UNIT - I

Introduction to MOS Transistor

MOS Transistor, CMOS logic, Inverter pass transistor, Transmission gate, layout design rules, Gate layouts, Stick diagrams, long-channel I-V characteristics, C-V characteristics, Non ideal I-V effects, DC Transfer characteristic, RC Delay Model, Elmore delay, Linear Delay Model, Logical effort, parasitic delay, delay in logic gate, scaling.

Introduction

Electronics is characterized by

1. Reliability
2. Low power dissipation
3. Extremely low weight and volume
4. Low cost.
5. Ability to cope easily with high degree of sophistication and complexity

* Integrated circuits has made possible the design of powerful and flexible processors which provide highly intelligent and adaptable devices for the user.

Four Generation of IC.

1. Small scale integration (SSI) → 10 transistors
2. Medium scale integration (MSI) → (100 - 1000) transistors
3. Large scale integration (LSI) → (1000 - 2000) transistors.
4. Very large scale integration (VLSI) → (20000 - 1,000,000) transistor.

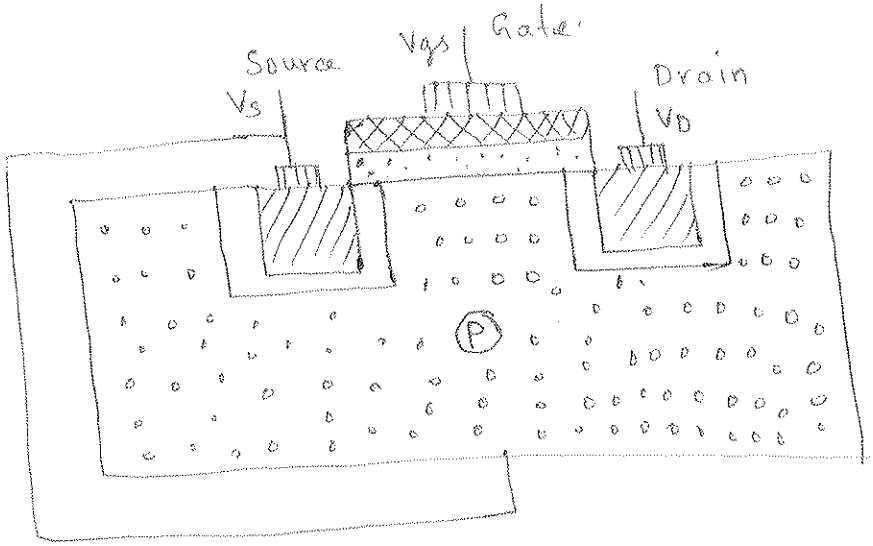
MOS TRANSISTOR

* n MOS enhancement mode transistor
* n MOS depletion mode transistor
* PMOS depletion and Enhancement mode.

n MOS :-

* n MOS transistors are formed in a p type material substrate of moderate doping level.
* Source and drain regions are formed by diffusing n type impurities.
* The source and drain are separated from each other by two diodes
* Connections to the source and drain are made by a deposited metal layer.
* To make a useful device there must be a capability for establishing and controlling the current between source and drain.

Enhancement mode :-



Metal

polysilicon

oxide

p substrate

n substrate

depletion

p diffusion.

* A polysilicon gate is deposited on a layer of insulation over the region between source and drain.

* In this mode, the channel is not established and device is in a non-conducting condition

$$V_D = V_S = V_{gs} = 0.$$

$V_D \rightarrow$ Drain Voltage

$V_S \rightarrow$ Source Voltage.

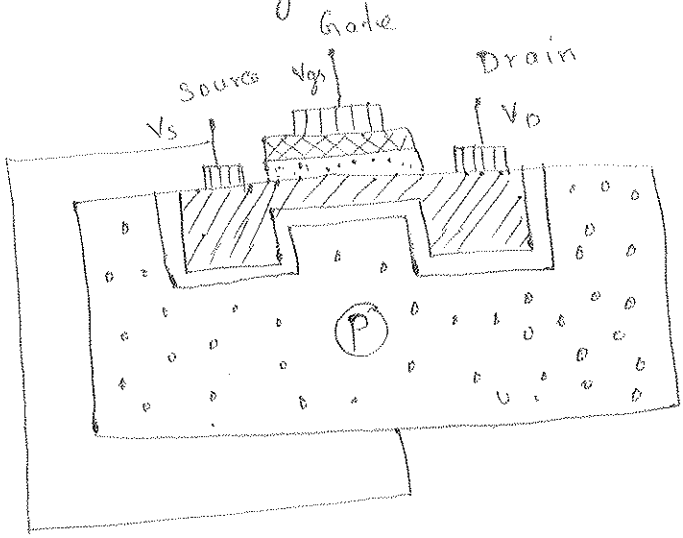
$V_{gs} \rightarrow$ Gate to substrate Voltage.

* If the gate is connected to the suitable positive voltage with respect to the source, then the electric field established between gate & substrate gives arise to a charge inversion region in the substrate

under the gate insulation & a conducting path or a channel is formed between source & drain.

Depletion Mode :-

* A channel may also be established so that it is present under the condition $V_{gs} = 0$ by implanting suitable impurities in the region between source and drain during manufacture and prior to depositing the insulation and gate.



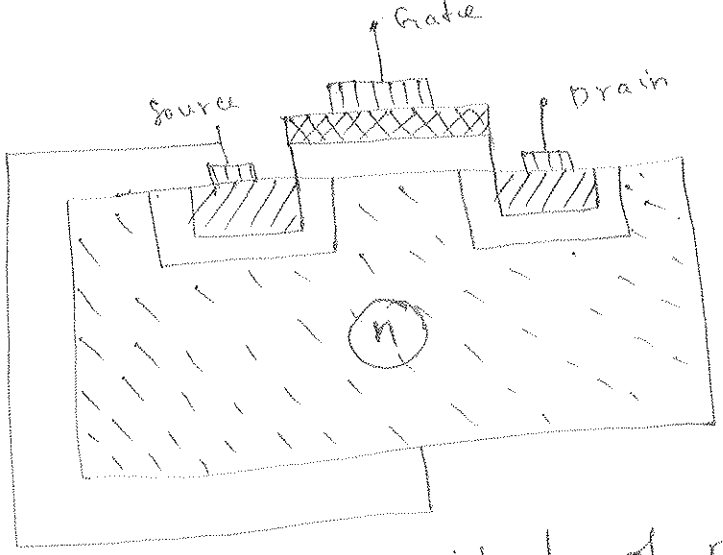
* Under these circumstances, source and drain are connected by a conducting channel, but the channel may now be closed by applying the suitable negative voltage to the gate.

PMOS :-

* PMOS transistors are formed in a n type material
* Source and drain regions are formed by diffusing p type impurities

Enhancement mode :-

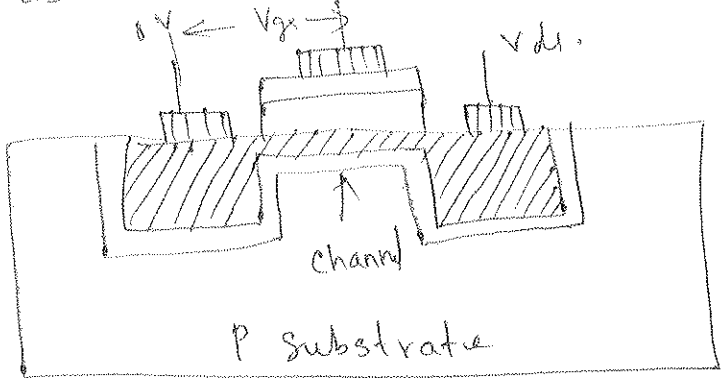
* Under unbiased condition, channel is not established between source and drain.



* When a suitable magnitude of negative voltage is applied between gate and source, a channel (p-type) is formed between source and drain, current may flow, if the drain is made negative with respect to the source.
* Current is carried by holes.

Enhancement mode Transistor action

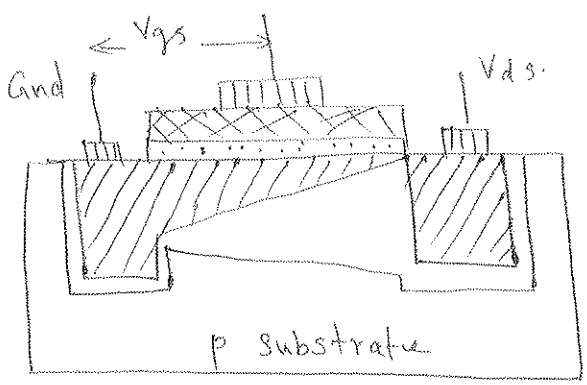
① $V_{ds} = 0$, $V_{gs} > V_t$



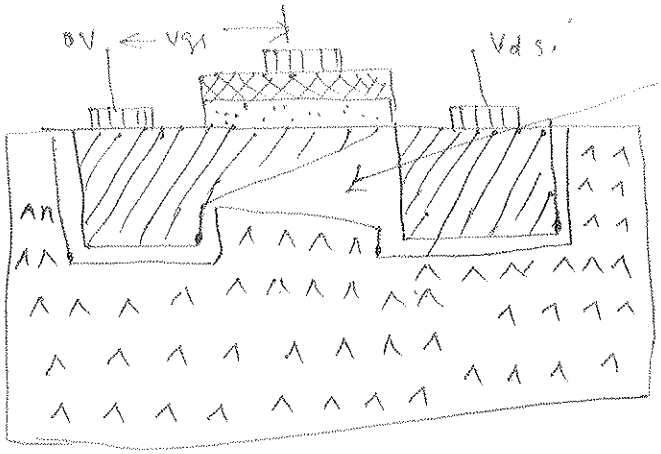
* To establish a channel in first place a minimum voltage level of threshold voltage V_t must be established between gate and source.
* A channel is established but no current is flowing through it $V_{ds} = 0$.
* Now consider a voltage V_{ds} is applied between the drain and source. Current is flowing through the channel. There must be corresponding IR drop equal to V_{ds} along the channel.
* This results in the voltage between gate and channel varying with distance along the channel with the voltage being maximum of V_{gs} at the source end.
* Effective voltage at gate is $V_g = V_{gs} - V_t$. there will be voltage available to invert the channel at the drain end

$$V_{gs} - V_t \geq V_{ds}.$$

II, $V_{ds} < V_{gs} - V_t$, the device is in non-saturated region



(ii) $V_{ds} > V_{gs} - V_t$



* In this case IR drop = $V_{gs} - V_t$ takes place over less than the whole length of the channel so that near the drain there is insufficient electric field available to give rise to an inversion layer to create a channel. The channel is therefore 'pinched off'

* Diffusion current completes the path from source to drain causing the channel to exhibit a high resistance and behave as a constant current source. This region is known as saturated region.

* For enhancement mode devices

$$V_E = 0.2\ V_{DD}.$$

$\therefore V_{DD} = 5V$

Depletion mode transistor action

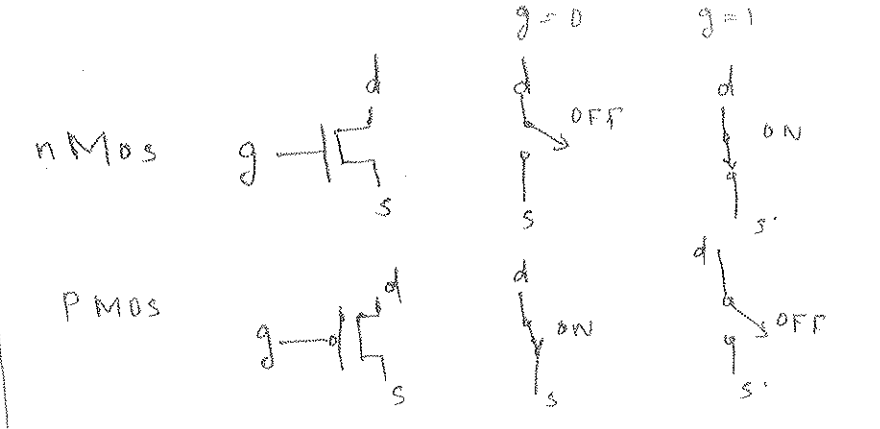
* Similar to enhancement mode of operation, but threshold voltage

$$V_t < -0.8\ V_{DD}$$

Complementary metal oxide semiconductor (CMOS) Logic

The Inverter

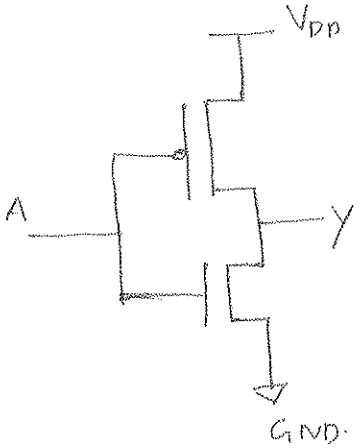
Below fig shows a CMOS inverter or NOT gate using one nMOS transistor and one pMOS transistor.



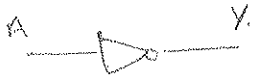
Transistor symbols and switch level models

The horizontal bar at the top indicates V_{DD} and the triangle at the bottom indicates GND. When the input A is '0' the nMOS transistor is OFF and the pMOS transistor is ON. Thus the output y is pulled up to '1' because it is connected to V_{DD} but no to GND.

///ly When A is '1' the nMOS is ON the pMOS is OFF and the y is pulled down to '0'.



Symbol



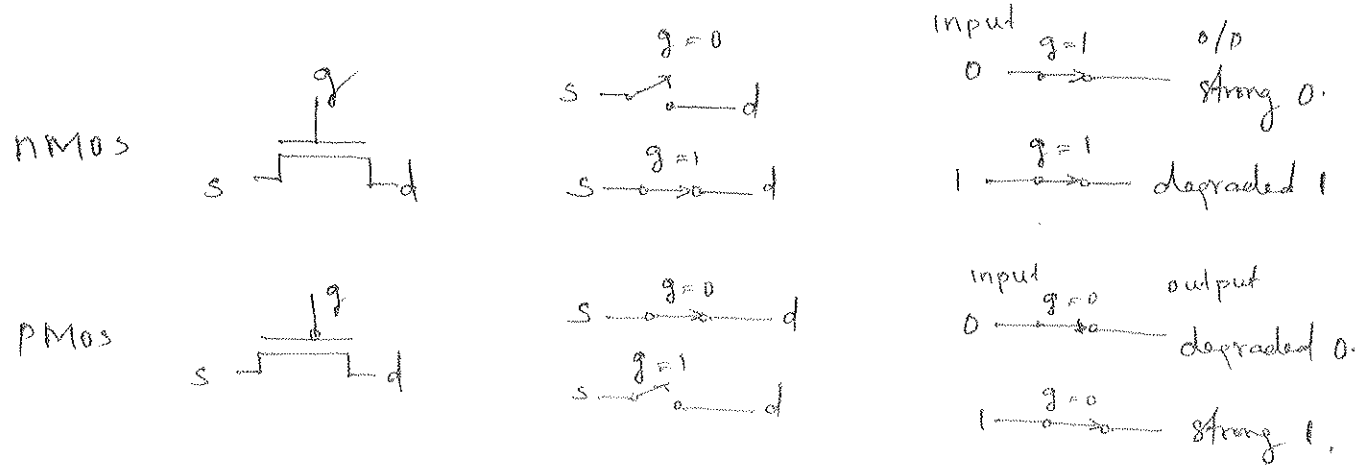
Truth table.

A	Y
0	1
1	0

Pass Transistors and Transmission gates:

The strength of a signal is measured by how closely it approximates an ideal voltage source. In general the stronger a signal, the more current it can source or sink.

* The power supplies or rails (V_{DD} and GND) are the source of the strongest '1's and '0's.
* An nMos transistor is an almost perfect switch when passing a '0' and thus we say it passes a strong '0'. However, the nMos transistor is imperfect at passing a '1'.
* A pMos transistor again has the opposite behavior passing strong '1's but degraded '0's.



When an nMos or pMos is used alone as an imperfect switch, we call it a pass transistor. By combining an nMos and a pMos transistor in parallel as shown in fig (a) we obtain a switch that turns on when a '1' is applied to g.

Fig (b) in which 0's and 1's are both passed in an acceptable fashion. Fig (c) we term this a transmission gate or pass gate.

In a circuit where only a '0' or a '1' has to be passed the appropriate transistor (n or p) can be deleted, reverting to a single nMOS or pMOS device. Note that both the control input and its complement are required by the transmission gate. This is called double rail logic.

Symbols for the transmission gate is shown in fig (d)

Layout design rules :-

* Rules followed to prepare the photomask is known as layout design rules.

① Representation of layers :

* By using layers, specification of IC is converted to mask.

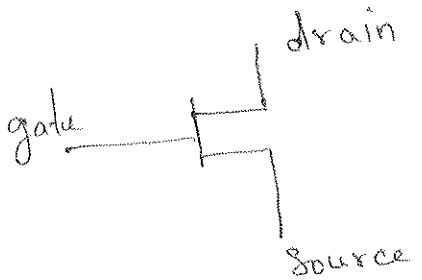
Name of the layer	Colour of the layer
n well	Brown
Thinoxide (n transistor)	Green
poly (poly silicon)	Red
p+ (p transistor)	yellow
Metal 1	Light blue

Metal 2	Dark blue
Metal 3	Gray
Via	Black
Contact-cut	Black.

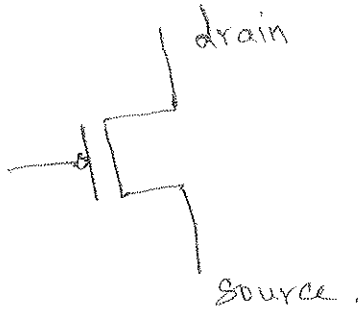
② stick diagram

* The diagram which conveys the layer information throught the use of a color is known as stick diagram.

n Mos Representation



p Mos Representation



* n Mos and p Mos are separated by demarcation line (.....) All p Mos placed above demarcation line. All n mos placed below the demarcation line.
* Diffusion paths should not cross demarcation line.
* n diffusion and p diffusion wires should not be joined n and p features joined by a metal.

* X is the symbol used to represent V_{DD} and V_{SS}. All n-MOS is close to V_{SS}. V_{DD} and V_{SS} lines are known as Rails.

Instructions to draw the stick diagram:

① Draw the V_{DD}, V_{SS} rails using blue colour

——— V_{DD} Rail
——— V_{SS} Rail.

② Draw the demarcation line in between V_{DD} and V_{SS} rail

——— V_{DD} Rail
- - - - - Demarcation line
——— V_{SS} Rail.

③ Draw n-diffusion, p-diffusion lines (green and yellow).

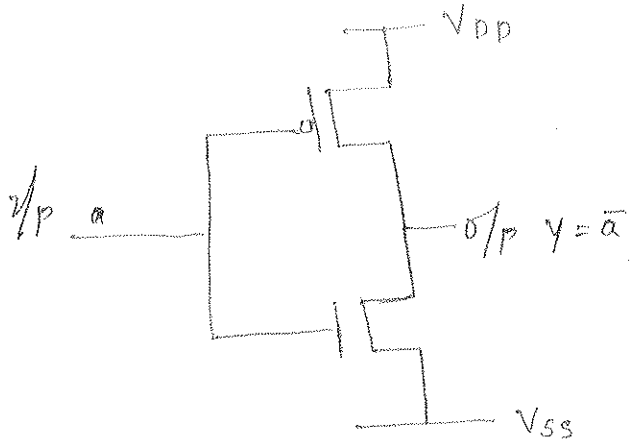
④ Draw i/p and o/p lines using red colour (poly).

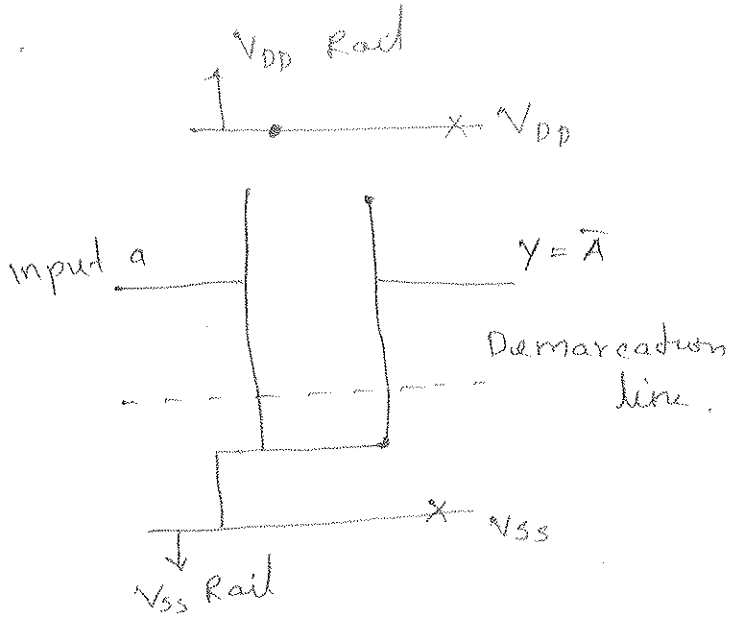
⑤ Draw the metal contact in blue colour.

⑥ Draw the contact in black colour and draw V_{DD} and V_{SS} using x symbol in black.

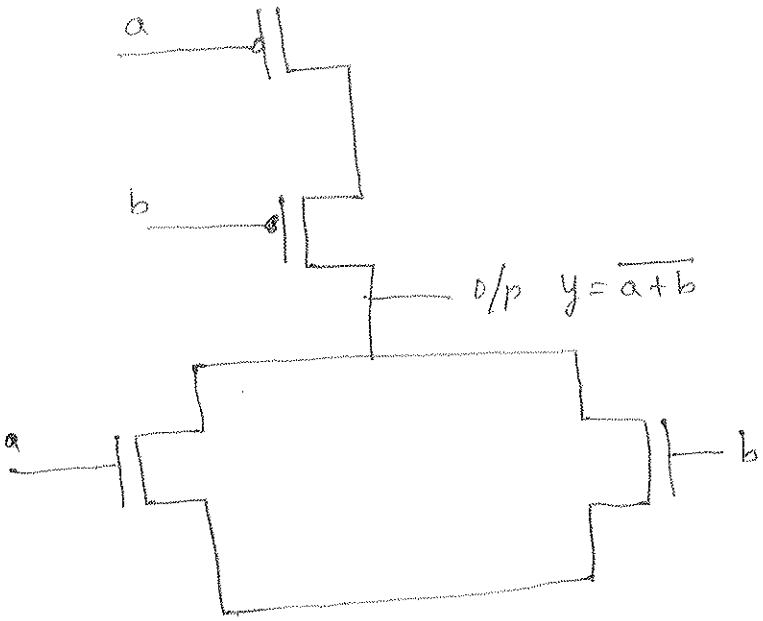
Colour	Layer Name
———	polysilicon
———	n diffusion
———	Metal 1
•	Contact cut via
	p diffusion
———	Metal 2
- - - - -	Demarcation line
x	V_{DD}, V_{SS} Contact

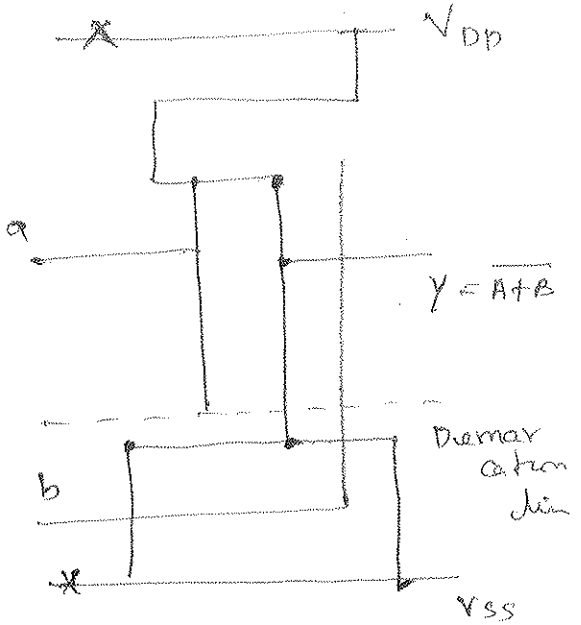
NOT Gate (inverter)





NOR Gate :-





Ideal I-V characteristics

* Current - voltage characteristics is known as I-V char's.

There are 3 operating regions in MOS

① Cut-off region or Sub threshold region

② Linear or Non saduration region

③ Saturadion region.

① Cut-off region :-

* In this region $V_{gs} < V_t$. There is no channel and almost zero current flows from drain to source ($I_{ds} = 0$)

② Non saturation region :-

* In this region, gate attracts the carriers to form a channel.

* Assume that the voltage along the channel varies linearly with distance x.

* When the device is not saturated the average value of IR drop is taken as $\frac{V_{ds}}{2}$.

