap is minimized.
- While the C^2MOS provides a skew-tolerant solution, it is possible to design registers that only use a single phase clock.

- The True Single-Phase Clocked Register (TSPCR) uses a single clock without an inverse clock.
- Figure 3.19 shows positive and negative latch concept.
- For the positive latch, when CLK is high, the latch is in the transparent mode and propagates the input to the output. Latch has two cascaded inverters, so latch is non-inverting.
- When CLK = 0, both inverters are disabled and the latch is, in hold-mode.
- Only the pull-up networks are still active, while the pull-down circuits are deactivated.
- As a result of the dual-stage approach, no signal can ever propagate from the input to the output.
- For the negative latch, when CLK is low, the latch is in the transparent mode and propagates the input to the output.
- When CLK = 1, both inverters are disabled and the latch is in hold-mode.
- A register can be constructed by cascading positive and negative latches.
- The main advantage is the use of a single clock phase.
- The disadvantage is, increase in the number of transistors (12 transistors are required).

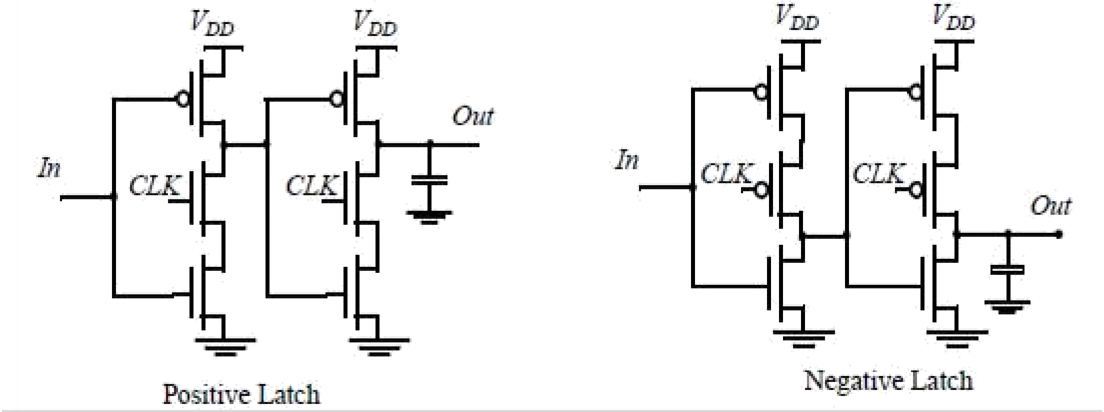


Figure 3.19 TSPC approach.

- The TSPC latch circuits can be reduced, as in Figure 3.20, where only the first inverter is controlled by the clock.
- Number of transistors are reduced and clock load is reduced by half.

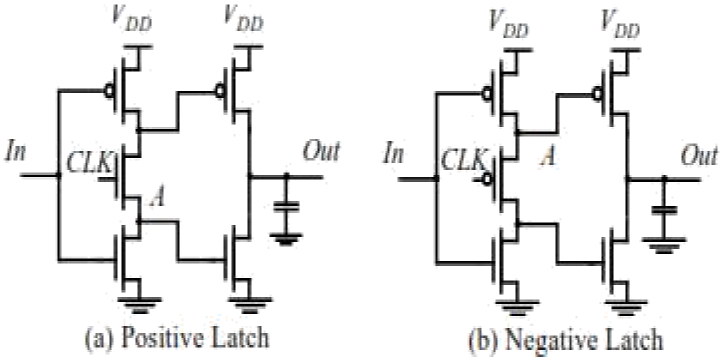


Figure 3.20 Simplified TSPC latch (also called split-output).

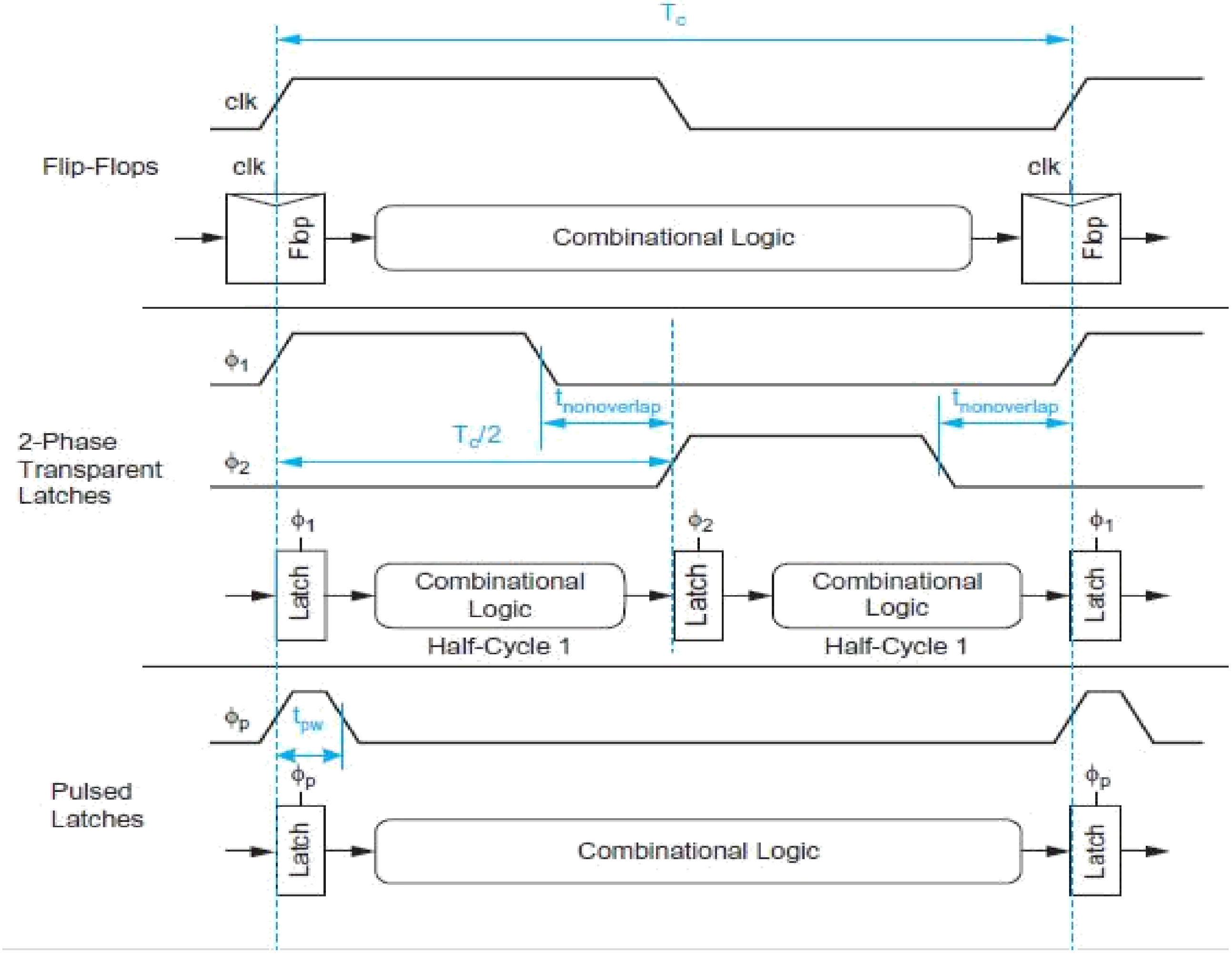
****

3.3 Timing Issues:

- **Explain in detail about timing issues needed for a logic operation. (April 2017)**
- **Explain the timing basics in synchronous design in detail. (Nov 2017)**

(A)Sequencing methods:-

- Three methods of sequencing block of combinational logic are possible, as shown in figure below.
- In flip-flop based system, one flip flop use one cycle boundary.
- Token (data) advances from one cycle to the next on the rising edge. If a token arrives too early, it waits at the flip flop until next cycle.
- In 2-phase system, phases may be separated by $t_{nonoverlap}$. [$t_{nonoverlap}>0$]
- In pulsed system, pulse with is t_{pw}.



- In 2-phase system, full cycle of combinational logic is divided into two phases, sometimes called “half-cycles”. Two latch clocks are called 1 and 2.
- Flip flop can be viewed as, a pair of back to back latches using clk and its complements.
- Table shows delay and timing notations of combinational and sequencing elements. These delays may differ for rising (with suffix „r‟) and falling (with suffix „f‟).

TERM	***NAME***
T_{pd}	logic propagation delay
T_{cd}	logic contamination delay
T_{pcq}	latch flop clock-Q propagation delay
T_{ccq}	latch flop clock- to Q contamination delay
T_{pdq}	latch flop D –to Q propagation delay
T_{cdq}	latch flop clock D to Q contamination delay
T_{setup}	latch flop setup time
T_{hold}	latch flop hold time

- The delay with timing diagram for all three sequencing elements are, as shown in figure below.
- In combinational logic, input A changing to another value, output Y cannot change instantaneously. After the contamination delay {t_{cd}}, Y may begin to change (or) glitch.
- Output Y settles to a value in propagation delay {t_{pd}} .
- Input D in flip flop must have settled by some setup time {t_{setup}} before the rising edge of clock and should not change again until, a hold time {t_{hold}} after the clock edge.

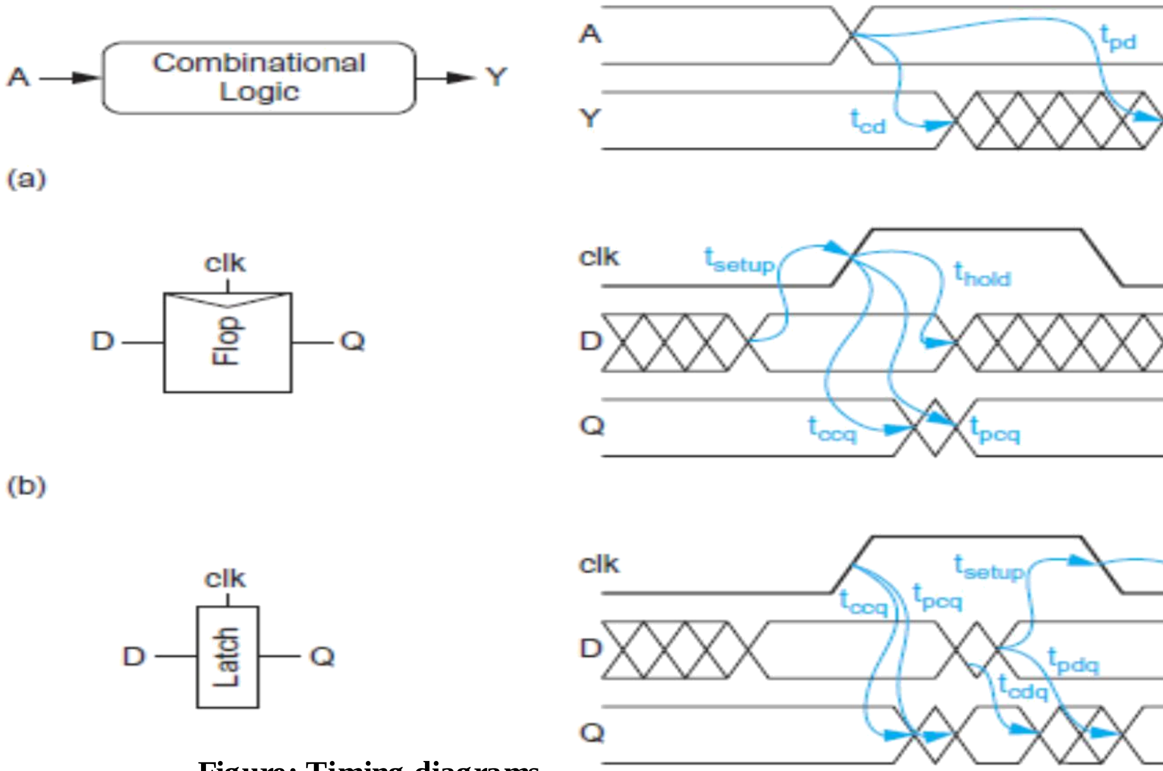


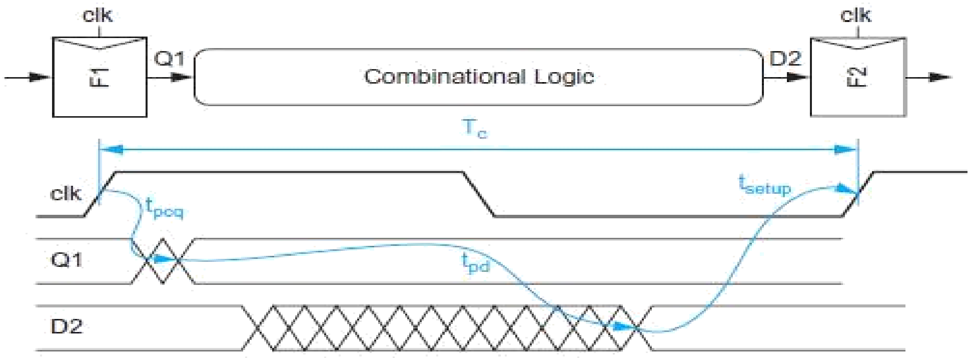
Figure: Timing diagrams

(c)

- The output begins to change after a clock-to-contamination delay {t_{ccq}} and completely settles after clock to-Q propagation delay {t_{pcq}}.

(B) Max Delay Constraints:-

- Ideally, the entire clock cycle will be available for computation in the combinational logic.
- If the combination logic delay is too high, the receiving element {next flop/latch} will miss its setup time and sample the wrong value.
- This is called as "setup-time failure" or "max-delay failure".
- It can be solved by redesigning the logic to be faster (or) by increasing the clock period.



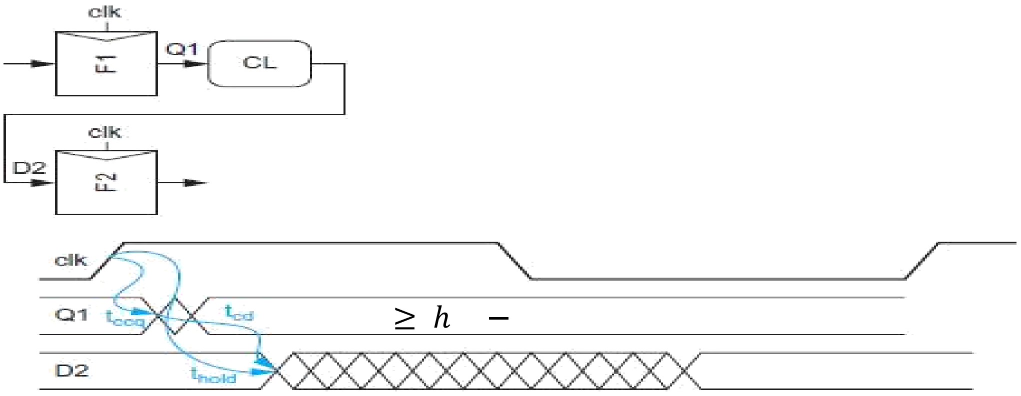
The clock period must be ≥ + + (or)

≤ − (+)

Where, tsetup+tpcq – sequencing overhead.

(C) Min-delay constraints:-

- Sequencing elements can be placed back to back without intervening combinational logic and still function correctly.
- If the hold time is large and contamination delay is small, data can incorrectly propagate through successive elements, on one clock edge.
- This corrupt the state of the system called, race condition (or) hold time failure (or) min-delay failure.
- It can be fixed by redesigning the logic and not by slowing the clock.

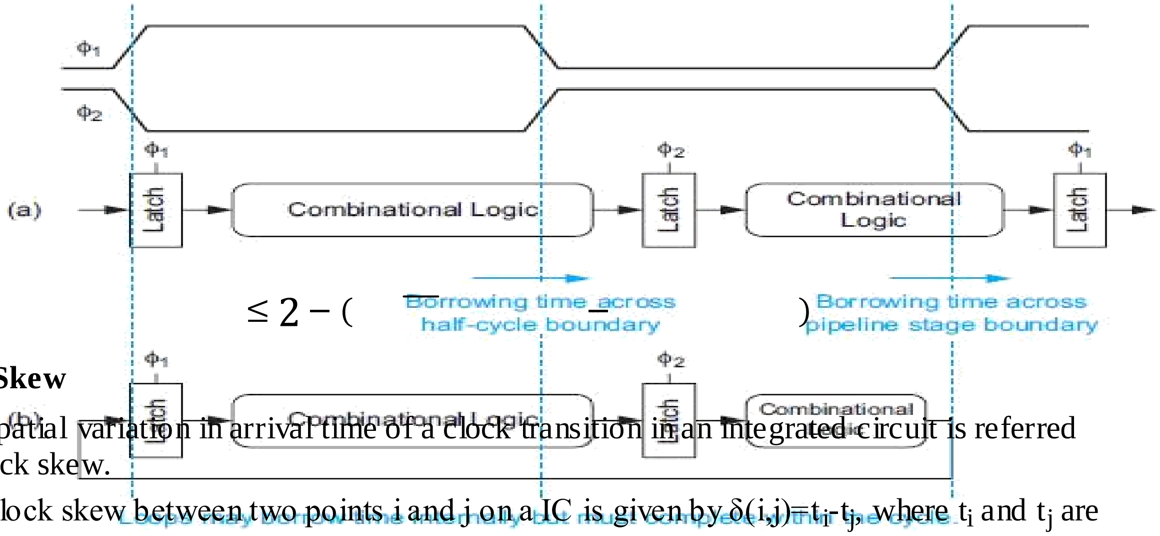


$\geq h \quad -$

FIGURE Flip-flop latch min-delay constraint

(D)Time Borrowing:

- In flip-flop, dada departs the first flip-flop on the rising edge of the clock and must set up at the second flip-flop before the next rising edge of the clock.
- If data arrives late, produces wrong result.
- If data arrives early, it is blocked until the clock edge arrives and the remaining time goes unused and clock imposes a “hard edge”.
- If one half cycle (or) stage of a pipeline has too much logic, it can borrow time in half-cycle (or) stage.
- This is called as “Time borrowing”, which can accumulate across multiple cycles.



(E) Clock Skew

- The spatial variation in arrival time of a clock transition in an integrated circuit is referred as clock skew.
- The clock skew between two points i and j on a IC is given by $\delta(i,j)=t_i-t_j$, where t_i and t_j are the position of the rising edge of the clock with respect to a reference.
- The clock skew can be positive or negative, depending upon the routing direction and position of the clock source.
- The timing diagram for the case with positive skew, is shown in figure.
- In the figure, the rising clock edge is delayed by a positive δ at the second register.

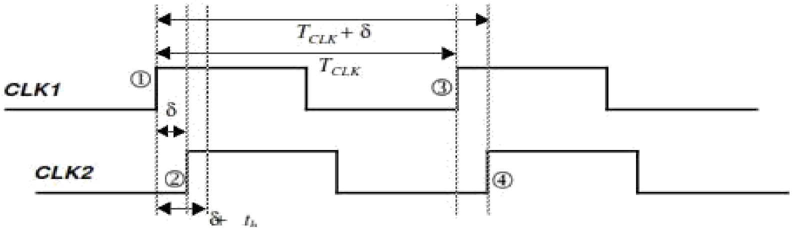


Figure: Timing diagram to study the impact of clock skew on performance and functionality. In this sample timing diagram, δ >0.

(F) Clock Jitter:

- Clock jitter is the temporal variation of the clock period at a given point. The clock period can reduce or expand on a cycle-by-cycle basis. It is a temporal uncertainty measure.
- Cycle-to-cycle jitter refers to time varying deviation of a single clock period.
- For a given spatial location, i is given as T_{jitter}, i(n) = $T_{i,n+1} – T_{i,n} – T_{CLK}$.

  Where Ti,n is the clock period for period n, $T_{i,n+1}$ is clock period for period n+1, and T_{CLK} is the nominal clock period.
- Jitter directly impacts the performance of a sequential system.
- Figure shows the nominal clock period as well as variation in period.

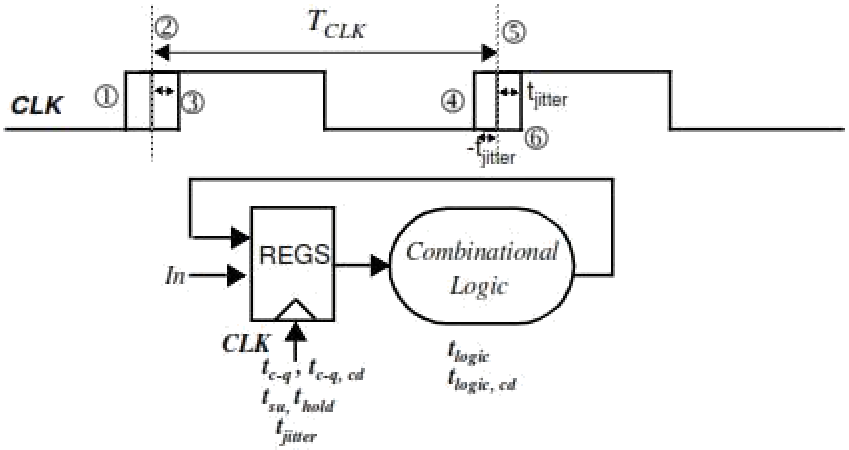


Figure: Circuit for studying the impact of jitter on performance.

****

3.4 Pipelining:

❖ **Explain in detail about pipelining structure needed for a logic operation. (April 2017, Nov 2017)**

❖ **Discuss in detail various pipelining approaches to optimize sequential circuits. (May 2013, 2016)**

- Pipelining is a design technique used to accelerate the operation of the datapaths in digital processors.
- The idea is explained with Figure 3.22a.
- The goal of the circuit is to compute log(|a - b|), where both a and b represent streams of numbers.
- The minimal clock period T_{min} necessary to ensure correct evaluation is given as:

$$T_{\min} = t_{c-q} + t_{pd,\text{logic}} + t_{su}$$

  Where, t_{c-q} and t_{su} are the propagation delay and the set-up time of the register respectively.
- Registers are edge-triggered D registers.
- The term $t_{pd,logic}$ stands for the worst-case delay path through the combinatorial network, which consists of the adder, absolute value and logarithm functions.

- In conventional systems, the delay is larger than the delays associated with the registers and dominates the circuit performance.
- Assume that each logic module has an equal propagation delay.
- Each logic module is then, active for only 1/3 of the clock period.
- Pipelining is a technique to improve the resource utilization and increase the functional throughput.
- Introduce registers between the logic blocks, as shown in Figure 3.22b.
- This causes the computation for one set of input data to spread over a number of clock periods, as shown in Table 1.
- The result for the data set (a1, b1) only appears at the output after three clock-periods.

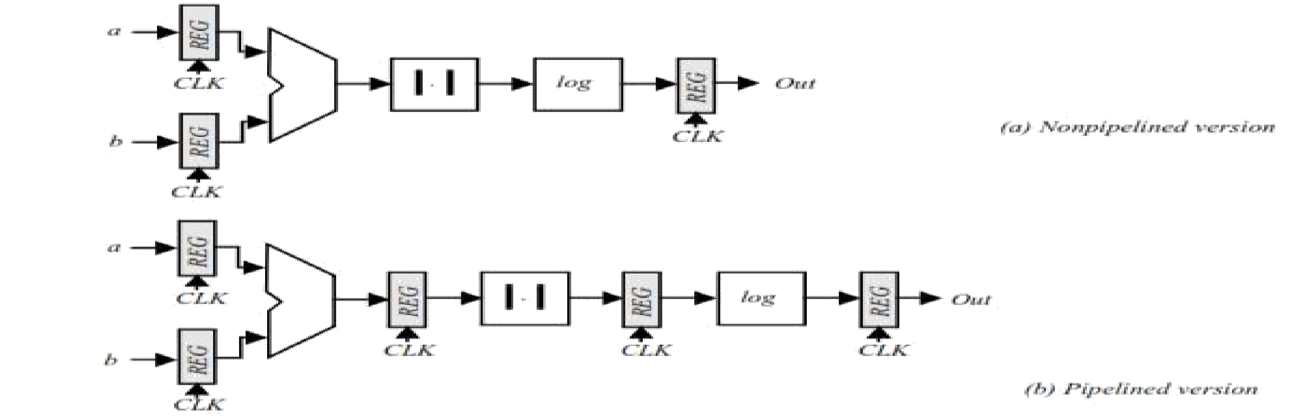
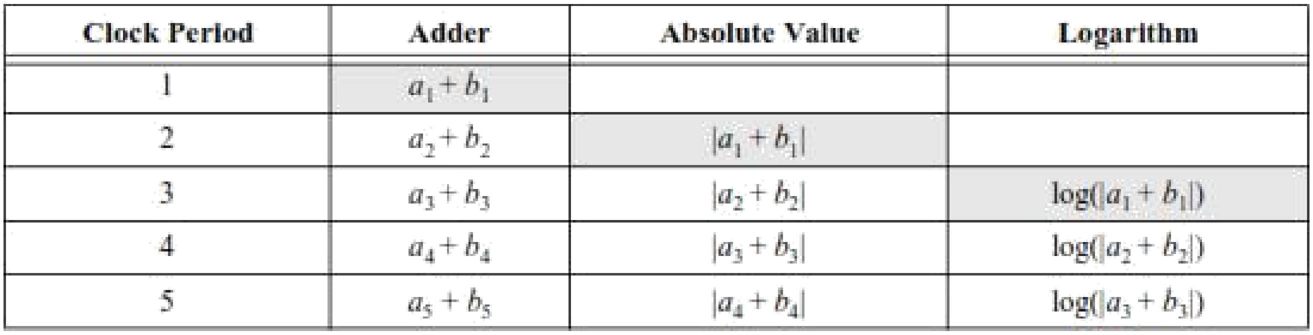


Figure 3.22 Data path for the computation of log(|a + b|).

- At that time, the circuit has already performed parts of the computations for the next data sets, (a2, b2) and (a3, b3).
- The computation is performed in an assembly-line fashion, hence the name pipeline.
- The combinational circuit block has been partitioned into three sections, each of which has a smaller propagation delay than the original function.



Clock Period	Adder	Absolute Value	Logarithm
1	$a_1 + b_1$		
2	$a_2 + b_2$	$\lvert a_1 + b_1 \rvert$	
3	$a_3 + b_3$	$\lvert a_2 + b_2 \rvert$	$\log(\lvert a_1 + b_1 \rvert)$
4	$a_4 + b_4$	$\lvert a_3 + b_3 \rvert$	$\log(\lvert a_2 + b_2 \rvert)$
5	$a_5 + b_5$	$\lvert a_4 + b_4 \rvert$	$\log(\lvert a_3 + b_3 \rvert)$

Table 1: Example of pipelined computations.

- This reduces the value of the minimum allowable clock period:

$$T_{\min, pipe} = t_{c-q} + \max(t_{pd, add}, t_{pd, abs}, t_{pd, \log})$$

- Suppose all logic blocks have same propagation delay and that the register overhead is small with respect to the logic delays.
- The pipelined network performs the original circuit by a factor of three, under these assumptions $T_{min,pipe}=T_{min}/3$.

- The increased performance comes at the relatively small cost of two additional registers, and an increased latency.
- Pipelining is implemented for very high-performance datapaths.

