VLSI DESIGN

UNIT-I

Introduction to Ic Technology:

Electronics as we know it today is characterized by reliability, low power dissipation, extremely low weight and volume & low cost, coupled with an ability to cope easily with a high degree of sophistication and complexity. Electronics & in particular the Integrated circuit, has made possible the design of Powerful and flexible processors which provide highly intelligent and adaptable devices for the user. IC memories have provided the essential elements to complement these processors & together with a wide range of logic & analog Integrated circuitry, they have provided the system designer with components of considerable capability & extensive application.

The invention of Transistor by William B. Shockley, Walter H. Brattain & John Bardeen of Bell Telephone Laboratories was followed by the development of the IC.

The very first IC emerged at the beginning of 1960 and since that time there have already been 4 generations of Ic's:

1) SSI (small scale Integration) (10-100)
2) MSI (mediam scale Integration) (100-1000)
3) LSI (Large scale Integration) (1K - 20K)
4) VLSI (Very Large Scale Integration

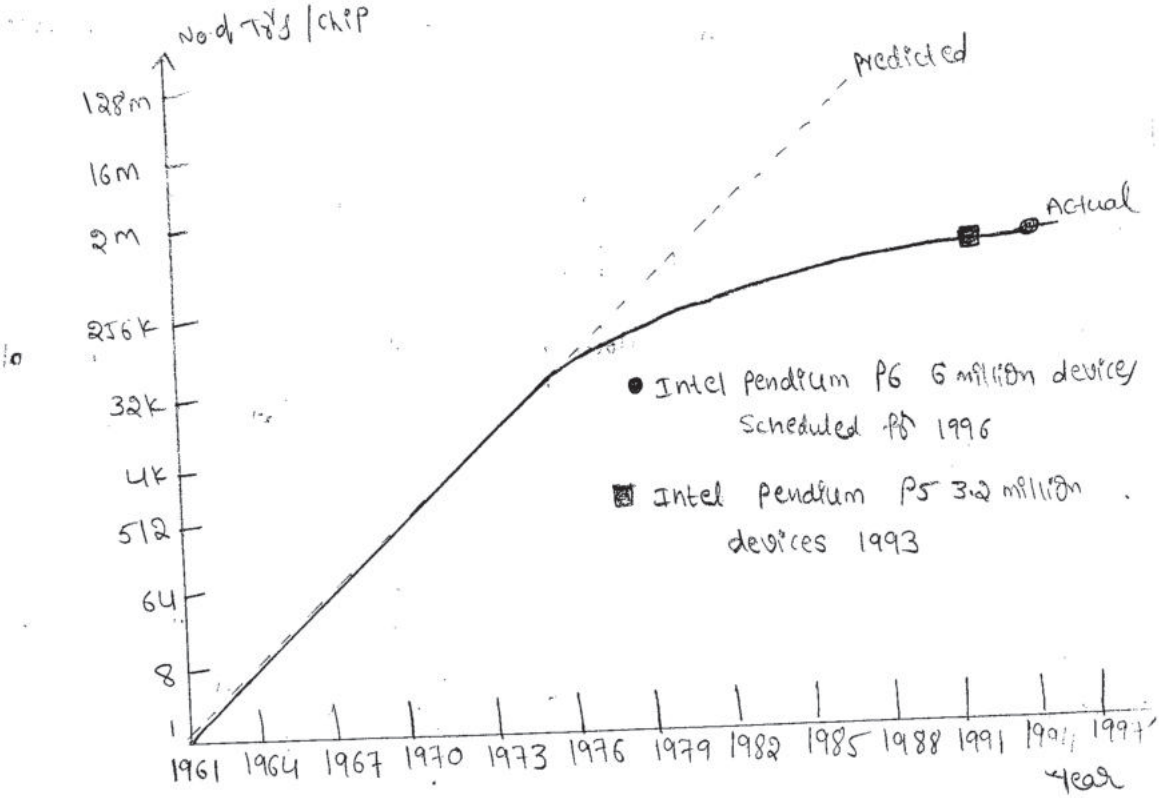
⇒ Upcoming 5th generation, ULSI (Ultra Large scale Integrati which is characterized by complexities in excess of 3 million devices on a single IC chip.

The IC Era :—

Such hay been the potential of silicon Integrated circuit that hey been an extremely rapid growth in the numbel of transistors being Integrated into circuits on a single silicon chip.

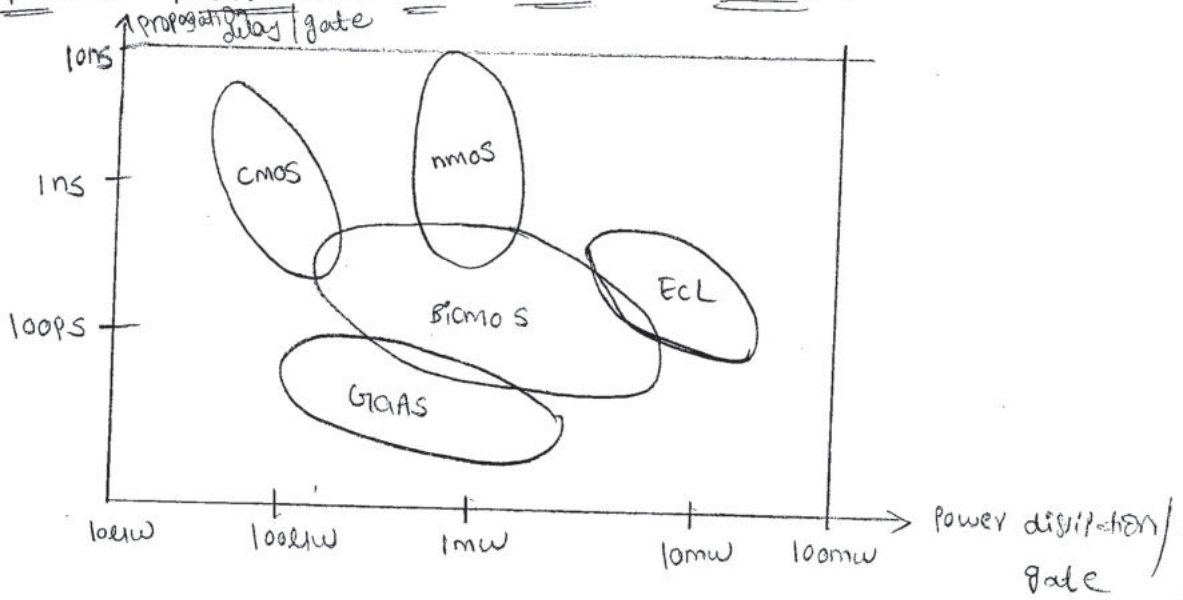
In lessthan 3 decades, this number hay risen from tens to millions.

This increase in number of Transistors per chip getting double for every 2 years is known as "moore's first Law". made by Gordon moore (of Intel) in the 1960's.



... predictions and actual over the last few years due to problems associated with the complexities involved in designing and testing such very large circuits.

Speed / Power performance of available technologies :-



over the past several years, silicon cmos technology has become the dominant fabrication process for relatively high performance & cost effective VLSI circuits.

→ The increase in no. of Trs/chip is highlighted by recent products such as RISC chips in which it is possible to process some 35 million instructions per second.

→ In order to improve on this throughput rate it will be necessary to improve the technology, both in terms of scaling and processing, & through the incorporation of other enhancements such as Bicmos.

→ In particular. The emerging Gallium Arsenide (GaAs) based technology will be most significant for ultra high speed logic / fast digital processors.

→ GaAs also has further potential as a result of its photo-electronic properties both as a Rx'er and as a Tx'er of light.

Microelectronics evolution:

Year	1947	1950	1961	1966	1971	1980	1990	2000	
Technology	Invention of Transistor	Discrete components	SSI	MSI	LSI	VLSI	VLSI*		GIGI
Approximate number of Transistors per chip in Commercial Products	1	1	10	100 – 1000	1000 – 20,000	20,000 – 1,000,000 (10 Lakh)	(1c Lakh) 1,000,000 – 10,000,000 (1 Crore)		> 1 Crore 10,000,000
Typical Products	–	Junction Transistor & diode	Planar devices, logic gates FF	Counters, MUX, Adders	8-bit µP, Rom, RAM	16 & 32 bit µP, Sophisticated Peripherals (Ultra Dram	special processors, Virtual reality machine, smart sensors		

GISI : Giant-scale Integration

Metal-oxide-semiconductor (MOS) and related VLSI Technology:-

Within the bounds of MOS Technology, the possible circuit realizations may be based on PMOS, NMOS, CMOS & BiCMOS devices. GaAs Technology is also one of the leading IC Technology.

Although CMOS is the dominant technology, some of the examples used to illustrate the design processes will be presented in NMOS form. The reasons for this are

(1) For NMOS Technology, the design methodology and the design rules are easily learned, thus providing a simple but excellent introduction to structured design for VLSI.

(2) NMOS technology and design processes provide an excellent background for other technologies. In particular some familiarity with NMOS allows a relatively easy transition to CMOS Technology and design.

(3) For GaAs Technology some arrangements in relation to logic design are similar to those employed in NMOS Technology.

Not only in VLSI Technology, providing the user with a new and more complex range of "off the shelf" circuits, but VLSI processes are such that system designers can readily design their own special circuits of considerable

→ This provides a new degree of freedom for designers and it is probable that some very significant advances will result.

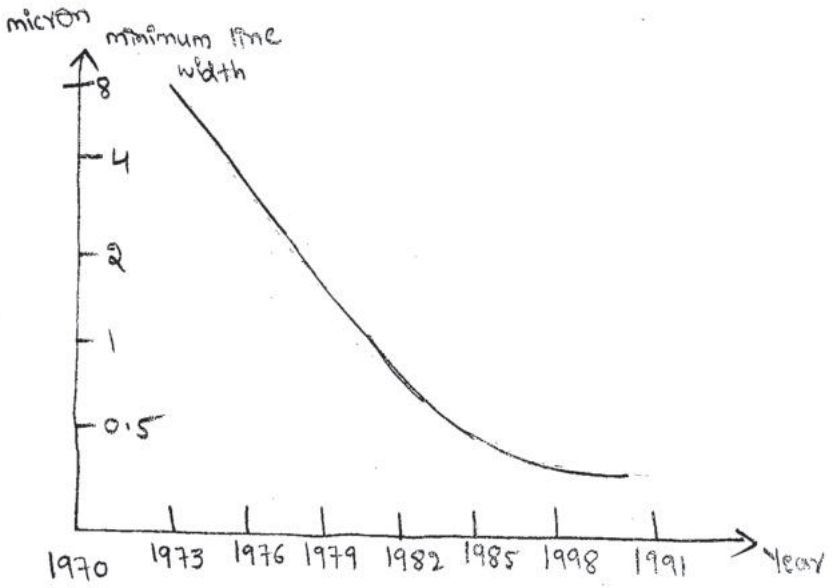


Fig. approximate minimum line width of commercial products Vs year

→ Couple this with the fact that integration density is increasing rapidly, as advances in technology shrink the feature size for circuits integrated in silicon.

The effectiveness of the circuits produced has increased with scaling down as shown in fig.

→ A common measure of effectiveness is the speed power product of the basic logic gate circuit of the technology (for nmos, the nor gate; with nand and nor gates for cmos).

→ speed power product is measured in picojouley (pJ) and

is the product of the gate switching delay in nanoseconds and the gate power dissipation in milliwatts.

Basic Mos transistors:-

nmos enhancement and depletion mode transistors:-

→ nmos devices are formed in a p-type substrate of moderate doping level.

→ The source and drain regions are formed by diffusing n-type impurities through suitable masks into these areas to give the desired n- impurity concentration and give rise to depletion regions which extend mainly in the more lightly doped p-region.

→ Thus source and drain are isolated from one another by 2 diodes.

→ connections to the source and drain are made by a deposited metal layer.

→ In order to make a useful device, there must be the capability to establishing and controlling a current b/w source and drain.

→ This is achieved in one of 2 ways, giving rise to the enhancement mode and depletion mode Transistors.

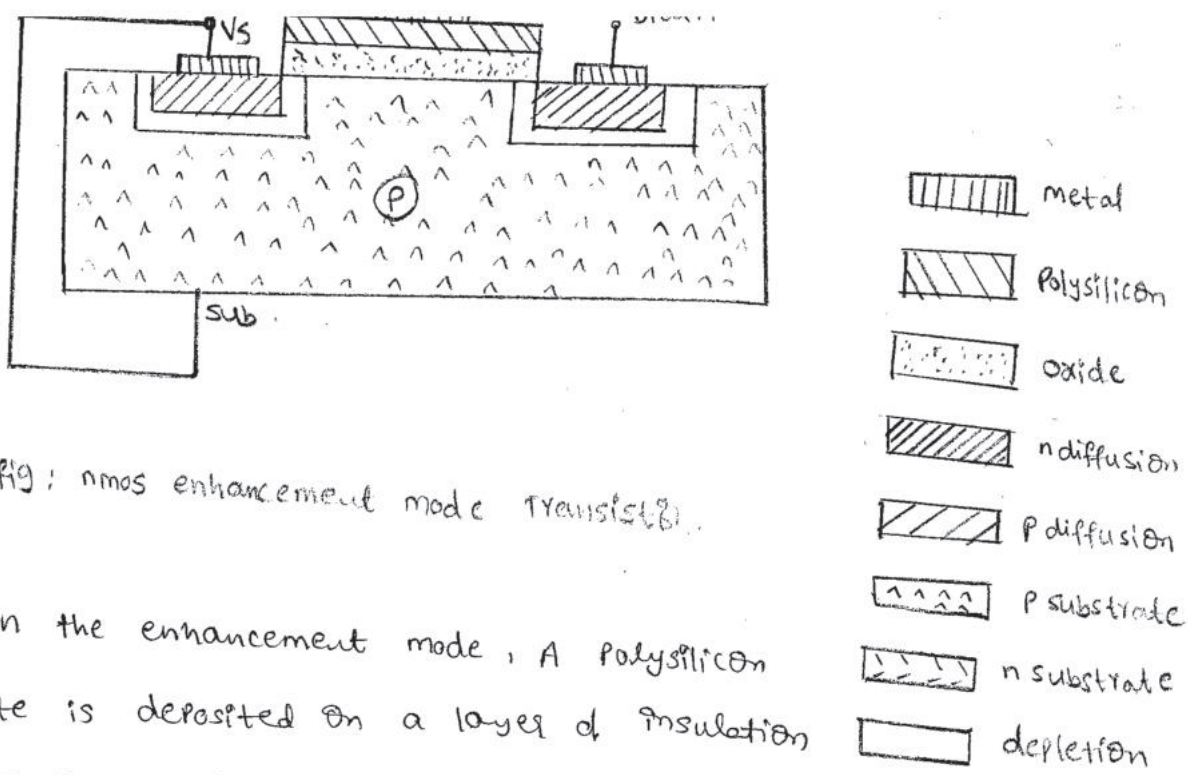


fig: nmos enhancement mode transistor.

In the enhancement mode, A polysilicon gate is deposited on a layer of insulation over the region b/w source and drain.

→ In enhancement mode device the channel is not established and the device is in a non-conducting condition, $V_D = V_S = V_{gs} = 0$.

→ If the gate is connected to a suitable positive voltage with respect to source, then the electric field established b/w the gate and substrate give rise to a charge inversion region in the substrate under the gate insulation and a conducting path of channel is formed b/w source and drain.

→

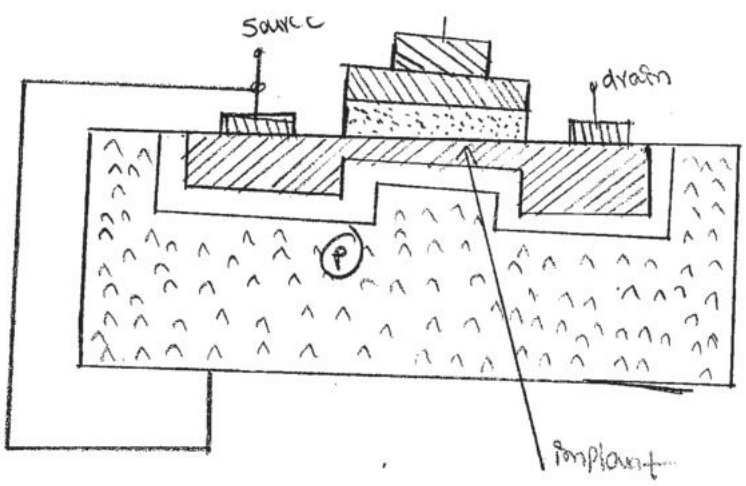
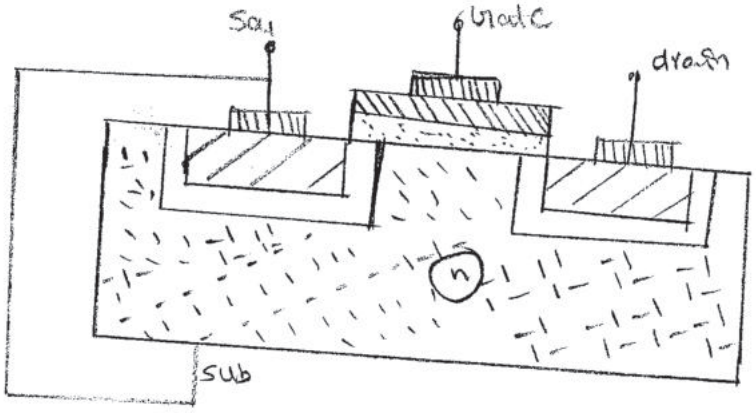


fig : nmos depletion mode transistor

In depletion mode, the channel may also be established so that it is present under the condition $V_{GS} = 0$ by implanting suitable impurities in the region b/w source and drain during manufacture and prior to depositing the insulation and the gate. Under these circumstances, source and drain are connected by a conducting channel, but the channel may now be closed by applying a suitable negative voltage to the gate.

In both cases, variations of the gate voltage allow control of any current flow b/w source and drain.



In Pmos enhancement mode Transistor, the substrate is of n-type material and the source and drain diffusions are p-type. By the application of a negative voltage of suitable magnitude ($>|V_t|$) b/w gate and source will rise to the formation of a channel (p-type) b/w source and drain and current may then flow if the drain is made negative with respect to the source.

→ In this case, the current is carried by holes as opposed to es.

→ The Pmos Transistors are inherently slower than nmos, since $\mu_n = 2.5\,\mu_p$.

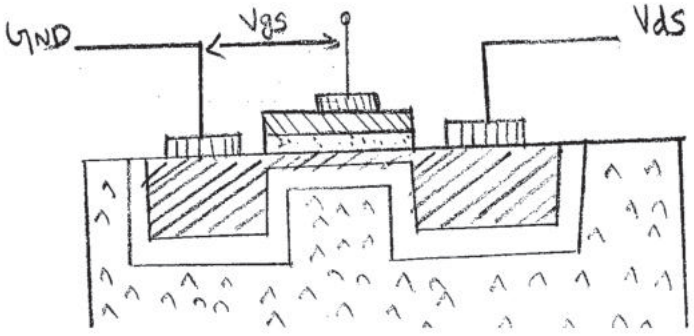
$\mu_n = 650\ cm^2/Vsec$

$\mu_p = 240\ cm^2/Vsec$.

Enhancement mode Transistor Action:

To understand the mechanism of Enhancemed mode we have to consider 3 conditions.

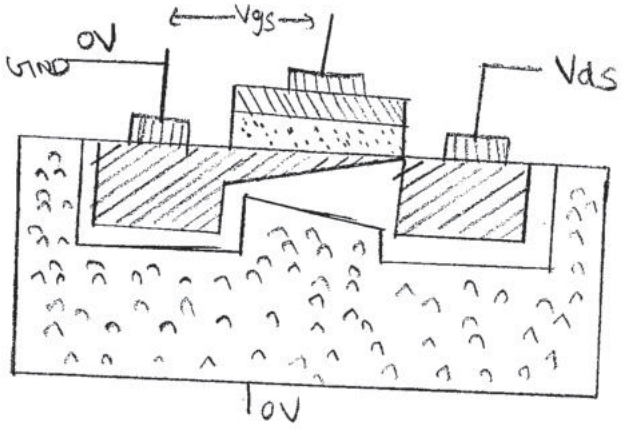
Threshold voltage (V_t):- The minimum voltage applied b/w gate and source to establish channel.



$V_{gs} > V_t$

$V_{ds} = 0V$

Fig. indicates the conditions prevailing with the channel established but no current flowing b/w source and drain ($V_{ds} = 0$).



$V_{gs} > V_t$

$V_{ds} < V_{gs} - V_t$

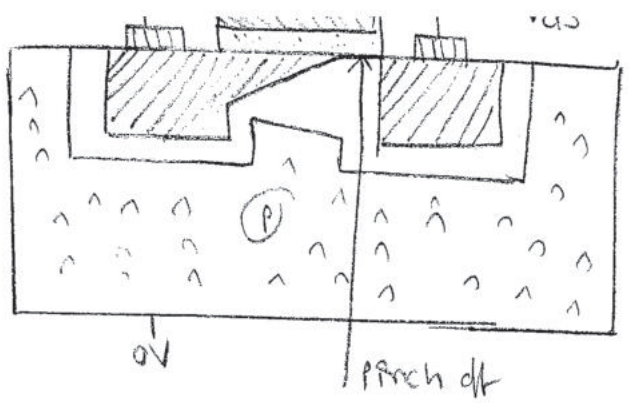
→ Now consider the conditions prevailing when current flows in the channel by applying a voltage V_{ds} b/w drain and source.

→ Along the channel, a corresponding IR drop = V_{ds} will be there.

→ This results in the voltage b/w gate and channel varying with distance along the channel with the voltage being a maximum of V_{gs} at the source end.

→ Since the effective gate voltage is $V_g = V_{gs} - V_t$ (no current flows when $V_{gs} < V_t$), there will be voltage available to invert the channel at the drain end so long as $V_{gs} - V_t \geq V_{ds}$.

→ The limiting condition come when $V_{ds} = V_{gs} - V_t$. For all voltages $V_{ds} < V_{gs} - V_t$, the device is in the non-saturation region of operation which is $\boxed{V_{ds} \leq V_{gs} - V_t}$.



$V_{gs} \geq V_t$

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

If V_{ds} is increased to a level greater than $V_{gs} - V_t$. In this case, an IR drop = $V_{gs} - V_t$ takes place over less than the whole length of the channel so that over part of the channel, near the drain, there is insufficient electric field available to give rise to an inversion layer to create the channel. The channel is therefore "pinch-off".

→ Diffusion current completes the path from source to drain in this case, causing the channel to exhibit a high resistance and behave as a constant current source.

→ This region is known as saturation, is characterized by almost constant current for increase of V_{ds} above $V_{ds} = V_{gs} - V_t$.

→ In all the cases the channel will cease to exist and no current will flow when $V_{gs} < V_t$.

→ The typical values for enhancement mode devices

$V_t = 1V$ for $V_{DD} = 5V$

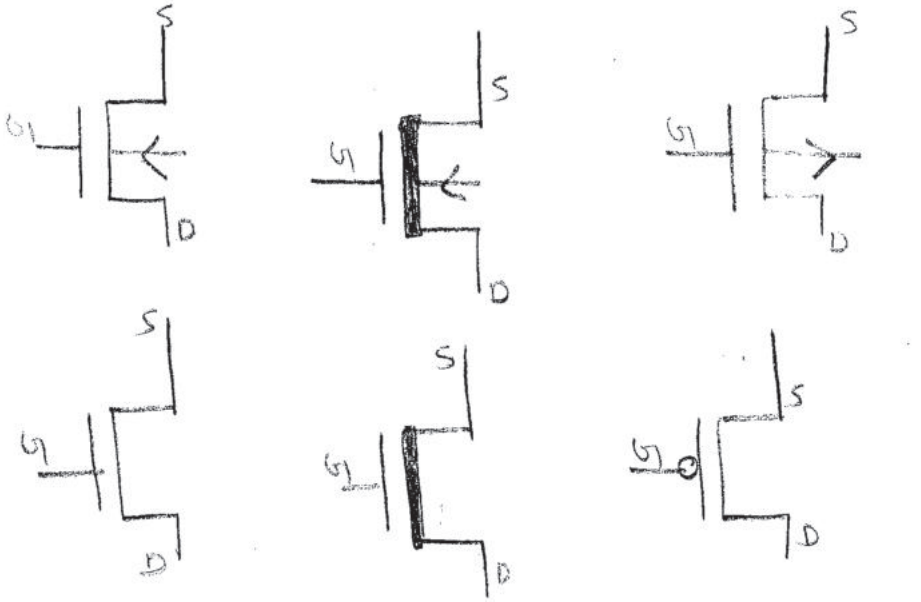
∴ In general $V_t = 0.2 V_{DD}$.

Depletion mode transistor Action:-

For depletion mode device the channel is established because of the implant, even when $V_{GS} = 0$ and to cause the channel to cease to exist a negative voltage V_{td} must be applied b/w gate and source.

V_{td} is typically $\angle -0.8 V_{DD}$, depending on implant and substrate bias, but threshold voltage difference aside the action is similar to that of enhancement mode T&R.

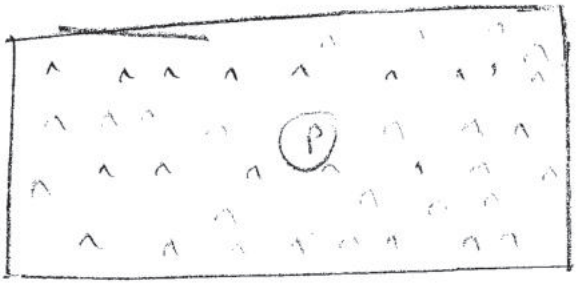
Symbols for nmos & pmos transistors:



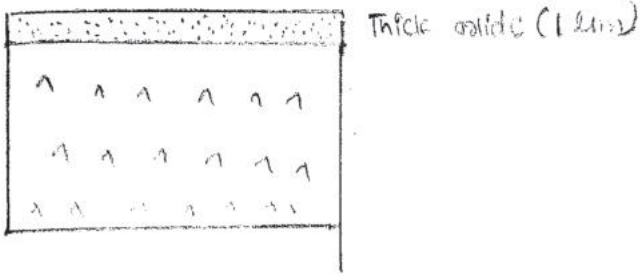
nmos enhancement | nmos depletion | pmos enhancement

nmos fabrication:—

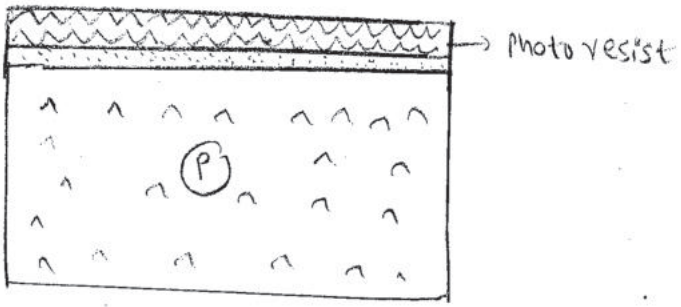
① Processing is carried out on a thin wafer cut from a single crystal of silicon of high purity into which the required P-impurities are introduced as the crystal is grown. Such wafers are typically 75 to 150mm in diameter and 0.4mm thick and are doped with, say boron to impurity concentration of $10^{15}/cm^3$ to $10^{16}/cm^3$, giving resistivity in the approximate range 25 ohm cm to 2 ohm cm.



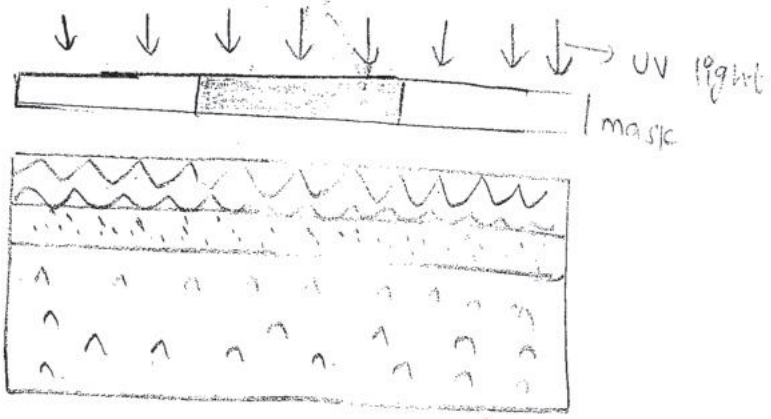
② A layer of silicon dioxide (SiO_2), typically 1 μm thick is grown all over the surface of the wafer to protect the surface, act as a barrier to dopants during processing and provide a generally insulating substrate onto which other layers may be deposited and patterned.



③ The surface is now covered with a photoresist which is deposited onto the wafer and spun to achieve an even



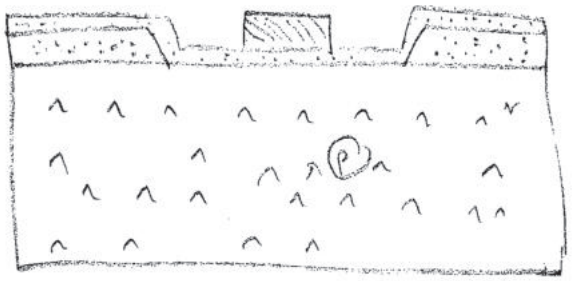
(4) The photoresist layer is then exposed to ultraviolet light through a mask which defines those regions into which diffusion is to take place together with transistor channels. For Example, that those areas exposed to ultraviolet radiation are polymerized (hardened), but that the areas required for diffusion are shielded by the mask and remain unaffected.



(5) These areas are subsequently readily etched away together with the underlying silicon dioxide so that the wafer surface is exposed in the window defined by the mask.

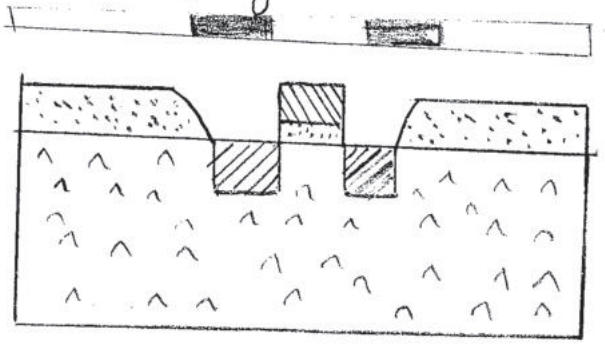
(6) The remaining photoresist is removed and a thin layer of SiO_2

is grown over the entire chip surface and then Polysi is deposited on top of this to form the gate structure. The polysi layer consists of heavily doped polysi, deposited by "Chemical Vapor Deposition (CVD)". In the fabrication of fine pattern devices, precise control of thickness, impurity concentration & resistivity is necessary.

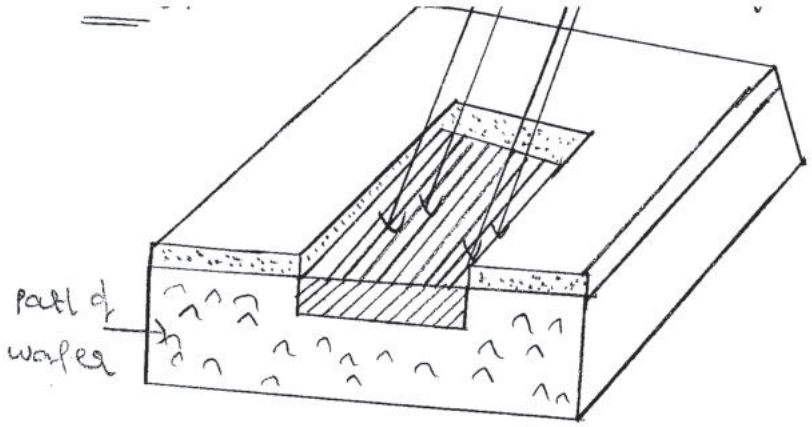


Patterned poly (1-2 μm) on thin oxide (800-1000 Å)

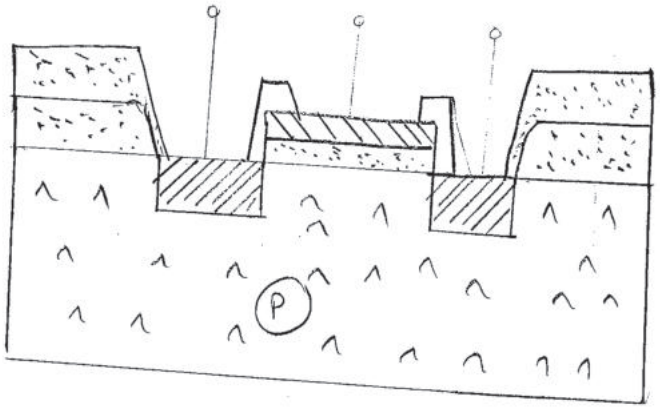
(7) Further photoresist coating and masking allows the polysilicon to be patterned, and then the thin oxide is removed to expose areas into which n-type impurities are to be diffused to form the source and drain. Diffusion is achieved by heating the wafer to a high temperature and passing a gas containing the desired n-type impurity (for example, phosphorus) over the surface. Note that the polysilicon with underlying thin oxide and the thick oxide act as masks during diffusion — the process is self-aligning.



n^+ diffusion (1 μm deep)

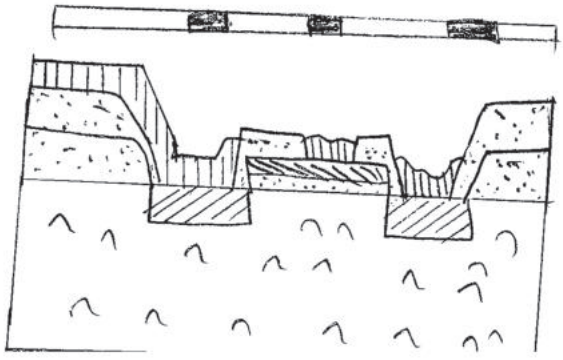


(8) Thick oxide (SiO_2) is grown over all again and is then masked with photoresist and etched to expose selected areas of the polysilicon gate and the drain and source areas where connections (ie contact cuts) are to be made.

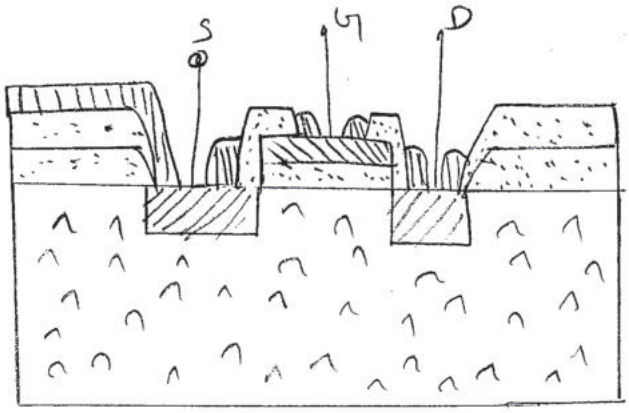


contact holes (cuts)

(9) The whole chip then has metal (aluminum) deposited over its surface to a thickness typically of 1 μm. This metal layer is then masked and etched to form the required interconnection pattern.



Patterned metalization (aluminum 1 μm)



The process revolves around the formation & deposition and patterning of 3 layers, seperated by silicon dioxide insulation. The layers are diffusion within the substrate, Polysilicon on oxide on the substrate & metal insulated again by oxide.

→ To form depletion mode devices it is only necessary to introduce a masked ion implantation step after forming window in oxide (b/w steps and 6). Again the thick oxide acts as a mask and this process stage is also "self-aligning".

→ Some extra process steps are necessary, including the overglassing of the whole wafer, except where contacts to the outside world are required.

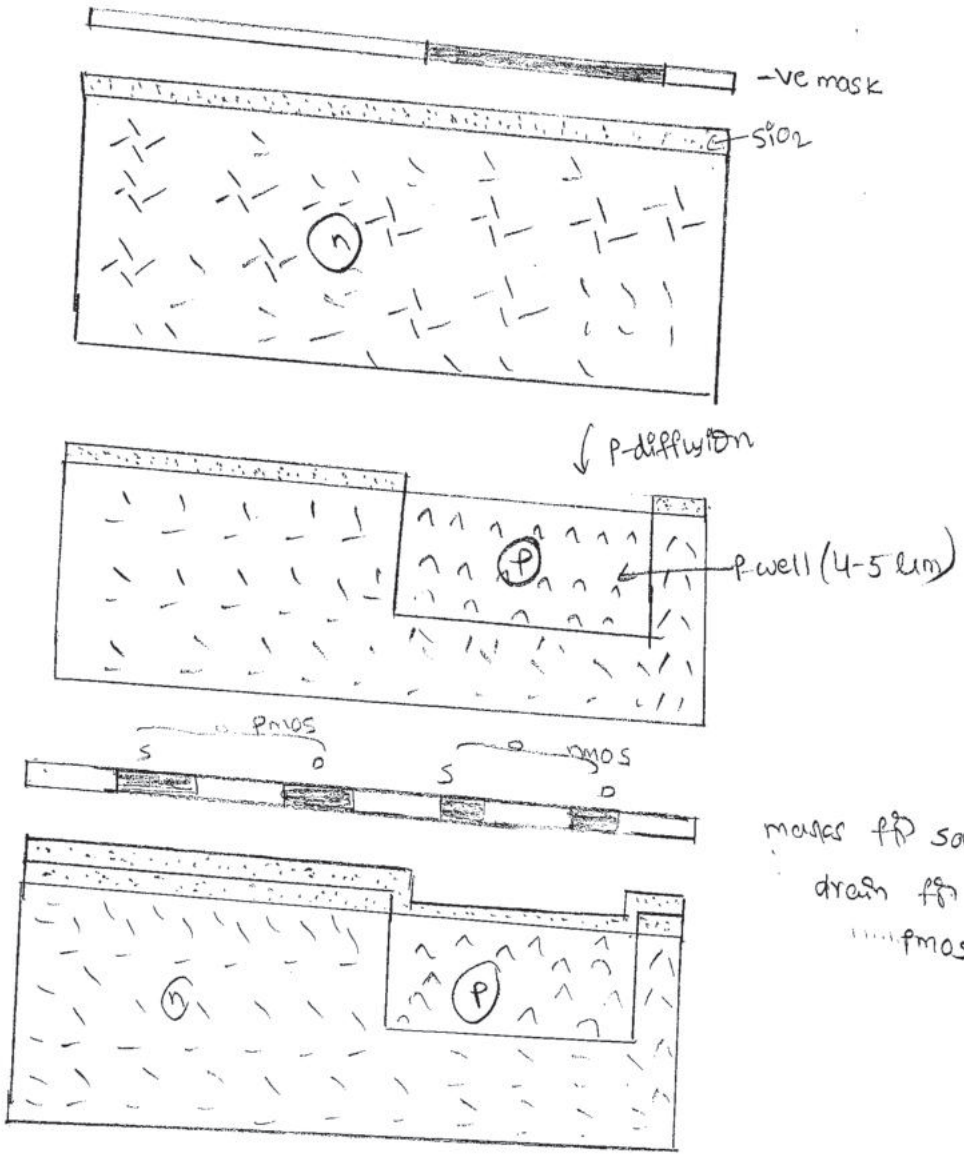
There are number of approaches to cmos fabrication.

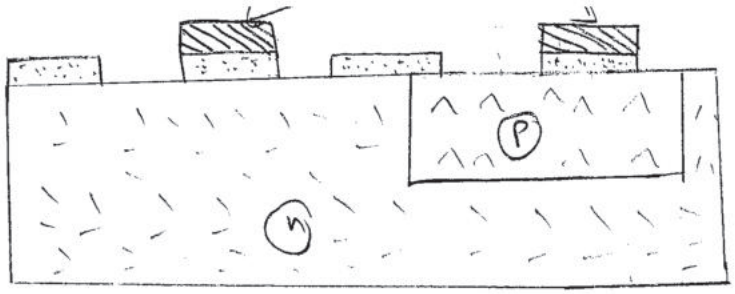
They are ① P-well process
② n-well process
③ Twin tub process
④ Silicon-On-Insulator (SOI) process.

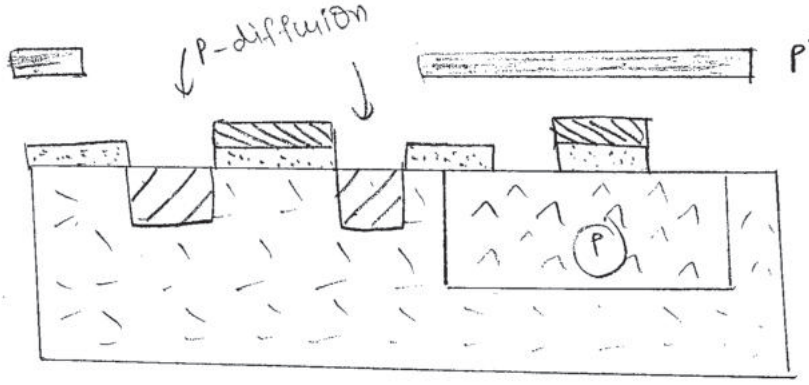
→ we will concentrate on well-bayed circuits. The P-well process is widely used in practice and the n-well process is also popular, particularly as it was an easy retrofit to existing nmos lines.

P-well Process:

①





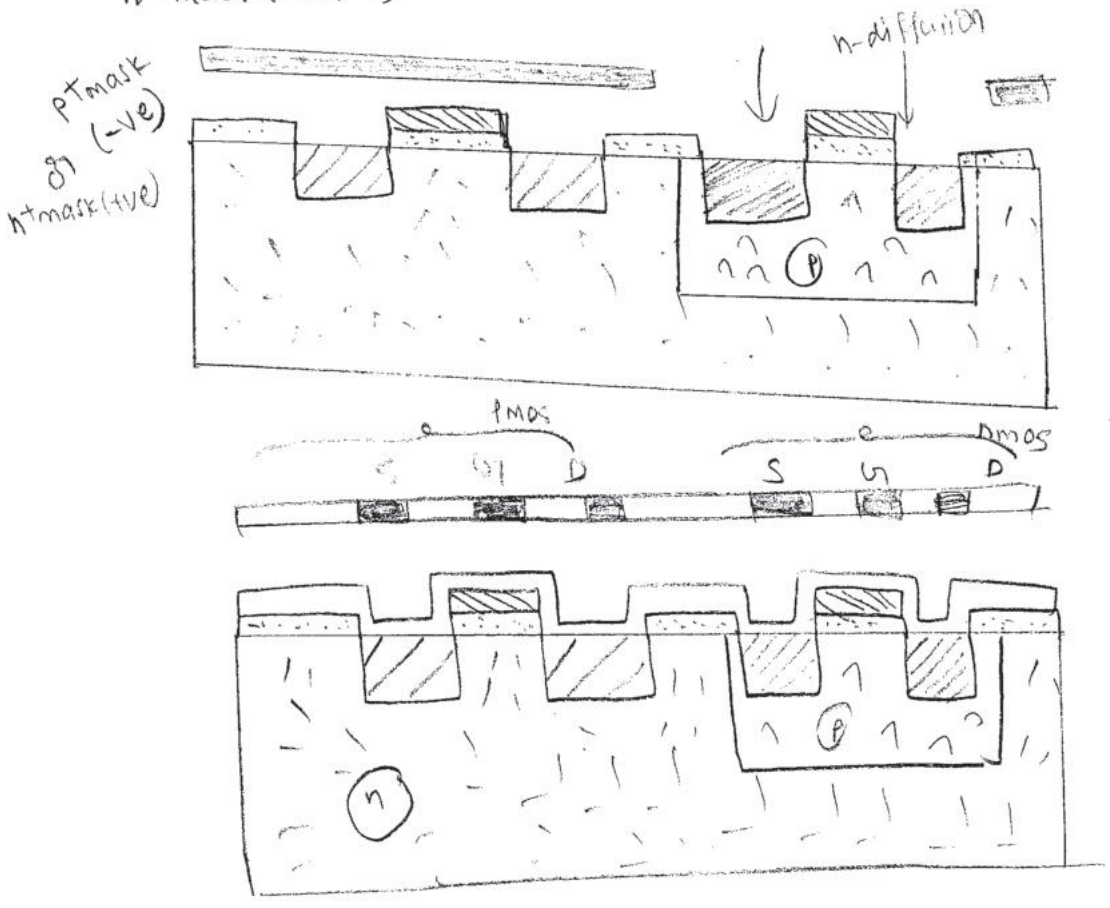


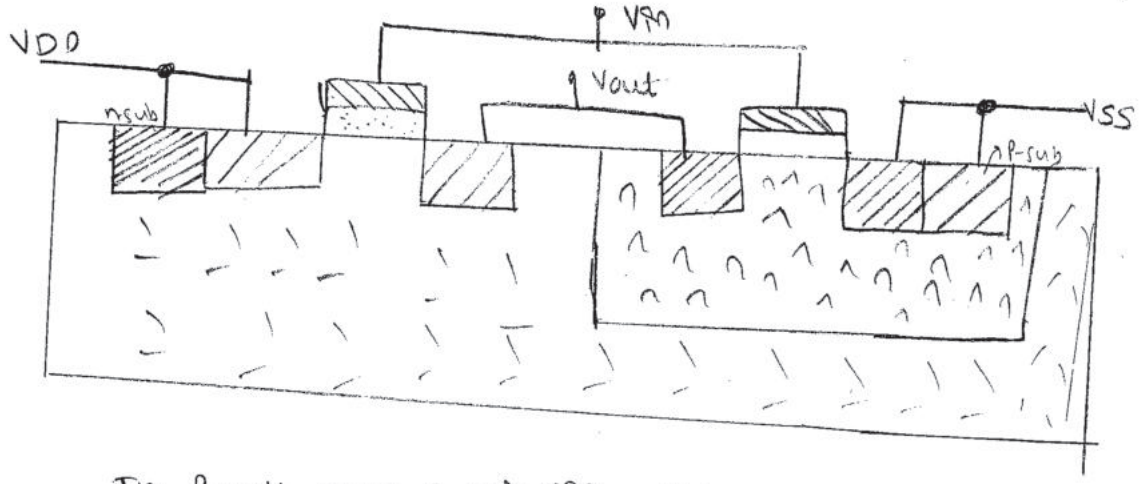
p+mask (+ve) ⇒ masked area not effected for P-diffusion.

p+mask (-ve) ⇒ masked area not effected for n diffusion.

n+mask (+ve) ⇒ masked area not effected for n-diffusion.

n+mask (-ve) ⇒ " " not effected for p-diffusion.