* Current I_{DS} depends on both V_{GS} and V_{DS}.

$$I_{ds} = -I_{sd} = \frac{\text{change induced in channel } Q_c}{\text{Electron transit time } (\tau)}$$

* Transit time

$$\tau = \frac{\text{Length of the channel}}{\text{Velocity}}$$

$$\tau = \frac{1}{v} = \frac{L}{\mu E_{ds}}$$

$\mu \Rightarrow e^-$ or hole mobility

$E_{ds} \rightarrow$ drain to source electric field

$$\boxed{\tau = \frac{L}{\mu E_{ds}}}$$

* Electric field $E_{ds} = \frac{V_{ds}}{L}$

$$I_{ds} = \frac{Q_c}{L/\mu E_{ds}} = \frac{Q_c}{L^2/\mu V_{ds}}$$

$$\boxed{I_{ds} = \frac{\mu V_{ds} Q_c}{L^2}} \quad \text{——— (1)}$$

In linear region,

* Effective gate voltage $\boxed{V_g = V_{gs} - V_t}$

* charge per unit area $= E_g \varepsilon_{ins} \varepsilon_0$

induced charge $Q_c = E_g \varepsilon_{ins} \varepsilon_0 WL$ ——— (2)

E_g → avg $\bar{e}$ field gate to channel

ε_{ins} → relative permitivity of insulation between gate and channel ≃ 4, for SiO_2

ε_0 → permitivity of free space = 8.85×10^{-14} $F cm^{-1}$

$$I_{ds} = \frac{\mu V_{ds} Q_c}{L^2}$$

To find Q_c :-

$$E_g = \frac{\text{Effective voltage} - \text{Average voltage drop}}{\text{Width of the oxide [oxide thickness]}}$$

$$= \frac{V_g - \frac{V_{ds}}{2}}{D}$$

$$E_g = \frac{(V_{gs} - V_t) - \frac{V_{ds}}{2}}{D}$$

$$Q_c = \frac{\varepsilon_{ins} \varepsilon_0 WL}{D} \left[(V_{gs} - V_t) - \frac{V_{ds}}{2} \right]$$

$$I_{ds} = \frac{\varepsilon_{ins} \varepsilon_0 W\cancel{L}}{D} \left[(V_{gs} - V_t) - \frac{V_{ds}}{2} \right] \cdot \frac{\mu V_{ds}}{\cancel{L^2}}$$

$$\boxed{I_{ds} = K \frac{W}{L} V_{ds} \left[(V_{gs} - V_t) - \frac{V_{ds}}{2} \right]} \quad (3)$$

$\therefore K = \frac{\varepsilon_{in} \varepsilon_0 \mu}{D}$

$\beta = \frac{KW}{L}$

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

$$\boxed{I_{ds} = \beta \left[V_{ds} (V_{gs} - V_t) - \frac{V_{ds}^2}{2} \right]} \quad (4)$$

* The gate channel capacitance

$$C_g = \frac{\varepsilon_{in}\varepsilon_0 WL}{D}, \quad K = \frac{C_g \mu}{WL}$$

Sub. C_g & K value in equ (3)

* $$I_{ds} = \frac{C_g \mu}{WL} \cdot \frac{W}{L} V_{ds}\left((V_{gs}-V_t) - \frac{V_{ds}}{2}\right)$$

$$\boxed{I_{ds} = \frac{C_g \mu}{L^2}\left[V_{ds}(V_{gs}-V_t) - \frac{V_{ds}^2}{2}\right]} \quad \text{——— (5)}$$

* By means of gate capacitance per unit area C_o

$$C_g = C_o WL \implies K = \frac{C_o WL\mu}{WL}$$

$$K = C_o \mu$$

Sub K value in equ (3)

$$\boxed{I_{ds} = C_o \mu \frac{W}{L}\left[(V_{gs}-V_t)V_{ds} - \frac{V_{ds}^2}{2}\right]} \quad \text{——— (6)}$$

(3) Saturation Region

* Saturation begins when

$$\boxed{V_{ds} = V_{gs} - V_t} \quad \text{——— (7)}$$

Sub (7) in (4)

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

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

$$\boxed{I_{ds} = \beta\left[\frac{(V_{gs}-V_t)^2}{2}\right]} \quad \text{——— (8)}$$

From (5) $I_{ds} = \frac{C_g \mu}{L^2}\left[(V_{gs}-V_t)(V_{gs}-V_t) - \frac{(V_{gs}-V_t)^2}{2}\right]$

$$\boxed{I_{ds} = \frac{C_g \mu}{L^2}\left[\frac{(V_{gs}-V_t)^2}{2}\right]} \quad \text{——— (9)}$$

Sub eq (7) in (6)

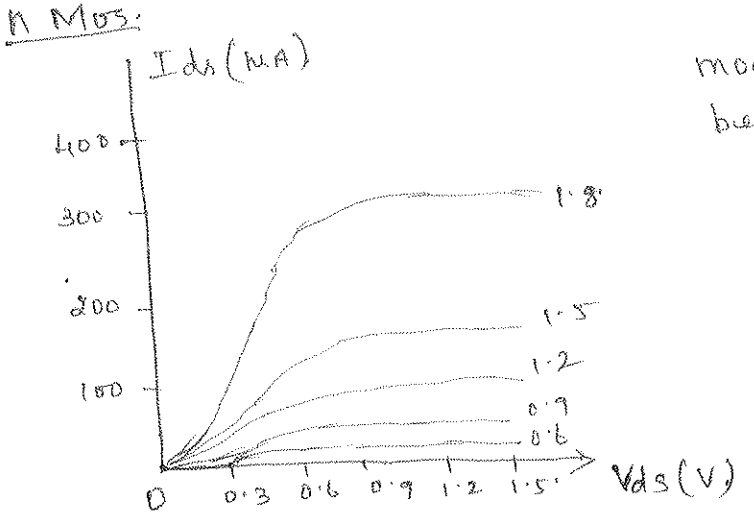
$$I_{ds} = C_0 \mu \frac{W}{L}\left[(V_{gs}-V_t)^2 - \frac{(V_{gs}-V_t)^2}{2}\right]$$

$$\boxed{I_{ds} = \frac{C_0 \mu W}{L}\left[\frac{(V_{gs}-V_t)^2}{2}\right]} \quad \text{——— (10)}$$

$$I_{ds} = \begin{cases} 0 & , \quad V_{gs} < V_t \quad \text{cut off.} \\ \beta\left[(V_{gs}-V_t) - \frac{V_{ds}}{2}\right]V_{ds} & , \quad V_{ds} < V_{gs}-V_t \quad \text{Non saturation} \\ \frac{\beta}{2}\left[V_{gs}-V_t\right]^2 & , \quad V_{ds} = V_{gs}-V_t \quad \text{Saturation.} \end{cases}$$

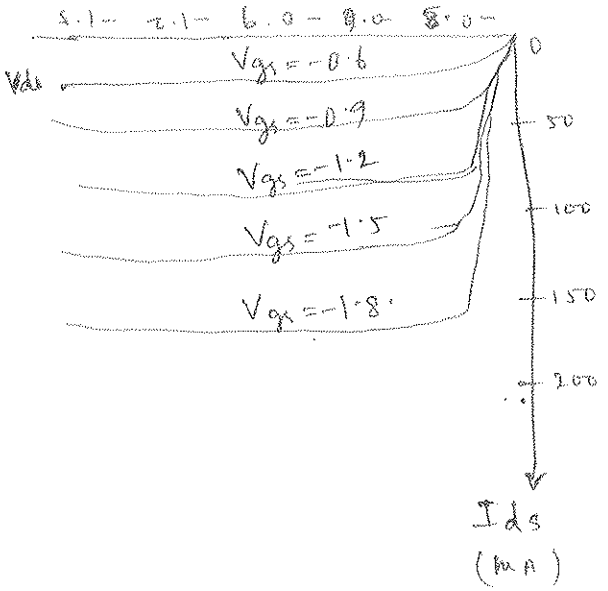
I-V characteristics for the transistor

n MOS:



* According to the first order model, current is zero for gate voltage below V_t
* For higher gate voltages, current increases linearly with Vds for small Vds.
* As Vds reaches the saturation pt. ie $V_{gs}-V_t$ current roll offs eventually becomes independent of Vds when the transistor is saturated.

PMOS transistor



* Same as nMOS, but with the signs reversed, IV characteristics in the third quadrant.
* Mobility of holes in Si is typically lower than that of e^-, so it provides less current than nMOS
* Mobility ratio $M = \frac{M_n}{M_p} = \frac{e^-}{\text{holes}}$.

C-V characteristics

* Each terminal of MOS has capacitance to the other terminals. These capacitances are non-linear and voltage dependent (C-V).

Simple MOS capacitance Model:

* In MOS, the gate is considered as parallel plate capacitor with thin oxide dielectric.
* The capacitance $\boxed{C_g = C_o WL}$
* MOS transistors are having minimum manufacturable L.

Let $C_g = C_p W$

Where, $C_p = C_o L$ → parasitic capacitance (or) diffusion capacitance

Detailed MOS-gate capacitance Model:

* The gate capacitances has two components.
  ① Intrinsic capacitance (over the channel)
  ② Overlap capacitance (to the source, drain)

* Intrinsic capacitance $\boxed{C_0 = C_{ox} WL}$

a) Cut-off

* If the MOS is OFF, the channel is not inverted and charge on the gate is matched with the opposite charge from the body. It is known as gate to body capacitance (C_{gb}).

b) Linear : $(V_{gs} > V_t)$.

* The channel is connected to the source and drain rather than body

$$\boxed{C_{gs} = C_{gd} = C_0/2}$$

c) Saturation $(V_{ds} > V_{gs} - V_t)$

* The channel is pinched off

$$\boxed{C_{gs} = \tfrac{2}{3} C_0}$$

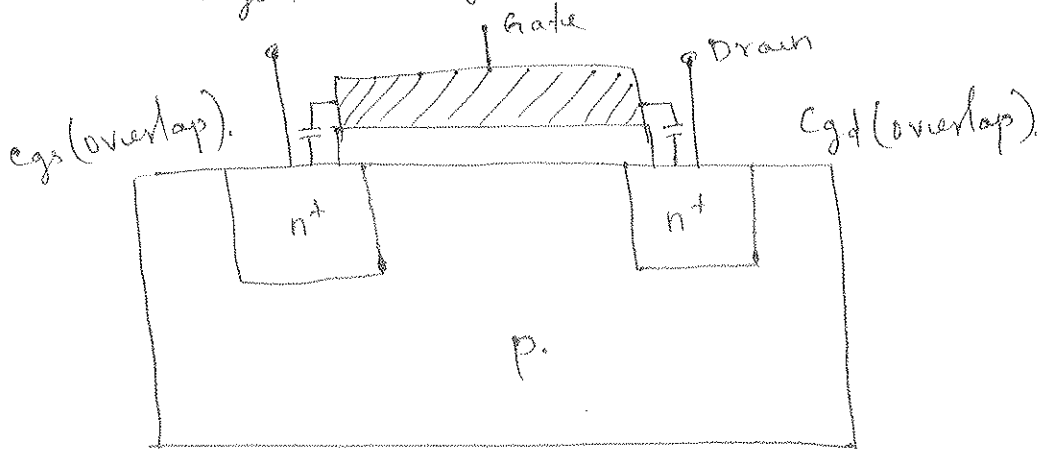
Intrinsic gate capacitance

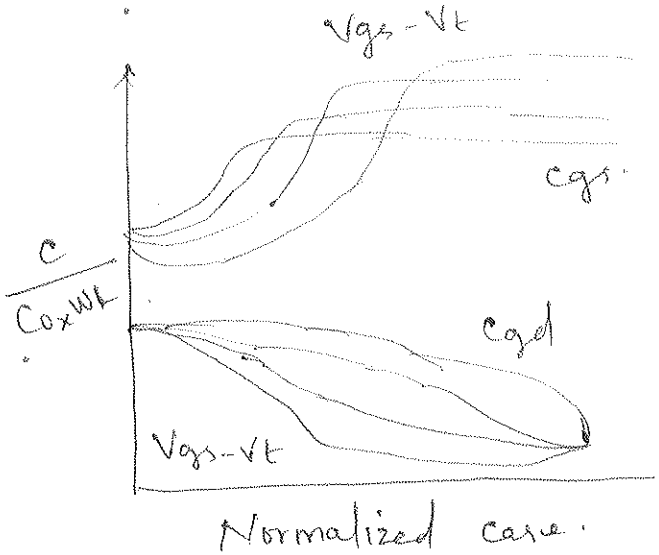
	cut off region	linear region	saturation
C_{gb}	C_0	0	0
C_{gs}	0	$C_0/2$	$\frac{2}{3} C_0$
C_{gd}	0	$C_0/2$	0
$C_g = C_{gs} + C_{gd} + C_{gb}$	C_0	C_0	$\frac{2}{3} C_0$

overall capacitance

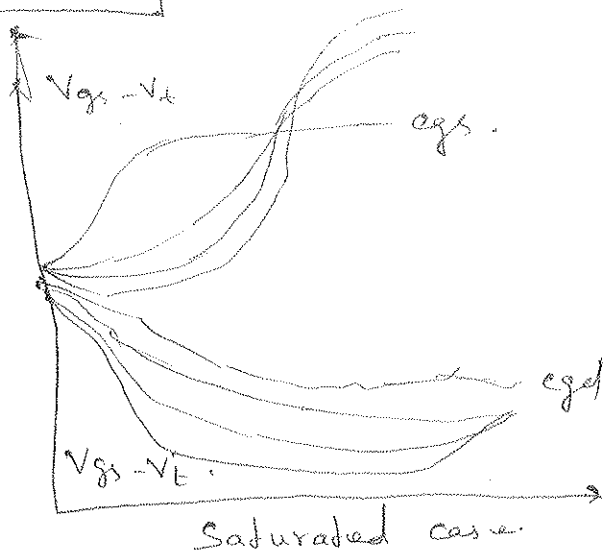
* The gate overlaps the source and drain by a small amount.

$$C_{gs01} = C_{gd01} = 0.2 - 0.4\ fF/\mu m.$$





Normalized case.

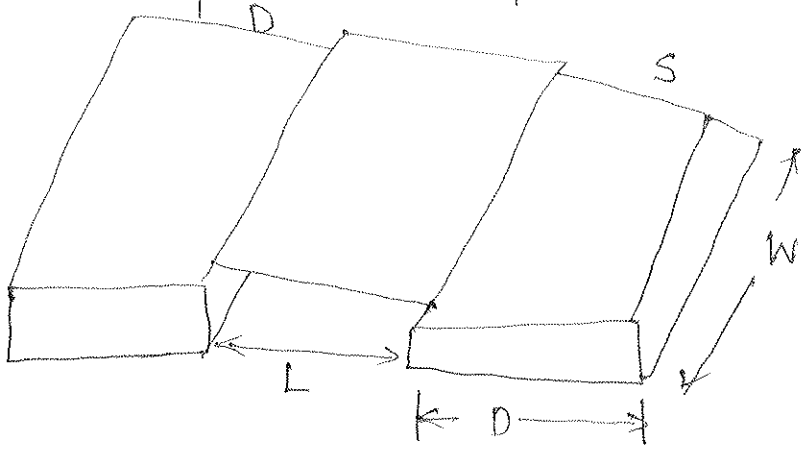


Saturated case.

$$C_{gs(overlap)} = C_{gs01} \cdot W$$
$$C_{gd(overlap)} = C_{gd01} \cdot W$$

MOS diffusion capacitance model

* The parasitic capacitance depends on area (As) and perimeter (ps).



* Area $AS = WD$

perimeter $PS = 2W + 2D$

* Total source parasitic capacitance.

$$C_{sb} = AS\, C_{jbs} + PS\, C_{jbssw}$$

Where $C_{jbs} \rightarrow C_j \left(1 + \frac{V_{sb}}{\phi_0}\right)^{-M_j}$

$C_j \rightarrow$ junction capacitance at zero bias

$M_j \rightarrow$ Junction grading coefficient

$\phi_0 \rightarrow$ built in potential depends doping level

$$\phi_0 = V_T \ln \frac{N_A N_D}{n_i^2}$$

$N_A \rightarrow$ doping level of body

$N_D \rightarrow$ doping level of source diffusion region

$n_i \rightarrow$ intrinsic carrier concentration in undoped silicon ($1.45 \times 10^{10}\ cm^{-3}$).

$V_T \rightarrow$ Thermal voltage $= \frac{KT}{q} \Rightarrow K = 1.38 \times 10^{-23}\ J/K$

* Side wall capacitance $C_{jbssw} = C_{JSW} \left(1 + \frac{V_{sb}}{\phi_0}\right)^{-M_{JSW}}$

* side wall capacitance abutting the gate.

$$C_{jbsswg} = C_{JSWG} \left(1 + \frac{V_{sb}}{\phi_0}\right)^{-M_{JSWG}}$$

Non ideal I-V Effects.

* I-V characteristics equations of MOS transistors are

$$I_{ds} = \begin{cases} 0 & ; V_{gs} - V_t \text{ , cut off.} \\ \beta\left(V_{gs} - V_t - \frac{V_{ds}}{2}\right)V_{ds} & ; V_{ds} < V_{dsat} \text{ , Linear region} \\ & 0 < V_{ds} < V_{gs} - V_t. \\ \beta/2\left(V_{gs} - V_t\right)^2 & ; V_{ds} > V_{dsat} \text{ Saturation} \\ & \text{region } 0 < V_{gs} - V_t < V_{ds}. \end{cases}$$

* Saturation current increases less than quadratically with increasing V_{gs} because of

① Velocity saturation:

* Carrier drift velocity and current increases linearly with the lateral electric field.

$$\boxed{E = \frac{V_{ds}}{L}} \longrightarrow \text{True for weak field.}$$

* At high field strength drift velocity rolls off due to carrier scattering and eventually saturates at V_{sat}

$$V_{sat} = \mu E_{sat}$$

$$V_{sat} = 6 - 10 \times 10^6 \text{ cm/s} \quad \text{for electrons.}$$

$$= 4 - 8 \times 10^6 \text{ cm/s} \quad \text{for holes.}$$

$$E_{sat} = 2 \times 10^4 \text{ V/cm} \quad \text{for nMOS.}$$

* The saturation current without velocity saturation is given by

$$I_{ds} = \frac{\mu C_o W}{L}\left(\frac{V_{gs} - V_t}{2}\right)^2$$

* With Saturation

$$I_{ds} = \frac{\mu C_o W}{2L} (V_{gs} - V_t)(V_{gs} - V_t).$$

$$I_{ds} = \frac{\mu C_o W}{L} (V_{gs} - V_t) \frac{(V_{gs} - V_t)}{2}.$$

$$= \frac{\mu C_o W}{2L} (V_{gs} - V_t) \cdot V_{ds} \qquad \therefore V_{ds} = V_{gs} - V_t.$$

$$= \frac{\mu C_o W}{2} (V_{gs} - V_t) \cdot \frac{V_{ds}}{L}$$

$$= \frac{\mu C_o W}{2} (V_{gs} - V_t) E$$

$$I_{ds} = C_o W (V_{gs} - V_t) \cdot V \qquad \therefore \frac{\mu E}{2} \simeq V$$

If $V = V_{sat}$

$$\boxed{I_{ds} = C_o W (V_{gs} - V_t) V_{sat}}$$

* The drain current is quadratically dependent on voltage without velocity saturation and linearly dependent when fully velocity saturated.

* Using alpha power model.

$$I_{dsat} = B P_c (V_{gs} - V_t)^{\alpha} \qquad \alpha \rightarrow \text{velocity modulation index}$$

$$V_{dsat} = P_v (V_{gs} - V_t)^{\alpha/2}$$

$V = V_{sat}$

$$I_{ds} = \begin{cases} 0 & ; \ V_{gs} < V_t \ \text{cut off.} \\ I_{dsat} \cdot \frac{V_{ds}}{V_{dsat}} & ; \ 0 < V_{ds} < V_{gs} - V_t \cdot \text{non Saturation} \\ I_{dsat} & ; \ 0 < V_{gs} - V_t < V_{ds} \ \text{Saturation.} \end{cases}$$

② Mobility Variation :- (μ).

$$\mu = \frac{\text{Average carrier drift velocity}(v)}{\text{Electric field.}} \quad = \frac{cm/s}{V/cm}$$

* Mobility varies with the type of charge carriers e^- in silicon have much higher mobility than holes.
* Mobility decreases with increasing doping concentration
* Mobility decrease with increasing temperature.

③ channel length Modulation

* The depletion region effectively shortens the channel length

$$L_{eff} = L - L_{short}.$$

Where $L_{short} = \sqrt{2 \frac{\varepsilon_{si}}{q N_A} (V_{ds} - (V_{gs} - V_t))}$

* If V_{ds} is increased the effective channel length is decreased

$$I_{ds} = \beta \frac{(V_{gs} - V_t)^2}{2} (1 + \lambda V_{ds}).$$

$\lambda \rightarrow$ empirical channel length modulation

$\lambda \rightarrow 0.02\ V^{-1}$ to $0.005\ V^{-1}$

④ Threshold voltage - Body effect :-

* Threshold voltage V_t is not const with respect to the voltage difference between the substrate and the source of MOS transistor. This is known as substrate bias effect or body effect.

$$V_t = V_{fb} + 2\phi_b + \frac{\sqrt{2 \varepsilon_{si} q N_A (2\phi_b + |V_{sb}|)}}{C_{ox}}.$$

$$V_t = V_{t0} + \gamma\left[\sqrt{(2\phi_b + |V_{sb}|)} - \sqrt{2\phi_b}\right]$$

Where V_{sb} → substrate bias

V_{t0} → threshold voltage for $V_{sb} = 0$

γ → Constant that describes substrate bias effect $= \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2q\varepsilon_{si}N_A}$

$$= \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{si}N_A}.$$

γ Ranges from 0.4 to 0.12.

ε_{ox} → dielectric constant of SiO_2

ε_{si} → dielectric constant of silicon substrate.

N_A → doping concentration density of the substrate.

⑤ Subthreshold Region

* Current flows from source to drain is zero when $V_{gs} > V_t$ ideally.

* Practically current doesnot abruptly cut off below threshold. It drops off exponentially. This conduction is known as leakage.

* Subthreshold case.

$$I_{ds} = I_{ds0}\frac{(V_{gs} - V_t)}{nV_T}\left[1 - e^{-\frac{V_{ds}}{V_T}}\right]$$

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

Where I_{ds0} → current at threshold

n → process dependent term affected by depletion region characteristics

V_t → thermal voltage.

(6) Junction leakage.

* The pn junction formed between diffusion and the substrate.
* Junction leakage current $I_D = I_s \left[e^{V_D/V_T} - 1 \right]$

I_s = diode reverse biased saturation current.

* If the junction is reversed biased by significantly more than the thermal voltage then the leakage is $-I_s$.

(7) Temperature dependence :-

* Carrier mobility is decreased with the temperature

$$\mu(T) = \mu(T_1) \left(\frac{T}{T_1} \right)^{-k_\mu}$$

T → absolute temperature

T_1 → room temperature

k_μ - fitting parameter.

* V_T is decreased with temperature. Junction leakage is increased with temperature and I_{dsat} is increased with temperature.

(8) Geometry dependence :-

* Effective transistor length and width

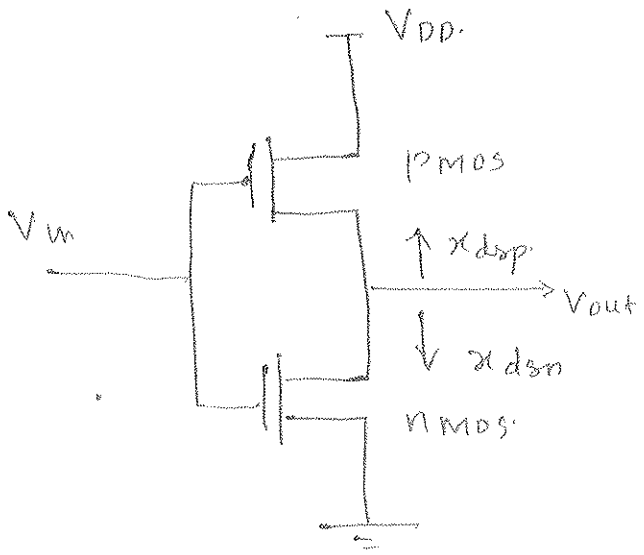
$$L = L_d + x_L - 2LD$$

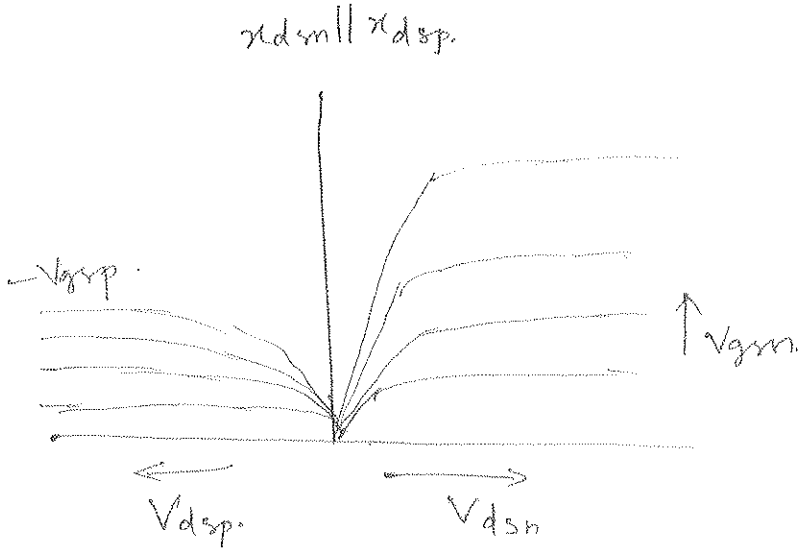
$$W = W_d + x_W - 2WD$$

x_L, x_n → Actual gate dimension.

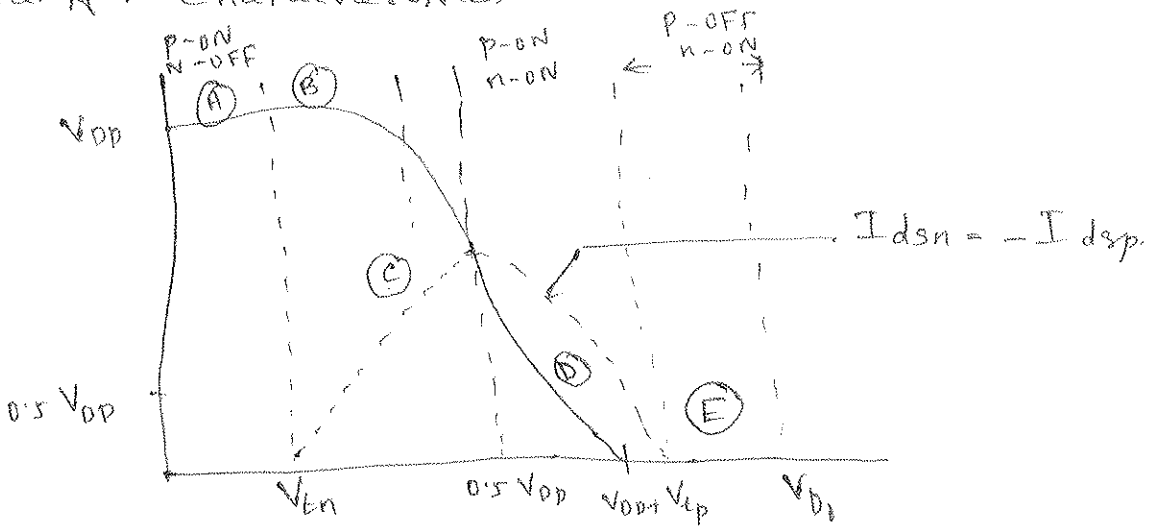
DC Transfer characteristics

* CMOS inverter is realized by the series connection of p and n device.





Transfer characteristics



* The switching point is designed to be 50% of the magnitude of the supply voltage = $\frac{V_{DD}}{2}$.

* During transition both transistor in the CMOS inverter is momentarily ON resulting in short pulse of current drawn from the power supply.

Delay Estimation

* Critical paths can be affected at four levels.
  1. Architectural level
  2. Logic level
  3. Circuit level
  4. Layout level.

These paths can be identified by using timing analyzer.

- Slope rate (or) Rise time (t_r):-

* The time taken by the waveform to rise from 20% to 80% of its steady state value.

Edge rate (or) Fall time (t_f):

* The time taken by the waveform to fall from 80% to 20% of its steady state value.

Edge rate (t_e)

$$t_e = \frac{t_r + t_f}{2}$$

Propagation delay time (t_{pd}) of Max time:

* The maximum time from the input crossing 50% to the output crossing 50%

Contamination delay time (t_{cd}) or Min time:-

* Minimum time from the input crossing 50% to the output crossing 50%

RC Delay Model

* The delay of the logic gates can be estimated as the RC product of the effective drivier resistance and the load resistance.

Effective resistance and capacitance:

* The unit width nMOS has effective resistance of R
* The unit width PMOS has effective resistance of 2R.
* If the PMOS has double unit width then it has effective resistance of R

Resistance $R = ZR_s$

Where $Z \rightarrow$ aspect ratio $R_s \rightarrow$ sheet resistance

* If transistors are in series then their resistance is the sum of individual resistance.
* If transistors are in parallel then the effective resistance is equal to the resistance of the single transistor

Diffusion capacitance

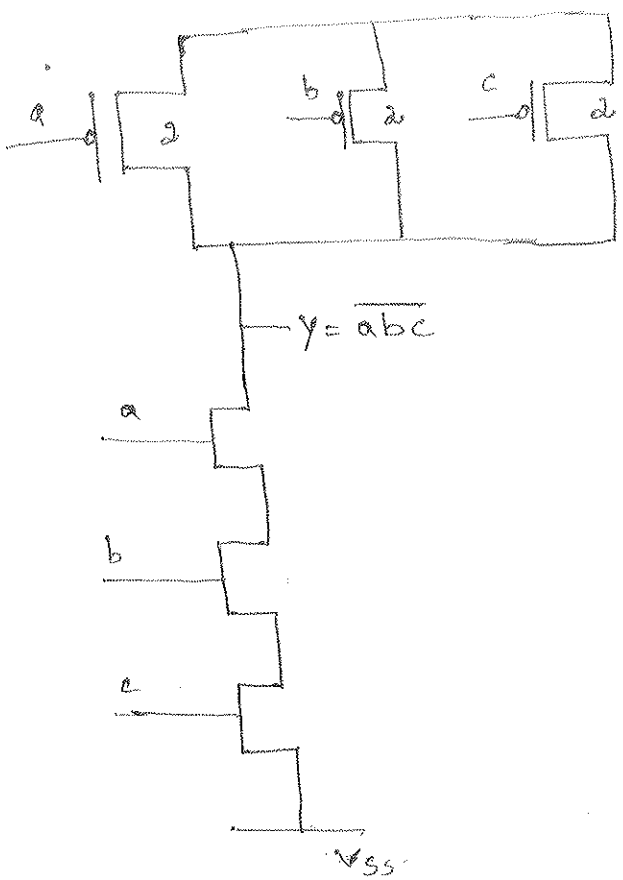
* If diffusion nodes are shared then diffusion capacitance is reduced.
* Uncontacted diffusion nodes between series transistors are usually smaller than those must be Contacted.
* Uncontacted nodes have less capacitance.

* The gate capacitance can be calculated by using transistor widths.