3.4.1 NORA-CMOS—A Logic Style for Pipelined Structures:

Discuss about the NORA–CMOS structure. (Nov 2016)

- The latch-based pipeline circuit can also be implemented using C^2MOS latches, as shown in Figure 3.24.
- This topology has one additional property:
  - *A C^2MOS-based pipelined circuit is race-free as long as all the logic functions F (implemented using static logic) between the latches are noninverting.*

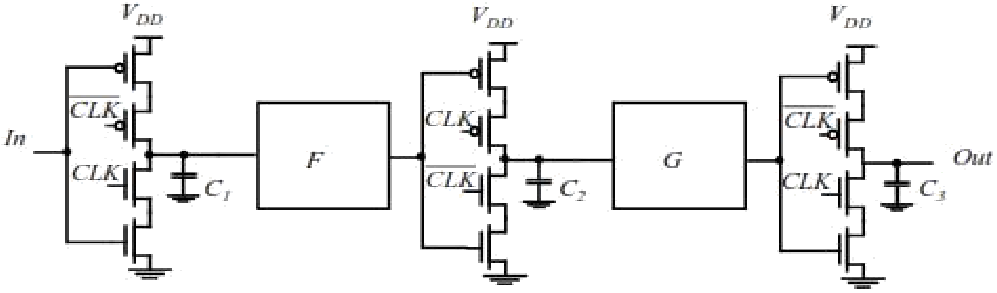


Figure 3.24 Pipelined datapath using C^2MOS latches

- The only way, a signal can race from stage to stage under this condition is, when the logic function F is inverting, as in Figure 3.25.
- Here F is replaced by a single, static CMOS inverter. Similar considerations are valid for the (1-1) overlap.
- It combines C^2MOS pipeline registers and NORA dynamic logic function blocks.
- Each module consists of a block of combinational logic, that can be a mixture of static and dynamic logic, followed by a C^2MOS latch.

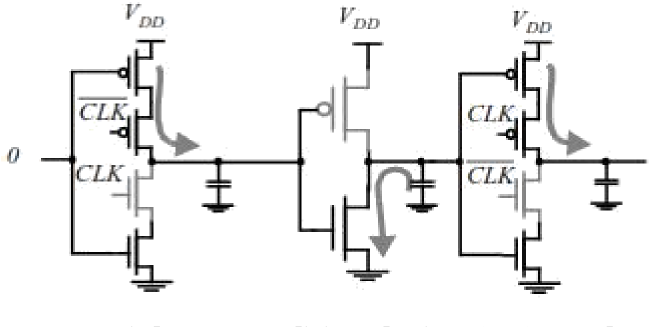
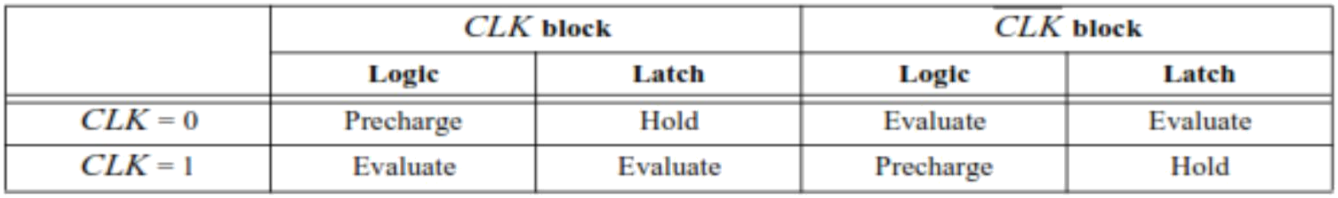


Figure 3.25 Potential race condition during (0-0) overlap in C^2MOS-based design.

- Logic and latch are clocked, in such a way that both are simultaneously in either evaluation or hold (precharge) mode.
- A block that is, in evaluation during CLK = 1 is called a CLK-module, while the inverse is called a $\overline{CLK}$ -module.
- The operation modes of the modules are summarized in Table 2.



	CLK block		$\overline{CLK}$ block	
	Logic	Latch	Logic	Latch
$CLK = 0$	Precharge	Hold	Evaluate	Evaluate
$CLK = 1$	Evaluate	Evaluate	Precharge	Hold

Table 2: Operation modes for NORA logic modules.

3.4.2 Latch- vs. Register-Based Pipelines:

Compare Latch and register based pipelines.

- Pipelined circuits can be constructed using level-sensitive latches instead of edge-triggered registers.
- Consider the pipelined circuit of Figure 3.23.
- The pipeline system is implemented based on pass-transistor-based positive and negative latches instead of edge triggered registers.
- Here logic is introduced between the master and slave latches of a master-slave system.
- When the clocks CLK and $\overline{CLK}$ are non-overlapping, correct pipeline operation is obtained.

- Input data is sampled on C_1 at the negative edge of CLK and the computation of logic block F starts.
- The result of the logic block F is stored on C_2 on the falling edge of $\overline{CLK}$ and the computation of logic block G starts.
- The non-overlapping of the clocks ensures correct operation.
- The value stored on C_2 at the end of the CLK low phase, is the result of passing the previous input through the logic function F.
- When overlap exists between CLK and $\overline{CLK}$, the next input is already being applied to F, and its effect might propagate to C_2 before $\overline{CLK}$ goes low.

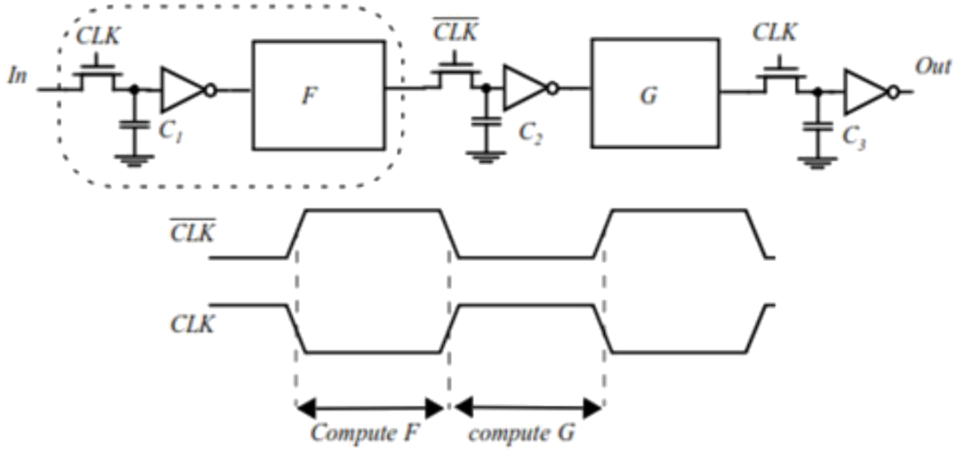


Figure 3.23 Operation of two-phase pipelined circuit using dynamic registers.

****

3.5 Choosing a Clocking Strategy:

Discuss about strategy required for choosing a clock signal.

- Choosing the right clocking scheme affects the functionality, speed and power of a circuit.
- The simple clocking scheme is the two-phase master-slave design.
- The predominant approach is use the multiplexer-based register and to generate the two clock phases locally, by simply inverting the clock.
- High-performance CMOS VLSI design is using simple clocking schemes, even at the expense of performance.

3.5.1 Static sequencing element methodology:-

- Many issues are related to static sequencing element methodology.

(a) Choice of element:-

Flip-flop: Flip-flop has high sequencing overhead. It is simple and easy to understand the operation of flip-flop.

Pulsed latches:-

- Faster than flip-flop.
- Provides some time borrowing option.
- Consumes law power.

Transparent Latch:-

- It has low sequencing overhead compared with flip-flop.
- It allows almost half cycle of time borrowing and it is good choice.

(b)Low power sequential design:-

- Pulsed latches are power efficient.
- Flip-flop consumes more power.
- Clock gating can be used to reduce power.

(c) Two-phase Timing types:-In this type, the signal can belong to phase 1 (or) phase 2. In each phase, 3 different clocks are stable, valid and qualified clock.

Stable clock:-

A signal is stable (in 1), if it settles to a value before rises and remains constant until after, 1 falls.

Valid clock: - A signal is valid (in 1), if it settles to a value before 1 falls and remains at that value after 1 falls.

Qualified signal: -A signal is said to be in qualified clock in 1 , if it either rises and falls like 1(or) remains low for the entire cycle.

****

3.6 Synchronous and Asynchronous circuits-Timing issues:

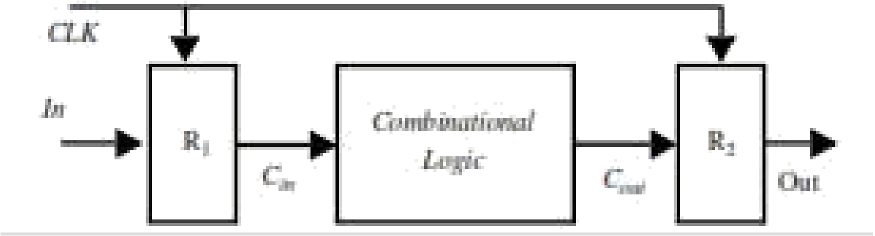
- A synchronous system approach is one, in which all memory elements in the system are simultaneously updated using a globally distributed periodic synchronization signal.
- Functionality is ensured by imposing some strict constraints on the generation of the clock signals and their distribution to the memory elements.
- Analyze the impact of spatial variations of the clock signal, called clock skew and temporal variations of the clock signal, called clock jitter.
- Asynchronous design avoids the problem of clock uncertainty by eliminating the need for globally-distributed clocks.
- The important issues of synchronization, which is required when interfacing different clock domains or when sampling an asynchronous signal.

Classification of Digital Systems:

- In digital systems, signals can be classified depending on, how they are related to a local clock.
- Signals that transition only at predetermined periods in time can be classified as synchronous, mesochronous and plesiochronous with respect to a system clock.
- A signal that can transition at arbitrary times is considered asynchronous.

3.6.1 Synchronous Interconnect

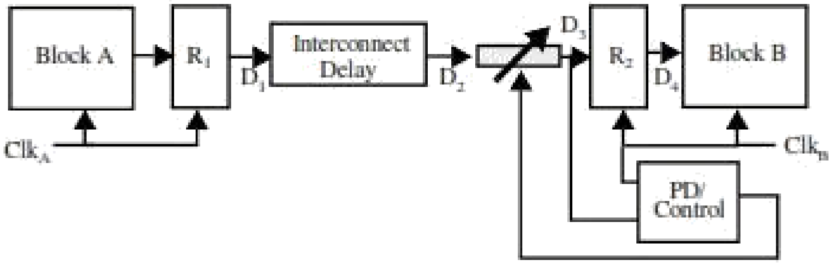
- A synchronous signal has the exact same frequency and a known fixed phase offset with respect to the local clock.
- In such a timing methodology, the signal is "synchronized" with the clock and the data can be sampled directly without any uncertainty.
- In digital logic design, synchronous systems are straight forward type of interconnect, where the flow of data in a circuit proceeds with the system clock as shown below.



3.6.2 Mesochronous interconnect:

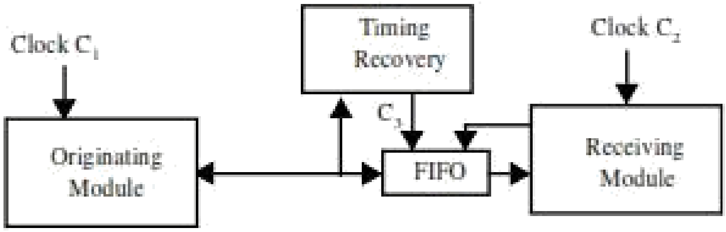
- A mesochronous signal has the same frequency but an unknown phase offset with respect to the local clock ("meso" from Greek is middle).
- For example, if data is being passed between two different clock domains, then the data signal transmitted from the first module can have an unknown phase relationship to the clock of the receiving module.
- In such a system, it is not possible to directly sample the output at the receiving module because of the uncertainty in the phase offset.
- A (mesochronous) synchronizer can be used to synchronize the data signal with the receiving clock as shown below.

- The synchronizer serves to adjust the phase of the received signal to ensure proper sampling.



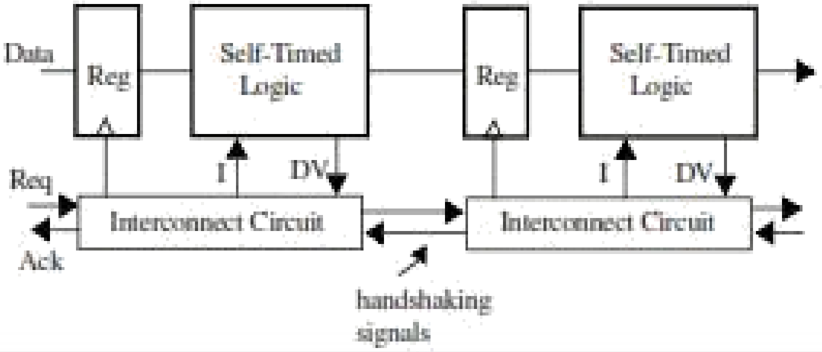
3.6.3 Plesiochronous Interconnect

- A plesiochronous signal has nominally the same, but slightly different frequency as the local clock ("plesio" from Greek is near).
- This scenario can easily arise when two interacting modules have independent clocks generated from separate crystal oscillators.
- Since the transmitted signal can arrive at the receiving module at a different rate than the local clock, one need to utilize a buffering scheme to ensure, all data is received.
- A possible framework for plesiochronous interconnect is shown in Figure below.



3.6.4 Asynchronous Interconnect:

- Asynchronous signals can transition at any arbitrary time and are not slaved to any local clock.
- As a result, it is not to map these arbitrary transitions into a synchronized data stream.
- Asynchronous signals are used to eliminate the use of local clocks and utilize a self-timed asynchronous design approach.
- In this approach, communication between modules is controlled through a handshaking protocol.



****

3.7 Clock-Distribution Techniques

❖

Explain the clock distribution techniques in synchronous design in detail. (Nov 2017)

- Clock skew and jitter are major issues in digital circuits and they limit the performance of a digital system.
- It is necessary to design a clock network, that minimizes skew and jitter.
- Another important consideration in clock distribution is the power dissipation.
- In most high-speed digital processors, a majority of the power is dissipated in the clock network.
- To reduce power dissipation, clock networks must support clock conditioning, the ability to shut down parts of the clock network.
- Unfortunately, clock gating results in additional clock uncertainty.

- Clock networks include a network that is used to distribute a global reference to various parts of the chip.
- A final stage is responsible for local distribution of the clock, while considering the local load variations.
- Most clock distribution schemes use the absolute delay from a central clock source to the clocking elements.
- Therefore one common approach to distributing a clock is, to use balanced paths (or called trees).
- The most common type of clock primitive is, the H-tree network (named for the physical structure of the network) in figure, where a 4x4 array is shown.
- In this scheme, the clock is routed to a central point on the chip and balanced paths.
- Include both matched interconnect as well as buffers, are used to distribute the reference to various leaf nodes.
- If each path is balanced, the clock skew is zero. It takes multiple clock cycles for a signal to propagate from the central point to each leaf node. The arrival times are equal at every leaf node.
- The H-tree configuration is particularly useful for regular-array networks, in which all elements are identical and the clock can be distributed as a binary tree.

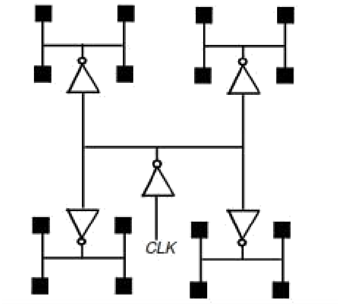


Figure: Example of an H-tree clock-distribution network for 16 leaf nodes.

Latch-Based Clocking:

- The use of a latch based methodology (in Figure) enables more flexible timing, allowing one stage to pass slack to or steal time from following stages.
- This flexibility allows an overall performance increase.
- In this configuration, a stable input is available to the combinational logic block A (CLB_A) on the falling edge of CLK1 (at edge2).
- On the falling edge of CLK2 (at edge3), the output CLB_A is latched and the computation of CLK_B is launched.
- CLB_B computes on the low phase of CLK2 and the output is available on the falling edge of CLK1 (at edge4).
- This timing appears, equivalent to having an edge-triggered system where CLB_A and CLB_B are cascaded and between two edge-triggered registers.
- In both cases, it appears that the time available to perform the combination of CLB_A and CLB_B are T_{CLK}.

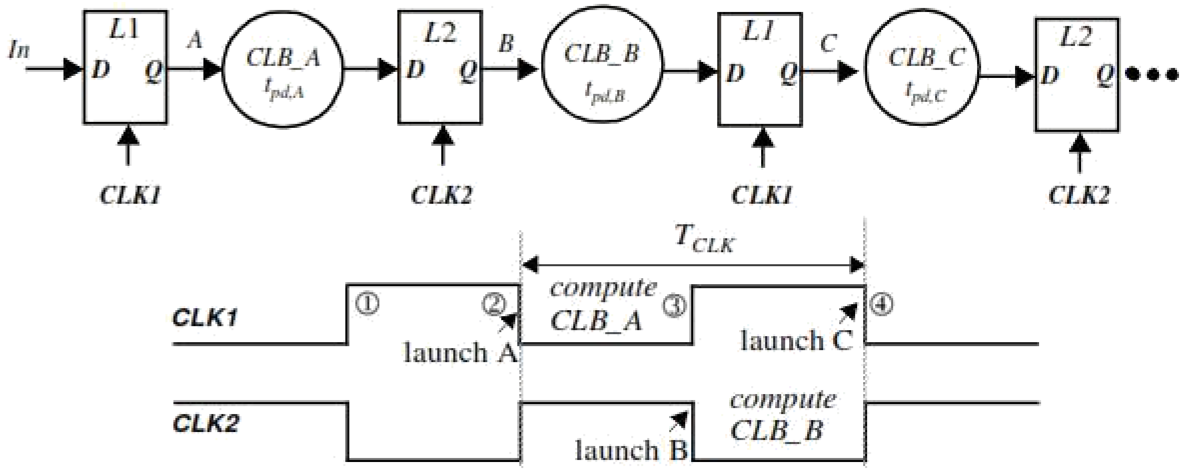


Figure: Latch-based design in which transparent latches are separated by combinational logic.

****

3.8 Self-Timed Circuit Design

Discuss about asynchronous design in logic design.

3.8.1 Self-Timed Logic : An Asynchronous Technique

- A more reliable and robust technique is the self-timed approach, which presents a local solution to the timing problem.
- Figure uses a pipelined datapath to illustrate how this can be accomplished.
- The computation of a logic block is initiated by asserting a Start signal.
- The combinational logic block computes on the input data.

- This signaling ensures the logical ordering of the events and can be achieved with the aid of an extra Ack(nowledge) and Req(uest) signal.

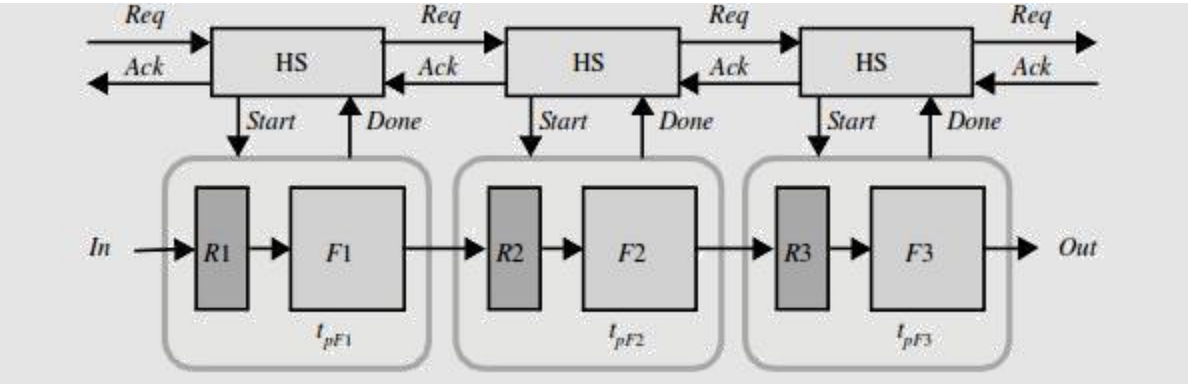


Figure: Self-timed, pipelined datapath.

In the case of the pipelined datapath, the scenario could proceed as follows.

1. An input word arrives, and a Req(uest) to the block F1 is raised. If F1 is inactive at that time, it transfers the data and acknowledges this fact to the input buffer.
2. F1 is enabled by raising the Start signal. After a certain amount of time, dependent upon the data values, the Done signal goes high indicating the completion of the computation.
3. A Req(uest) is issued to the F2 module. If this function is free, an Ack(nowledge) is raised, the output value is transferred and F1 can go ahead with its next computation.

3.8.2 A simple synchronizer

How do eliminates metastability problem in sequential circuit and explain?

- A synchronizer accepts an input D and a clock ϕ. It produces an output Q that should be valid for some bounded delay after the clock.
- The synchronizer has an aperture, defined by a setup and hold time around the rising edge of the clock.
- If the data is stable during the aperture, Q should equal D. If the data changes during the aperture, Q can be chosen arbitrarily.

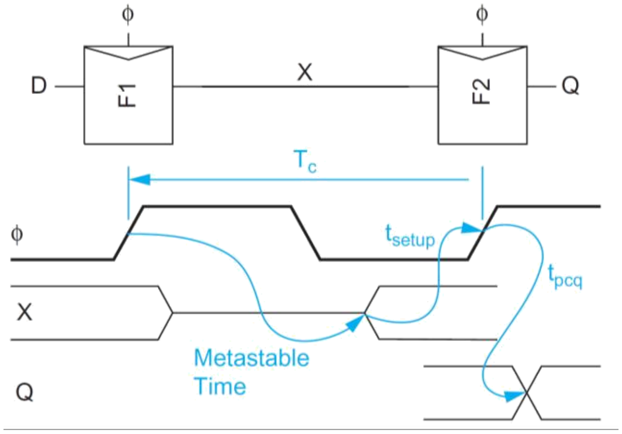


Figure: Simple Synchronizer

- Figure shows a simple synchronizer built from a pair of flip-flops. F1 samples the asynchronous input D.
- The output X may be metastable for some time, but will settle to a good level with high probability, if we wait long enough.
- F2 samples X and produce an output Q, that should be a valid logic level and be aligned with the clock.
- The synchronizer has a latency of one clock cycle Tc .

3.8.3 Communicating between asynchronous clock domains

- A common application of synchronizers is in communication between asynchronous clock domains, i.e., blocks of circuits that do not share a common clock.
- Suppose System A is controlled by clkA that needs to transmit N-bit data words to System B, which is controlled by clkB, as shown in Figure.
- The systems can represent separate chips or separate units within a chip using unrelated clocks.
- System A must guarantee that the data is stable, while the flip-flops in System B sample the word.
- It indicates when new data is valid by using a request signal (Req), so System B receives the word exactly once rather than zero or multiple times.
- System B replies with an acknowledge signal (Ack), when it has sampled the data, so System A knows when the data can safely be changed.
- If the relationship between clkA and clkB is completely unknown, a synchronizer is required at the interface.

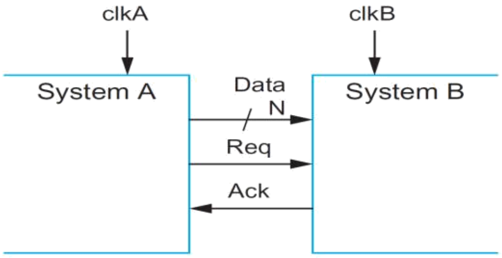


Figure: Communication between asynchronous systems

3.8.4 Arbiter

- The arbiter of Figure (a) is related to the synchronizer. It determines which of two inputs arrived first.
- If the spacing between the inputs exceeds some aperture time, the first input should be acknowledged.
- If the spacing is smaller, exactly one of the two inputs should be acknowledged, but the choice is arbitrary.
- For example, in a television game show, two contestants may hit buttons to answer a question.

- If one presses the button first, should be acknowledged. If both presses the button at times too close to distinguishes, the host may choose one of the two contestants arbitrarily.

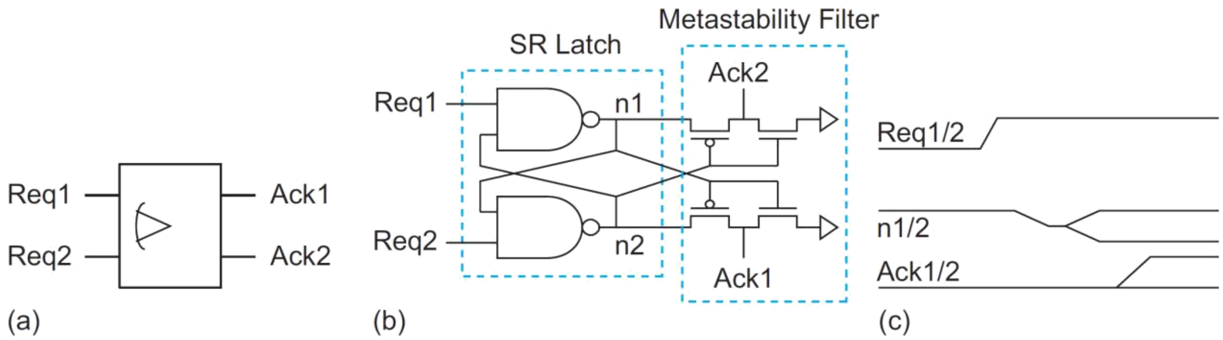


Figure: Arbiter

- Figure (b) shows an arbiter built from an SR latch and a four-transistor metastability filter.
- If one of the request inputs arrives well before the other, the latch will respond appropriately.
- If they arrive at nearly the same time, the latch may be driven into metastability, as shown in Figure (c).
- The filter keeps both acknowledge signals low, until the voltage difference between the internal nodes n1 and n2 exceeds Vt , indicating that a decision has been made.
- Such an asynchronous arbiter will never produce metastable outputs.

3.8.5 Synchronous versus Asynchronous Design:

Compare synchronous and asynchronous design.

- The self-timed approach offers a potential solution to the growing clock-distribution problem.
- It translates the global clock signal into a number of local synchronization problems.
- Handshaking logic is needed to ensure the logical ordering of the circuit events and to avoid race conditions.
- In general, synchronous logic is both faster and simpler since the overhead of completion-signal generation and handshaking logic is avoided.
- Skew management requires extensive modeling and analysis, as well as careful design.
- It will not be easy to extend this methodology into the next generation of designs.
- This observation is already reflected in the fact that the routing network for the latest generation of massively parallel supercomputers is completely implemented using self-timing.
- For self timing to become a mainstream design technique however (if it ever will), further innovations in circuit and signaling techniques and design methodologies are needed.

****

3.9 Memory Architecture and Memory Control Circuits:

Discuss Memory classification and its architecture and building blocks.

3.9.1 Memory Classification:

- Parameters used to characterize a memory device are area, power and speed.
- **Area:** area is important for its physical implementations by VLSI technology. Smaller the area per bit and more devices can be accommodated. So cost per bit is reduced.
- **Speed:** speed of operation plays a very important role. Memory can communicate at speed with processors.
- **Power:** Power is important, because MOS memories are used in many battery operated portable systems. Power dissipation of memory plays an important role. Memory devices will consume less power.

1. ROM
2. RAM

1. Volatile
2. Non-volatile

- Volatile memory devices will store information, as long as power is it.
- As soon as power is turned off, information is lost. Static RAM and dynamic RAM belong to the category of volatile memory.
- EPROM and mask programmable ROM are non-volatile memory devices.
- If the power is turned off information will not be lost.

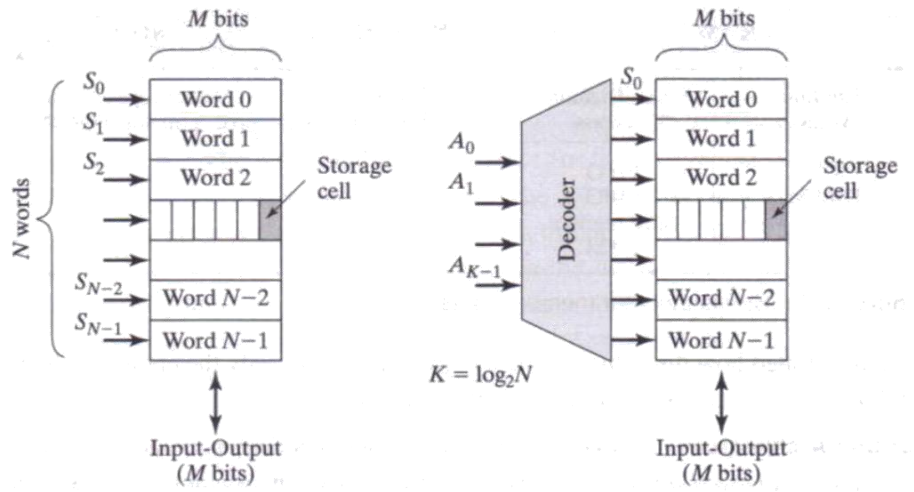
<u>Classification based on access method</u>:

1. Random access
2. Non-random access

3.9.2 Memory Architecture and Building Blocks:

When n x m memory is implemented, then, n memory words are arranged in a linear fashion. One word will be selected at a time by using select line.