In P-well cmos fabrication, structure consists of an n-type substrate in which P-device may be formed by suitable masking and diffusion and in order to accommodate n-type device, a deep P-well is diffused into the n-type substrate.

→ The diffusion must be carried out with special care since the Pwell doping concentration & depth will affect the threshold voltages as well as the breakdown voltages of n-Transistors.

→ TO achieve low threshold voltages (0.6 to 1V), we need either deep well diffusion or high well resistivity.

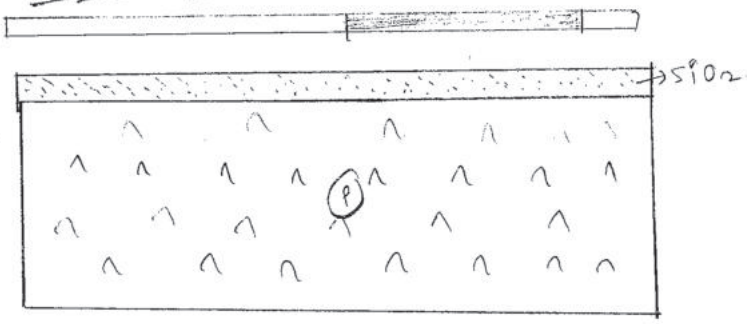
→ However deep wells require larger spacing b/w n- and P-type transistors and wires because of lateral diffusion & therefore a larger chip area.

CMOS n-well Process :-

n-well CMOS circuits are also superior to P-well because of the lower substrate bias effects on transistor threshold voltage and inherently lower parasitic capacitances associated with source and drain regions.

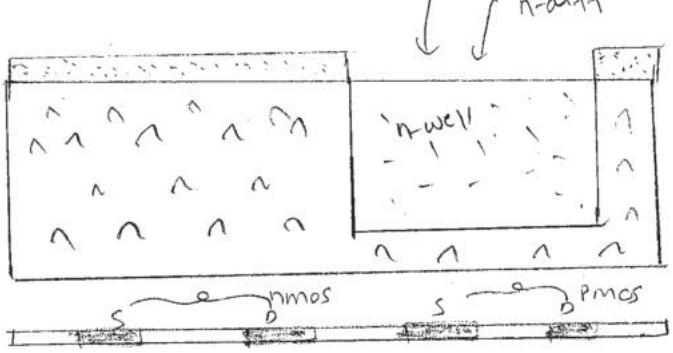
Fabrication steps in Typical n-well process :-

(1)



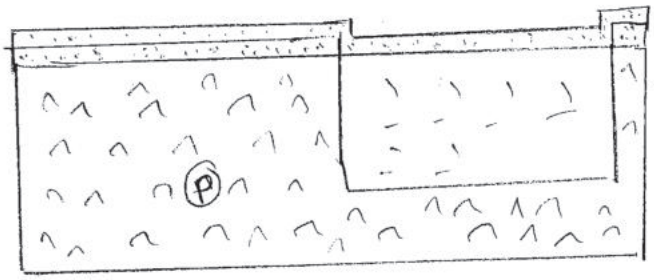
→ Wafer
→ doping with P-sub,
→ deposite oxide layer.

(2)



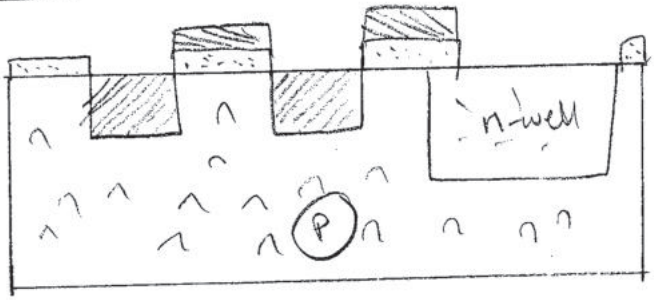
→ etching.
→ n-diffusion for well.

(3)

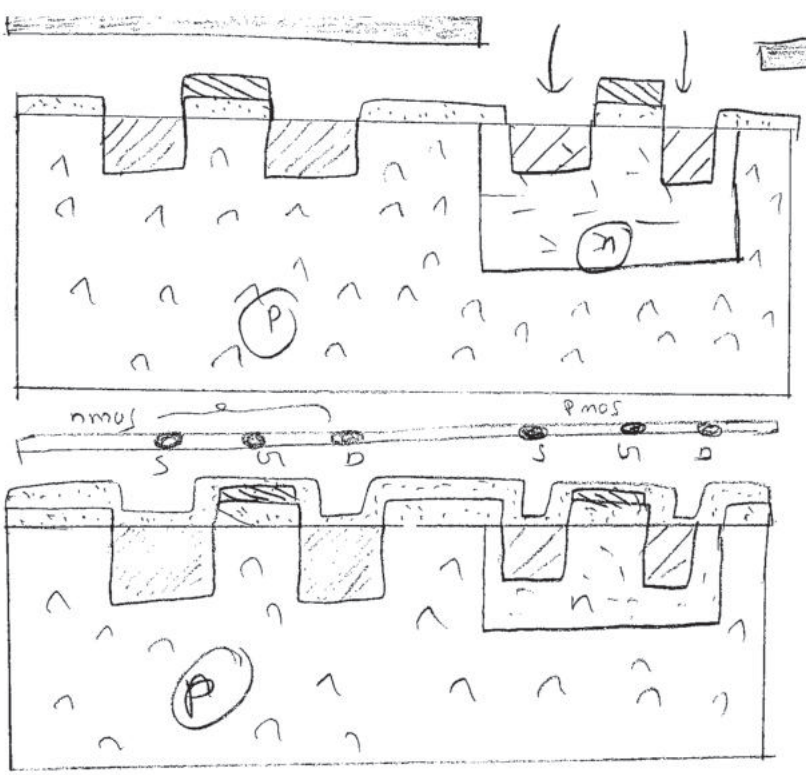


→ again deposit oxide
→ mask areas for source & drain for both Trs.

(4)



→ etching areas for S & D for both Trs.
→ form gate by depositing Polysilicon
→ For S & D of nmos by using



→ form S & D for P-MOS.

→ deposit SiO_2

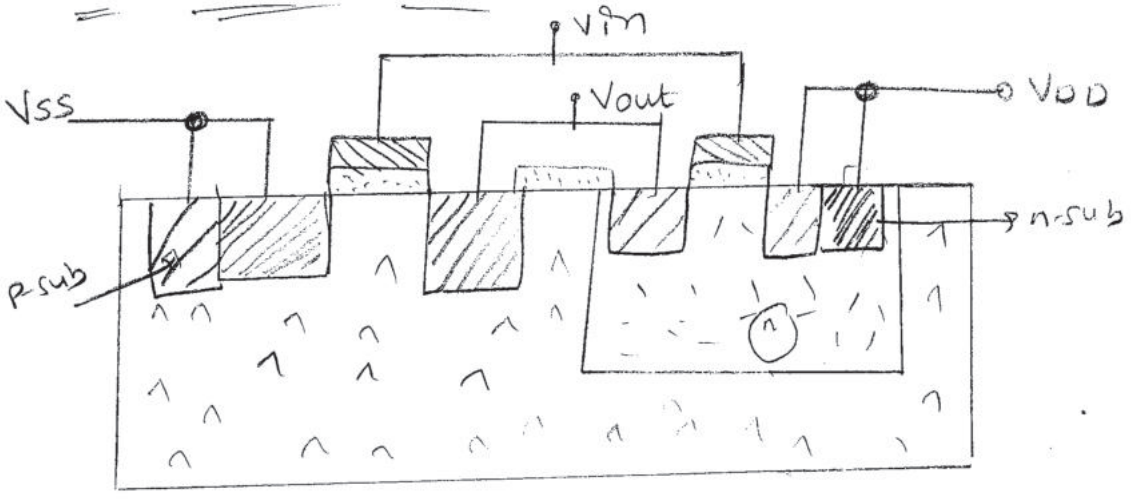
→ mask areas for S, G, D for both T's.

→ Then etch the masked areas & take contact cuts.

→ Finally deposit metal layer & then etch areas for contact cuts.

→ overall glass with cuts for bonding pads.

CMOS n-well inverter :-



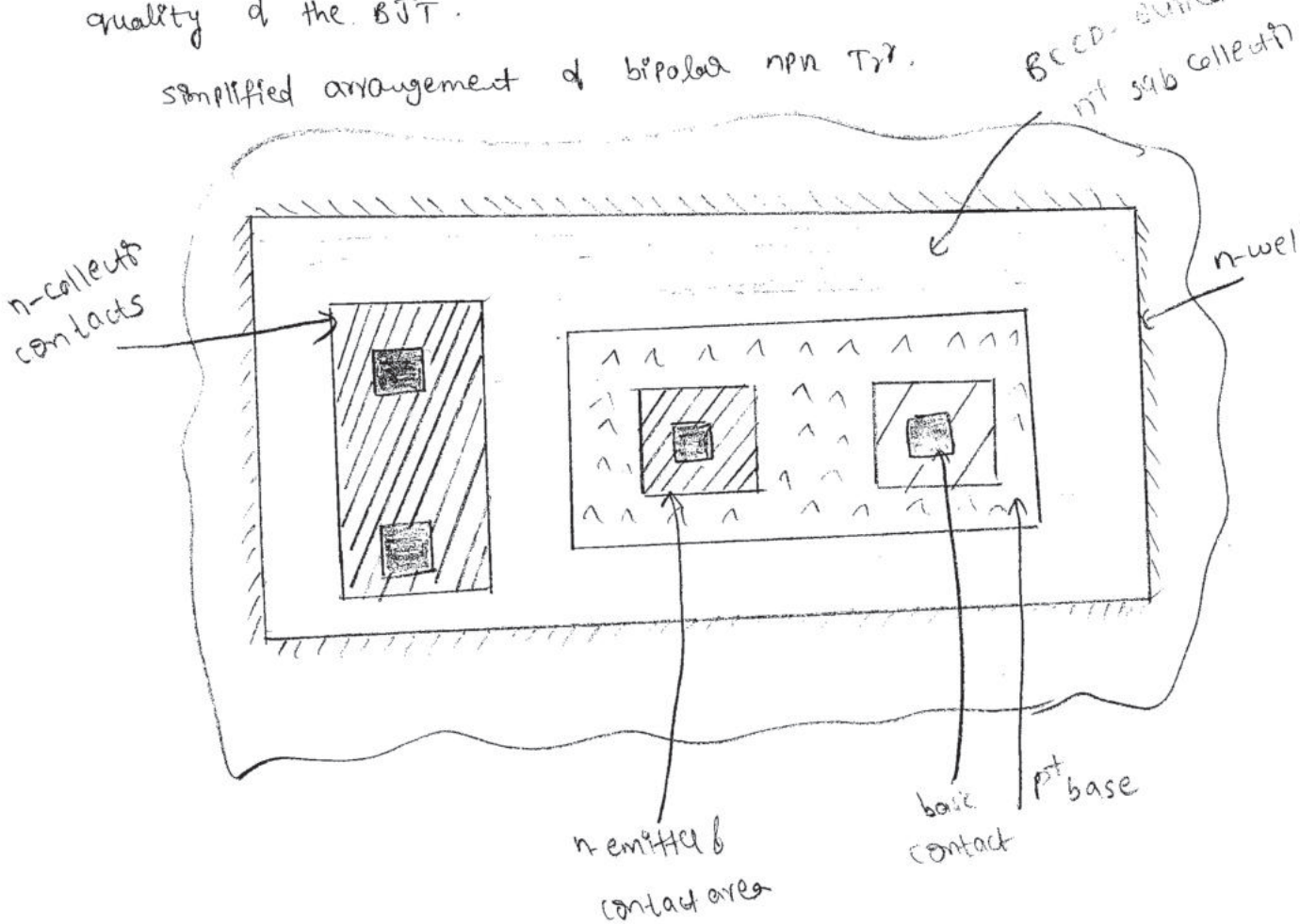
Bi-cmos Technology :-

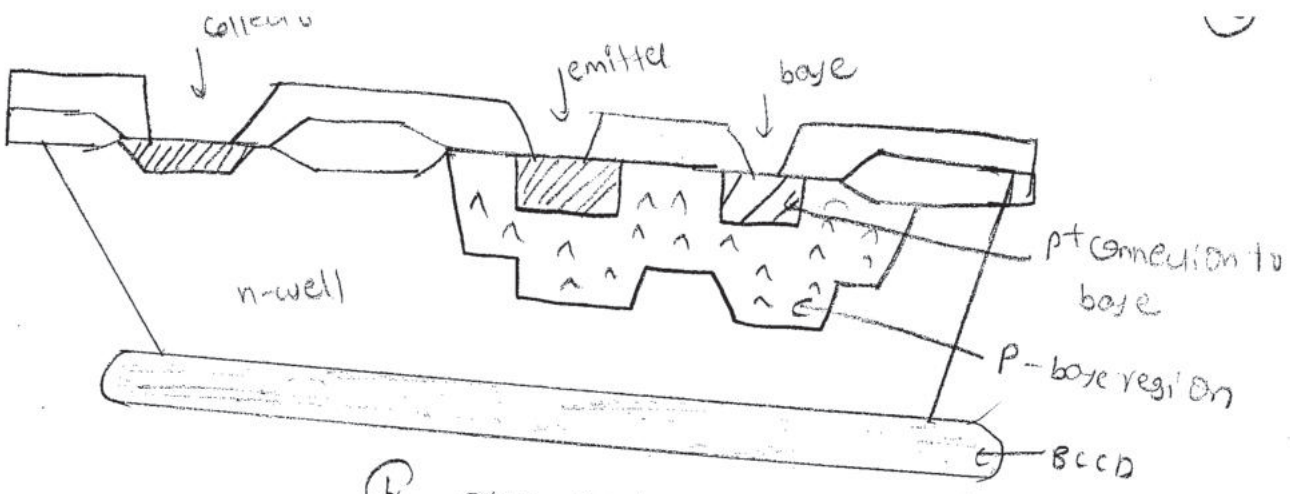
The production of npn bipolar Transistors with good performance ch's can be achieved for example, by extending the standard n-well cmos processing to include further masks to add 2 additional layers — The n^+ subcollector
— p^+ base layer

The npn T_r is formed in an n-well and additional p^+ base region is located in the well to form the p-base region of T_r.

→ The 2nd layer (additional), the Buried n^+ subcollector (BCCD) is added to reduce n-well (collector) resistance & thus improve quality of the BJT.

simplified arrangement of bipolar npn T_r.





(b) Cross section

Bicmos fabrication in an n-well process:

→ Form n-well & form source, gate, drain for both pmos & nmos.

→ Then additional masks to define

(1) pt base region
(2) nt collector region
(3) Buried sub collector.

Cmos

(1) Form n-well

(3) Define active area

(5) Vt adjustment

(6) Define poly/gate area

(7) Form nt active area

(8) Form pt active area

(10) Contact cuts

(11) metal area

Additional steps for Bipolar.

(2) Form buried nt layer

(4) Form deep nt collector

(9) Form pt base for bipolars

Bicmos Technology goes some way toward combining the virtues of both Technologies.

Comparison b/w CMOS & BJT:-

CMOS	BJT
(1) Low static power dissipation	(1) high power dissipation.
(2) High i/p impedance	(2) Low i/p impedance (high drive current)
(3) High packing density	(3) Low packing density
(4) High delay sensitivity to load	(4) Low delay sensitivity to load.
(5) Low o/p drive current	(5) High o/p drive current
(6) Low gm (gm α Vin)	(6) High gm (gm α eVin)
(7) Bidirectional capability (source & drain interchangeble)	(7) Essentially Unidirectional
Ideal switching device	
→ Scalable threshold voltage	
→ High noise margin	

To take maximum advantage of si Technology one might envisage the following mix of Technologies in a si system

CMOS for logic

Bi-CMOS for I/o & driver circuits

ECL for critical high speed parts of the system.

Diffusion:- In order to achieve selective doping, the technique most commonly used in silicon processing is called as "diffusion".

→ The basic principle underlying this process is that the dopant atoms migrate from a region of high concentration to the region of low concentration.

→ In simple diffusion is the process of introducing controlled amounts of dopants into the semiconductors.

Three kinds of situations arise in the process of diffusion.

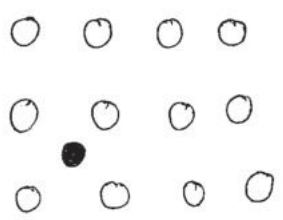
1. Substitutional diffusion
2. Interstitial "
3. Interstitialcy "

Substitutional Diffusion:—

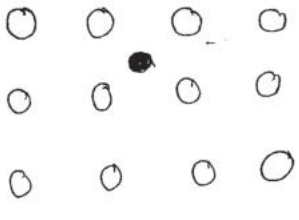
An impurity atom wanders through the crystal by jumping from one lattice site to the next, thus substituting for the original host atom. It is necessary that this adjacent site be vacant. i.e, vacancy must be present to allow substitutional diffusion to occur.

Note: Due to high temperature, silicon atoms get removed from its lattice position in the crystal & the impurity atom takes its place.

Interstitial diffusion :— An impurity atom moves through the crystal lattice by jumping one interstitial site to the next. Interstitial diffusion requires that jump motion occurring from one interstitial site to another adjacent interstitial site. This process is relatively fast, because of large number of vacant interstitial place in the semiconductor crystal.

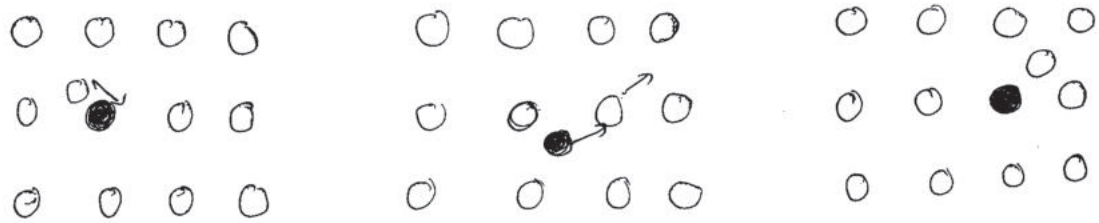


Before



After.

Interstitialcy diffusion :— This is modified version of substitutional diffusion. Interstitial host atoms can be annihilated by pushing substitutionally located impurity atoms in interstitial sites. These impurities can now diffuse to adjacent substitutional sites & create new self-interstitials.



Diffusion Equation :

The basic diffusion process is similar to that of charge carriers (Es & holes). we define flux 'F' as the no. of dopant atoms passing through a unit area in unit time and 'C' as

per unit volume.

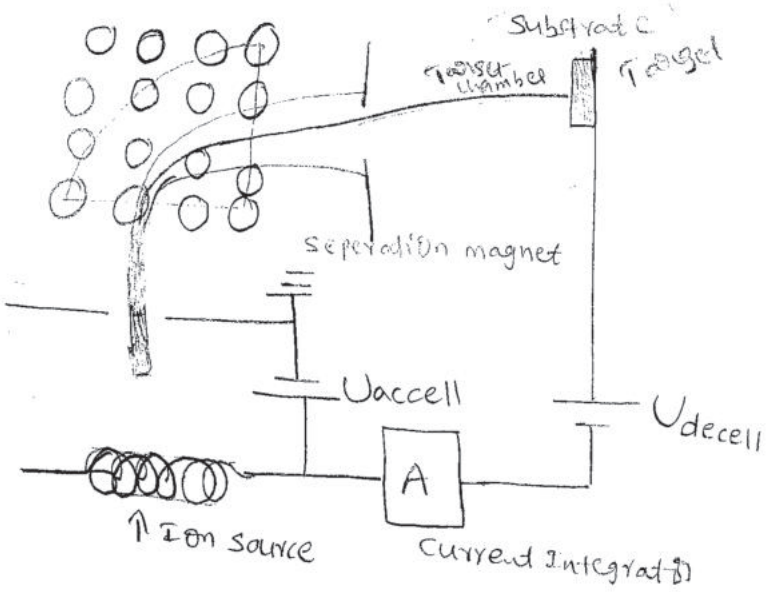
Then F= -D dc/dx

C = Carrier concentration

D = diffusion coefficient (or) diffusivity.

Ion implantation :-

Ion implantation is alternative technique for selective doping of semiconductors. The doping profile is more precisely controlled by this Technique.



Ion implantation setup with mass seperator.

→ Ion implantation equipement typically consists of an ion source, where ions of the desired element are produced, an accelerator where the ions are electrostatically accelerated to a high energy & a target chamber where the ions impinge on a target, which is the material (silicon substrate/ wafer) to be implanted.

→ Each ion is typically a single atom & thus the actual amount of material implanted in the target is the integral over time of the ion current. This amount is called "dose".

→ when these ions enter into the semiconductor, they loose the kinetic energy through a series of collisions with the electrons as well as nuclei of the lattice atoms.

→ Finally the impurity atom comes to rest when its K.E Fall to zero.

[material implanted: $\int_t I(\text{ion current})]$ dose

→ ion energy: 10 to 500 eV

implantation dose $\Rightarrow Q_0 = \frac{It}{q}$ ion beam current density & time t,

Thermal oxidation :— → By varying energy & ion beam, we can change dose

In micro fabrication, "Thermal oxidation" is a way to produce a thin layer of oxide on the surface of a wafer. This is usually carried out in an open tube furnace at the temperature range of 900–1200°C. A single furnace accepts many wafers at the same time, in a specially designed quartz rack called a "boat". The boat enters the oxidation chamber from one side & holds the wafers vertically, beside each other.

Types

① Dry oxidation ② wet oxidation.

Dry oxidation:- If oxidation is carried out in an atmosphere of dry oxygen, it is known as "dry oxidation".

⇒ Oxygen is passed through the tube where it reacts with the silicon to form SiO_2

$$Si(s) + O_2(g) \rightarrow SiO_2(s)$$

Wet oxidation :— If oxidation is carried out in water vapour, it is known as "wet oxidation".

→ high purity de-ionized water kept in quartz bubbler at the inlet of the furnace is heated to a temperature close to the boiling point. High purity oxygen or nitrogen is passed through it so that the gas flowing into the furnace is saturated with the water vapour.

$$Si(s) + 2H_2O(l) \rightarrow SiO_2(s) + 2H_2(g).$$

(liquid)

Kinetic of Thermal oxidation :—

Deal-Grove model :—

The Deal-Grove model mathematically describes the growth of an oxide layer on the surface of a material.

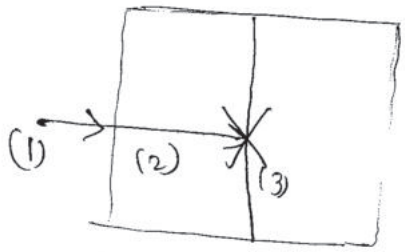


Fig. oxidation process

→ It is used to analyze the thermal oxidation of silicon in semiconductor device fabrication.

→ The model assumes that oxidation occurs at the interface b/w the oxide and substrate, rather than b/w the oxide and the ambient gas.

(1) Indicates, It diffuses from the bulk of the ambient gas to the surface.

(2) It diffuses through the existing oxide layer to the oxide—

(3) It reacts with the substrate.

→ The oxide thickness grown on the silicon by the process of dry & wet oxidation is dependent on the oxidation time and temperature. This thickness can be expressed as a linear-parabolic relation

$$d^2 + Ad = B(t + \tau)$$

d = oxide thickness

A & B are co-efficients that depend on the oxidation time and temperature.

t = oxidation time

τ = parameter for fitting the initial value of the oxide thickness.

For short duration of time, $d \ll A$ then

$$Ad = \frac{B(t+\tau)}{d^2} \quad \text{(we can neglect } d^2\text{)}.$$

$$\frac{B(t+\tau)}{d^2} \simeq B(t+\tau)$$

$$d = \frac{B(t+\tau)}{A}.$$

Above equation says, when oxidation is carried out for short duration of time, the oxide thickness increases linearly with time.

For longer duration of oxidation $(t+\tau) >> \left(\frac{A^2}{4B}\right)$, the oxide thickness can be expressed as,

$$d = \sqrt{Bt}.$$

lithography :-

Semiconductor lithography is a process of drawing patterns on a silicon wafer. The patterns are drawn with a light-sensitive polymer called "photoresist".

→ These patterns define various regions in an Integrated circuit such as implantation regions, contact windows etc.

→ The choice for performing this patterning is "optical lithography". It is basically a photographic process by which the light-sensitive polymer (photoresist) is exposed and developed to form three-dimensional images on the substrate.

Optical lithography systems-

Photomask :- The photomask is an essential component in semiconductor lithography. It contains the detailed blueprint of the designed circuit.

→ Using the photomask, specific images of detailed devices are transferred onto the surface of the silicon wafers by means of "photolithography".

→ A photomask is used just like the negative in photography that captures specific image for later reproduction.

→ In photography, multiple copies of photos are reproduced using the original image captured on the negative.

Likewise a photomask produces duplicate images of patterns onto

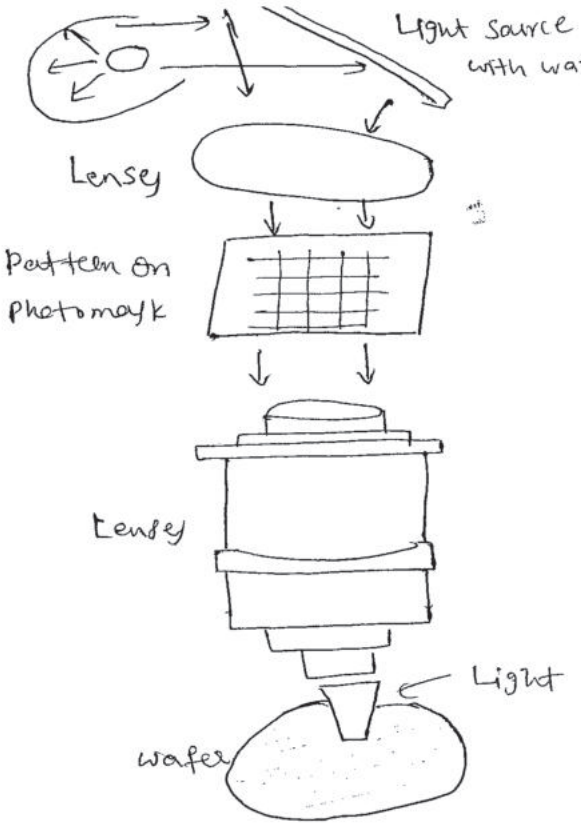


Fig: Optical lithography system.

→ A single photomask plate produces identical images on thousands of wafers.

→ As the quality of the finished photograph is determined by the quality of the original film, the quality of the photomask determines the ultimate quality of semiconductor chips.

→ The material used for building photomasks is a quartz plate upon which detailed images of patterns are formed.

→ The Patterns & image are then transferred onto the wafer surfaces by shining light through the quartz plate.

General methods used to expose photoresist are

① contact photolithography
② Proximity "
③ Projection "

Contact photolithography :-

In this, the mask is kept in direct contact with the substrate and exposure is done.

→ It offers high resolution. But practical problems such as mask damage and resultant low yield make this process unusable in most production environments.

Proximity photolithography :-

→ Similar to contact lithography & except that there is a small gap of a few microns b/w wafer and mask during exposure.

→ This results in optical diffraction which results a limit on the resolution.

Projection photolithography :-

The most common method of exposure is projection printing.

→ An image of the mask is projected onto the wafer.

→ 2 major classes of projection lithography tools

① Scanning
② Step-and-Repeat systems

Electron Beam lithography:

This is a direct analogy with photolithography, where electron beam exposure alters the chemistry of the resist instead of light exposure.

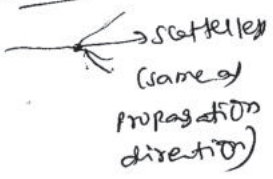
→ The most common EBL resist is Polymethyl — PMMA.

Ion-Beam lithography :-

→ It can achieve higher resolution than optical, X-ray and E beam lithography because ions have higher mass and therefore scatter less than electrons.

→ Ion-beam lithography scans an ion beam across a surface to form a pattern.

Scatter



2 Types of ion-beam lithography

① Scanning focussed beam systems

② mask beam system.

→ Former system is similar to beam lithography in which ion source can be Gas or HT.

→ 2nd method, similar to an optical 5x reduction projection step and repeat system in which project 100 keV light ions such as H_2^+ through a stencil mask.

Metalization :-

In semiconductor process technology, metal refers to a material with very high electrical conductivity.

→ metal is evaporated on the substrate and is patterned into wires whenever a connection is needed on an IC.

→ Aluminium is the most process - friendly metal & has low resistivity.

Group of metal layers

1. metal1, first level of interconnect
2. via1, to connect metal1 & metal2
3. metal2, second level of interconnect
4. via2, to connect metal2 & metal3
5. metal3
6. via3, 3 & 4
7. metal4
8. via4 4 & 5
9. metal5
10. via5 5 & 6
11. metal6
12. via6 6 & 7
13. metal7, 7th level of interconnect.

Encapsulation (Packaging) :-

A package is the housing of a semiconductor chip. It protects and preserves the performance of the semiconductor device from electrical, mechanical & chemical

Corruption or impairment.

→ It electrically interconnects the chip with outside circuitry.

→ It is designed to dissipate heat generated by the chip.

A package is a plastic, ceramic, laminate, or metal seal that encloses the chip or die inside.

Packages classified into 2 categories.

① pin-through-hole (PTH) package:

have pins are inserted into through-holes in the board and soldered in place from the opposite side of the board.

→ The through-hole mounting approach offers a mechanically reliable & sturdy connection.

② surface-mount-technology (SMT) package:-

have leads that are soldered directly to the metal leads on the surface of the circuit board.

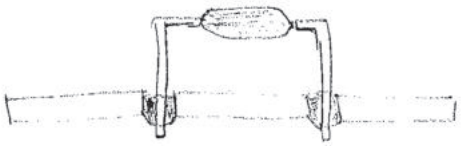
→ Preferred one (SMT)

→ Packing density is increased for the following reasons

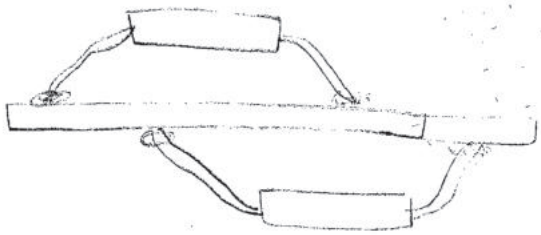
① through holes are eliminated, which provide more wiring space.

② lead pitch is reduced

③ Chips can be mounted on both sides of the board.



① through-hole mounting



surface mount

Resistors :-

Resistors have been available for use in IC's for many years.

→ Some of these are made in silicon, so they are directly integrated with the rest of the IC process.

→ Usually resistors in IC's are characterised in terms of their sheet resistance rather than their absolute resistance value.

→ Sheet Resistance, R_{sheet} is defined as the resistance of a resistive strip with equal length and width

$$R = R_{sheet} * \text{number of squares where } w = L.$$

$\rho \rightarrow$ material resistivity (Ωm)

$t \rightarrow$ thickness (m).

Integrated semiconductor Resistors :-

→ In this category, the existing semiconductor is used as the resistive material.

→ The resistor may be fabricated at a no. of stages during the IC process giving rise to different resistors with different characteristics.

Types

(1) Diffused Resistors :-

Formed during either the base or emitter diffusion of bipolar process.

→

→ For an npn process the base diffusion resistor is a P-type of moderate sheet resistivity typically in the range of 100 – 200 Ω/m. This can provide resistors in 50–10kΩ range.

→ The heavily doped n+ emitter diffusion will produce an n+ type resistor with low sheet resistivity of 2–10 Ω/m. Provide resistors in 1–100Ω range.

Pinched resistors

→ variation to the diffused resistor that is used to increase the sheet resistivity of base region is to use the n+ type emitter as a means to reduce the cross-sectional area of the base region thereby increasing sheet resistivity.

Epitaxial Resistors :–

→ High resistor values can be formed using the epitaxial layer since it has higher resistivity than other regions.

→ These resistors can have sheet resistances around 5kΩ/m.

Mos Resistors

A MOSFET can be biased to provide a non-linear resistor.

→ such a resistor provides much greater value than diffused one while occupying much smaller area.

→ with the gate shortened to the drain in a MOSFET a quadratic relation b/w current and voltage

conduct current only when the

voltage exceeds the threshold voltage.

→ Under these circumstances, the current flowing in this resistor (i.e, the MOSFET drain current) depends on the $\frac{W}{L}\mu_n C_{ox}$ of the channel.

→ To ↑ resistor value, aspect ratio of the MOSFET should be reduced to give longer channel length & narrower channel width.

Capacitors :-

most integrated capacitors are either junction capacitors or MOS capacitors.

Junction capacitors :-

→ Is formed when a P-n junction is reverse-biased.

→ This can be formed using the base-emitter, base-collector or collector-substrate junctions of an npn structure in bipolar IC's.

→ Particular Jn must be maintained in reverse bias to provide desired capacitance.

→ The capacitance is voltage dependent decreasing with increased Reverse bias. Capacitence depends on Reverse voltage.

→ base-emitter Jn provides 1000 pF/mm² highest capacitance per unit with low breakdown voltage. (~5V)

→ base-collector jn provides ~100 pF/mm² with higher breakdown voltage (~40V)

Mos capacitors:-

→ mos capacitors are usually formed as parallel plate devices with a top metallization & high conductivity n+ emitter diffusion as the 2 plates with a thin oxide dielectric sandwiched in b/w.

→ mos capacitors can provide around 1000pF/mm² with breakdown voltage upto 100V.

→ These are voltage independent & can be biased either positively or negatively.

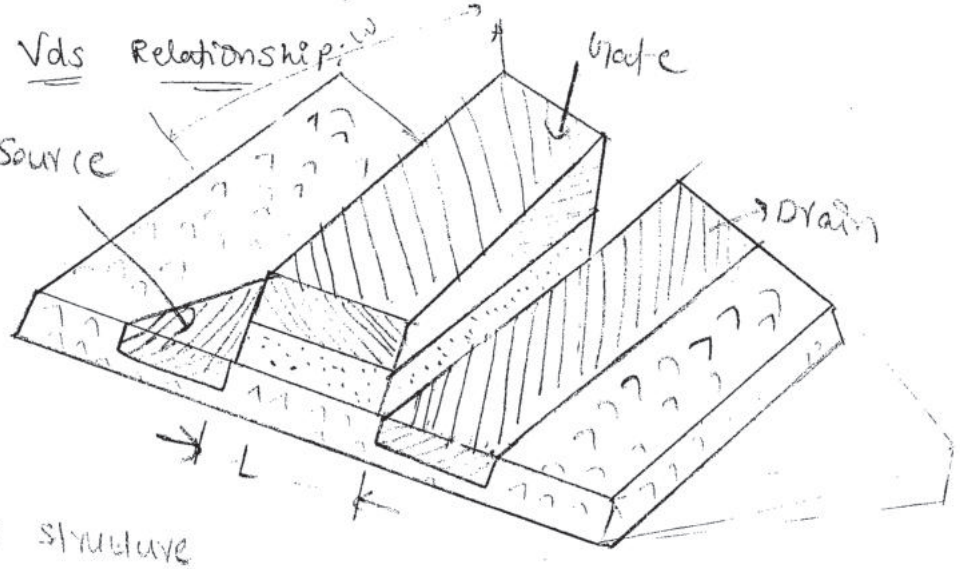
A wafer prober is a machine used to test IC's.

→ For electrical testing a set of microscopic contacts or probes called a "Probe card" are held in place whilst the wafer, vacuum-mounted on a wafer chuck, is moved into electrical contact.

→ The wafer prober is responsible for loading or unloading the wafers from their carrier and is equipped with automatic pattern recognition optics capable of aligning the wafer with sufficient accuracy to ensure accurate Registration b/w contact pads on the wafer & the tips of the probes.

BASIC ELECTRICAL PROPERTIES

Ids Vs Vds Relationship:



nmos structure

The whole concept of mos Tr involves from the use of a voltage on the gate to ~~include~~ induce a charge in the channel b/w source and drain which may then be caused to move from source to drain under the influence of an electric field created by voltage Vds applied b/w drain and source. Since the charge induced is dependent on gate to source voltage Vgs then Ids is dependent on both Vgs & Vdy.

$$I_{ds} = -I_{sd} = \frac{\text{charge induced in channel } (Q_c)}{\text{Transit Time } (\tau)}$$

$$\text{Transit time} \Rightarrow \tau_{sd} = \frac{\text{Length of channel } (L)}{\text{Velocity } (v)}$$

$$\text{Velocity } V = \mu E_d$$

$\mu \leftarrow$ mobility $cm^2/V\text{-}s$

E_{dy} = electric field (drain to source)

(B) $E_{dy} = \frac{V_{dy}}{L}$

So, $V = \frac{\mu V_{dy}}{L}$

$\tau_{sd} = \frac{L^2}{\mu V_{dy}}$

$\mu_n = 650 \text{ cm}^2/\text{V-sec}$

$\mu_p = 240 \text{ cm}^2/\text{V-sec}$

NON – Saturation region: –

Charge induced in channel due to gate voltage is due to the voltage difference b/w gate and channel, V_{gs}.

→ Voltage along channel varies linearly with distance from source due to IR drop in channel & assuming that device is in non-saturation region. then drain side voltage $\frac{V_{dy}}{2}$.

Effective gate voltage $(V_g) = V_{gs} - V_t$.

V_t → voltage needed to invert charge under the gate & established channel.

$Q = C_g V$.

Potential difference $V = \underbrace{V_{gs} - V_t}_{\text{source}} - \underbrace{\frac{V_{dy}}{2}}_{\text{drain}}$

$C_g = \frac{\epsilon_0 \epsilon_r A}{D}$

$$Q_c = \frac{\epsilon_0 \epsilon_r A}{D} \left(V_{gs} - V_t - \frac{V_{ds}}{2} \right)$$ (b)

$$I_d = \frac{Q_c}{T} \quad \text{(c)}$$

$$= \frac{\frac{\epsilon_0 \epsilon_r A}{D} \left(V_{gs} - V_t - \frac{V_{ds}}{2} \right)}{\frac{L^2}{\mu V_{ds}}}$$

$$= \frac{\epsilon_0 \epsilon_r \omega L}{D} \times \frac{\mu V_{ds}}{L^2} \left(V_{gs} - V_t - \frac{V_{ds}}{2} \right) \quad (A = \omega L)$$

$$= \frac{\epsilon_0 \epsilon_r w}{D} \times \frac{\mu V_{ds}}{L} \left((V_{gs} - V_t) - \frac{V_{ds}}{2} \right)$$

$$I_d = \frac{\epsilon_0 \epsilon_r w}{DL} \cdot \mu \left((V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right)$$

when $V_{ds} < V_{gs} - V_t$

let $k = \frac{\epsilon_0 \epsilon_r \mu}{D}$

$$I_d = \frac{kw}{L} \left((V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right)$$

Let $\beta = \frac{kw}{L}$

$$\therefore I_d = \beta \left((V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right)$$

$$k = \frac{\epsilon_0 \epsilon_r \mu}{D} \times \frac{A}{A}$$

Sub $k = \frac{\epsilon_g \mu}{\omega L}$ in I_d

$$I_d = \frac{C_g \mu}{\omega L} \cdot \frac{w}{L} \left((V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right)$$

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

Gate capacitance / unit area C_o

(D)

$$C_o = \frac{C_g}{WL}$$

$$C_g = C_o \cdot WL$$

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

Saturation region :—

$$V_d = V_{gs} - V_t$$

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

$$= \beta\left(V_d \cdot V_d - \frac{V_d^2}{2}\right)$$

$$= \beta\left(\frac{V_d^2}{2}\right)$$

$$= \frac{\beta}{2}(V_{gs} - V_t)^2$$

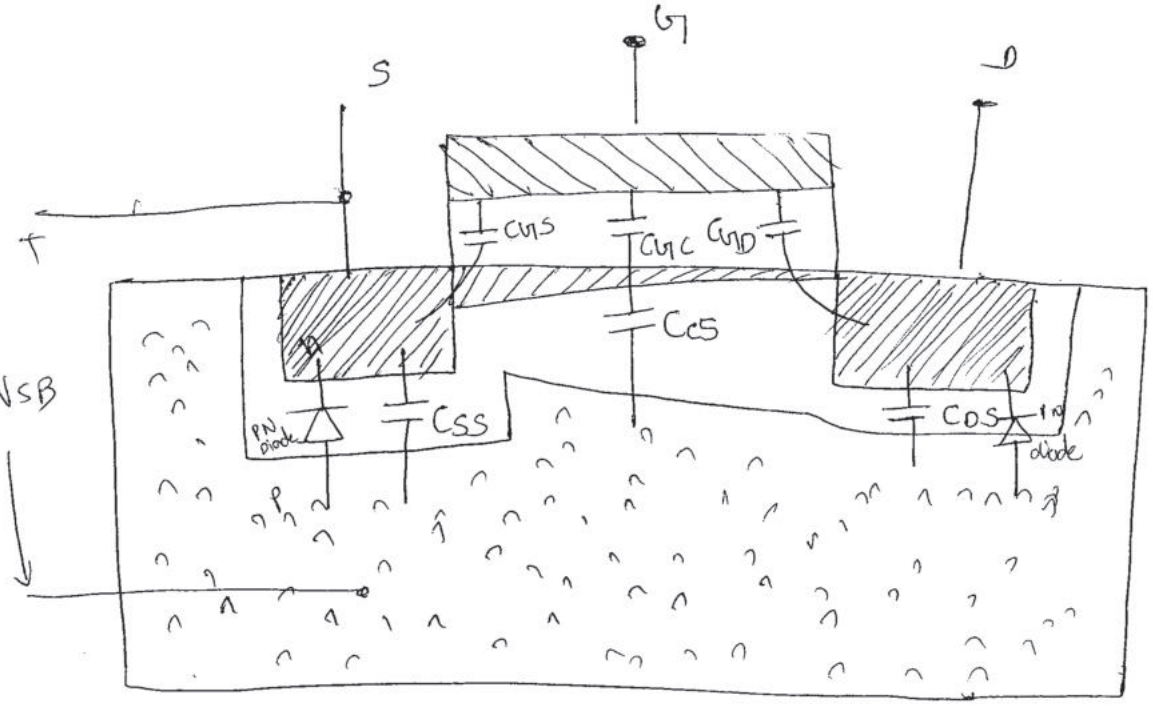
(8) $$I_d = \frac{C_g \mu}{2L^2}\left(\frac{V_d^2}{2}\right)$$

$$\text{or } I_d = \frac{C_o \mu W}{2L}(V_{gs} - V_t)^2$$

For both enhancement mode & depletion mode.

MOS Tr^y circuit model :-

→ The MOS Tr^y can be modeled with varying degrees of complexity. (X) consideration of actual physical construction of the device leads to some understanding of various components of the model.



nmos Tr^r model

C_{GC} → gate to channel capacitance
C_{GS} → gate to source
C_{GD} → gate to drain

} small for self-aligning mod process.

Remaining capacitances are effected with depletion layer & voltage dependent.

C_{SS} → source to substrate capacitance
C_{DS} → drain to substrate
C_{CS} → channel to substrate

(4)

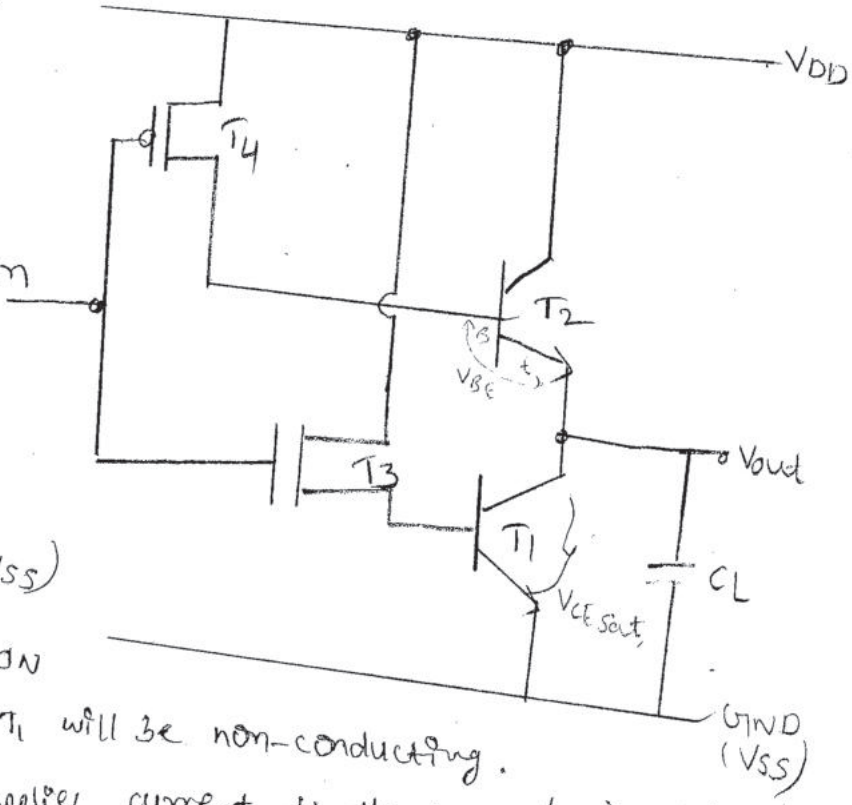
Bi-cmos inverter:—

→ It consists of 2 Bi-Polar Transistors T_1, T_2.

→ And 1 nmos T_{8}^{γ} – T_3,

1 Pmos T_{8}^{γ} – T_4.

both are enhancement mode $T_{8}^{\gamma 1}$ s.



Case 1:- Logic 0 i/p.

With Vin = 0 (VSS)

T_3 → off & T_4 – ON

So, As T_3 is off T_1 will be non-conducting.