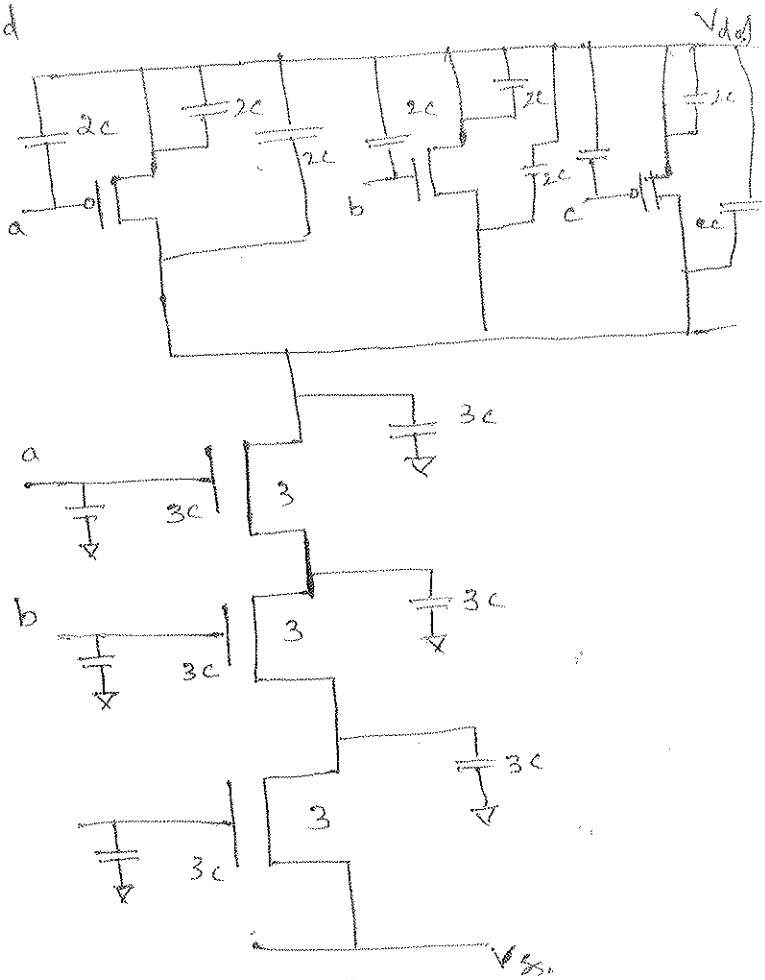
* n MOS transistor is n times unit size and has gate-Source capacitance = nc.

* pMOS transistor is 2n times unit size and has gate Source capacitance = 2nc.

Eg: Effective resistance



Effective capacitance



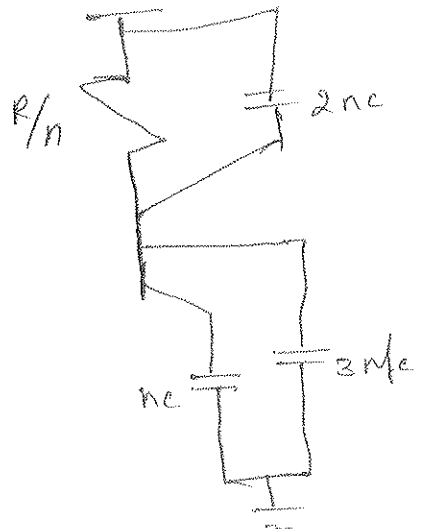
Calculation of fall and rise delay.

* Total input capacitance = nc + 2nc
= 3nc

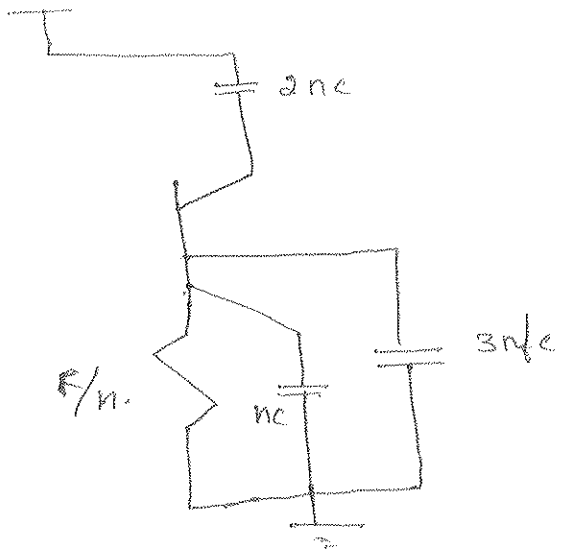
* For a fanout of f load capacitance = 3nfc

* The resistance of nMOS with unit size $\left(\frac{4}{2}\lambda\right)$ = R/n.

The resistance of pmos with unit size $\left(\frac{4}{3}\lambda\right) = \frac{2R}{n}$.



Raising circuit



Falling circuit

The propagation delay = R/n [input capacitance + load capacitance]

$= R/n\,[nc + 2nc + 3nfc]$

$= 3RC\,[f+1]$

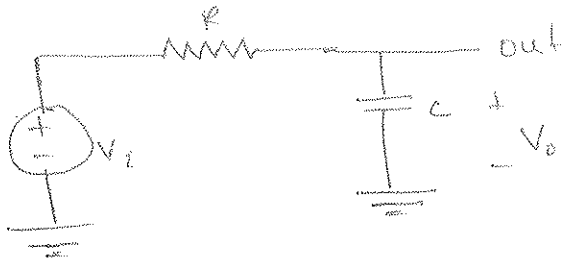
$= \boxed{3\tau(f+1)}$

$f\,\tau = RC$

Elmore Delay model

* Elmore delay model network has
  1. Single input node
  2. All the capacitors are between a node and ground
  3. Network doesn't contain any resistive loops.

Consider a single RC ladder network.



for the input side

$$V_i(t) = Ri(t) + \frac{1}{C}\int_0^t i(t)\,dt.$$

By taking laplace transform

$$V_i(s) = RI(s) + \frac{1}{C} * \frac{1}{s} I(s)$$

by taking laplace transform

$$V_i(s) = RI(s) + \frac{1}{C}\cdot\frac{1}{s} \cdot I(s)$$

$$V_i(s) = I(s)\left[R + \frac{1}{Cs}\right]$$

$$\boxed{I(s) = \frac{V_i(s)}{R + \frac{1}{Cs}}}$$

For that output side

$$V_o(s) = \frac{1}{Cs} I(s)$$

Sub. value of $I(s)$

$$V_o(s) = \frac{1}{Cs} \cdot \frac{V_i(s)}{R + \frac{1}{Cs}}.$$

* Transfer function of RC network

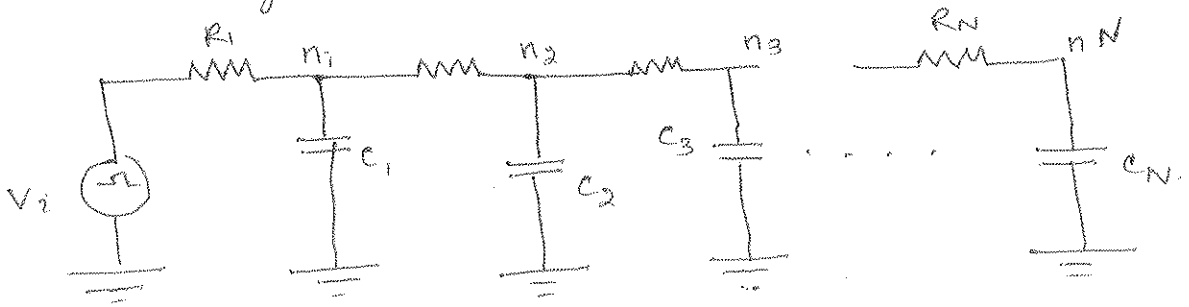
$$H(s) = \frac{V_o(s)}{V_i(s)} = \frac{1}{RCs + 1}$$

$T = RC$ is the change in the output voltage $V_o(t)$ is delayed by the time constant.

Delay at node 1 : $T_1 = R_1 C_1$

Delay at node 2 : $T_2 = (R_1 + R_2) C_2$

Delay at node 3 : $T_3 = (R_1 + R_2 + R_3) C_3$.



$$T_n = RC + 2RC + \dots + nRC.$$

* Propagation delay $t_{pd} = \sum_i R_{n-1} C_i$

$$t_{pd} = \sum_{i=1}^{N} C_i \sum_{j=1}^{i} R_j$$

Linear delay model

* propagation delay d = Effort delay (f) + parasitic delay (p).

Logic effort = g

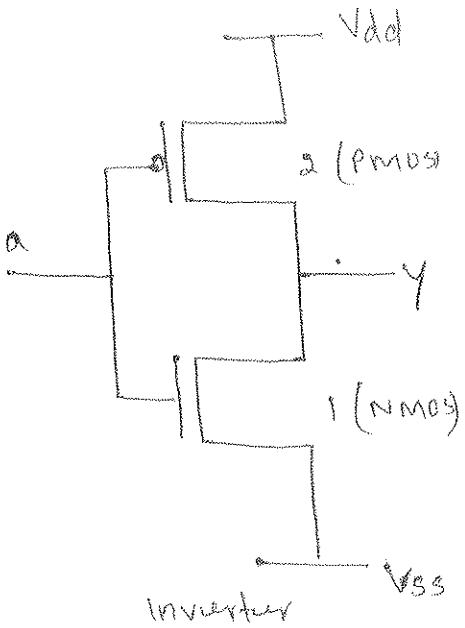
$f = gh$

$h = \frac{C_o}{C_{in}}$

Where C_o → Capacitance of external load

C_{in} → input capacitance of the gate.

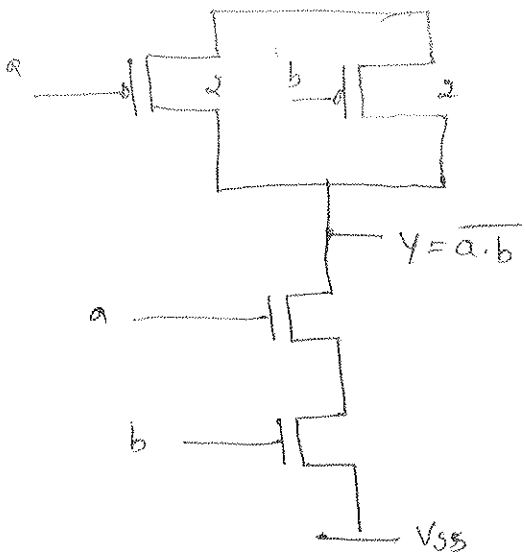
Logical effort (g) is defined as the ratio of the input capacitance of the gate to the input capacitance of an inverter that can deliver the same output current.



$n \rightarrow$ no of inputs

$$g = \frac{(n+2)}{3} = \frac{3}{3} = 1$$

$\boxed{g = 1}$



$$g = \frac{(n+2)}{3} = \frac{2+2}{3}$$

$\boxed{g = \frac{4}{3}}$

Parasitic delay

* It is the delay of the gate when it drives zero load it can be estimated with RC delay models.

* Normalised parasitic delay is the ratio of diffusion capacitance to gate capacitance for a process. It is designated as P_{inv} and approximately equal to 1.

Gate	No. of inputs 1	2	3	4
Inverter	1	—	—	—
NAND	—	2	3	4
NOR	—	2	3	4
MUX	2	4	6	8

Parasitic delay of

Inverter = 1

NAND = n

NOR = n

MUX = 2n.

Delay in a logic gate:

* propagation delay interms of complexity of a gate is given by g and interms of capacitive fanout is expressed as h

propagation delay = Effort delay + parasitic delay

$$d = f + p$$

$$fd = gh$$

$h = \frac{C_0}{C_{in}}$; C_0 → Capacitance of the external load

C_{in} → input capacitance of the gate.

Scaling

The design of high density chips in MOS VLSI technology requires high packing density and consequently the sizes of the transistors are as small as possible. The reduction of size (i.e) the dimensions of MOSFET is commonly referred to as scaling.

Scaling of MOS transistors is concerned with systematic reduction of overall dimension of the devices as allowed by the available technology which preserving the geometric ratios found in larger devices. The proportional scaling of all devices in a circuit would certainly result in a reduction of total silicon area

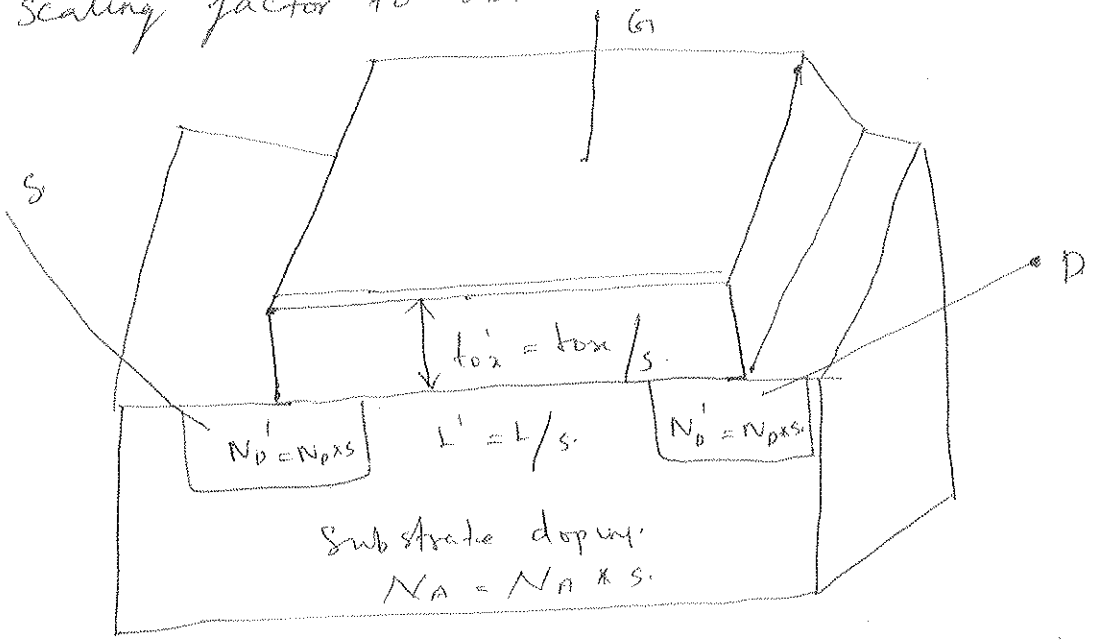
occupied by the circuit, thereby increasing the overall functional density of the chip.

Some physical limitations eventually restrict the extent of scaling that is practically achievable. They are

(I) Full scaling (constant field scaling)

(II) Constant voltage scaling

To describe device scaling we introduce a constant scaling factor $S > 1$. All horizontal and vertical dimensions of the large size transistor are then divided by this scaling factor to obtain the scaled device.



Scaling of a typical MOSFET by a scaling factor S.

Full scaling (constant field scaling)

To preserve the magnitude of internal electric field in the MOSFET, while the dimensions are scaled down by a factor S the scaling is done

To achieve this goal, all potentials must be scaled down proportionally by the same scaling factor. At last the poission equation is used to describe the relationship between the charge densities and electric field. The charge density must be increased by a factor S in order to maintain the field conditions.

UNIT II Combinational MOS Logic circuits

Circuit Families : Static CMOS, Ratioed circuits, Cascode Voltage switch logic, Dynamic circuits, pass Transistor Domino, Dual Rail Domino, CPL, DCVSPG, DPL, circuit pitfalls. power: Dynamic power, Static power, low power architecture.

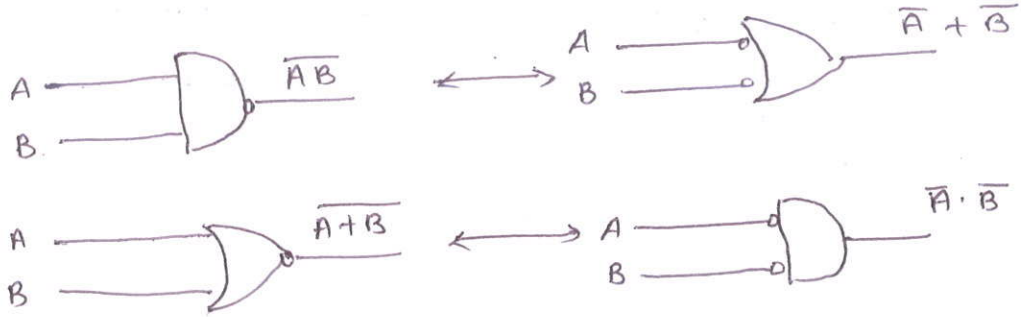
Circuit Families

* Alternative CMOS Logic configurations can be called as circuit families.

Static CMOS :-

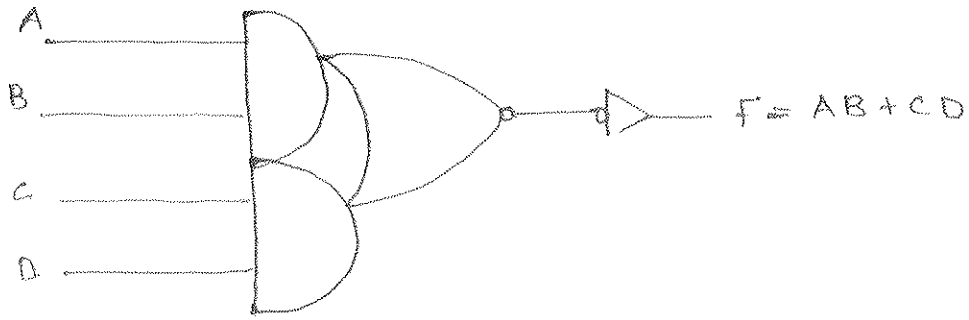
① Bubble pushing

* CMOS stages are inherently inverting. So AND and OR functions must be built from NAND and NOR gates.
* NAND gate is equivalent to an OR of inverted inputs.
* NOR gate is equivalent to an AND of inverted inputs.
* Switching between these representations is easy to do on a white board and is often called bubble pushing

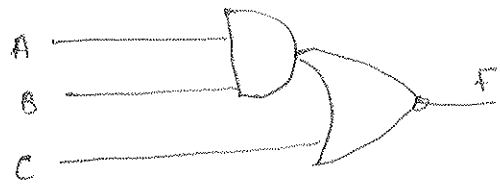


② Compound Gates.

The function AB+CD can be computed with an AND-OR-INVERTER 22 (AOI 22) gate and an inverter.



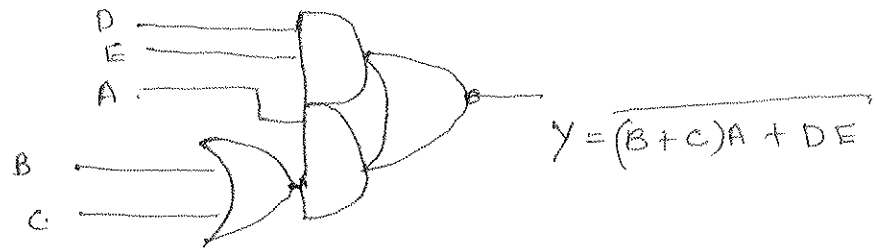
The transistor widths are chosen to give the same drive as a unit inverter.



AOI 21

$F = \overline{AB+C}$

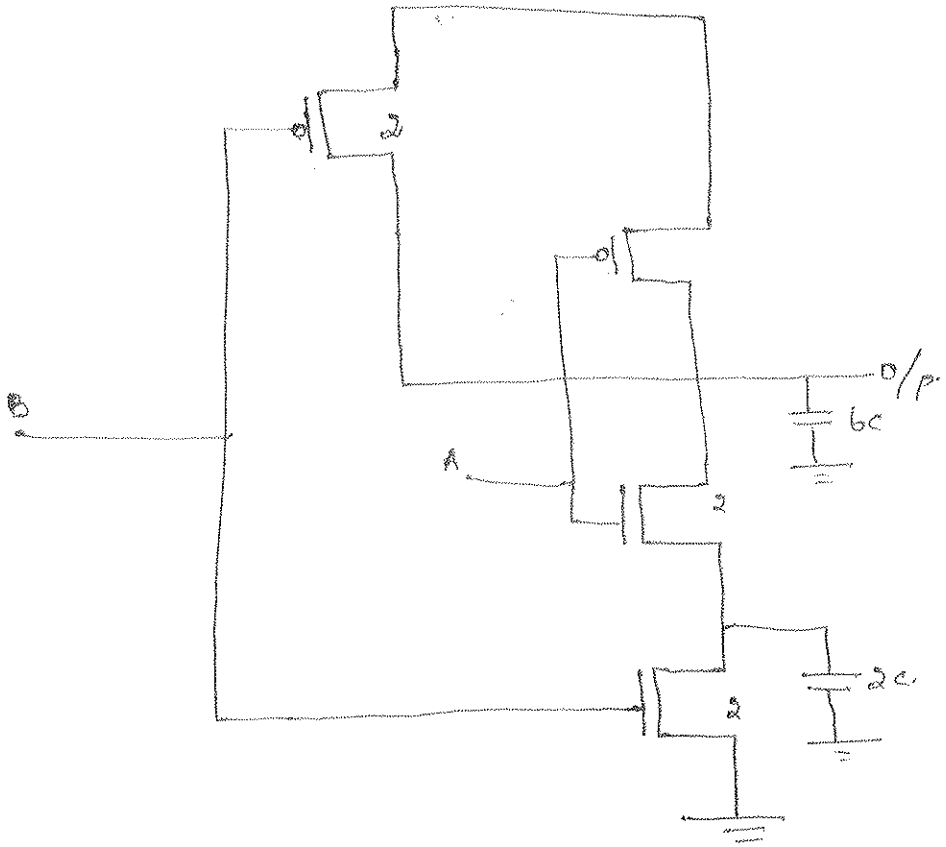
Complex AOI



③ Input ordering delay effect:

* The logic effort and parasitic delay of different gate input is often different.
* The outer input to be the input closer to the supply rail and the inner input to be the input closer to the output (2/p A).

* The parasitic delay is smaller when the inner i/p switches last because the intermediate nodes have already been discharged.



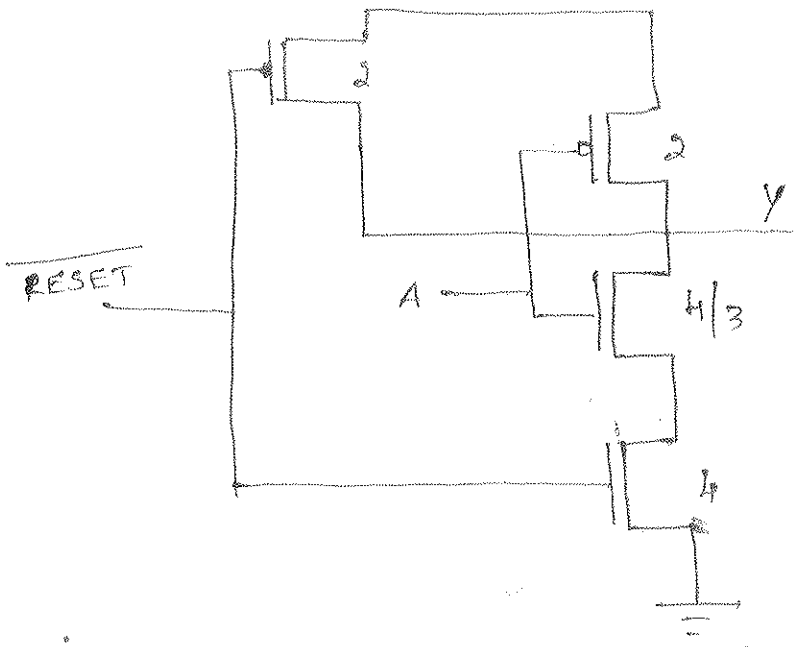
NAND gate delay estimation

④ Asymetric Gates.

* When one input is far less critical than another, Symmetric gates can be made asymmetric to favor the late i/p.

* Consider the input reset is slow, the circuit should be optimized for i/p to o/p delay in the expense of reset

Asymmetric NAND Gate.



* pull down resistance $= R/4 + R/4/3 = R$

  Capacidance at A $= 10/3$

  logical effort $g_A = 10/9$

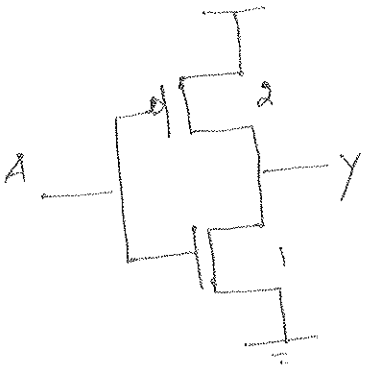
* This g_A is better than the logical effort of ordinary NAND gate.

(5) Skewed Gates.

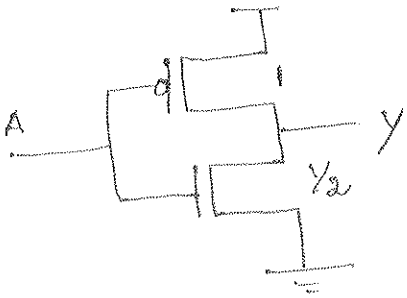
* If one input transition is more important than the others, the skewed gates are used.
* Hi skew gates are used to favour the rising output transition
* Lo skew gates are used to favour the falling output transition
* This can be done by decreasing the size of non-critical transistor.
* Logical effort for falling transition $= g_d$

Logical effort for rising transition = g_u

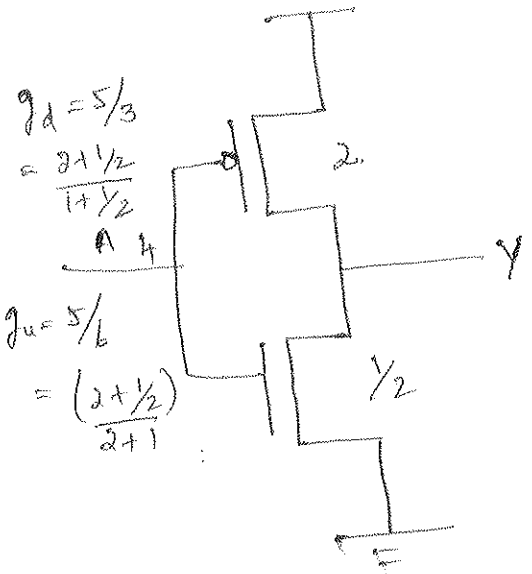
$$g_d \text{ (or) } g_u = \frac{\text{i/p capacitance of skewed gate}}{\text{i/p capacitance of unskewed inverter.}}$$



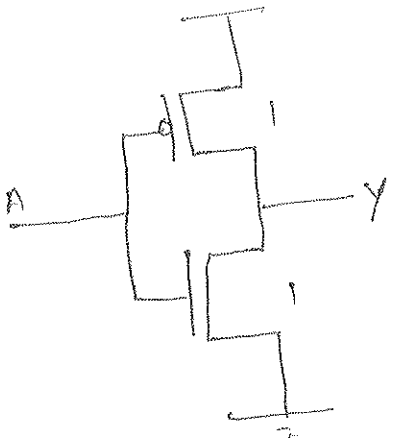
unskewed inverter
(equal rise resistance)



unskewed inverter
(equal fall resistance).



HI skew inverter



$g_d = 2/3$

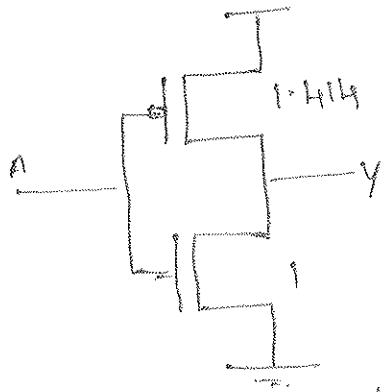
$g_u = 4/3$.

Lo skew inverter

P/N Ratio:

* Reducing the PMOS size from 2 to $\sqrt{2} = 1.414$ for the inverter gives the fastest average delay. But this delay improvement is only 2%

* This reduces the PMOS area, capacitance, power consumption and this moves the switch point lower and reduces the noise margin.



$g_u = 1.15$

$g_d = 0.81$

Multiple threshold voltages.

* Some CMOS process offers two or more threshold voltages. Transistors with lower threshold voltage produces more ON current, but leaks more OFF current exponentially

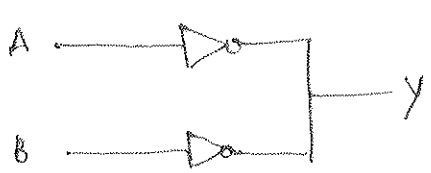
Ratioed circuits

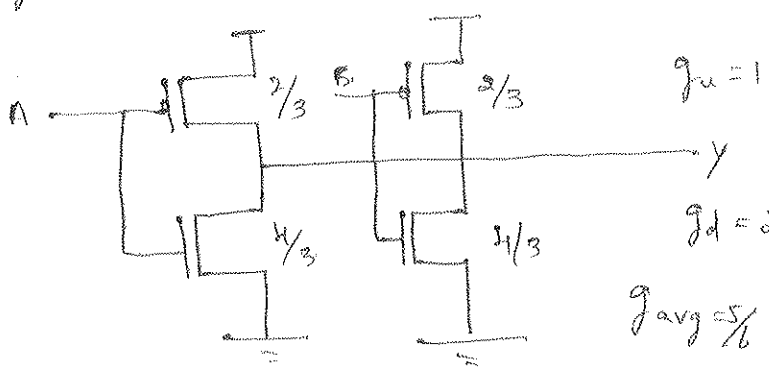
ⓐ Ganged CMOS

ⓑ Source follower pullup

ⓒ pseudo nMOS.