- If we want to implement the memory 8X8, n=8, m=8(number of bits).
- Then we need 8 select signals (one for each word).
- But by using decoder we can reduce the number of select signals.
- In case of 3 to 8 decoder, if 3 inputs are given to decoder, then we can get 8 select signals.
- If n=220, then we can give only 20 inputs to the decoder.



(a) Intuitive architecture for N × M memory

(b) A decoder reduces the number of address bits

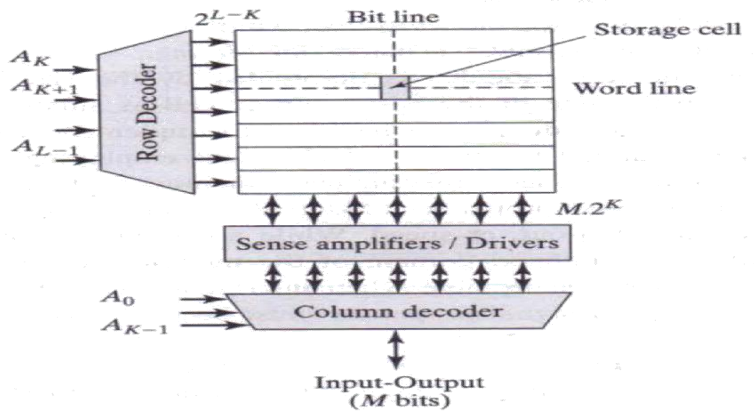


Figure: Array structured memory organization

- If basic storage cell size is approximately square, then the design is extremely slow. The vertical wire, which connects the storage cells to I/O will be excessively large.
- So, memory arrays are organized in such a way that vertical and horizontal dimensions are the same.
- The words are stored in a row. These words are selected simultaneously.
- The **column decoder** is used to route the correct word to the I/O terminals.
- The row address is used to select one row of memory and column address is used to select particular word from that selected row.
- **Word line**: The horizontal select line which is used to select the single row of cell is known as word line.
- **Bit line**: The wire which connects the cell in a single column to the input/output circuit is known as bit line.
- **Sense amplifier**: It requires an amplification of the internal swing to full rail-to-rail amplitude.
- **Block address**: the memory is divided into various small blocks.
- The address which is used to select one of the small blocks to be read or written is known as block address.

- **Advantages:**

  1. Access time is fast
  2. Power saving is good, because blocks not activated are in power saving mode.

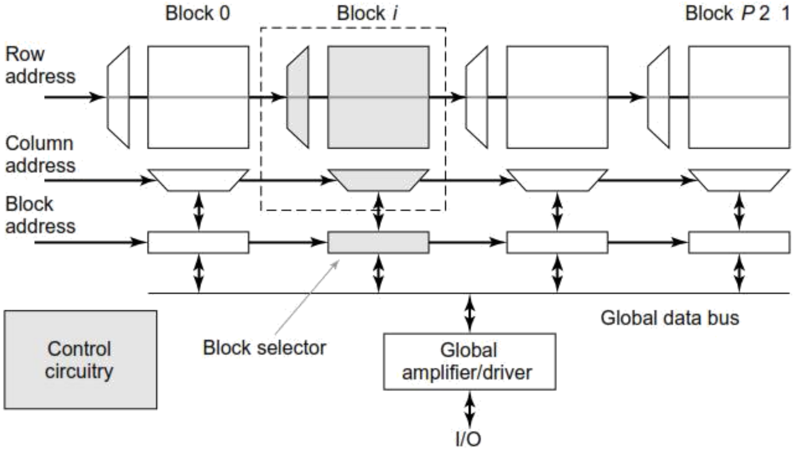


Figure: Hierarchical memory architecture

3.9.3 Memory Core

Discuss Memory core its types in detail.

3.9.3.1 Read Only Memory (ROM):

ROM is a memory where code is written only one time.

Diode ROM:

- It is simple where presence of diode in between bit line and word line is considered as logic 1 and absence of diode as logic 0.
- Disadvantage is used for small memories and no isolation between word line and bit line.

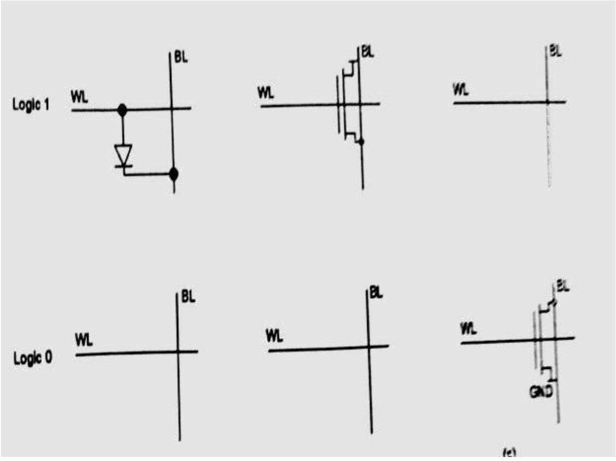


Figure: Diode ROM

MOS ROM:

- Diode is replaced by gate source connection of nMOS. Drain is connected to V_{DD}.
- The charging and discharging of word line capacitance has been taken care by the word line driver.

- Absence of a transistor between word line and bit line means logic 1 is stored and if presence then logic 0 is stored.

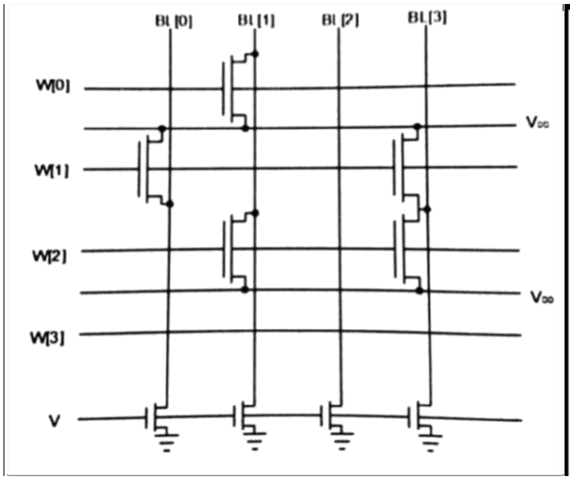


Figure: 4 x 4OR ROM cell array

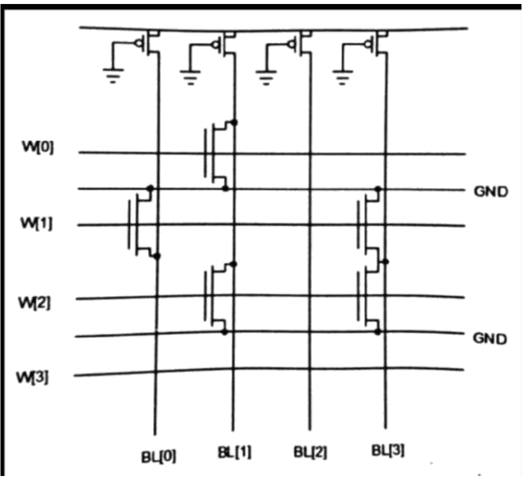


Figure: 4 x 4 MOS NOR ROM

Programming ROM

- The transistor in the intersection of row and column is OFF when the associated word line is LOW. In this condition, we get logic 1 output.

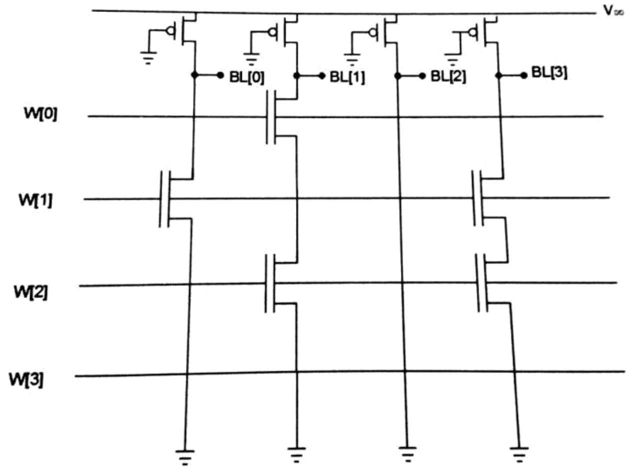


Figure: 4 x 4 MOS NAND ROM

Advantage: basic cell only consists of transistor. No need of connection to any of the supply voltage.

Disadvantage: As it has pseudo nMOS, it is ratioed logic and consumes static power.

- To overcome this, precharged MOS NOR ROM logic circuit is used.
- This eliminate static dissipation ratioed logic requirement.

3.9.3.2 Non-Volatile READ-WRITE Memory:

- It consists of array of transistors. We can write the program by enabling or disabling these devices selectively.
- To reprogram, the programmed values to be erased, then the new programming is started.

Floating gate transistor:

- It is mostly used in all the reprogrammable memories.
- In floating gate transistor, extra polysilicon strip is used in between the gate and the channel known as floating gate.
- Floating gate doubles the gate oxides thickness and hence device transconductance is reduced and threshold voltage is increased.
- The threshold voltage is a programmable.
- If high voltage is (>10V) is applied between the source terminals and gate-drain terminals, then high electric field is generated. So, avalanche injection occurs.
- After acquiring energy, electron becomes hot and transverse through the first oxide insulator . They get trapped on the floated gate.
- The floating gate transistor is known as floating gate avalanche injection MOS or FAMOS. **Disadvantage:** High programming voltage is need.

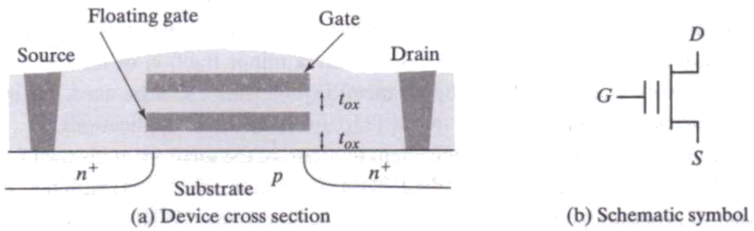


Figure (a) floating gate transistor (b) symbol

EPROM – Erasable Programmable Read Only Memory:

- Erasing is done by passing UV rays on the cell by using transparent window.
- This process will take some seconds to some minutes.
- It depends on intensity of UV source. The programming takes 5-10microseconds/word.
- During programming, chip is removed from the board and placed in EPROM programmer. **Advantages:** simple and large families are fabricated with low cost.

Disadvantages:

- Number of erase/program cycle is limited upto 1000.
- Reliability is not good.
- Threshold voltage of the device may be varied with repeated program.

EEPROM – E^{2}PROM:

- Electrically Erasable Programmable ROM. Here Floating gate tunneling oxide (FLOTOX) is used.
- It is similar to floating gate except that the portion of the floating gate is separated from the channel at the thickness of 10nm or <10nm.
- If 10V is applied, electron travels to and from the floating gate through Fowler-Nordheim tunneling.
- Erasing can be done by revering applied voltage which is used for writing.

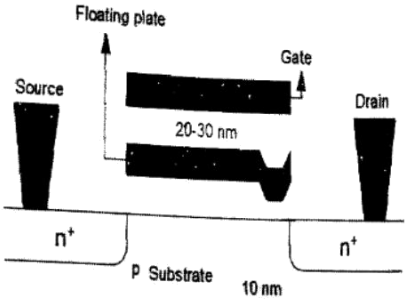


Figure: FLOTOX transistor

Advantage: High versatility and possible for 105 erase/write cycle.

Disadvantages: Larger than FAMOS transistor, Costly, Repeated programming causes a drift in threshold voltage.

Flash Memory – Flash Electrically Erasable Programmable ROM

- It is a combination of density of EPROM and versatility of EEPROM.
- Avalanche hot electron injection mechanism is used.
- Erasing can be done by Fowler-Nordheim tunneling concept. Here erasing is done in bulk.

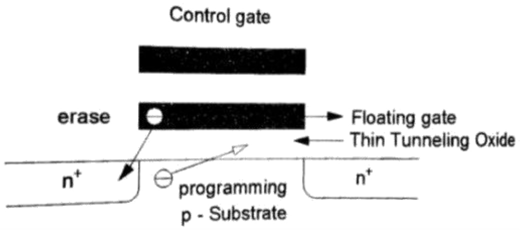


Figure: ETOX device

- It is similar to FAMOS gate.
- A very thin tunneling oxide layer (10nm thickness) is there.
- ***Erasing operation:*** Erasing can be performed when gate is connected to the ground and the source is connected to 12V.
- ***Write operation:*** High voltage pulse is applied to the gate of the selected device. Logic 1 is applied to the drain and hot electrons are injected into the floating gate.
- ***Read operation:*** To select a cell, its word line is connected to 5V. It causes conditional discharge of the bit line.

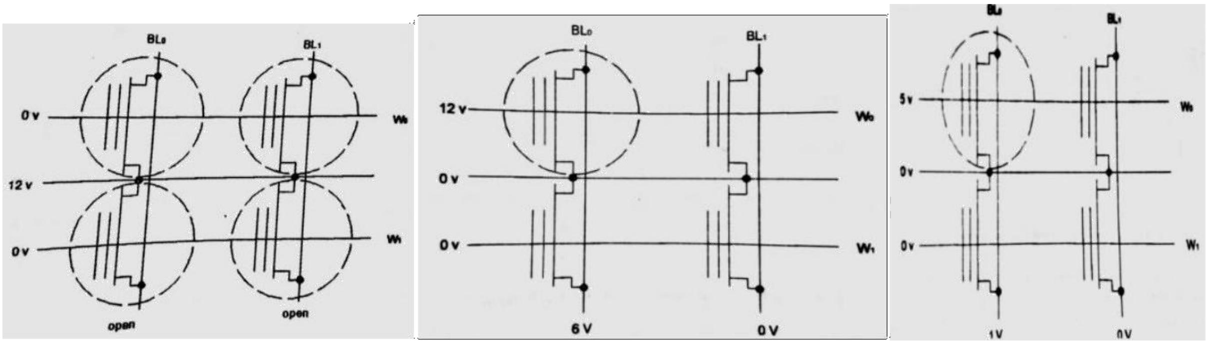


Figure: (a) Erase (b) Write (c) Read operation of NOR flash memory

3.9.3.3 RAM – Random Access Memory

Explain about static and dynamic RAM.

3.9.3.3.1 Static RAM:

- SRAM cell needs 6 transistors per bit.
- M5 and M6 transistors are shared between read and write operations.
- Bit line (BL) and inverse Bit Line signals are used to improve the noise margin during read and write operations.

***Read operation**:*

- Let us assume logic 1 is stored at Q and BL and inverse BL are precharge to 2.5V before starting read operation.
- The read cycle is started by asserting word line then M_5 and M_6 transistors are enabled.
- After the small initial word line delay then the values stored at Q and inverse Q are transferred to the bit lines by leaving BL at 2.5V and the value at inverse Q is discharge through M_1, M_5.

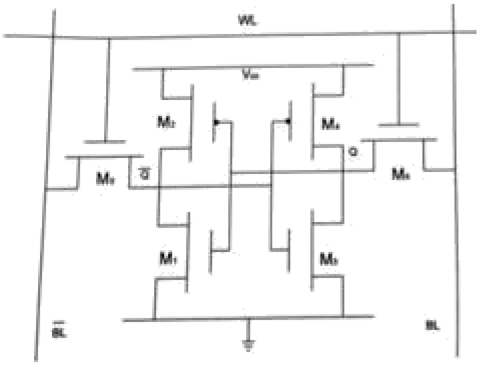


Figure: CMOS SRAM cell

***Write operation**:*

- Assume that Q=1, now logical 0 is to be written in the cell.
- Then inverse BL is set to 1 and BL is set to 0.
- The gate of M_1 is at V_{DD} and gate of M_4 is at ground as long as the switching is not commenced.
- Inverse Q is not pulled high enough to ensure the writing of logic 1.
- Cell voltage is kept below 0.4V. The new value of the cell is written through M_6.

3.9.3.3.2 Dynamic RAM:

Three transistors DRAM

- Content in the cell can be periodically rewritten through a resistive load, called as refresh operation.
- This refresh occurs for every 1-4ms. Dynamic memory has refresh operation.
- For example, logic 1 is to be written, and then BL1 is asserted high and write word line (WWL) is asserted.
- This data is retained as charge on the capacitor once WWL is low.

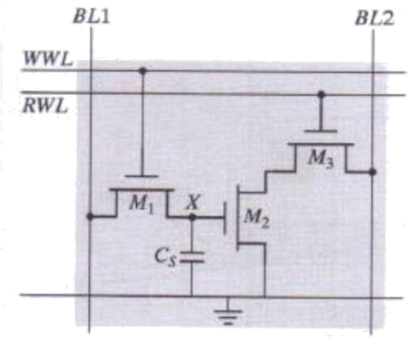


Figure: Three transistor dynamic memory cell

- To read the cell, the read word line (RWL) is raised. M_2 transistor is either ON or OFF depends upon the stored value.
- BL2 bit line is connected to V_{DD} or it is precharged to V_{DD} or V_{DD}-V_t.
- When logic 1 is stored, the series combination of $M_2$and M_3 pulls BL2 line low.
- If logic 0 is stored, then BL2 line is high.
- To refresh the cell, first the stored data is read, and its inverse is placed on BL1 and WWL line is asserted.

- In this cell, to write logic 1 then it is placed on bit line and word line is asserted high.
- The capacitor is charged or discharged depending upon the data. Before performing read operation, bit line is precharged.

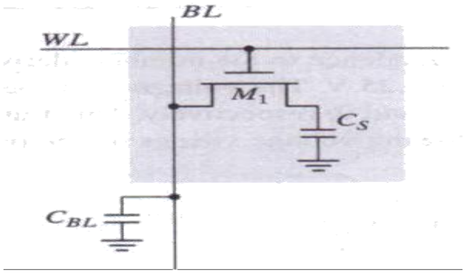


Figure: One transistor DRAM

3.9.3.3.3 CAM – Content Addressable or Associate Memory

Explain about CAM.

- It supports 3 operating modes,
  - Read
  - Write
  - Match
- In this memory, it is possible to compare all the stored data in parallel with the incoming data. It is not power efficient.
- Figure shows a possible implementation of a CAM array.
- The cell combines a traditional 6T RAM storage cell *(M_4-M_9)* with additional circuitry to perform a l-bit digital comparison *(M_1-M_3)*.
- When the cell is to be written, complementary data is forced onto the bit lines, while the word line is enabled a s in a standard SRAM cell.

- In the compare mode, stored data are compared using bit line. The match line is connected to all CAM blocks in a row. And it is initially precharged to V_{DD}.
- If there is some match occurs, then internal row is discharged. If even one bit in a row is mismatched, then the match line is low.

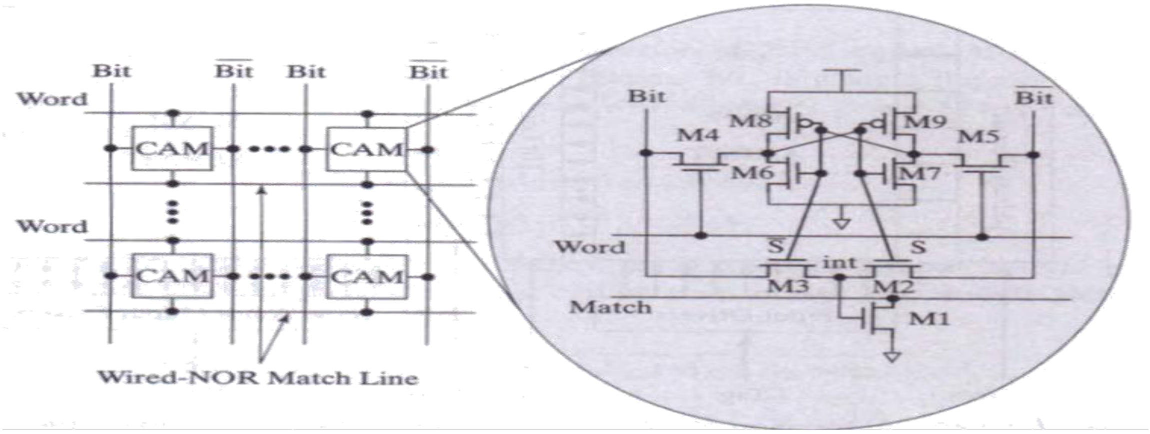


Figure: CAM cell

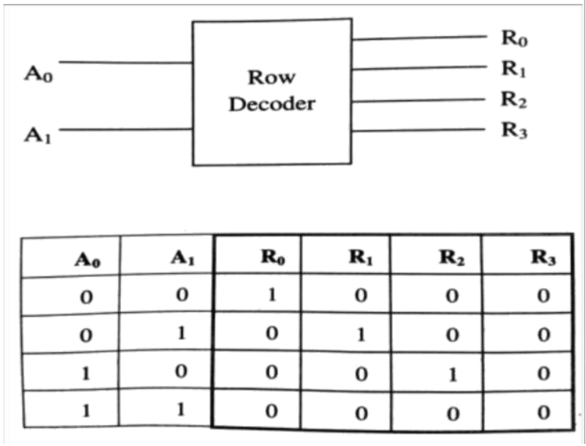
****

3.10 Memory peripheral (control) Circuits:

(i) Address & Block Decoders:

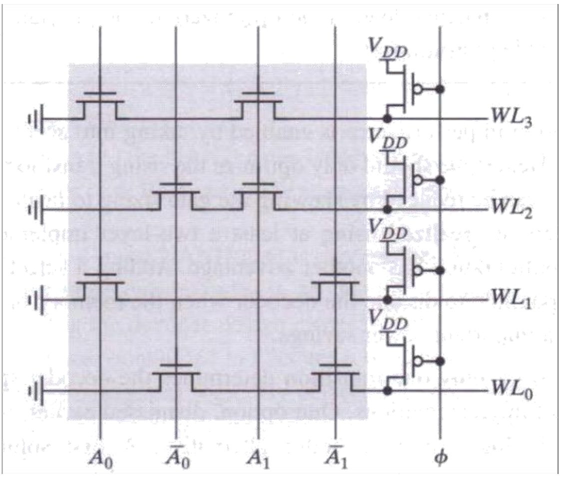
Row Decoder:

- Row and column address decoder are used to select the particular memory location in an array.
- Row decoder is used to drive NOR ROM array. It selects one of 2^n word lines.
- Dynamic 2 to 4 decoder reduces the number of transistors and propagation delay.



A_0	A_1	R_0	R_1	R_2	R_3
0	0	1	0	0	0
0	1	0	1	0	0
1	0	0	0	1	0
1	1	0	0	0	0

Symbol and Truth table



Dynamic 2-to-4 NOR decoder

Column Decoder

- It should match the bit line pitch of the memory array.
- In column decoder, decoder outputs are connected to nMOS pass transistors.
- By using this circuit, we can selectively drive one out of m pass transistors.

- Only one nMOS pass transistor is ON at the time.

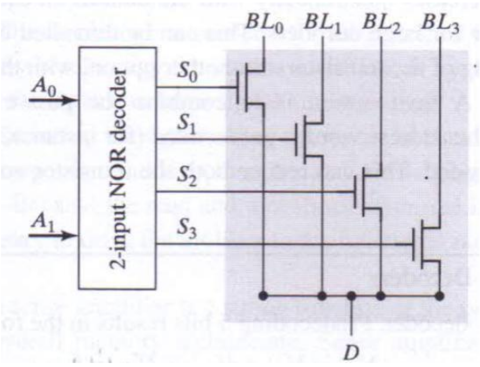
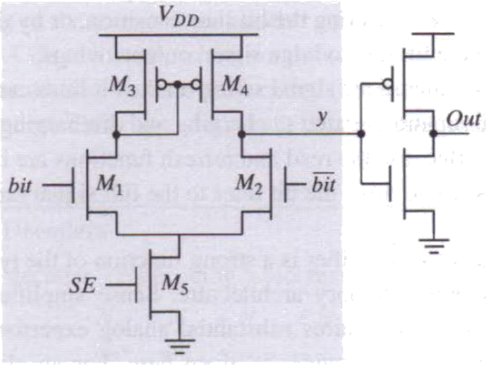


Figure: Four-input pass-transistor-based column decoder using a NOR predecoder

(ii) Sense Amplifier

- Sense amplifiers play a major role in the functionality, performance and reliability of memory circuits.
- Basic differential sense amplifier circuit shown in below figure.
- It performs the following performances



Amplification:

- In memory structures such as the 1T DRAM, amplification is required for proper functionality.

Delay Reduction:

- The amplifier compensates for the fan-out driving capability of the memory cell by detecting and amplifying small transitions on the bit line to large signal output swings.

Power reduction:

- Reducing the signal swing on the bit lines can eliminate large part of the power dissipation related to charging and discharging the bit lines.

(iii) Drivers/ Buffers

- The length of word and bit lines increases with increasing memory sizes.
- Large portion of the read and write access time can be attributed to the wire delays.
- A major part of the memory-periphery area is allocated to the drivers (address buffers and I/O drivers).

****

3.11: Low Power Memory design:

Discuss about Low power memory design.

(i) Active Power Reduction:

- Voltage reduction done by either an increase in the size of the storage capacitor and/or a noise reduction.

Techniques for power reductions:

- Half- V DD precharge:
  - Precharging a bit line to $V_{DD}/2$. It helps to reduce the active power dissipation in DRAM memories by a factor of 2.
- Boosted word line:
  - Raising the value of the word line above V_{DD} during a write operation, eliminates the threshold drop over the access transistor, yielding a substantial increase in stored charge.
- Increased capacitor area or value:
  - Keeping the "ground" plate of the storage capacitor at $V_{DD/}2$ reduces the maximum voltage over Cs, making it possible to use thinner oxides.
- Increasing the cell size:
  - Ultra-low-voltage DRAM memory operation might require a sacrifice in area efficiency.

(ii) Retention current Reduction:

- SRAM array should not have any static power dissipation. But the leakage current of the transistor will be the major problem and this is the main source of the retention current.
- This retention current can be reduced by the following factors.

1. Turn OFF unused memory blocks
2. Negative biasing voltage of the cells which are not active, thus reduce the leakage current.
3. If low threshold voltage transistor is inserted between V_{DD} and SRAM array, leakage reduces.
4. Leakage is a function of V_{DD}, thus if supply rail is lowered, then leakage current is reduced.