→ T_4 is on and supplies current to the base of T_2 which will conduct & act as a current source to charge the load C_L toward +5V (VDD). [a current source to charge the load C_L towards]

The o/p of inverter will rise to +5V. less the Base-to-emitter voltage V_{BE} of T_2. ie. $V_{DD} - V_{BE}$ = Logic 1

Case 2; With Vin = +5V (VDD)

T_4 → off so that T_2 is non-conducting.

T_3 is on & will supply current to the base of T_1 which will conduct & act as a current sink to the load C_L discharging it toward 0V.

The o/p of inverter will fall to 0V plus the saturation voltage V_{CEsat} from the collector to the emitter of T_1.

(2) i.e., $0V + V_{CEsat}$ = Logic 0.

→ T_1 & T_2 will present low impedance when turned on into saturation and the load C_L will be charged & discharged rapidly.

→ The o/p logic levels will be good and will be close to the rail voltages (VDD, Vss) since V_{CEsat} is quite small & $V_{BE} \cong +0.7V$.

$(V_{ss})\,0 \cong V_{CEsat}$ $\quad 5 - 0.7 \cong V_{DD}$.

→ The inverter has high i/p impedance.

→ The inverter has low o/p impedance.

→ The inverter has high noise margins.

→ Due to the presence of a DC path from VDD to GND through T_3 & T_1, this is not a good arrangement to implement since there will be significant static current flow whenever V_{in} = logic 1.

→ Problem: There is no discharge path for current from the base of either BJT when it is being turned off. This will slow down the action of this circuit.

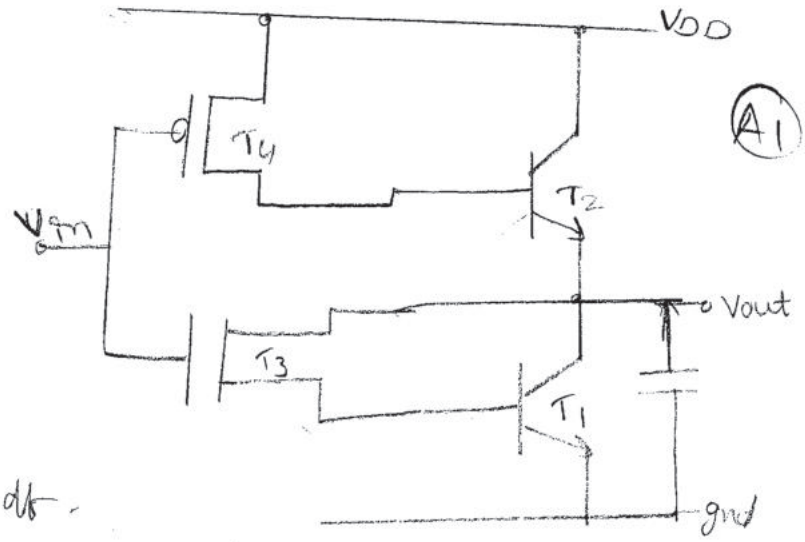
TO avoid this, improved version of the circuit used:

⇒ DC path through T_3 & T_1 is eliminated.

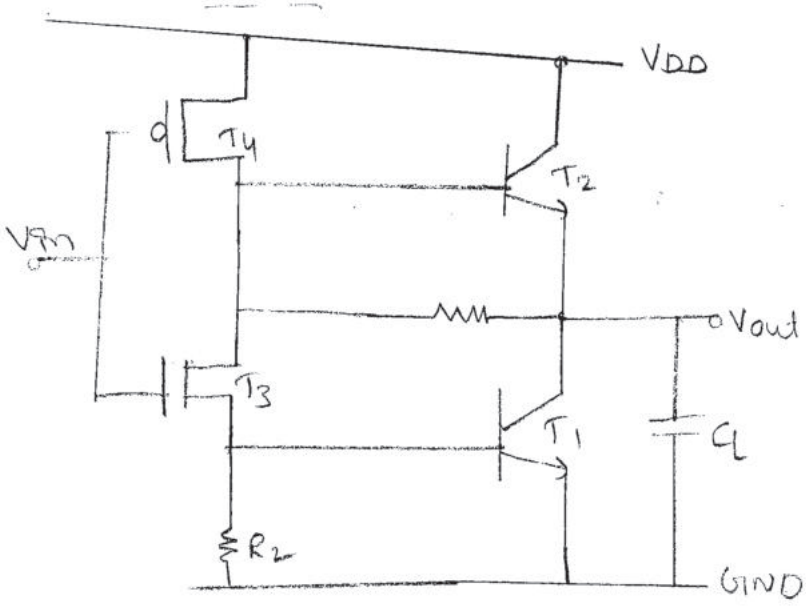
⇒ But, o/p voltage swing is now reduced, since the o/p

cannot fall below the V_{BE} of T_1. (B)

→ An improved inverter arrangement with resistor provide the improved swing of o/p voltage when each Bipolar Tr is off & also provide discharge path for base current during turn-off.

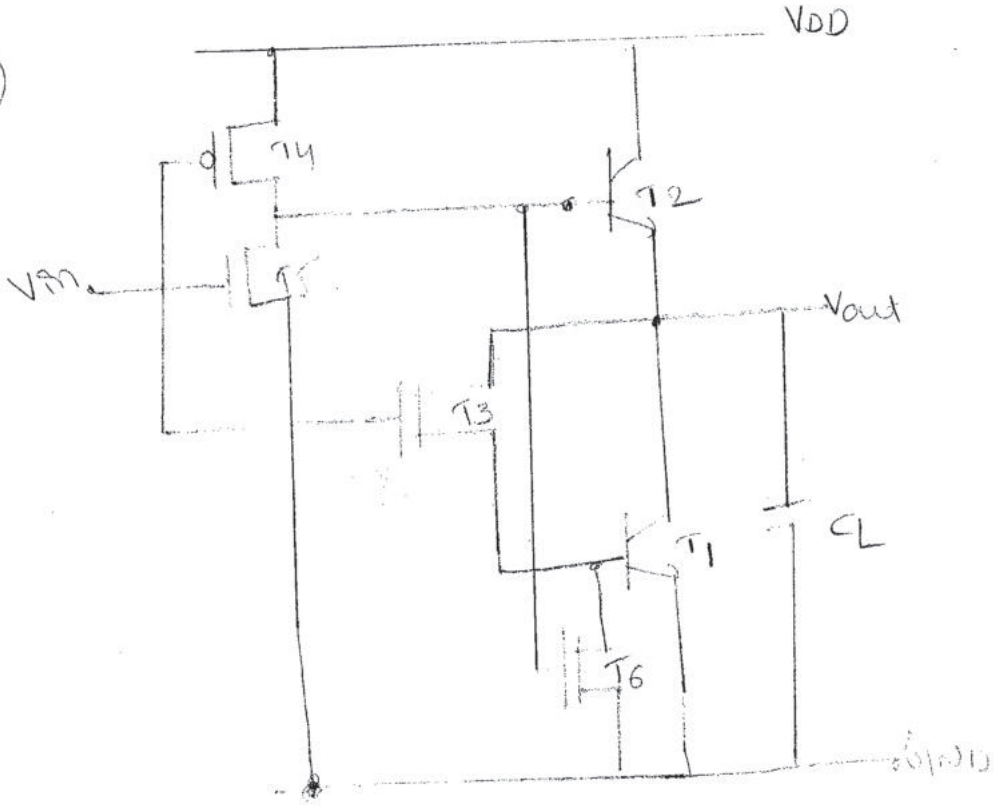


(A1)

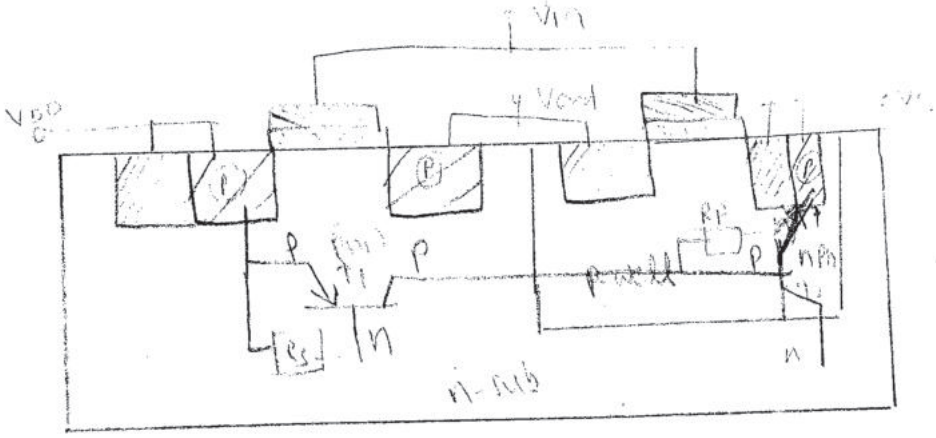


→ The provision of On-chip resistor of suitable value is not always convenient and may be space-consuming. To avoid this, another arrangement in that, T_5 & T_6 are arranged to turn on when T_2 & T_1 respectively are being turned off.

A2



Latch up in cmos circuits :-



→ A problem which is inherent in p-well & n-well process due to the large no. of junctions which are formed in these structures. The consequent presence of parasitic T's & diodes.

Latch up : is a condition in which the parasitic component give rise to the establishment of low-resistance conducting path between Vdd & Vss.

Problems :-

(1) Find g_m and r_{ds} for an n-channel transistor with $V_{gs} = 1.2V$, $V_{tn} = 0.8V$, $\frac{W}{L} = 10$, $\mu_n C_{ox} = 92 \mu A/V^2$ & $V_{ds} = V_{eff} + 0.5V$. The output impedance constant $= 95.3 \times 10^{-3}/V$.

Sol:

$$\lambda = 95.3 \times 10^{-3}/V$$

$$V_{ds} = V_{eff} + 0.5V$$

where

$$V_{eff} = V_{gs} - V_{tn}$$

$$= 1.2 - 0.8$$

$$= 0.4V$$

$\therefore V_{ds} = 0.4 + 0.5 = 0.9V$

$0.9 > 0.4$

$V_{ds} > V_{gs} - V_{tn}$

So, Tr is in saturation region,

$$I_d = \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{gs} - V_{tn})^2$$

Transconductance, $g_m = \frac{C_g \mu}{L^2} V_{ds}$

$C_g = C_{ox} WL$ (C_{ox} = unit capacitance)

$\therefore g_m = \mu_n C_{ox} WL$ V_{ds}

$= \mu_n C_{ox} \frac{W}{L} V_{ds}$

$= 92 \times 10^{-6} \times 0.9 \times 10$

$\boxed{g_m = 8.28 \times 10^{-4} \, \mho}$

Drain to Source resistance, $r_{ds} = \frac{1}{\partial I_{ds}}$

$I_{ds} = \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{gs} - V_{th})^2$

$= \frac{92 \times 10^{-6}}{2} \times 10 \, (0.4)^2$

$= 7.36 \times 10^{-5} \, A$

$\therefore r_{ds} = \frac{1}{95.3 \times 10^{-3} \times 7.36 \times 10^{-5}}$

$= 142.57 \, k\Omega$

If $\beta_n = \beta_p$ & $V_{tn} = -V_{tp}$

then

$$V_{in} = \frac{V_{DD} + V_{tp} - V_{tp}\left(\sqrt{\frac{\beta_n}{\beta_n}}\right)}{1 + \left(\sqrt{\frac{\beta_n}{\beta_p}}\right)}$$

$$V_{in} = \frac{V_{DD}}{1+1} = \frac{V_{DD}}{2} = 0.5 V_{DD}.$$

→ change over b/w logic levels is symmetrically disposed about the point at which

$$\boxed{V_{in} = V_{out} = 0.5 V_{DD}}$$

since only at this pt will the 2 'β' factors be equal.

But for $\beta_n = \beta_p$, the device geometries must be such that

$$\beta_n = \beta_p$$

$$\frac{\text{Electron mobility} \cdot W_n}{\text{oxide} \cdot L_n} = \frac{\text{Electron mobility} \cdot W_p}{L_p}$$

$$\frac{\mu_n W_n}{L_n} = \frac{\mu_p W_p}{L_p}.$$

we know, $\mu_n = 2.5 \mu_p$.

$$\frac{2.5 \mu_p W_n}{L_n} = \frac{\mu_p W_p}{L_p}$$

$$\boxed{\frac{W_p}{L_p} \simeq 2.5 \frac{W_n}{L_n}}$$

width to length ratio of p-device is to be 2 to 3 time that of n-device.

→ mobility μ affected by transverse electric field [illegible]

dependent on V_{gs}.

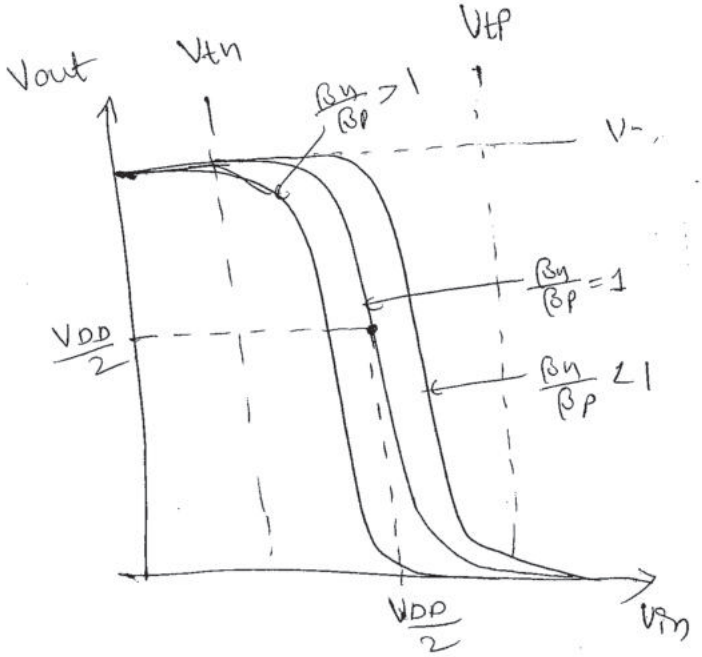
actual mobility if, $\mu = \mu_Z (1 - \phi (V_{gs} - V_t))^{-1}$

(a) $\phi \rightarrow$ constant $\cong 0.05$.

$\mu_Z \rightarrow$ mobility with zero transverse field.

β ratio of 1 ($\frac{\beta_n}{\beta_p} = 1$) will only hold good around the pt of symmetry when $V_{out} = V_M = 0.5 V_{DD}$.

→



Trends in transfer c/v with β ratio

Region 4 :- the input voltage of P-Tr exceed threshold voltage. The P-Tr conducts but large voltage b/w source and drain.

So, P-Tr saturation,

→ The n-Tr conducting with a small voltage across it, it operates in resistive region.

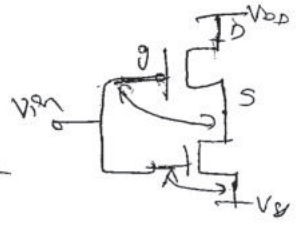
→ Currents in region 2 & 4 are small. Most of energy consumed in switching from one state to another.

In region 3 both Tr's are in saturation. The currents in each device must be same since Tr's are in series,

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

$$I_{ds} = \frac{\beta}{2}(v_{gs} - V_t)^2$$

$$I_{dsp} = \frac{\beta_p}{2}(V_{in} - V_{DD} - V_{tp})^2$$



& $I_{dsn} = \frac{\beta_n}{2}(V_{in} - V_{SS} - V_{tn})^2$

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

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

$$V_{in} - V_{DD} - V_{tp} = -\sqrt{\frac{\beta_n}{\beta_p}}(V_{in} - V_{tn})$$

$$V_{in} + \sqrt{\frac{\beta_n}{\beta_p}} V_{in} = V_{DD} + V_{tp} + \sqrt{\frac{\beta_n}{\beta_p}} V_{tn}$$

$$V_{in} = \frac{V_{DD} + V_{tp} + V_{tn}\sqrt{\frac{\beta_n}{\beta_p}}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}$$

CMOS inverter :—

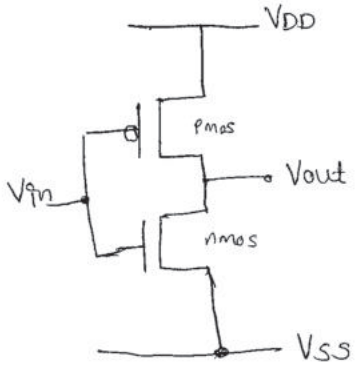
we have seen,
current - voltage relationships for MOS Tr's,

$$Ids = \frac{kw}{L}\left((Vgs - Vt)Vds - \frac{Vds^2}{2}\right)$$

In resistive region or non-saturation region,

$$Ids = \frac{kw}{L}\frac{(Vgs - Vt)^2}{2}$$

In saturation region.



(5)

circuit

→ In both cases, the factor K is a technology - dependent parameter such that;

$$K = \frac{\epsilon_0 \epsilon_r \mu}{D}.$$

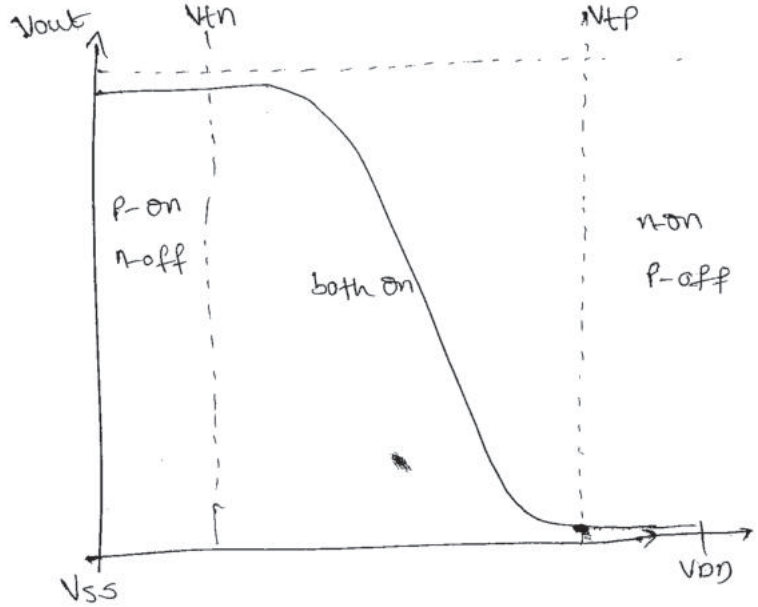
& $$\beta = \frac{kw}{L}$$

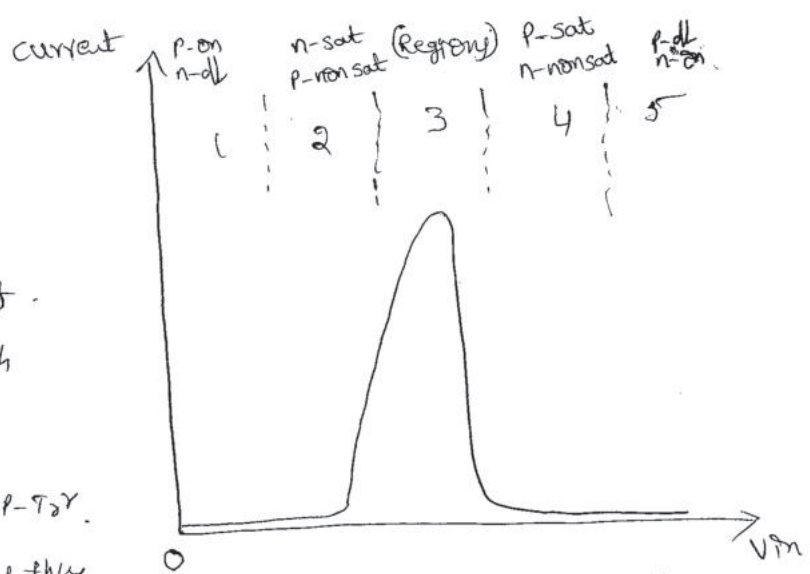
β may be applied to both nmos & pmos Tr's as follows,

$$\beta_n = \frac{\epsilon_0 \epsilon_r \mu_n}{D}\frac{w_n}{L_n}$$

$$\beta_p = \frac{\epsilon_0 \epsilon_r \mu_p}{D}.\frac{w_p}{L_p}$$

where ~~β_n & β_p~~ w_n, L_n & w_p, L_p are n-Tr & p-Tr dimensions





In region 1 :—

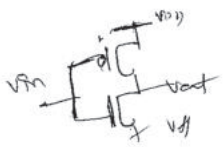
For which Vin = Logic 0.

We have P-Tr fully Turned On while n-Tr fully Turned off.

Thus no current flows through inverter & o/p is directly connected to VDD through P-Tr.

A good logic 1 o/p voltage is thus present at the o/p.

Vin	P	n
Logic 0	on	dl
i2>Vtp	non-sat	sat



In region 5 :—

Vin = Logic 1, the n-Tr is fully on while P-Tr is fully off. Again, no current flows & good logic 0 appears at o/p.

In region 2 :— i/p voltage has increased to a level which just exceeds the threshold voltage of n-Tr. ($V_{in} > V_{tn}$), ie, it is in saturation region. ($V_{ds} \geq V_{gs} - V_t$)

The n-Tr conducts & a large voltage b/w source & drain (V_{ds}).

→ The P-Tr also conducts but a small voltage across it, ($V_{ds} < V_{gs} - V_t$)

it is in non-saturated or resistive region.

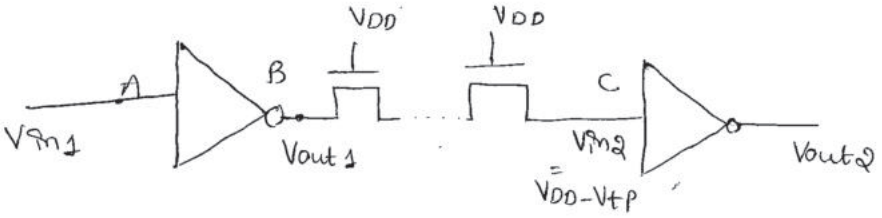
→ A small current flows through the inverter from VDD to Vss.

→ If we with to analyze behaviour in this region,

P-device non-saturation current = n device saturation current.

In region 3 :— is the region in which inverter exhibits gain & both Tr's are in saturation region.

Pull-up to Pull-down ratio for an nmos inverter driven through one or more pass transistors :-



Input 2 to inverter 2 comes from the o/p of inverter 1 but passes through one or more nmos transistors used as switches in series (Pass Tr)

By the connection of Pass Transistors in series will degrade the logic 1 level into inverter 2 so that o/p will not be a proper logic 0 level..

When Vin1 = 0V (point A), the o/p of 1st inverter is VDD (Pt B). The voltage into inverter 2 at point C is reduced from VDD by voltage of series pass Transistor.

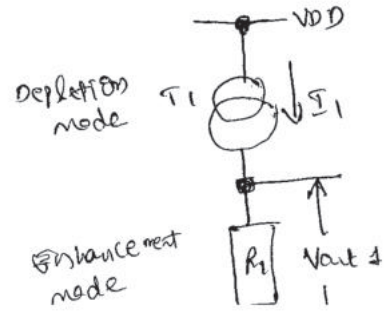
The i/p voltage to inverter 2 $V_{in2} = V_{DD} - V_{tp}$.

V_{t1} → Threshold voltage of a Pass Tr.

Consider inverter 1 with i/p VDD. If i/p is at VDD, then Pd Tr is conducting but with a slow voltage across it.

∴ It is in its resistive region (non-saturation) represented by R_1.

meanwhile the Pu Tr is in saturation & is represented by current source.



$I_{ds} = \frac{k w}{L}\left((V_{gs}-V_t)V_{ds} - \frac{V_{ds}^2}{2}\right)$

For P.d Tr, $(V_{ds} < V_{gs}-V_t)$

$$I_{ds} = \frac{k' W_{pd1}}{L_{pd1}}\left((V_{DD}-V_t)V_{ds1} - \frac{V_{ds1}^2}{2}\right) \quad (V_{in} = V_{DD})$$

(P)

$$\frac{I_{ds}}{V_{ds1}} = \frac{k W_{pd1}}{L_{pd1}}\left((V_{DD}-V_t) - \frac{V_{ds1}}{2}\right)$$

$$R_1 = \frac{V_{ds1}}{I_{ds}} = \frac{L_{pd1}}{k W_{pd1}}\left[\frac{1}{V_{DD}-V_t-\frac{V_{ds1}}{2}}\right] \quad (V_{ds} < V_{gs}-V_t)$$

V_{ds} small

V_{ds1} is small & $\frac{V_{ds1}}{2}$ we can ignore.

$$R_1 = \frac{1}{k}\cdot Z_{pd1}\left[\frac{1}{V_{DD}-V_t}\right] \quad \left(\because \frac{L_{pd1}}{W_{pd1}} = Z_{pd1}\right) \quad \frac{V_{ds1}}{2} \to \text{ignored.}$$

——— (1)

For depletion mode P.U in saturation with $V_{gs}=0$,

$$I_1 = I_{ds} = k\frac{W_{PU1}}{2L_{PU1}}\left[V_{gs}-V_{td}\right]^2 \qquad \left[I_{ds} = \frac{kw}{L}\frac{(V_{gs}-V_t)^2}{2}\right]$$

But $V_{gs}=0$.

$$I_1 = k\frac{W_{PU1}}{L_{PU1}}\frac{(-V_{td})^2}{2}$$

$$I_1 = k\cdot\frac{1}{Z_{PU1}}\frac{(-V_{td})^2}{2} \quad \text{——— (2)}$$

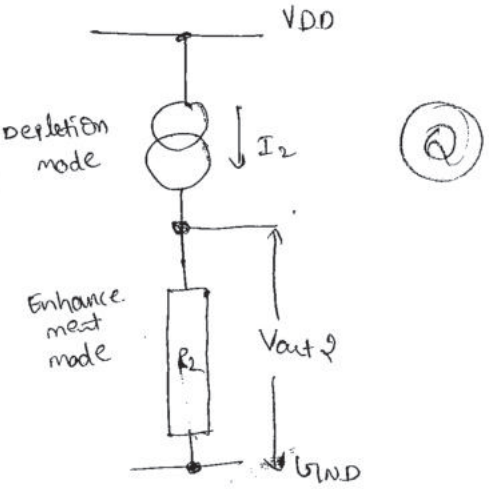
sub (1) & (2) in below eqn

$$V_{out1} = I_1 R_1$$

$$V_{out1} = \frac{k}{Z_{PU1}}\frac{(-V_{td})^2}{2} \times \frac{1}{k} Z_{pd1}\left[\frac{1}{V_{DD}-V_t}\right]$$

$$V_{out1} = \frac{Z_{pd1}}{Z_{PU1}}\left(\frac{1}{V_{DD}-V_t}\right)\left(\frac{-V_{td}}{2}\right)^2 \quad \text{——— (3)}$$

CONSIDER INVERTER 2



Inverter 2 with i/p = $V_{DD} - V_{tp}$

Input $V_{DD} - V_{tp}$.

For p-d Tr, $I_{d2} = \frac{k w_{pd}}{L_{pd}} \left(((V_{DD} - V_{tp}) - V_t) V_{ds2} - \frac{V_{ds2}^2}{2} \right)$

$$R_2 = \frac{1}{k} \cdot Z_{pd2} \cdot \frac{1}{((V_{DD} - V_{tp}) - V_t)} \quad \text{—(4)}$$

$$I_2 = k \cdot \frac{1}{Z_{pu2}} \frac{(-V_{td})^2}{2} \quad \text{—(5)}$$

$$V_{out2} = I_2 R_2$$

$$= k \frac{1}{Z_{pu2}} \frac{(-V_{td})^2}{2} \cdot \frac{1}{k} Z_{pd2} \cdot \frac{1}{V_{DD} - V_{tp} - V_t}$$

$$V_{out2} = \frac{Z_{pd2}}{2 Z_{pu2}} \cdot \frac{(-V_{td})^2}{V_{DD} - V_{tp} - V_t} \quad \text{—(6)}$$

If Inverter 2 is to have same o/p voltage under these conditions then $V_{out1} = V_{out2}$.

$$I_1 R_1 = I_2 R_2$$

(3) = (6)

$$\frac{Z_{pd1}}{Z_{pu1}} \frac{(-V_{td})^2}{2} \frac{1}{V_{DD} - V_t} = \frac{Z_{pd2}}{Z_{pu2}} \frac{(-V_{td})^2}{2 (V_{DD} - V_{tp} - V_t)}$$

$$\frac{Z_{pu1}}{Z_{pd1}} (V_{DD} - V_t) = \frac{Z_{pu2}}{Z_{pd2}} (V_{DD} - V_{tp} - V_t)$$

Typical values,

$V_t = 0.2 V_{DD}$

$V_{tp} = 0.3 V_{DD}$

(R)

$$\frac{Z_{pu2}}{Z_{pd2}} = \frac{Z_{pu1}}{Z_{pd1}} \cdot \frac{(V_{DD} - V_t)}{(V_{DD} - V_{tp} - V_t)}$$

$$\frac{Z_{pu2}}{Z_{pd2}} = \frac{Z_{pu1}}{Z_{pd1}} \cdot \frac{V_{DD} - 0.2V_{DD}}{V_{DD} - 0.3V_{DD} - 0.2V_{DD}}$$

$$= \frac{Z_{pu1}}{Z_{pd1}} \cdot \frac{0.8V_{DD}}{0.5V_{DD}}$$

$$= \frac{Z_{pu1}}{Z_{pd1}} \times (1.6)$$

$$\simeq \frac{4}{1} \times 2 \qquad (1.6 \simeq 2).$$

$$\frac{Z_{pu2}}{Z_{pd2}} \simeq \frac{8}{1}$$

$$\frac{Z_{pu1}}{Z_{pd1}} = \frac{4}{1}$$

Summarizing for nmos inverter

1) An inverter driven directly from o/p of another should have a Z_{pu}/Z_{pd} ratio of $\geq \frac{4}{1}$

2) An inverter driven through one or more pass transistors should have $\frac{Z_{pu}}{Z_{pd}}$ ratio of $\geq \frac{8}{1}$.

Aspects of MOS Tr threshold Voltage (V_t):-

(5)

The threshold voltage V_t may be expressed as,

$$V_t = \phi_{ms} \frac{Q_B - Q_{ss}}{C_o} + 2\phi_{fN}$$

Q_B = charge per unit area in depletion layer beneath the oxide.

Q_{ss} = charge density at $Si : SiO_2$ interface.

C_o = capacitance per unit gate area.

ϕ_{ms} = work fn difference b/w gate and Si.

ϕ_{fN} = fermi level potential b/w inverted surface & bulk Si.

For polySi gate & Si substrate,
ϕ_{ms} is -ve. but negligible.

the magnitude, sign of V_t determined by negative term $-\frac{Q_{ss}}{C_o}$ & remaining +ve terms.

To evaluate V_t, each term is determined as follows:

$$Q_B = \sqrt{2\epsilon_0 \epsilon_{si} q N (2\phi_{fN} + V_{SB})} \text{ Coulomb/m}^2.$$

$$\phi_{fN} = \frac{KT}{q} \ln \frac{N}{n_i} \text{ volts.}$$

$$Q_{ss} = (1.5 \text{ to } 8) \times 10^{-8} \text{ Coulomb/m}^2.$$

Where V_{SB} = substrate bias voltage
(-ve w.r.t source for nmos
+ve " " pmos)

$q = 1.6 \times 10^{-19}$ Coulomb.

$\epsilon_{ins} = 6 \text{ or } 4$ for SiO_2

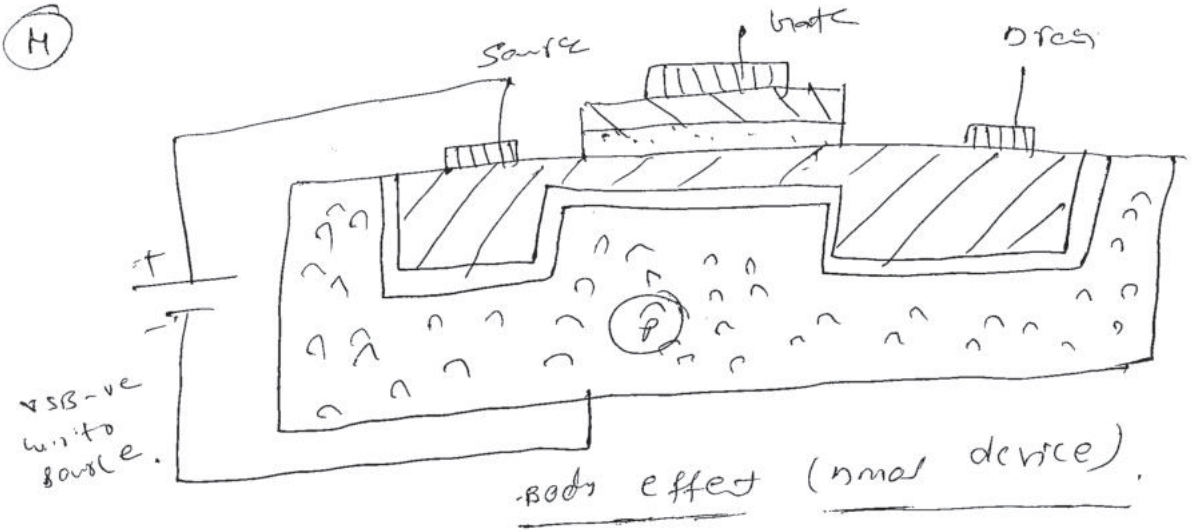
N = Impurity concentration in substrate (N_A or N_D)

ϵ_{si} = relative permittivity of Si

n_i = intrinsic concentration ($1.6 \times 10^{10}/cm^3$ at 300°K)

k = Boltzmann's constant = 1.4×10^{-23} joule/°K.

(H)



Body effect (nmos device).

Body effect may also be taken into account since substrate may be biased w.r.t source.

change in Vt : $\Delta V_t \cong \gamma(\sqrt{V_{SB}})$

γ is constant depends on substrate doping.

So, that more lightly doped the substrate, smaller will be body effect

we may write,

$$V_t = V_t(0) + \left[\frac{D}{\varepsilon_{ins} t_o}\right)\sqrt{2\varepsilon_o \varepsilon_{si} q N} \cdot \sqrt{V_{SB}}$$

$V_t(0)$ → Threshold voltage for $V_{SB} = 0$.

For nmos enhancement mode type

$V_{SB} = 0V$, $V_t = 0.2\, V_{DD}$ (1V for $V_{DD} = 5V$)

$5 \times 0.2 = 5 \times \frac{2}{10} = 1V$

$V_{SB} = 5V$, $V_t = 0.3\, V_{DD}$ [1.5V for $V_{DD} = 5V$)

For PMOS negative values.

nmos depletion mode :-

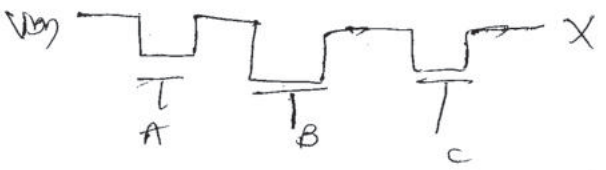
$V_{SB} = 0V$, $V_t = -0.7\, V_{DD}$ (-3.5 V for $V_{DD} = 5V$)

Pass Tr :-

Unlike bipolar Tr, the isolated nature of the gate allows MOS Tr to be used as switches in series with lines carrying logic levels. (1)

This application of MOS device is called as Pass Tr and switching logic array can be formed.

EX: AND array



$X = A.B.C$ (Logic 1 = $V_{DD} - V_t$).

$\overline{X} = ?$ $\qquad \overline{X} = \overline{ABC}$

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

$=$

A assumes gnd potential when $A+B+C=0$.

fig:- pass Tr AND gate

Mos ckt figure of merit (ω_0):—

An indication of freq response may be obtained from the parameter ω_0.

$$\frac{1}{\text{Time}} \propto g_m = \omega_0 = \frac{g_m}{C_g} = \frac{\mu}{L^2}(V_{gs} - V_t) = \frac{1}{\tau_{sd}}.$$

$$\begin{cases} \tau_{sd} = \dfrac{L^2}{\mu V_{ds}} \\ \tau_{sd} = \dfrac{L^2}{\mu (V_{gs} - V_t)} \\ \dfrac{1}{\tau_{sd}} = \dfrac{\mu (V_{gs} - V_t)}{L^2} \end{cases}$$

less time → high speed.

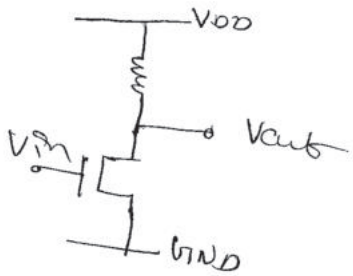
(1) switching speed depends on gate voltage (V_{gs}) above threshold and on carrier mobility (μ) & inverse of square of channel length ($\frac{1}{L^2}$)

For ckt required that g_m be as high as possible.

→ e mobility less ≅ 3.4p.

Surface mobility is also dependent on effective gate voltage $V_{gs} - V_t$.

The Basic inverter ckt requires a Tr with source connected to gnd and a load resistor of some sort connected from drain to the positive supply VDD. The o/p is taken from drain and the i/p is applied b/w gate and gnd.



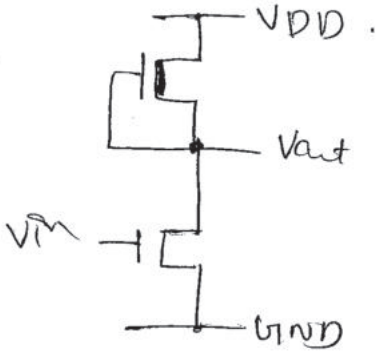
Basic inverter.

Resistors are not conveniently produced on Si substrate. Even modest values occupy excessively large areas so that some other form of load resistance is required.

Convenient way to solve this problem is to use a a depletion mode Tr as load.

(1) With no current drawn from o/p, the current Ids for both Trs must be equal.

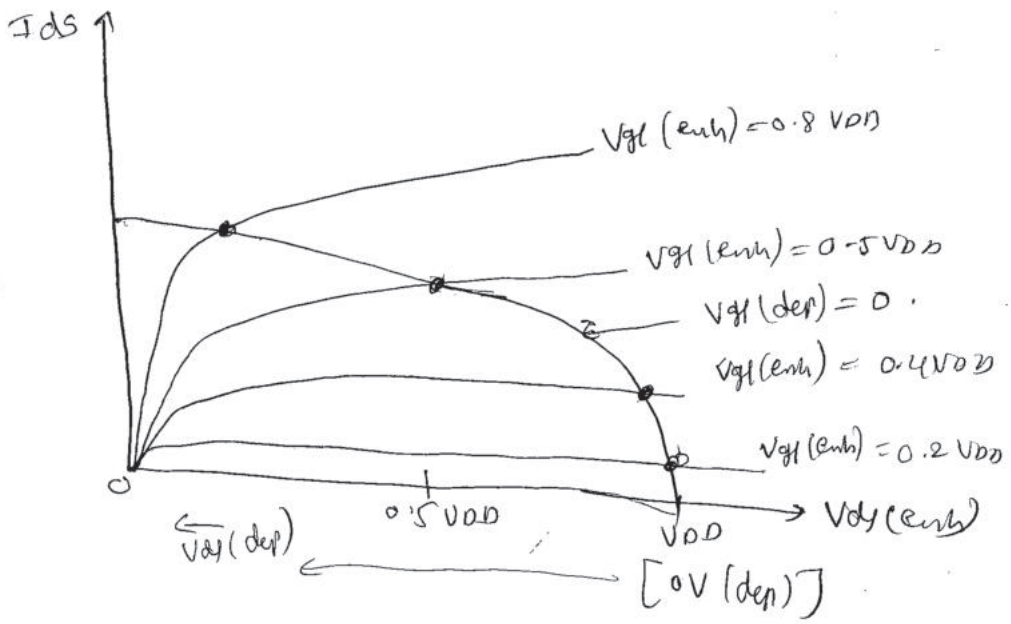
(2) For depletion mode Tr, the gate is connected to source so it is always on.



(3) depletion mode Tr is called pull up (PU) & enhancement mode Tr " pull down (PD)

(4) To obtain inverse transfer ch's we superimpose the $V_{gs} = 0$, depletion mode ch curve on the family of curves for enhancement mode device, noting that max. voltage

across the ...



(L)

$V_{ds}(enh) = V_{DD} - V_{ds}(dep)$
$= V_{out}$.

$V_{gs}(enh) = V_{in}$

→ V_{in} ($= V_{gs}$ P.d Tr) exceeds the P.d threshold voltage current begins to flow. $(V_{gs} > V_t)$

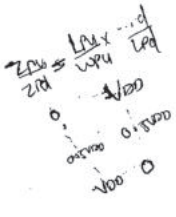
The o/p voltage thus decreases and subsequent increase in V_{in} will cause P.d Tr to come out of saturation and become resistive. $(V_{ds} < V_{gs} - V_t)$

i/p volts ↑ → o/p V ↓

Slope of Transfer ch's determines gain.

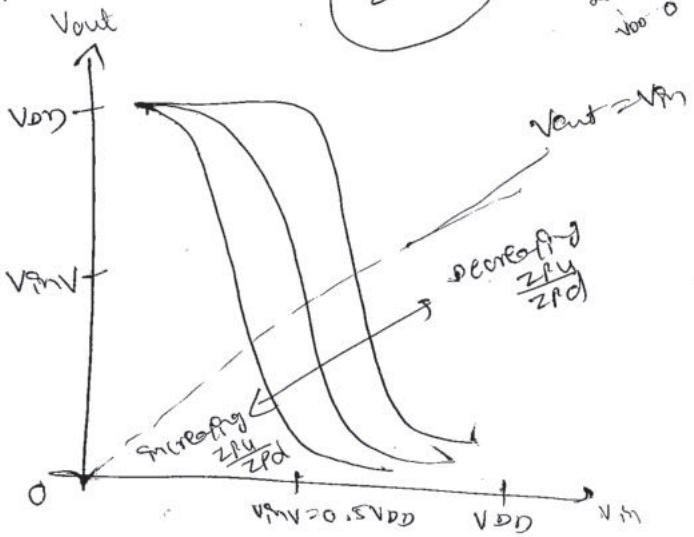
$$\text{Gain} = \frac{\delta V_{out}}{\delta V_{in}}$$

$Z = \frac{L}{W}$



→ The point at which $V_{out} = V_{in}$ is denoted as V_{inv} and it will noted that Transfer ch's and V_{inv} can be shifted by variation of ratio of pull-up to pull down resistances. $\left(\frac{Z_{pu}}{Z_{pd}}\right)$

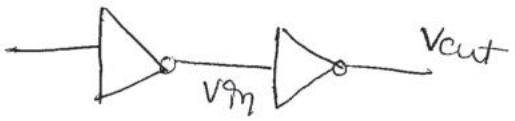
Z → $\frac{L}{W}$ ratio.



Determination of Pull up to Pull down ratio (Z_{pu}/Z_{pd}) for an nmos inverter driven by another nmos inverter?

(M)

→ an inverter is driven from the o/p of another similar inverter.



→ consider the depletion mode TrY for which $V_{gs} = 0$ under all conditions,

→ Assume that in order to cascade inverters without degradation of levels we are aiming to meet the requirement

$$V_{in} = V_{out} = V_{inv}$$

⇒ we set $V_{inv} = 0.5 V_{DD}$. At this pt both TrYs are in saturation. $(V_{ds} > V_{gs} - V_t)$

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

In depletion mode $V_{gs} = 0$.

$$I_{ds} = k \cdot \frac{w_{pu}}{L_{pu}} \frac{(-V_{td})^2}{2} \quad - (1)$$

For Enhancement mode (Pull down TrY)

$$I_{ds} = k \cdot \frac{w_{pd}}{L_{pd}} \frac{(V_{inv} - V_t)^2}{2} \quad - (2) \quad (V_{gs} = V_{inv})$$

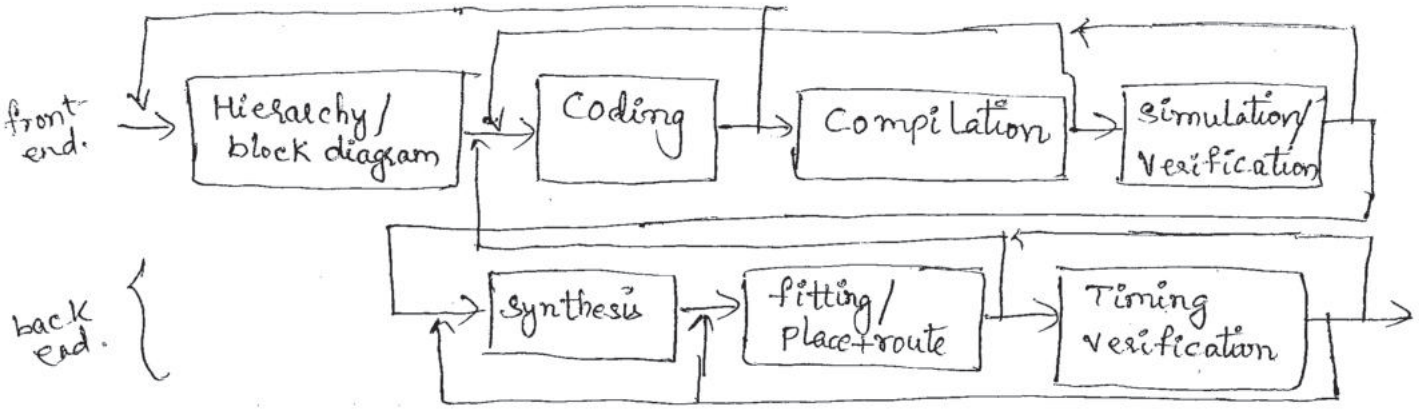
Both current are same.

$$\therefore \quad (1) = (2)$$

$$\frac{k w_{pu}}{L_{pu}} (-V_{td})^2 = k \cdot \frac{w_{pd}}{L_{pd}} \frac{(V_{inv} - V_t)^2}{2}$$

VLSI CIRCUIT DESIGN PROCESS

VLSI DESIGN FLOW



The "front-end" begins with figuring out basic approach & building blocks at the block diagram level.

The next level is writing HDL code for modules, their interfaces, & their internal details. The HDL compiler analyzes code for syntax errors and also checks it for compatibility with other modules on which it relies.