ⓐ Ganged CMOS or Symmetric NOR gate.



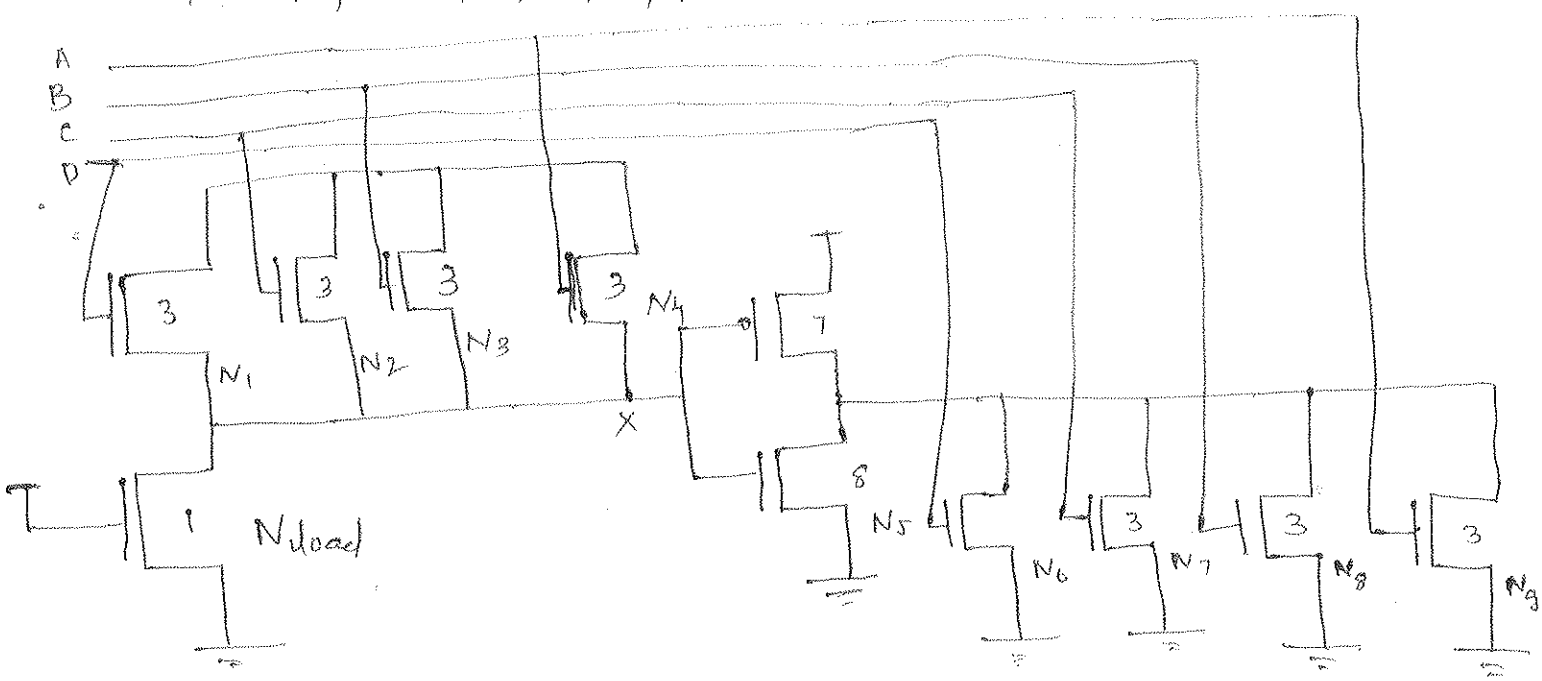


* When one input is '1' and the other a '0' the circuit operate as a pseudo nMOS.

* When both inputs are '0' both PMOS turns ON in parallel pulling the o/p faster than in pseudo nMOS.
* When both inputs are 1, both PMOS are off reducing static power dissipation.

(b) Source follower pull up logic (SFPL)

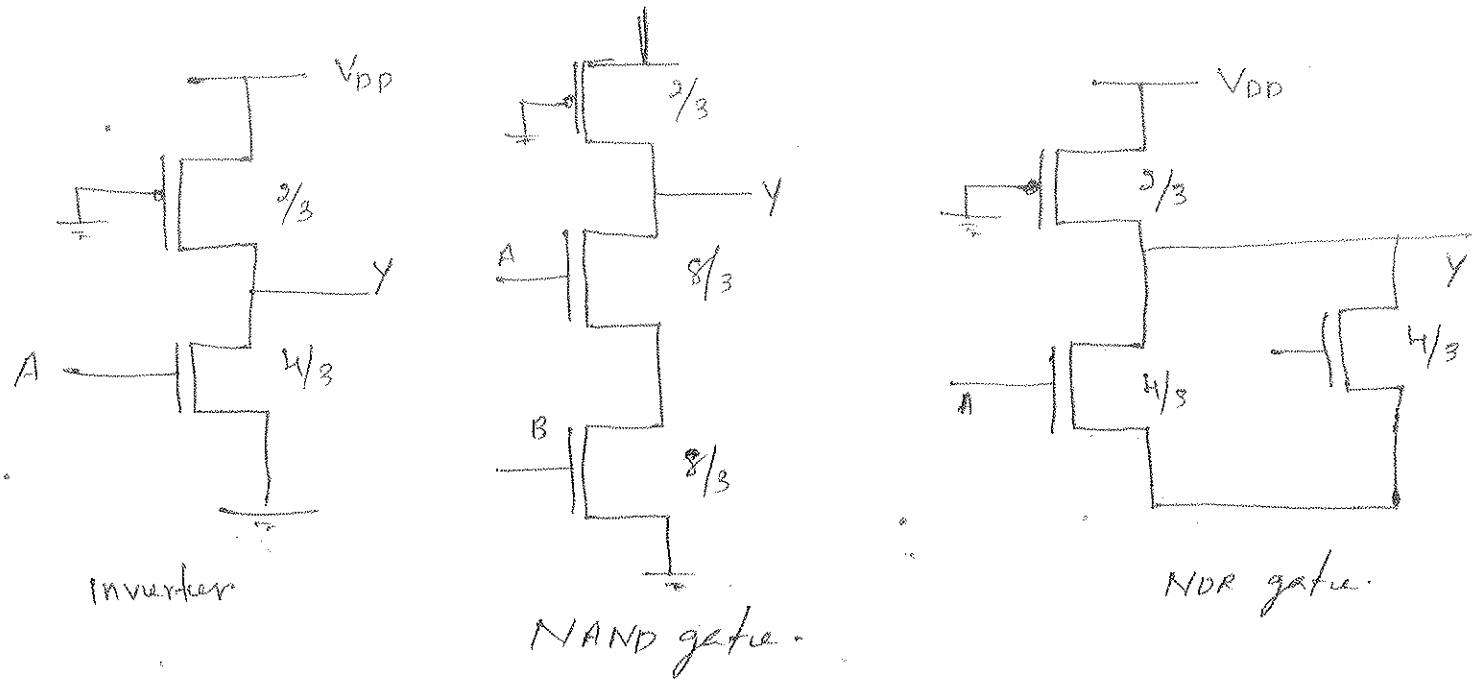
* 4 input NOR gate.
* pull up is controlled by input.
* $N_6 - N_9$ and P_1 form pseudo nMOS NOR
* Gate of pull up 'P_1' is controlled by source follower $N_1 - N_4$ and Nload.



* When one input turns ON source follower pulls node x to approximately $V_{DD}/2$.
* This partially turns off P_1. Which allows smaller nMOS pulls down $N_5 - N_9$ to be used.
* SFPL is used for constructing wide NOR gates.

© pseudo n-MOS.

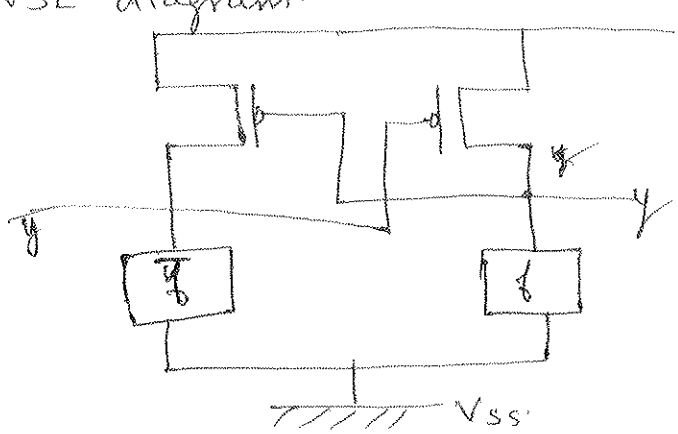
* In pseudo nMOS, the pull down network is like that of a static gate, but the pull up network is replaced with a single PMOS transistor that is grounded so it is always ON.
* pseudo nMOS is usefull for fast wide NOR gates or NOR based structures like ROM and PLA.



Cascade voltage switch logic

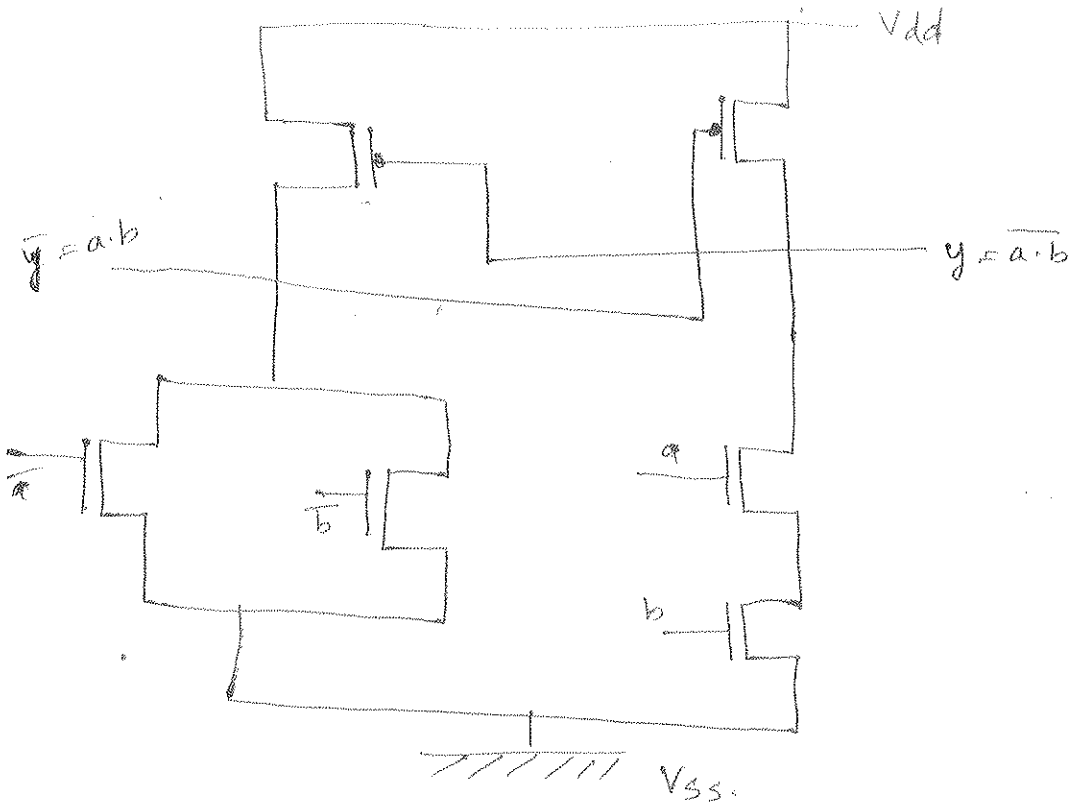
* It is also known as differential cascade voltage switch logic (DCVSL). cascade means that the transistors are connected in series

General CVSL diagram.



for any i/p one pull down network is ON and other is OFF.

CVSL AND and NAND gate.



Advantages

(i) Its speed is good, because all logic performed with n-MOS so input capacidance is reduced.

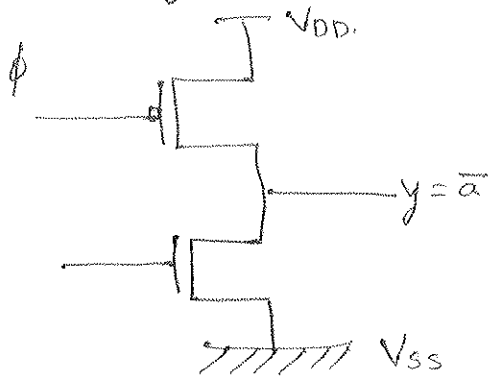
Dynamic circuits

Advantages

(i) It has lower input capacidance
(ii) No contention during switching
(iii) It has zero static power dissipation.

Operating Mode

1) Precharge Mode:

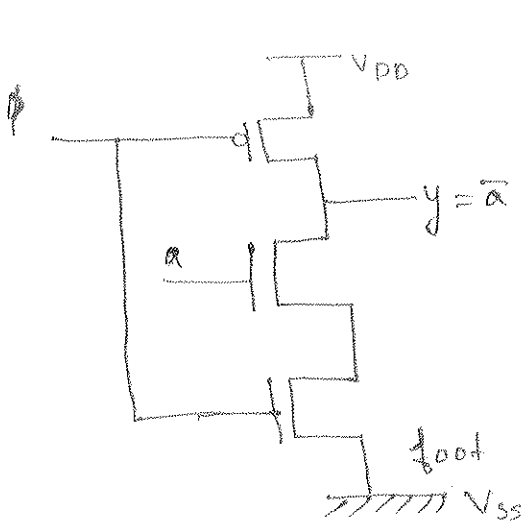


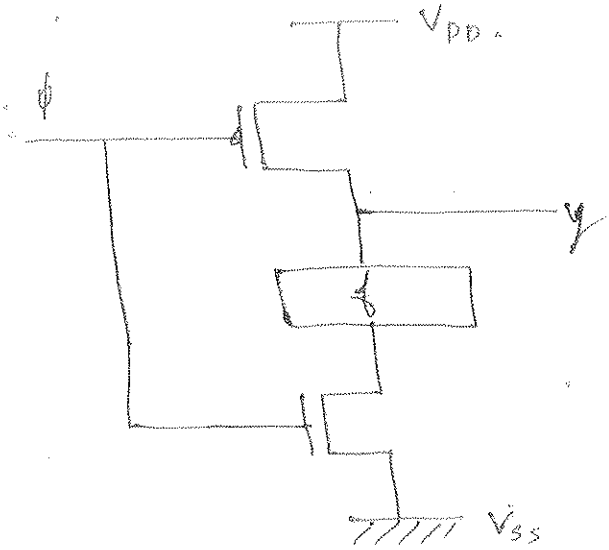
Here, ϕ is '0'

So clocked PMOS is ON

$y = 1$

(ii) Evaluate Mode:





* Here clock ϕ is 'l' clocked PMOS is OFF. The O/p may remain high or discharged low.

* If the input cannot be '0' during precharge, an extra transistor can be added at the bottom of the nMOS stack. It is known as foot.

* Monotonicity requirement is one of the difficulty in the dynamic circuits. While a gate is in evaluation mode, the inputs must be monotonically rising.

It means, the input can

→ start low and remain low (00)

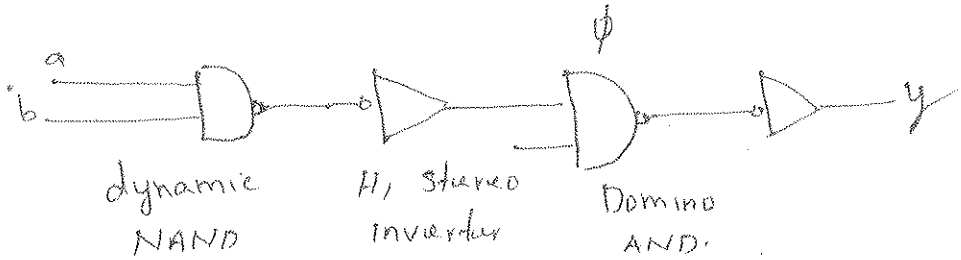
→ start low and rise high (01)

→ start high and remain high (11).

but not start high and fall low (10).

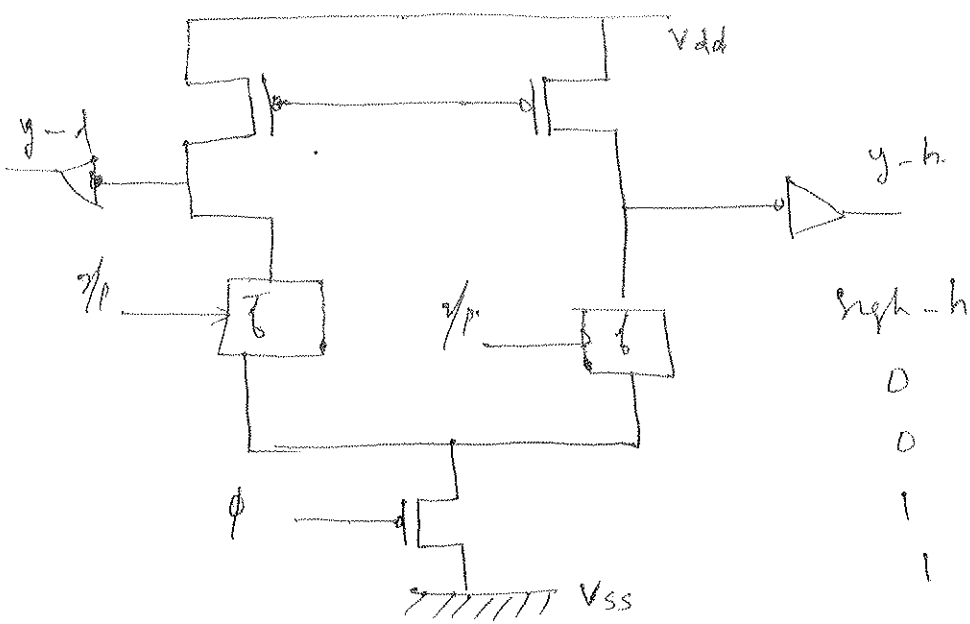
(a) Domino logic :-

* The monotonicity problem can be solved by connecting a static cmos inverter between the dynamic gate.
* It is used to convert monotonically falling output into a monotonically rising signal suitable for the next gate
* The dynamic static pair is known as domino gate.
* Hi skew gate is used as static inverter.



(b) Dual - Rail Domino logic.

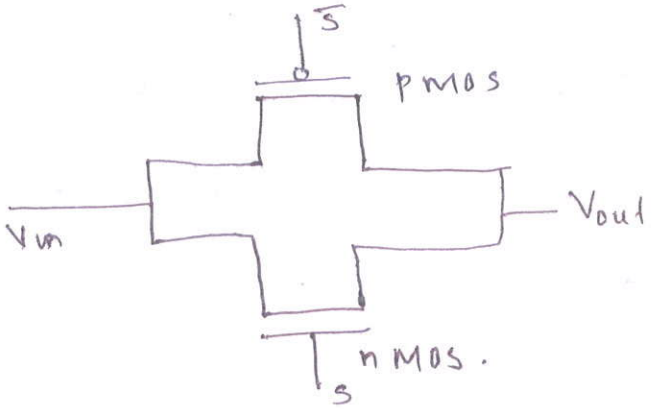
* It is a dynamic form of cascade voltage switch logic
* It encodes each signal with a pair of wires. The input and output signal pairs are $-h$ and l.
* If the o/p is '1' then $-h$ is asserted.
  If the o/p is '0' then $-l$ is asserted.

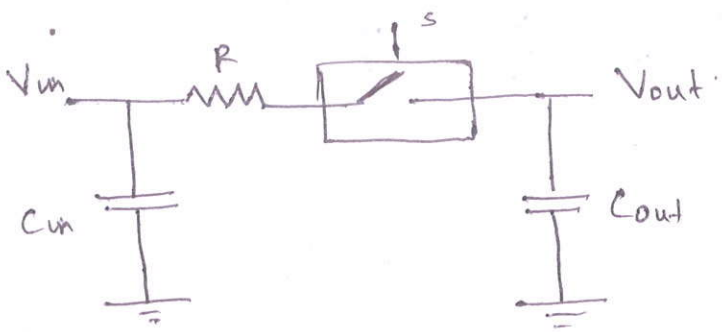


Sig-h	Sig-l	Meaning
0	0	Precharged
0	1	0
1	0	1
1	1	Invalid.

Pass transistor circuits

* Transmission gate is nothing but the parallely connected p-MOS and n-MOS.





Switching model

$R \rightarrow \max(R_n, R_p)$

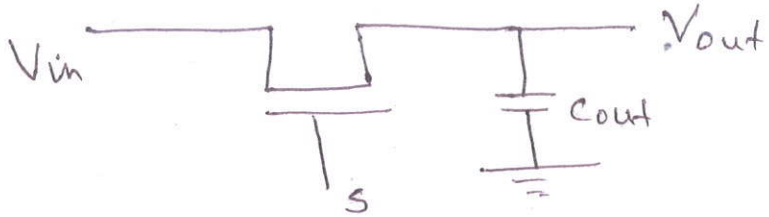
$R_n \rightarrow$ (n refers nMOS) n MOS Resistance.

$R_p \rightarrow$ (prefers p-MOS) pMOS Resistance.

* The transmission gate acts as voltage controlled resistor connecting the input and output.
* It can be used as a logic structure, switch, latch element etc.

Operation of transmission gate.

① nMOS pass transistor.



→ When S=0, nMOS is off

→ When S=1, $V_{in} = V_{DD}$, initial condition of Vout (at t=0) Σ_0

→ $V_{out}(t) = V_{max} \dfrac{t/\tau_n}{1 + t/\tau_n}$

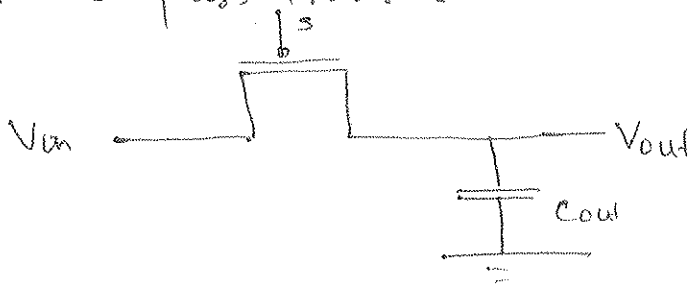
→ $V_{max} = V_{DD} - V_{tn}$ ⟹ maximum voltage that can be transferred through nMOS.

→ When limit $t \to \infty$ then

$$\lim_{t\to\infty} V_{out}(t) = V_{max}$$

→ t_r = Rise time = $1.8\tau_n = 1.8 R_n C_{out}$.

(ii) pMOS pass transistor



$\tau_p = R_p \cdot C_{out}$.

When $\bar{S} = 1$ (S=0), $V_{in} = V_{DD} \Rightarrow V_{out} = V_{SS}$.

When $\bar{S} = 0$ (S=1) $V_{in} = V_{SS} \Rightarrow V_{out} = V_{DD}$.

At this stage Cout discharges through pMOS until

$V_{out} = V_{tp}$ (Trans. stop conducting)

$t_r = 2.94\tau_p$ $\tau_p = R_p \cdot C_{out}$.

* At this stage Cout discharges through pMOS until $V_{out} = V_{tp}$ (At this point transistor stops conducting)

$t_r = 2.94\tau_p$, $\tau_p = R_p \cdot C_{out}$.

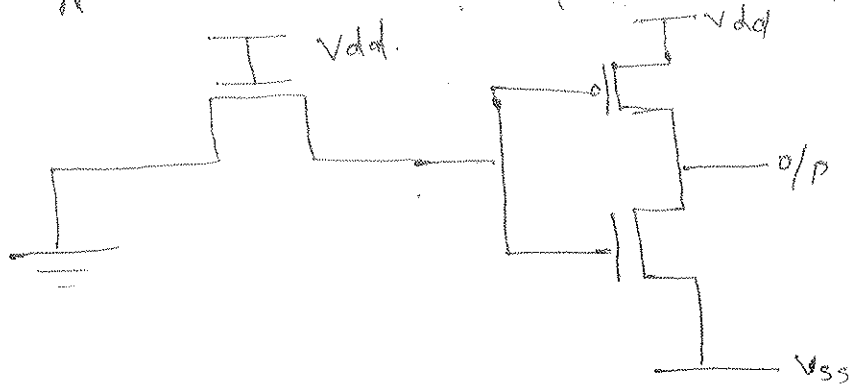
Circuit pitfalls:-

* Various pitfalls cause the chip to fail.

(a) Threshold drops.

* The output of the pass transistor swing to within V_t of the rail and output of pass transistors rise to $V_{DD} - V_t$.

* If voltage rise to $V_{DD} - V_t$, it is not sufficient to turn off PMOS, so static power dissipation occur.



(b) Leakage

* This problem is due to sub threshold conduction, gate tunneling and reverse biased diode leakage

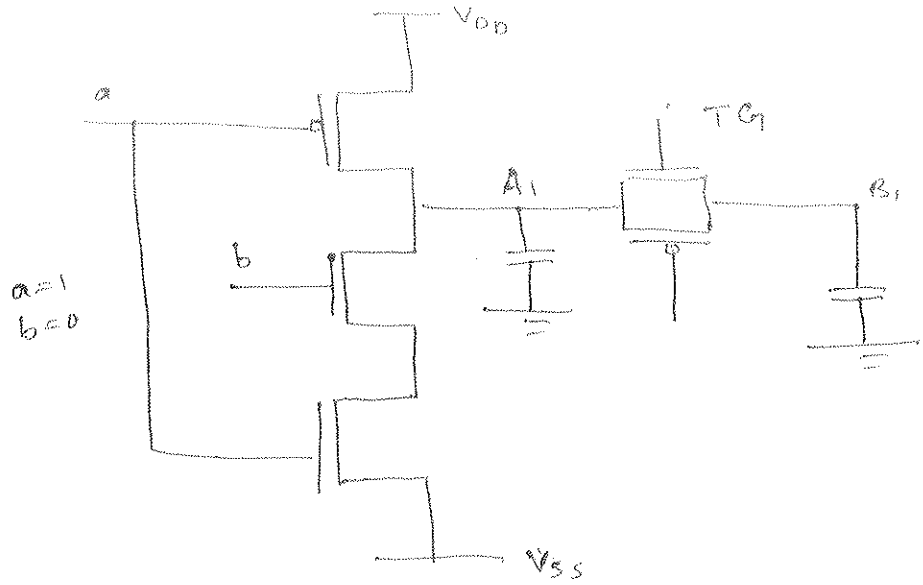
* The time needed for leakage to disturb a dynamic node is

$$t = \frac{C_{node} \cdot \Delta V}{I_L}$$

I_L -> Leakage current.

(c) Charge sharing.

* If dynamic gates drive pass transistors then Charge sharing is occured.



A_1, B_1 → nodes

* Transmission gate is OFF, $B_1 = 0$
* If TG is ON, charge is shared between A_1 and B_1 it affects the dynamic output.

(d) Ratio failures:

* It can occur if a node is simultaneously pulled up and down.
* Ratio failure can occur in the circuit which has feedback.

(e) power supply noise :-

* It occurs due to IR drop occur across the resistance R
* It affects the performance and degrade the noise margin

(f) Hotspots.

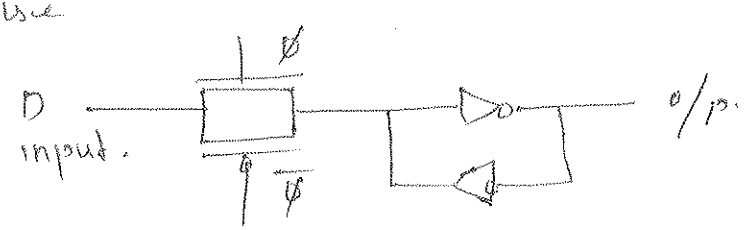
* The performance of the transistor can be affected by hot spots. It is due to non uniform power dissipation.

(g) Minority carrier injection.

* The junctions between drain and body may be forward biased so the current is flowing into the substrate. It is known as minority carrier injection.

⑤ Diffusion input noise sensitivity

* Here exposed diffusion input is applied & is sensitive to noise



Power dissipation

* Static cmos gates are very efficient because they dissipate nearly zero power while idle.