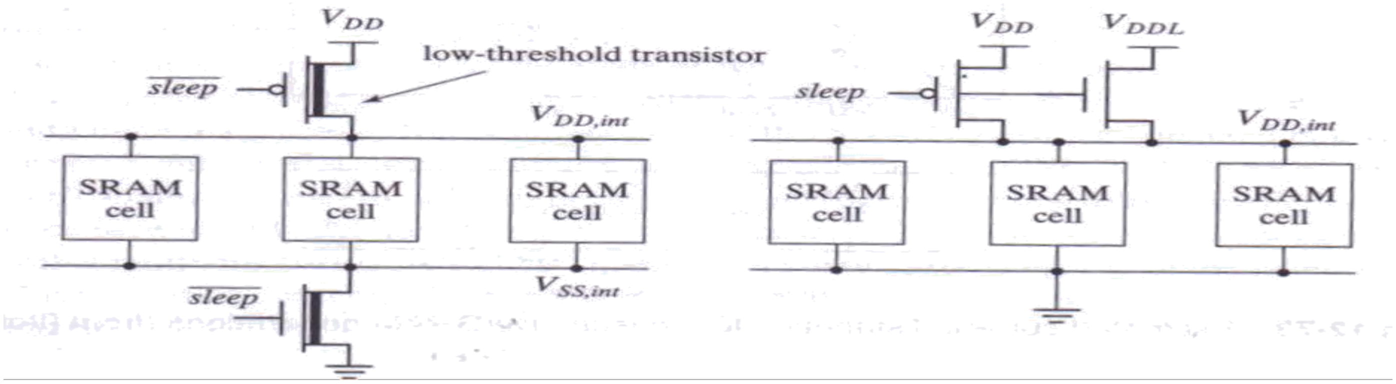


Figure: (a) Insertion of low threshold device (b) Reducing supply Voltage

UNIT – IV
DESIGNING ARCHITECTURE BUILDING BLOCKS

Data path circuits, Architectures for ripple carry adders, carry look ahead adders, High speed adders, accumulators, Multipliers, dividers, Barrel shifters, speed and area tradeoff.

I. Design of Data path circuits:

Discuss about data path circuits.

- Data path circuits are meant for passing the data from one segment to other segment for processing or storing.
- The data path is the core of processors, where all computations are performed.
- It is generally defined with general digital processor. It is shown in figure.

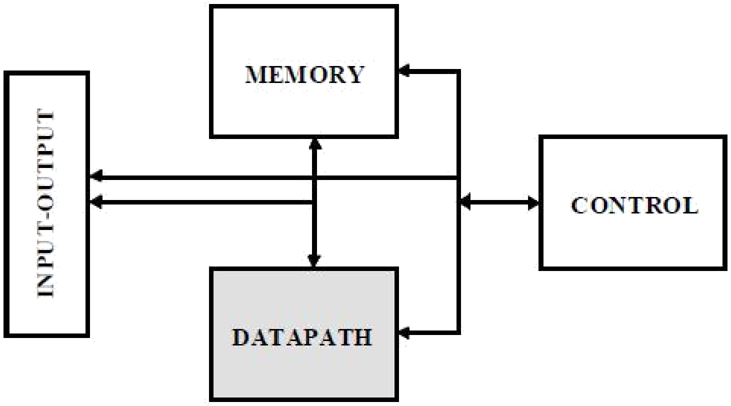
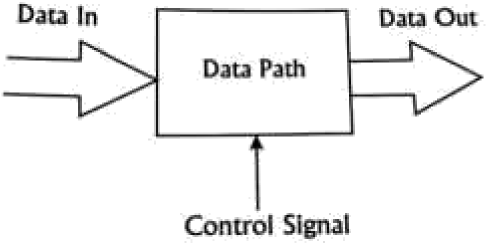
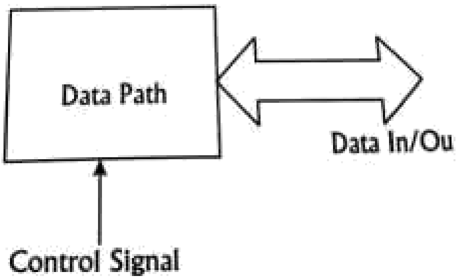


Figure: General digital processor

- If only data path and its communication is shown as



- In this, data is applied at one port and data output is obtained at second port.



- Data path block consists of arithmetic operation, logical operation, shift operation and temporary storage of operands.

- Data paths are arranged in a bit sliced organization.
- Instead of operating on single bit digital signals, the data in a processor are arranged in a word based fashion.
- Bit slices are either identical or resemble a similar structure for all bits.
- The data path consists of the number of bit slices (equal to the word length), each operating on a single bit. Hence the term is ***bit-sliced.***

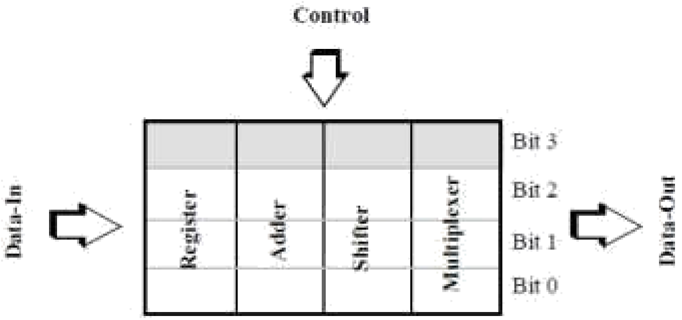


Figure: Bit-sliced datapath organization

****

II. Ripple Carry Adder:

❖ **Draw the structure of ripple carry adder and explain its operation. (Nov 2017)**

❖ **Explain the operation of a basic 4 bit adder. (Nov 2016)**

Architecture of Ripple Carry Adder:

- AOI Full adder circuit (AND OR INVERT)
- An AOI algorithm for static CMOS logic circuit can be obtained by using the equation.

$$C_{i+1} = a_i\, b_i + c_i\,.(a_i + b_i)$$

$$\overline{S}_i = (a_i + b_i + \overline{c_i})c_i + (a_i\,.b_i\,.c_i)$$

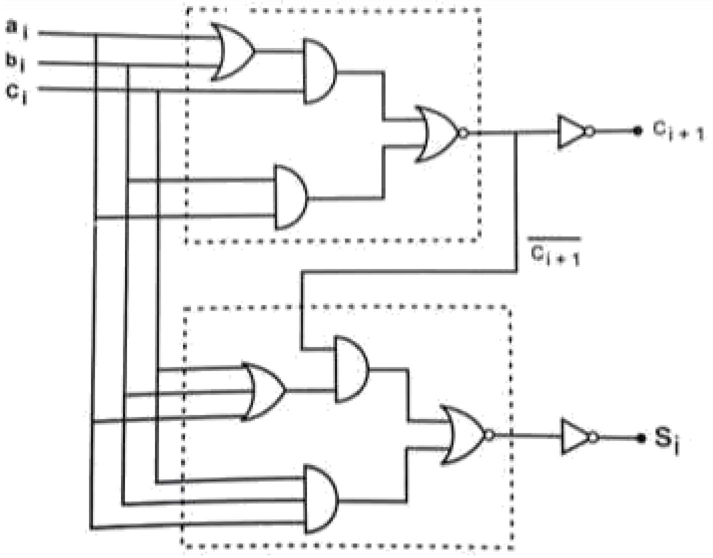


Figure: AOI Full adder

- If n bits are added, then we can get n-bit sum and carry of C_n. C_i= Carry bit from the previous column.
- N bit ripple carry adder needs n full adders with C_{i+1} carry out bit.

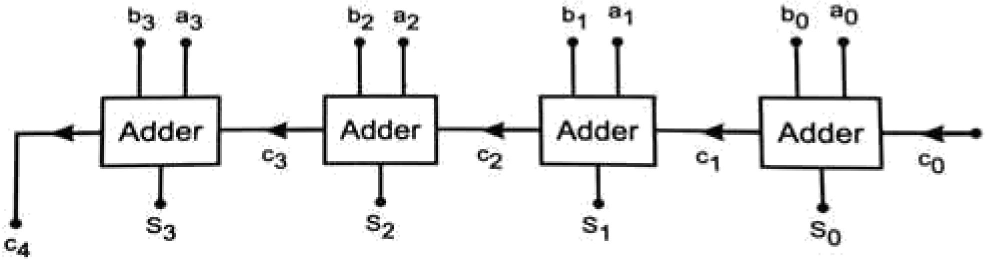


Figure: Ripple carry adder

- The overall delay depends on the characteristics of full adder circuit. Different CMOS implementation can produce different delay parts.
- t_{di}- worst case delay through the i^{th} stage. We can calculate the total delay using the following equation

$$t_{4b} = t_{d3}+t_{d2}+t_{d1}+t_{d0} \text{ and } t_{d0} = t_d (a_0, b_0 \rightarrow c_1)$$

- This is the time for the input to produce the carry out bit.

$$t_{d1} = t_{d2} = t_d (c_{in} \rightarrow c_{out})$$

$$t_{d3} = t_d (c_{in} \rightarrow S_3)$$

$$t_{4b} = t_d (c_{in} \rightarrow S_3) +2t_d (c_{in} \rightarrow c_{out}) + t_d (a_0, b_0 \rightarrow c_1)$$

- If it is extend to n-bit, then the worst case delay is
- Worst case delay linear with the number of bits

$$t_d = O(N)$$

$$t_{adder} = (N-1)t_{carry} + t_{sum}$$

- The figure below shows 4-bit adder/subtractor circuit.
- In this, if add/sub=0, then sum is a+b. If add/sub=1, then the output is a-b.

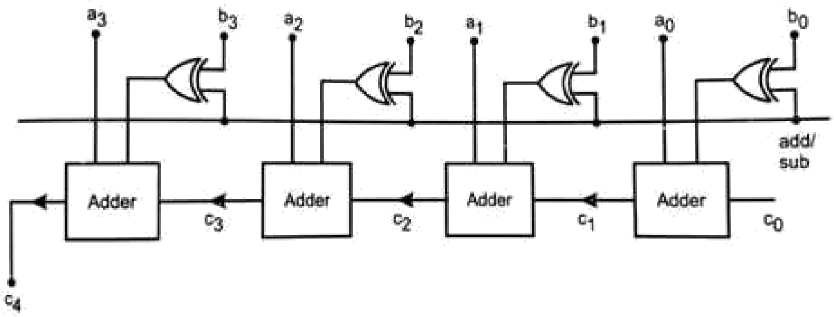


Figure: 4-bit adder/subtractor circuit

- Sum and carry expressions are designed using static CMOS.
- It requires 28 transistors which lead large area and circuit is slow.
  Sum, S= + $\bar{0}$(+ +) and Carry, C0= + +

Drawbacks:

- Circuit is slower.

- In ripple carry adder, carry bit is calculated along with the sum bit. Each bit must wait for calculation of previous carry.

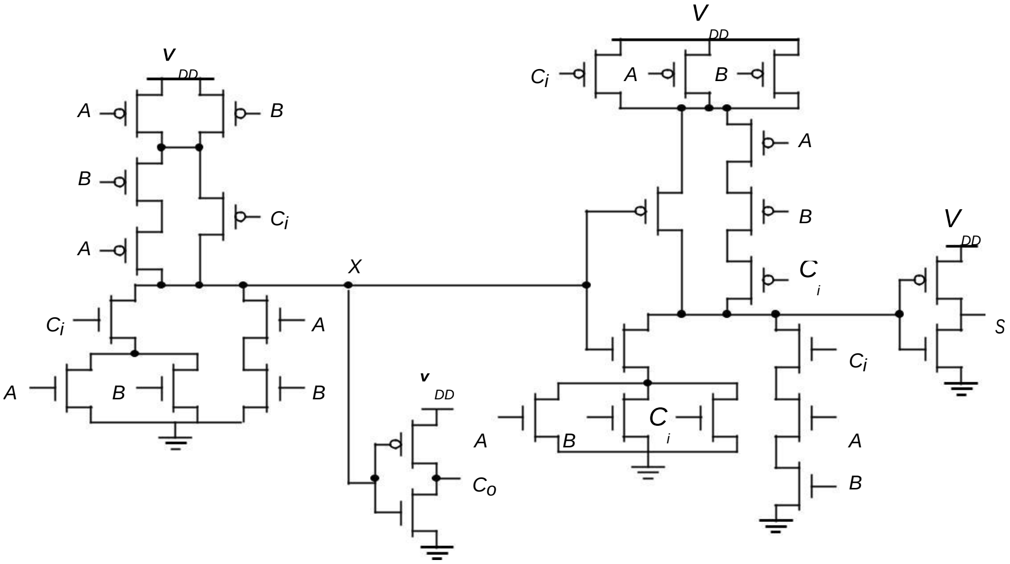


Figure: Complimentary Static CMOS Full Adder

****

III. Carry Look Ahead Adder (CLA):

❖ **Explain the operation and design of Carry lookahead adder (CLA). (May 2017, Nov 2016)**

❖ **How the drawback in ripple carry adder overcome by carry look ahead adder and discuss. (Nov 2017)**

- A carry-lookahead adder (CLA) is a type of adder used in digital circuit.
- A carry-lookahead adder improves speed by reducing the amount of time required to determine carry bits.
- In ripple carry adder, carry bit is calculated alongwith the sum bit.
- Each bit must wait until the previous carry is calculated to begin calculating its own result and carry bits.
- The carry-lookahead adder calculates one or more carry bits before the sum, which reduces the wait time to calculate the result of the larger value bits.
- A ripple-carry adder works starting at the rightmost (LSB) digit position, the two corresponding digits are added and a result obtained. There may be a carry out of this digit position.
- Accordingly all digit positions other than LSB. Need to take into account the possibility to add an extra 1, from a carry that has come in from the next position to the right.
- Carry lookahead depends on two things:
  - ✓ Calculating, for each digit position, whether that position is going to propagate a carry if one comes in from the right.
  - ✓ Combining these calculated values to be able to realize quickly whether, for each group of digits, that group is going to propagate a carry.

- Theory of operation:
  - ✓ Carry lookahead logic uses the concept of generating and propagating carry.
  - ✓ The addition of two 1-digit inputs A and B is said to generate if the addition will carry, regardless of whether there is an input carry.
- *Generate:*
  - ✓ In binary addition, A + B generates if and only if both A and B are 1.
  - ✓ If we write G(A,B) to represent the binary predicate that is true if and only if A + B generates, we have:

$$\boldsymbol{G(A,B) = A . B}$$

- *Propagate:*
  - ✓ The addition of two 1-digit inputs A and B is said to propagate if the addition will carry whenever there is an input carry.
  - ✓ In binary addition, A + B propagates if and only if at least one of A or B is 1.
  - ✓ If we write P(A,B) to represent the binary predicate that is true if and only if A + B propagates, we have:

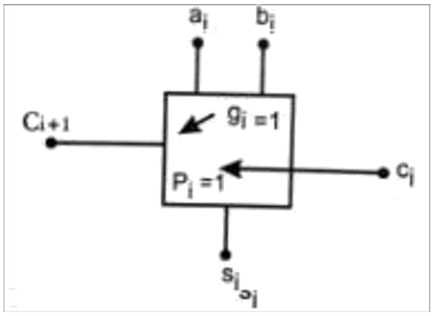
$$P (A, B) = A \oplus B$$

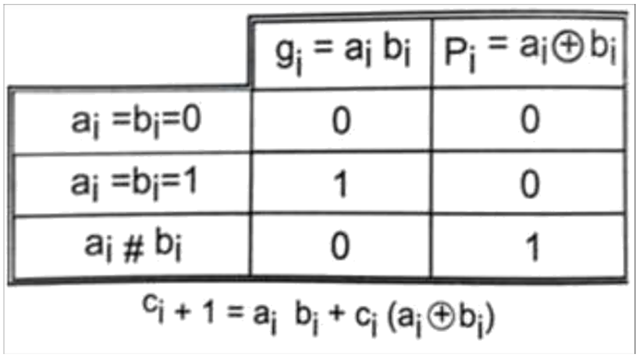
- These adders are used to overcome the latency which is introduced by the rippling effect of carry bits.
- Write carry look-ahead expressions in terms of the generate g_i and propagate p_i signals. The general form of carry signal c_i thus becomes

$$c_{i+1} = a_i .b_i + c_i .(a_i \oplus b_i) = g_i + c_i .p_i$$

- If $a_i .b =1$, then $c_{i+1} = 1$, write generate term as, $g_i = a_i .b_i$
- Write the propagate term as, $p_i = a_i \oplus b_i$
- Sum and carry expression are written as,

$$S_i = a_i \oplus b_i$$
$$c_1=g_0+p_0.c_0$$
$$c_2=g_1+p_1.c_1= g_1+p_1.(g_0+p_0.c_0)$$
$$c_3=g_2+p_2.c_2$$
$$c_4=g_3+p_3.c_3 = g_3+p_3.g_2+ p_3.p_2.g_1+ p_3.p_2.p_1.g_0 + p_3.p_2.p_1.p_0.c_0$$





	$g_i = a_i b_i$	$P_i = a_i \oplus b_i$
$a_i = b_i = 0$	0	0
$a_i = b_i = 1$	1	0
$a_i \# b_i$	0	1

$c_i + 1 = a_i\ b_i + c_i (a_i \oplus b_i)$

Figure: Symbol and truth table of generate & propagate

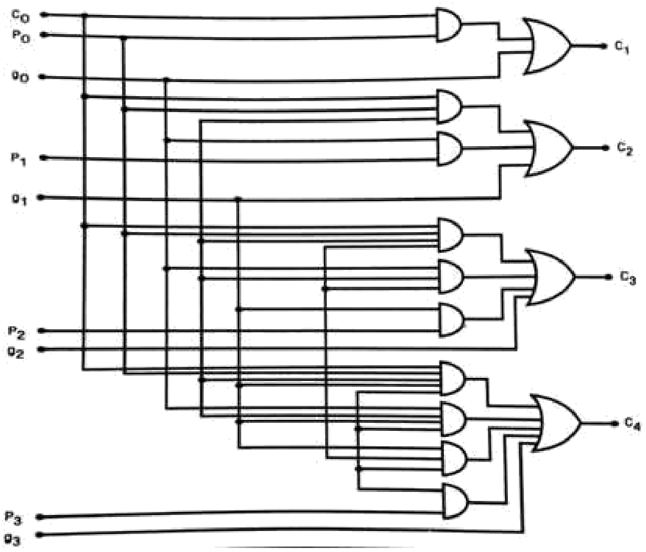


Figure – Logic network for 4-bit CLA carry bits

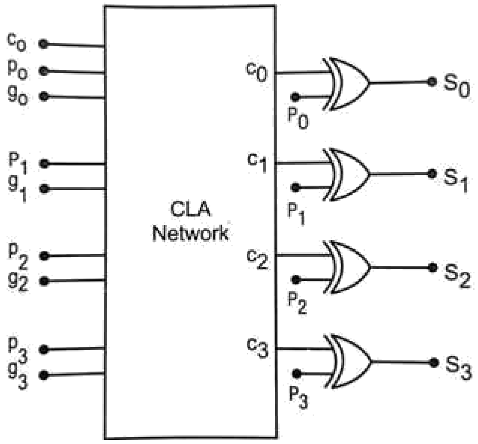


Figure – Sum calculation using the CLA network

- The symmetry in the array is shown in mirror. It allows more structured layout at the physical design level.

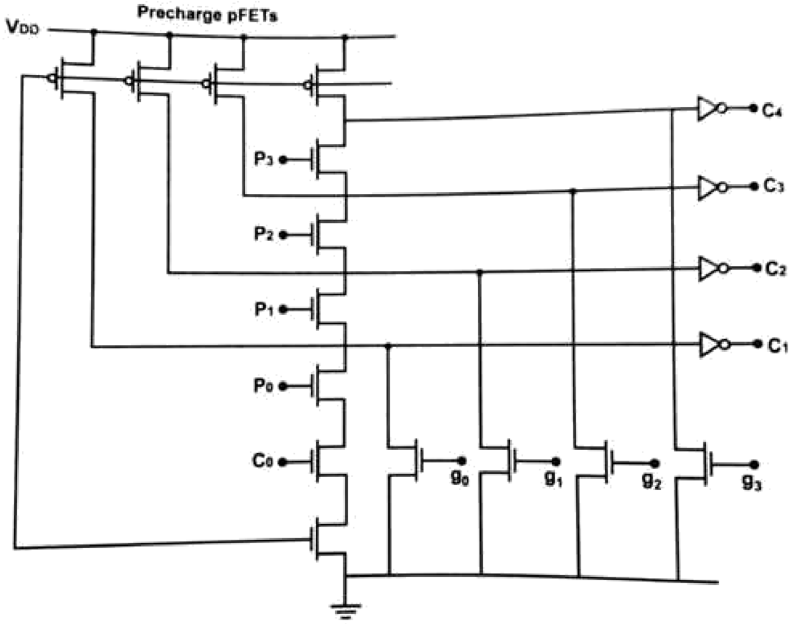


Figure – MODL carry circuit

- MODL-Multiple Output Domino Logic.
- MODL is non-inverting logic family and is a dynamic circuit technique.
- Its limitations are
  - i. Clocking in mandatory
  - ii. The output is subject to charge leakage and charge sharing.
  - iii. Series connected nFET chains can give long discharge times.

****

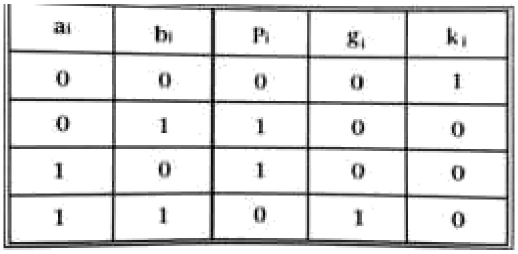
IV. Manchester Carry Chain Adder:

Discuss about Manchester Carry Chain Adder.

- The Manchester carry chain is a variation of the carry-lookahead adder that uses shared logic to lower the transistor count.
- A Manchester carry chain generates the intermediate carries by tapping off nodes in the gate that calculates the most significant carry value.
- Dynamic logic can support shared logic, as transmission gate logic.
- One of the major drawbacks of the Manchester carry chain is increase the propagation delay.
- A Manchester-carry-chain section generally won't exceed 4 bits.
- In this adder, the basic equation is $c_{i+1} = g_i + c_i . p_i$

Where $p_i = a_i \oplus b_i$ and $g_i = a_i . b_i$

- Carry kill bit $k_i = \overline{a_i + b_i} = \overline{a_i} . \overline{b_i}$
- If $K_i=1$, then $p_i=0$ and $g_i=0$. Hence, k_i is known as carry kill bit.



a_i	b_i	P_i	g_i	k_i
0	0	0	0	1
0	1	1	0	0
1	0	1	0	0
1	1	0	1	0

Table

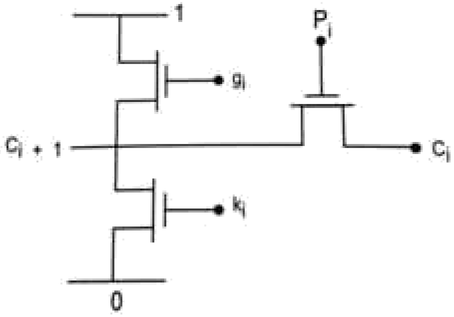
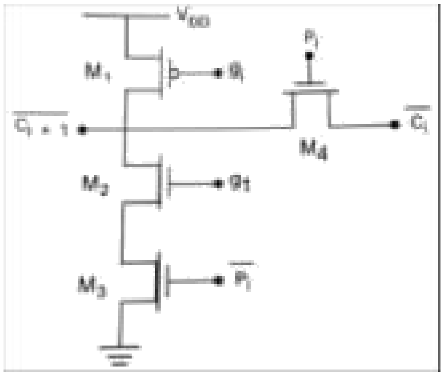


Figure – switch level circuit

- In the circuit shown below $\overline{C_l}$ is used as an input if $P_i = 0$, then M_3 is ON, M_4 is OFF.
- If $g_i=0$, then M_1 is ON, M_2 is ON

- If g_i=1, then M_2 is OFF, M_4 is ON and output equal to zero.
- If P_i=1, then this case is a complicated one.



- In dynamic circuit figure
- If ϕ = 0, then recharge occur and output is 1
- If ϕ = 1, then evaluation occur.

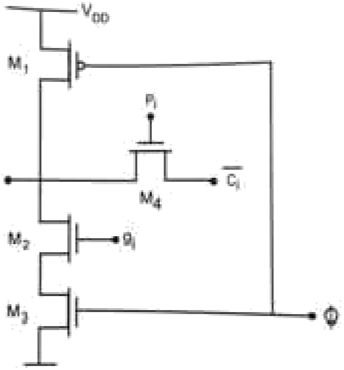
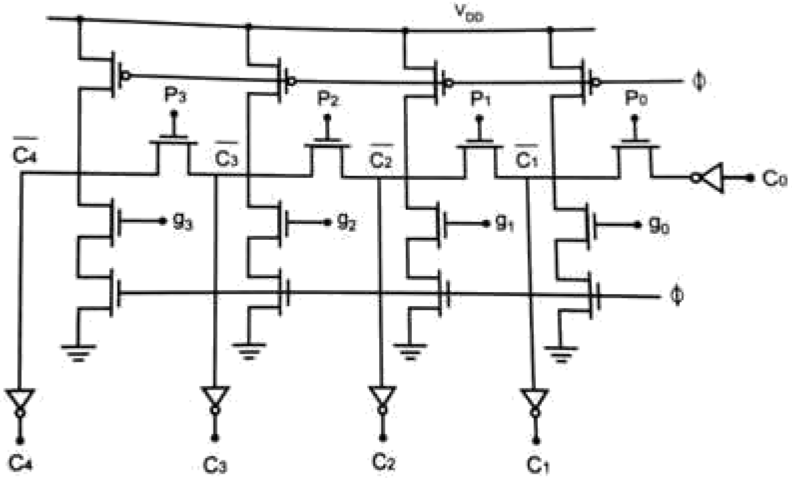


Figure dynamic circuit

- Dynamic Manchester carry chain for the carry bit upto C_4 is shown below. C_1, C_2, C_3, C_4 can be taken by using inverters. The carry input is given as C_0



****

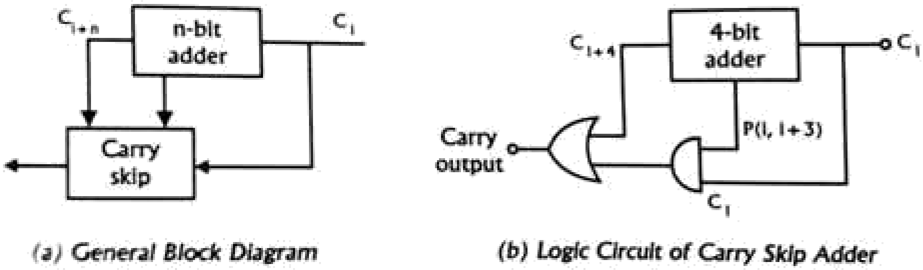
V. HIGH SPEED ADDERS:

- ❖ **Discuss about different types of high speed adders. (Apr. 2016)**
- ❖ **Describe the different approaches of improving the speed of the adder. (Nov 2016)**

(i) Carry Skip(bypass) Adder:

Design a carry bypass adder and discuss its features. (May 2016)

- It is high speed adder. It consist of adder, AND gate and OR gate.
- An incoming carry $C_{i,0}=1$ propagates through the complete adder chain and an outgoing carry $C_{0,3}=1$.
- In other words, if ($P_0P_1P_2P_3=1$) then $C_{0,3}= C_{i,0}$ else either DELETE or GENERATE occurred.
- It can be used to speed up the operation of the adder, as shown in below fig (b).



(a) General Block Diagram (b) Logic Circuit of Carry Skip Adder

Figure: Carry Skip Adder.

- When BP= $P_0P_1P_2P_3=1$, the incoming carry is forwarded immediately to the next block.
- Hence the name carry bypass adder or carry skip adder.
- Idea: if (P_0 and P_1 and P_2 and $P_3=1$) the $C_{03} = C_0$, else “kill” or “generate”.