The HDL simulator allows to define and apply I/ps to design and to observe its o/ps without having to build physical circuits. (functional verification).

Back end stage starts with synthesis, thats converts HDL description into a set of primitives or components that can be assembled in the target technology.

This is called netlist that specifies how they should be interconnected.

In fitting step, a fitter maps the synthesized components onto available device resources.

Place & route process lays components and finds ways to connect them. The designer can usually specify additional constraints at this stage, like placement of modules within a chip or the pin assignments of external I/p & O/p pins.

The final step is post-fitting timing verification of the fitted circuit. At this stage actual ckt delays due to wire lengths, electrical loading and other factors can be calculated.

MOS LAYERS

MOS circuits are formed on 4 basic layers:

(i) n-diffusion
(ii) p-diffusion
(iii) Poly Si
(iv) Metal.

→ The thinox mask region includes n-diff, p-diff & transit channels.

→ Poly Si & thinox regions interact so that a transistor is formed where they cross one another.

→ Contacts formed by joining layers.

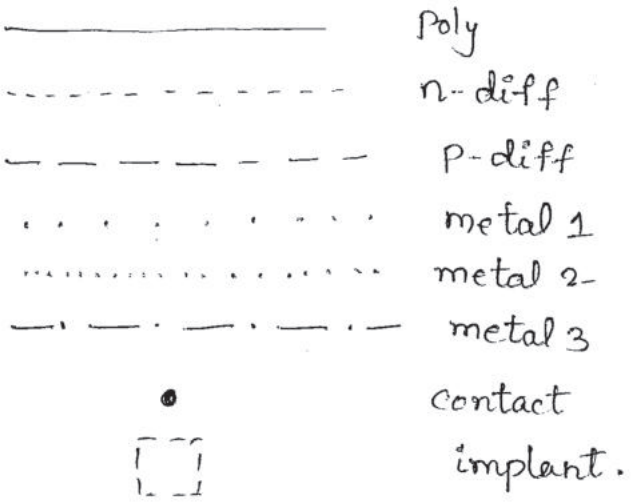
→ some processes includes 2nd metal layer, also 2nd polysi layer.

→ Bipolar transistors can be included in design by addition of extra layers to CMOS process.

STICK DIAGRAMS :

A stick diagram is a cartoon of a chip layout. It represents rectangles with lines which represent wires and component symbols.

Representation



Colors :-

(nMOS design style).

Red : Poly
Green : n-diff
Blue : Metal
Black : Contact
Yellow : implant

(CMOS design style)

Red : Poly
Green : n-diff
Yellow : P-diff
Blue : Metal
Black : Contact

→ A Tr is formed wherever poly crosses diffusion.

→ Area and aspect ratio are difficult to estimate from stick diagrams.

→ Faster to design.

→ Important tool for layouts built from large cells & testing connections b/w cells.

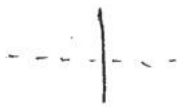
→ A stick diagram is interface b/w symbolic ckt & the actual layout.

→ Often used to solve routing problems.

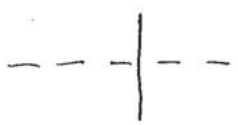
→ Rules:- (1) When two or more sticks of same type cross or touch each other represent electrical contact.

(2) When two or more sticks of different type cross or touch each other there is no electrical contact.

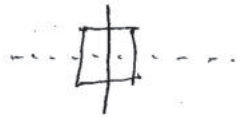
(3) When poly crosses diff^n, it represents MOSFET.



nMOS



pMOS



depletion nMOS.

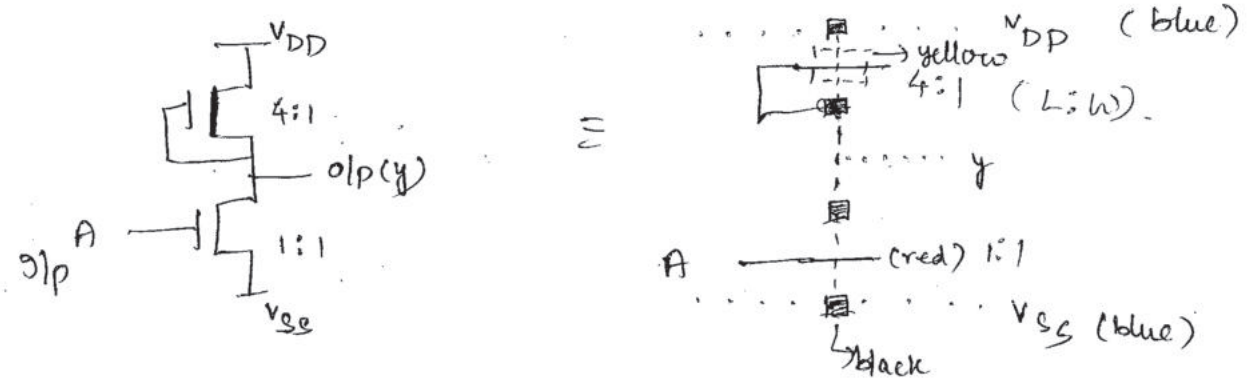
nMOS Design Style :-

A transistor is formed wherever poly crosses n-diff (red over green) and all diffusion wires (interconnections) are n-type (green).

Draw V_{DD} & gnd rails in parallel using metal (blue) allowing enough space b/w them for other circuit elems.

ex:- (1) nMOS inverter

$y = \bar{A}$

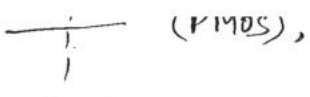


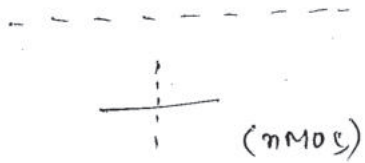
implant in yellow.

CMOS Design Style :-

→ Logical extension of nMOS approach.

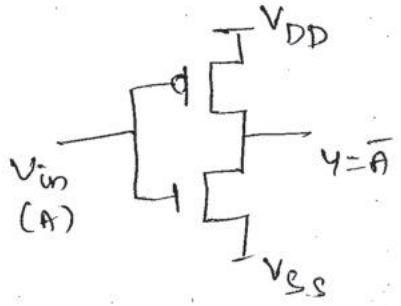
→ The two types of trs used 'n' & 'p' are separated in the stick layout by the demarcation line above which all p-type devices are placed. The n-devices (green) are placed below the demarcation line. and are thus located in the p-well.

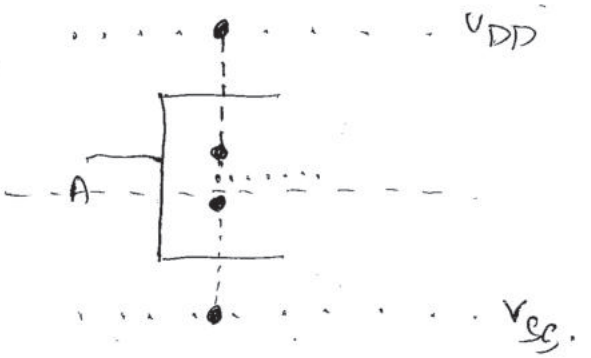




Diffusion paths must not crose the demarcation line and n-diff & p-diff wires must not join.
The 'n' & 'p' features are normally joined by metal where a connection is required.

ex:- CMOS inverter.

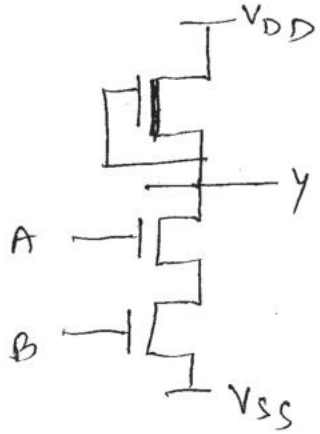


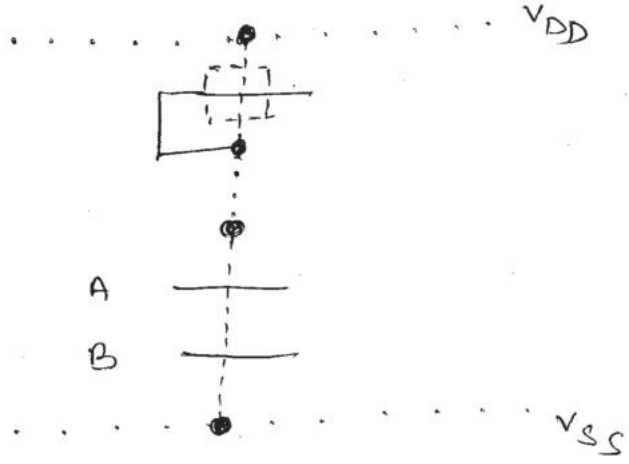


More Examples

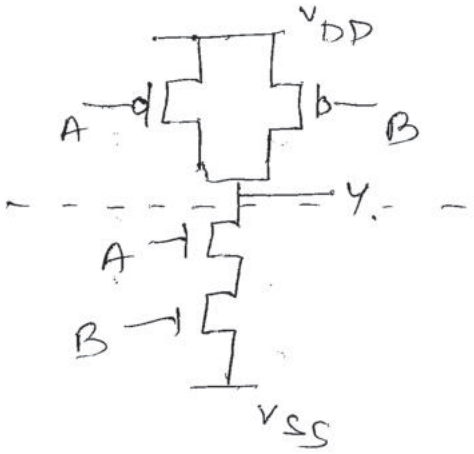
1) $Y = \overline{A.B}$ (Nand Gate)

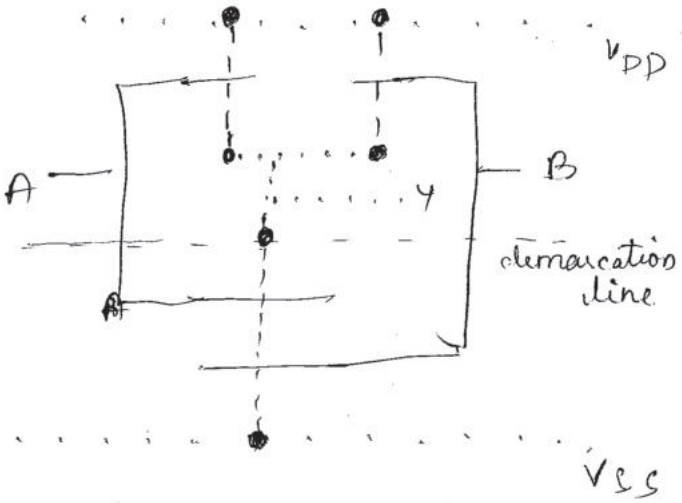
nMOS Logic





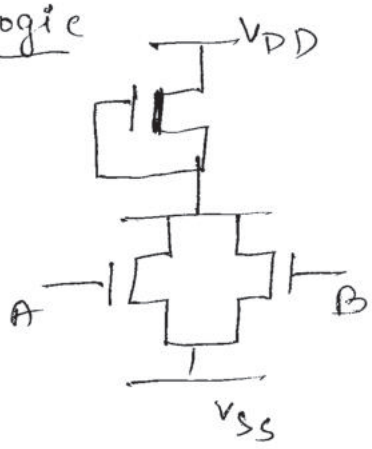
CMOS logic

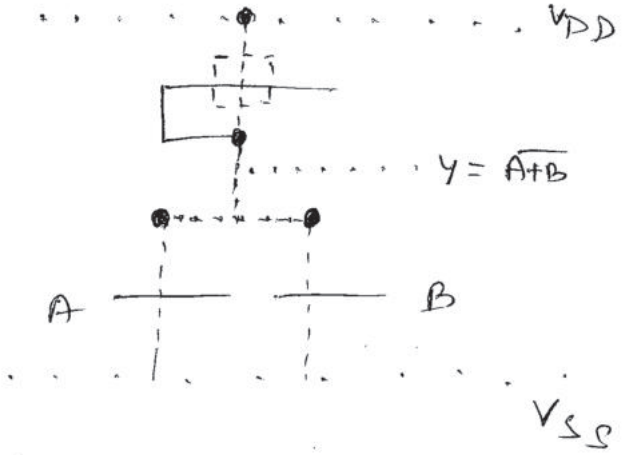




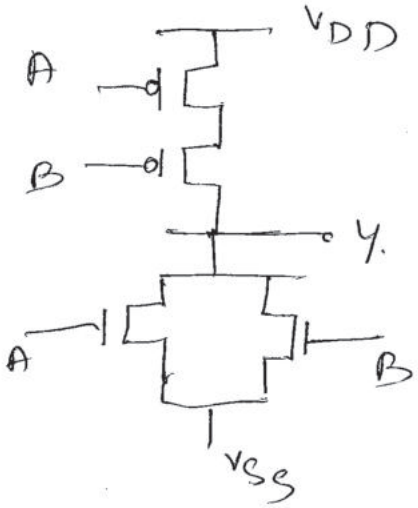
2) $Y = \overline{A+B}$ (NOR).

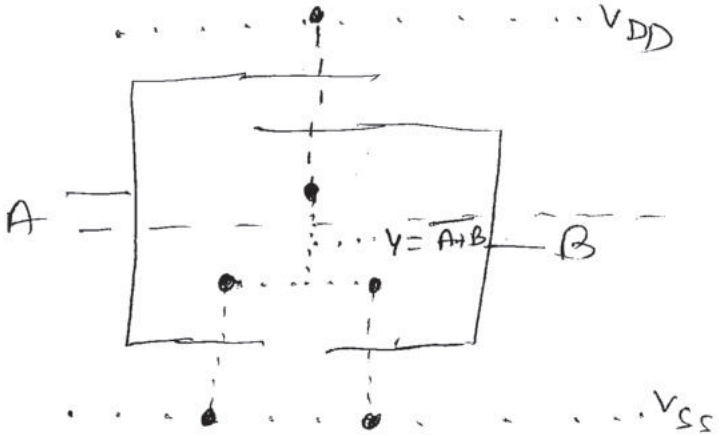
nMOS logic





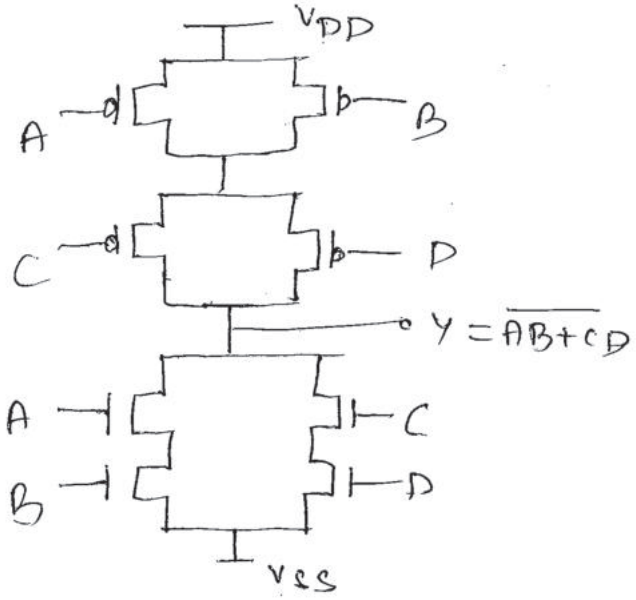
3) $Y = \overline{A+B}$ CMOS logic

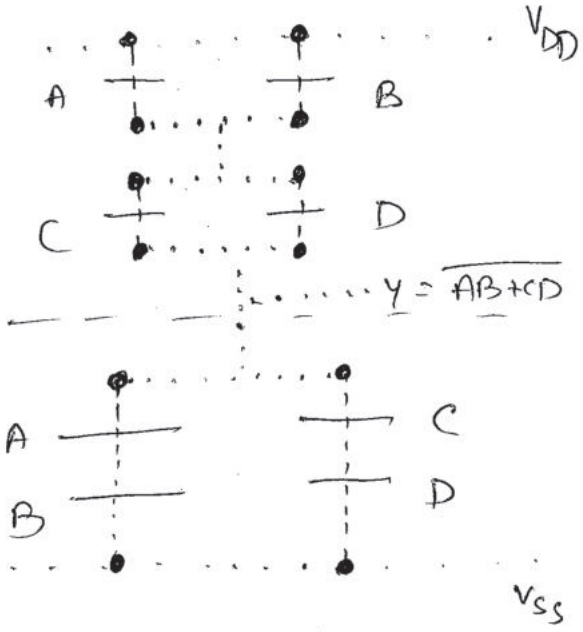




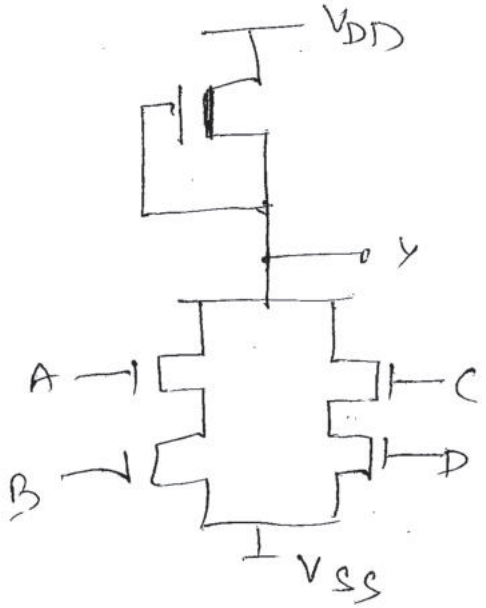
3) $Y = AB + CD$

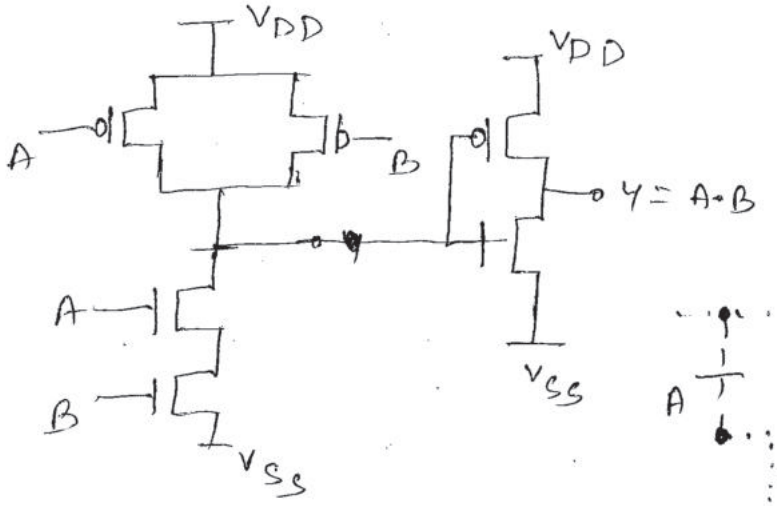
CMOS logic

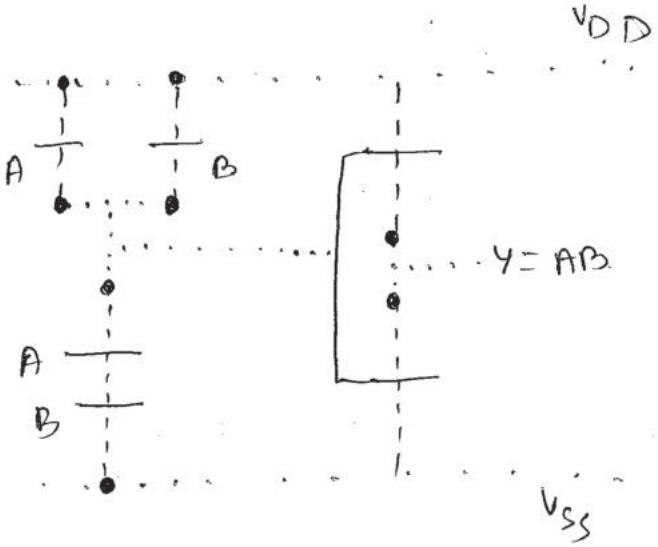




similarly nMOS logic.

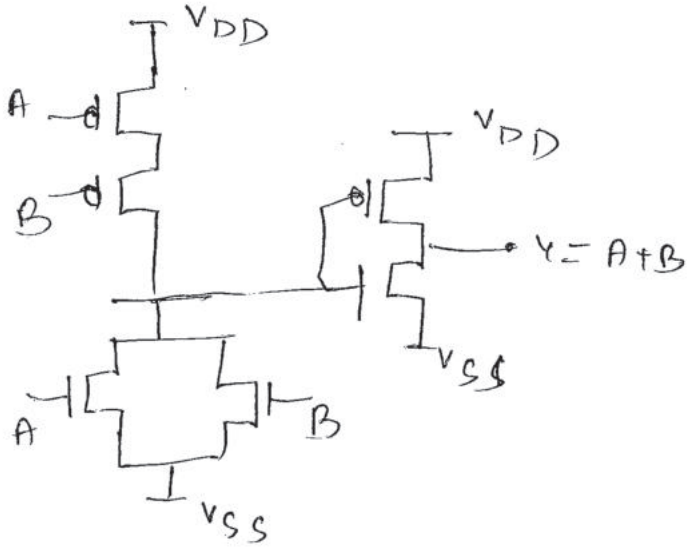


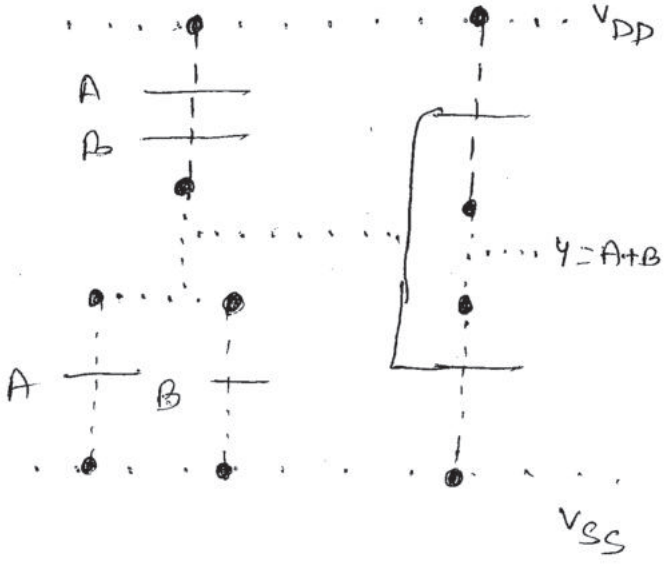




* similarly nMos. logic

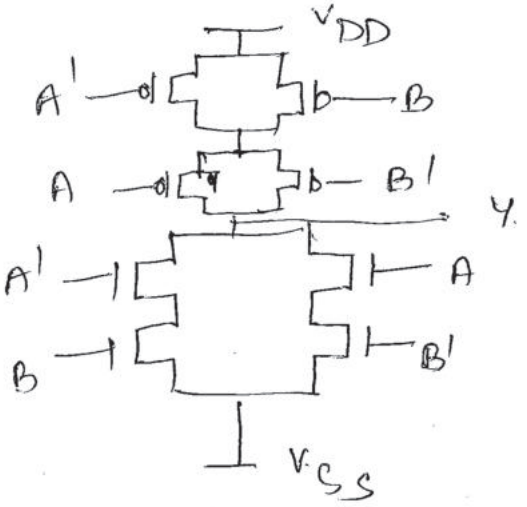
5) $Y = A+B = \overline{\overline{A+B}}$

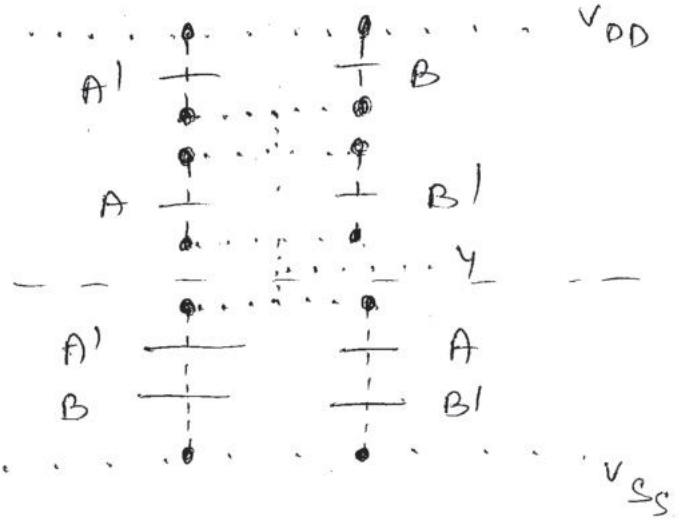




* Do nMos logic.

6) $Y = A'B + AB'$ (xnor)





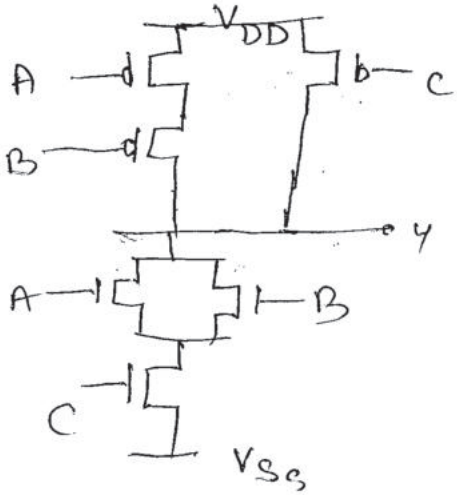
* for nMOS logic: Replace Pull up with nMOS depletion mode transistor with gate connected to source.

7) Similarly $Y = \overline{A'B + AB'}$ (xor gate).

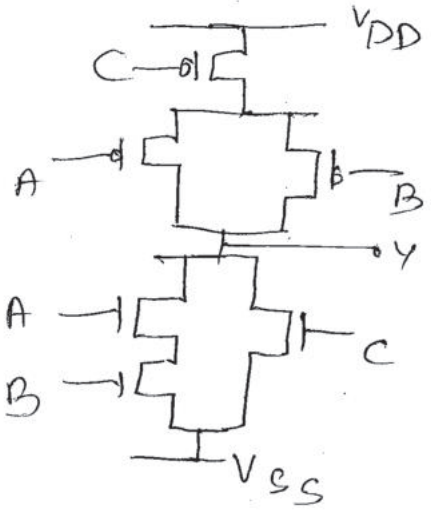
$Y = \overline{A'B + AB'}$

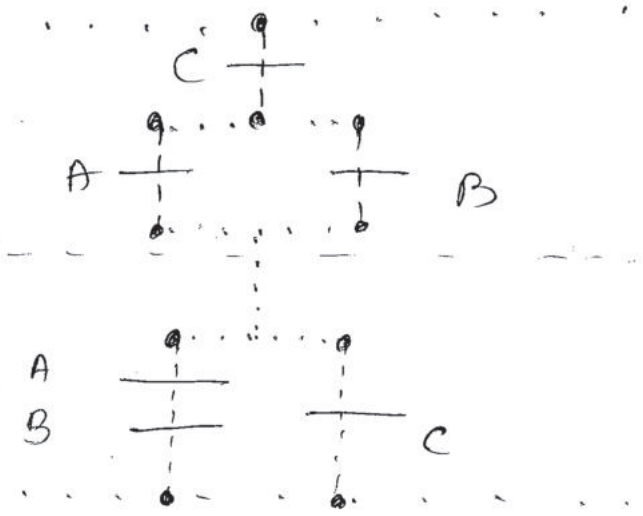
connect inverter to o/p of xnor gate

8) $Y = \overline{(A+B)C}$



9) $Y = AB + C$





DESIGN RULES AND LAYOUTS

→ Design rules allow the translation of circuit design concepts, usually in stick diagrams or symbolic form into actual geometry in silicon.

→ Design rules govern the layout of individual components and interactions - spacings & electrical connections b/w the components.

→ Design rules are specific to a particular semiconductor manufacturing process. It determines low-level props of chip designs (how small individual logic gates are made, how small can the wires be).

→ As small a component size as possible is desired to increase the no. of functions in the chip. But fabrication error arise such as shorting together of wires, or absence of connection b/w wires, faulty transistors etc.

Design rules are used to minimize the occurrence of common fabrication problems and to bring yield of correct chips to acceptable level.

One of fabrication problem is that a wire or any feature being made too wide or too narrow. A too narrow wire may never conduct or may burn off when conducting. A wide wire may short itself with other wires. If poly crosses or cuts diffusion, then it is formation of a new element.

Remedy

1) Introduction of spacing rules
2) Introduction of min - width rules.

Min width rule:

Gives min size for layout element. It also ensures that even with minutest variations, the elem will be of acceptable size.

Spacing rule:-

Gives min distance b/w the edges of layout elems, so that even with minor variations it will not cause the element to overlap nearby layout elems.

Composition rules:-

Ensures that components are well- formed.

Construction rules (vias)

→ Material on both layers to be connected must extend beyond SiO_2 cut and cut must be at least

Scalable Design Rules

→ Design rules can be scaled in terms of 'λ', which is the size of the smallest elem in the layout.

→ When devices shrink, layouts need not be completely redesigned. All features can be measured in integral multiples of 'λ'.

→ By choosing a value for λ, all dimensions set at a scalable layout.

→ Scalable layouts are advantageous as chips become faster as size shrinks.

→ Digital ckt designs scale, b'coz the cap loads that must be driven by logic gates shrink faster than the currents supplied by the Trs.

→ Assuming that the basic physical paramtrs of chip are shrunk by a factor of $1/x$.

Length = $L \rightarrow L/x$; Width = $W \rightarrow W/x$.

Thickness = $D \rightarrow D/x$

Supply $V_t = V_{DD} - V_{SS} \rightarrow (V_{DD} - V_{SS})/x$

Doping $\rightarrow N_d \rightarrow N_d/x$

Tr transconductance:

$$\hat{g_m} = x \cdot g_m$$

Threshold v_{tg}:

$$\boxed{\hat{V_t} = \frac{V_t}{x}}$$

Satⁿ Drain Current:

$$I_{ds} = K \frac{W}{L} [(V_{gs} - V_t)^2] .$$

$$K = \frac{\mu \epsilon_0 \epsilon_{ins}}{D}$$

$$\boxed{\hat{K} = K \cdot x}$$

$$\Rightarrow \boxed{\frac{\hat{I_{ds}}}{I_{ds}} = \frac{1}{x}}$$

&

$$\boxed{\frac{\hat{C_g}}{C_g} = \frac{1}{x}}$$

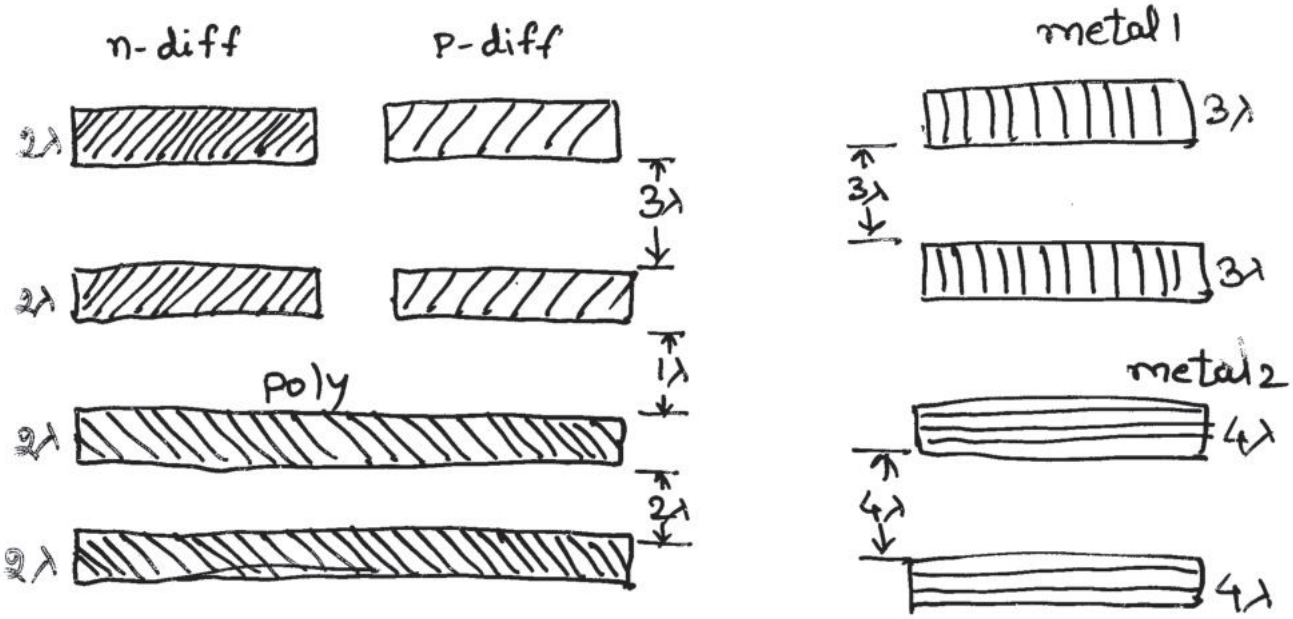
→ $\frac{CV}{I}$ is measure of speed of ckt over scaling.

$$\therefore \frac{\hat{C}\hat{V}/\hat{I}}{CV/I} = \frac{1}{x}$$

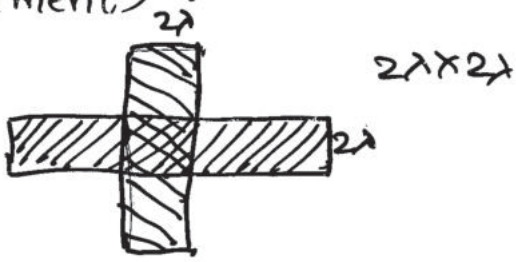
⇒ Scaling is done on λ, thus $\frac{\hat{\lambda}}{\lambda} = \frac{1}{x}$.

(thus speed up by factor x)

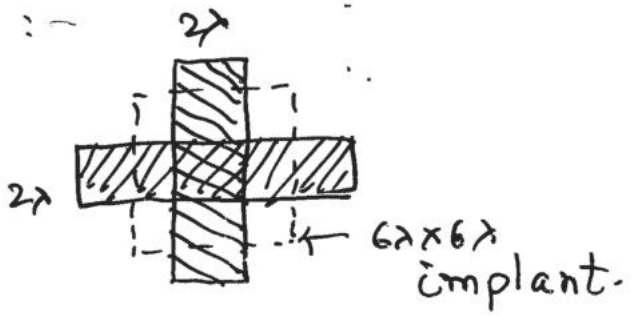
Design rules for wires (nMOS & CMOS)



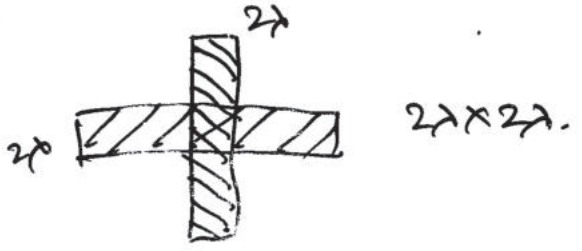
nMOS (enhancement) :-



nCMOS (Depletion) :-



pMOS (Enhancement)



Summary:

1) Metal 1 : min-width = 3λ
min-sep^n = 3λ

2) Metal 2 : min-width = 4λ
min-sep^n = 4λ

3) Poly : min-width = 2λ
min poly-poly sep^n = 2λ

4) P & n diffusion: min width = 2λ
min sep^n b/w same diff = 2λ.

5) Tubs: 10λ wide.

Min Separation b/w tub & src/drain = 5λ.

Tub Tie: P-tub tie : 2λ × 2λ cut, 4λ × 4λ metal, 4λ × 4λ p$^+$ diff
n-tub tie.

Construction rules:-

1) Transistors : width = 2λ
Length = 2λ

2) Poly extends 2λ beyond active region.

3) Diffusion extends 2λ.

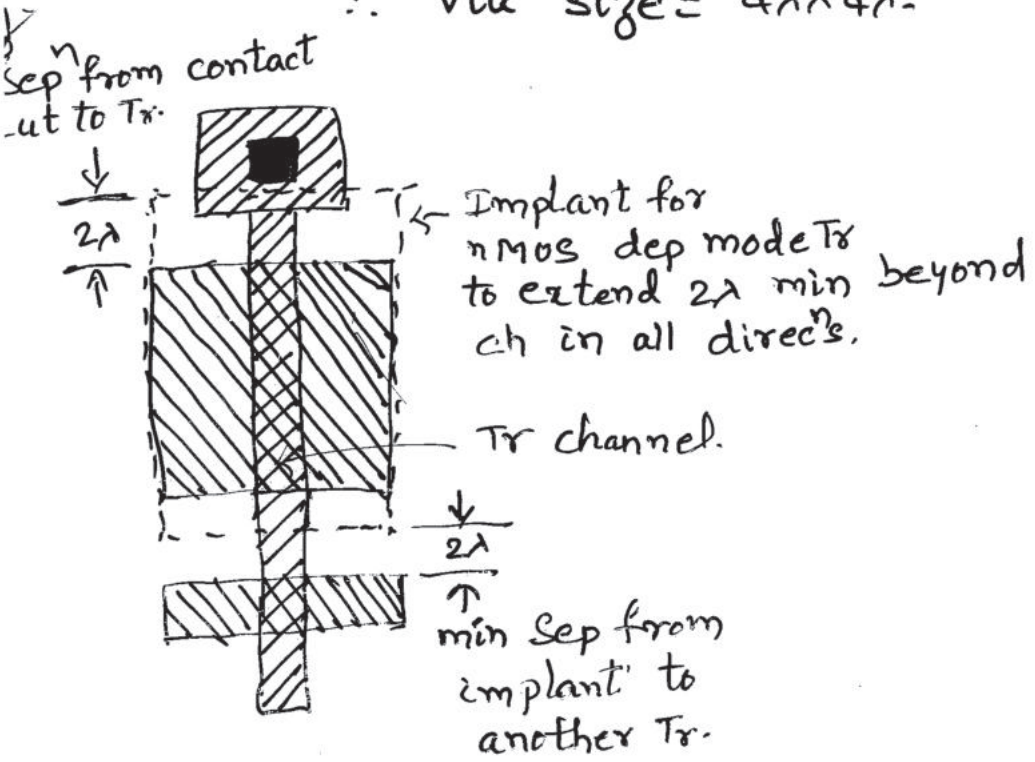
4) Active region must be atleast 1λ from poly-metal via, 2λ from another Tr,

Vias :

→ cuts : $2\lambda \times 2\lambda$

→ Material on both layers extend 1λ in all dirⁿ from cut.

∴ via size = $4\lambda \times 4\lambda$.



Extentions & Separations (Trs.)

Contact Cuts :

3 ways to make contacts b/w poly & diffusion in nMOS ckts :

(i) poly to metal then metal to diffusion
(ii) buried contact (poly to diff).
(iii) butting contact (poly to diff using metal).

The $2\lambda \times 2\lambda$ contact cut indicates an area in which the oxide is to be removed down to the underlying polysi or diff surface

When deposition of metal layer takes place the metal is deposited thru contact cut areas onto underlying area so that contact is made b/w the layers.

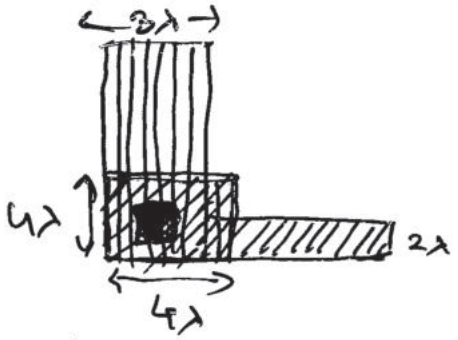
ex:-

1) Metal1 to polysi or to diffusion.

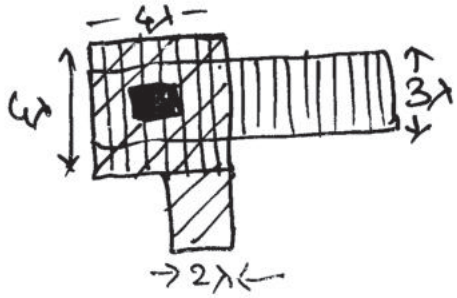
Metal1 to poly.

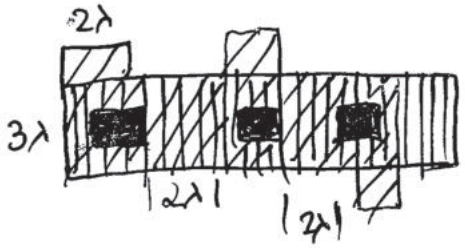
$2\lambda \times 2\lambda$: cut centered on $4\lambda \times 4\lambda$ superimposed areas of layers to be joined in all cases.

Metal1 to n-diff



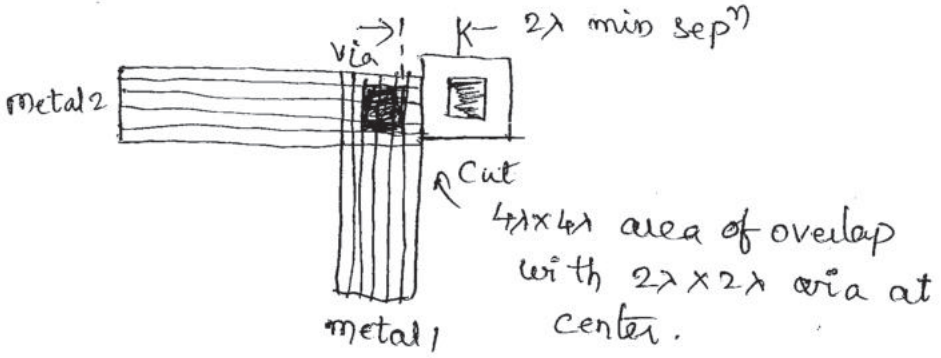
Metal1 to P-diff





Min Sep, Multiple cut

2) Via (Contact from metal 2 to metal 1)



NOTE: min sepⁿ b/w ~~n~~diffⁿ wire and poly wire = 1λ.

→ Contact cuts are also known as via cuts.

→ 4λ×4λ size.

→ Contact cut types:

(i) n/p diffⁿ to polysi
(ii) poly to metal 1
(iii) n/p diffⁿ to metal 1
(iv) metal 1 to metal 2.

Contact b/w polysi and diffusion wires can be done in 3 ways:-

(a) Polysi to the metal and then metal to polysi.

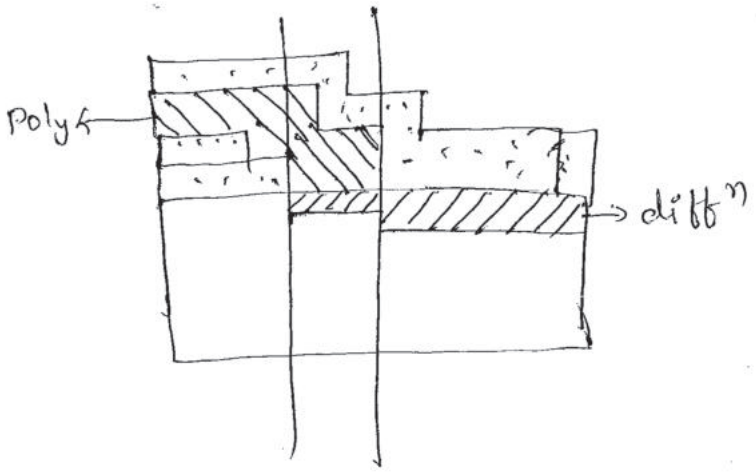
* Oxide is removed from 2λ×2λ contact cut down to underlying polysi wire. Then metal is deposited. It flows thru the oxide etched area to polysi area. Then polysi is deposited on the surface, which acts as conduction path.

(b) Buried Contact

Before starting the process, there is oxide layer on Si surface, oxide is etched to expose the underlying Si & polysi is deposited on the surface.

In the next step, diffⁿ is carried out on the exposed surface. When diffⁿ takes place impⁿs will diffuse into polysi as well as diffused area within the contact area.

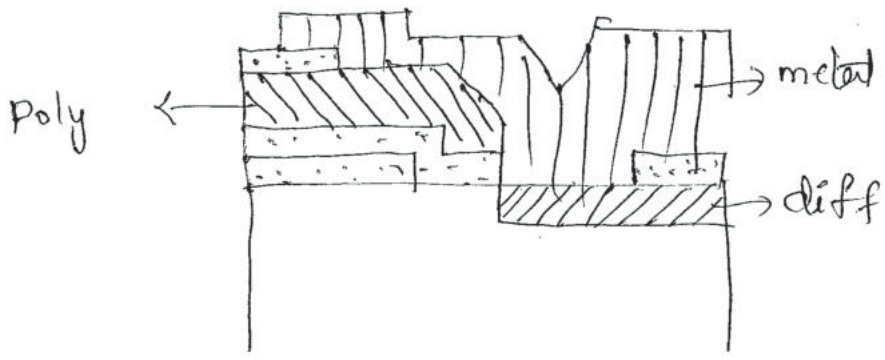
This ensures a satisfactory connecⁿ b/w polysi & diffⁿ. Buried contacts are smaller than butting contacts.



(c) Butting Contact:-

→ a complex process.

→ 2λ×2λ contact cut is made down to each layers to be joined. Layers are butted together so that two contact cuts become contiguous. The poly & diffⁿ outlines overlap & thin oxide under poly acts as mask in the diffⁿ process. Poly & diffused layers are butted together. The contact b/w two layers is then made by metal overlay.



Double metal MOS process rules

In this process a second metal layer is used so that V_{DD} & V_{SS} (gnd) rails in the system are distributed more flexibly on the chip. vias are used to establish connection b/w metal 2 to other layers thru metal 1.

The first level metal can be used for local distribution of power & for signal lines.