① Instantaneous power:-

* It is drawn from the power supply is proportional to the supply current $i_{DD}(t)$ and supply voltage V_{DD}.

$$P(t) = i_{DD}(t) \cdot V_{DD}$$

② Energy consumed over some interval T is the integral of the instantaneous power.

$$E = \int_0^T i_{DD}(t) \cdot V_{DD} \cdot dt$$

③ Average power:-

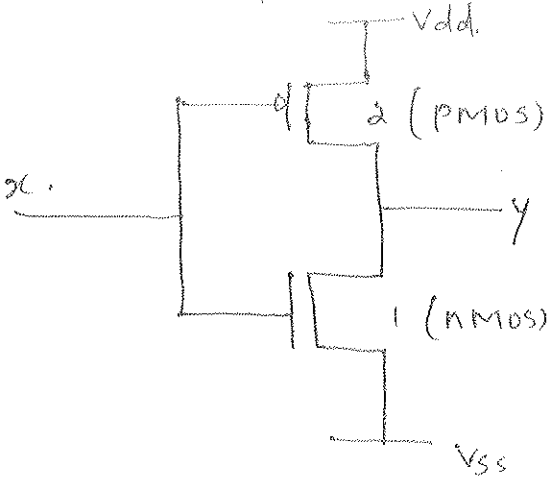
$$P_{avg} = \frac{E}{T} = \frac{1}{T} \int_0^T i_{DD}(t) \cdot V_{DD} \cdot dt.$$

Static power dissipation

It is due to

* Subthreshold conduction through off transistors
* Tunneling current through gate oxide
* Leakage through reverse biased

* Contention current in ratioed circuit



* If $x = 0$, NMOS is OFF and PMOS is ON
* If $x = 1$, NMOS is ON and PMOS is OFF.
* In these 2 cases, at any time one transistor is OFF

Ideally no current flows through the OFF transistor.

* But due to subthreshold conduction through OFF transistors and tunneling current through gate oxide, leakage through reverse biased diodes and contention current in ratioed circuit.

* Static power dissipation

$$P_s = I_s \cdot V_{DD}$$

Where,

$P_s \longrightarrow$ Static power dissipation

$I_s \longrightarrow$ Static current.

Dynamic dissipation

* The primary dynamic dissipation component is charging the load capacitance. Suppose a load is switched between GND and V_{DD} at an average frequency of f_{sw}.
* Over any given interval T, the load will be charged and discharged $T f_{sw}$ times.
* Total charge transferred from V_{DD} to ground

$$\boxed{Q = C V_{DD}}$$

* Average dynamic dissipation is

$$P_{dynamic} = \frac{1}{T} \int_0^T i_{DD}(t)\, V_{DD}\, dt.$$

$$= \frac{V_{DD}}{T} \int_0^T i_{DD}(t)\, dt.$$

$$= \frac{V_{DD}}{T} \int_0^T \frac{Q}{t_{sw}} \cdot dt \qquad \left[\because i_{DD}(t) = \frac{Q}{t_{sw}}\right]$$

$$= \frac{V_{DD}}{T} \cdot Q f_{sw} \int_0^T dt$$

$$= \frac{V_{DD}}{T} \cdot f_{sw} Q \cdot [t]_0^T$$

$$= \frac{V_{DD}}{T} \cdot f_{sw} Q \cdot T$$

$$P_{dynamic} = V_{DD} \cdot f_{sw} \cdot C V_{DD}.$$

$$\boxed{P_{dynamic} = C V_{DD}^2 f_{sw}}$$

* dynamic dissipation interms of activity factor α

$$P_{dynamic} = \alpha C V_{DD}^2 f_{sw}$$

$\alpha = 1$, for clock

$\alpha = 0.5$, for most data

$\alpha = 0.1$, for static CMOS logic.

Low power design

* Total power = static power + dynamic power.

Dynamic dissipation is far greater than static dissipation when systems are active

(a) Dynamic power reduction

* To reduce the dynamic power dissipation decreases.
  1. Activity factor.
  2. Switching capacitance.
  3. power supply
  4. operating frequency.

(1) Activity factor.

* static logic has low activity factor.
* dynamic logic have clocked nodes and high internal activity factor.
* clock gating can be used to stop portions of the chip that are idle
* Eg: If integer codes are executed then floating point until is OFF.

② switching capacitance

* Reduce the capacitance by choosing small transistor
* use minimum sized gates.
* Reduce the interconnect switching capacitance.

③ power supply

* Voltage has a quadratic effect on dynamic power
* Choosing a low power supply reduces the power consumption.
* As many transistors operate in velocity saturation mode, voltage will not reduce performance as much.
* Voltage can be adjusted based on operating mode.

④ Operating frequency

* frequency can also be traded for power
* two multipliers running at half speed may be replaced by a single multiplier running at full speed.
* This reduces the area and power consumption.

Common metrices for low power design

① power
② power delay product.
③ Energy delay product.

Static power reduction

* Static power involves minimizing I static
* Analog current sources and pseudo nMos gates draw static power.

* Subthreshold current for $V_{gs} < V_t$ a

$$I_{ds} = I_{ds}\, e^{(V_{gs} - V_t)/nV_T}\left(1 - e^{-V_{ds}/V_T}\right)$$

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

η -> drain induced barrier lowering

γ -> Body effect.

* The term in brackets is 1 for any V_{ds} values. To reduce the other terms.

(a) Increase V_{t0}, V_{sb}

(b) Reduce V_{gs}, V_{ds}.

* Subthreshold leakage power is a major problem

(1) To control leakage voltage apply a body voltage using body effect.

Bias	Devices	Uses
RBB - Reverse body bias	Low V_t devices	to reduce the leakage
FBB - Forward body bias	high V_t devices	improve the performance

* Too much RBB leads to greater junction leakage through a mechanism called "band to band tunneling"

* Too much FBB leads to substantial current through the substrate to source diodes

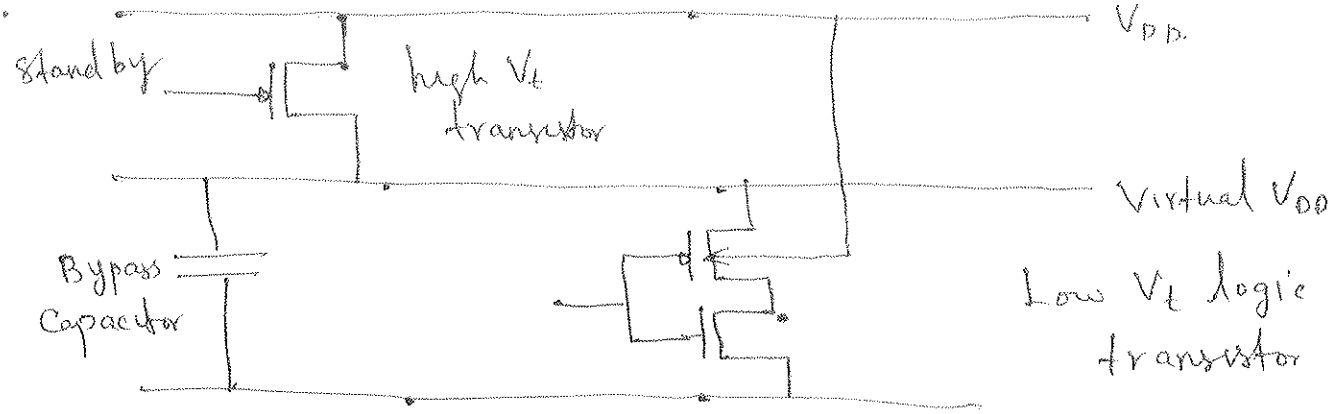
(2) To reduce the leakage raise the source voltage in sleep mode.

(3) Reduce V_{DD} in stand by mode.

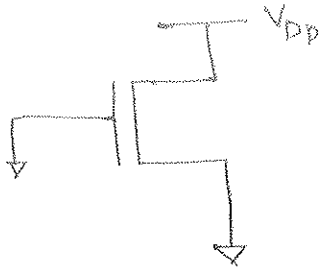
(4) Turn off the power supply entirely

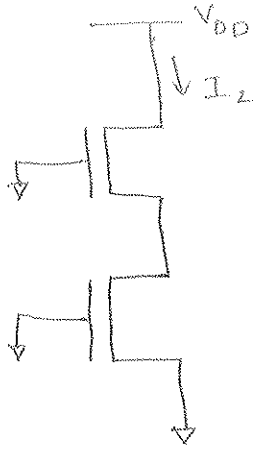
MTCMOS (Multi-Threshold CMOS).

* Uses low V_t transistors for computation
* Uses high V_t transistor for switching



* High V_t device is connected between true V_{DD} and virtual V_{DD}. Thus increases the power supply noise and delay
* By pass capacitance stabilizes the supply but the capacitor discharges each time V_{DD} is disconnected.
* pMOS body should be tied to V_{DD} and nMOS body should be tied to ground.
* Leakage through two series OFF transistor is much lower than that of a single transistor because of stack effect.





$I_1 > I_2$

(Stack effect)

↓

When two or more stacked transistors turned together, Leakage power is suppressed.

UNIT III Sequential circuit design

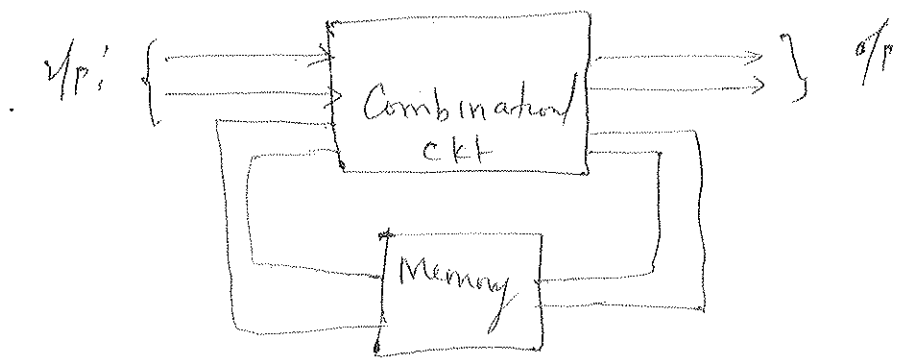
Static latches and Registers, Dynamic latches and registers, pulse registers, Sense amplifier based register, pipelining, Schmitt trigger, Monostable Sequential circuits, Astable Sequential circuits

Timing Issues: Timing classification of digital system, synchronous design.

---

* In Combinational logic circuits, the o/p a a function of the current inputs, whereas in Sequential logic circuits the output depends upon both the current inputs as well as previous inputs.

* It requires memory to store the previous input values called states. In this circuit the output is given as feedback to the input. So these circuits are called regenerative circuits



Finite state machine (FSM) that consists of combinational logic and registers, which stores the system state information. Here all registers are under the control of a single global clock. The outputs of the FSM are a function of the current inputs and the current state.

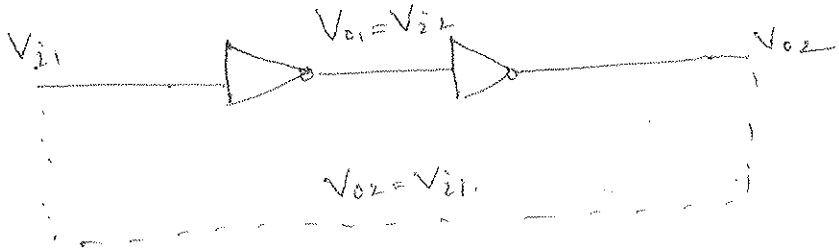
The next state is determined based on the current state and the current inputs and it is fed in to the inputs of the registers.

Static latches and Registers

Bistability principle

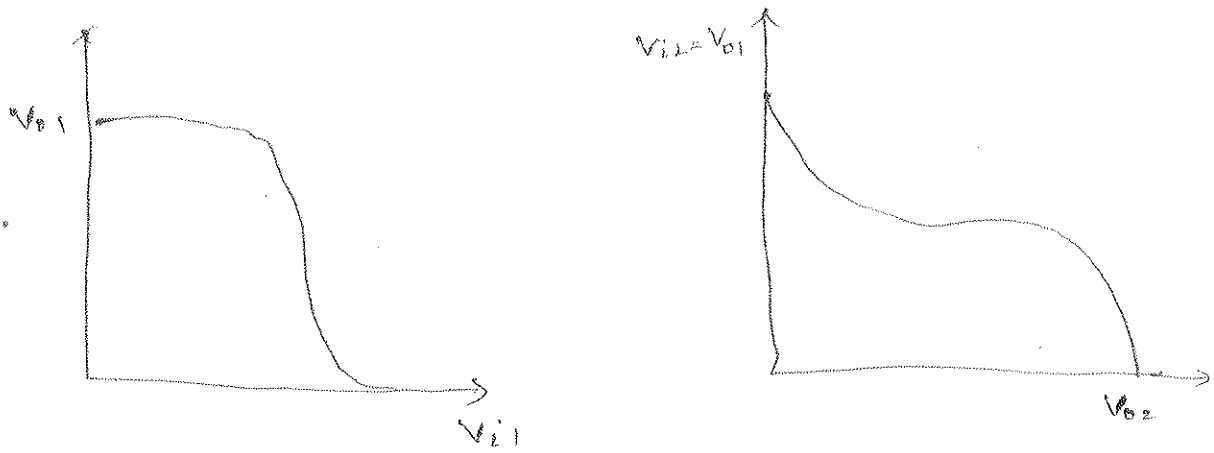
Static latches use positive feedback so that bistable circuits can be formed. It has two stable states that represent 0 and 1.

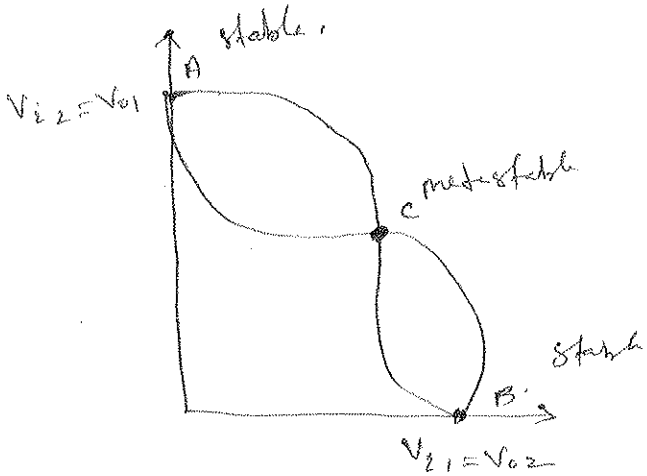
Two inverters are connected in cascade to form a basic bistable circuit as shown below.



Voltage - transfer characteristics

The VTC of the first inverter is V_{o1} versus V_{i1} is shown in figure (a) and the second inverter is V_{o2} versus V_{i2} is given in fig b. This is plotted by considering that $V_{i2} = V_{o1}$





* Assume that the output of the second inverter V_{o2} is connected to the input of the first inverter V_{i1}
* the resulting VTC of two cascade inverter shows three possible operating points.
* To change its operating point a sufficiently large external voltage is required to make voltage gain of

inverter loop greater than unity.

Suppose this cross-coupled inverter pair is biased at point C. A small deviation from this bias point is caused by noise. It is amplified and regenerated around the circuit loop. As a results the gain around the loop being larger than 1.

* Now, A and B are the only stable operation points and C is a metastable operating point. Metastable means every deviation causes the operation point to run away from its original point.

* The cross coupling of two inverters results in a bistable circuit that is a circuit with two stable states, each corresponding to a logic state. The circuit serves as a memory, storing either a 1 (or) a 0.

Dynamic latches and Registers.

The stored value remains valid as the supply voltage is applied to the circuit in static latches. The main drawbacks are

(i) its complexity

(ii) When registers are used in computational structures that are constantly clocked.

(iii) If the memory requires holding state for extended period of time then static latches cannot be used.

* Dynamic gates are used to decrease complexity, increase speed and low power dissipation.

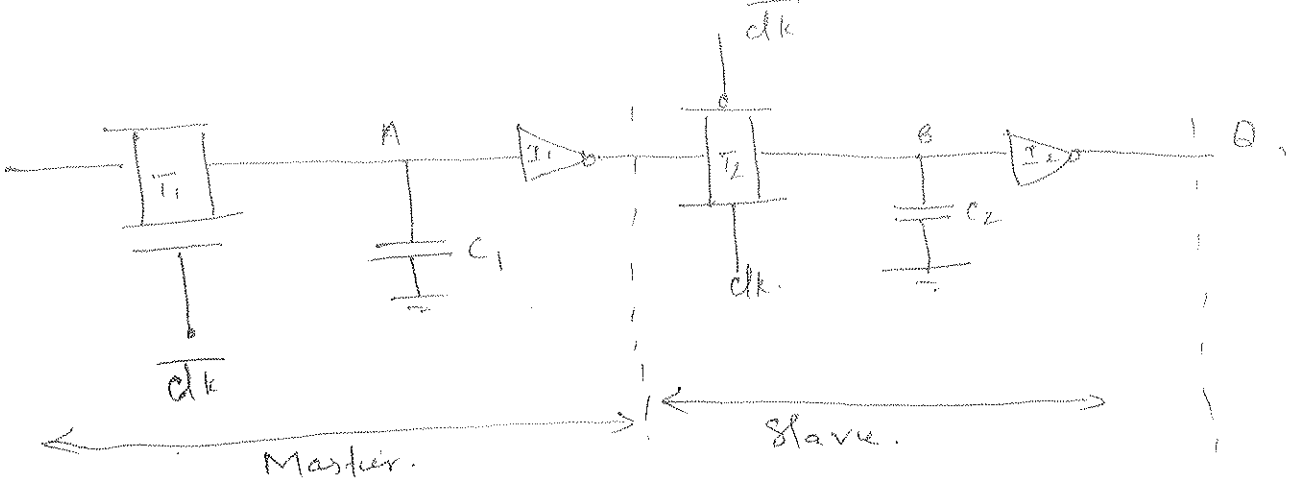
* In dynamic latches the charge is stored on parasitic capacitors temporarily. If the charge is present it represents '1' and if charge is absent, then it represent 0.

* Due to charge leakage in capacitor, the charge is stored for a limited amount of time. The capacitor should be periodically refreshed to obtain signal integrity. So it is called a dynamic storage. There are two types of dynamic registers.

(i) Dynamic Transmission-gate edge-triggered register
(ii) Clocked CMOS (C^2MOS) register.

Dynamic Transmission-gates edge-Triggered register

* A dynamic transmission gate positive edge-triggered register based on the master slave concept. In figure T_1 and T_2 are transmission gates and I_1 and I_2 are inverters, C_1 and C_2 are equivalent capacitance at node-1 and node-2.



clk = 0.

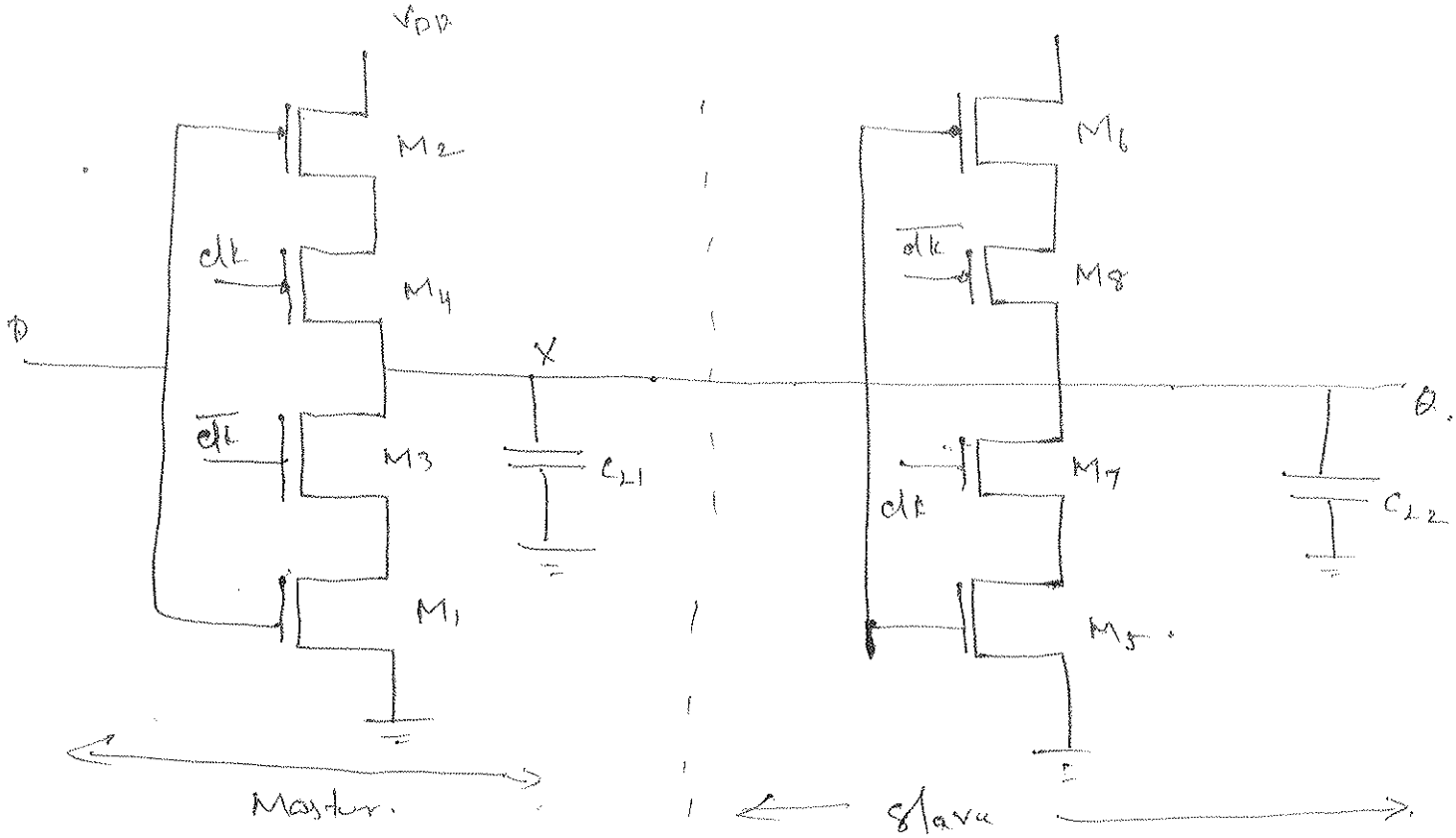
* Transmission gate T_1 is ON and T_2 is OFF.
* Input data D is sampled on storage node 1, which has the equivalent capacitance C_1, consisting the combination of gate capacitance of I_1, junction capacitance of T_1 and the overlap gate capacitance of T_1
* During the period the slave is in a hold mode and node 2 is in high impedance (floating) state.

clk = 1

* On the rising edge of clock, transmission gate T_2 is ON and T_1 is OFF. Hence the value sampled on node 1 at clk = 0 is propagated to the output Q.
* Thus node - 2 stores the inverted value of node-1

clocked CMOS (C^2MOS) Register

* clocked CMOS (C^2MOS) is used to reduce power dissipation and layout size and to increase speed. Following figure shows a positive edge triggered register based on the master-slave concept which is insensitive to clock overlap. This circuit is also called C^2MOS register



clk = 0 ($\overline{clk}$ = 1)

* Master stage is ON and acts as an inverter & samples the inverted input D on the internal node x. This is called evaluation mode, whereas the slave is a hold mode (or) high impedance mode.

* In slave stage, M_7 and M_8 transistors are OFF, decoupling the output from the input. Thus output Q retains its previous value stored on the output capacitor C_{L2}.

clk = 1

* The transistors M_3 and M_4 are OFF and the master is in the hold mode. The transistor M_7 and M_8 are ON and the slave is said to be in evaluation mode.

* The value stored on capacitor C_{L1} is transmitted to the output node through the slave stage which acts as an inverter. The output on Q is actually an inverted form of input

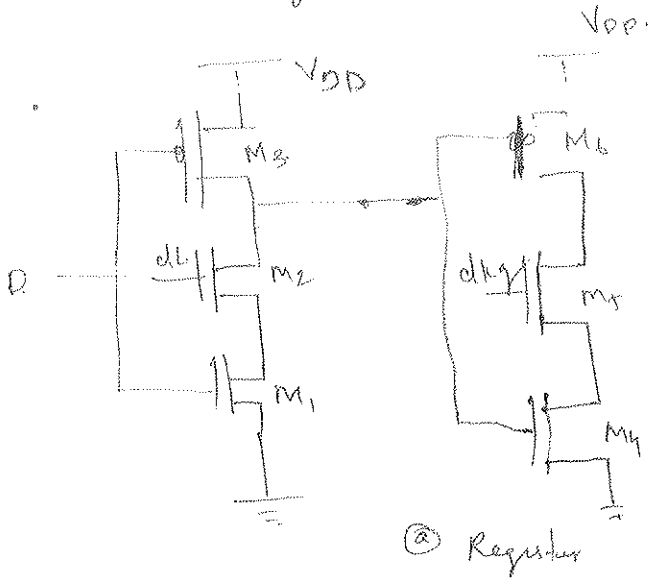
* A C^2MOS register with clk & $\overline{clk}$ is insensitive to overlap as long as the rise and fall times of the clock edges are sufficiently small.

pulse register

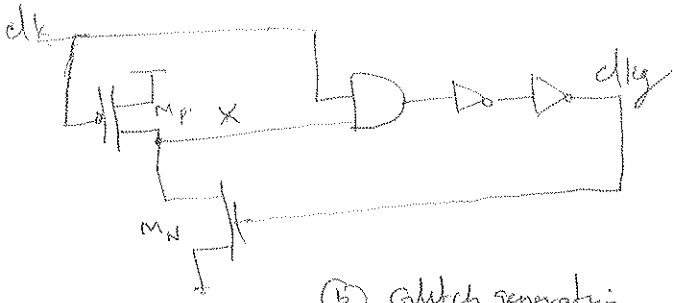
The idea is to construct a short pulse around the rising (or falling) edge of the clock. This pulse acts as the clock input to a latch, sampling the input only in a short window. Race conditions are thus avoided by keeping the opening time (ie the transparent period) of the latch very short.

The combination of the glitch generation circuitry and the latch results in a positive edge-triggered register.

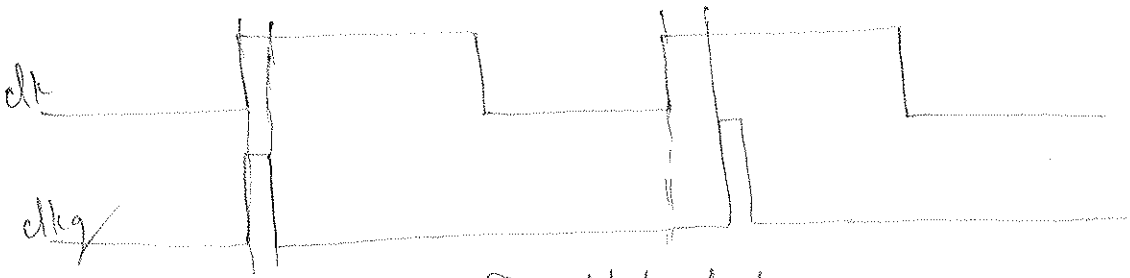
Fig shows an example circuit for constructing a short intentional glitch on each rising edge of the clock. When clk=0, node X is charged up to V_{DD} (M_N is off since clkg is low).



(a) Register



(b) Glitch generation



(c) Glitch clock.