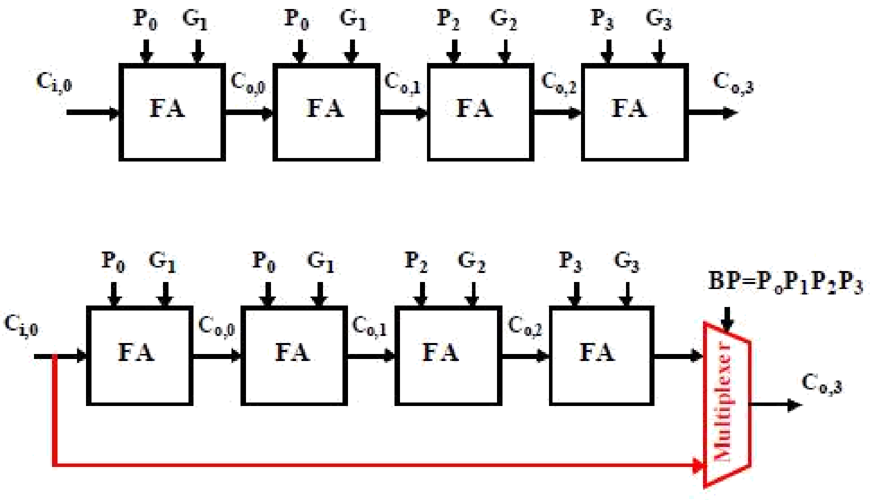
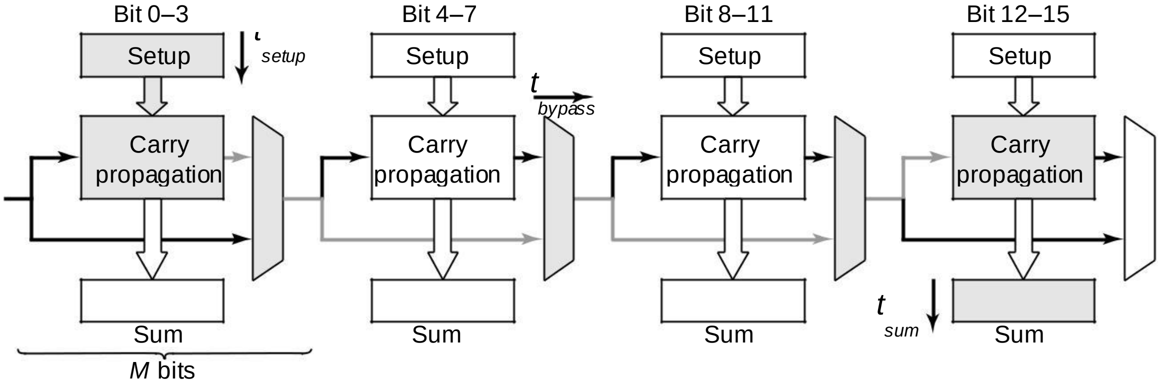


Figure: (a) Carry propagation (b) Adding a bypass

- The below figure shows n no. of bits carry skip adder.



$t_{adder} = t_{setup} + Mt_{carry} + (N/M-1)t_{bypass} + (M-1)\ t_{carry} + t_{sum}$ (worst case)

t_{setup}*: overhead time to create G, P, D signals*

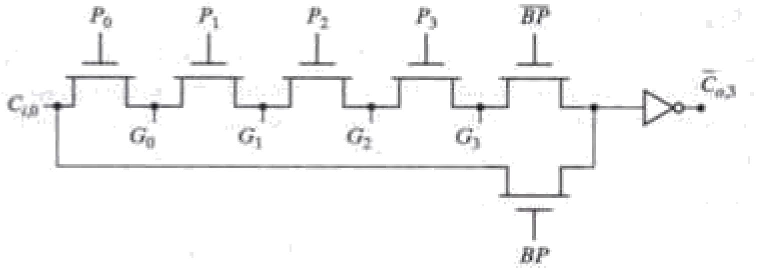


Figure: Manchester carry-chain implementation of bypass adder

(ii) Carry Select Adder:

Design a carry select adder and discuss its features. (May 2016)

- A carry-select adder is a particular way to implement an adder, which is a logic element that computes the (n+1)-bit sum of two n-bit numbers.
- The carry-select adder is simple but rather fast, having a gate level depth of O($\sqrt{n}$).
- The carry-select adder generally consists of two ripple carry adders and a multiplexer.
- Adding two n-bit numbers with a carry-select adder is done with two adders in order to perform the calculation twice.
- One time with the assumption of the carry-in being zero and the other assuming it will be one.
- After the two results are calculated (the correct sum as well as the correct carry-out), it is then selected with the multiplexer once the correct carry-in is known.
- The number of bits in each carry select block can be uniform, or variable.
- In the uniform case, the optimal delay occurs for a block size of $\sqrt{n}$.
- The O($\sqrt{n}$) delay is derived from uniform sizing, where the ideal number of full-adder elements per block is equal to the square root of the number of bits being added.
- Propagation delay, P is equal to $\sqrt{2}$ where N = N-bit adder
- Below is the basic building block of a carry-select adder, where the block size is 4.
- Two 4-bit ripple carry adders are multiplexed together, where the resulting carry and sum bits are selected by the carry-in.

-

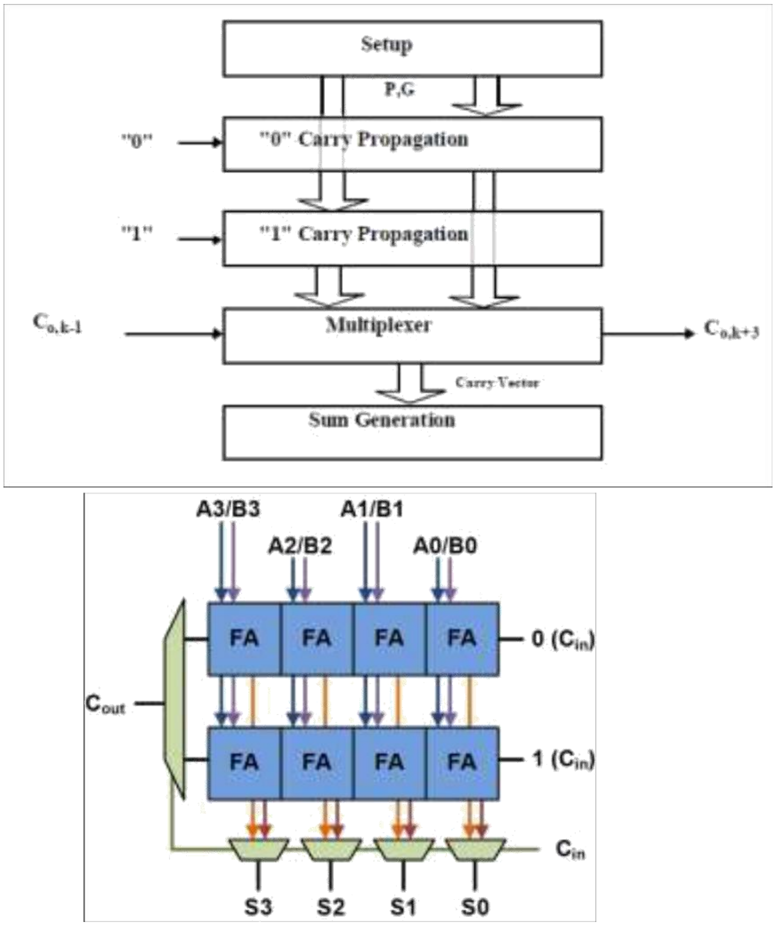
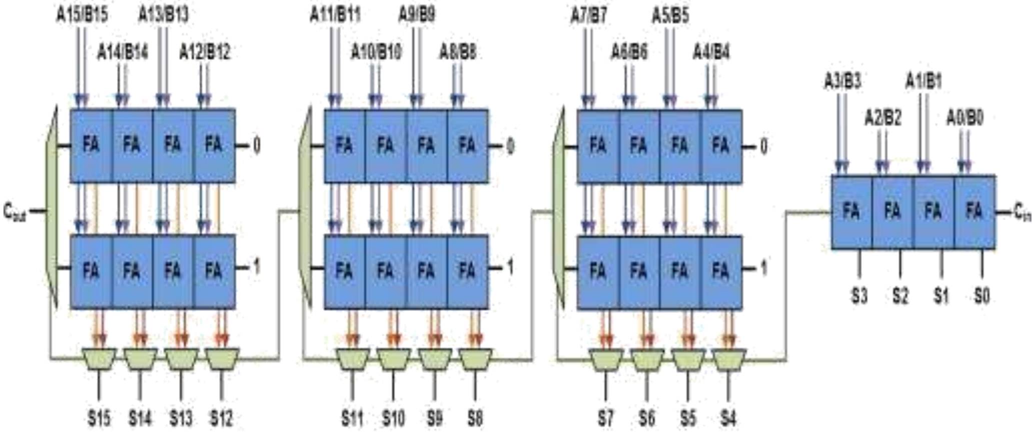


Figure: Building blocks of a carry-select adder

Uniform-sized adder:

- A 16-bit carry-select adder with a uniform block size of 4 can be created with three of these blocks and a 4-bit ripple carry adder.
- Since carry-in is known at the beginning of computation, a carry select block is not needed for the first four bits.
- The delay of this adder will be four full adder delays, plus three MUX delays.
- $t_{adder} = t_{setup} + Mt_{carry} + (N/M)t_{mux} + t_{sum}$



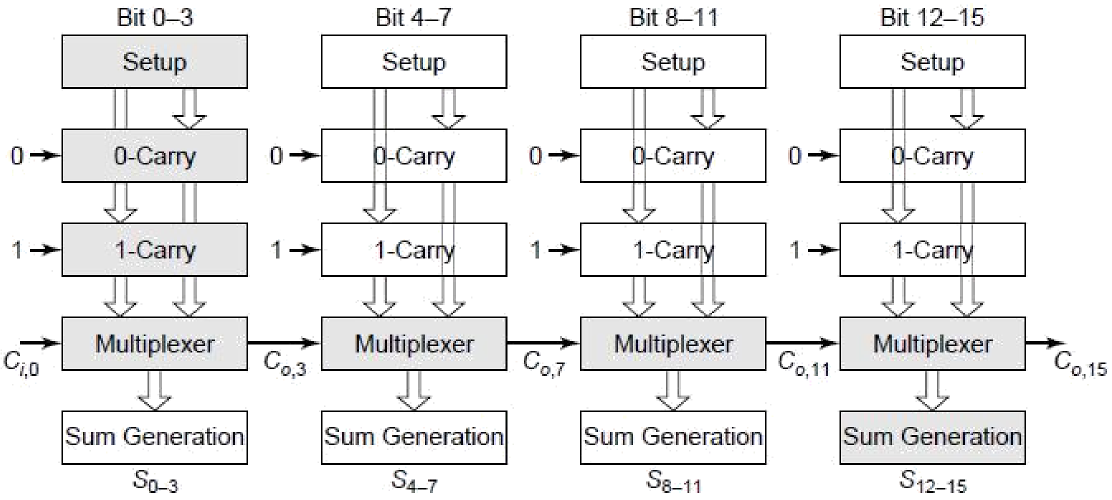
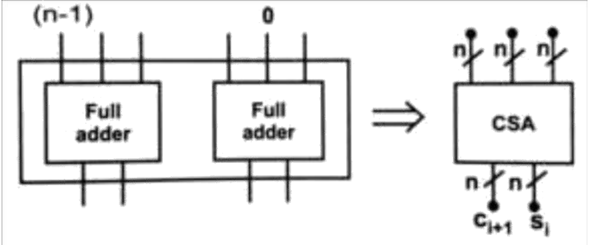


Figure: general structure of 16 bit adder

Disadvantage: hardware cost is increased.

(iii) Carry Save Adder:

- Carry save adder is similar to the full adder. It is used when adding multiple numbers.
- All the bits of a carry save adder work in parallel.
- In carry save adder, the carry does not propagate. So, it is faster than carry propagate adder.
- It has three inputs and produces 2 outputs, carry-out is saved. It is not immediately used to find the final sum value.



n-bit Carry Save Adder

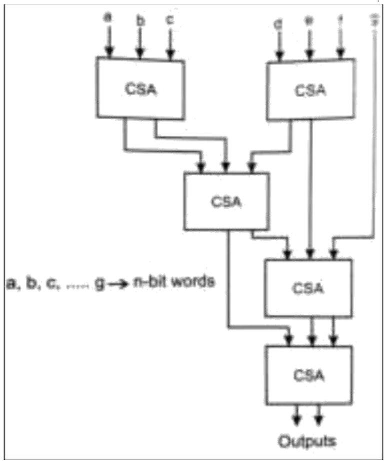


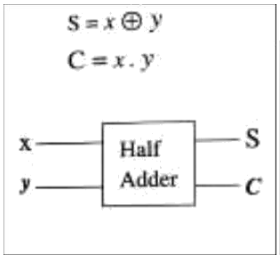
Figure: ***Carry Save Adder***

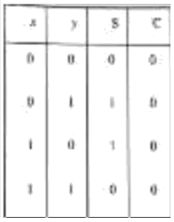
****

V. ACCUMULATOR:

Briefly discuss about accumulators.

- Accumulator acts as a part of ALU and it is identified as register A. The result of an operation performed in the ALU is stored in the accumulator.
- It is used to hold the data for manipulation (arithmetic and logical)
- Arithmetic functions are very important in VLSI. Ex: multiplication.
- Half adder circuit has two inputs and two outputs. S = $x \oplus y$, C = x.y.

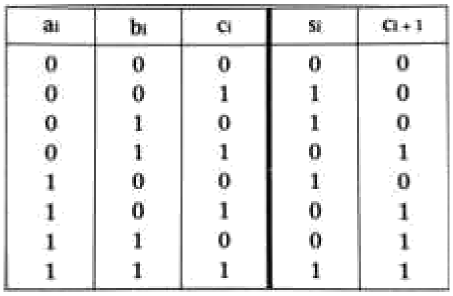




x	y	S	C
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	0

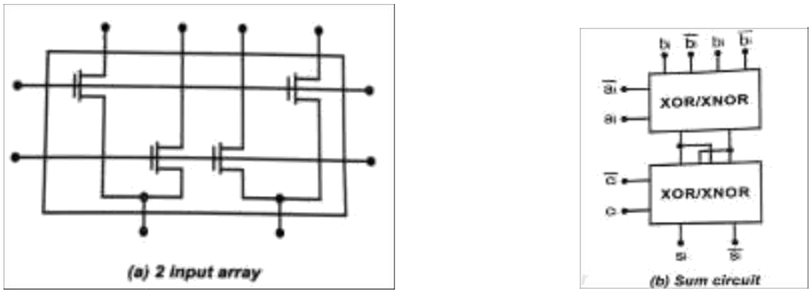
Figure: Half adder and Truth table

- Full adder circuit has three inputs and two outputs



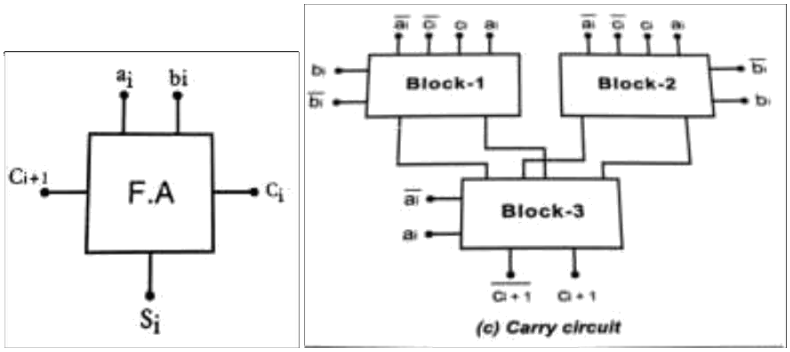
a_i	b_i	c_i	s_i	c_{i+1}
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Figure : Full adder and truth table CPL --- Complementary Pass Logic



(a) 2 input array

(b) Sum circuit



(c) Carry circuit

Figure : CPL Full adder design

****

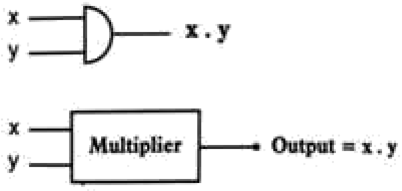
VI. MULTIPLIERS:

❖ **Explain the design and operation of multiplier circuit. (Apr. 2016, Nov 2016, May 2017)**

❖ **Design a multiplier for 5 bit by 3 bit. Explain its operation and summarize the numbers of adders. Discuss it over Wallace multiplier. (Nov 2017)**

- A study of computer arithmetic processes will reveal that the most common requirements are for addition and subtraction.
- There is also a significant need for a multiplication capability.
- Basic operations in multiplication are given below.

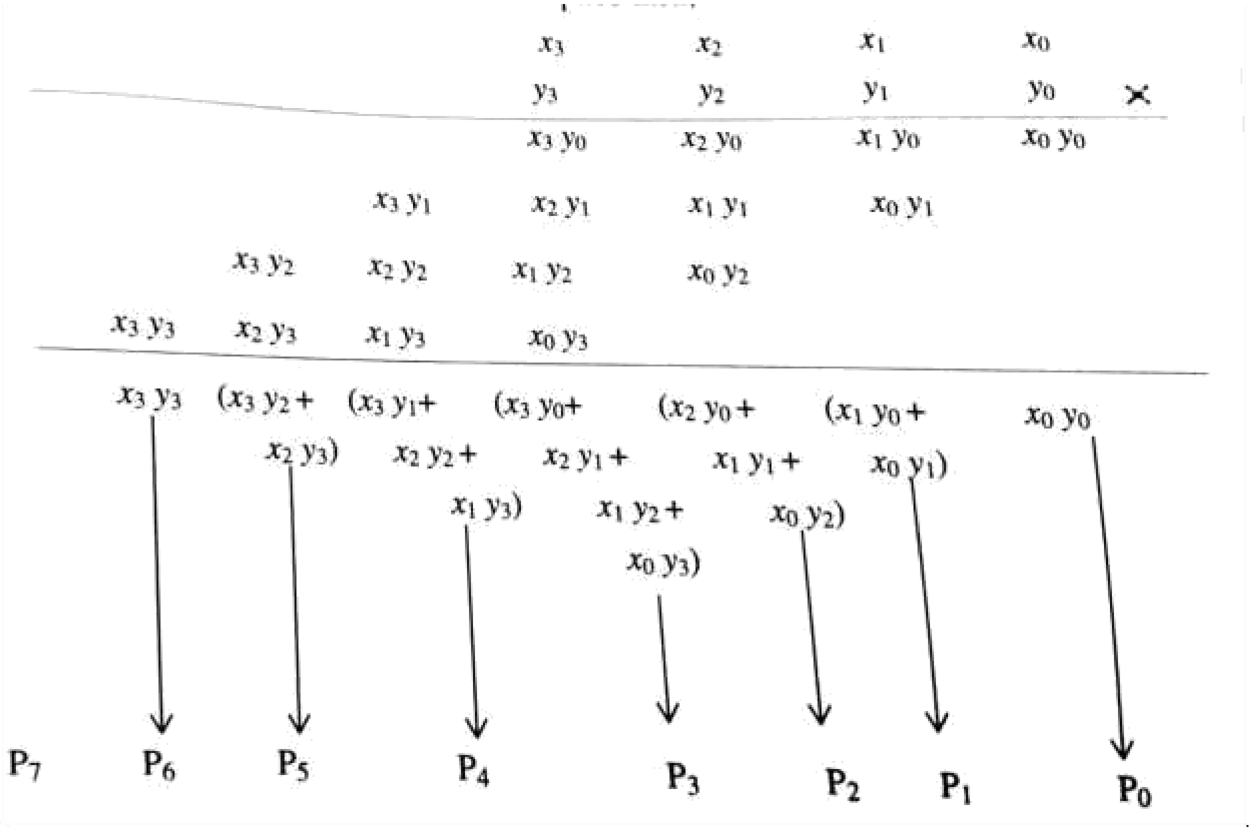
  0 x 0 = 0, 0 x 1 = 0, 1 x 0 = 0, 1 x 1 = 1



x	y	x . y
0	0	0
0	1	0
1	0	0
1	1	1

				1	0	1	0	1	0	Multiplicand
x						1	0	1	1	Multiplier
				1	0	1	0	1	0	Partial products
			1	0	1	0	1	0		
		0	0	0	0	0	0			
+	1	0	1	0	1	0				
	1	1	1	0	0	1	1	1	0	Result

- If two different 4-bit numbers (x_0, x_1, x_2, x_3 & y_0, y_1, y_2, y_3)are multiplied then



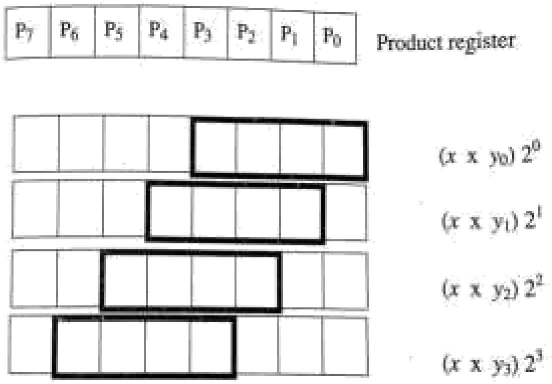
Product bits → ($P_7\ P_6\ P_5\ P_4\ P_3\ P_2\ P_1\ P_0$)

$$P_i = \sum_{i=j+k} x_j y_k + c_{i-1}$$

$$c_{i-1} = 0 \text{ for } (i-1) \le 0$$

Multiplication by shifting:

- If $x=(0020)_2 = (2)_{10}$
- If it is to be multiplied by 2, then we can shift x in left side. $x = (0100)_2 = (4)_{10}$
- If it is to be divided by 2, then we can shift in right side. $x = (0001)_2 = (1)_{10}$.
- So, shift register can be used for multiplication or division by 2.



- A practical implementation is based on the sequence. The product is obtained by successive addition and shift right operations

(i) Array multiplier:

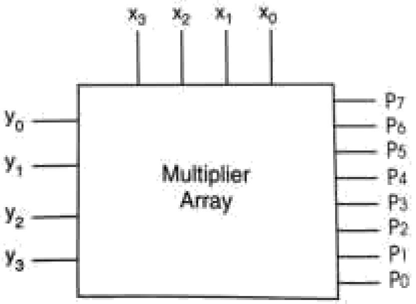


Figure: General block diagram of multiplier

- Array multiplier uses an array of cells for calculation.
- Multiplier circuit is based on repeated addition and shifting procedure. Each partial product is generated by the multiplication of the multiplicand with one multiplier digit.
- The partial products are shifted according to their bit sequences and then added.
- N-1 adders are required where N is the number of multiplier bits.
- The method is simple but the delay is high and consumes large area by using ripple carry adder for array multiplier. Product expression is given below

$$P_i = \sum_{i=j+k} x_j y_k + c_{i-1}$$

$$P = X.Y = \left(\sum_{j=0}^{n-1} x_j 2^j\right)\left(\sum_{k=0}^{n-1} y_k 2^k\right)$$

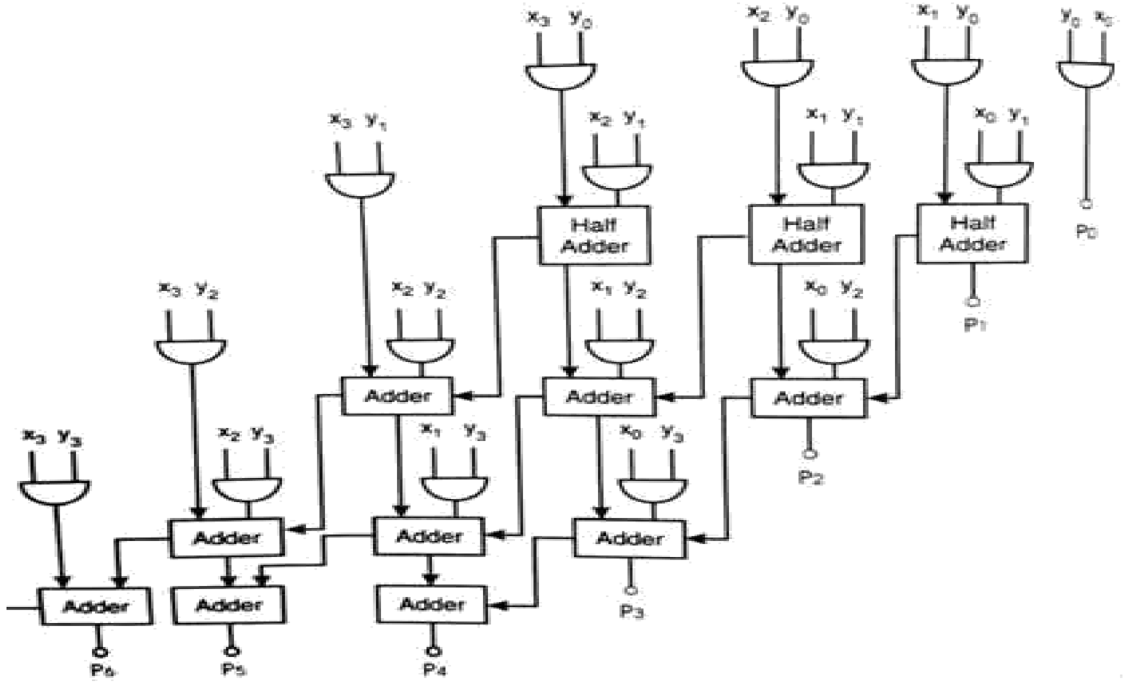


Figure: 4 x 4 array multiplier

- This multiplier can accept all the inputs at the same time. An array multiplier for n-bit word need n(n-2) full adders, n-half adder and n^2 AND gates.

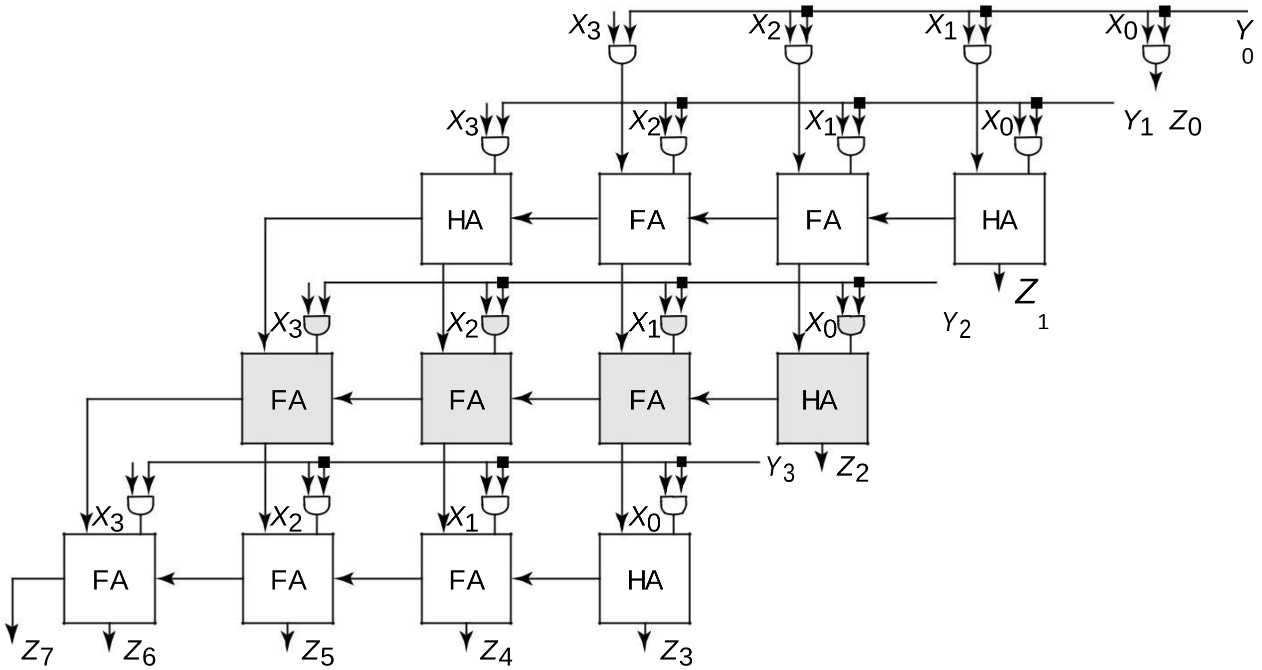


Figure: 4 x 4 array multiplier using Fulladder, Halfadder and AND gate.