CMOS Lambda-based design rules :-

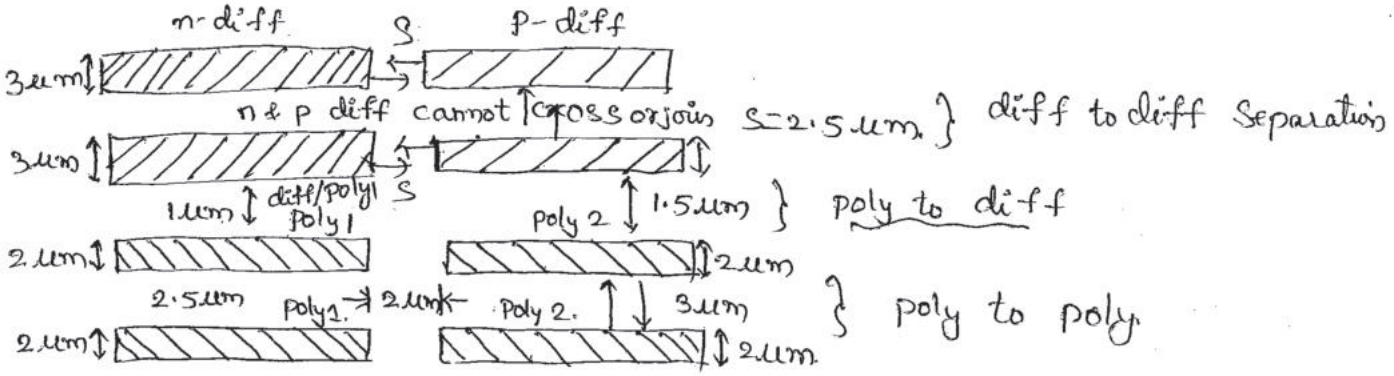
The rules of n-well (PMOS Tr), p-wires & special substrate contacts are added to the existing NMOS rules.

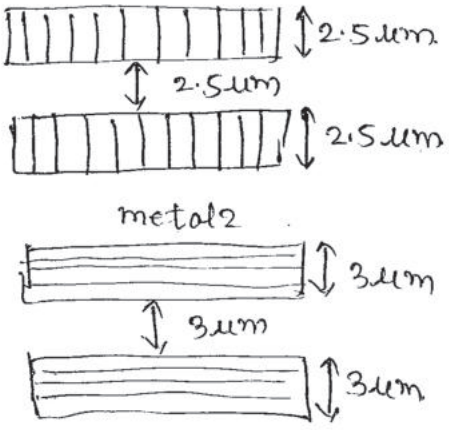
2 μm CMOS DESIGN RULES

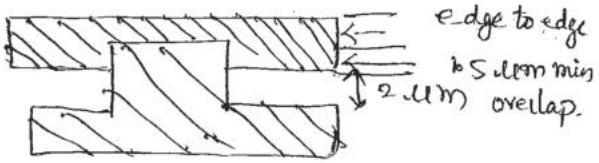
* 2 μm double metal, double poly

→ n-well : brown
Poly 1 : red
Poly 2 : Orange
n-diff : Green
P-diff : Yellow
} CMOS

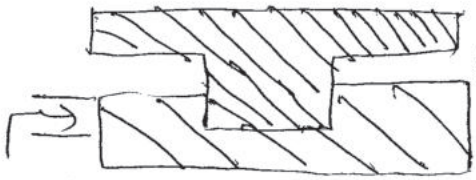
buried n+ subcoll : pale green
p-base : pink
} BiCMOS.







poly 2 overlapping poly 1.



1.5um min. overlap.

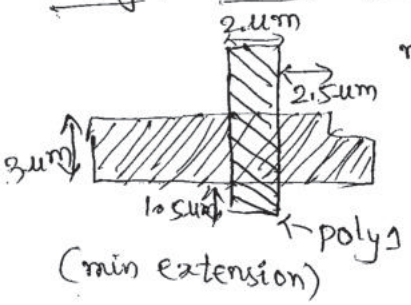
poly 1 overlapping poly 2.

fig: Design rules for wires (2 um CMOS).

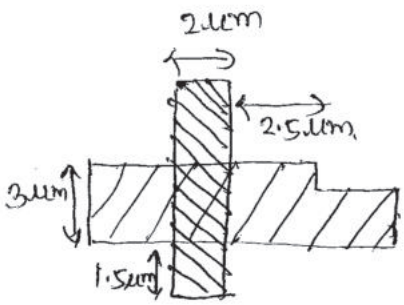
NOTE:-

For p-well CMOS, n-diff can only exist inside & p-diff wires outside p-well. For n-well CMOS, p-diff wires can only exist inside & n-diff wires outside n-well.

Design rules for transistors:-



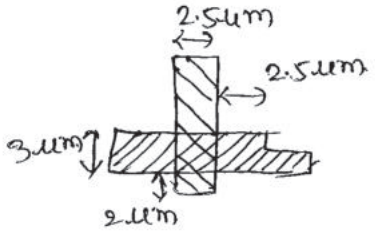
min overlap of n-diff beyond gate.



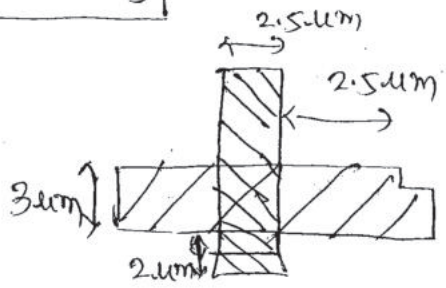
(i) n-type enhancement

(ii) p-type Tr

fig:- PolySi Transistors.



(i) n-type Tr

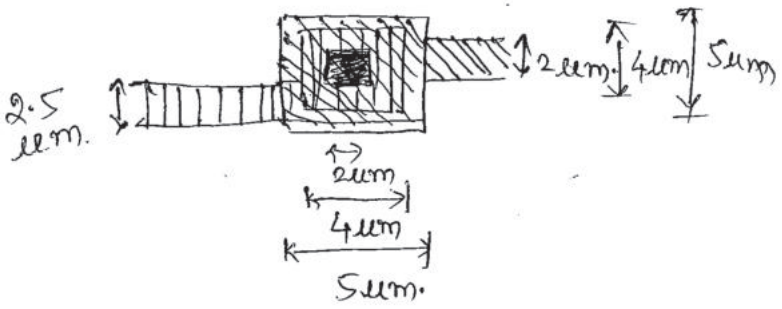


(ii) p-type Tr

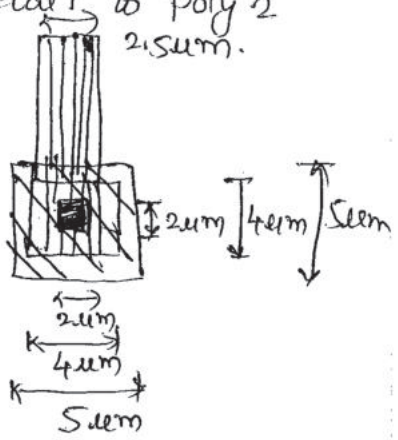
fig:- Polysi 2 Transistors.

Design rules

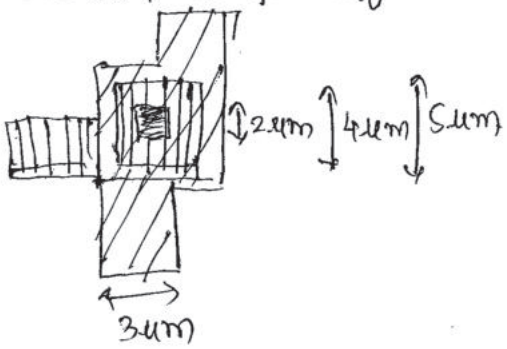
(a) Metal 1 to poly-1



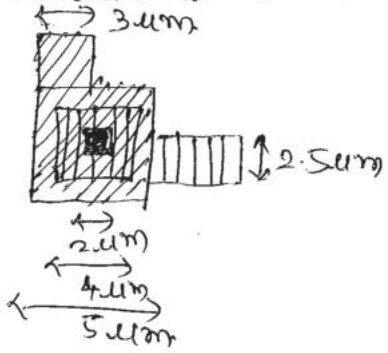
(b) Metal 1 to poly 2



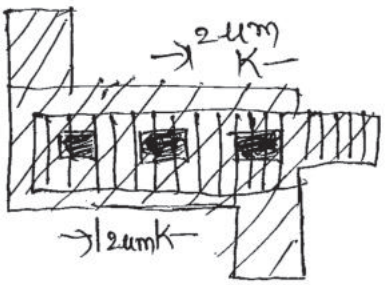
(c) Metal 1 to p diff.



(d) Metal 1 to n diff.

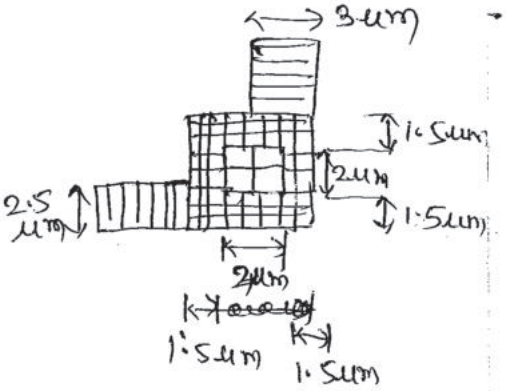


(e) Multiple contact cuts.



min spacing b/w contact cuts = 2µm

(f) Via metal 1/metal 2



Limitations of Scaling

(1) Substrate doping. $d = \sqrt{\frac{2\epsilon_{si}\epsilon_0 V}{qN_B}}$ (where $V = V_a + V_B$)

(2) Limits on miniaturization

(3) Limits of interconnect & contact resistance.

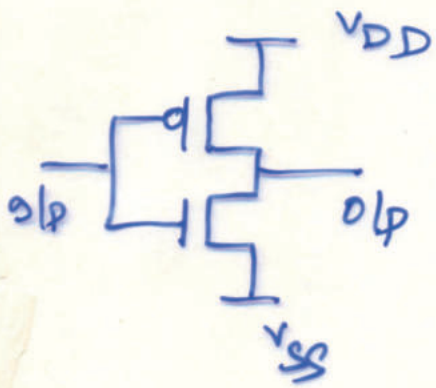
(4) Limits due to subthreshold currents.

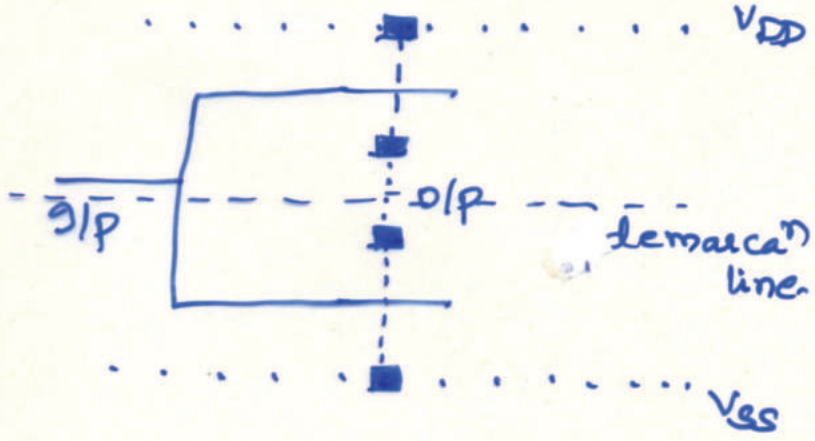
(5) Limits on logic levels & supply vtg due to noise.

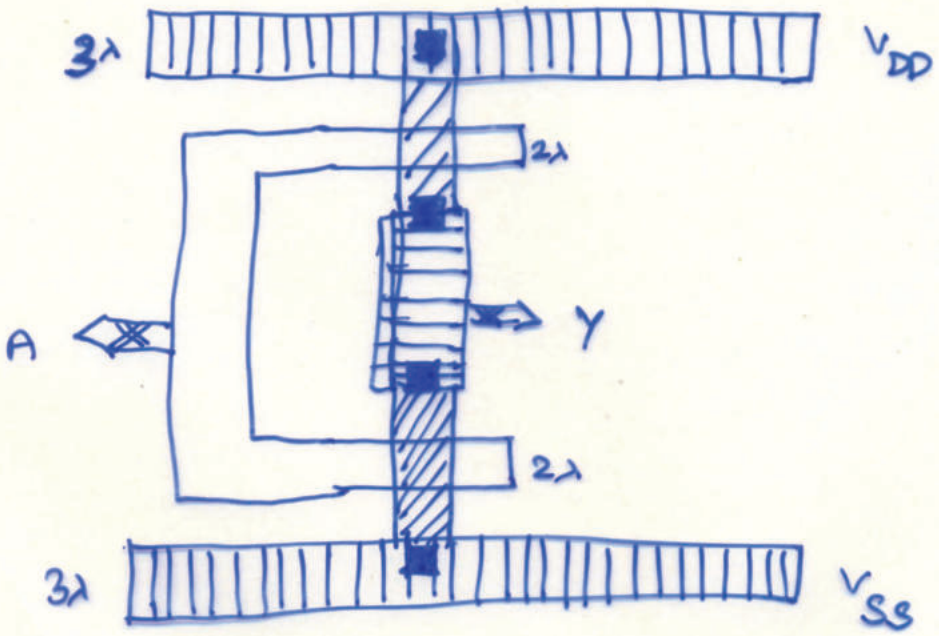
(6) Limits due to current density.
[$J = 1$ to $2\,mA/\mu m^2$]

Examples:

1) NOT (INVERTER)





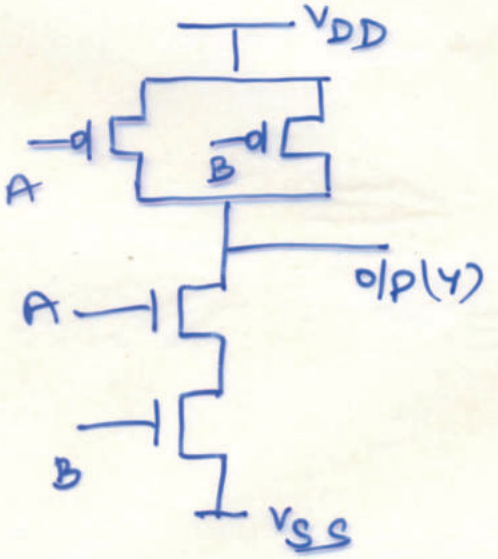


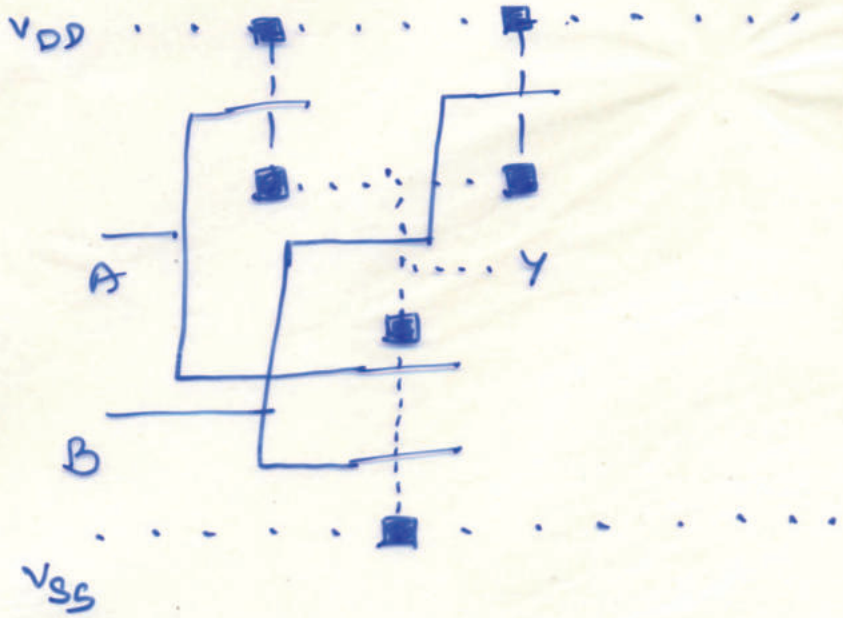
I/p: poly
O/p: Metal
V_{DD}, V_{SS}: Metal

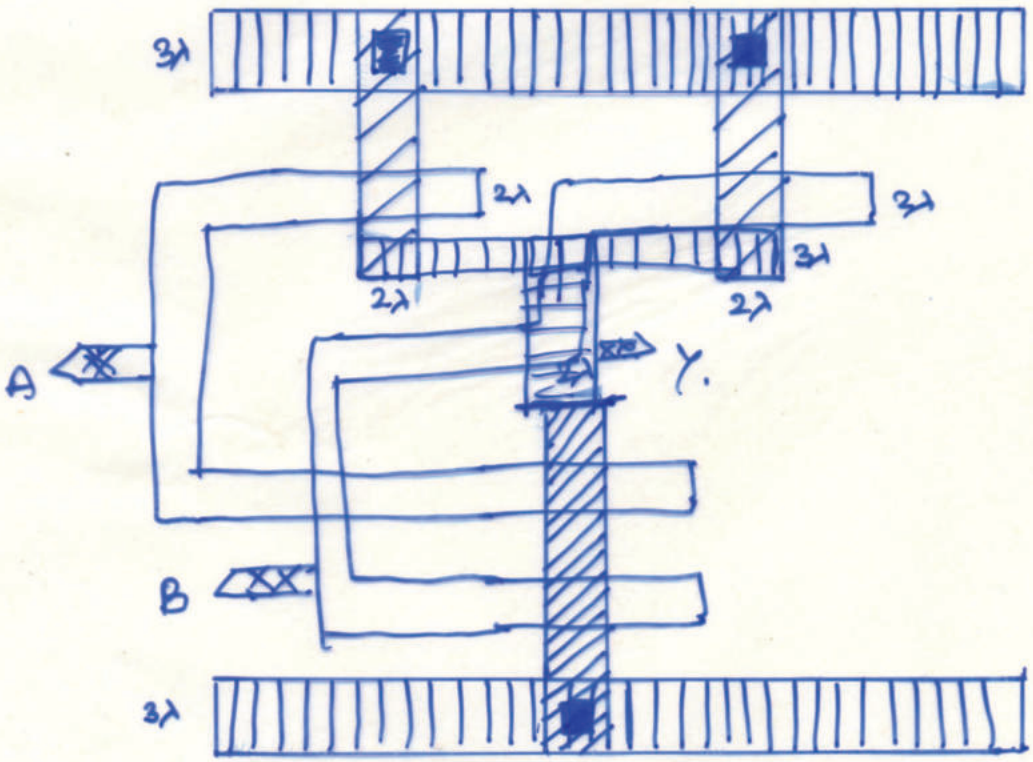
Color Codes:

Metal - Blue
Poly - Red.
n-diff - Green
P-diff - Yellow
Via - Black.

2) $Y = \overline{A \cdot B}$





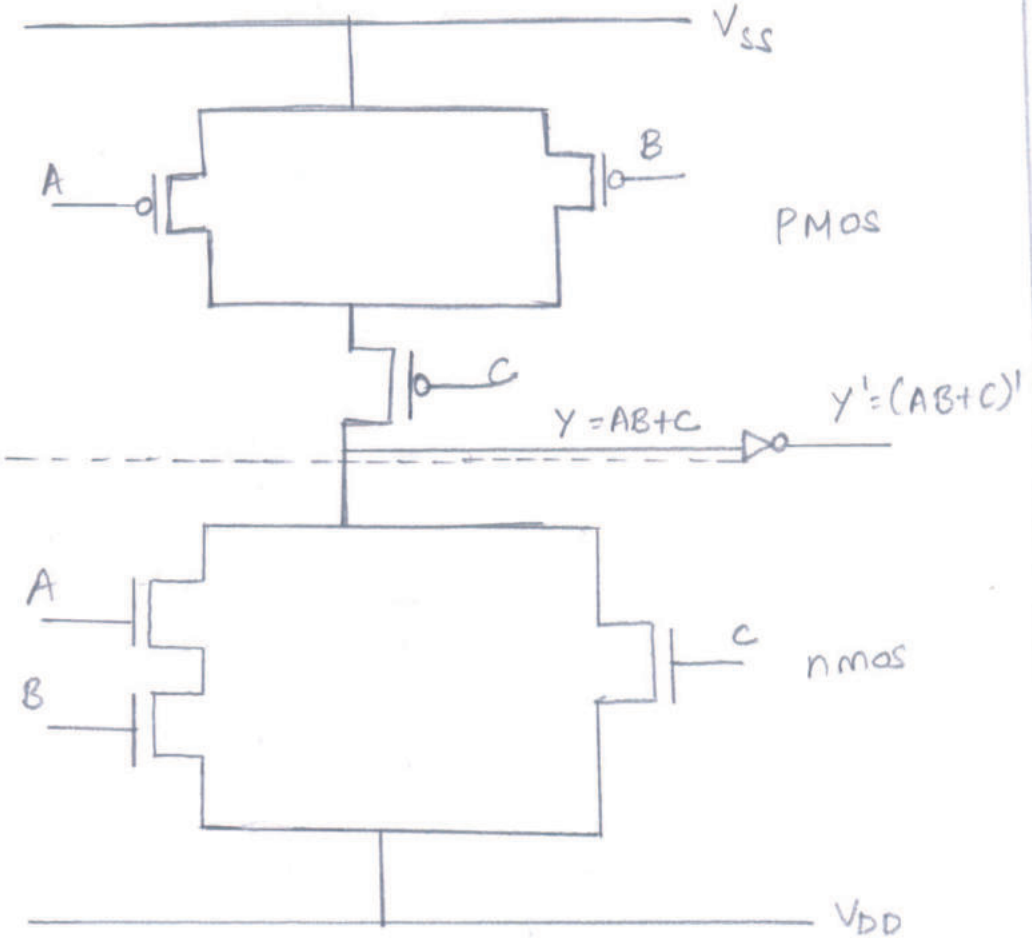


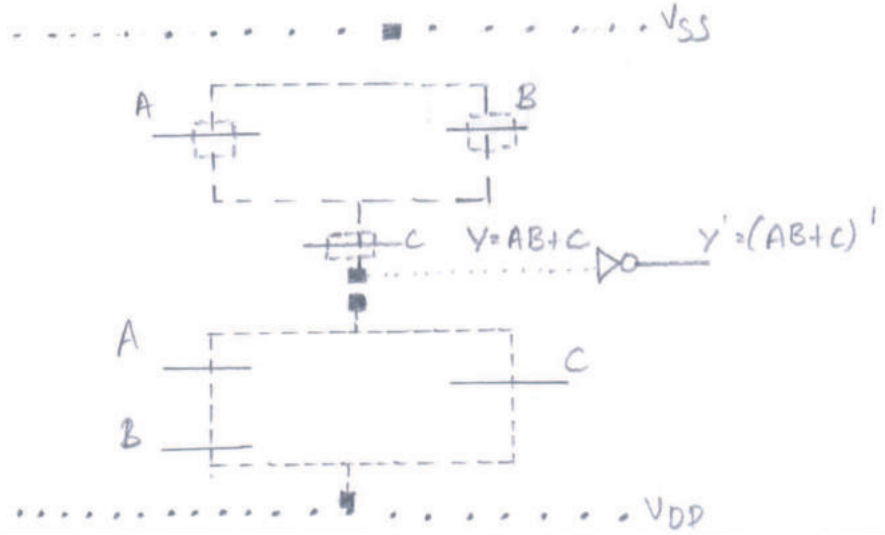
NOTE :- Via types :

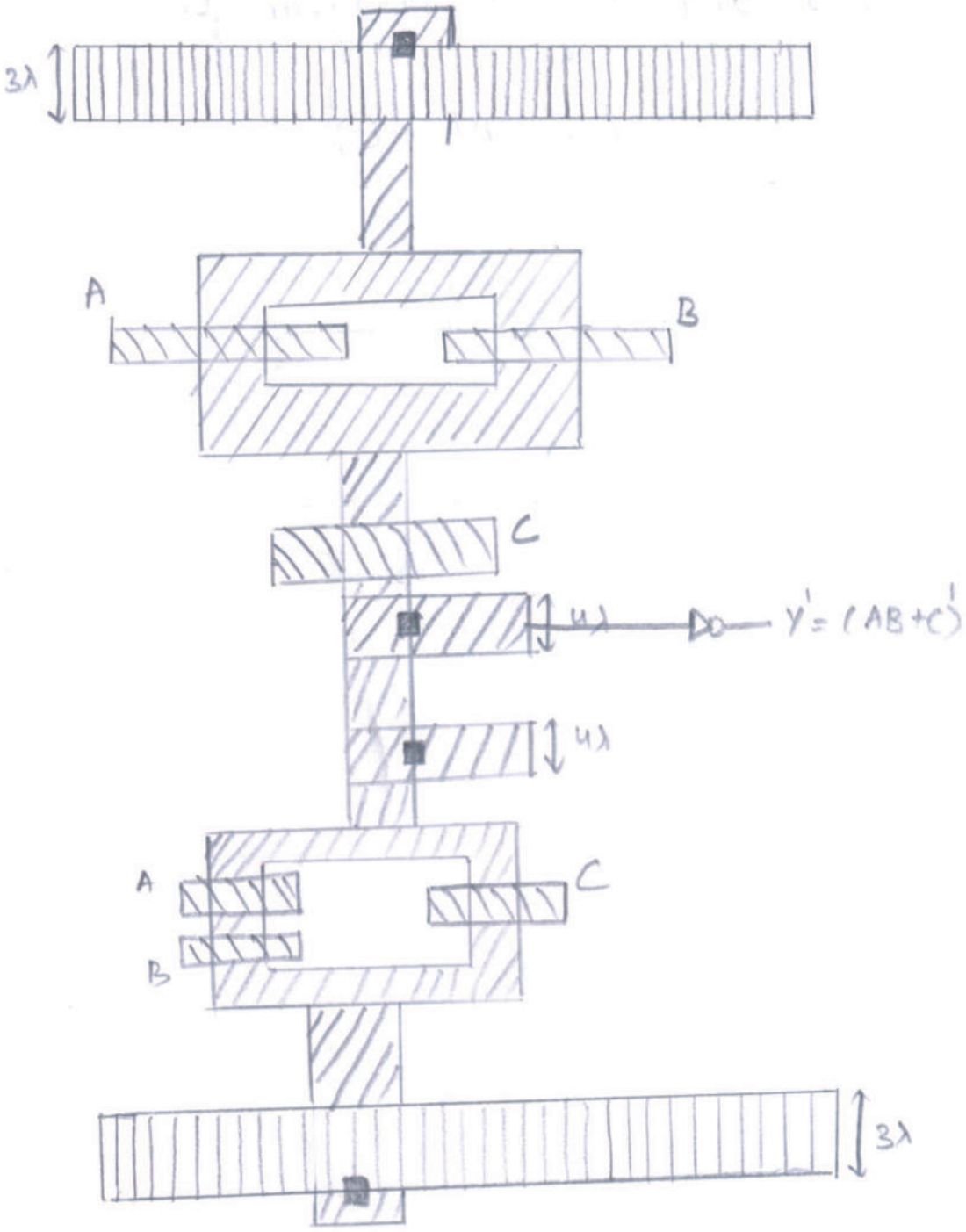
(i) n/p diff - poly
(ii) poly - metal 1
(iii) n/p diff - metal 1
(iv) metal 1 - metal 2

3) Stick and layout diagram for

$$Y = (AB + C)'$$






3λ
A
B
C
4λ
Y' = (AB+C)'
4λ
A
B
C
3λ

Scaling Factors for device Parameters:

(1) Gate Area (A_g):

$$A_g = W \cdot L$$

W & L scaled by $1/\alpha$.

$\therefore$ A_g scaled by $1/\alpha^2$

(2) Gate Cap Per Unit Area C_o or C_{ox}.

$$C_{ox} = \frac{\epsilon_{ox}}{D} = \frac{\epsilon_0 \epsilon_{ins}}{D}$$

$$\because \hat{D} = \frac{1}{\beta} D$$

$$\Rightarrow \boxed{\hat{C_{ox}} = \beta C_{ox}} \quad \text{(scaled by } \beta\text{)}$$

(3) Gate Cap C_g

$$C_g = C_{ox} \cdot W \cdot L$$

$$\hat{C_g} = \beta C_{ox} \cdot \frac{W}{\beta\alpha} \cdot \frac{L}{\alpha}$$

$$\hat{C_g} = \frac{\beta}{\alpha^2} \cdot C_{ox} \cdot W \cdot L$$

$$\boxed{\hat{C_g} = \frac{\beta}{\alpha^2} C_g.} \quad \text{(Scaled by } \frac{\beta}{\alpha^2}\text{)}.$$

(4) Parasitic Capacitance C_x:

$$C_x \propto \frac{A_x}{d}$$

where d: depletion width around S & D

$$\hat{d} = d/\alpha.$$

A_x: Area of depletion region around src or drain

$$\hat{A_x} = A_x/\alpha^2.$$

$$\therefore \hat{C_x} = \frac{A_x}{\alpha^2} \cdot \frac{1}{d/\alpha} = \frac{A_x}{\alpha \cdot d}$$

$\therefore$ C_x scaled by $1/\alpha$

(5) Carrier Density in Channel Q_{on}:

$$Q_{on} = C_{ox} \cdot V_{gs}.$$

where Q_{on}: avg charge per unit area in ch in 'on' state.

C_{ox}: Scaled by β

V_{gs}: " " $1/\beta$

$$\therefore \hat{Q_{on}} = \beta \cdot C_{ox} \cdot \frac{1}{\beta} \cdot V_{gs} = Q_{on}$$

$\Rightarrow$ Q_{on} is scaled by 1.

(6) Channel Resistance R_{on}:

$$R_{on} = \frac{L}{W} \cdot \frac{1}{\varepsilon_{on} \cdot \mu}$$

where μ = carrier mobility (const).

R_{on} scaled by $\frac{1}{\alpha} \cdot \frac{1}{1/\alpha} \cdot 1 = 1$

(7) Gate Delay T_d.

$$T_d \propto R_{on} \cdot C_g$$

$$\hat{T_d} \propto 1 \cdot R_{on} \cdot \frac{\beta}{\alpha^2} \cdot C_g$$

$$\hat{T_d} \propto \frac{\beta}{\alpha^2} \cdot T_d$$

$\therefore$ scaled by $\frac{\beta}{\alpha^2}$

(8) Max Operating Frequency f_o:

$$f_o = \frac{W}{L} \cdot \frac{\mu C_o V_{DD}}{C_g}$$

$$f_o \propto \frac{1}{T_d}$$

$\therefore$ f_o scaled by $\frac{\alpha^2}{\beta}$

(9) Saturation Current I_{dss}.

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

$\because$ V_{gs} & V_t scaled by $1/\beta$ & C_o by β

$\Rightarrow$ I_{dss} scaled by $\beta \left(\frac{1}{\beta}\right)^2 = 1/\beta$

(10) Current Density J:

$$J = \frac{I_{dss}}{A}$$

'A' (area) of ch scaled by $1/\alpha^2$

$$\hat{J} = \frac{I_{dss}/\beta}{A/\alpha^2} = \frac{\alpha^2}{\beta} \cdot \frac{I_{dss}}{A} = \frac{\alpha^2}{\beta} J$$

$\therefore$ J is scaled by $\frac{\alpha^2}{\beta}$.

(11) Switching Egy Per Gate E_g.

$$E_g = \frac{C_g}{2} (V_{DD})^2$$

E_g scaled by $\frac{\beta}{\alpha^2} \cdot \frac{1}{\beta^2} = \frac{1}{\alpha^2 \beta}$

UNIT – III

Logic Gates and other complex Gates :–

Cmos static Logic :–

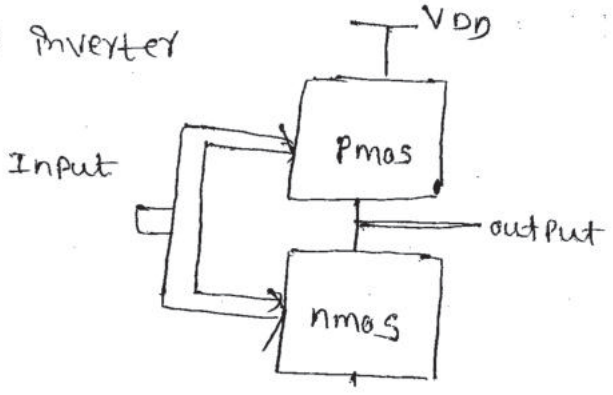
Static, fully complementary cmos gate designs using Inverter, NAND and NOR gates can build more complex functions.

→ These cmos gates have good noise margins and low static power dissipation at the cost of more transistors when compared with other cmos logic designs.

→ Cmos gate have 2 tansistor nets (nmos and Pmos) whose topologies are related.

Pmos transistor net is connected between the Powersupply and logic gate output, whereas the nmos transistor topology is connated between the output and ground.

Ex: Cmos Inverter



The transistor network is related to the Boolean function with a straightforward design procedure:

(1) Derive the nmos transistor topology with following rules:

- Product terms in the Boolean function are implemented with series-connected nmos transistors.
- Sum terms are mapped to nmos transistors connected in parallel.

(2) The pmos transistor network has a dual or complementary topology with respect to the nmos net.

(3) Add an inverter to the output to complete the function if needed. some functions are inherently negated such as NAND, NOR etc, and donot need an inverter at the output state.

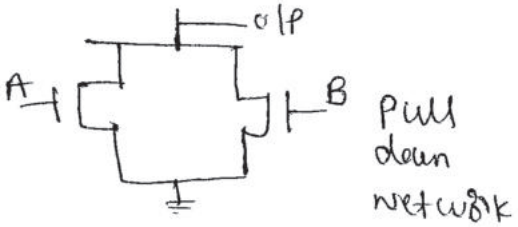
An inverter added to a NAND or NOR function produces the AND and OR function.

Examples which require inverter to fulfil the function:

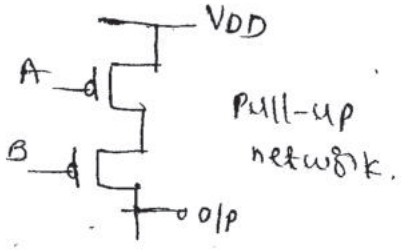
(1) OR: $F=(A+B)$

(1) nmos transistor topology:

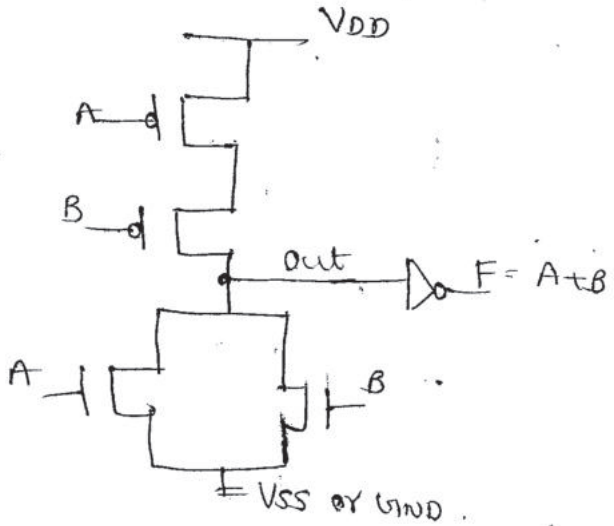
Connect 2 transistors in parallel which indicates NOR function of A & B



(2) Implement pmos net as a dual topology to nmos net. i.e, Connect 2 pmos transistors in series.



(3) Finally add an inverter to obtain the function, so that $F = \overline{out}$

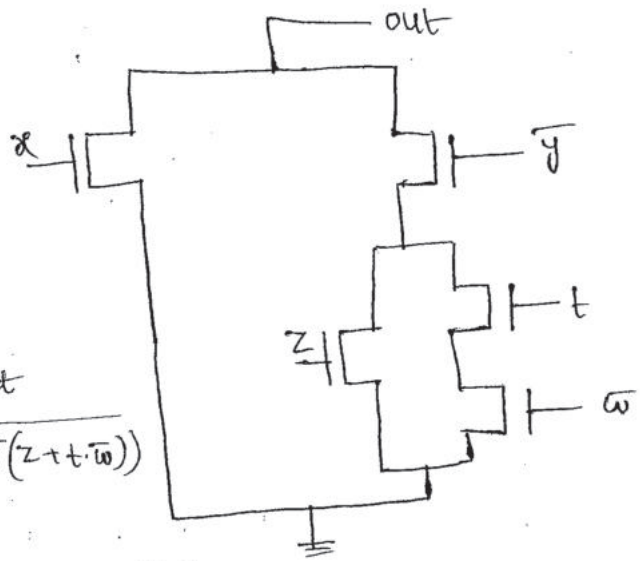


Example 2 :-

Design nmos transistor net for a Boolean function

$F = x + \{\bar{y} \cdot [z + (t \cdot \bar{w})]\}$.

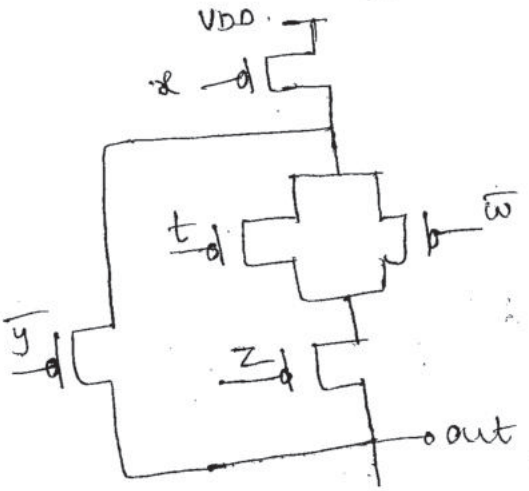
Soln: We design this gate with a top-down approach. The nmos transistor network is connected between the output and ground terminal.



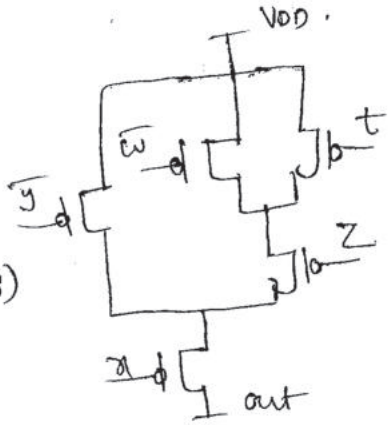
nmos net

for $F = \overline{x + (y(z + t \cdot \bar{w}))}$

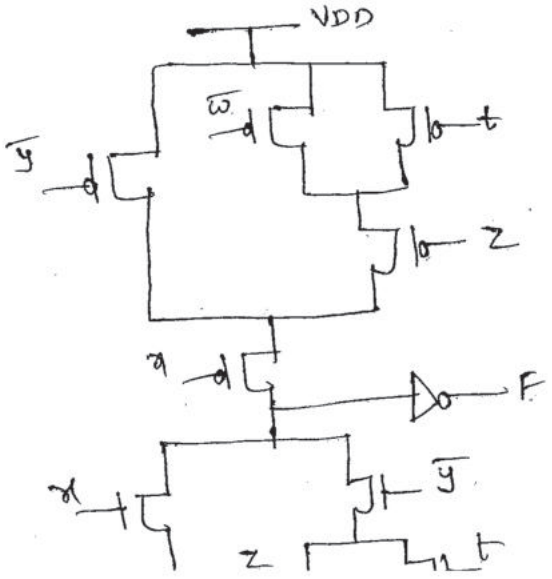
Pmos net



(or)



→ Then connect both pmos net & nmos net, then connect an inverter to its output.



Driving large capacitive loads :-

The problem of driving comparatively capacitive loads arises when signals must be propagated from the chip to off chip destinations.

Generally, off-chip capacitances may be several orders higher than on-chip C_g value.

Ex: C_L denotes off chip load then

$$C_L \geq 10^4 C_g \text{ (typically)}$$

→ capacitances of this order must be driven through low-resistances, otherwise excessively long delays will occur.

Cascaded Inverters as drivers :-

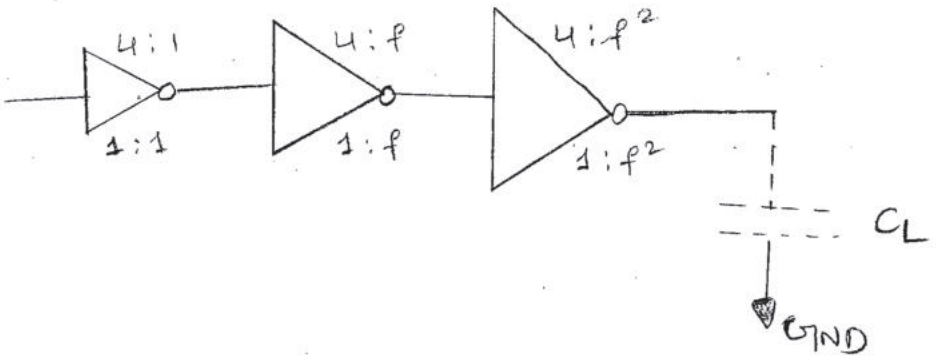
Inverters intended to drive large capacitive loads must therefore present low pull-up & pull-down resistance.

Low resistance values for Z_{pd} and Z_{pu} imply low L:w ratio's.

→ channel must be made very wide to reduce resistance value and in consequence, an inverter to meet this need occupies a larger area.

→ moreover, because of large L:w ratio and since length L cannot be reduced below the minimum feature size, the gate region area L×w becomes significant and a comparatively large capacitance is presented at the input, which in turn slows down the rate of change of voltage which can take place at the input.

Remedy: Use N cascade inverters, each one of which is larger than the preceding stage by a width factor f.



Driving Large capacitive loads

As the width factor increases, the capacitive load presented at the inverter input increases and area occupied increases also. Equally clearly, the rate at which the width increases (ie, value of f) will influence the number N of stages which must be cascaded to drive a particular value of C_L.

With large f, N decreases but delay per stage increases.

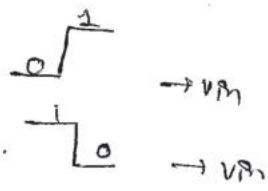
For 4:1 nmos inverters

delay per stage = fT for ΔV_{in}

or = $4fT$ for ∇V_{in}

ΔV_{in} = indicates logic 0 to 1 transition.

∇V_{in} = " logic 1 to 0 " of V_{in}.



∴ Total delay per nmos pair = $5fT$.

My treatment yields delay per cmos pair = $7fT$.

Let $y = \frac{C_L}{\square C_g} = f^N$

so that the choice of f and N are interdependent.

We need to determine the value of f which will minimize delay for a given value of y & from the definition of y

$$\ln(y) = N \ln(f)$$

i.e, $$N = \frac{\ln(y)}{\ln(f)}$$

Thus for N even

$$\text{total delay} = \frac{N}{2} 5fT = 2.5 N f T \text{ (nmos)}$$

$$\text{or} = \frac{N}{2} 7fT = 3.5 NfT \text{ (cmos)}$$

In all cases,

$$\text{delay} \propto NfT = \frac{\ln(y)}{\ln(f)} fT$$

total delay minimized if f assumes the value e (base of natural logarithms).

i.e, each stage should be $\simeq$ 2.7* times wider than its predecessor.

assuming that $f = e$, we have

$N = \ln(y)$ & overall delay t_d.

N even: $t_d = 2.5 e_{NT}$ (nmos)

$t_d = 3.5 e_{NT}$ (cmos)

N odd: $t_d = [2.5(N-1)+1] e_T$ (nmos)

$t_d = [3.5(N-1)+2] e_T$ (cmos)

} for ΔV_{in}

or $t_d = [2.5(N-1)+4] e_T$ (nmos)

$t_d = [3.5(N-1)+5] e_T$ (cmos)

} for ∇V_{in}

Super buffers :-

The asymmetry of the conventional inverter is clearly undesirable, and gives rise to significant delay problem when an inverter is used to drive more significant capacitive loady.

→ A common approach used in nmos technology to alleviate this effect is to make use of supper buffers.

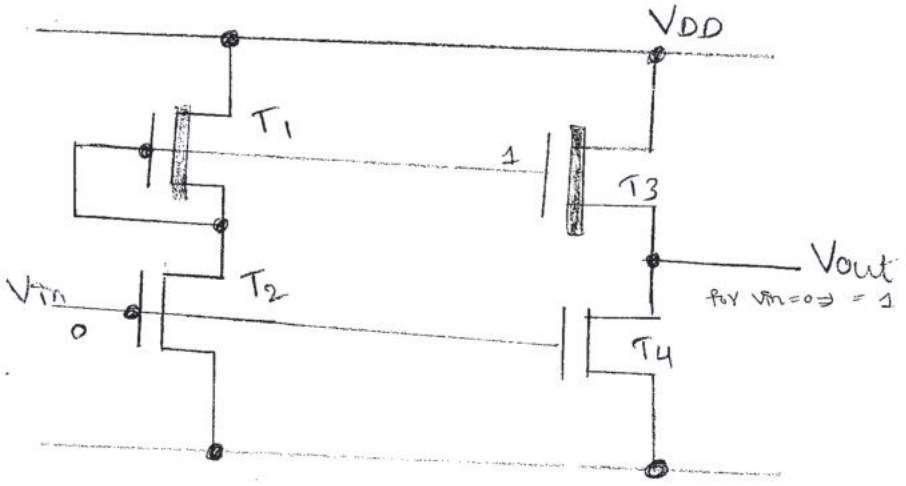


fig: Inverting type nmos super buffer

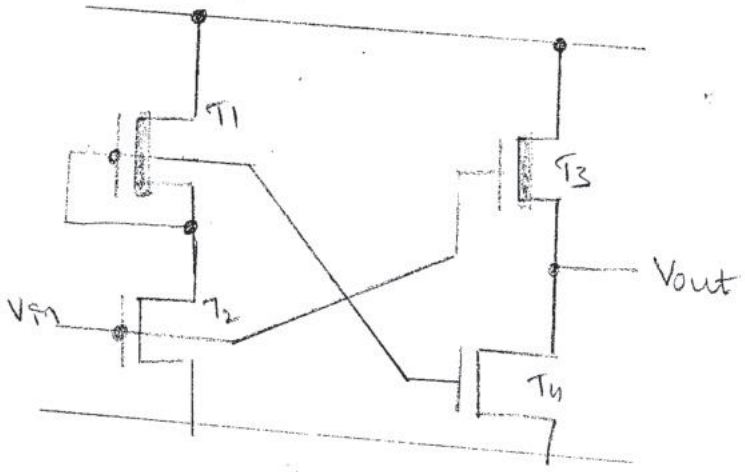


fig: Non-Inverting type nmos super buffer

Inverting type;-

Considering a positive going logic transition Vin at the input, (logic1), it will be seen that the inverter formed by T_1 and T_2 is turned on & thus the gate of T_3 is pulled down toward 0V with a small delay. Thus T_3 is cut off while T_4 (the gate of which is also connected

to Vin) is turned on and output is pulled down quickly.