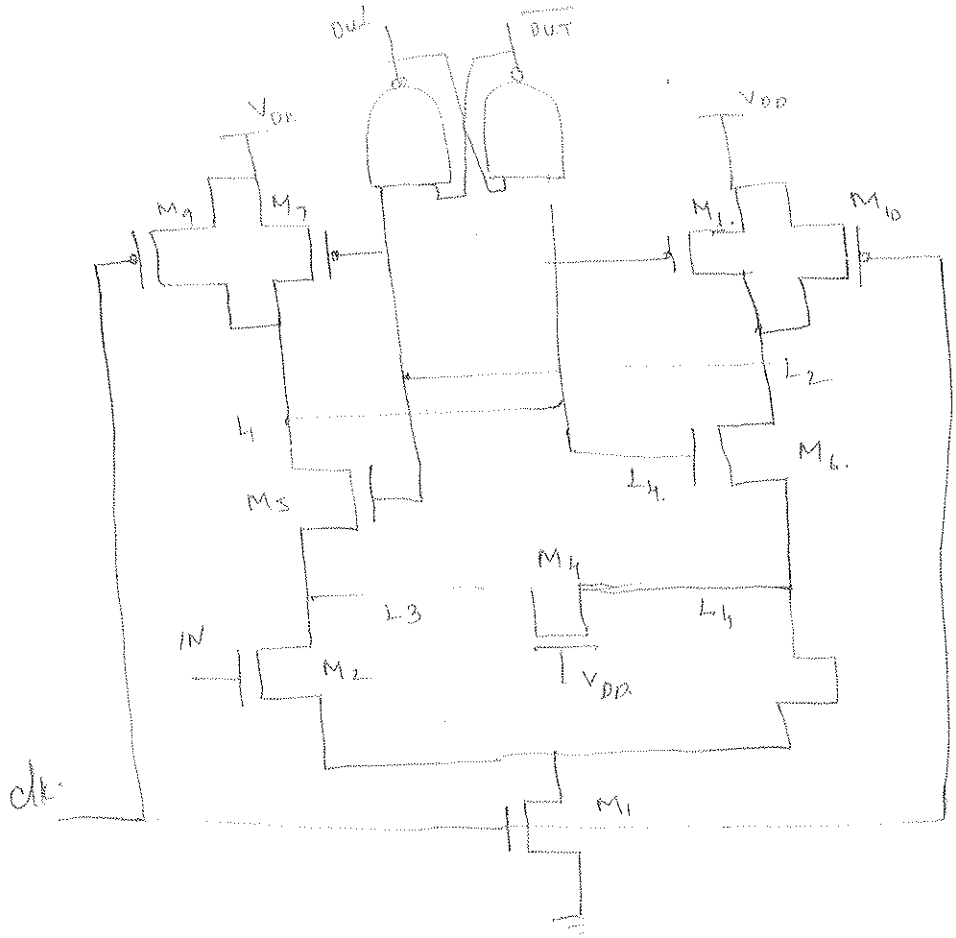
On the rising edge of the clock there is a short period of time when both inputs of the AND gate are high, causing clkg to go high. This in turn activates M_N, pulling X and eventually clkg low. The length of the

pulse is controlled by the delay of the AND gate and the two inverters. Note that there exists also a delay between the rising edges of the input clock (clk) and the glitch clock (clkg) also equal to the delay of the AND gate and the two inverters.

If setup time and hold time are measured in reference to the rising edge of the glitch clock, the setup time is essentially zero, the hold time is equal to the length of the pulse, and the propagation delay equals two gate delays. The advantage of the approach is the reduced clock load and the small number of transistors required. The glitch-generation circuitry can be amortized over multiple register bits. The disadvantage is a substantial increase in verification complexity.

Sense-Amplifier-based Registers.

Sense-amplifier circuits accept small input signals and amplify them to generate rail-to-rail swings. They are used extensively in memory cores and in low swing bus drivers to either improve performance or reduce power dissipation. There are many techniques to construct these amplifiers.



A common approach is to use feedback - for instance, through a set of cross-coupled inverters. The circuit uses a precharged front-end amplifier that samples the differential input signal on the rising edge of the clock signal. The outputs of front end are fed into a NAND cross-coupled SR flipflop that holds the data and guarantees that the differential outputs switch only once per clock cycle. The differential inputs in this implementation don't have to have rail-to-rail swing.

The shorting transistor M_4 is used to provide a DC-leakage path from either node L_3 or L_4 to ground.

This is necessary to accommodate the case in which the inputs change their value after the positive edge of clk has occurred, resulting in either L_3 or L_4 being left in a high-impedance state with a logical low voltage level stored on the node. Without the leakage path, that node would be susceptible to charging by leakage currents. The latch could then actually change state prior to the next ~~rising~~ rising edge of CLK.

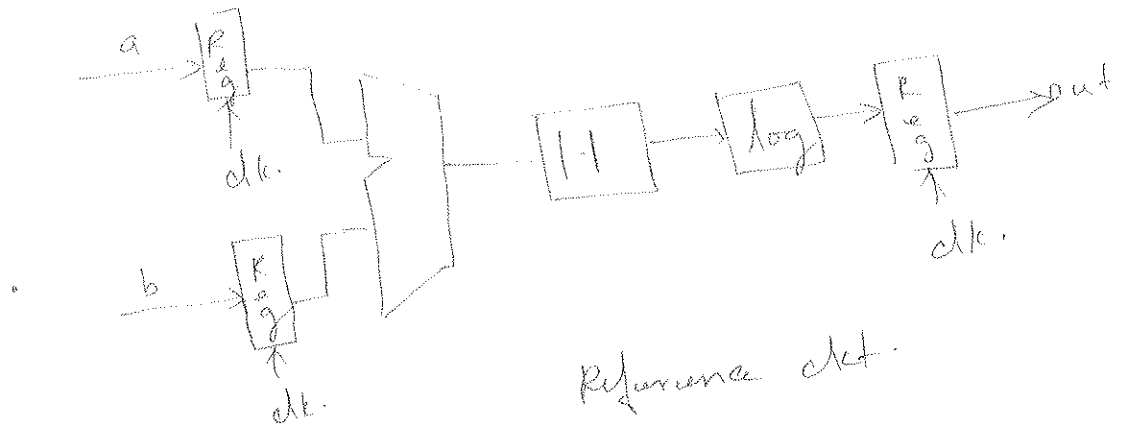
Pipelining

Pipelining is a popular design technique often used to accelerate the operation of datapaths in digital processors. The goal of the presented circuit is to compute $\log(|a+b|)$ where both a and b represent streams of numbers. The minimal clock period T_{min} is necessary to ensure correct evaluation is given as

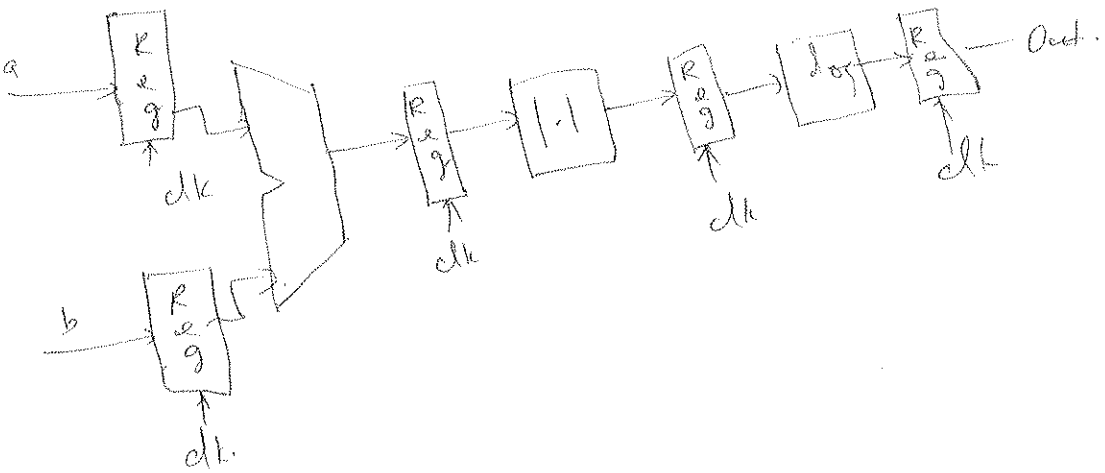
$$T_{min} = t_{c-q} + t_{pd,logic} + t_{su}.$$

Where t_{c-q} and t_{su} are the propagation delay and the setup time of the register. We assume that the

registers are edged-triggered D registers. The term t_{pd} logic stands for the worst case delay path through the combinational network which consists of the adder, absolute value and logarithm functions.



Reference ckt.



Pipeline computation

clk 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	$a_5 + b_5$	$\lvert a_4 + b_4 \rvert$	$\log(\lvert a_3 + b_3 \rvert)$.

In conventional systems, the latches delay is generally much larger than the delays associated with the registers and dominates the circuit performance. Assume that each logic module has an equal propagation delay. Pipelining is a technique to improve the resource utilization and increase the functional throughput.

The advantage of pipelined operation becomes apparent when examining the minimum clock period of the modified ckt. The combinational circuit block has been partitioned into three sections, each of which has a smaller propagation delay than the original function. This effectively reduces the value of the minimum allowable clock period.

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

The pipelined network outperforms the original circuit by a factor of three under these assumptions.

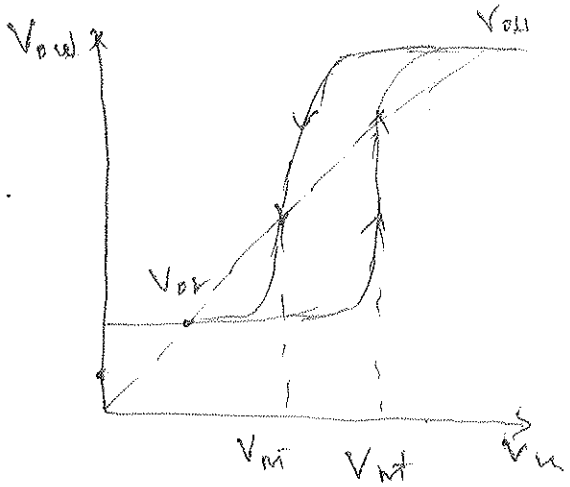
i.e. $(T_{min,pipe} = T_{min}/3)$.

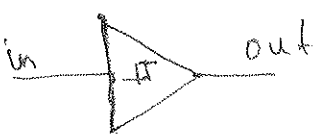
Schmitt trigger

A Schmitt trigger is a device with two important properties:

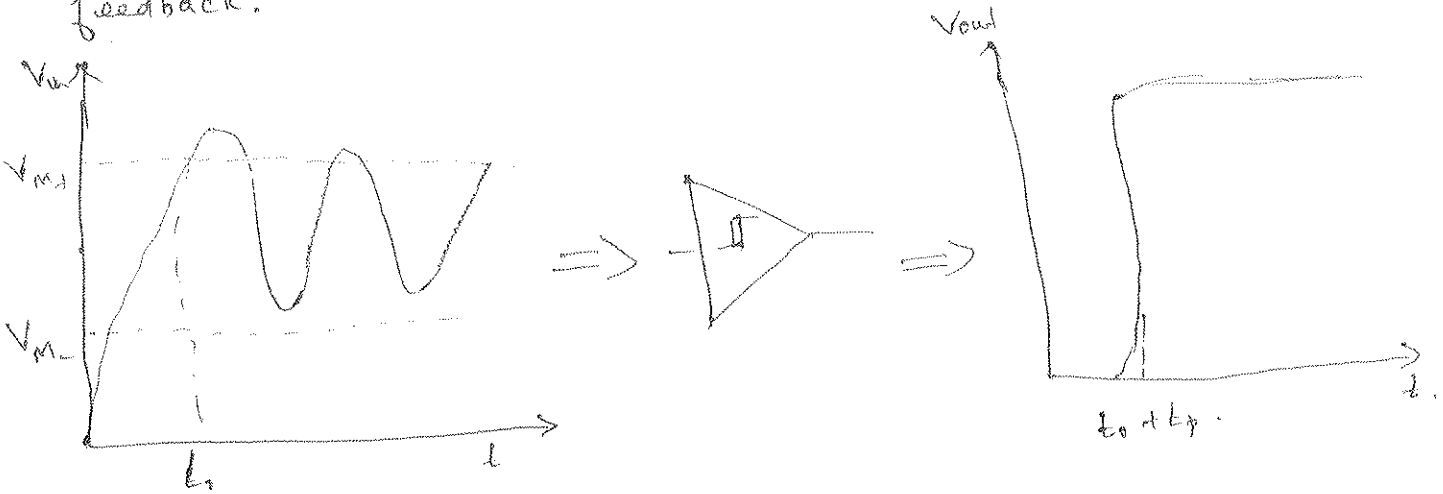
1. It responds to a slowly changing input waveform with a fast transition time at the output.
2. The voltage-transfer characteristic of the device displays different switching thresholds for positive- and negative-going input signals. The switching thresholds for the low-to-high and high-to-low transitions are called V_{M+} and V_{M-}, respectively. The hysteresis voltage is defined as the difference between the two.

One of the main use of the Schmitt trigger is to turn a noisy or slowly varying input signal into a clean digital output signal. This is shown in fig. Notice how the hysteresis suppresses the ringing on the signal.



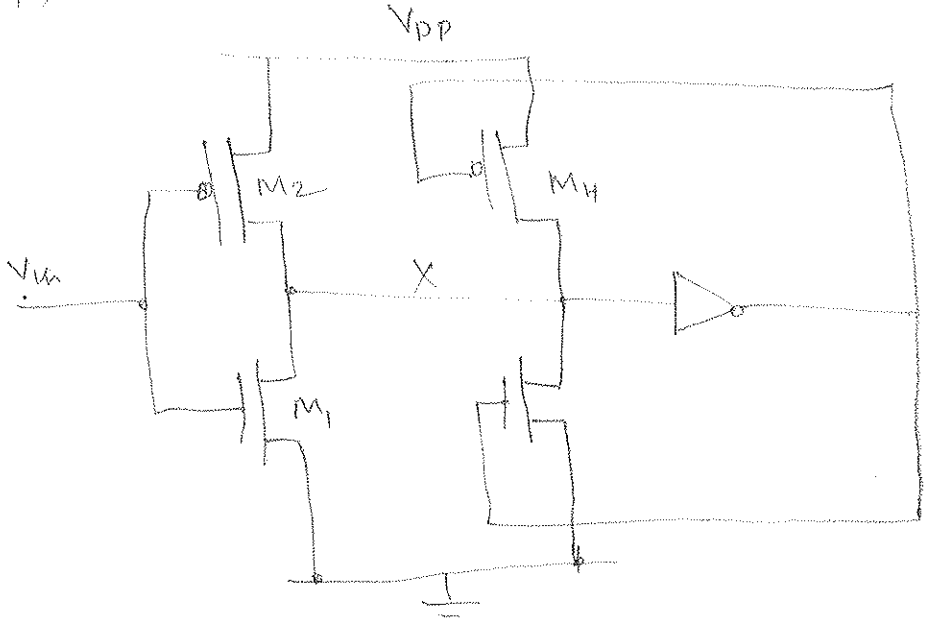


At the same time, the fast low-to-high transitions of the output signal should be observed. Steep signal slopes are beneficial in general for instance for reducing power consumption by suppressing direct-path currents. The "secret" behind the schmitt trigger concept is the use of positive feedback.



CMOS implementation

The idea behind this circuit is that the switching threshold of a CMOS inverter is determined by the (k_n/k_p) ratio between the PMOS and NMOS transistors.



Increasing the ratio raises the threshold, while decreasing it lowers V_M. Adapting the ratio depending upon the direction of the transition results in the switch threshold and a hysteresis effect. This adaptation is achieved with the aid of feedback.

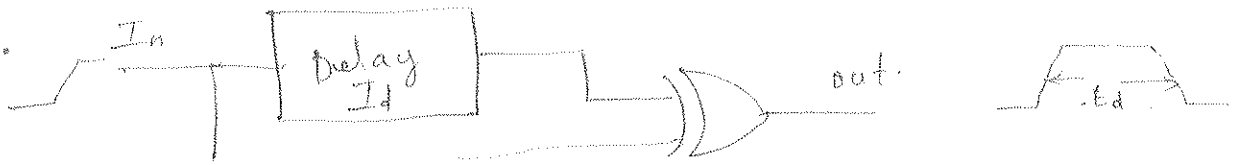
Once the inverter switches the feedback loop turns off M_4 and the NMOS device M_3 is activated. This extra pull down device speeds up the transition and produces a clean output signal with steep slopes.

Monostable Sequential circuits

A monostable element is a circuit that generates a pulse of a predetermined width every time the quiescent circuit is triggered by a pulse or transition event. It is called monostable because it has only one stable state. A trigger event, which is either a signal transition or a pulse causes the circuit to go temporarily into another quasi-stable state. This means that it eventually returns to its original state after a time period determined by the circuit parameters. This circuit also called one shot is useful in generating pulses of a known length.

This functionality is required in a wide range of applications.

The most common approach to the implementation of one shots is the use of a simple delay element to control the duration of the pulse. This concept is shown in the below figure.

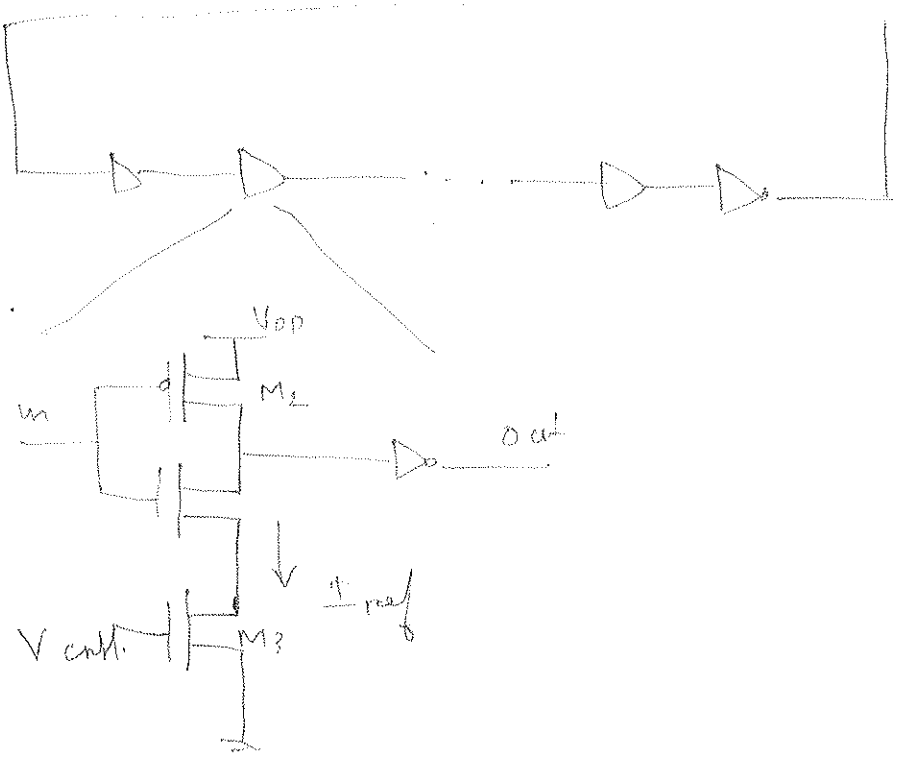


In the quiescent state both inputs to the XOR are identical and the output is low. A transition on the input causes the XOR inputs to differ temporarily and the output to go high. After a delay Td this disruption is removed, and the output goes low again. A pulse of length Td is created. The delay circuit can be realized in many different ways such as an RC network or a chain of basic gates.

Astable circuits

An astable circuit has no stable states. The output oscillates back and forth between two quasi stable states with a period determined by the circuit topology and parameters. One of the main applications of oscillators is the on chip generation of clock signals.

The ring oscillator is a simple example of an astable circuit. It consists of an odd number of inverters connected in a circular chain. Due to the odd number of inversion, no stable operation point exists and the circuit oscillates with a period equal to $2 \times t_p \times N$, where N is the number of inverters in the chain and t_p is the propagation delay of each inverter.



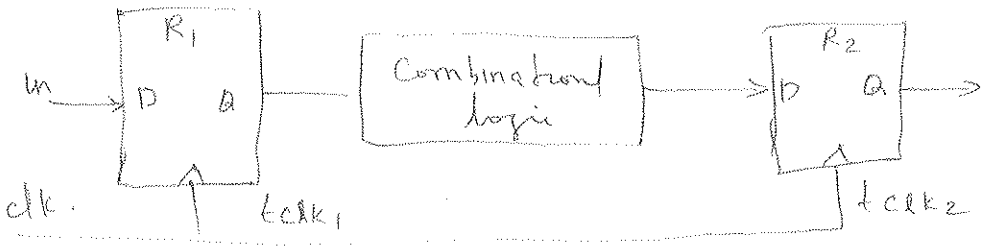
The ring oscillator composed of cascade inverters produces a waveform with a fixed oscillating frequency determined by the delay of an inverter in the CMOS process. In may application it is necessary to control the frequency of the oscillator. An example of such a circuit is the voltage controlled oscillator (VCO) whose oscillation frequency is a function of a control voltage. The standard ring oscillator can be modified in to a VCO by replacing the standard inverter with a current-starved inverter. The mechanism for controlling the delay of each inverter is to limit the current available to discharge the load capacitance of the gate.

Timing issues : Synchronous design

All the systems virtually designed today use a periodic synchronization signal (or) clock. The generation and distribution of a clock has a significant impact on the performance and power dissipation of the system.

The fig shows the basic structure of a synchronous pipeline datapath. In an ideal case the clock

at registers 1 and 2 have the same period and transition at the exact time.



t_{c-q}
$t_{c-q,cd}$
t_{su} t_{hold}

t_{logic}
$t_{logic,cd}$.

The contamination (or) minimum delay ($t_{c-q,cd}$) and the maximum propagation delay of the register (t_{c-q}).

(a) The setup (t_{su}) and hold times (t_{hold}) for the registers.

(b) The contamination delay ($t_{logic,cd}$) and the maximum delay (t_{logic}) of the combinational logic

(c) The positions of the rising edges of the clocks clk, and clk_2 (t_{clk_1} and t_{clk_2} respectively) relative to a global reference.

clock skew

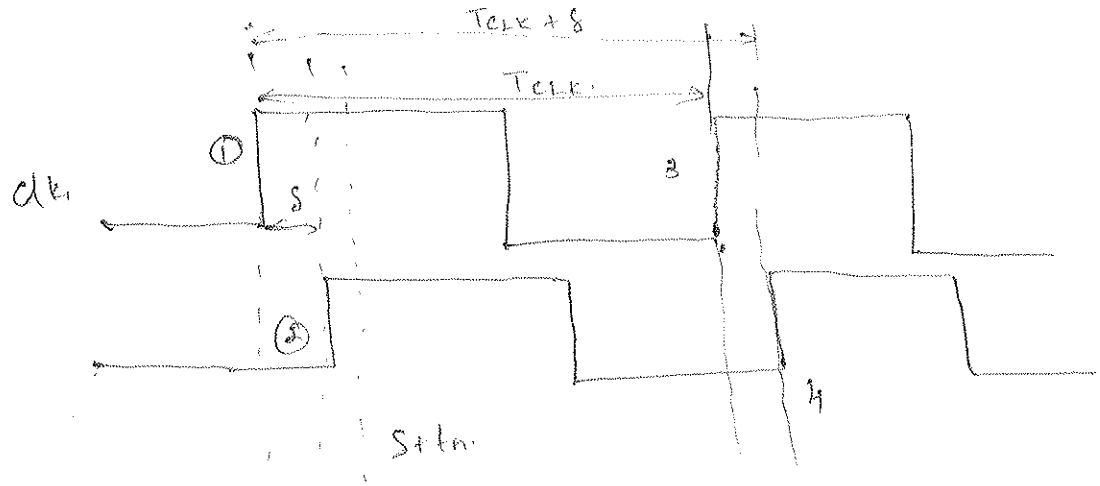
The spatial variation in arrival time of a clock transition on an integrated circuit is commonly referred as clock skew.

The clock skew between two points i and j on an IC expressed as

$$\delta(i,j) = t_i - t_j$$

Where t_i and t_j are the positions of the rising edge of the clock with respect to the reference.

Consider the data transfer between R_1 and R_2 shown in fig. The clock skew can be positive (or) negative depending upon the routing direction and position of the clock source.



* Fig shows that the rising clock edge is delayed by a positive δ at the second register
* The clock-skew is caused by static mismatches in the clock paths and differences in the clock load.
* The clock-skew is caused by static mismatches in the clock paths and differences in the clock load.
* The clock-skew has strong implications for both the performance and the functionality of sequential systems.

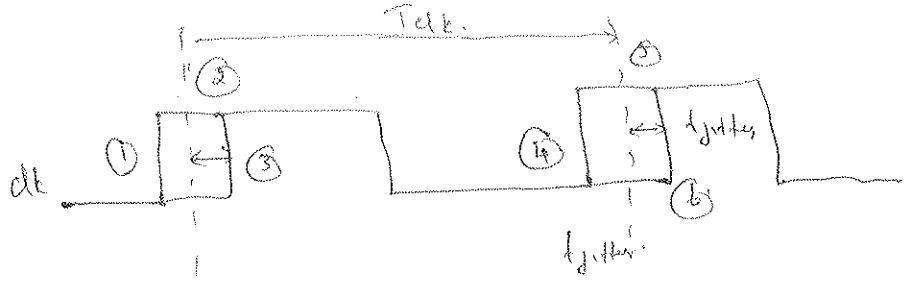
clock jitter

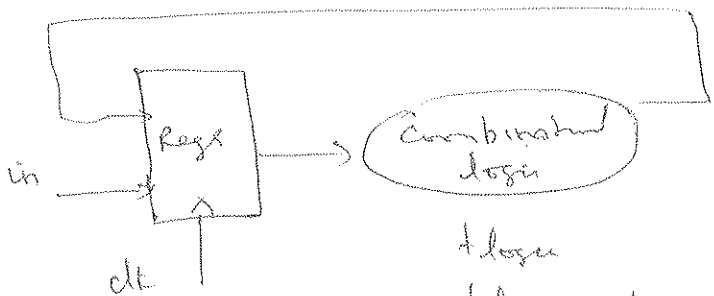
* clock jitter refers to the temporal variation of the clock period at a given point on the chip that is the clock period can be reduced (or) expand on a cycle-by-cycle basis.
* It is strictly a temporal uncertainty measure and it is specified at a given point. Jitter can be measured and characterized in a number of ways and it is a zero mean random variable.
* The absolute jitter (t_{jitter}) refers to the worst case variation (absolute value) a clock edge at a given location with respect to an ideally periodic reference clock edge.
* The cycle-to-cycle jitter (T_{jitter}) typically refers to the time-varying deviation of a single clock period relative to an ideal reference clock. For a given spatial location 'i' and it is expressed as

$$T^{i}_{jitter}(n) = t^{i}_{clk,n+1} - t^{i}_{clk,n} - T_{clk}$$

Where, $t^{i}_{clk,n+1}$ and $t^{i}_{clk,n}$ represents the arrival time of $n+1^{th}$ and n^{th} clock edges at node i respectively and T_{clk} is the nominal clock period.

under the worst case condition the magnitude of the cycle-to-cycle jitter equals twice the absolute jitter ($2t_{jitter}$).





t_{c-q}, $t_{c-q,cd}$
t_{su}, t_{hold}
t_{jitter}.

t_{logic}
$t_{logic,cd}$.

Ideally the clock period start at edge ① and ends at edge ⑤ with a nominal clock period of Tclk. In the worst case scenario, the leading edge of the current clock is delayed by jitter (edge ③), while jitter causes the leading edge of the nex clock period to occur early (edge ⑪).

* As a result, the total time available to complete the operation is reduced by $2t_{jitter}$ in the worst case and it is given as.

$$T_{clk} - 2t_{jitter} \geq t_{c-q} + t_{logic} + t_{su} \text{ (or)}$$

$$T_{clk} \geq t_{c-q} + t_{logic} + t_{su} + 2t_{jitter} \quad \text{——} (11)$$

Equation (A) represents that jitter directly reduces the performance of a sequential circuit.

Unit IV

Design of Arithmetic building blocks and subsystems

Arithmetic Building blocks: Data paths, Adders, Multipliers shifters, ALUs, power and speed trade offs, case study: Design as a trade off. Designing memory and array structure: Memory architectures and building blocks, memory core, memory peripheral circuitry.

Arithmetic building blocks:

Data paths.

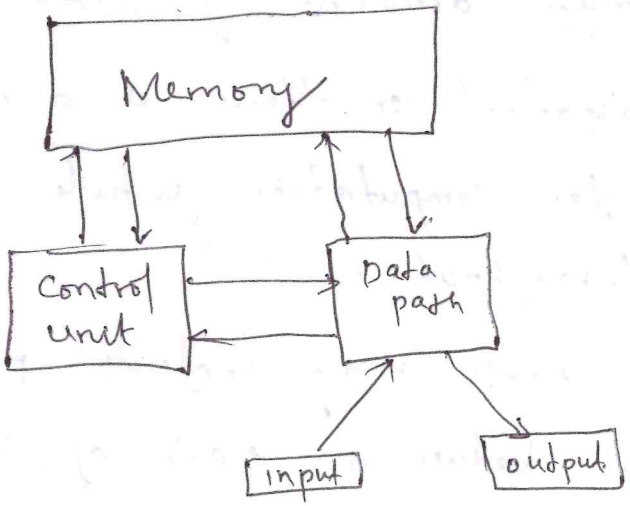


Figure shows a digital processor architectures and its components consists of the datapath, memory, control and input/output blocks.