(iv) Booth (encoding) multiplier:

- Booth's algorithm is an efficient hardware implementation of a digital circuit that multiplies two binary numbers in two's complement notation.
- Booth multiplication is a fastest technique that allows for smaller, faster multiplication circuits, by recoding the numbers that are multiplied.

- The Booths multipliers widely used in ASIC oriented products due to the higher computing speed and smaller area.
- In the binary number system, the digits called bits are to the set of {0,1}.
- The result of multiplying any binary number by binary bit is either 0 or original number.
- This makes the formation of partial products are more efficient and simple.
- Then adding all these partial products is time consuming task for any binary multipliers.
- The entire process consists of three steps partial product generation, partial product reduction and addition of partial products as shown in figure.

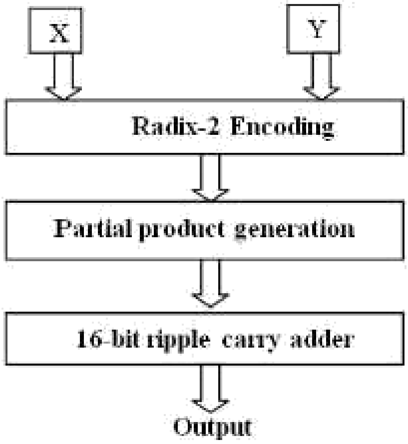


Figure: Block diagram of Booth multiplier

- But in booth multiplication, partial product generation is done based on recoding scheme e.g. radix 2 encoding.
- Bits of multiplicand (Y) are grouped from left to right and corresponding operation on multiplier (X) is done in order to generate the partial product.
- In radix-2 booth multiplication partial product generation is done based on encoding which is as given by Table.

Q_n	Q_{n+1}	Recoded Bits	Operation
0	0	0	Shift
0	1	+1	Add X
1	0	-1	Subtract X
1	1	0	Shift

Table: Booth encoding table with RADIX-2

- RADIX-2 PROCEDURE:
  1) Add 0 to the LSB of the multiplier and make the pairing of 2 from the right to the left which shown in the figure.

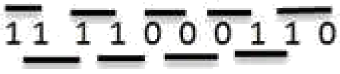


Figure: 2- Bit pairing as per Booth recoding using Radix- 2.

2) 00 and 11: do nothing according to the encoding table.

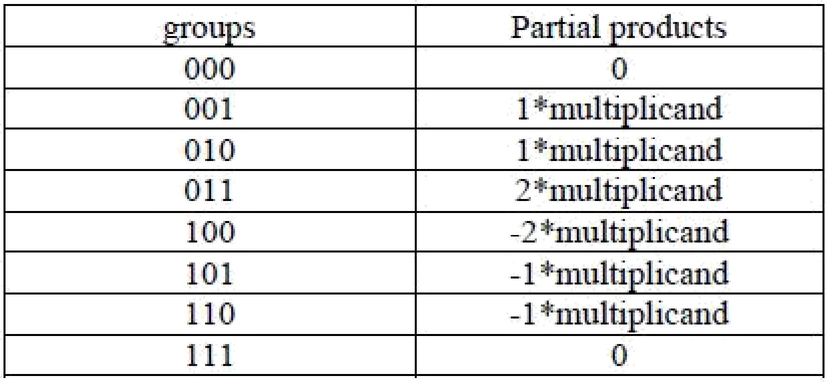
3) 01: mark shows the end of the string' of 1and add multiplicand to the partial product.
4) 10: mark shows beginnings of the string of 1 subtract multiplicand from partial product.

Modified Booth Multiplier using Radix -4:

- **The disadvantage of Booth Multiplier with Radix-2 is increasing partial products.**
- **Modified Booth Multiplier with Radix-4 is reducing the half of the partial products in multipliers.**
- Modified Booth multiplication is a technique that allows for smaller, faster circuits by recoding the numbers that are multiplied.
- In Radix-4, encoding the multiplicands based on multipliers bits. It will compare 3-bits at a time with overlapping technique.
- Grouping starts from the LSB and the first block contains only two bits of the multipliers and it assumes zero for the third bit.

Figure. Grouping of 3-bit as per booth recoding

- These group of binary digits are according to the Modified Booth Encoding Table and it is one of the numbers from the set of (-2, 2, 0, 1, -1).



groups	Partial products
000	0
001	1*multiplicand
010	1*multiplicand
011	2*multiplicand
100	-2*multiplicand
101	-1*multiplicand
110	-1*multiplicand
111	0

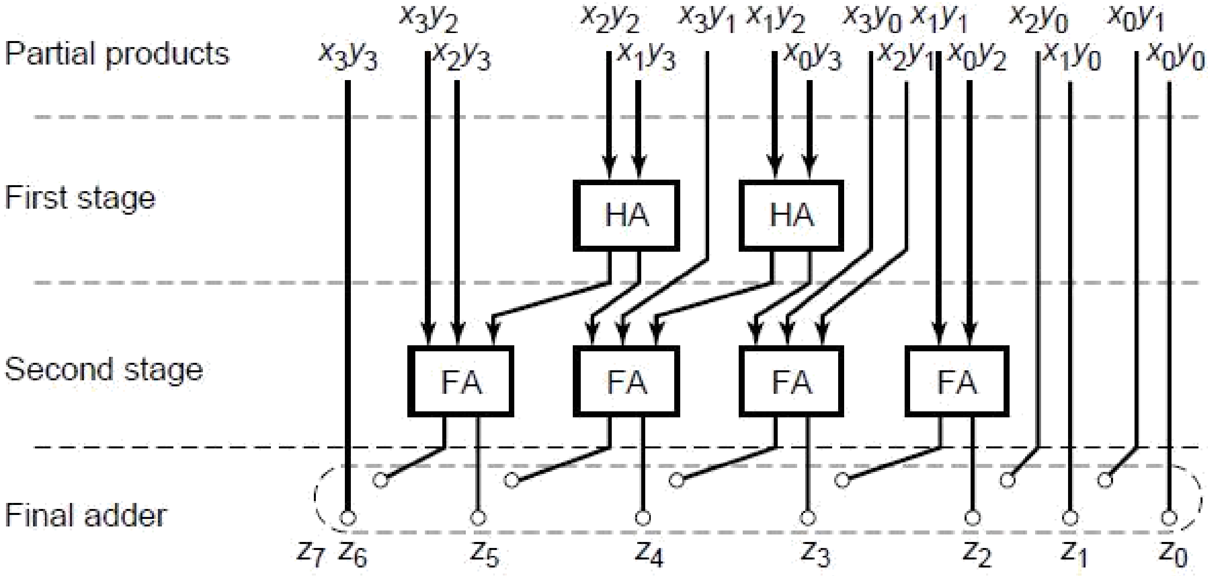
Table: Booth encoding table with RADIX-4

- **RADIX-4 PROCEDURE:**
  1) Add 0 to the right of the LSB of the multiplier.
  2) Extend the sign bit 1 position if it is necessary when n is even.
  3) Value of each vector, the partial product is coming from the set of (-2, 2, 0, 1, -1).

(v) Wallace tree Multiplier:

- A Wallace tree is an efficient hardware implementation of a digital circuit that multiplies two integer numbers.
- The Wallace tree multiplier has three steps to be followed,
  (a) Multiply each bit of one of the arguments, by each bit of the other, yielding n^2 results.
  (b) Reduce the number of partial products to two by layers of full and half adders.
  (c) Group the wires in two numbers and add them with a conventional adder.

- The second section works as follows,
  - (a) Take any three wires with same weights and input them into a full adder. The result will be an output wire of the same weight and an output wire with a higher weight for each three input wires.
  - (b) If there are two wires of the same weight left, input them into a half adder.
  - (c) If there is just one wire left and connects it to next layer.
- The Wallace tree multiplier output structure is tree basis style. It reduces the number of components and reduces the area.
- The architecture of a 4 x4 Wallace tree multiplier is shown in figure.



****

VII. DIVIDERS

Explain in detail about the design and procedure for dividers.

- There are two types of dividers, Serial divider and Parallel divider. Serial divider is slow and parallel divider is fast in performance.
- Generally division is done by repeated subtraction. If 10/3 is to be performed then,

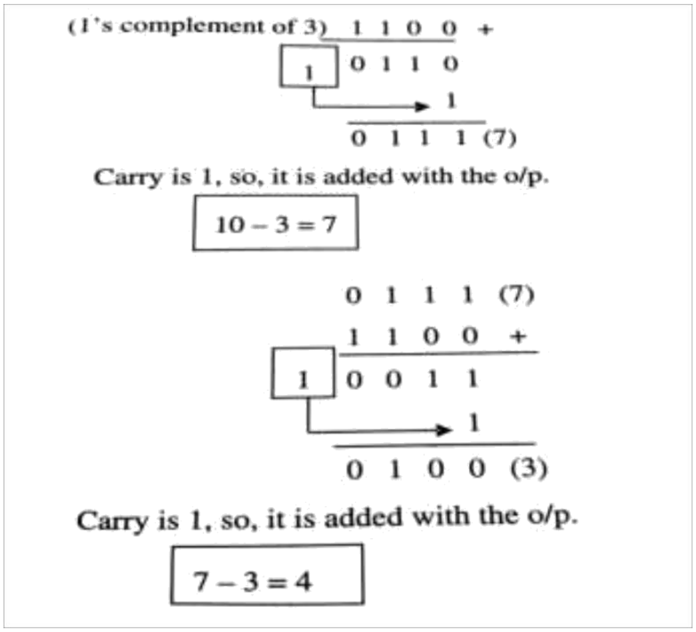
10 -3 =7, (divisor is 3, dividend is 10)
7 – 3 = 4,
4 – 3 = 1

- Here, repeated subtraction has been done, after 3 subtractions, the remainder is 1. It is less than divisor. So now the subtraction is stopped.
- Let see the example of binary division with use of 1's complement method

1010 (10_d) / 0011 (3_d)

Step1: find 1's complement of divisor
Step2: add this with the dividend
Step3: if carry is 1, then it is added with the output to get the difference output
Step4: the same procedure is repeated until we are get carry 0.
Step5: then the process is stopped.

1 0 1 0 (10)



$$\begin{array}{lccccc} \text{(1's complement of 3)} & 1 & 1 & 0 & 0 & + \\ \hline \boxed{1} & 0 & 1 & 1 & 0 & \\ & \xrightarrow{} & & & 1 & \\ \hline & 0 & 1 & 1 & 1 & (7) \end{array}$$

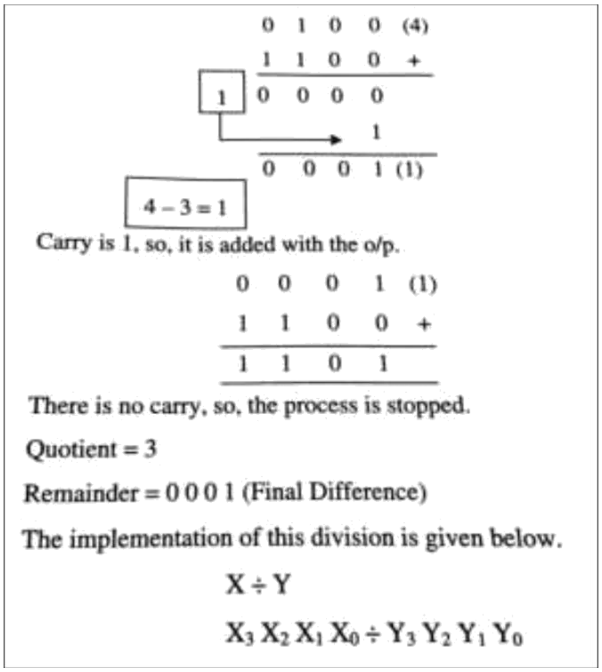
Carry is 1, so, it is added with the o/p.

$\boxed{10 - 3 = 7}$

$$\begin{array}{lccccc} & 0 & 1 & 1 & 1 & (7) \\ & 1 & 1 & 0 & 0 & + \\ \hline \boxed{1} & 0 & 0 & 1 & 1 & \\ & \xrightarrow{} & & & 1 & \\ \hline & 0 & 1 & 0 & 0 & (3) \end{array}$$

Carry is 1, so, it is added with the o/p.

$\boxed{7 - 3 = 4}$



$$\begin{array}{lccccc} & 0 & 1 & 0 & 0 & (4) \\ & 1 & 1 & 0 & 0 & + \\ \hline \boxed{1} & 0 & 0 & 0 & 0 & \\ & \xrightarrow{} & & & 1 & \\ \hline & 0 & 0 & 0 & 1 & (1) \end{array}$$

$\boxed{4 - 3 = 1}$

Carry is 1, so, it is added with the o/p.

$$\begin{array}{ccccc} 0 & 0 & 0 & 1 & (1) \\ 1 & 1 & 0 & 0 & + \\ \hline 1 & 1 & 0 & 1 & \\ \hline \end{array}$$

There is no carry, so, the process is stopped.

Quotient = 3

Remainder = 0 0 0 1 (Final Difference)

The implementation of this division is given below.

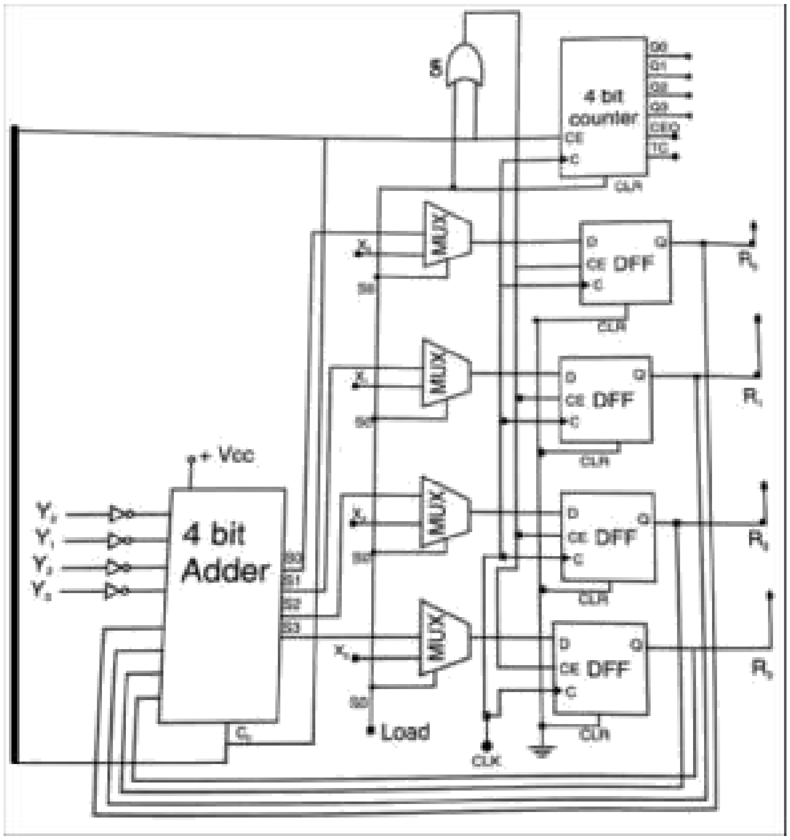
$X \div Y$

$X_3\, X_2\, X_1\, X_0 \div Y_3\, Y_2\, Y_1\, Y_0$

- Basic building blocks of serial adder are given below.
  1. 4 bit adder
  2. 4 bit binary up counter
  3. 2:1 MUX (4 MUXs are used)
  4. D flipflop

- Y_0 Y_1 Y_2 Y_3 are complemented and given to 4 bit adder block (figure shown below)
- X_0 X_1 X_2 X_3 are given to MUXs and MUX output is given to D flipflop. Select signal of MUX is high. It is connected to clear input of counter.
- Carry output of adder is connected with clock enable pin of counter. The same is given to OR gate. The output of this OR gate is given to clock enable signal of flipflops.
- The other input of OR gate is tied with select signal of MUX.

- If X > Y, C_0 of adder is high.
- After first subtraction, the counter output is incremented by 1.
- For each subtraction, the counter output is incremented.
- If C_0 of adder is low, then clock of counter and FF is disabled. Counting is stopped.
- $Q_3 Q_2 Q_1 Q_0$ is the counter output (Quotient)
- $R_3 R_2 R_1 R_0$ is the flipflop output (remainder)

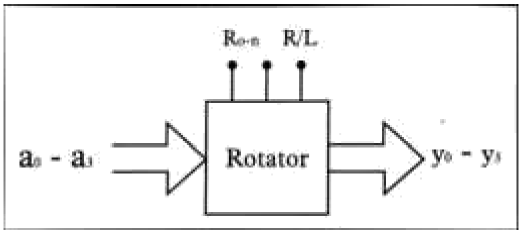


****

VIII. SHIFT REGISTERS:

Discuss about shifter and barrel shift adder.

- An n-bit rotation is specified by using the control word R_{0-n} and L/R bit defines a left or right shifting.



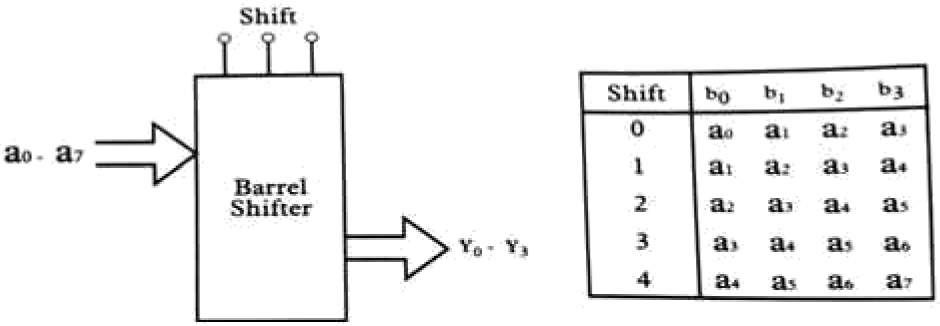
- For example $y_3\, y_2\, y_1\, y_0 = a_3\, a_2\, a_1\, a_0$

  If it is rotated 1-bit in left side, we get $y_3\, y_2\, y_1\, y_0 = a_2\, a_1\, a_0\, \mathbf{a_3}$

  If it is rotated 1-bit in right side, we get $y_3\, y_2\, y_1\, y_0 = \mathbf{a_0}\, a_3\, a_2\, a_1$

Barrel Shifter:

- A barrel shifter is a digital circuit that can shift a data word by a specified number of bits in one clock cycle.
- It can be implemented as a sequence of multiplexers (MUX), and in such an implementation the output of one MUX is connected to the input of the next MUX in a way that depends on the shift distance.
- For example, take a four-bit barrel shifter, with inputs A, B, C and D. The shifter can cycle the order of the bits ABCD as DABC, CDAB, or BCDA; in this case, no bits are lost.
- That is, it can shift all of the outputs up to three positions to the right (thus make any cyclic combination of A, B, C and D).
- The barrel shifter has a variety of applications, including being a useful component in microprocessors (alongside the ALU).



Shift	b_0	b_1	b_2	b_3
0	a_0	a_1	a_2	a_3
1	a_1	a_2	a_3	a_4
2	a_2	a_3	a_4	a_5
3	a_3	a_4	a_5	a_6
4	a_4	a_5	a_6	a_7

Figure: 8 X 4 barrel shifter

- General symbol for barrel shifter is shown in figure. The outputs are given as y_3 y_2 y_1 y_0. S_0, S_1, S_2, S_3 are known as shift lines.
- A barrel shifter is often implemented as a cascade of parallel 2×1 multiplexers.
- For a 8-bit barrel shifter, two intermediate signals are used which shifts by four and two bits, or passes the same data, based on the value of S[2] and S[1].
- This signal is then shifted by another multiplexer, which is controlled by S[0].
- A common usage of a barrel shifter is in the hardware implementation of **floating-point arithmetic**.

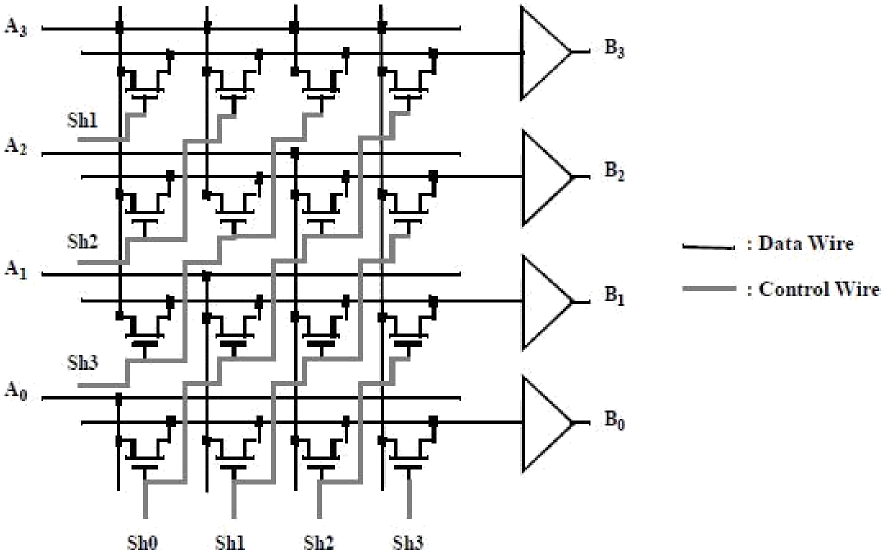
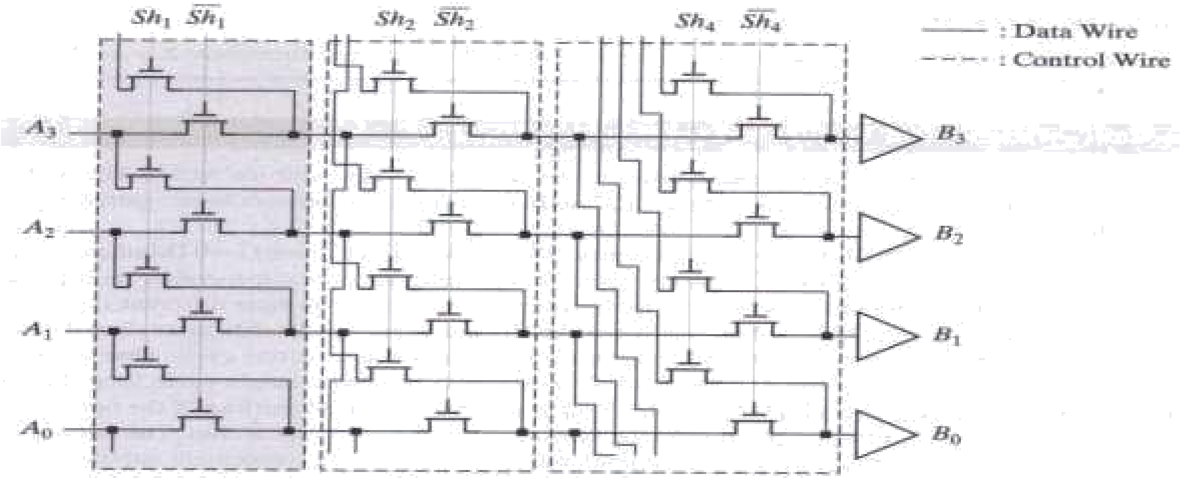


Figure: Barrel Shifter

- For a floating-point add or subtract operation, requires shifting the smaller number to the right, increasing its exponent, until it matches the exponent of the larger number.
- This is done by using the barrel shifter to shift the smaller number to the right by the difference, in one cycle.
- If a simple shifter were used, shifting by n bit positions would require n clock cycles.
- The disadvantages of FET array barrel shifter are the threshold voltage drop problem, parasitic limited switching time problem.
- The figure shown is known as a barrel shifter and a 8 x 4-bit barrel shifter circuit.

Logarithmic Shifter:

- A Shifter with a maximum shift width of M consists of a $\log_2 M$ stages, where the i^{th} stage either shifts over 2^i or passes the data unchanged.
- Maximum shift value of seven bits is shown in figure, to shift over five bits, the first stage is set to shift mode, the second to pass mode and the last again to shift.
- The speed of the logarithmic shifter depends on the shift width in a logarithmic wa, M-bit shifter requires $\log_2 M$ stages.
- The series connection of pass transistors slows the shifter down for larger shift values.
- Advantage of logarithmic shifter is more effective for larger shift values in terms of both area and speed.



****

IX. SPEED AND AREA TRADE OFF:

Discuss the details about speed and area trade off. (May 2017)

Adder:

- The tradeoff in terms of power and performance is shown below.
- The performance is represented in terms of the delay(speed).
- The area estimations for each of the delays are given based on the fact that area is in relation to the power consumption.
- The area of a carry lookahead adder is larger than the area of a ripple carry for a particular delay.

- This is because the computations performed in a carry lookahead adder are parallel, which requires a larger number of gates and also results in a larger area.

CLA –Carry Lookahead Adder, RC, R – Ripple carry adder

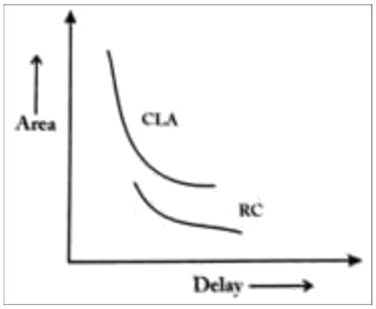


Figure: Area Vs Delay for 8 bit adder

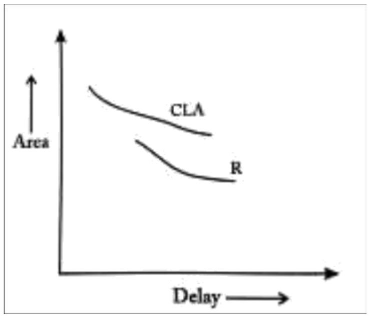


Figure: Area Vs Delay for 16 bit adder

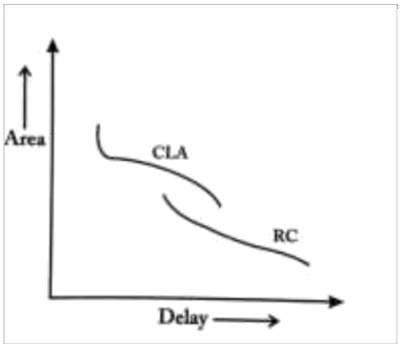


Figure: Area Vs delay for 32 bit adder

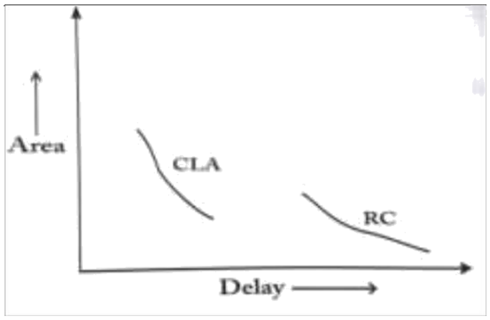


Figure: Area Vs delay for 64 bit adder

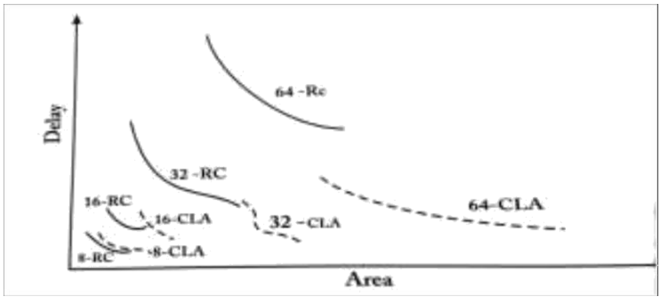


Figure: Delay Vs Area for all adders

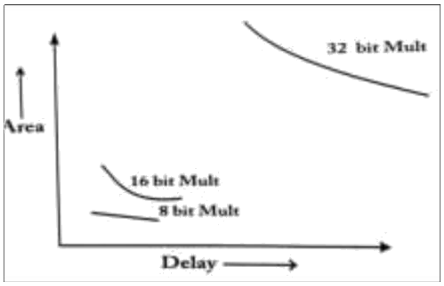
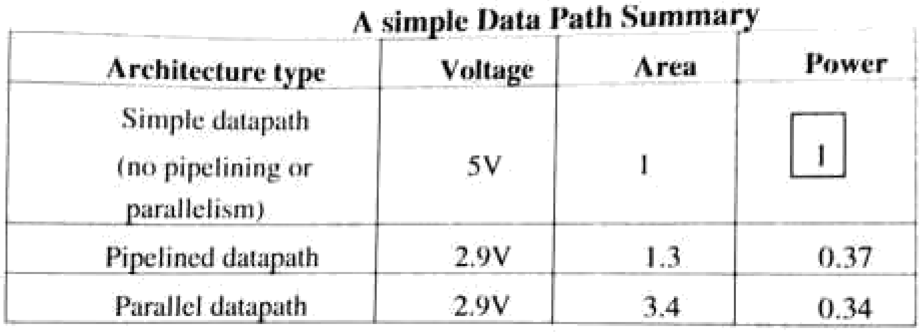


Figure: Area Vs Delay for all multiplier

- Above figures shows that the delay of the ripple carry adder increases much faster when compared to the carry lookahead adder as the number of bits is increased.
- In the carry lookahead adder, the cost is in terms of the area because computations are in parallel, and therefore more power is consumed for a specific delay.