When Vin drops to 0V,

then the gate of T_3 is allowed to rise quickly to V_{DD}. Thus as T_4 is also turned off by Vin, T_3 is made to conduct with V_{DD} on its gate i.e, with twice the average voltage that would apply if the gate was tied to the source as in the conventional nmos inverter.

Now, since I_{ds} & V_{gs} then doubling the effective V_{gs} will increase the current and thus reduce the delay in charging any capacitance on the o/p, so that male symmetrical transitions are achieved. $I = C \cdot \frac{dv}{dt}$

Non-inverting :-

Vin = 0V.

then T_2 open & T_1 conduct & it will turn on T_4 with V_{DD}. Then T_3 is nonconducting and T_4 is connected to o/p. Hence we get o/p = 0V through T_4.

When Vin = logic 1.

then T_2 ON

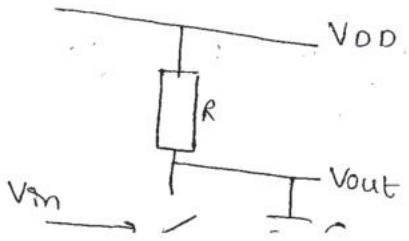
Bicmos drivers :-

The availabilpty of bipolar transistors in Bicmos technology presents the possibility of using bipolar transistor drivers of the output stage of inverter and logic gate circuits.

Bipolar transistors have an exponential dependence of the output current I_C on the input base to emitter voltage V_{be}. This means that the device can be operated with much smaller input voltage swing than Mos transistors and still switch relatively large currents.

Important thing to consider is the possible effect of temperature T on the required input voltage V_{be}. Although V_{be} is logarithmically dependent on base width W_B, doping level N_A, $\bar{e}$ mobility μ_n & collector current I_C it is only linearly dependent on T.

The switching performance of a transistor driving a capacitive load may be visualized initially from the simple model.



The time necessary to change the output voltage by an amount that is equal to the input change is given by

$$\Delta t = \frac{C_L}{g_m}$$

g_m = device transconductance

→ The time Δt necessary to change the output voltage Vout by an amount equal to the input voltage Vin is given by

$$\Delta t = \frac{C_L}{g_m}$$

g_m — transconductance of bipolar transistor.

→ Transconductance of bipolar transistor relatively high, hence the value of Δt is small.

A more exacting appraisal of the bipolar transistor delay reveals that it comprises 2 main components.

(1) T_{in} — initial time necessary to charge the base-emitter junction of the npn transistor. Typically, for Bicmos Transistor-based driver we are considering T_{in} is in the region of 2ns.

→ Similarly, consideration of a cmos transistor driver in the same Bicmos Technology would reveal a figure of 1ns for T_{in}, this being the time taken to charge the input gate capacitance.

⇒ Another significant parameter contributing to delay is the collector resistance (R_C) of bipolar Transistor.

→ High value of R_C will mean a long propagation delay through a transistor when charging a capacitive load.

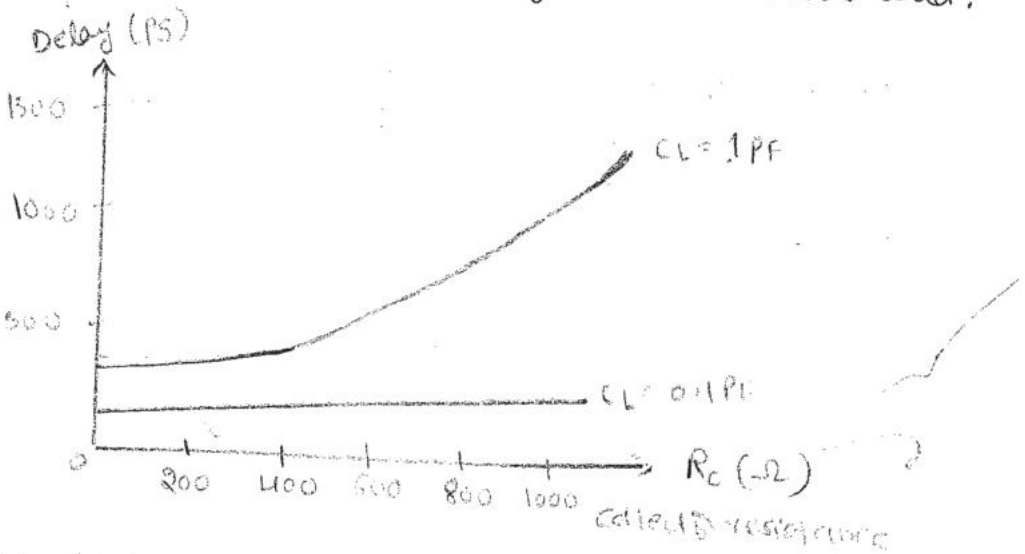


Fig: Gate delay as a function of collector resistance

→ The reason for including the buried subcollector region in the Bicmos process is to keep R_C as low as possible.

→ Bicmos fabrication processes produce reasonably good bipolar transistors – high g_m, high β, high h_{fe} & Low R_C – without compromising or overelaborating the basic cmos process.

→ The availability of bipolar transistors in logic gate and driver/buffer design provides a great deal of scope and freedom for VLSI designer.

(2) T_L – the time taken to charge the output load capacitance C_L & it will be noted that this time is less for the bipolar driver by a factor of hfe, where hfe is bipolar transistor gain.

Combined Effect of T_{in} & T_L :-

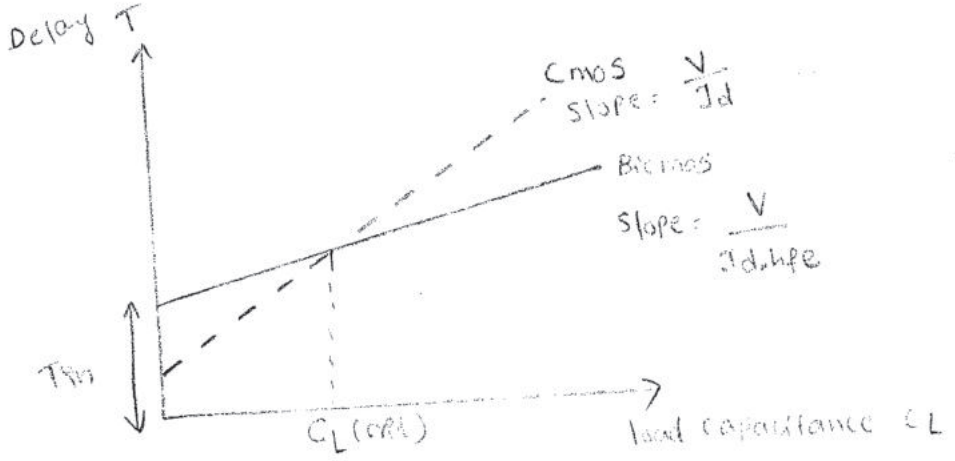


Fig: Delay estimation

Delay of Bicmos inverter can be described by

$$T = T_{in} + \left(\frac{V}{I_d}\right)\left(\frac{1}{h_{fe}}\right) C_L$$

where T_{in} – time to charge up base/emitter junction

hfe = Transistor current gain (CE)

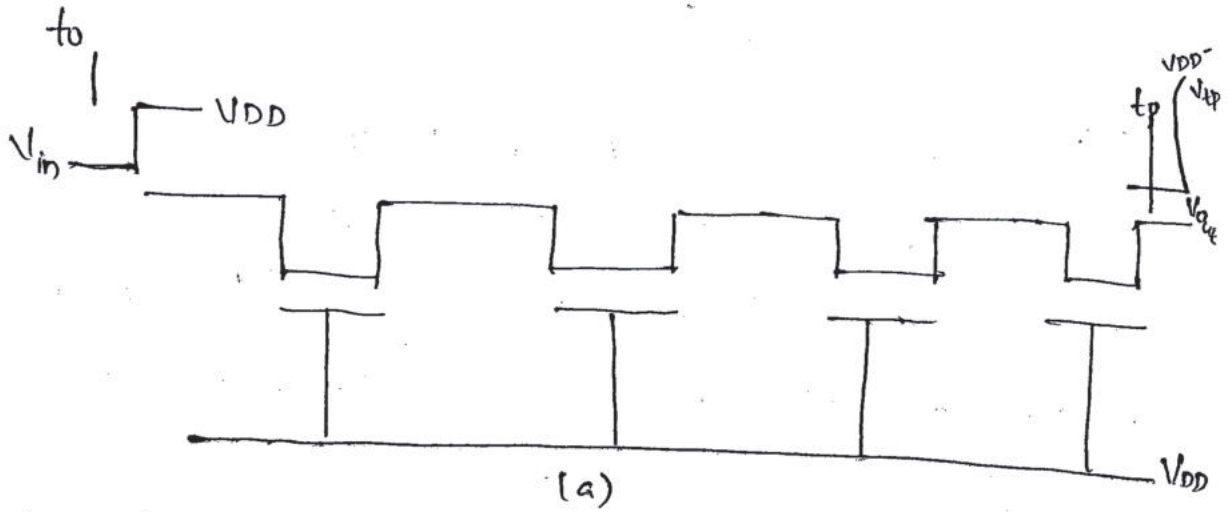
Hence, delay for Bicmos inverter is reduced by a factor of hfe compared with a cmos inverter.

C_L (crit) : The value of load capacitance below which the Bicmos driver is slower than a comparable cmos driver.

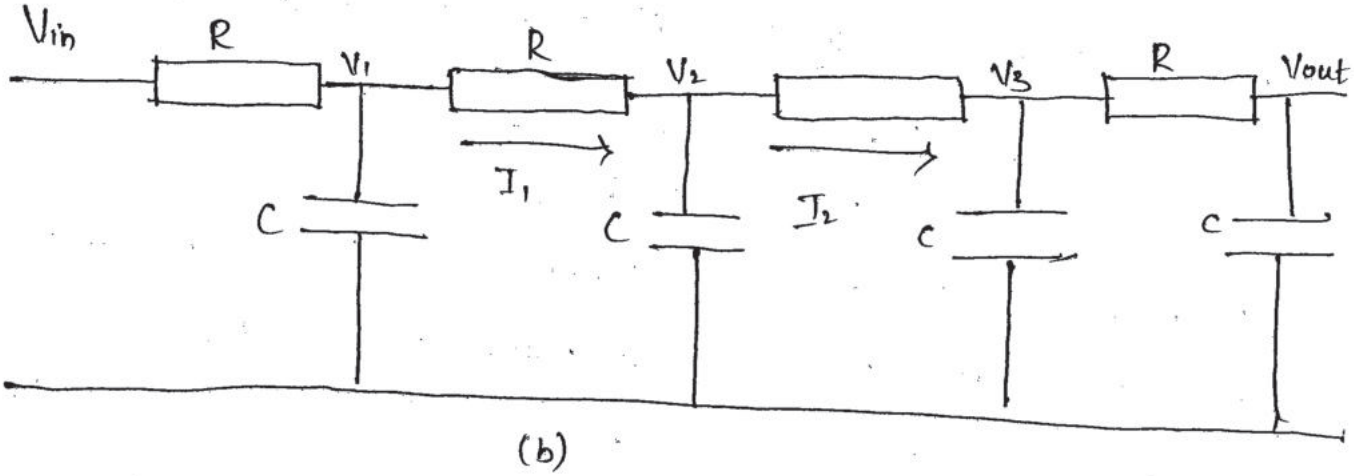
Propagation Delays :-

Cascaded pass Transistors :-

A degree of freedom offered by Mos technology is the use of pass transistors as series / parallel switches in logic arrays. Quite frequently, therefore, logic signals must pass through a number of pass transistors in series. A chain of four such transistors is shown in fig. 5.29 (a)



(a)



(b)

fig : 5.29 (a) and (b) propagation delays in pass transistor chain

in which all gates [illegible]
which would be the case for a signal to be propagated to the output. The circuit thus formed may be modelled as in fig. 5.29 (b) and it is then possible to evaluate the delay through the network.

$$\frac{c\,dv_2}{dt} = (I_1 - I_2) = \frac{(V_1 - V_2) - (V_2 - V_3)}{R}$$

In the limit as the number of sections in such a network becames large, this expression reduces to

$$Rc\,\frac{dv}{dt} = \frac{d^2v}{dx^2}$$

where,

R = Resistance per unit length

c = Capacitance per unit length

x = distance along network from input

The propagation time t_p for a signal to propagate a distance x is such that $t_p \propto x^2$

The analysis can be simplified if all R_s and C_s are lumped together, then

$$R_{total} = n r R_s$$

$$C_{total} = n c \square c_g$$

Where r gives the relative resistance per section in terms of R_s & c gives the relative capacitance per section in terms of $\square c_g$. Then, it may be shown that over all delay t_d for n sections - is given by

$$t_d = n^2 rc(\tau)$$

Thus, the over all delay increases rapidly as n increases & in practice no more than four pass transistors should be normally connected in [illegible] however this number can be exceeded if a buffer is [illegible]

5.4.2 Design of long polysilicon wires :

Long polysilicon wires also contribute distributed series R and C as was the case for cascaded pass transistors and in consequence signal propagation is slowed down. This would also be the case for wires in diffusion where the value of C may be quite high, and for this reason the designer is discouraged from running signals in diffusion except over very short distances.

for long polysilicon runs, the use of buffers is recommended In general, the use of buffers to drive long polysilicon runs the use of buffers is recommended. In general, the use of buffers to drive long polysilicon runs have two desirable effects. first, the signal propagation is speeded up & , Second there is a reduction is Sensitivity to noise.

The reason why noise may be a problem with slowly rising signals may be deduced by Considering fig. 5.30. In the diagram, the slow rise-time of the signal at the input of the inverter (to which the signal emerging from the long polysilicon line is connected) means that the input voltage spends a relatively long time in the vicinity of V_{inv} so that small disturbances due to noise will switch the inverter state between '0' and '1' as shown at the output point.

Thus it is essential that the long polysilicon wires be driven by suitable buffers to guard against the effects of noise and to stop up the rise time of propagated signal

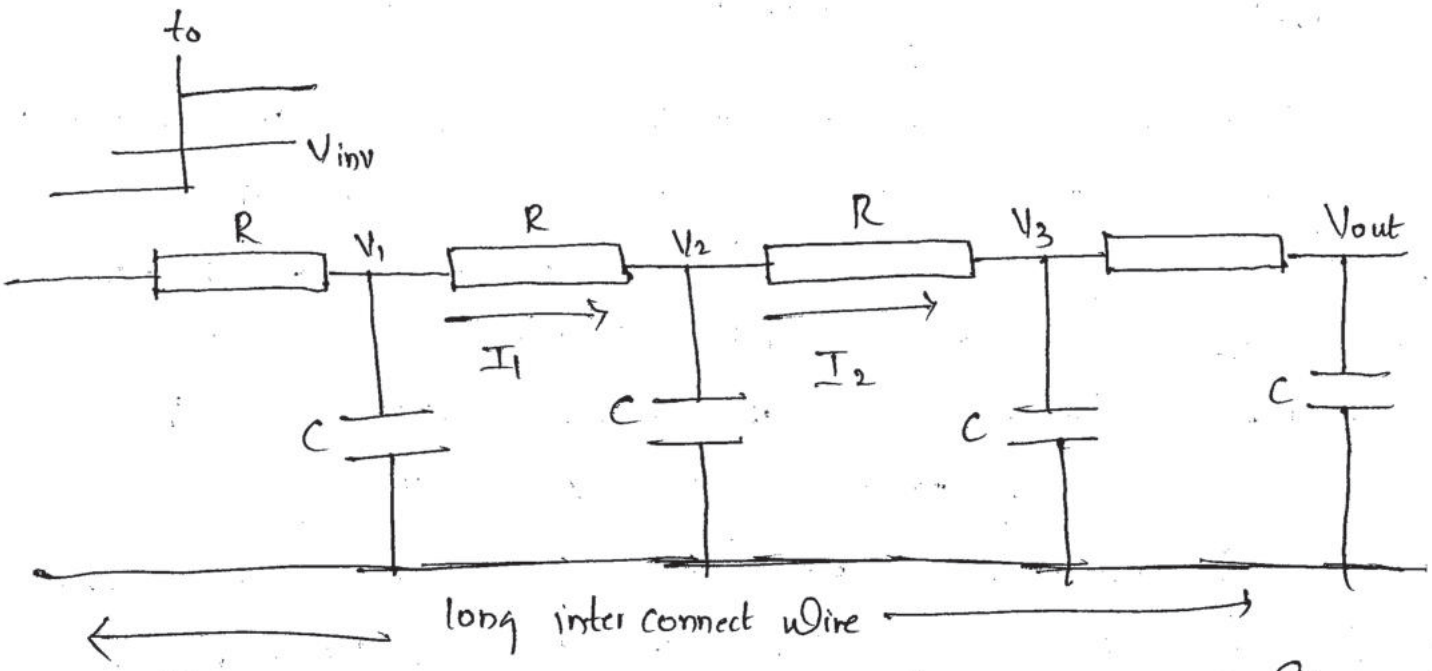


fig: 5-30 Possible effects of delays in Polysilicon wires

Wiring capacitances :-

There are significant sources of capacitance which contribute to the overall wiring capacitance. Three such sources are

1. Fringing fields
2. Interlayer capacitances.
3. Peripheral capacitance.

Fringing fields :-

Capacitance due to fringing fields effects can be a major component of the overall capacitance of interconnect wires.

→ For fine line metallization, the value of fringing field capacitance (C_{ff}) can be of the same order as that of the area capacitance.

$$C_{ff} = \epsilon_{sio_2} \epsilon_0 l \left[\frac{\pi}{\ln\left\{1 + \frac{2d}{t}\left(1+\sqrt{\left(1+\frac{t}{d}\right)}\right)\right\}} - \frac{t}{4d} \right]$$

l – wire length

t – thickness of wire

d – wire to substrate seperation.

Then, total wire capacitance, $C_w = C_{area} + C_{ff}$

Interlayer capacitances:-

→ Obviously, the parallel plate effects are present between one layer and another.

→ For example, some thought on the matter will confirm the fact that, for a given area,

metal to polysilicon capacitance > metal to substrate capacitance.

→ The reason for not taking such effects into account for simple calculations is that the effects occur only when layers cross or when one layer underlies another, & therefore interlayer capacitance is highly dependent on Layout.

→ However for regular structures it is readily calculated & contributes significantly to the accuracy of circuit modeling and delay calculation.

Peripheral capacitance:-

→ The source and drain n-diffusion regions form junctions with the p-substrate or p-well at well-defined and uniform depths.

Similarly, for P-diffusion regions in n-substrate or n-wells.

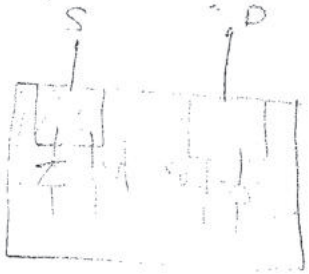
→ For diffusion regions, each diode thus formed has associated with it a perpheral capacitance in pF (picofarady) per unit length.

→ This can be considerably greater than the area capacitance of the diffusion region to substrate.

→ The smaller the source or drain area, the greater becomes the relative value of the peripheral capacitance.

In order to calculate the total diffusion capacitance we must add the contributions of area and peripheral components,

$$C_{total} = C_{area} + C_{periph.}$$



Typical values for diffusion capacitance:

Diffusion capacitance	Typical value		
	5 lim	2 lim	1.2 lim
Area c (C_{area})	1.0×10^{-4} pF/lim^2	1.75×10^{-4} pF/lim^2	3.75×10^{-4} pF/lim^2
Periphery capacitance (C_{periph})	8.0×10^{-4} pF/lim^2	negligible	negligible

Switch Logic :-

To build switches from MOS Transistors one way is "Transmission gate", built from parallel n-type and p-type transistors.

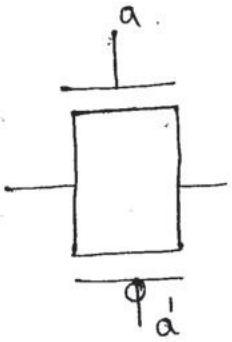


Fig: Complementary Transmission gate

→ This switch is built from both types of Transistors so that it transmits logic 0 and 1 from drain to source equally well.

→ when we put a V_{DD} or V_{SS} at the drain, we get V_{DD} or V_{SS} at the source.

→ But it requires 2 transistors and their associated tubs. equally damning, it requires both true and complement forms of the gate signal.

An alternative to buid switches from MOS Transistors is the "n-type switch" — a solitary n-type transistor.

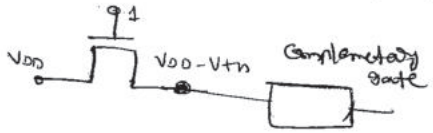
→ It requires only one transistor and one gate signal,

→ but it is not as forgiving electrically.

→ It transmits a logic 0 well, but when V_{DD} is applied to the drain, the voltage at the source is $V_{DD} - V_{th}$.

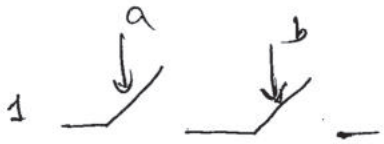
→ When switch logic drives gate logic, n-type switches can cause electrical problems.

→ An n-type switch driving a complementary gate causes the complementary gate to run slower when switch input is 1. since the n-type pulldown current is weaker when a lower gate voltage is applied, the complementary gate's pulldown will not suck current off the output capacitance as fast.

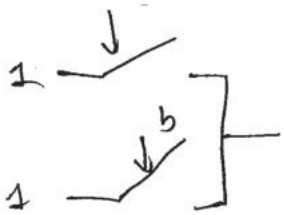


→ A pseudo-nmos g is driven by n-type switch, disaster may occur. A pseudo-nmos gate's ratioed transistors depend on logic 0 and 1 inputs to occur within a prescribed voltage range.

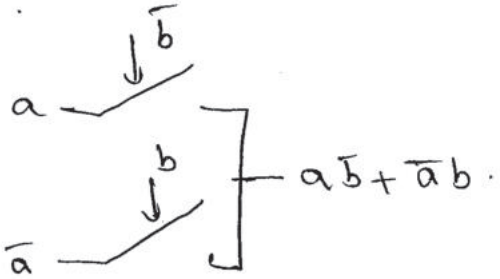
→ If the n-type switch doesn't turn on the pseudo-nmos pulldown strongly enough, the pulldown may not divert enough current from the pull-up to force the output to a logic 0, even if we wait forever.



Pseudo-AND



Pseudo - OR



switch n/w with non-constant source inputs.

Several important alternative CMOS gate topologies. Each has important uses in chip design. But it is important to remember that they all have their limitations and caveats. particular care must be taken when mixing logic gates designed with different circuit topologies to ensure that one's output meets the requirments of the next's input's

i) Pseudo - nMOS Logic:

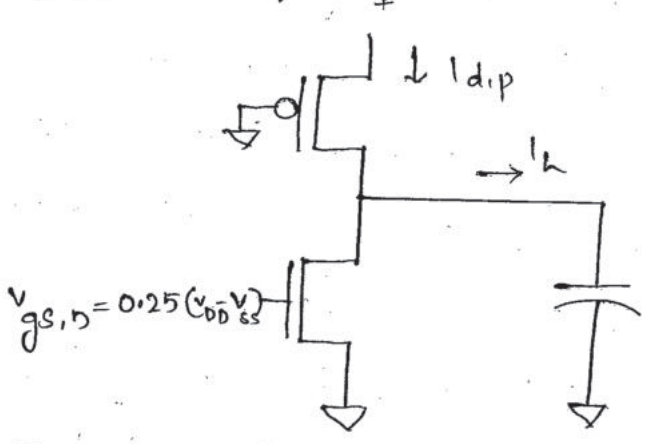


Fig: Pseudo nmos NOR gate

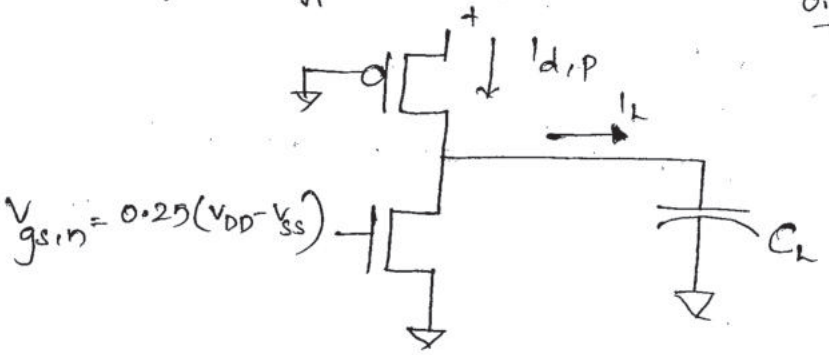
The simplest non-standard gate topology is pseudo-nMOS, so called because it mimics the design of an nMOS logic gate. The pulldocon network of the gate is the same as for a fully complementary gate. The pullup network is replaced by a single p-type transistor whose gate is connected to V_{ss}, leaving the transistor permanently on. The p-type transistor is used as a resistor. When the gate's inputs are ab=00, both n-type transistores are off and the p-type

transistor pulls the gate's output up to V_{DD}. [illegible] is 1, both the p-type and n-tpye transistor are on and both are fighting to determine the gate's output voltage.

We need to determine the realationship between the W/L ratio of the pullup & and the pulldowns which provide reasonable output voltage for the gate. For simplicity, assume that only one of the pulldown transistors & is on; then the gate circuit's output voltage depends on the ratio of the effective resistance of the pullup and the operating pulldown. The high output voltage of the gate is V_{DD}, but the output low voltage V_{OL} will be some voltage above V_{ss}. The chosen V_{OL} must be low enough to activate the next logic gate in the chain. For pseudo-nMOS gate which feed static or pseudo-nMOS gate, a value of $V_{OL} = 0.15(V_{DD} - V_{SS})$ is a reasonable value, though others could be chosen. To find the transistor sizes which give reasonable output voltages, one must consider the simultaneous operation of the pullup and pulldown. When the gate's output has just switched to a logic 0, the n-type pulldown is in saturation with $V_{gs,n} = V_{in}$. The p-type pullup is in it's linear region: it's $V_{gs,p} = V_{DD} - V_{SS}$ and it's $V_{ds,p} = V_{out} - (V_{DD} - V_{SS})$. We need to find Vout in terms of the W/Ls of the pullup and pulldown. To solve this problem, we set the currents through the saturated pulldown and the linear pullup to be equal.

$\frac{W_n}{L_n}(V_{gs,n} - V_{tn}) = \frac{1}{2} k'_p \left[2(V_{gs,p} - V_{tp}) V_{ds,p} - V^2_{ds,p}\right]$ — (1)

The pulldown network must exhibit this effective resistance in the worst case combination of inputs. Therefore, if the network contains series pulldowns, they must be made larger to provide the required effective resistance.



Tech
0.5 µm: Sub, $V_{dd} = 3.3V$,

$V_{gs,n} = V_{DD} - V_{SS}$ in (1)

we find that

$$\frac{W_p/L_p}{W_n/L_n} \simeq 3.9$$

As shown in figure so long as the pulldown drain current is significantely less than the pullup drain current, there will be enough current to charge the output capacitance and bring the gate output to the desired level.

The ratio of the pullup & pulldown sizes also ensures that the times for $0 \rightarrow 1$ & $1 \rightarrow 0$ transitions are asymmetric. Since the pullup transistor has about three times the effective resistance of the pulldown, the $0 \rightarrow 1$ transition occures much more slowely then the $1 - 0$ transition and dominates the gate's delay. The long pullup time makes the pseudo-nMOS gate slower than the static complementary gate.

The main advantage of pseudo-nMOS gate is

the small size of the pullup network, both in terms of number of devices and wiring complexity. The pullup network of a static complementary gate can be large for a complex function. The input signals do not have to be routed to the pullup, as in a static complementary gate. The pseudo-nMOS gate is used for circuits where the size and wiring complexity of the pullup network are major concerns but speed and power are less important.

ii) DCVS logic (Differential Cascode Voltage Switch logic)

→ DCVS logic is a static logic family that has a very different structure.

→ It uses a latch structure for the pullup which both eliminates static power consumption and provides true and complement outputs

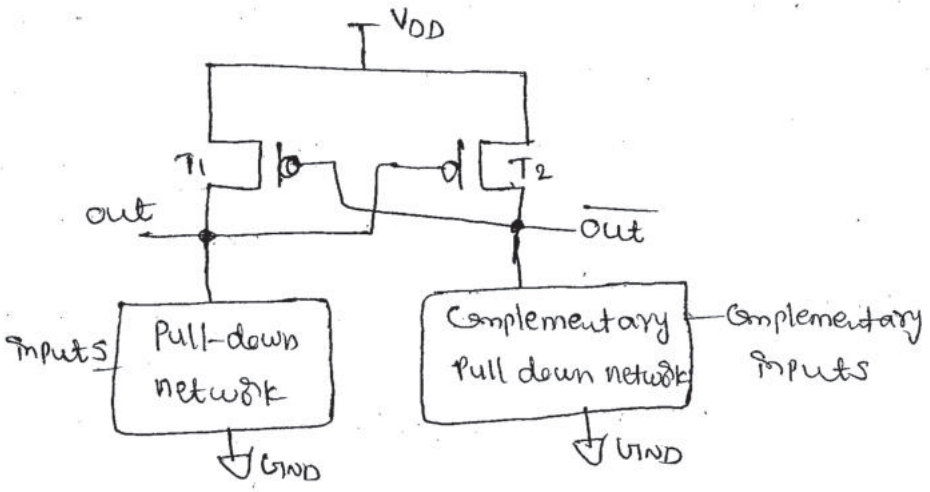
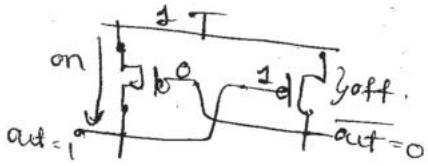


Fig: Structure of a DCVS gate.

→ There are 2 pull-down networks which are dualy of each other.

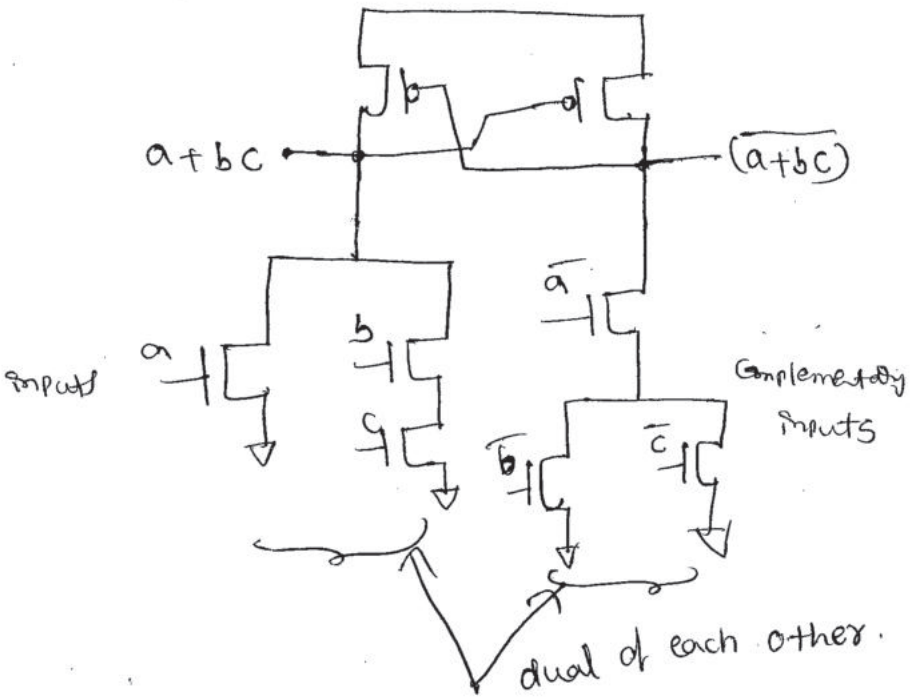
→ Each pull-down network has a single p-type pull-up, but the pull-ups are crosscoupled. Exactly one of the pulldown n/w will create a path to ground when the gate's inputs change, causing the output nodes to switch to the required values.

→ The cross-coupling of the pull-ups helps to speedup the transition, Ex, If the complementary n/w forms a path to ground, the $\overline{out}$ goes toward Vss, which turns on the true output's pullup (T_1), raising the True output, which in turn lowers the gate voltage on the complementary output's pullup. (T_2).



→ This gate consumes no DC power, since neither side of gate will ever have both its pullup and pull down network at once.

DCVSL gate, which computes $(a+bc)$ on one output & $\overline{(a+bc)} = \bar{a}\bar{b}+\bar{a}\bar{c}$ on its other output.



$= \overline{a+bc}$

$= \overline{(a+b)(a+c)}$

(demorgan's law $\overline{ab} = \bar{a}+\bar{b}$)

$= \overline{a+b} + \overline{a+c}$

$= (\bar{a}\bar{b}) + (\bar{a}\bar{c})$

$= \bar{a}(\bar{b}+\bar{c})$

$\bar{a}$ is seriesed

iii) Domino Logic :-

Precharged circuits offer both low area and higher speed than static complementary gates. Precharged gates introduce functional complexity because they must be operated in 2 distinct phases, requiring introduction of a clock signal.

The canonical precharged logic gate circuit is the domino circuit.

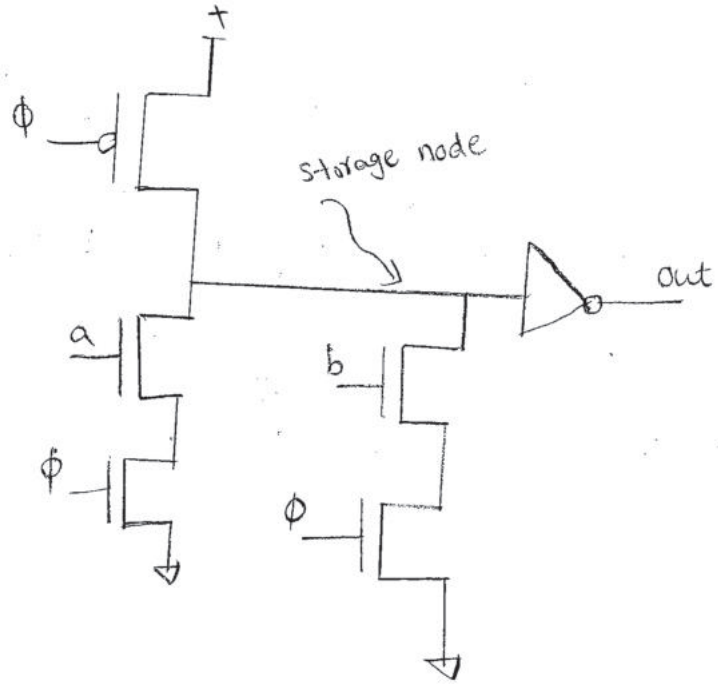


fig: domino OR gate

→ The gate works in 2 phases, first to precharge the storage node, then to selectively discharge it. The 2 phases are controlled by the clock signal ϕ.

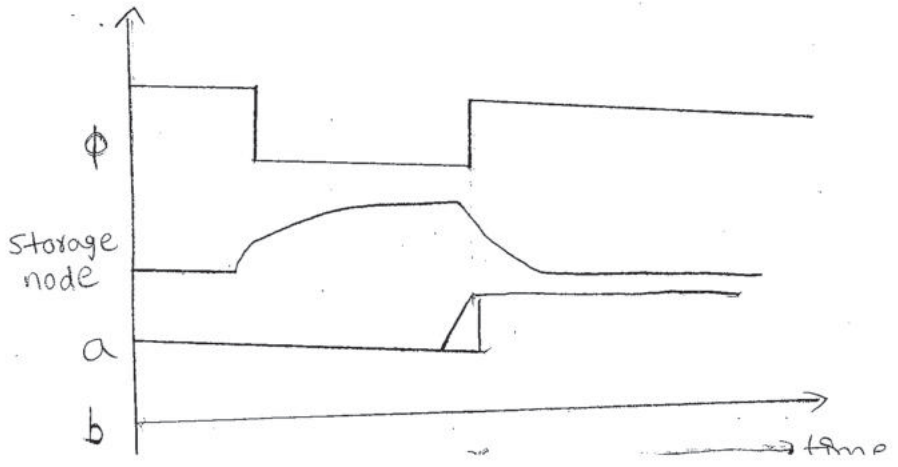
Precharge: When ϕ goes low, the p-type transistor starts charging the precharge capacitance. The pull down transistors controlled by the clock keep that precharge node from being drained. The length of the $\phi=0$ phase is adjusted to ensure that the storage node is charged to a solid logic 1.

Evaluate: when ϕ goes high, precharging stops (P-mos off) and the evaluation phase begins (n type pulldown ON).

→ The logic inputs a and b can now assume their desired value of 0 or 1.

→ The input signals must monotonically rise – if an input goes from 0 to 1 and back to 0.

→ If the inputs create a conducting path through the pulldown network, the precharge capacitance is discharged forcing its value to 0 and the gate's o/p is 1 (through the inverter).



→ If neither a nor b is 1, then the storage node would be left charged at logic 1 and the gate's output would be 0.

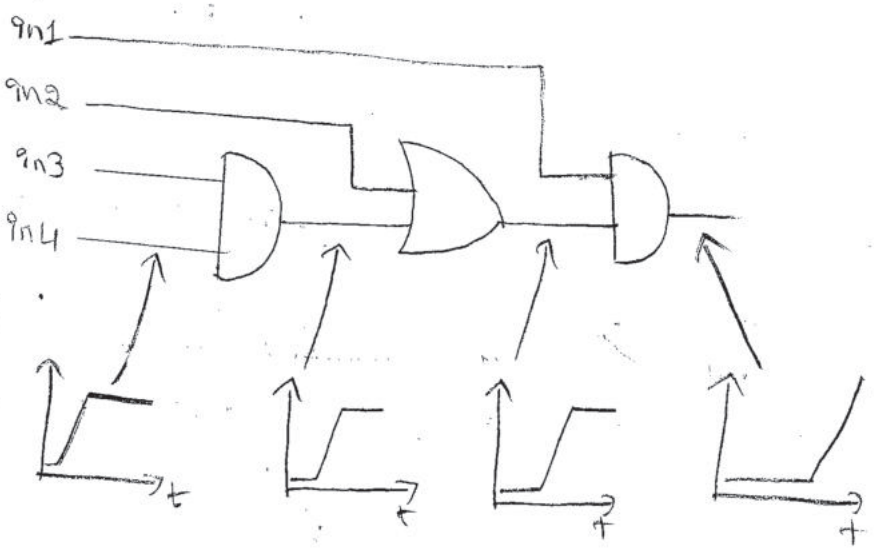


Fig: Successive evaluations in a domino logic network

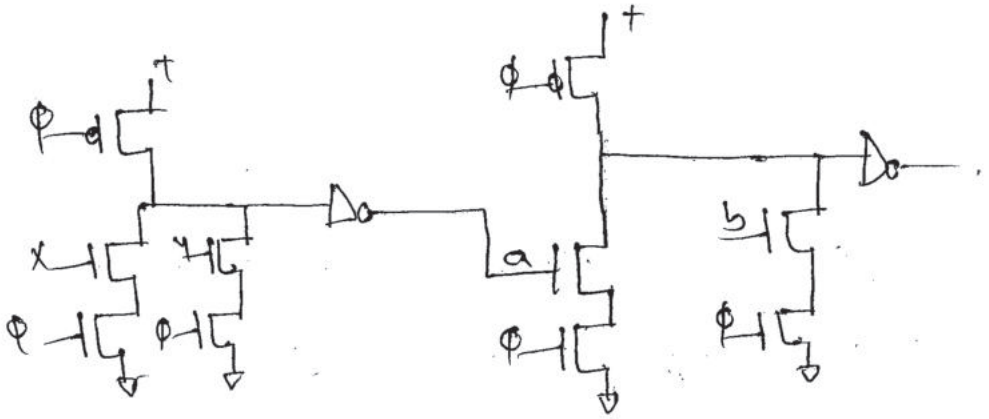
The figure illustrates the phenomenon which gave the domino gate its name. Since each gate is precharged to a low output level before evaluation, the change at the primary inputs ripple through the domino network from one end to another.

→ signals at the far end of the network change last, with each change to a gate output causing a change to the next output. This sequential evaluation resembles a string of falling dominos.

Need of inverter at the output of the domino gate:

Reason 1: logical operation and circuit behaviour.

If output of one domino gate is fed into an input of another domino gate, then during precharge phase, if the inverter were not present, the intermediate signal would rise to 1, violating the requirement that all inputs to the second gate be '0' during precharging.



Reason 2: To increase the reliability of the gate.

MOST chips are built from collection of subsystems like adders, regs, state machines etc.

Digital fns can be divided into:

- datapath operators
- memory elems
- ctrl structures
- I/o cells.

→ Area & delay costs can be reduced by optimizatn of diff levels of abstraction:

layout, circuit, logic, Register transfer.

SHIFTERS

→ shifters are imp elems in many microprocr designs for arithmetic shifting, logical shifting & rotation functions.

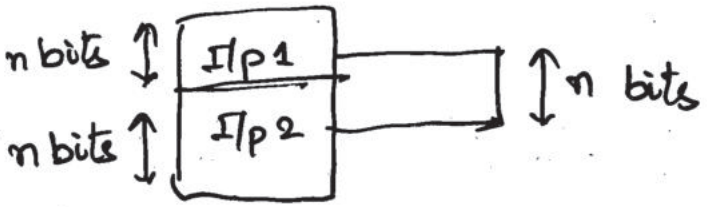
→ A barrel shifter can perform n-bit shifts in a single combinational fn.

→ Barrel shifter accepts '2n' data bits & 'n' ctrl signals and produces 'n' o/p bits.

→ It shifts by transmitting an n-bit slice of the '2n' data bits to the o/ps.

→ The position of the transmitted slice is determined by the ctrl bits; exact opern is determined by values placed at the data I/ps.

ex:- Right shift with zero fill.



data word 'd' into top I/p.
all 0's into bottom I/p.

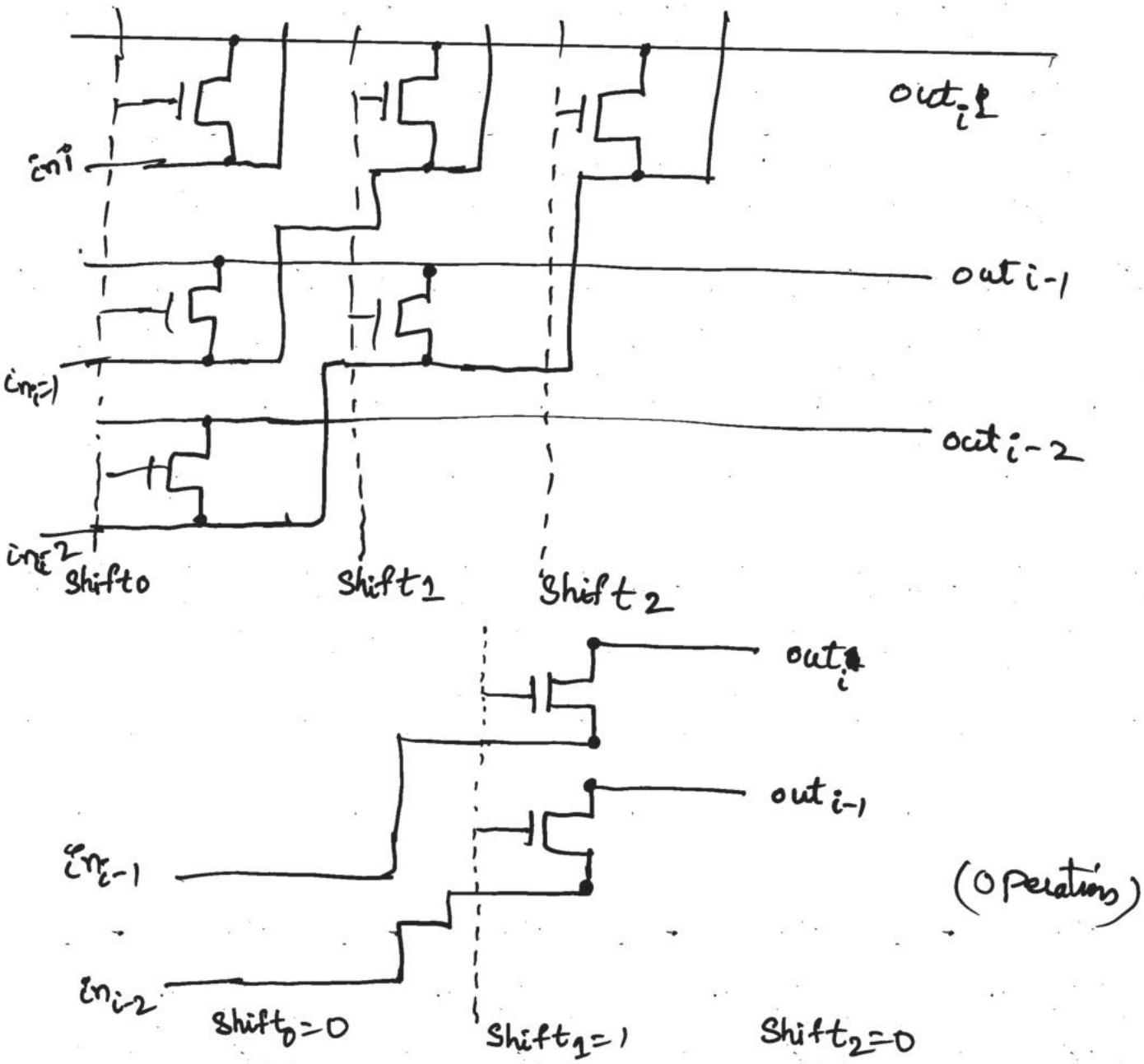
The o/p is right shift with zero fill.

→ by setting ctrl bits to select top-most 'n' bits is shift of zero.

Selecting bottom 'n' bits is an n-bit shift pushing entire word out of the shifter.

(2) Rotate Operation

When both top & bottom I/ps have same data. Rotate operation shifting out top bits of word causing those bits to reappear at the bottom of o/p



(Operation)

→ Barrel shifter with 'n' o/p bits is built from '2n' vertical by 'n' horz'l array of cells, each has single Tr & few wires.