* The data path is the core of the processor that is all computations are performed. The other blocks in the processor are support units that either store the results

produced by the datapath (or) help to determine what will happen in the next cycle.

* A typical datapath consists of an interconnection of basic combinational elements may be shifters, adders multipliers such as arithmetic operators which can performs addition, multiplication comperison and shift (or) logic functions that is AND, OR and XOR

Main constraints on the datapath design

* The processing speed is the main constraint in personal computers. In most of the applications there is a maximum amount of power that is allowed to be dissipated or there is a maximum energy available for computation while maintaining the desried throughput.

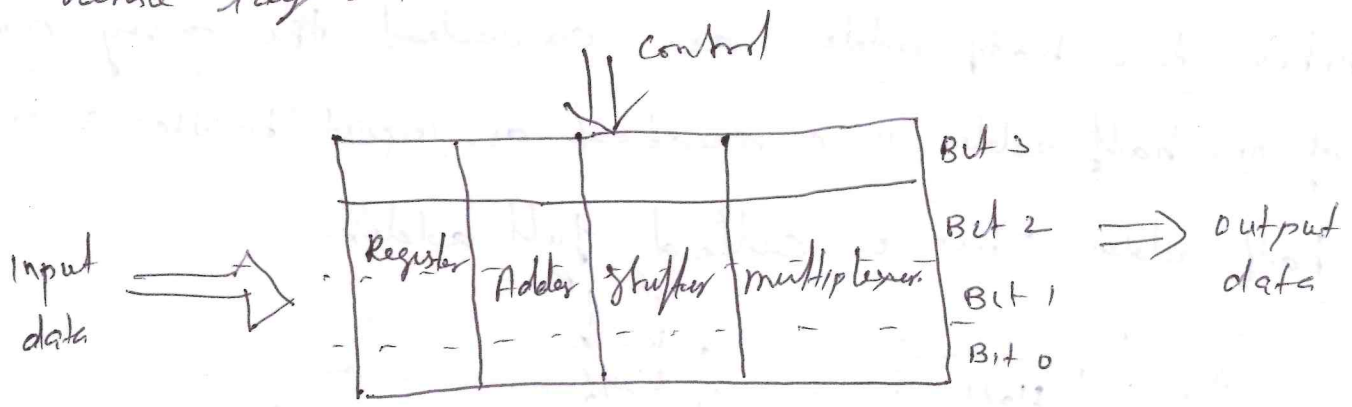
* A data path may include point-to-point connections between pair of components. Data pass between components on one (or) more buses. The numbers of buses determine the maximum number of data transfers on a clock cycle.

Bit-sliced data path organization

A bit-slice is a one-bit version of a complete datapath which is shown in figure ①.

A bit-slice is replicated to design an n-bit data path. The data flows horizontally through the bit-slice along point-to-point connections (or) buses.

* A datapath is best implemented in a bit-sliced fashion. A single layout is used repetitively for every bit in the data world. This regular approach eases the design effort and it results in fast and dense lay out.



Consider a 32-bit processor which operates on data words with 32 bits wide. The datapath consists of 32-bit slices and each slice operates on a single bit. Hence it is called as bit sliced.

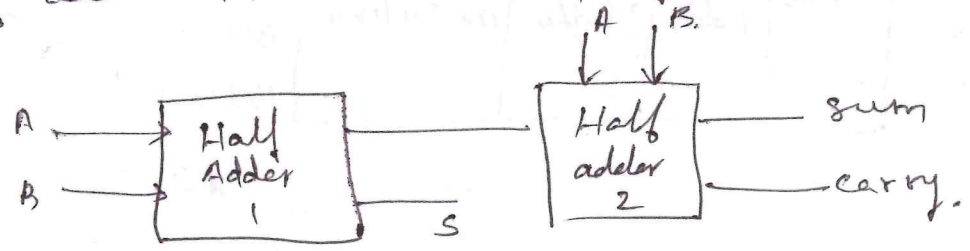
* Bit slices are either identical (or) resemble a similar structure for all bits. The datapath designer can concentrate on the design of a single slice that is repeated 32 times.

Adders.

Addition is one of the basic operation in data processing. It is used in every stage, starting from counting to multiplication and filtering. Adders can be implemented in various forms which suits different speed and density requirements.

Single - bit binary adder.

When two half adders are cascaded the carry output of one half adder is connected as input to the second half adder. This is called full adder.



The logical expression for sum and carry

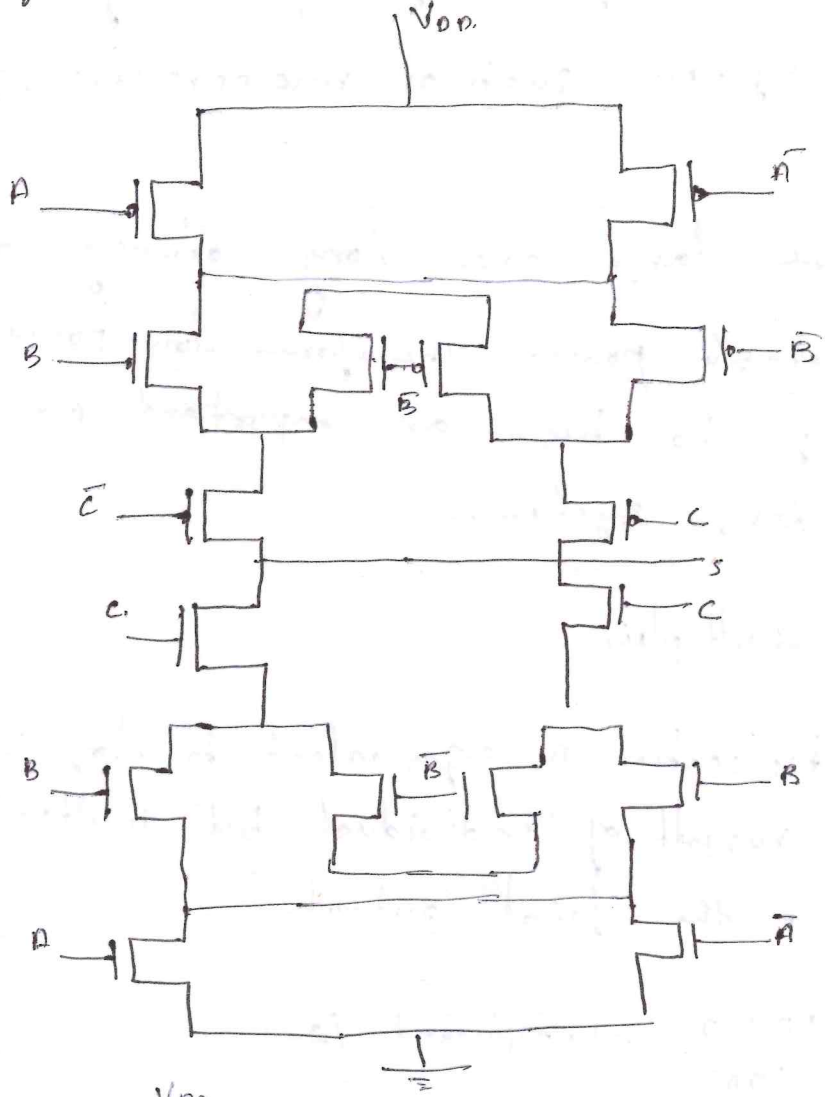
$$S = A \oplus B \oplus C$$

$$\text{Carry} = AB + BC + CA.$$

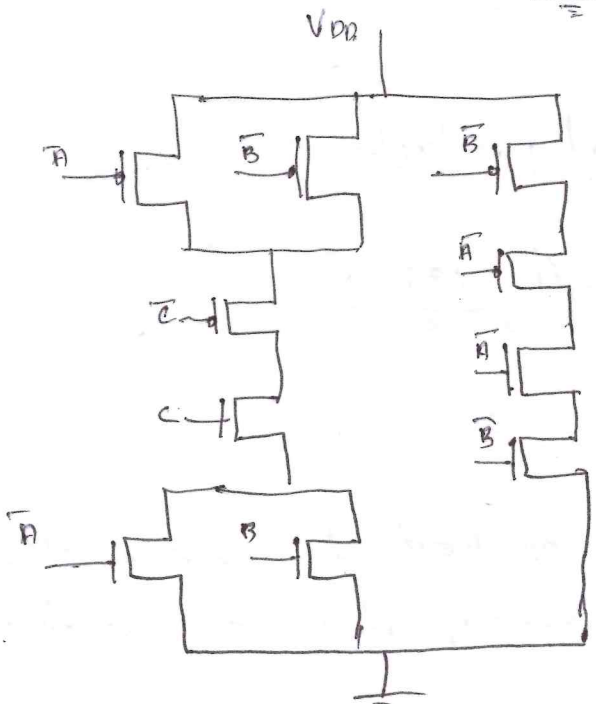
The carry gate is also called majority gate.

CMOS implementation of full adder

* An implementation of 1-bit adder for 3-input XOR gate using transistor is shown below.



Sum



Cerry.

Multipliers

* Multipliers plays a major role in computation process. Multiplication are expensive and also slow operations. It is widely used in digital signal processing and in high performance systems such as microprocessor, graphic engine etc.
* Multipliers have large area, long latency and consumes considerable power. Therefore low-power multipliers design has been an important one in low power VLSI design system.

Definitions of multiplier

Binary multiplication is equivalent to logical AND operation. The result of individual bit multiplication is added to get the final output.

```
    10110   Multiplicand (22)
     1001
   -------
    10110
   00000
  00000
 10110
-----------
 11000110      Result 198.
```

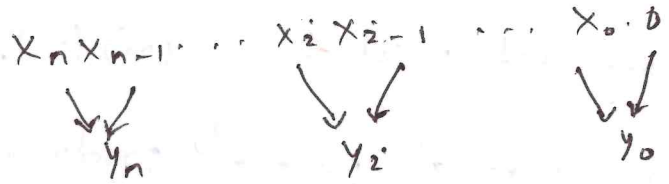
N ↕ partial products

Booth's multiplication

Booth encoding is a method to reduce the number of partial product. Simply it is a multiplication

algorithm that multiplies two signed binary numbers in two's complement notation

* For booth - n (number of bits)
  * Examines n+1 bits of the multiplier
  * Encode n bits



$$y_i = x_{i-1} - x_i$$

x_i	$x_i - 1$	operation	y_i
0	0	Shift only	0
0	1	shift only	0
1	0	subtract shift	-1
1	1	Addition shift	1

Booth Encoding

Assume for example an eight-bit multiplier of the form 01111110 which produces six non zero partial-product rows. To reduce the number of non zero rows by recoding this number into a different format. Here we are using booth encoding table as follows.

Given 0 1 1 1 1 1 1 0

1 0 0 -1

-1 (or) $\bar{1}$

Add '0' at 0 1 1 1 1 1 1 0 0

0 0 0 0

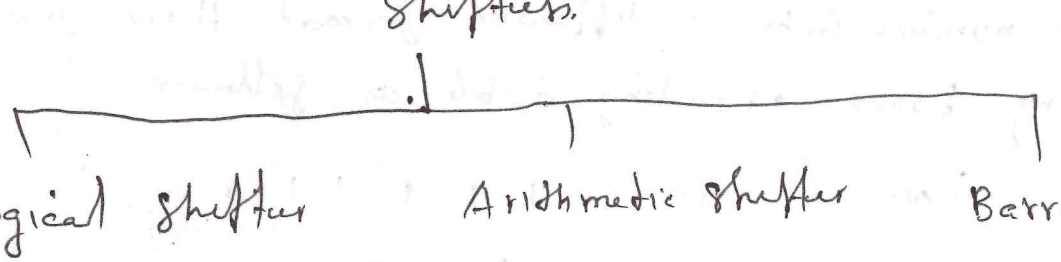
1 - - - - - $\bar{1}$ -

using this format we have to add only two partial products but the final adder has to be able to perform subtraction as well. This type of transformation is also called booth's recording

* It reduces the number of partial products almost one half. It ensures that for every two consecutive bit at most one bit will be 1 (or) (-1).
* Reducing the number of partial products is equivalent to reducing the number of additions which leads to a speedup as well as an area reduction.

Shifters

Shifters are used to shift the numbers from one bit position to the other. Shifts can neither be performed by a constant (or) variable amount. Constant shifts are trivial in hardware requiring only wires. The various types of shifters are as follows



Logical shifter

It is used to shift the number to the left (or) right and empty spots are filled with 0's.

Example : 1011 : LSL1 = 0110 & LSR1 = 0101

Arithmetic shifter:

Similar to logical shifter but sign bit is inserted in MSB

Example : 1011 ; ASR1 = 1101 & ASL1 = 0110

Barrel shifter

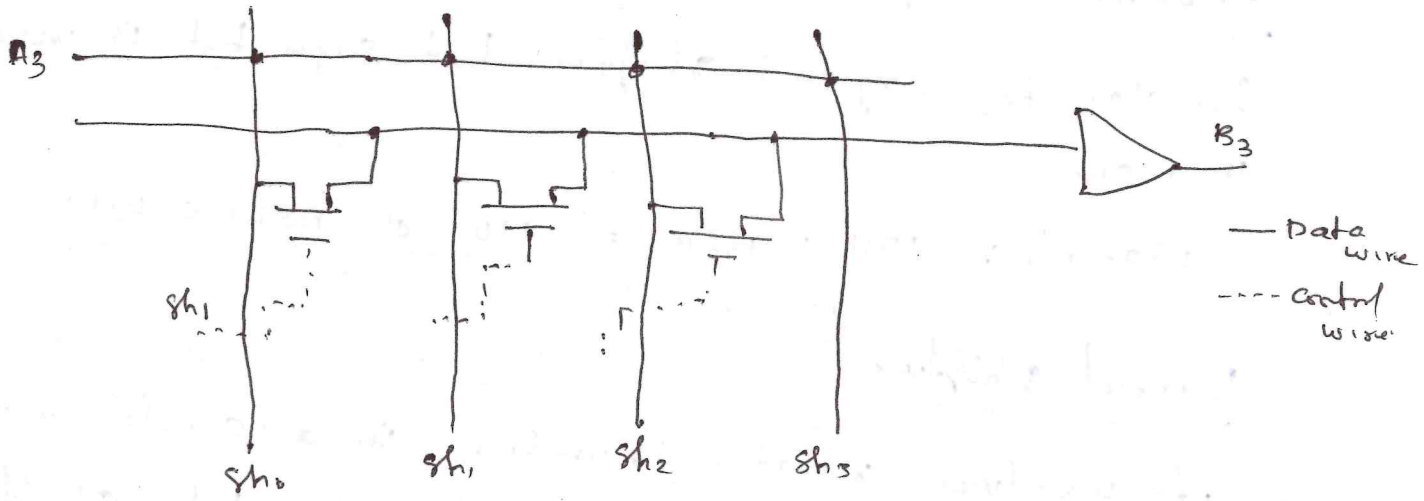
It involves rotates numbers in a circle such that empty spots are filled with bits shifted at other end.

Example : 1011 ; ROR1 = 1101 & ROL1 = 0111.

Barrel shifter.

The structure of the barrel shift is shown in the figure ⊙. It consists of an array transistors, in which the number of rows and equals the word length of the data and the number of columns equals the maximum shift width. In this case, both are set to four. The control wiries are routed diagonally through the array.

Figure shows the barrel shifter with a programmable shift width from zero to three bits to the right. The structure supports automatic repetition of the sign bit (A_3) also called sign-bit extension.



* A major advantage of this shifter is that the signal has to pass through at most one transmission gate. The capacitance at the input of the buffers rise linearly with the maximum shift width.
* Barrel shifter needs a control wire for every shift bit. For example a four-bit shifter needs four control signals To shift over three bits the signals $sh_3 : sh_0$ take on the value 1000 where only one of the signals is high.
* A barrel shifter performs a right rotate operation it can also handle left rotation using complementary shift operation.

By using suitable masking hardware, barrel shifters can also perform shifting. Barrel shifters are in two forms as

① Array form

② Logarithmic form

* Logarithmic barrel shifter is widely used and they are better suitable for large shifts.

Speed, Area and power trade off

* Ideally standard cell design would be used for higher volume applications. Different levels of design skills are required in 'back end' design.

* The reduced level of verification is required for sending to the factory and thus low cost is prepared. However front end design is virtually identical for each implementation style.

* The important parameters which impact the successful design of a VLSI chip are

Area - Size of the die, which relates to cost and profit.

Speed - How fast the transistor can switch

power - How much energy does it takes to do the work.

The metrics produce the chip attributes such as frequen. Which is inverse of the clock cycle and performance.

Speed Metric

* The critical paths are identified and the speed can be measured in ns, ps etc. Noise, power delivery, cross-die transistor variations can cause timing variations.
* The speed of the performance increases when the size of the chip decreases and also it consumes very less power.
* But the drawback is the error percentage it will increase due to the reduction in the size of the chip.

Power Metric

The standard power is measured which may affects the performance.

There are two types of power

(I) Static power: Dc current that does not depend on signal activity

(II) Dynamic power: Ac current proportional to signal transitions.

The main issues associated with power are

Power delivery : Ability to deliver the voltage and current needed to run the chip

Power Extraction : Ability to remove the heat generated by the chip.

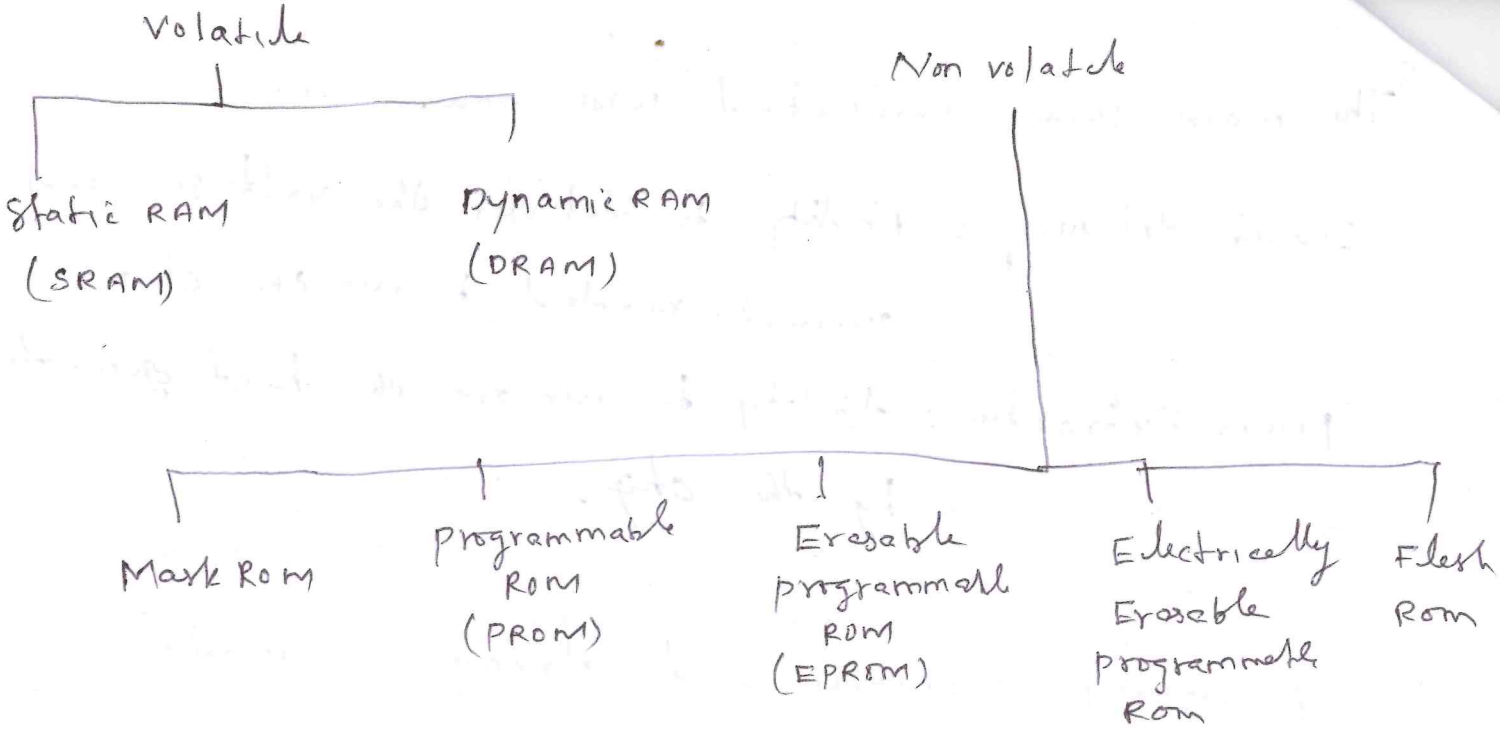
Designing memory and Array structure

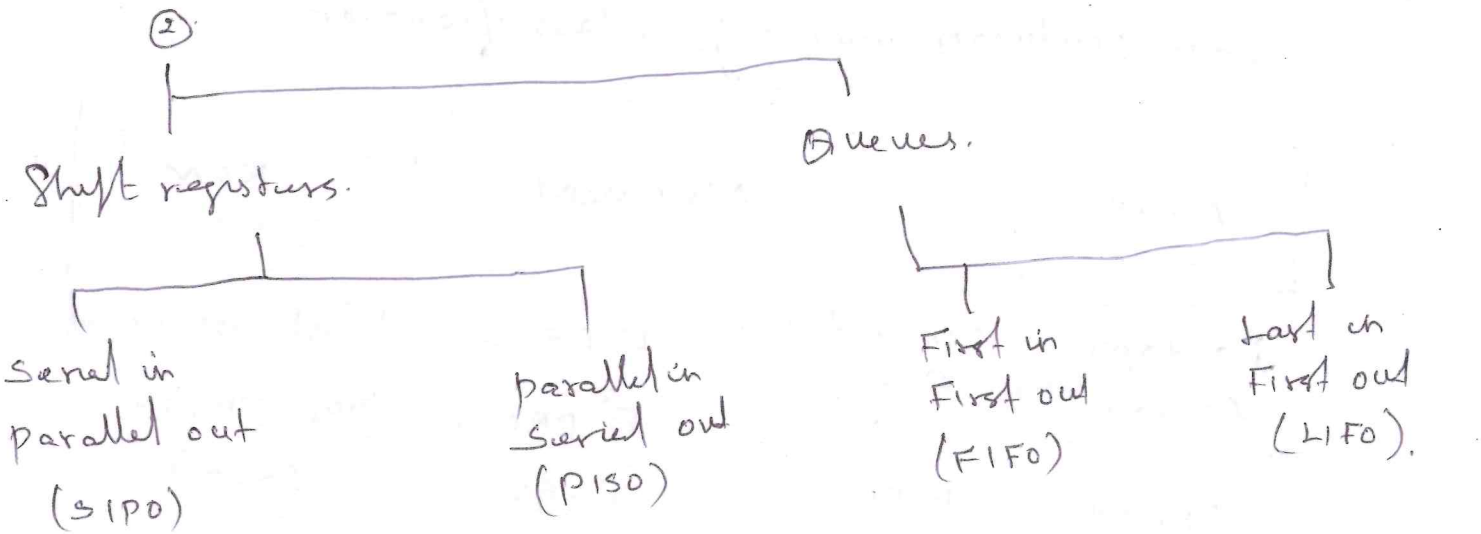
Semiconductor memory classification

RWM		NVRWM	ROM
Random Access	Non-random access	EPROM E^2PROM FLASH	Mask-programmed programmable (PROM)
SRAM DRAM	FIFO LIFO Shift register CAM		

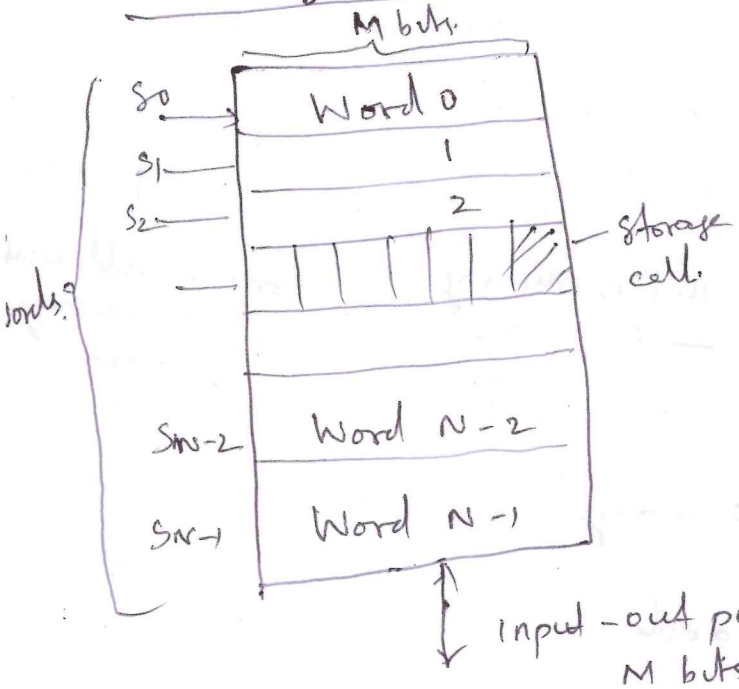
Memory Array

- RAM
  - R/W Memory (RAM)
  - Read only memory (ROM) (Non volatile
- Serial Access Memory — ②
- Content addressable memory (CAM)



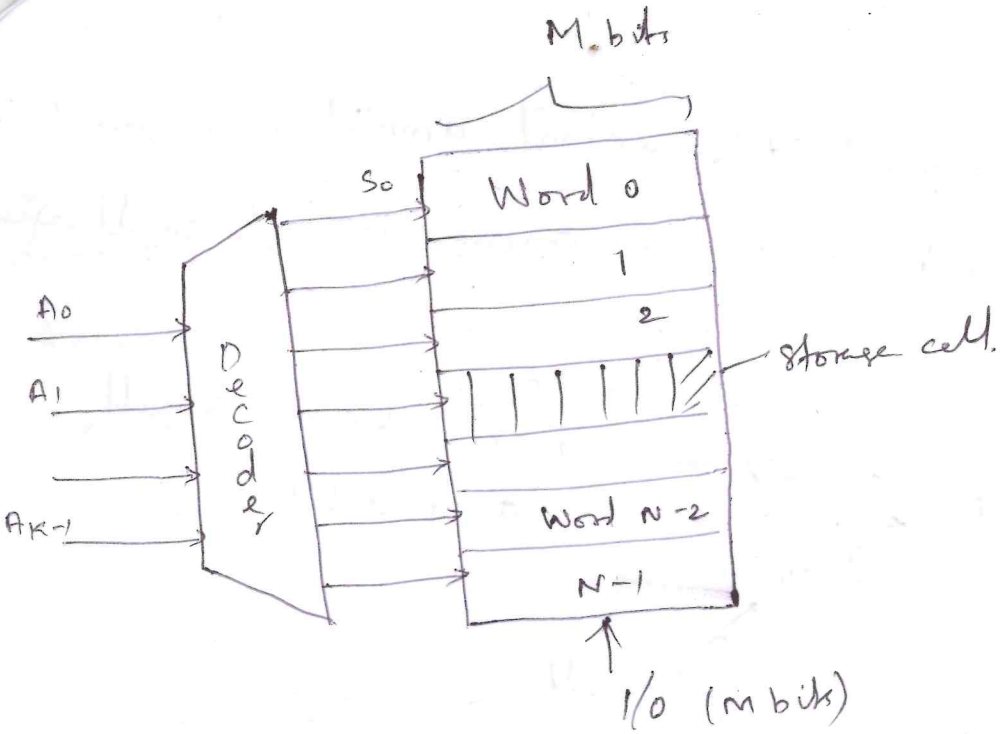


Memory Architecture : Decoders.



N words ⇒ N select signals

Too many select signals



Decoder reduces # of select signals

$K = \log_2 N$.