A simple Data Path Summary

Architecture type	Voltage	Area	Power
Simple datapath (no pipelining or parallelism)	5V	1	1
Pipelined datapath	2.9V	1.3	0.37
Parallel datapath	2.9V	3.4	0.34

UNIT V - IMPLEMENTATION STRATEGIES

Full custom and Semi custom design, Standard cell design and cell libraries, FPGA building block architectures, FPGA interconnect routing procedures.

Introduction:

- **ASIC** - Application Specific Integrated Circuit is an Integrated Circuit (IC) designed to perform a specific function for a specific application.
- **Levels of integration:**
  The levels of integration are:
  - ✓ SSI - Small scale integration
  - ✓ MSI - Medium scale integration
  - ✓ LSI - Large scale integration
  - ✓ VLSI - Very large scale integration
  - ✓ USLI - Ultra large scale integration
- **Implementation technology**
- The implementation technologies used in ASIC are:
  - ✓ TTL – Transistor Transistor Logic
  - ✓ ECL – Emitter Coupled Logic
  - ✓ MOS – Metal Oxide Semiconductor (NMOS, CMOS)

5.1: Types of ASICs

- ❖ **Explain about different types of ASICs with neat diagram. (April 2016, April 2017)**
- ❖ **Write brief notes on: (a) Full custom ASIC (b) Semi custom ASIC (May 2010,May 2016)**
- ❖ **Compare the different types of ASICs. (Nov 2007, Nov 2008)**

- The ASICs are classified as follows:

I. Full-Custom ASICs II.
Semi-custom ASICs
  a. Standard-Cell–Based ASICs (CBIC)
  b. Gate-Array–Based ASICs (MPGA)
    i. Channeled Gate Array
    ii. Channelless Gate Array
    iii. Structured Gate Array
III. Programmable ASICs
  a. Complex Programmable Logic Devices (CPLD)
  b. Field-Programmable Gate Arrays (FPGA)

5.1.1:Full-Custom ASICs

Explain the full custom ASICs.

- In full custom ASIC, engineer can design full logic cells in IC. So, this technique is known as Full custom ASIC technique.

- Engineer uses mixed analog and digital technique to manufacture IC. All the logic cells are specifically designed for one ASIC.

- The characteristics of bipolar components in the same IC are matched very well.
- But the characteristic of components in different IC are not matched well

- This is widely used technology to manufacture IC.
- Mixing of analog and digital function are integrated in the same IC for which CMOS technology suits well.
- Designers give importance to performance.
- When large volume is manufactured, overall cost will be reduced.
- In super computer, quality is important so this design is implemented.
- All mask layers are customized in a full-custom ASIC
- Generally, the designer lays out all cells by hand
- Some automatic placement and routing may be done
- Critical (timing) paths are usually laid out completely by hand
- Full-custom design offers the ***highest performance and lowest part cost*** (smallest die size) for a given design.
- The ***disadvantages*** of full-custom design include ***increased design time, complexity, design expense***, and ***highest risk.***
- Microprocessors (strategic silicon) were exclusively full-custom, but designers are increasingly turning to semicustom ASIC techniques in this area as well.

5.1.2: Semi-custom ASICs – Design

❖ **Briefly explain the semi-custom Asics with its classification. (May 2016, NOV 2016)**

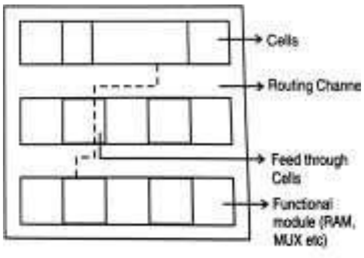
I. Standard cell based design:

- Standard cells are referred to AND gate, OR gate, multiplexer, flip flop, NOR gate etc.
- Standard cells can be used with larger predefined cells.
- This approach standardizes design entry level at logic gate.
- A design is generated automatically from HDL language.
- Then layout is created. In standard cell design, cells are placed in rows, and rows are separated by routing channel.
- All cells in library are in identical heights, widths of the cells can be varied to accommodate for variations in complexity between cells.
- A substantial fraction of area is allotted for signal routing.
- The minimization of interconnect overhead is most important goal of standard-cell placement routing tools. It is done by **feed through cells**.
- By using **feed through cell**, cells in different rows can be connected through vertical routing. So length of wire is reduced by feed through cells.

Semi-custom ASICs – CBIC

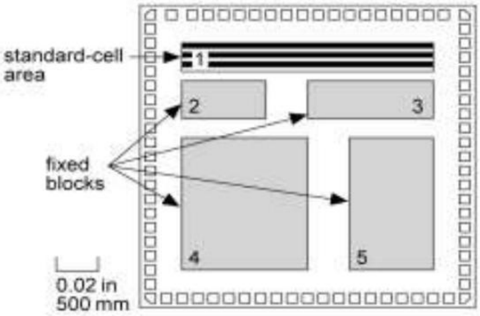
- CBIC means Cell Based ASICs.
- All the mask layers of CBIC are customized.

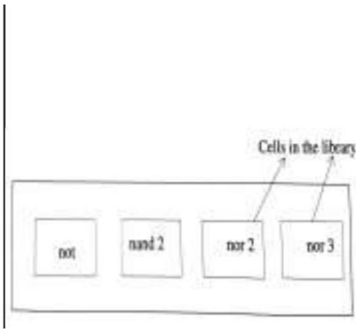
- It allows mega cells (SRAM, MPEG, decoder etc) to be placed in the same IC with standard cells (adder, gates etc).
- Mega cells are supplied by ASIC Company.
- Data path logic means the logic that operates on multiple signals across a data bus.
- Some of the ASIC library companies provide data path compiler which automatically generate data path logic.
- Data path library contains cells like adders, multiplexer, simple ALUs.
- ASIC Library Company provide data book which has functional description.



Features:

- It is a cell-based ASIC (CBIC —"sea-bick")
- It has Standard cells. Standard cell is logic elements used CMOS technology.
- Possibly megacells , megafunctions , full-custom blocks , system-level macros (SLMs), fixed blocks , cores , or Functional Standard Blocks (FSBs)
- All mask layers are customized - transistors and interconnect
- Automated buffer sizing, placement and routing. And custom blocks can be embedded.





- A "wall" of standard cells forms a flexible block.

II.Gate Array Based ASICs:

❖ **Explain gate array based ASICs with diagrams. (April 2008, May 2009)**

- Gate array is known as GA.
- In GA based ASIC, the transistors are predefined on the silicon wafer.
- *Base array*: the predefined pattern of transistors on a gate is known as base array.
- *Base cell:* the small element which is replicated to make the base array is known as base cell or primitive cell.
- *Masked Gate array*: Interconnect is defined by using top few layers of metal. •
This type of gate array is known as masked gate array.
- Gate array library is provided by ASIC Company.
- The designer can choose the predefined logic cells from a gate array library. These logic cells are known as **Macros**.
- Cell-layout is same for each logic cell. But interconnect is customized.
- It is also called as pre-diffused array because the transistors are diffused at first.

Types of MPGAs (Mask Programmable Gate Arrays):

- ✓ Channeled Gate Array
- ✓ Channel less Gate Array
- ✓ Structured Gate Array

(a) Channeled Gate Array:

- It is similar to CBIC (cell based ASIC).
- In the both types, rows of cells are separated by channels. These channels are used for interconnect.
- Space between rows of cells is fixed in a channeled gate array. But space between rows of cells may be adjusted in a CBIC.
  - ✓ Only interconnect is customized.
  - ✓ The interconnect uses predefined spaces between rows of base cells.
  - ✓ Manufacturing lead time is between two days and two weeks.

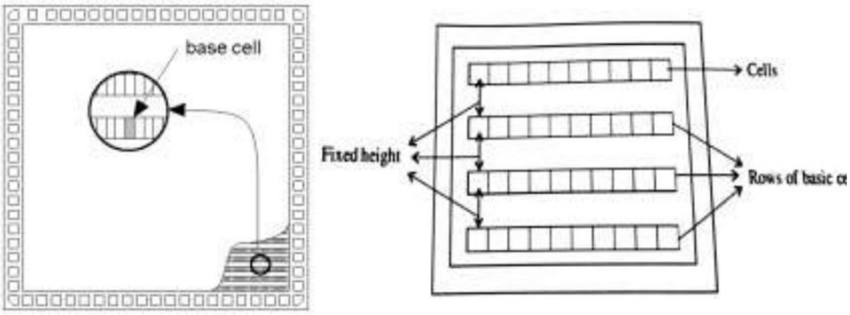


Figure: Channel Gate Array

(b) Channel less Gate Array:

- Channel less Gate Array is also called as channel free GA.
- In this array, there is no predefined space between rows for routing.
- Top few layers are used for defining interconnect connections.
- There are no predefined areas set aside for routing - routing is over the top of the gate-array devices.
- Achievable logic density is higher than for channeled gate arrays.
- Each logic cell or macro in a gate-array library is predesigned using fixed tiles of transistors known as the gate-array base cell (or just base cell).

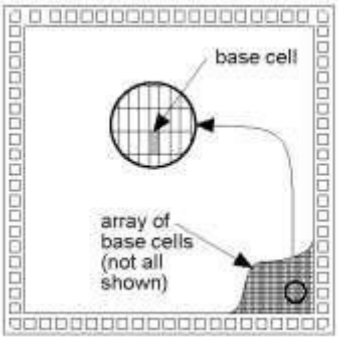


Figure: Channel less Gate Array

- Channeled and channelless gate arrays may use either gate isolation or oxide isolation.
- Isolate the transistors on a gate array from one another either with thick field oxide or by using other transistors that are wired permanently off.

(c) Structured Gate Array:

- Structured Gate Array is also called as embedded gate array or master slice or master image gate.
- It combines some of the features of CBIC and Masked gate array (MGA).
- In this array, some of the area is used for implementation of specially designed embedded block.
- Embedded area either can contain a different base cell that is more suitable for building memory cells, or a complete circuit block, such as a microcontroller.
    - Only the interconnect is customized
    - Custom block can be embedded
    - Manufacturing lead time is 2 days to 2 weeks
    - Area efficiency is increased
    - Performance is increased with low cost
- For ex: if embedded block has 32K bit memory. But the customer needs only 18K bit, the 16K memory is wasted.

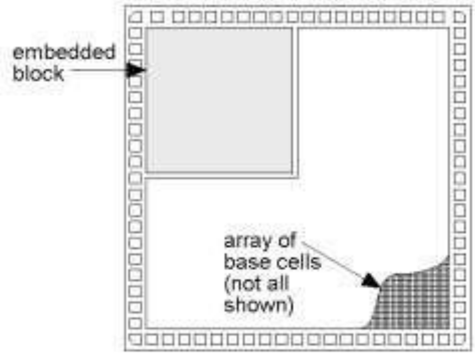


Figure: Structured Gate Array

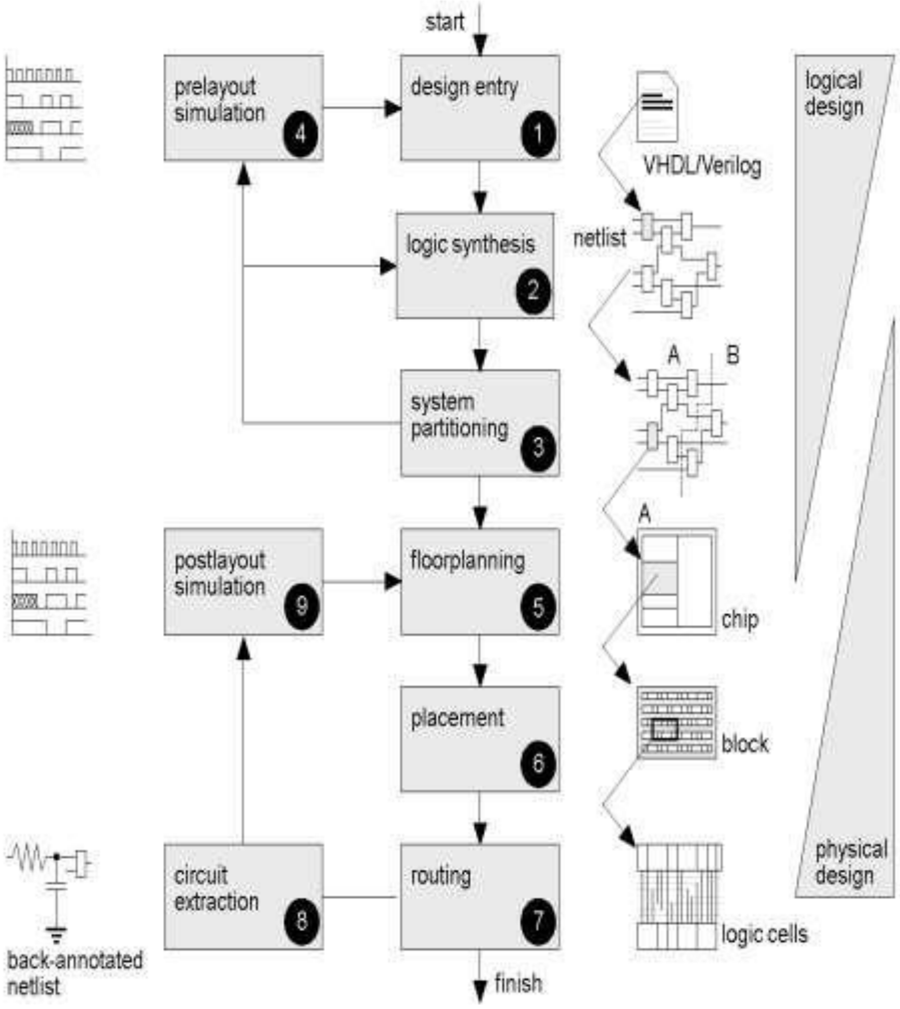
****

5.2: ASIC Design Flow / Cycle

Explain the ASIC design flow with a neat diagram. (Nov 2007, April 2008, Nov 2008)

1. **Design entry** - Using a hardware description language (HDL) or schematic entry
2. ***Logic synthesis*** - Produces a netlist - logic cells and their connections
3. ***System partitioning*** - Divide a large system into ASIC-sized pieces
4. **Prelayout simulation** - Check to see if the design functions correctly
5. ***Floorplanning*** - Arrange the blocks of the netlist on the chip
6. ***Placement*** - Decide the locations of cells in a block
7. ***Routing*** - Make the connections between cells and blocks
8. ***Extraction*** - Determine the resistance and capacitance of the interconnect
9. **Post layout simulation** - Check to see the design still works with the added loads of

the interconnect



****

5.3: ASIC Cell Libraries

Explain the standard Cell libraries in ASIC.

- The cell library is the key part of ASIC design.
- For a programmable ASIC the FPGA Company supplies a library of logic cells in the form of a design kit.
- For MGAs and CBICs you have three choices:
- ASIC vendor will supply a cell library, or you can buy a cell library from a third-party library vendor, or you can build your own cell library.
- The first choice, using an ASIC-vendor library, requires you to use a set of design tools approved by the ASIC vendor to enter and simulate your design.
- Some ASIC vendors (especially for MGAs) supply tools that they have developed in-house.
- An ASIC vendor library is normally a phantom library the cells are empty boxes, or phantoms, but contain enough information for layout.
- After you complete layout you hand off a netlist to the ASIC vendor,
- The second and third choices require making a buy-or-build decision.
- If complete an ASIC design using a cell library that you bought, you also own the masks that are used to manufacture your ASIC. This is called customer-owned tooling.

- A library vendor normally develops a cell library using information about a process supplied by an ASIC foundry.
- An ASIC foundry only provides manufacturing, with no design help. If the cell library meets the foundry specifications, we call this a qualified cell library.
- These cell libraries are normally expensive, but if a library is qualified at several foundries.
- The third choice is to develop a cell library in-house. Many large computer and electronics companies make this choice.
- However created, each cell in an ASIC cell library must contain the following:
  - ✓ A physical layout
  - ✓ A behavioral model
  - ✓ A Verilog/VHDL model
  - ✓ A detailed timing model
  - ✓ A test strategy
  - ✓ A circuit schematic
  - ✓ A cell icon
  - ✓ A wire-load model
  - ✓ A routing model

- The ASIC designer needs a high-level, behavioral model for each cell.
- Because simulation at the detailed timing level takes too long for a complete ASIC design.
- The designer may require Verilog and VHDL models in addition to the models for a particular logic simulator.

****

5.4: Library-Cell Design

Explain the important of Library –cell design in detail.

- Design rules for each ASIC vendor are slightly different even for the same generation of technology.
- For example, two companies may have very similar 0.35 nm CMOS process technologies, but the third-level metal spacing might be slightly different.
- A library constructed in this fashion may not be competitive with one that is constructed specifically for each process.
- ASIC vendors prize their design rules as secret, it turns out that they are similar except for a few details.
- We would like all vendors to agree on a common set of design rules.
- The reason that most vendors have similar rules is because most vendors use the same manufacturing equipment and a similar process.
- Layout of library cells is either hand-crafted or uses some form of symbolic layout.
- Symbolic layout is usually performed in one of two ways: using either interactive graphics or a text layout language.
- Shapes are represented by simple lines or rectangles, known as sticks, in symbolic layout.
- The actual dimensions of the sticks are determined after layout is completed in a Post processing step.
- Graphical symbolic layout uses a text layout language, similar to a programming language such as C that directs a program to assemble layout.

****

5.5: Programmable devices (Programmable ASIC):

- Programmable devices can be divided into three areas
  1. Programmable logic structure
  2. Programmable interconnect
  3. Reprogrammable Gate array
- A programmable logic device (PLD) is an electronic component used to build reconfigurable digital circuits.
- Unlike a logic gate, which has a fixed function, a PLD has an undefined function at the time of manufacture.

1. Programmable Logic Structure:

❖ **Describe in detail the chip with programmable logic structures. (Nov 2009)**

(a) Programmable Logic Array:

- Programmable logic arrays (PLAs) is a type of fixed architecture logic devices with *programmable AND gates followed by programmable OR array.*
- Logic array is the structure unit which can be programmed to perform various functions.
- Programmable Logic Array (PLA) can be implemented as AND-OR plane devices.
- Structure of AND-OR PLA is shown below.

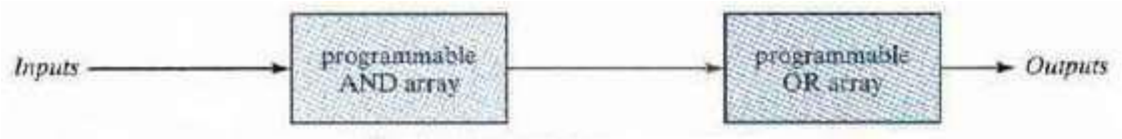


Figure: Programmable logic array

- PLA is used to implement a complex combinational circuit.
- The AND and OR gates inside the PLA are initially fabricated with fuses among them.
- The specific Boolean functions are implemented in sum of products (SOP) form by blowing appropriate fuses and leaving the desired connections.
- For an example, the Boolean expressions are,

$$F_1 = A\bar{B} + AC + \bar{A}B\bar{C}$$

$$F_2 = \overline{(AC + BC)}$$

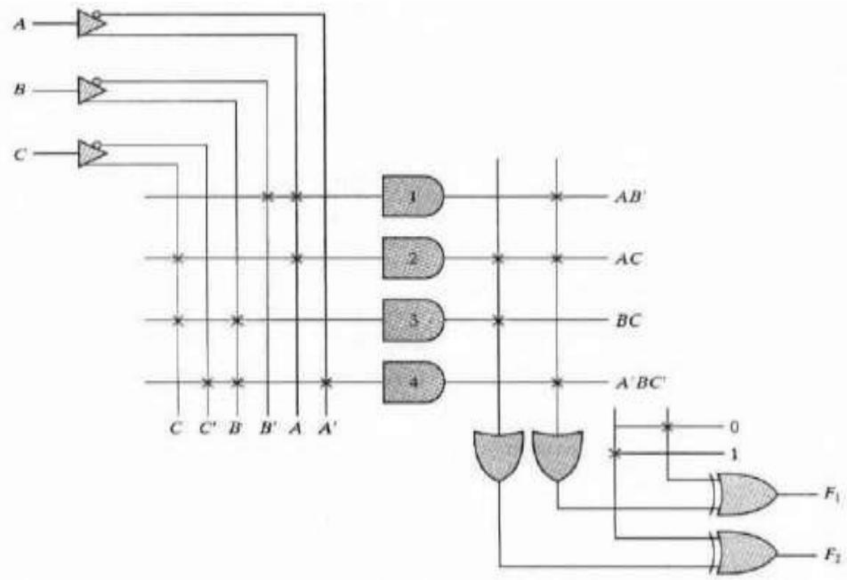


Figure: PLA with three inputs, four product terms and two outputs

(b) PAL (Programmable Array Logic) Architecture:

- The PAL is a programmable logic device with a *fixed OR array and a programmable AND* array.

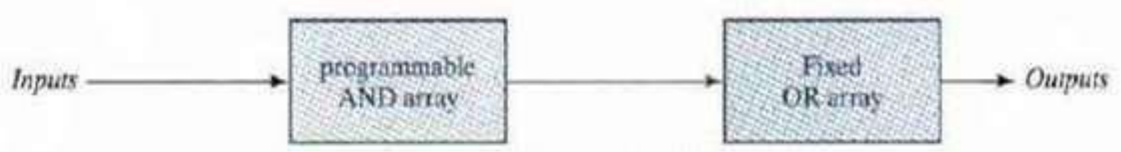


Figure: Programmable Array Logic

- Because only the AND gates are programmable, the PAL is easier to program than but is not as flexible as the PLA.
- The PAL is a programmable logic device with a fixed OR array and a programmable AND array.

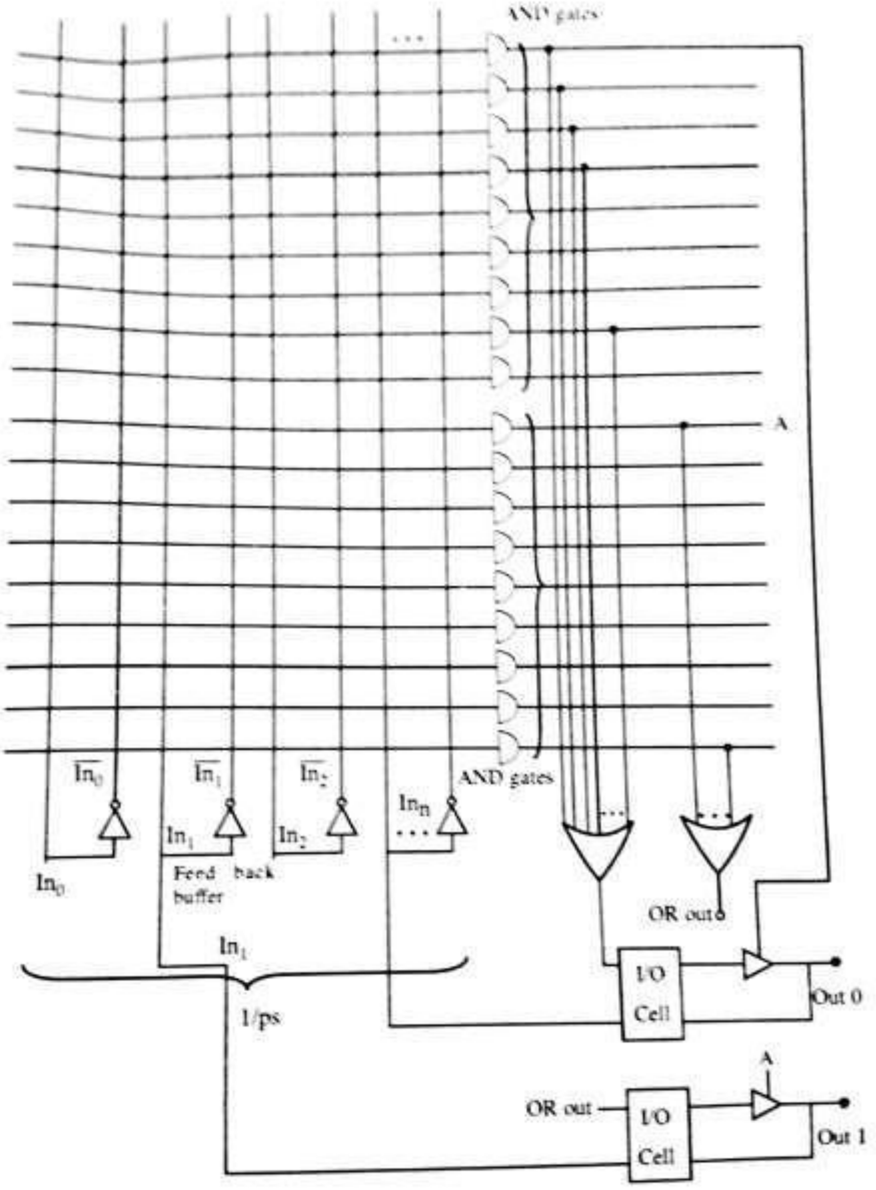


Figure: Example of PAL circuit

Reprogrammable Gate array:

- A field programmable gate array (FPGA) is a VLSI circuit that can be programmed at the user's location.
- A typical FPGA consists of an array of millions of logic blocks, surrounded by programmable input and output blocks and connected together via programmable interconnections.
- There is a wide variety of internal configurations within this group of devices.

- The performance of each type of device depends on the circuit contained in its logic blocks and the efficiency of its programmed interconnections.

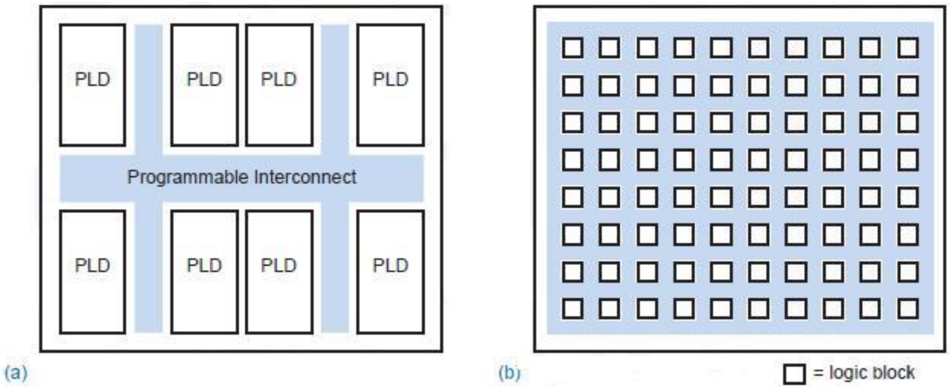


Figure: Programmable-logic-device approaches: (a) CPLD (b) FPGA.