→ Cell is Tx^n gate formed using single n-type Tr.

→ Ctrl lines run vertically; I/p run diagnally upward thru the system; o/p horizontally.

Ctrl lines set to '1', turns T_x^n gate in single column.

The T_x^n gate connect diagonal I/p wires to horz o/p wires; when col is turned on, all I/ps are shunted to the o/ps.

The length of the shift is determined by position of the selected column.

ADDERS:

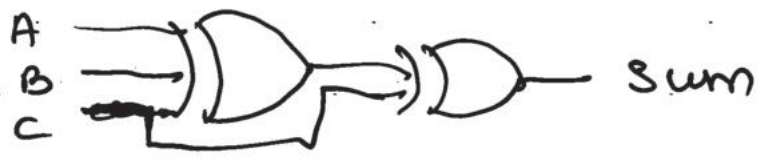
1) Full adder: 1-bit sum; 1-bit carry.

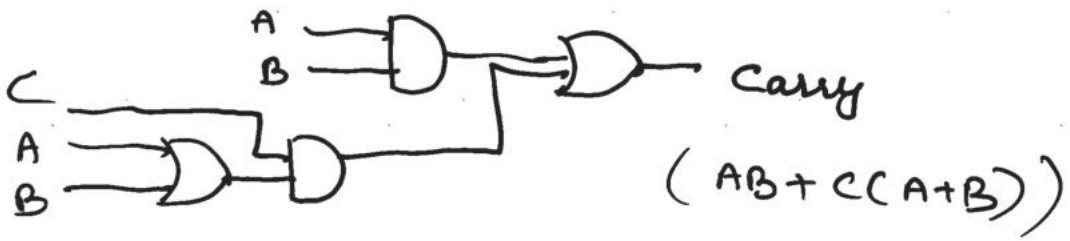
$$Sum = A\bar{B}\bar{C} + A\bar{B}C + \bar{A}\bar{B}C + \bar{A}B\bar{C}$$

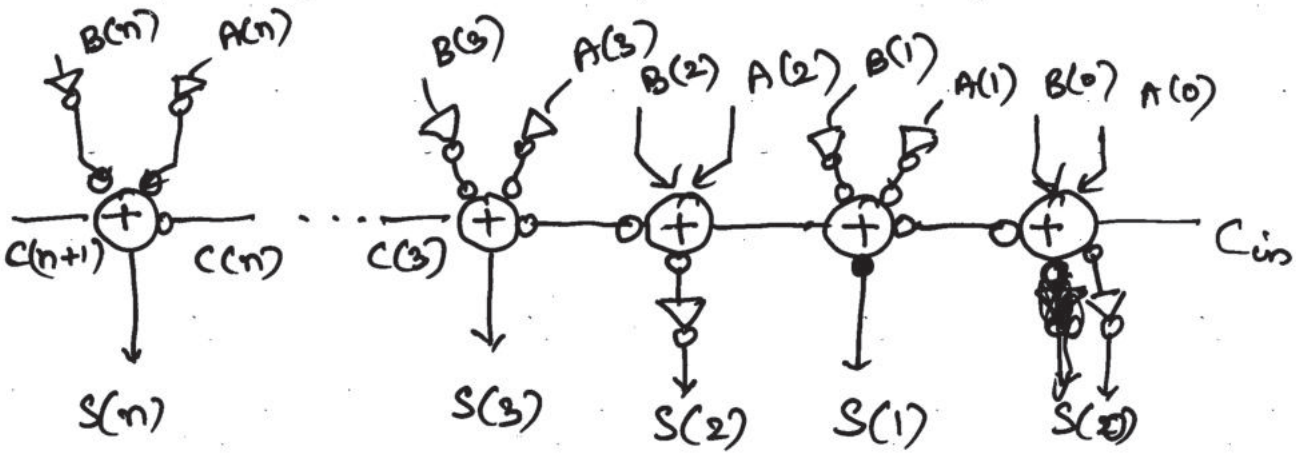
$$Sum_i = A_i \oplus B_i \oplus C_i$$

$$C_{i+1} = Carry_i = A_iB_i + B_iC_i + A_iC_i$$

Sum_i = Sum at ith stage, C_{i+1} = Carry.



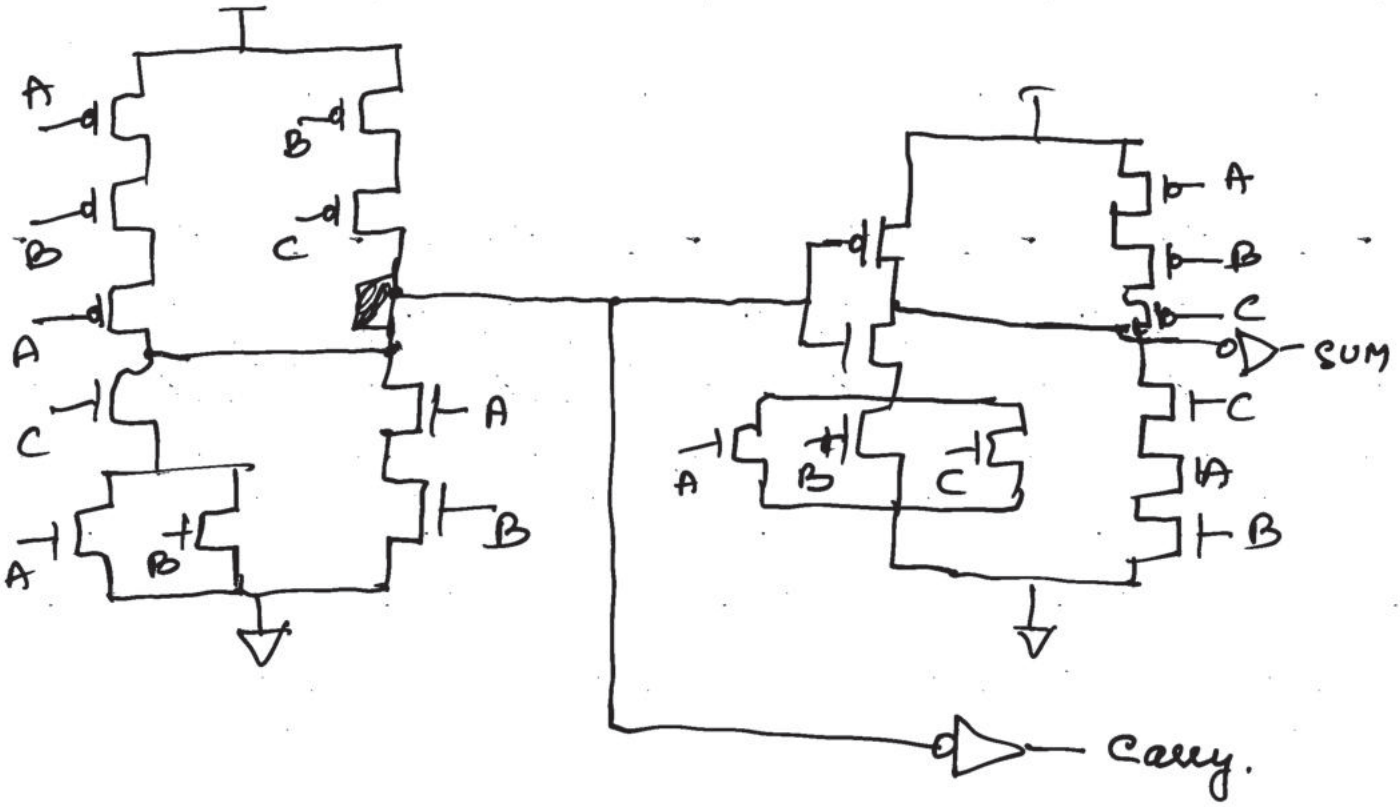




Single-bit adder (28 Trs)

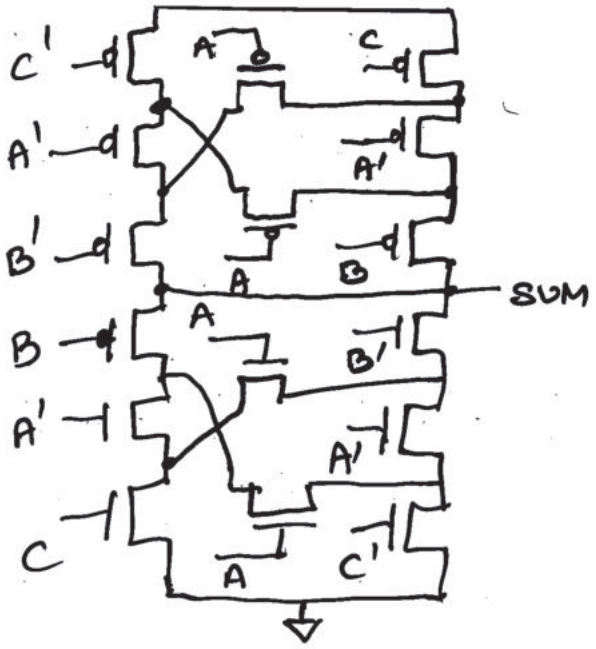
$$SUM = ABC + (A+B+C)\,\overline{CARRY}$$

$$R = ABC + (A+B+C)\left(\overline{AB + C(A+B)}\right)$$

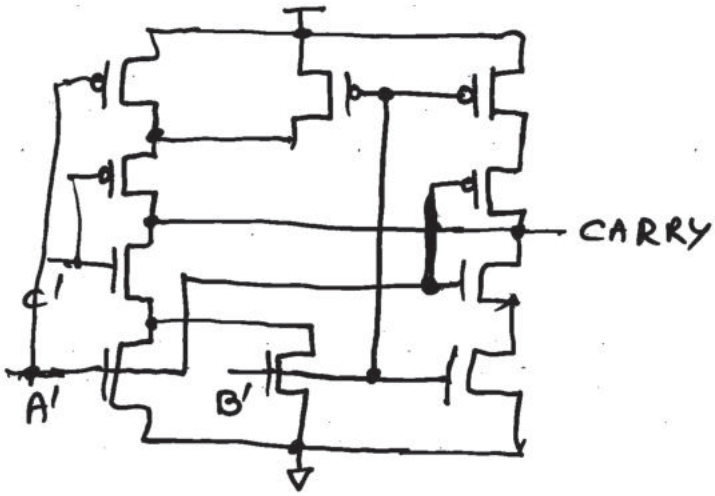


FA using XOR gates uses 32 transistors.

FA using cascaded logic gates uses 28 transistors.



$S = (A \oplus B) \oplus C = P \oplus C$



$Carry = AB + AC + BC$

$CY = \overline{A'B' + C'(A' + B')}$

(2) Bit-Parallel Adder (Ripple Carry Adder).

⇒ n-bit adder built by cascading 'n' 1-bit adders.

→ Ripple-carry adder is easy to design but is slow when 'n' is large.

→ The I/ps are n-bit A & B. The carry signal of stage 'i' is fed to the 'c' signal of stage 'i+1' & 'SUM' signal forms n bit o/p.

(3) n-bit subtractor.

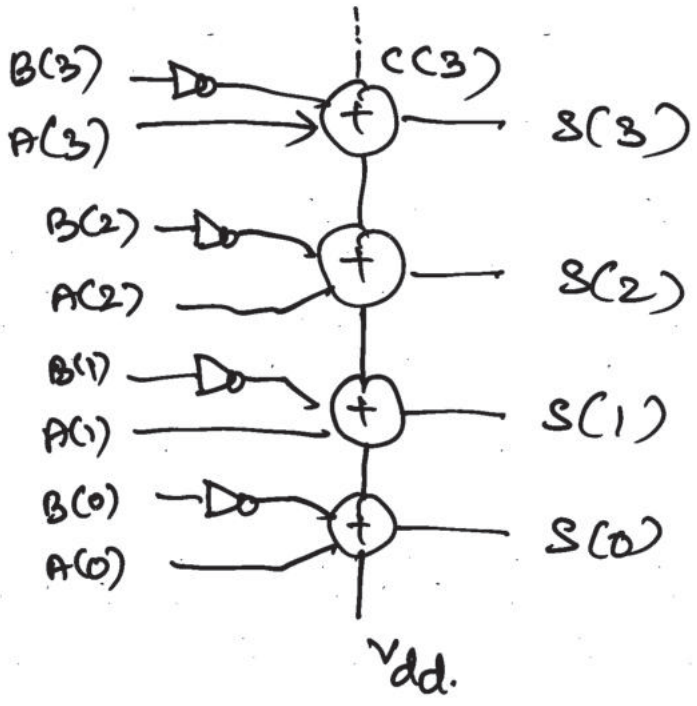
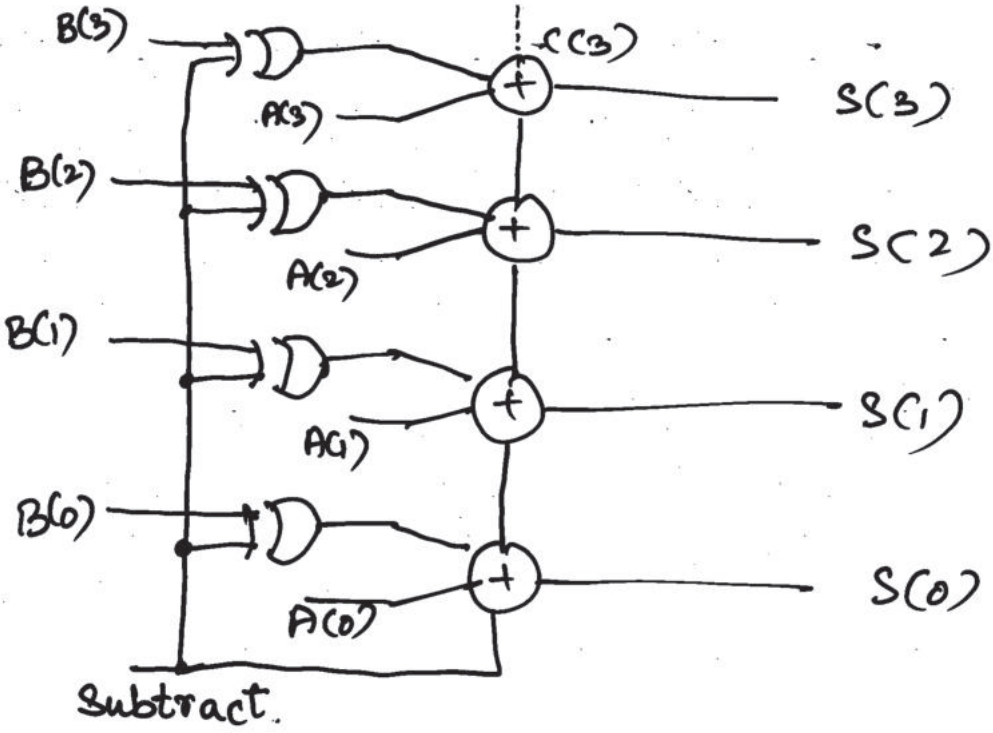


fig: A-B.

(4) Adder/Subtractor.

if subtract=0
S = A+B
else
S = A-B



(3) 4-bit subtractor.

fig: A-B.

(4) Adder/subtractor.

if subtract=0
S= A+B
else
S= A-B

(5) Carry-look ahead adder.

If a_i's, b_i's are I/Ps, P = Propagate
G = Generate

where

$$P_i = a_i + b_i$$
$$G_i = a_i b_i$$

Sum $S_i = c_i \oplus P_i \oplus G_i$

$$c_{i+1} = G_i + P_i c_i$$

$\therefore$

$$c_{i+1} = G_i + P_i (G_{i-1} + P_{i-1} \cdot c_{i-1})$$

$$c_{i+1} = G_i + P_i G_{i-1} + P_i P_{i-1} G_{i-2} + P_i P_{i-1} P_{i-2} c_{i-2}$$

Thus c_{i+1} depends on c_{i-2} & not on 'c_i' or 'c_{i-1}'.

→ The carry-look ahead units can be recursively connected to form a tree; each unit generates its own 'P' & 'G' values, which are used to feed carry-lookahead unit at the next level of the tree.

(7) Carry-Select Adder.

→ computes two versions of the addition with diff carry-ins, then selects the right one.

→ m-bit stages.

→ The 2nd stage computes 2 values: one assuming carry-in is '0' & another '1'.

→ The carry out of prev stage is used to select which version is correct: Muxes ctrlled by prev stage's carry-out choose correct sum & carry-out.

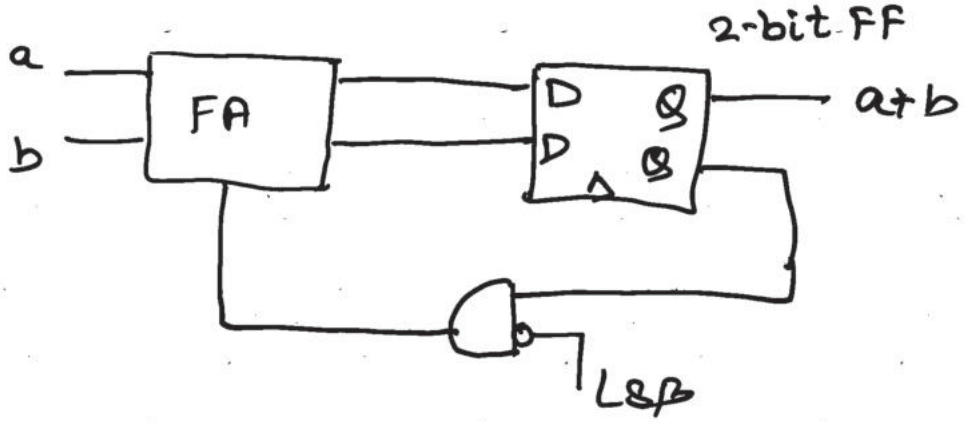
→ This speeds up addn coz ith stage can be computing two versions of the sum in parallel with i-1th's computation of its carry.

(8) Serial Adders:

→ Require many clk cycles to add two n-bit nos.

→ Small

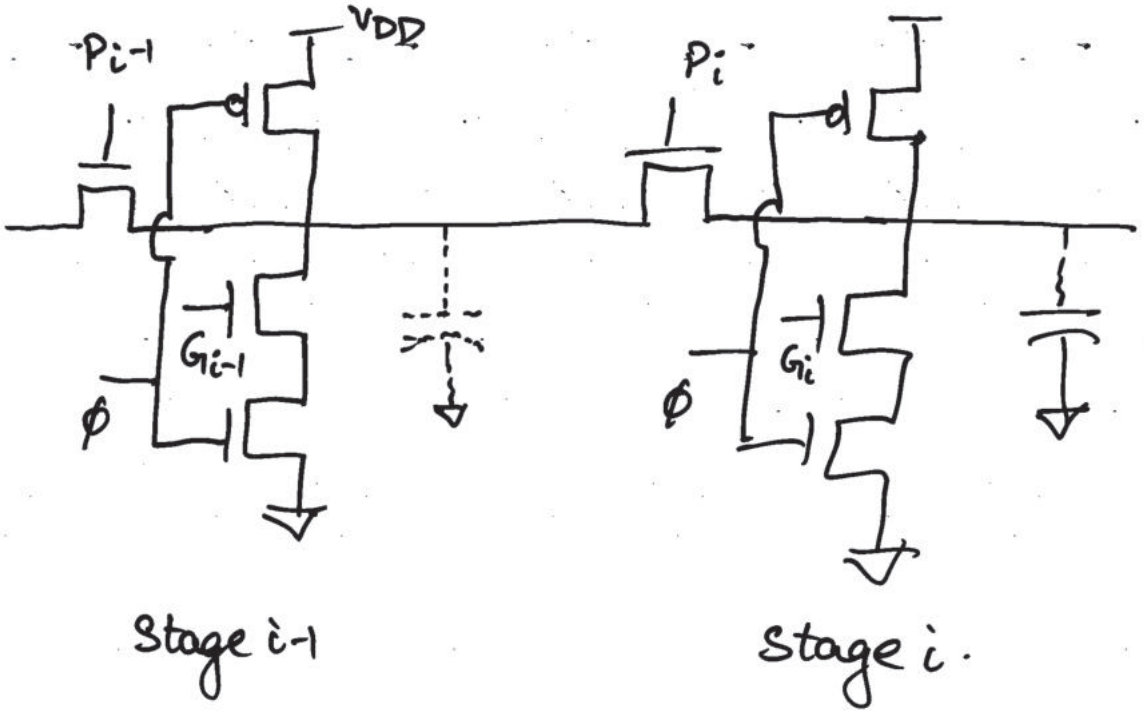
← data format



LSB = high when curr data bits are LSBs of the addends.

(9) Manchester Carry Chain:

→ Precharged adder.



Stage i-1 Stage i.

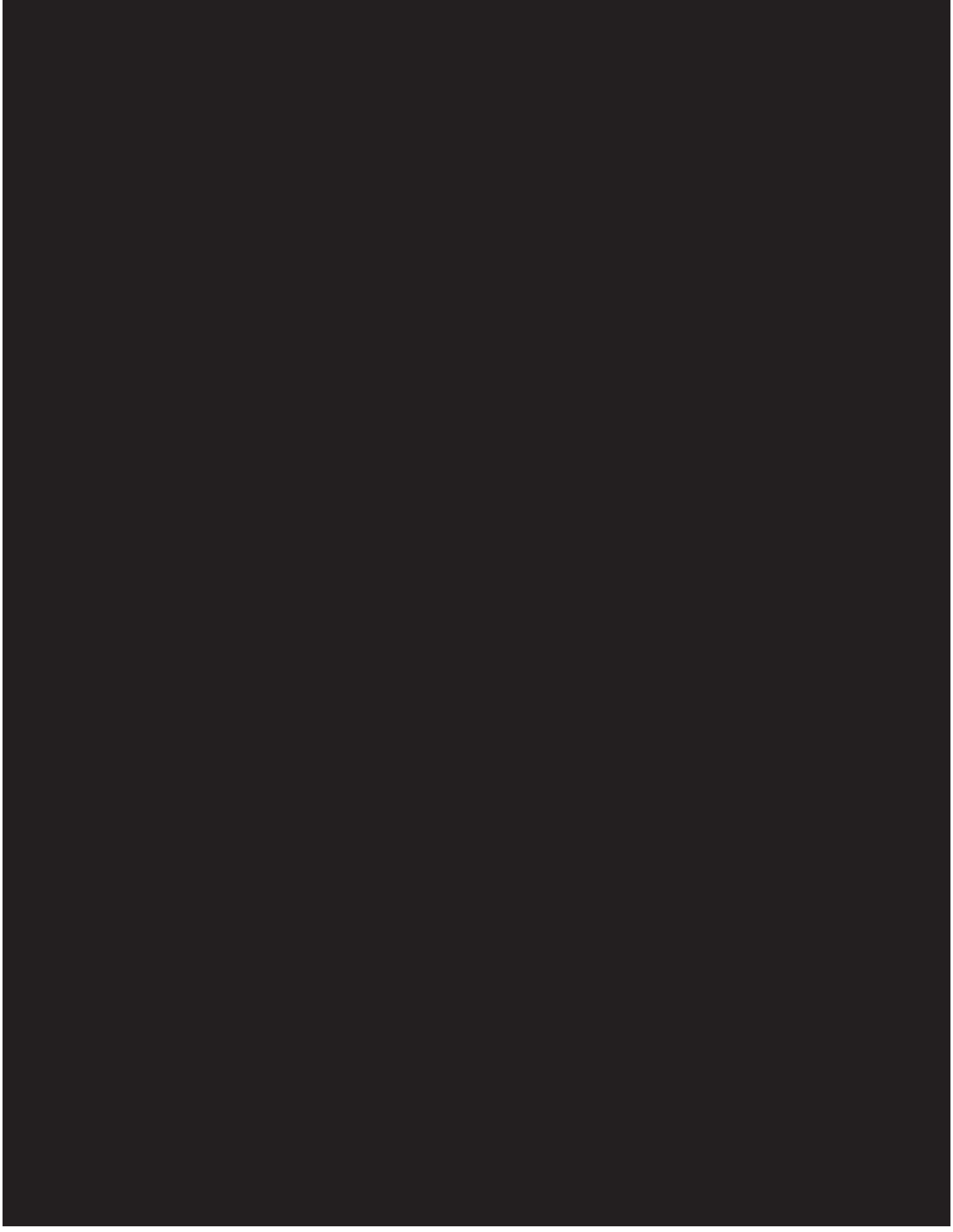
→ Storage node holds complementary (C_i') charged to '1' during precharge phase.

→ If $G_i = 1$, during evaluate stage, storage node is discharged, producing carry into next stage.

If $P_i = 1$, then ith storage node is connected to $(i-1)^{th}$ storage node; ith storage node can be discharged by P_{i-1} P.d.

→ 16-bit adders. (Pwr Consumpn)

adder	Pwr (mv)
* Ripple-Carry	0.117
* Const width carry-skip	0.109
* Carry-lookahead	0.171
* Carry-select	0.216



Carry-Save Adder :

⇒ Save carry & sum for each cycle.

→ n-bit adders: 'n' carries & 'n' sums.

→ n-bit CSA reqs '2n' registers.

→ reduces critical path. of reg, adder delay set up time into reg.

Tx - Gate Adder (TG)

(6 Trs)

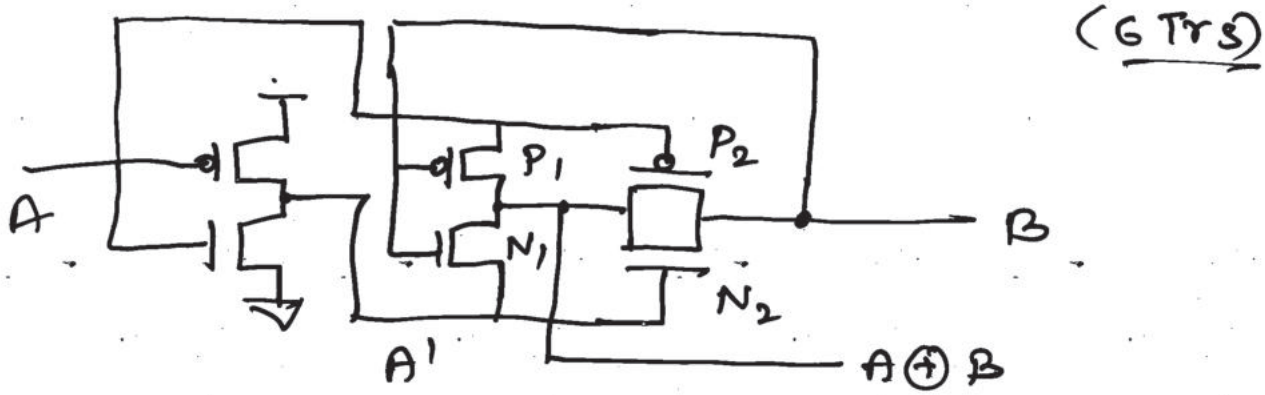


fig: Txn gate XOR. (Tiny XOR)

→ A= logic 1, A' = low, P_1 & N_1 : inverter with B' at o/p. Tx gate by pair P_2 & N_2 is open.

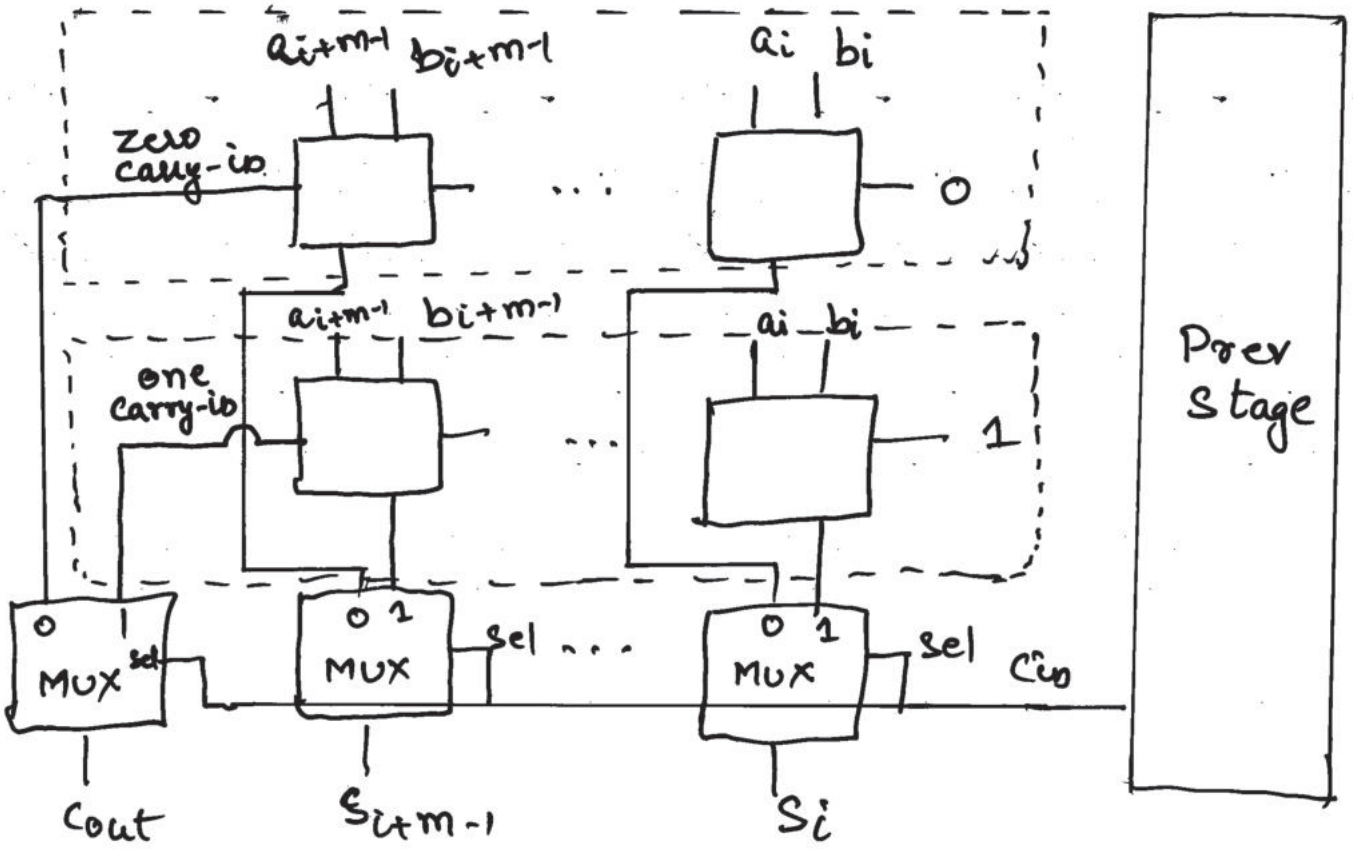
→ A = low, A' = high, (P_2, N_2) = closed, passing 'B' to o/p. The inv^r (P_1, N_1) is Partially disabled

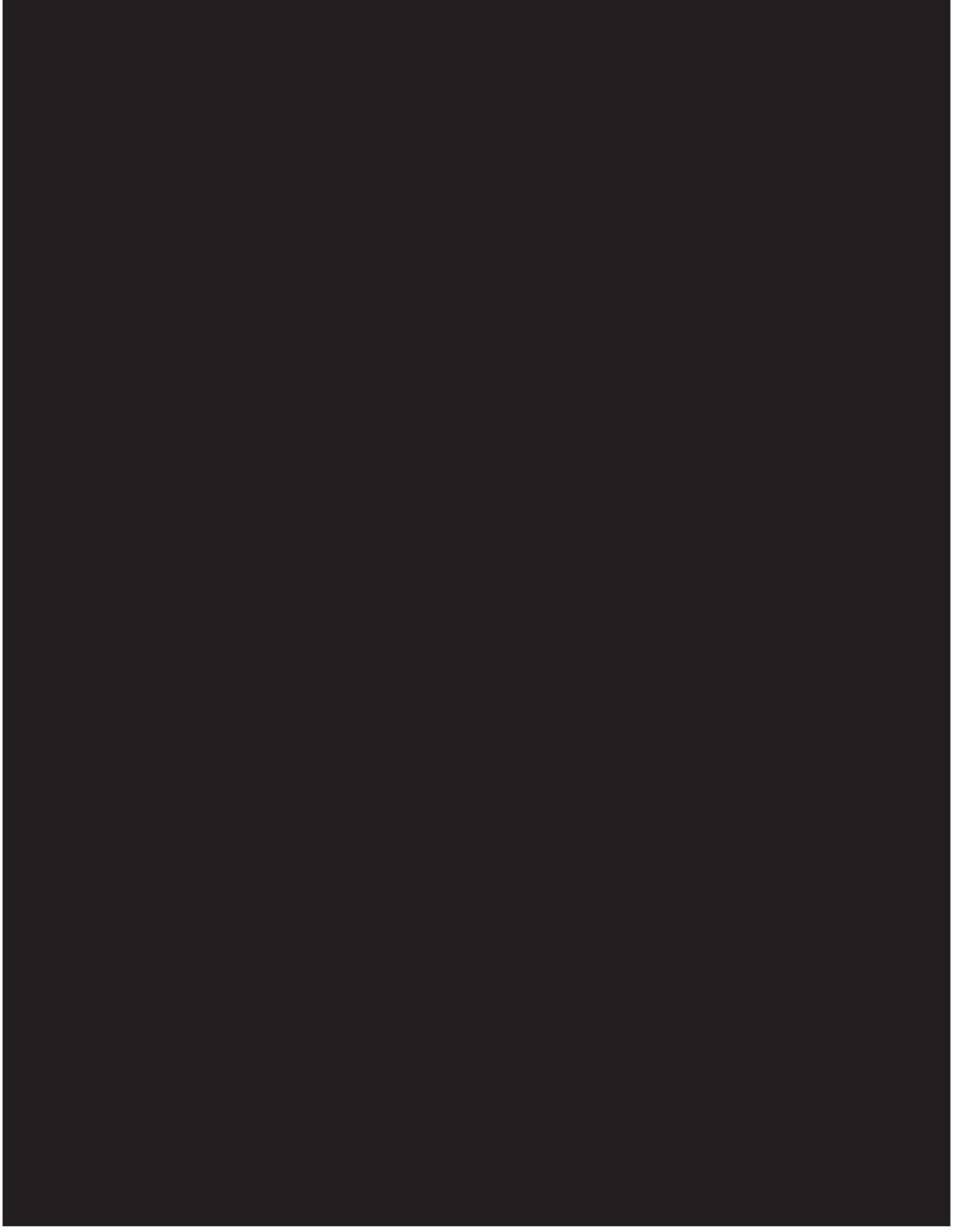
(6) Carry-Skip Adder:

→ This looks for cases where carry out of a set of bits is same as the carry in to those bits.

→ This adder is organized into m-bit groups; if the carry is propagated for every bit in the stage, then bypass gate sends the g stage's carry I/p directly to carry o/p.

→ divided into 2 groups of bits;
- * True carry into group
- * True Propagate cond^n 'P' at every bit in the group is needed to cause carry to skip.





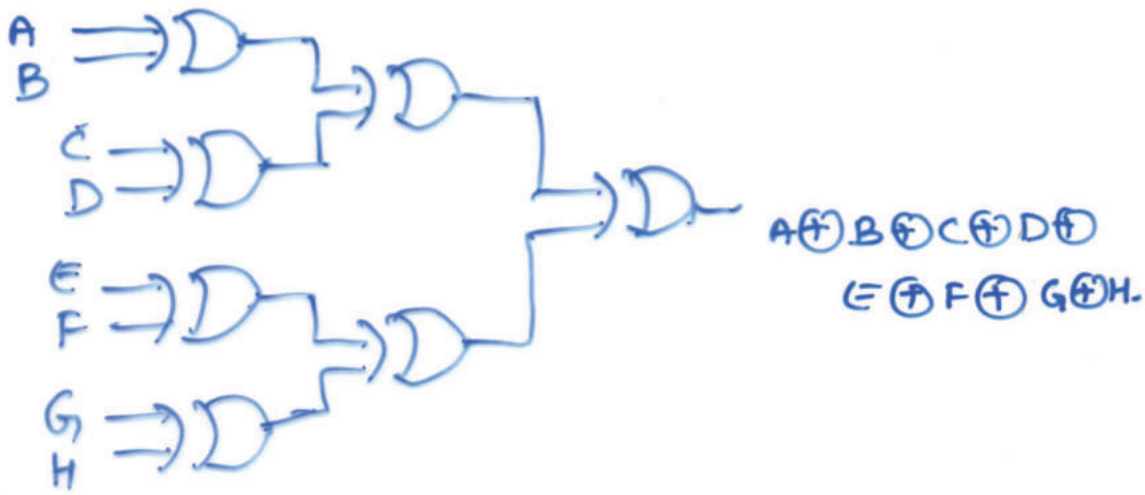
PARITY GENERATORS

Detects no. of 1s in s/p word as odd or even.

Function is

$$\text{Parity} = A_0 \oplus A_1 \oplus \cdots \oplus A_n.$$

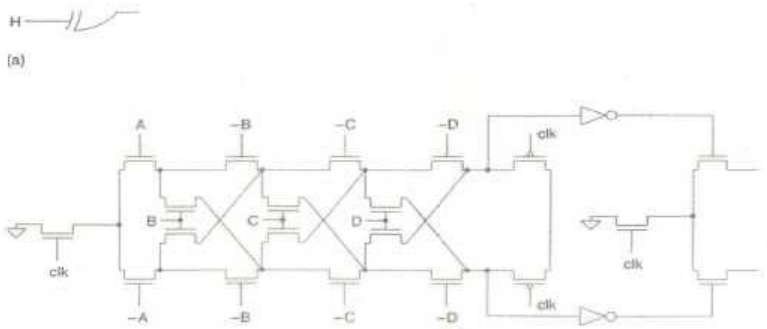
(a) Static XOR tree.



COMPARATOR:

→ Used to compare the magnitude of two binary nos.

→ It can be built using an adder and complementer.



(dual-rail)

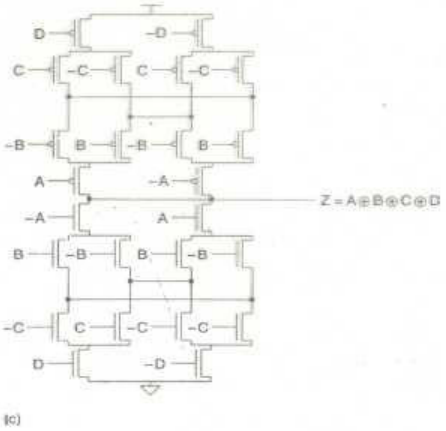


Figure Parity generation: (a) static XOR tree; (b) dynamic version; (c) static 4-input XOR

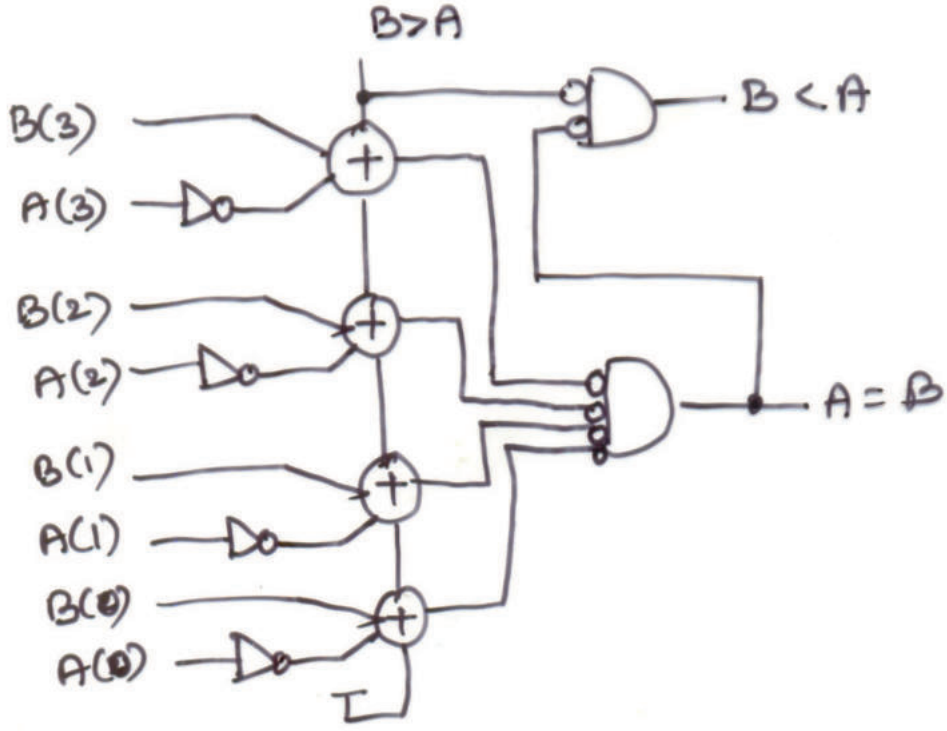


fig: Comparator using adder.

→ A zero detect (NOR gate) provides A=B signal while final carry o/p provides B>A signal.

→ other signals like A<B or A≤B, can be generated by logical combination of these signals.

→ For equality, need only XNOR & AND gate.

→ fig(b): Pass-gate logic. (Tx n gate)
draws no DC curr, slow for long comps.

→ fig(c): merged XNOR/NOR gate.
draws DC curr, bt small & fast.

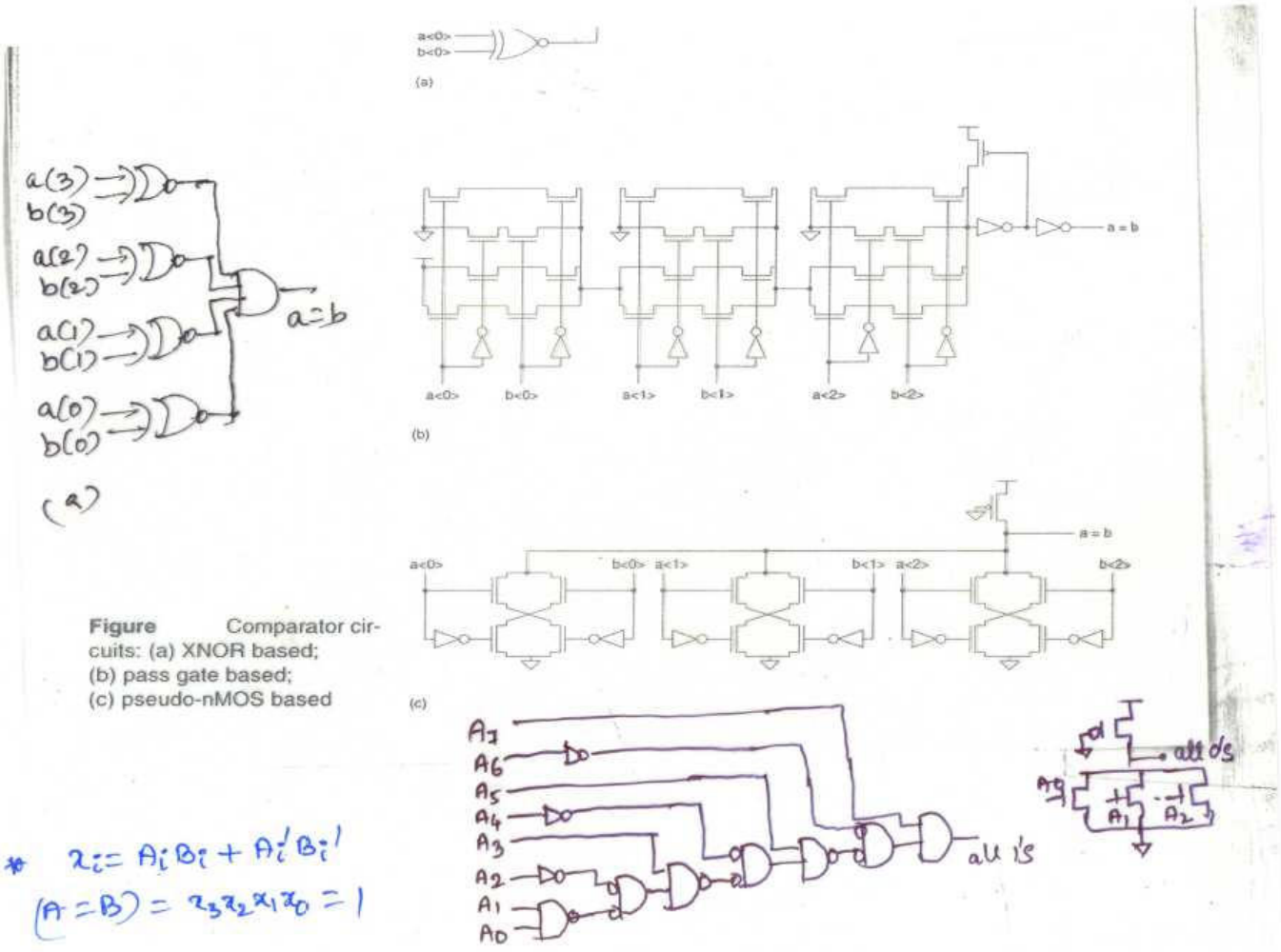


Figure Comparator circuits: (a) XNOR based; (b) pass gate based; (c) pseudo-nMOS based

* $z_i = A_i B_i + A_i' B_i'$

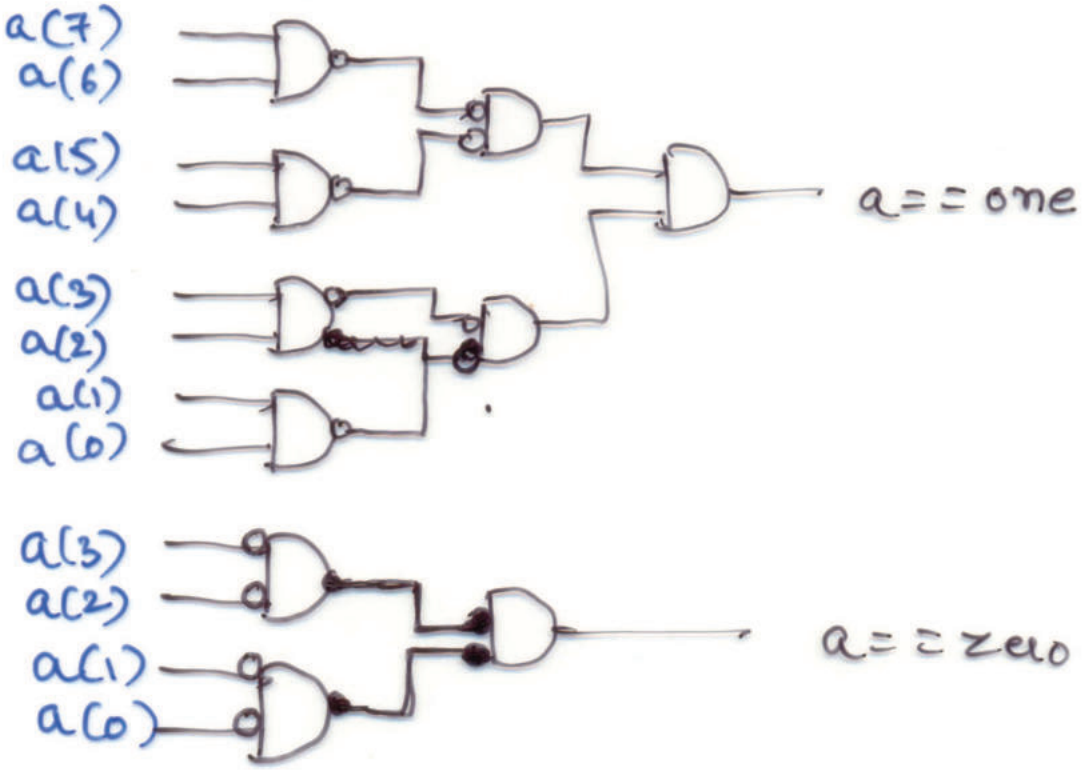
$(A = B) = z_3 z_2 z_1 z_0 = 1$

ZERO / ONE DETECTOR :

→ Detecting all ones or all zeros on wide words req large fan-in AND (or) OR gates.