Memory peripheral circuitry

The important periphery circuits are

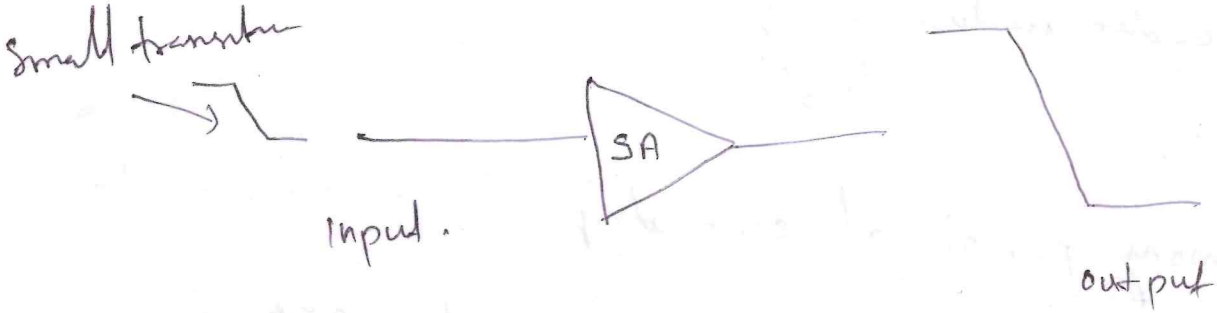
1. Decoders
2. Sense amplifiers
3. input/output buffers.
4. control/timing circuit.

Sense amplifiers are used to detect a signal difference on data lines such as current-mirror differential, semi-latch and full latch types.

The cross coupled differential amplifier is used to achieve speed, low power consumption and small area penalty

$t_p = C \cdot \Delta V$ / I_{av} ← make ΔV as small as possible

large ↑ small ↓



The figure shows voltage sense amplifiers The data due to invalid signal difference can be corrected with speed penalty unlike the cmos latch type.

UNIT - V Implementation Strategies and Testing

FPGA Building block Architectures, FPGA interconnect Routing procedures. Design for testability: Ad Hoc Testing, Scan Design, BIST, IDDQ Testing, Design for Manufacturability, Boundary scan.

Field-programmable Gate Arrays (FPGAs)

FPGA's are prefabricated silicon devices that can be electrically programmed in the field to become almost any kind of digital circuit or system. For low to medium volume productions, FPGAs provide cheaper solution and faster time to market as compared to Application specific integrated circuits (ASIC) which normally requiries a lot of resources in terms of time and money to obtain first device.

FPGAs on the other hand take less than a minute to configure and they cost anywhere around a few hundred dollars to a few thousand dollars. Also for varying requirements, a portion of FPGA can be

partially reconfigured while the rest of an FPGA is still running. Any future updates in the final product can be easily upgraded by simply downloading a new application bitstream. However, the main advantage of FPGAs is flexibility is also the major cause of its drawback. Flexible nature of FPGAs makes them significantly larger, slower, and more power consuming than their ASIC counterparts. These disadvantages arise largely because of the programmable routing interconnect of FPGAs present a compelling alternative for digital system implementation due to their less time to market and low volume cost.

Normally FPGAs comprise of

* programmable logic blocks which implement logic functions.
* programmable routing that connects these logic functions.
* I/O blocks that are connected to logic blocks through routing interconnect and that make off-chip connections.

Figure Simplified FPGA floorplan

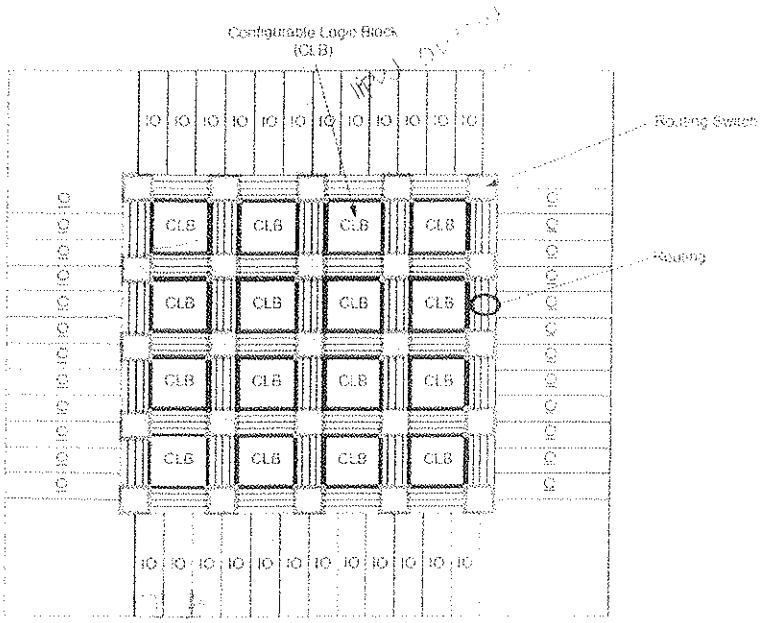
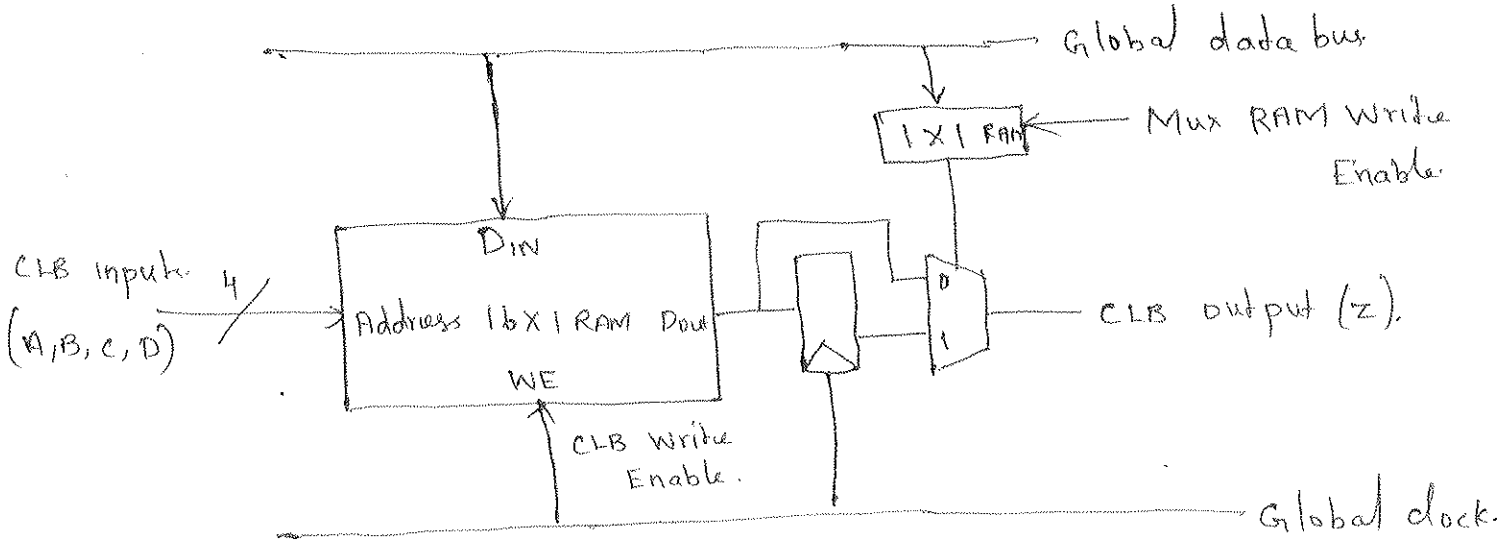


Figure shows the floorplan of a simplified FPGA. The chip is composed of an array of configurable logic blocks (CLBs). Metal routing tracks run vertically and horizontally between the array of CLBs. These terminate at the gray blocks, which are routing switches that can be implemented using antifuses, CMOS transmission gates or tristate buffers. The routing resources can also be connected to the inputs and outputs of the adjacent CLBs. CLBs use programmable lookup tables to compute any functions of several variables. Configurable I/O cells that can be used as input, output or bidirectional pads surround the core array of CLBs.

Two basic versions of FPGA exist. The first uses a special process option such as a fuse or antifuse to permanently program interconnect and personalize logic. These are one-time programmable. The second type uses static RAM or flash memory to configure routing and logic functions. In general, an FPGA chip consists of an array of logic cells surrounded by programmable routing resources. As an example of the first type of FPGA devices manufactured by Actel embed an array of logic modules within an interconnect matrix that is formed on the top metal layers.

A simple SRAM-based FPGA logic cell is shown below.



It is composed of a 16x1 static RAM as the logic element. This provides for any logic function of

four variables merely by loading the RAM with the appropriate content. A full adder can be implemented in two CLBs (one for carry and one for sum). The CLB shown also provides an optional output register. While it may seem inefficient or slow to use a RAM to perform logic, specially designed single-data line RAMs are small and fast in current processes and resources such as the routing tend to dominate modern design from a density and speed viewpoint.

FPGAs have matured to the stage where they boast millions of logic gate equivalents supported by megabits of RAM. I/Os can operate in excess of 10GHz. FPGAs frequently have embedded microprocessor cores and DSP accelerator hardware. Their low up-front cost and ease of correcting design errors makes them the best choice now for many low-to-medium-volume custom logic applications.

Configurable logic blocks (CLBs)

CLBs implement most of the logic in an FPGA. The principal CLB elements are shown in figure ⊙

Two 4-input function generators (F and G) offer unrestricted versatility. Most combinational logic functions need four or fewer inputs. However, a third function generator (H) is provided. The H function generator has three inputs. Either zero, one or two of these inputs can be the outputs of F and G; the other input(s) are from outside the CLB. This CLB can, therefore, implement certain functions of up to nine variables, like parity check or expandable - identity comparison of two sets of four inputs. Thirteen CLB inputs and four CLB outputs provide access to the function generators and storage elements. These inputs and outputs connect to the programmable interconnect resources outside the block.

A CLB can be used to implement any of the following functions

* any function of up to four variables plus any second function of up to four unrelated variables plus any third function of up to three unrelated variables.
* any single function of five variables
* any function of four variables together with some functions of six variables.
* some functions of up to nine variables.

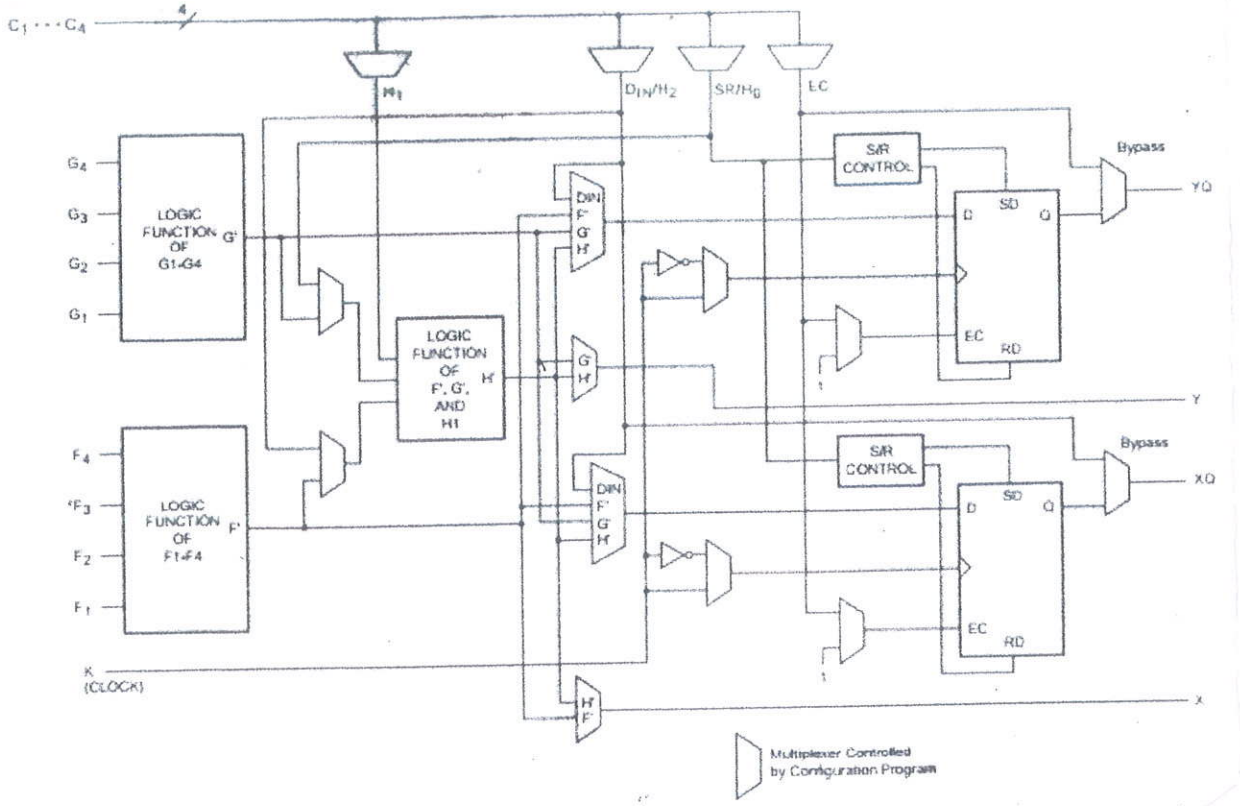


Figure Simplified Block Diagram of XC4000 Series CLB.

Input/output Blocks (IOBs).

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.

IOB input signals: Two paths, labeled I1 and I2 bring input signals into the array. Inputs also connect to an input register that can be programmed as either an edge-triggered flip-flop or a level sensitive latch.

Registered inputs: The I1 and I2 signals that

exit the block can each carry either the direct or registered input signal. The input and output storage elements in each IOB have a common clock enable input which through configuration can be activated individually for the input or output flip-flop or both. This clock enable operates exactly like EC pin on the XC4000 series CLB. It can not be inverted within the IOB.

IOB output signals : output signals can be optionally inverted within the IOB and can pass directly to the pad or be stored in an edge-triggered flip flop.

output slew Rate. The slew rate of each output buffer is by default reduced to minimize power bus transients when switching non-critical signals.

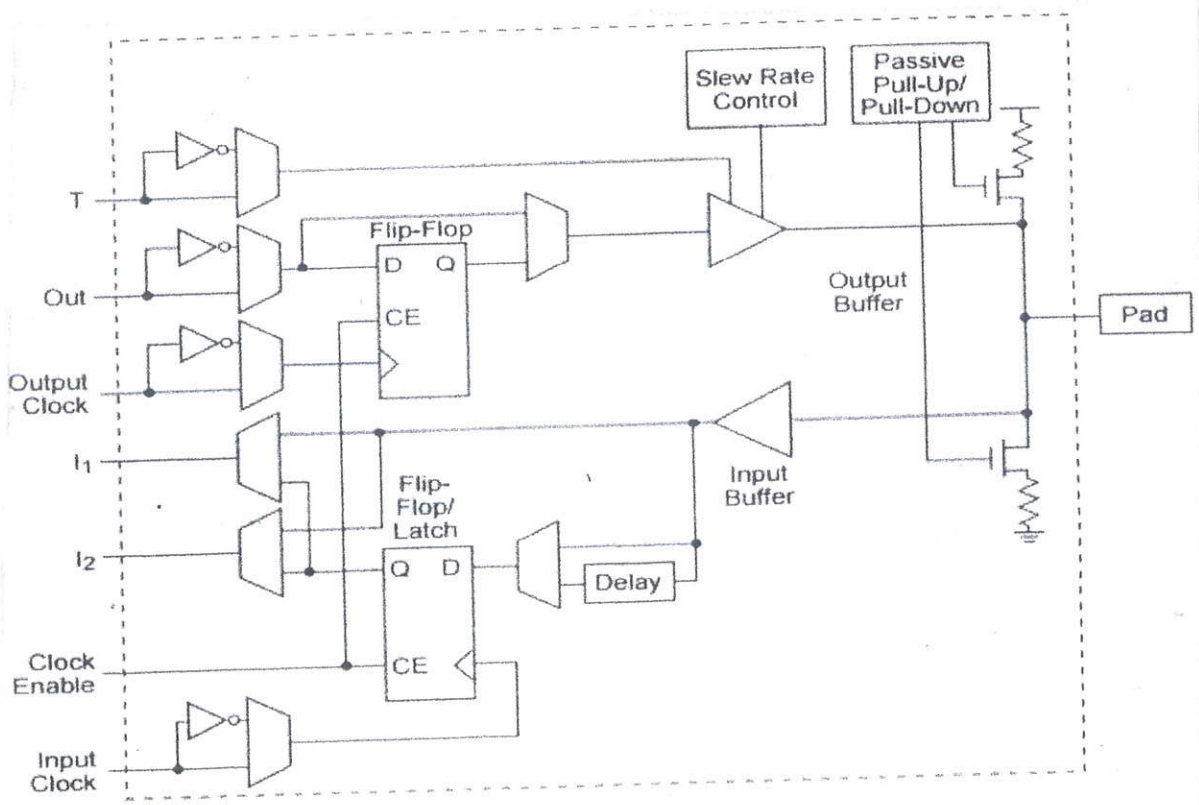


Figure Simplified Block Diagram of XC4000E IOB

pull-up and pull-down Resistor: programmable pull-up and pull-down resistors are useful for tying unused pins to Vcc or Ground to minimize power consumption and reduce noise sensitivity.

FPGA Interconnect Routing procedures:

All internal connections are composed of metal segments with programmable switching points and switching matrices to implement the desired routing. A structured hierarchical matrix of routing resources is provided to achieve efficient automated routing.

There are several types of interconnect.

* CLB routing is associated with each row and column of the CLB array.
* IOB routing forms a ring (called a VersaRing) around the outside of the CLB array. It connects the I/O with the internal logic blocks.
* Global routing consists of dedicated networks primarily designed to distribute clocks throughout the device with minimum delay and skew. Global routing can also be used for other high-fanout signals.

Five interconnect types are distinguished by the relative length of their segments: single length lines, double-length lines, quad and octal lines

(XC4000EX only) and longlines. Extra routing is included in the IOB pad ring. The XC4000EX also includes a ring of octal interconnect lines near the IOBs to improve pin-swapping and routing to locked pins

programmable wiring.

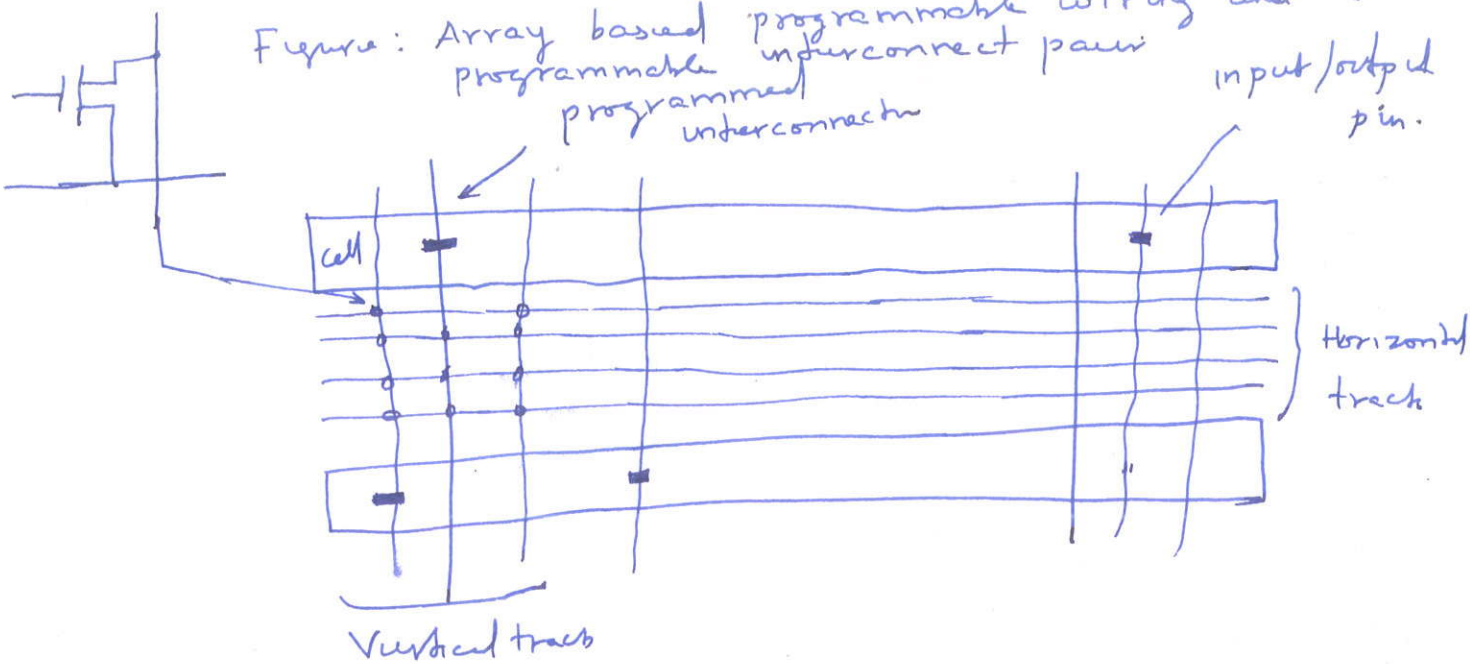
To fully utilize the available logic cells the interconnect network must be flexible and routing bottlenecks must be avoided. Speed is another prerequisite, since interconnect delay tends to dominate the performance in this style of design. programmable wiring can be classified into two major groupings: array and switch box routers.

Array-Based programmable wiring

In this approach, wiring is grouped into routing channels each of which contains a complete grid of horizontal and vertical wires.

An interconnect wires can then be programed into the structure by short circuiting some of the intersections between horizontal and vertical wires. This can be accomplished by providing a pass transistor at each of the cross points.

closing the interconnection means raising the control signal - by programming a 1 into the



connected memory cell M. This approach is expensive as it requires more transistors and control signals leading to high fan-out, delay and power consumption. A fuse is blown whenever a connection a connection is not needed but results in long programming times. An antifuse represents a small fraction of the overall grid and can be used when a connection is required. Circuit corrections or extension is not possible and new components are required for every design change.

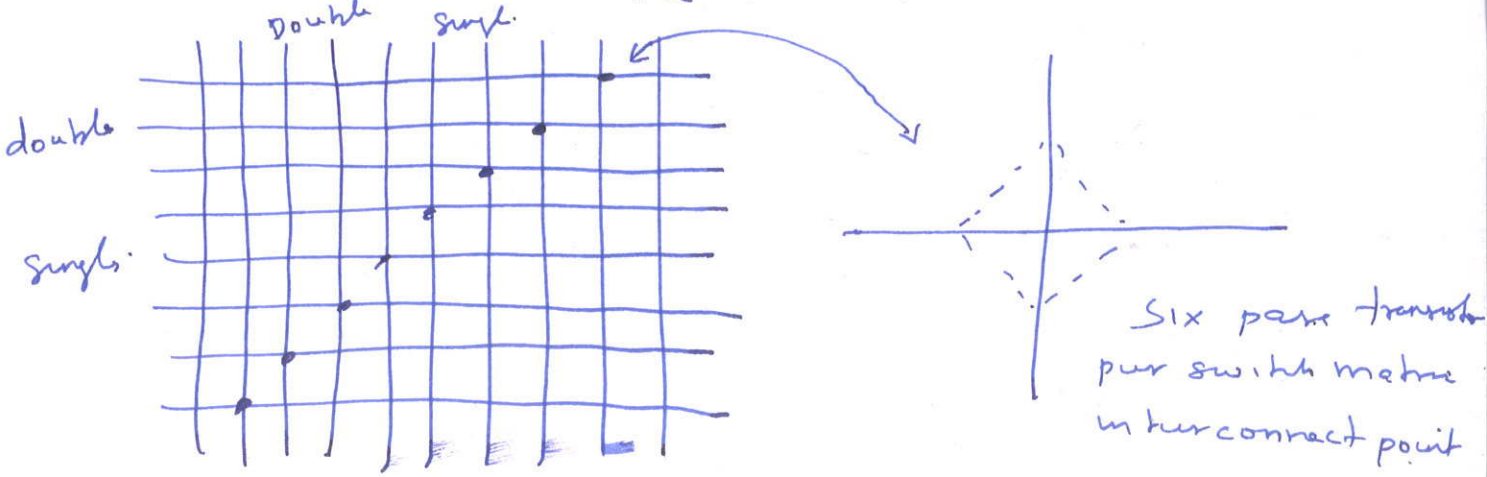
Switch-box based programmable wiring

Many local interconnection can be accounted for providing a mesh-like interconnection between neighbouring cells. For instance the output of each logic cell (CLB) can be distributed to its neighbours

in all directions. To account for interconnections between dugout cells or to provide global interconnection routing channels are placed between the cells containing a fixed number of uncommitted vertical and horizontal routing wires. RAM-programmable switching matrices (PSM) are provided that direct the routing of the data.

Programmable Switch Matrices

The horizontal and Vertical single and double length lines intersect at a box called a programmable switch matrix (PSM). Each matrix consists of programmable pass transistor used to establish connections between the lines.



For example a single-length signal entering on the right side of the switch matrix can be routed to a single-length line on the top, left or bottom sides or any combination thereof if multiple branches are required. //ly a double-length signal can be

routed to a double-length line on any or all of the other three edges of the programmable switch matrix.

Single-Length lines

Single-length lines provide the greatest interconnect flexibility and offer fast routing between adjacent blocks. There are eight vertical and eight horizontal single-length lines associated with each CLB. These lines connect the switching matrices that are in every row and a column of CLBs. Single-length lines are connected by way of the programmable switch matrices as shown in figure.

Single-length lines incur a delay whenever they go through a switching matrix. Therefore they are not suitable for routing signal for long distances. They are normally used to conduct signals within a localized area and to provide the branching for nets with fanout greater than one.

Double-length lines : consist of a grid of metal segments each twice as long as the single length lines: they run past two CLBs before entering a switch matrix. Double-length lines are grouped in pairs with the switch matrices staggered so that each line goes through a switch matrix at every other row or column of CLBs'.

There are four vertical and four horizontal double length lines associated with each CLB. These lines provides faster signal routing over intermediate distances, while retaining routing flexibility. Double length line are connected by way of programmable switch matrices. Routing connectivity is shown in the figure.

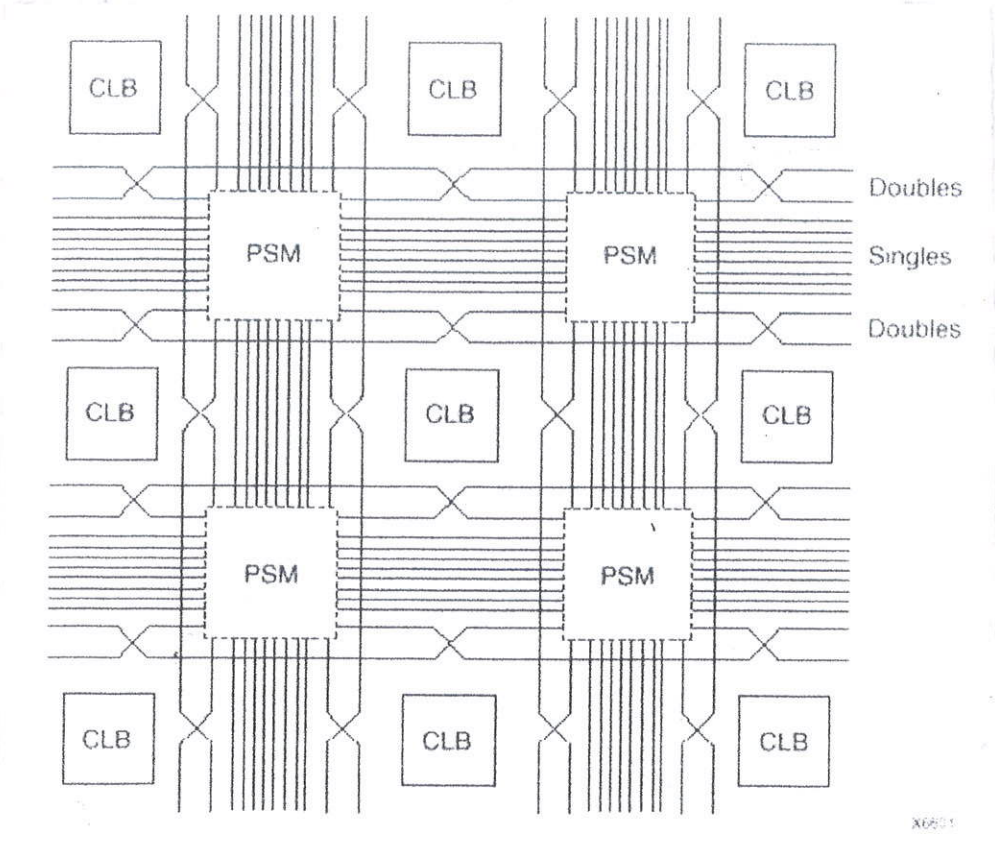


Figure Single- and Double-Length Lines, with Programmable Switch Matrices