- A typical FPGA logic block consists of lookup tables, multiplexers, gates, and flip-flops.
- A lookup table is a truth table stored in an SRAM and provides the combinational circuit functions for the logic block.
- The combinational logic section, along with a number of programmable multiplexers, is used to configure the input equations for the flip-flop and the output of the logic block.
- The *advantage* of using RAM instead of ROM to store the truth table is that the table can be programmed by writing into memory.
- The *disadvantage* is that the memory is volatile and presents the need for the lookup table's content to be reloaded in the event that power is disrupted.
- The program can be downloaded either from a host computer or from an onboard PROM.
- The program remains in SRAM until the FPGA is reprogrammed or the power is turned off. The device must be reprogrammed every time power is turned on.

****

5.6: Programming technology used in FPGA :

Discuss the different types of programming technology used in FPGA design. (NOV 2016)

- There are three types of programming technology.
  - ✓ Fusible link programming (Anti fuse)
  - ✓ SRAM Programming
  - ✓ EPROM and EEPROM programming

5.6.1: Fusible link programming:

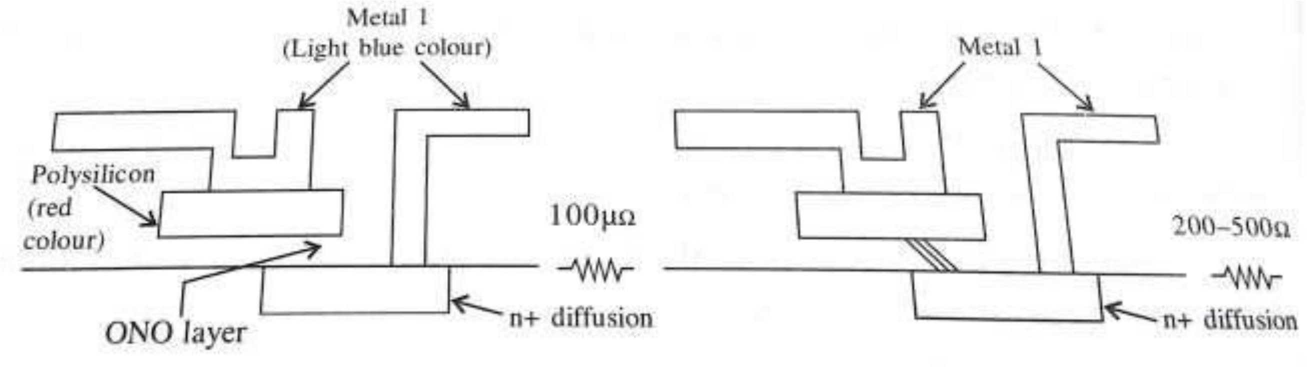
- In this type, Platinum, Titanium tungsten is used to form link.
- It is blown when certain current is exceeded in the fuse. Higher voltage is applied to device to blow the fuses.

Draw and explain the operation of metal-metal antifuse and EPROM transistor. (June 2012)

5.6.1.1: ANTIFUSE:

- In FPGA, the device is programmed by changing the characteristic of switching element (or) we can write the program for routing.

- Programming routing can be explained by using the product of ACTEL, Quick Logic Companies etc.
- In ACTEL, interconnect is done by PLICE (or) Antifuse.
- PLICE means Programmable Low Impedance Circuit Element.
- Antifuse is high resistance (>100MΩ) is changed into low resistance (200-500Ω) by applying programming voltage.
- It consists of ONO (Oxide-Nitride-Oxide) layer which is sandwiched between polysilicon layer and n+ diffusion.
- Antifuses separate interconnect wires on the FPGA chip and the programmer blows an antifuse to make a permanent connection.
- Once an antifuse is programmed, the process can't be reversed. This is an **OTP Technology.**



(a) Before applying programming voltage *(b) After applying programming voltag*

In-system programming (ISP):

- Possibility to program the chip after it has been assembled on the PCB.
- In Quick logic company, programmable interconnect is provided with Vialink (metal-metal anti-fuse).

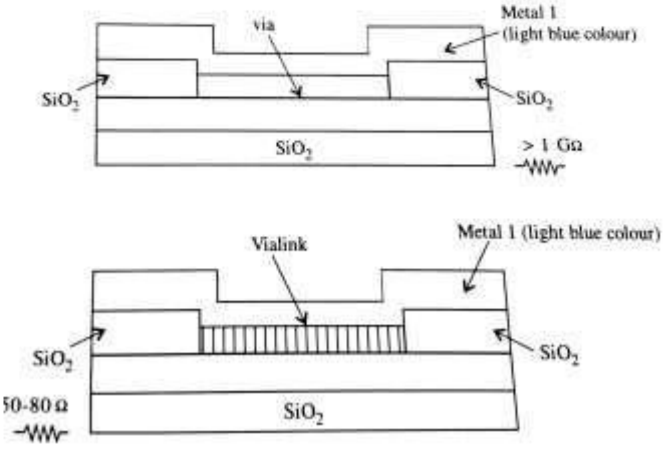


Figure: Metal-metal anti-fuse

Advantages of metal-metal antifuse:

- Advantages of metal-metal antifuse over poly diffusion antifuse are:
    1. The connections are direct to metal wiring layers.
    2. It is easier to use larger programming currents to reduce the antifuse resistance.

5.6.2: UV-Erasable programming:

Find the reason for referring EPROM technology as floating gate avalanche MOS. (Dec. 2013)

5.6.2.1: EPROM programming:

- In this type floating gate transistor is used.
- We can reprogram by using UV-light.
- High electric field causes electrons flowing towards drain to move across the insulating gate oxide, where they trapped on the bottom, floating gate.
- These energetic electrons are HOT and this effect is known as Hot-electron injection (or) avalanche injection.
- EPROM technology is sometimes called floating –gate avalanche MOS (FAMOS).

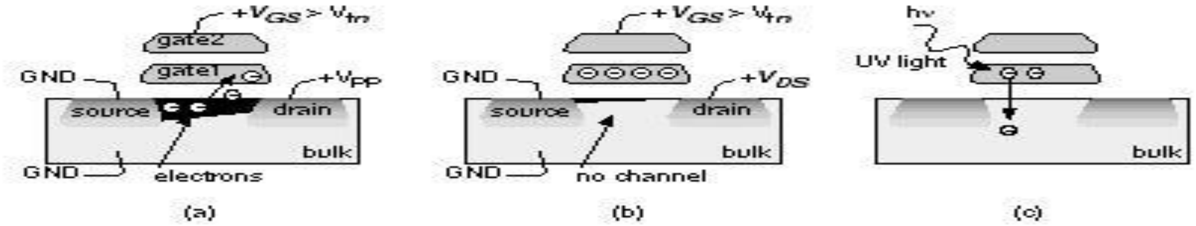


Figure: EPROM transistor

- (a) With a high (>12V) programming voltage, V_{PP} applied to the drain. Electrons gain enough energy to jump onto the floating gate (gate1).
- (b) Electrons stuck on gate1 raise the threshold voltage so that the transistor is always off for normal operating voltages.
- (c) Ultraviolet light provides enough energy for electrons stuck on gate1 to jump back to the bulk, allowing the transistor to operate normally.

5.6.2.2: EEPROM programming:

- Electrically Erasable programming is most popular CMOS technology.
- A very thin oxide between floating gate and the drain allow the electrons to tunnel to or from the floating gate (gate is charged or discharged).
- Thus enabling writing and erasing operation.

Advantages:

- The advantages of EEPROM technology are:
  - ✓ faster than using a UV lamp
  - ✓ chips do not have to be removed from the system
  - ✓ if the system contains circuits to generate both program and erase voltages, it may use ISP

5.6.3: SRAM Programming

- SRAM programming is shown in figure.
- SRAM configuration cell is constructed from two cross-coupled inverters and uses a standard CMOS process.
- The configuration cell drives the gates of other transistors on the chip (using pass transistors or transmission gates) to make a connection or off to break a connection.
- The cell is programmed using the WRITE and DATA lines.

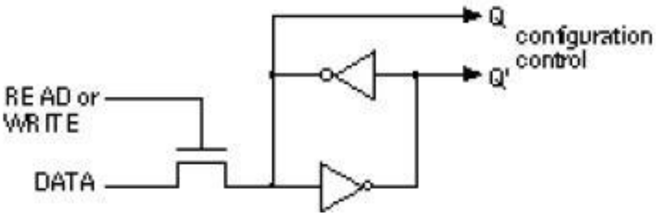


Figure: SRAM programming

Advantages:

- Designers can reuse chips during prototyping.
- Designers can update or change a system on the fly in reconfigurable hardware.

Disadvantage:

- Need to keep power supply for retaining the connection information.

****

- **Explain the reprogrammable device architecture with neat diagrams.**
- **With neat diagram explain the functional blocks in PDA (Programmable Device Architecture). (AU:June 2015, June 2016)**
- **With neat sketch explain the CLB, IOB and Programmable interconnects of an FPGA device. (May 2016)**
- **Explain about building block architecture of FPGA. (April 2017)**

5.7: Re-Programmable Devices Architecture

- FPGA provide the next generation in the programmable logic devices.
- It refers to the ability of the gate arrays to be programmed for a specific function by the user.
- The word Array is used to indicate a series of columns and rows of gates that can be programmed by the end user.
- As compared to standard gate arrays, the field programmable gate arrays are larger devices.
- The basic cell structure for FPGA is complicated than the basic cell structure of standard gate array.
- The programmable logic blocks of FPGA are called Configurable Logic Block (CLB).
- The FPGA architecture consists of **three types of configurable elements**-
    - (i) IOBs –Input/output blocks
    - (ii) CLBs- Configurable logic blocks
    - (iii) Resources for interconnection
- The IOBs provide a programmable interface between the internal, array of logic blocks (CLBs) and the device's external package pins.
- CLBs perform user-specified logic functions.
- The interconnect resources carry signals among the blocks.
- A configurable program stored in internal static memory cells.
- Configurable program determines the logic functions and the interconnections.
- The configurable data is loaded into the device during power-up reprogramming function.
- FPGA devices are customized by loading configuration data into internal memory cells.

The structure of FPGA:

The basic elements of the FPGA structure:

1. Logic blocks
    - Based on memories *(Flip-flop & LUT – Lookup Table)* Xilinx
    - Based on multiplexers *(Multiplexers)*-Actel
    - Based on PAL/PLA - Altera
    - Transistor Pairs
2. Interconnection Resources
    - Symmetrical FPGA-s
    - Row-based FPGA-s
    - *Sea-of-gates* type of FPGA-s
    - Hierarchical FPGA-s (*CPLD)*
3. Input-output cells (*I/O Cell)*
    - Possibilities for programming :
        a. Input
        b. Output
        c. Bidirectional

RE-PROGRAMMABLE DEVICE ARCHITECTURE:

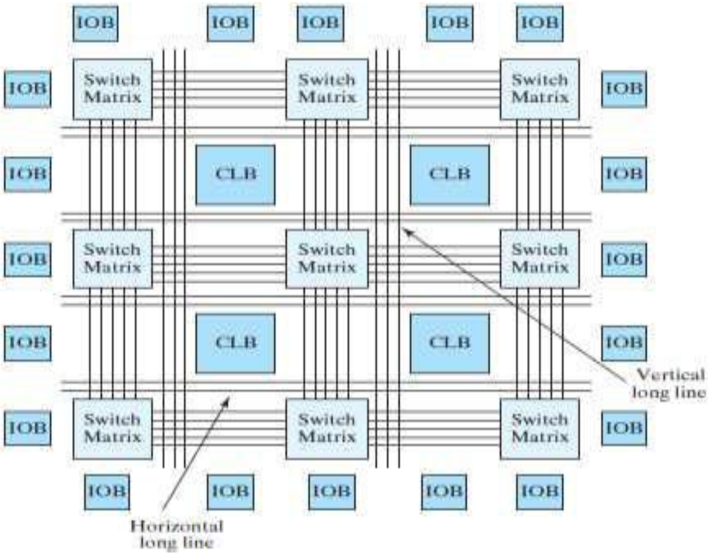
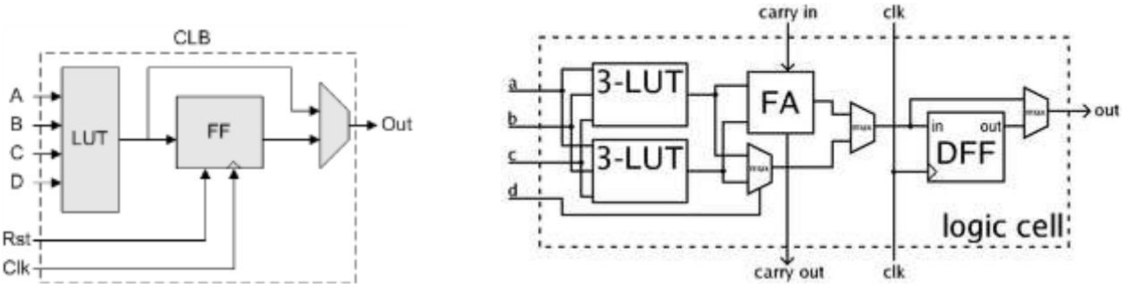


Figure: FPGA building blocks structure

- The figure shows the general structure of FPGA chip.
- It consists of a large number of programmable logic blocks surrounded by programmable I/O block.

5.7.1: Configurable Logic Block:



- **Figure: Various configurable Logic Block**

- The programmable logic blocks of FPGA are smaller and less capable than a PLD, but an FPGA chip contains a lot more logic blocks to make it more capable.
- As shown in figure the logic blocks are distributed across the entire chip.
- These logic blocks can be interconnected with programmable inter connections.
- The programmable logic blocks of FPGAs are called Configurable Logic Blocks (CLBs).
- CLBs contain LUT, FF, logic gates and Multiplexer to perform logic functions.
- The CLB contains RAM memory cells and can be programmed to realize any function of five variables or any two functions of four variables.
- The functions are stored in the truth table form, so the number of gates required to realize the functions is not important.

5.7.2: Interconnection resources:

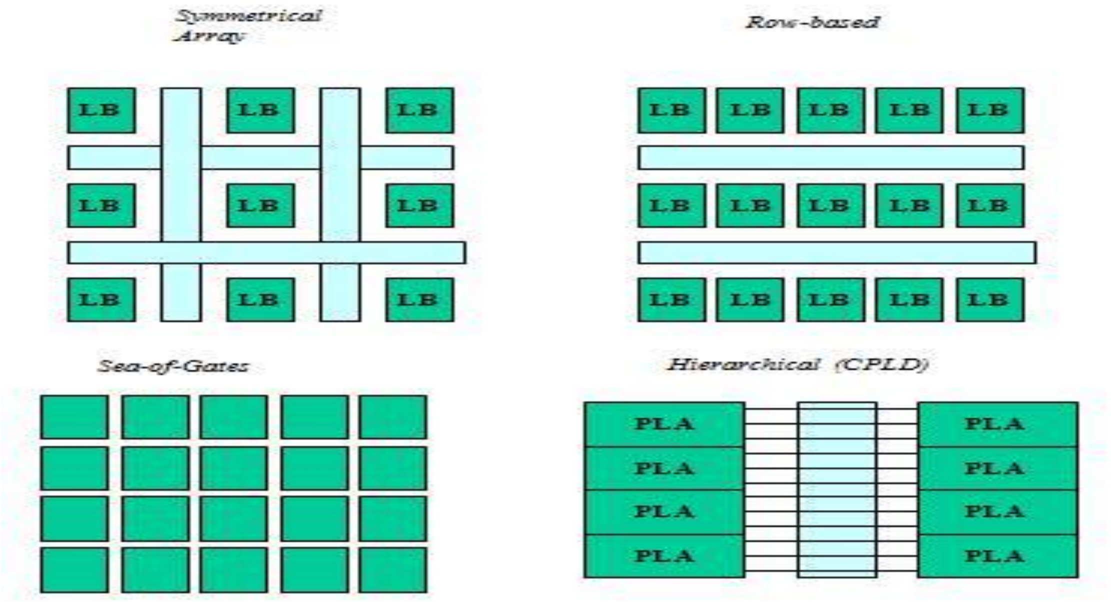


Figure: Types of interconnection resources

(a) Symmetrical Arrays

- It consists of logic elements (CLBs) arranged in rows and columns of a matrix and interconnect laid out between them.
- This symmrtrical martrix is surrounded by I/O blocks which connect it to outside world.

(b) Row based architecture:

- It consists of alternating rows of logic modules and programmable interconnect tracks.
- Input output blocks is located in the periphery of the rows.
- One row may be connected to adjacent rows via vertical interconnect.

(c) Hierarchical CPLD:

- This architecture is designed in hierarchical manner with top level containing only logic blocks and interconnects.
    1.Connections within macrocells
    2.Local connection resource within the logical block.
    3.Global connection resource *(Switch Matrix)*

(d) Sea of gates structure:

- It consists of logic elements (CLBs) arranged in rows and columns of a matrix inthe channel less gate arrays module.

5.7.3: I/O cells(Blocks):

- User-configurable input/output blocks (IOBs) provide the interface between external package pins and the internal logic.
- Each IOB controls one package pin and can be configured for input, output, or bidirectional signals.
- Figure shows a three-state bidirectional output buffer.
- When the output enable, OE is ‘1’ the output section is enabled and drives the I/O pad.
- When OE is ‘0’ the output buffer is placed in a high-impedance state.

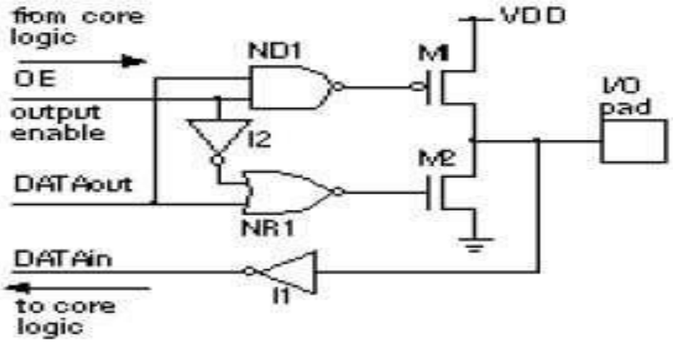


Figure: A three-state bidirectional output buffer

- We can limit the number of I/O drivers that can be attached to any one V_{DD} and GND pad.
- It allows employ the same pad for input and output bidirectional I/O.
- When we want to use the pad as an input, set OE low and take the data from DATAin.
- We can build output-only or input-only pads.

****

*5.8: FPGA(*PROGRAMMABLE ASIC *)interconnect routing procedures (Architectures):*

❖ **Give short notes on FPGA interconnect routing procedures. (May 2016)**

- Routing architecture comprises of programmable switches and many wires.
- Routing provides connection between logic blocks, I/O blocks, and between one logic block and another logic block.
- The type of routing architecture decides area consumed by routing as well as density of logic blocks.
- Routing techniques decide the amount of area used by wire segments and programmable switches as compared to area consumed by logic blocks.

Types of FPGA interconnect routing procedures:

- ✓ Hierarchical Routing Architecture
- ✓ Island-Style Routing Architecture
- ✓ Xilinx Routing Architecture
- ✓ Altera Routing Architecture
- ✓ Actel Routing Architecture

(a) Hierarchical Routing Architecture:

- Hierarchical routing architectures separates FPGA logic blocks into distinct groups.
- Connections between the logic blocks within a group can be made using wire segments at the lowest level of the routing hierarchy.

- Connections between the logic blocks in distant groups require the traversal of one or more levels of routing segments.
- As shown in Figure, only one level of routing directly connects to the logic blocks.
- Programmable connections are represented with the crosses and circles.

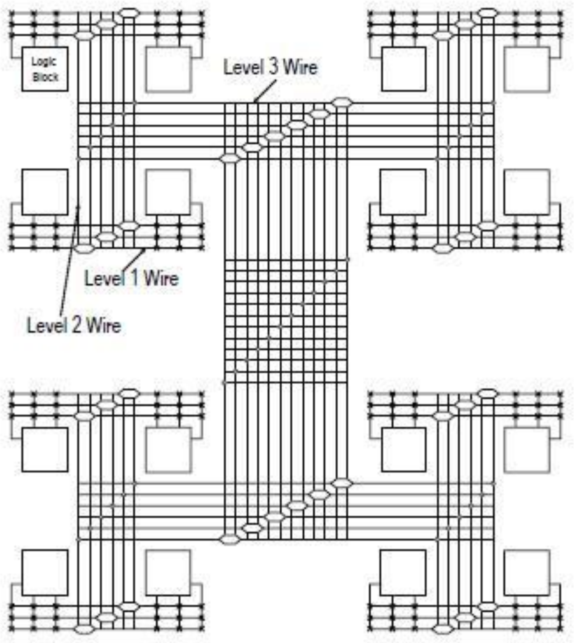


Figure: Example of Hierarchical

FPGA (b) Xilinx Routing Architecture:

- In Xilinx routing, connections are made from logic block into the channel through a connection block.
- As SRAM technology is used to implement Lookup Tables, connection sites are large.
- A logic block is surrounded by connection blocks on all four sides.

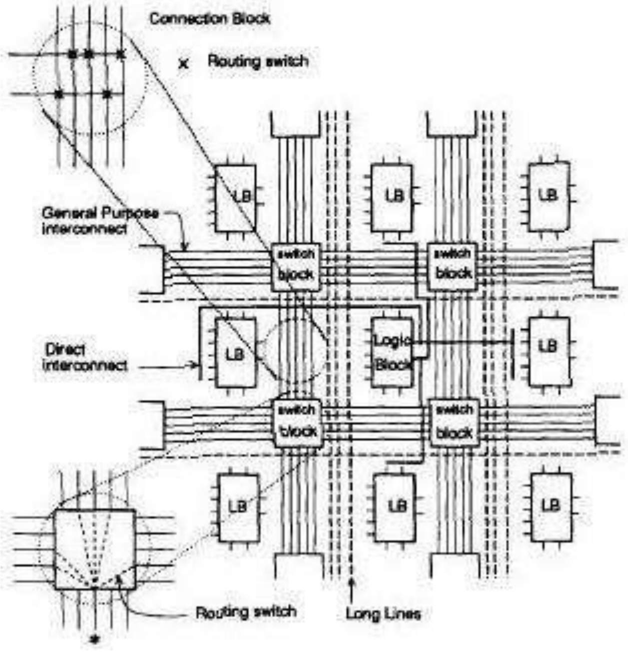


Figure: Xilinix Routing Architecture

- They connect logic block pins to wire segments.
- Pass transistors are used to implement connection for output pins, while use of multiplexers for input pins saves the number of SRAM cells required per pin.

- The logic block pins connecting to connection blocks can then be connected to any number of wire segments through switching blocks.
- Figure shows the Xilinx routing architecture.
- There are four types of wire segments available:
  - ✓ General purpose segments that pass through switches in the switch block.
  - ✓ Direct interconnect connects logic block pins to four surrounding connecting blocks
  - ✓ Long line: high fan out uniform delay connections
  - ✓ Clock lines: clock signal provider which runs all over the chip.

(c) Altera Routing Architecture :

- Altera routing architecture has two level hierarchies.
- At the first level of the hierarchy, 16 or 32 of the logic blocks are grouped into a Logic Array Block (LAB).
- The channel here is set of wires that run vertically along the length of the FPGA.
- Figure shows Alter Max 5000 routing architecture.
- Tracks are used for four types of connections:
  - ✓ Connections from output of all logic blocks in LAB.
  - ✓ Connection from logic expanders.
  - ✓ Connections from output of logic blocks in other LABs
  - ✓ Connections to and from Input output pads

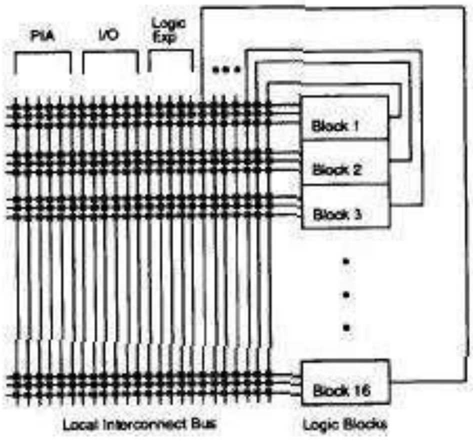


Figure: Altera Max 5000 Routing Architecture

- All four types of tracks connect to every logic block in the array block.
- Any track can connect to into any input which makes this routing simple.
- Advantage: It allows to be packed tightly and efficiently.
- Disadvantage: Large number of switches required, which adds to capacitive load.

(d) Island-Style Routing Architecture:

- As shown in Figure, island-style FPGAs logic blocks are arranged in a two dimensional mesh with the routing resources evenly distributed throughout the mesh.
- An island-style global routing architecture typically has the routing channels on all four sides of the logic blocks.
- The number of wires contained in the channel, W, is pre-set during fabrication, and is one of the key choices made by the architect.

- It employs wire segments of different lengths in each channel to provide the most appropriate length for each given connection.

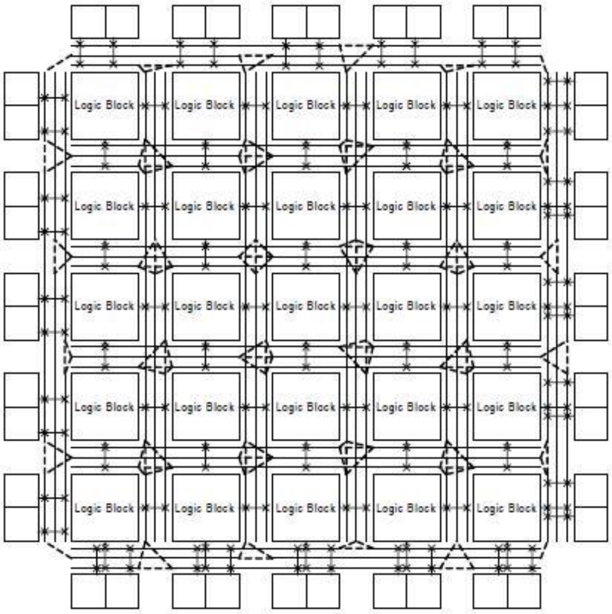


Figure: Island-Style Routing Architecture

(e) Actel Routing Architecture:

- Actel's design has more wire segments in horizontal direction than in vertical direction.
- The input pins connect to all tracks of the channel that is on the same side as the pin.
- The output pins extend across two channels above the logic block and two channels below it.
- Output pin can be connected to all 4 channels that it crosses.
- The switch blocks are distributed throughout the horizontal channels.
- All vertical tracks can make a connection with every incidental horizontal track.
- This allows for the flexibility that a horizontal track can switch into a vertical track, thus allowing for horizontal and vertical routing of same wire.
- The drawback is more switches are required which add up to more capacitive load.

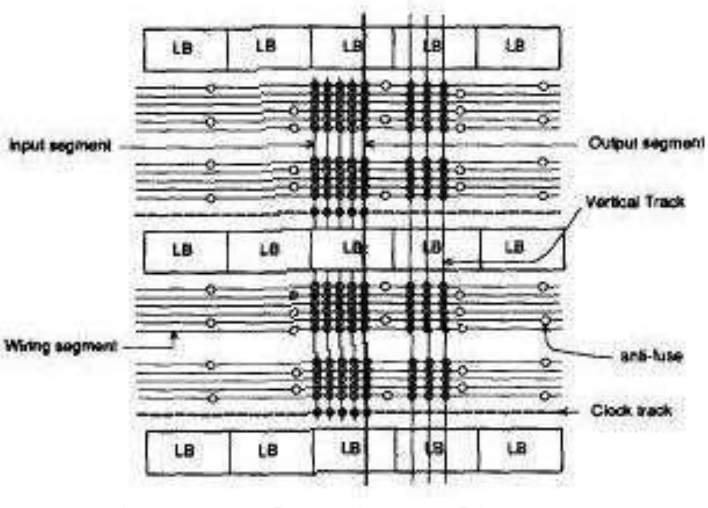


Figure: Actel Routing Architecture

****