→ Delay to o/p $\propto \log N$ (N = bit width of word)

→ pseudo-nMOS NOR gate implementation of zero/one detector is very fast & small for word widths less than 32 bits.



(using alternate NAND & NOR)

COUNTERS:

→ Count goes thru a sequence of binary nos.

Asynchronous Counters

→ o/ps change at varying times w.r.t clk cycle or edge.

ex:- ripple counter.

→ The clking of each stage is carried out by the previous counter stage, & thus the time it takes the last counter stage to settle can be quite large for a long counter chain.

Synchronous Counters:

→ o/ps change at substantially same time.

ex:- up/ down counter.

→ up/ down counter basically uses adder & a D-FF.

→ The speed is determined by the ripple-carry time from LSB to the MSB.

COUNTERS :

→ Count goes thru a sequence of binary nos.

Asynchronous Counters

→ o/ps change at varying times w.r.t clk cycle or edge.

ex:- ripple counter.

→ The clking of each stage is carried out by the previous counter stage, & thus the time it takes the last counter stage to settle can be quite large for a long counter chain.

Synchronous Counters :

→ o/ps change at substantially same time.

ex:- up/ down counter.

→ up/ down counter basically uses adder & a D-FF.

→ The speed is determined by the ripple-carry time from LSB to the MSB.

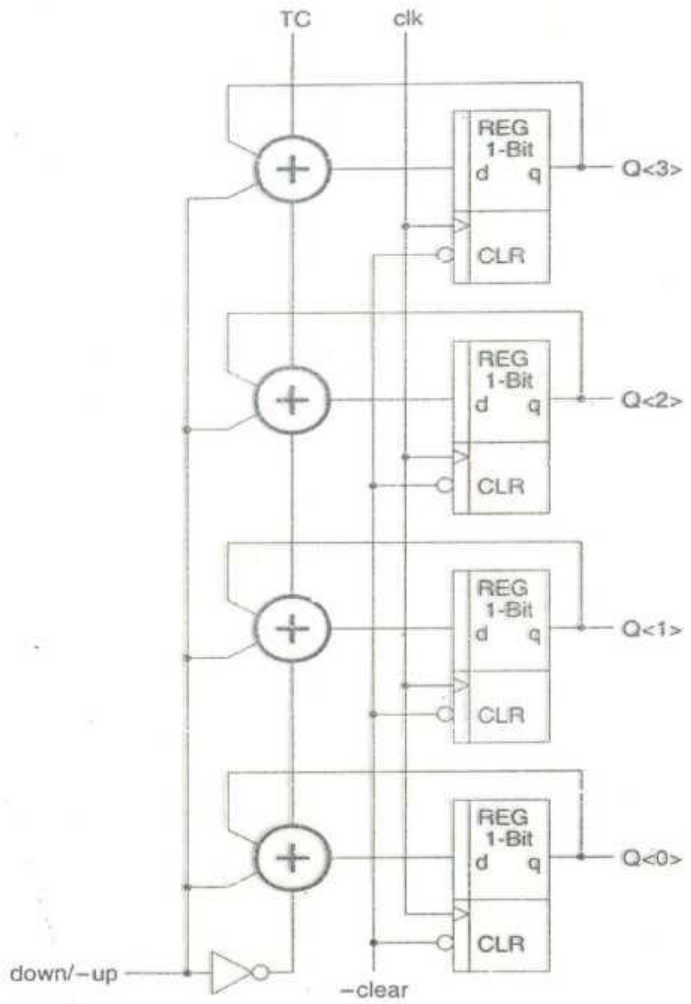


Figure 8.30 Synchronous up/down counter using adders and registers

ALUs:

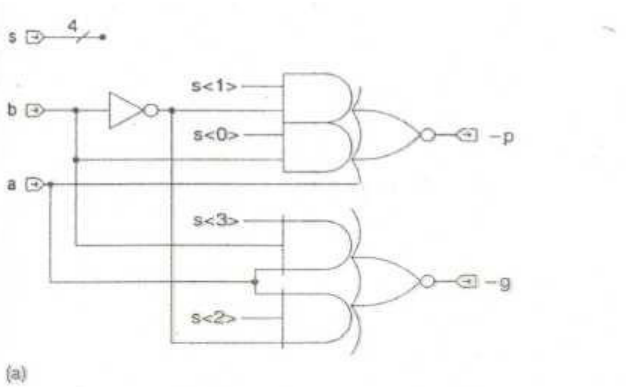
→ Arithmetic Logic Unit

→ Performs both arithmetic & logical opers.

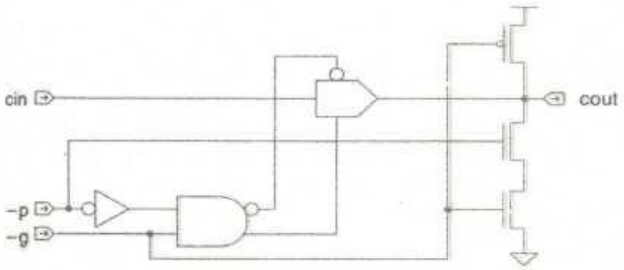
→ Basic ALU takes two data I/ps & a set of ctrl signals, called opcode.

→ The opcode along with ALU's carry-in, determine ALU's function.

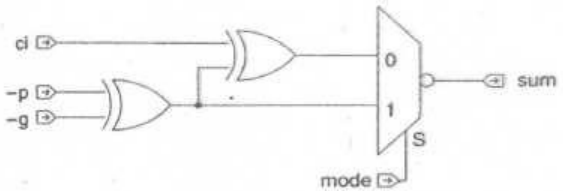
→ Logic to compute all possible ALU fns can be large unless carefully designed.



(a)



(b)



(c)

FIGURE 8.34 181 ALU

Signal mode = false (arithmetic)
= true (boolean).

ex:- S(0)=1, 'b' passed to o/p.

MULTIPLIERS

→ Used in Conv, Correlation, filtering, freq analysis.

→ Shift & add.

→ 2 steps: (i) Evaluation of partial products
(ii) Accumulation of shifted partial products.

→ partial products = ANDing of multiplicand & multiplier bit.

→ Classification of multipliers:
- Serial form
- Serial/parallel form
- Parallel form.

→ In Parallel multiplier, partial products computed in parallel.

ex:- $P = X \times Y = \sum_{i=0}^{m-1} X_i 2^i \cdot \sum_{j=0}^{n-1} Y_j 2^j$

$$P = \sum_{k=0}^{m+n-1} P_k 2^k$$

{P_k = Partial Prod = Summands}

For 4-bit Multiplier.

x_3 x_2 x_1 x_0 — Multiplicand

y_3 y_2 y_1 y_0 — Multiplier

x_3y_0 x_2y_0 x_1y_0 x_0y_0

x_3y_1 x_2y_1 x_1y_1 x_0y_1

x_3y_2 x_2y_2 x_1y_2 x_0y_2

x_3y_3 x_2y_3 x_1y_3 x_0y_3

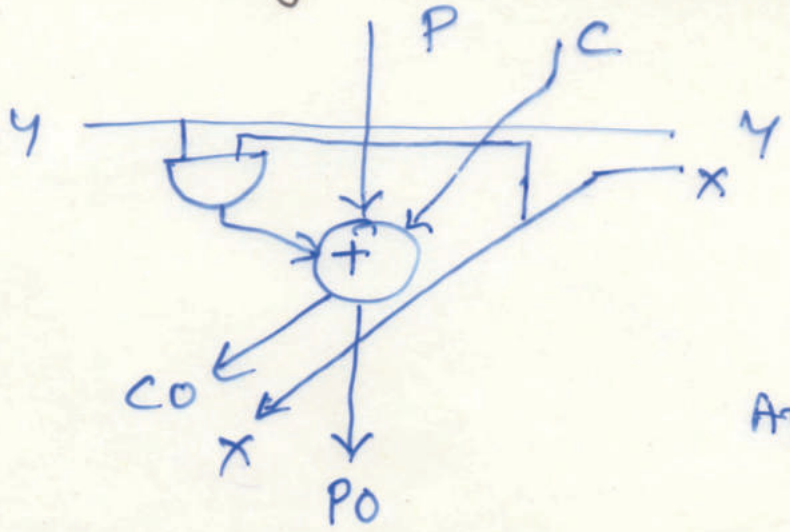
P_7 P_6 P_5 P_4 P_3 P_2 P_1 P_0 — Product

→ $n \times n$ Multiplier reqs:

$n(n-2)$ FAs, n HAs, n^2 AND gates.

→ Worst-Case delay = $(2n+1)\tau_g$

τ_g = worst-case adder delay.



Array Multiplier Cell.

x_i : propagated diagonally from top right to bottom left.

Y_j : horizontally.

→ The bit-wise AND is performed in the cell, & SUM is passed to next cell below.

→ "Carry-out" is passed to the bottom left of the cell.

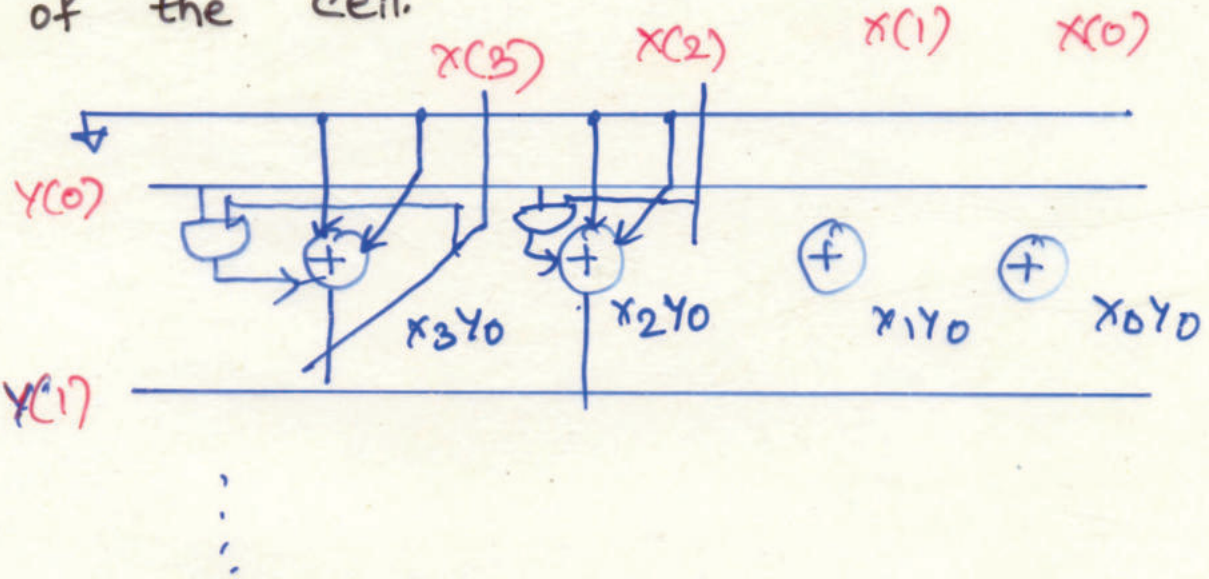


fig:- 4x4 array multiplier

Radix-n Multiplier :-

→ Above is radix-2 multiplier, coz computation done by one bit of multiplicand at a time.

→ High-radix multipliers reduce no. of adders & delay reqd to compute partial sums.

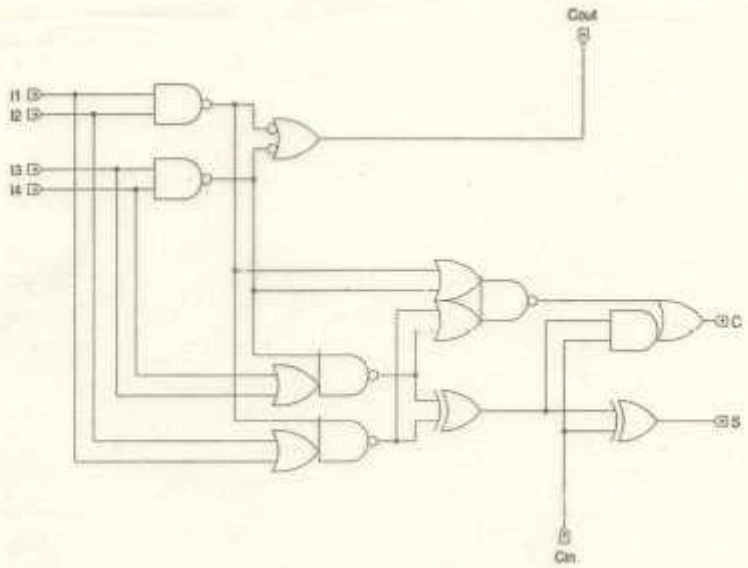


Figure 8.41 A 4:2 (5:3) compressor circuit

The two numbers X and Y are presented serially to the circuit (at different rates to account for multiplier and multiplicand word-lengths). The partial product is evaluated for every bit of the multiplier, and a serial addition is performed with the partial additions already stored in the register. The AND gate ($G2$) between the input to the adder and the output of the register is used to reset the partial sum at the beginning of the multiplication cycle. If the register is made of $N-1$ stages, then the 1-bit shift required for each partial product is obtained automatically. As far as the speed of operation is concerned, the complete product of $M+N$ bits can be obtained in MN intervals of the multiplicand clock.

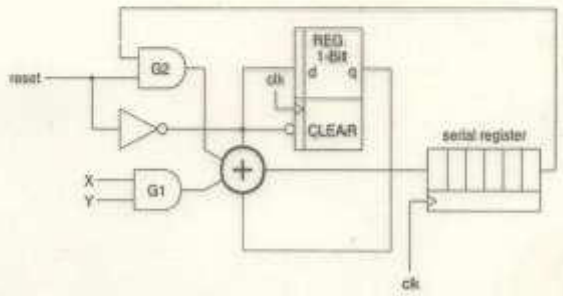


FIGURE 8.42 Serial multiplier

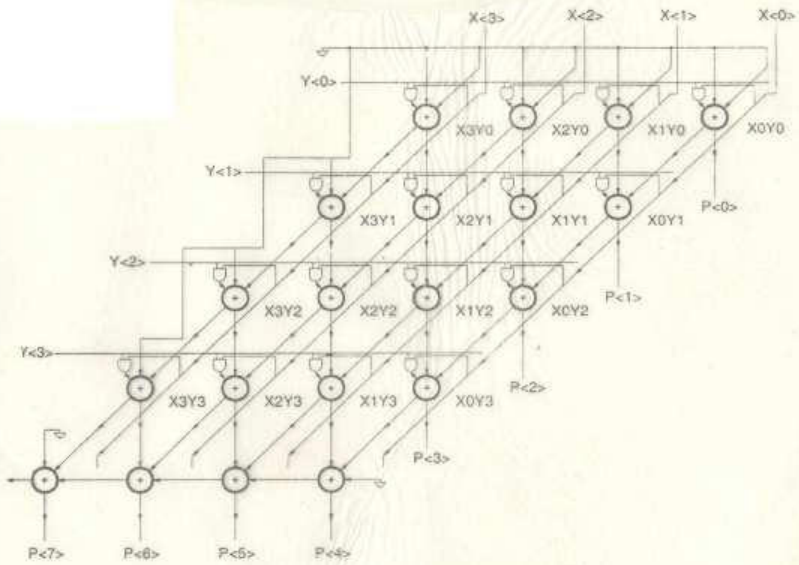


FIGURE A 4 × 4 array multiplier

array cell). In this case the appropriate inputs to the first and second row would be connected to ground, as shown in Fig. 8.36.

The cell design for this multiplier is relatively straightforward, with the main attention paid to the adder. An adder with equal carry and sum propagation times is advantageous, because the worst-case multiply time depends on both paths.

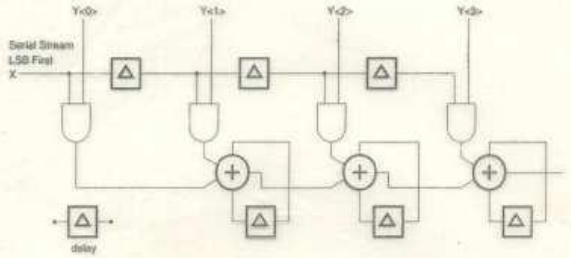


FIGURE Serial/parallel multiplier

Using the general approach discussed previously, it is possible to realize a serial/parallel multiplier with a very modular structure that can easily be modified to obtain a pipelined system. The basic implementation is illustrated by Fig. 8.43. In this structure, the multiplication is performed by means of successive additions of columns of the shifted partial products matrix. As left-shifting by one bit in serial systems is obtained by a 1-bit delay element, the multiplier is successively shifted and gates the appropriate bit of the multiplicand. The delayed, gated instances of the multiplicand must all be in the same column of the shifted partial-product matrix. They are then added to form the required product bit for the particular column.

This structure requires $M + N$ clock cycles to produce a product. The main limitation is that the maximum frequency is limited by the propagation through the array of adders. The structure of Fig. 8.43 can be pipelined with the introduction of two delay elements in each cell, as shown in Fig. 8.44. If rounding or truncation of the product term to the same word length as the input is tolerated, then the time necessary to produce a product is 2M clock cycles. In this case the multiplier accumulates partial product sums, starting with the least significant partial product. After each addition, the result is an

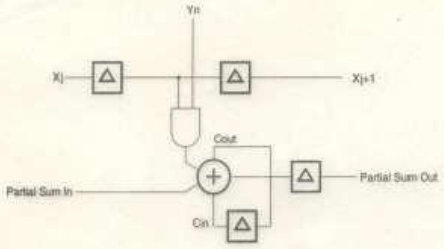


FIGURE Pipelined serial/parallel multiplier

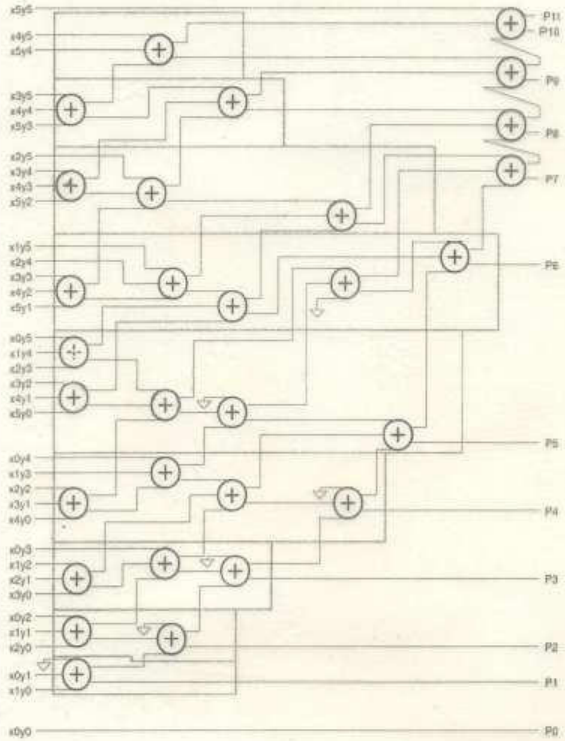


FIGURE Wallace adder tree (for 6 × 6 multiplier)

layout for a 54-by-54 bit multiplier using the compressor shown in Fig. 8.41 may be found in Goto et al.[16]

8.2.7.4 *Serial Multiplication*

Multiplication may be performed serially. The simplest form of serial multiplier, shown in Fig. 8.42, uses the successive addition algorithm and is implemented using a full adder, a logical AND circuit, a delay element (i.e., either static or dynamic flip-flop), and a serial-to-parallel register.

Booth Multiplier examines 3 bits of multiplicand at a time to determine whether to add zero, 1*, -1*, 2*, -2* of that rank of multiplicand.

Booth-recoding table:

x_{i-1}	x_i	x_{i+1}	Operation	NEG	ZERO	TWO
0	0	0	add 0	1	1	0
0	0	1	add 2	0	0	1
0	1	0	sub 1	1	0	0
0	1	1	add 1	0	0	0
1	0	0	sub 1	1	0	0
1	0	1	add 1	0	0	0
1	1	0	sub 2	1	0	1
1	1	1	add 0	0	1	0

NOTE: ZERO : zeroes the operand.
NEG : inverts operand
TWO : multiplies value by 2 (left shift)

Booth Multiplier examines 3 bits of multiplicand at a time to determine whether to add zero, 1*, -1*, 2*, -2* of that rank of multiplicand.

Booth-recoding table:

x_{i-1}	x_i	x_{i+1}	Operation	NEG	ZERO	TWO
0	0	0	add 0	1	1	0
0	0	1	add 2	0	0	1
0	1	0	sub 1	1	0	0
0	1	1	add 1	0	0	0
1	0	0	sub 1	1	0	0
1	0	1	add 1	0	0	0
1	1	0	sub 2	1	0	1
1	1	1	add 0	0	1	0

NOTE: ZERO : zeroes the operand.
NEG : inverts operand
TWO : multiplies value by 2 (left shift)

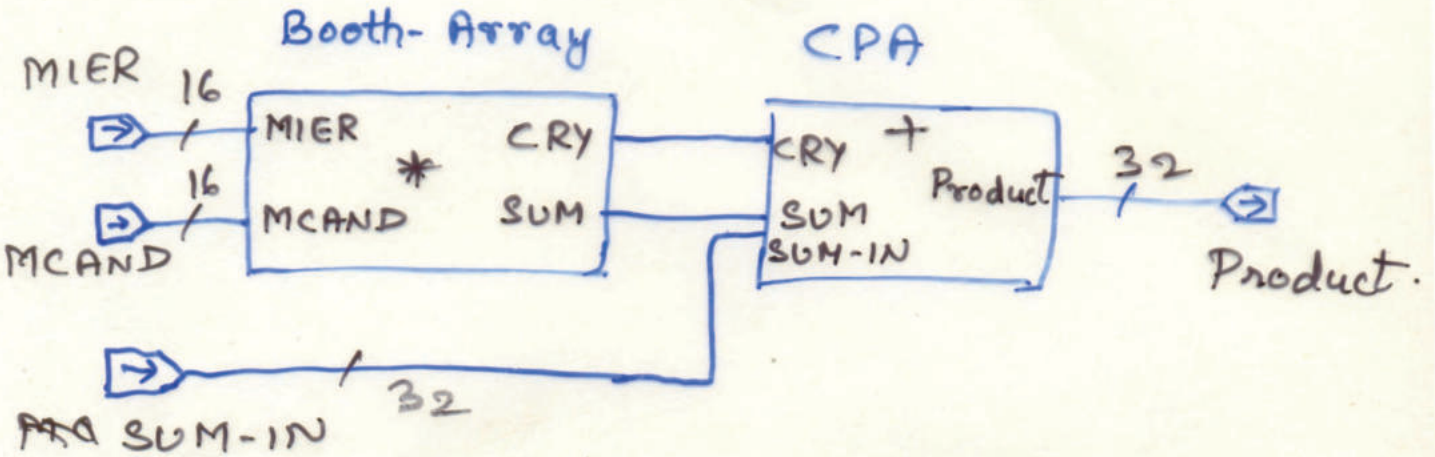


fig: 16x16 booth recoded multiplier (Schematic)

→ Booth-array accepts two 16-bit I/ps, & feeds CPA. (Carry Propagate Array).

→ The CPA also accepts 32-bit I/p to perform multiple accumulates.

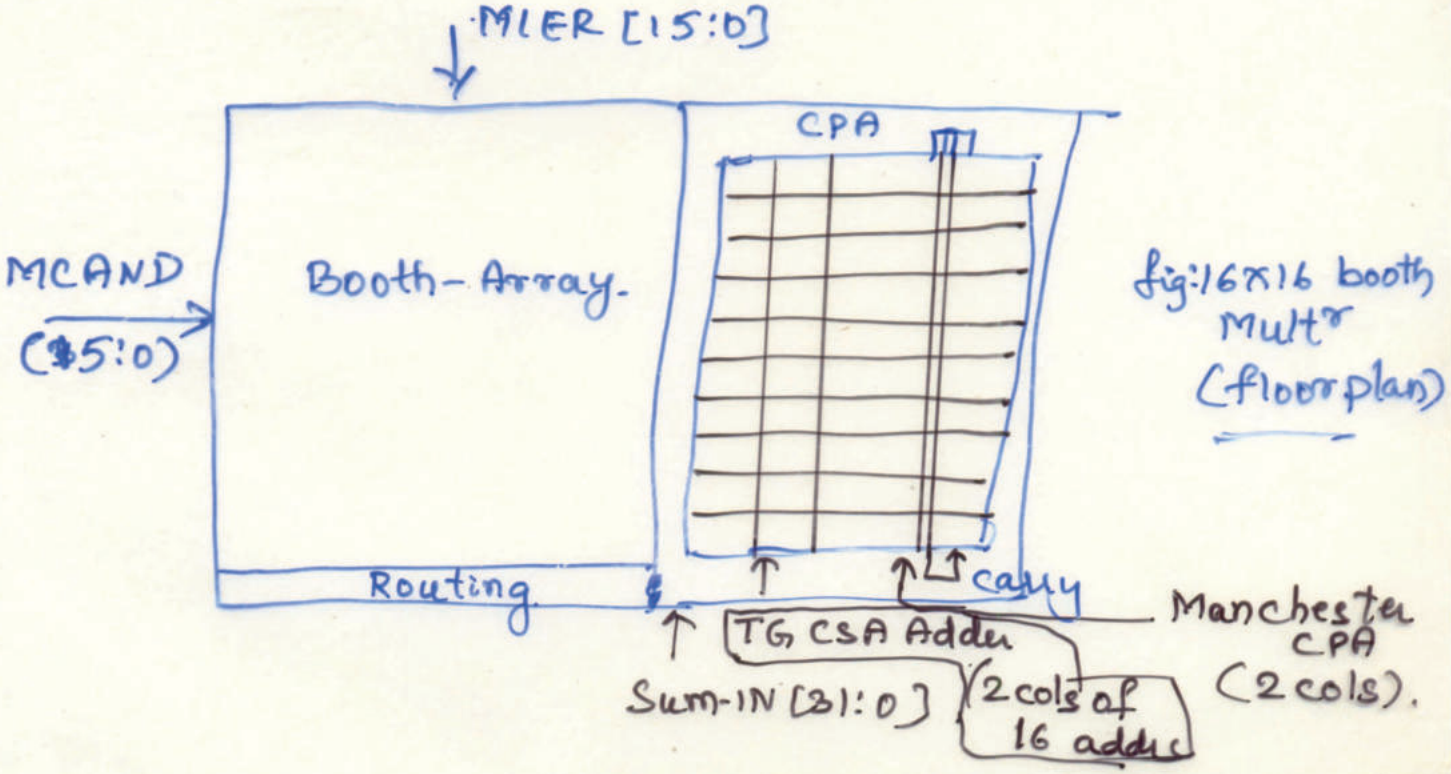


fig: 16x16 booth Multr (floorplan)

ARRAY SUBSYSTEMS

Memory elements may be divided into :

(i) Random access memory
(ii) Serial access memory
(iii) Content access memory.

RAM at chip level is classed as memory that has access time independent of the physical location of the data. Where as serial-access memories have some latency associated with reading or writing of particular data and with content addressable memories.

ROMs have write time greater than read times. But RAMs have very similar read and write times These are divided into

* Static-load : reqs no clock.
* Synchronous : require clk' edge to enable memory operation.
* Asynchronous : asyn. RAMs recognize address changes & o/p new data after any such change.

→ Static-load & synchronous memories are easier to design & best choice for system-level building block.

→ Memory cells in RAMs further divided into :
(a) Static structures (b) Dynamic structures.

Static cells use some form of latched storage, while dynamic cells use dynamic storage of charge on a capacitor.

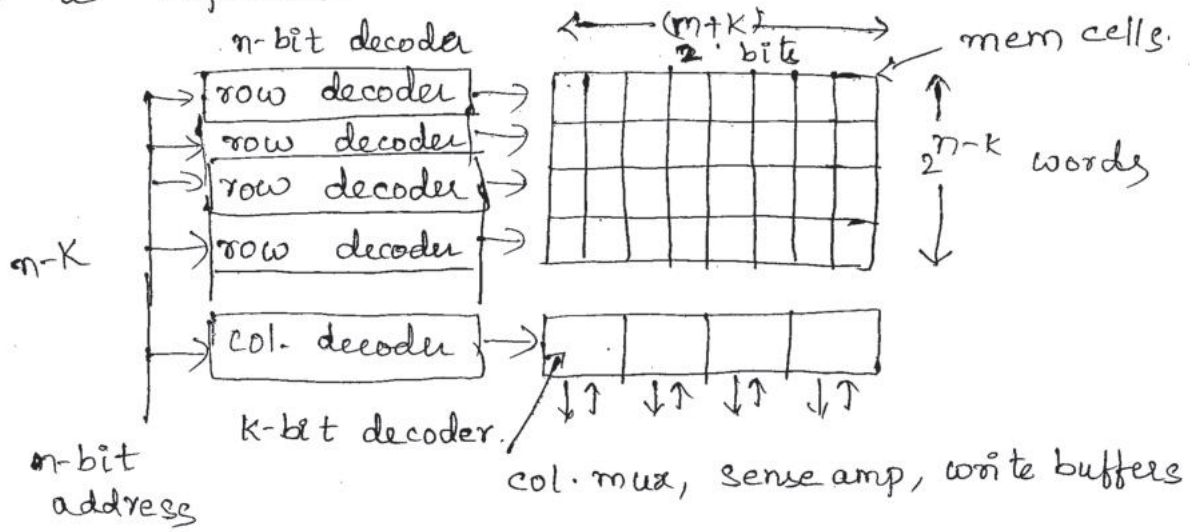


fig: Memory-chip architecture.

RAM :-

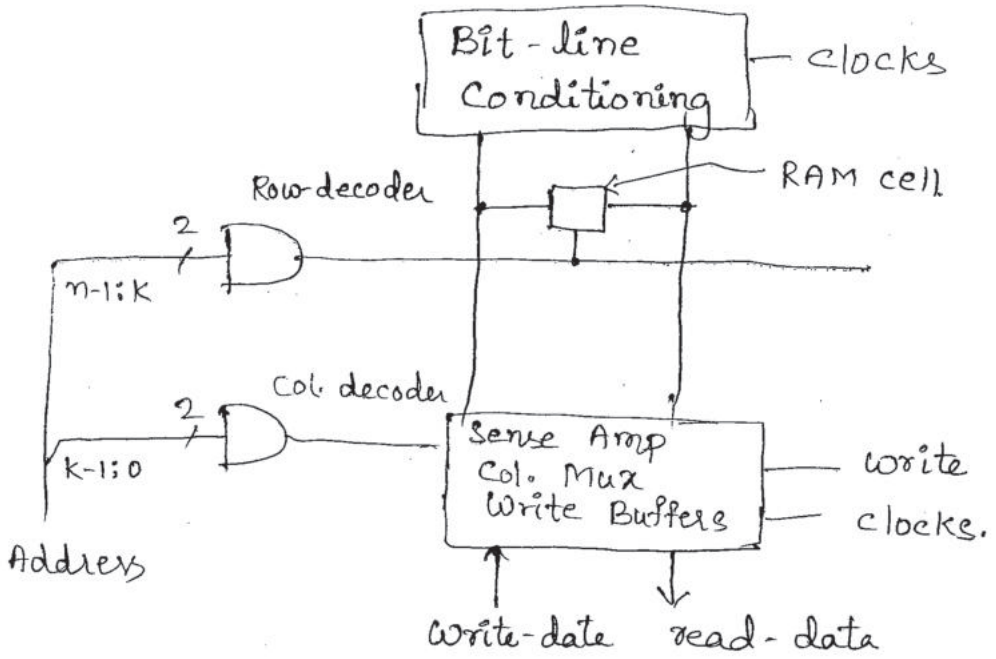
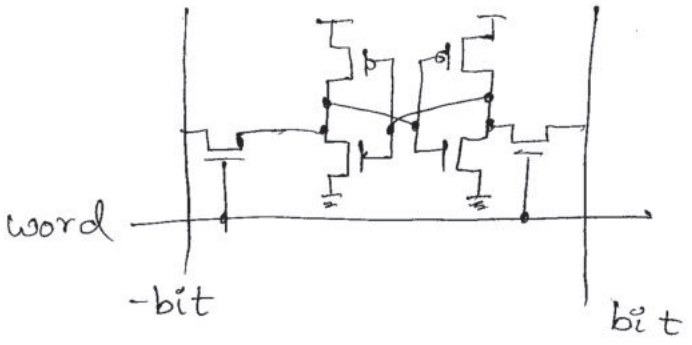


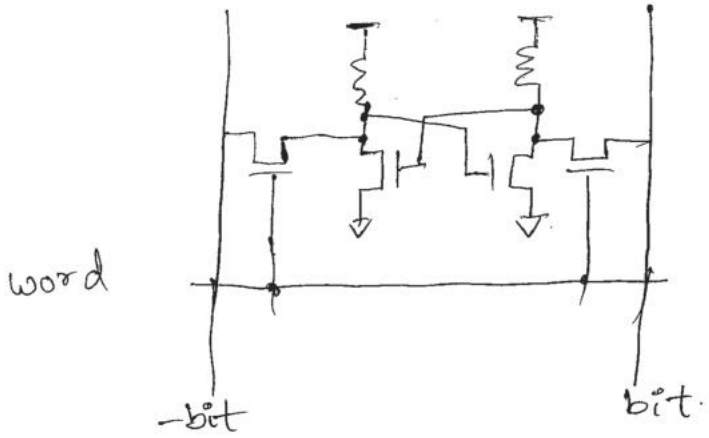
fig: Generic RAM circuit.

cross- coupled inverter ckt.

The p- Trs may be replaced with high value polysilicon resistors. The value of resistor has to be ∋ it prevents leakage from changing any value stored in RAM cell. Generally 100's to 1000's of MΩ.



(a) SRAM



(b) SRAM

HIGH DENSITY MEMORY ELEMENTS

→ On-chip memory is important as levels of integration increase to allow both processors and useful amts of mem to be integrated on a single chip.

→ ROM

→ Flash memory : dominant form of electrically erasable PROM memory.

→ RAM — SRAM
DRAM

→ SRAM is faster, Larger & uses more pow.

→ DRAM has smaller layout & uses less pow. DRAM cells require periodically refreshing of dynamically stored values.

→ A design that reqs high-density ROM or RAM is partitioned into several chips.

→ Medium density memory, on the order of one 'K' bytes, often be put on same chip with logic that uses it, giving faster access times, as well as greater integration.

→ The bit lines are typically precharged, so the cell discharges one of the lines.

→ The bit lines are read by ckts that sense the value on the line, amplify to speed it up & restore the signals to the proper vtg levels.

→ A write is performed by setting the bit lines to the desired values & driving that value from the bit lines into the cell.

→ Row decoders typically use NOR gates to decode the addrs, followed by chain of buffers to allow ckt to drive large cap'ce of the word line.

→ ckts to implement NOR :
  → Pseudo-nMOS
  → Precharged.

→ Precharged ckts provide better performance for large memory arrays compared to Pseudo-nMOS.

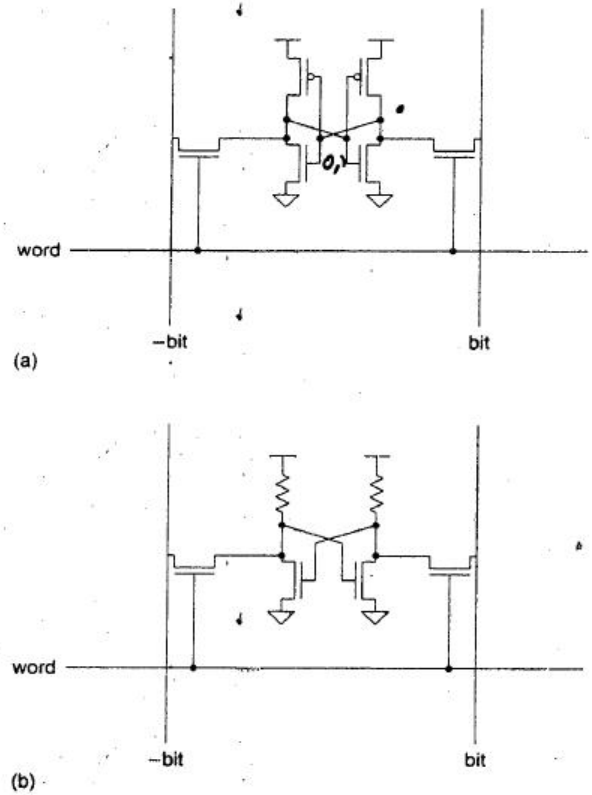


Figure Static RAM cell circuits

merged read and write data busses may be used. A 1-transistor cell is shown in Fig. 8.52(c).[20] The memory value is again stored on a capacitor. The capacitor can be implemented as a transistor as shown in Figs. 8.52(d) and 8.52(e). Sense amplifiers sense the small change in voltage that results when

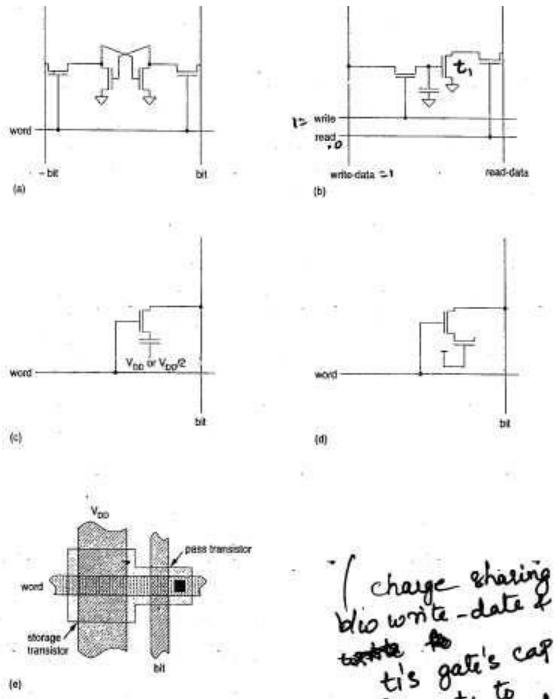


Figure Dynamic RAM circuits: (a) 4-transistor; (b) 3 transistor; (c) 1 transistor with capacitor; (d) 1 transistor with transistor capacitor; (e) representative layout for (d)

charge sharing b/w write-data & t1's gate's cap forces t_1 to desired value

the RAM circuit concentrates on pulling the bit line from high to low. Thus one method of reading a RAM cell would be to precharge the bit lines high and then enable the word-line decoder. For a given pair of bit lines, one RAM cell will attempt to pull down either the *bit* or *–bit* line depending on the stored data. The bit-line pull-up circuit may use p-channel transistors to precharge each bit line (Fig. 8.53a). In this example, the sense amplifier is an inverter that forms a single-ended sense amplifier. The sense time is roughly

* To read, read-data precharged to V_{DD} & set read=1. If t_1's gate has stored charge, t_1 p.d read-data ... read-data has complement of value stored

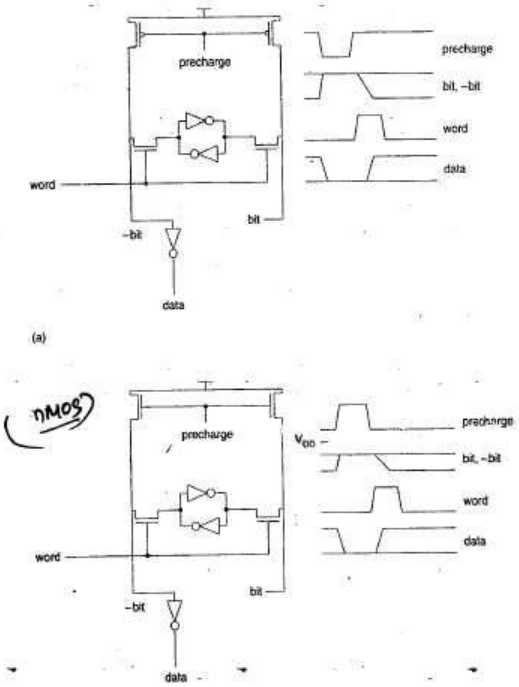


FIGURE RAM read options: (a) V_{DD} precharge; (b) $V_{DD}-V_{tn}$ precharge

the time it takes one RAM cell pull-down and access transistor to reach the inverter threshold. To optimize speed, one might set the inverter threshold above the V_{DD} midpoint, but below an adequate noise margin down from the V_{DD} rail. Alternatively, one can precharge the bit lines with n-channel transistors, which results in the bit lines being precharged to an n threshold down from V_{DD} (Fig. 8.53b). This can dramatically improve the speed of the RAM cell access. In addition, it reduces power dissipation because the bit lines do not change by the supply voltage. The key aspect of the precharged RAM read cycle is the timing relationship between the RAM addresses, the precharge pulse, and the enabling of the row decoder. If the word-line assertion precedes the end of the precharge cycle, the RAM cells on the active word-line will see both bit lines pulled high and the RAM cells may flip state. If

to amplify this bit-line change. Design margins must be valid over all process, temperature, and voltage extremes. Figure 8.55 shows the zero bit voltage ($V_{bit(0)}$) and the pull-down voltage ($V_{pulldown}$) for various ratios of pull-up beta to pull-down betas. As the pull-up becomes weaker, the $V_{bit(0)}$ voltage approaches V_{SS} and the differential voltage between a high and a low on the bit lines increases. However, as the pull-down transistors are limited in size by the desire to keep the RAM cell small, a design trade-off has to be made between speed and the differential bit voltage, which affects the noise

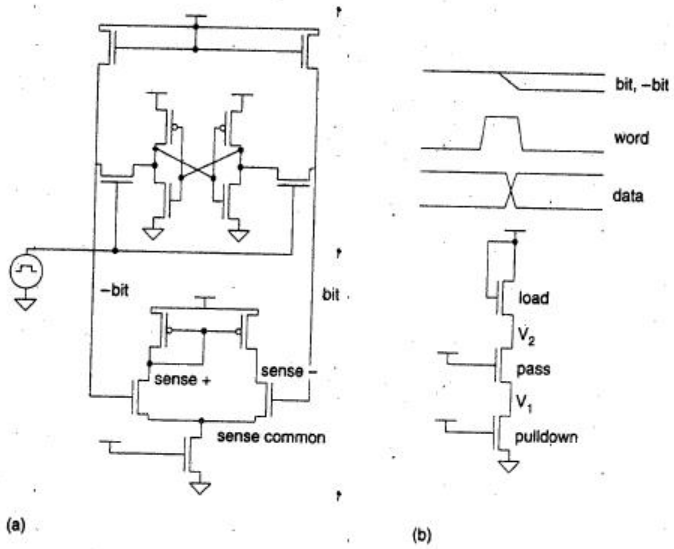


Figure RAM read operation model

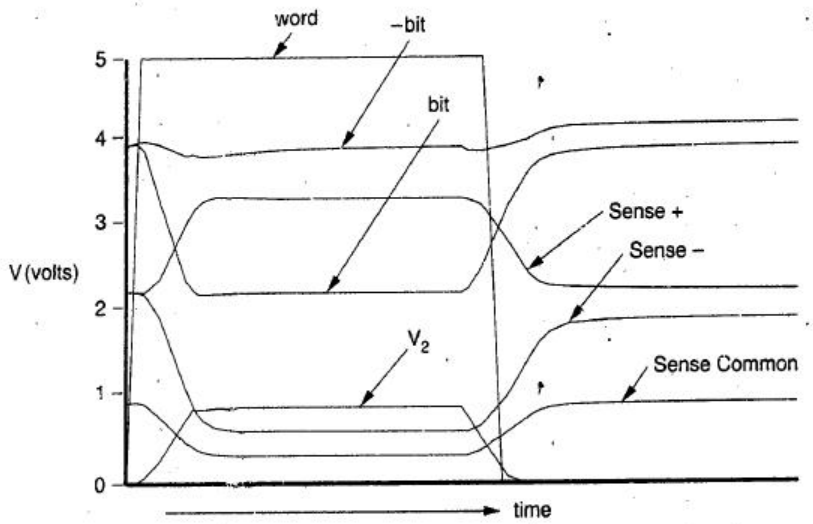


Figure Static RAM–read waveforms

Current mode sensing may also be used.[29,30,31] In this technique, the current change in the bit lines is detected using special circuits. The theory is that by using low-impedance circuits, the RC delay inherent in driving the bit lines may be decreased.

8.3.1.1.2 Static RAM—write

The objective of the RAM write operation is to apply voltages to the RAM

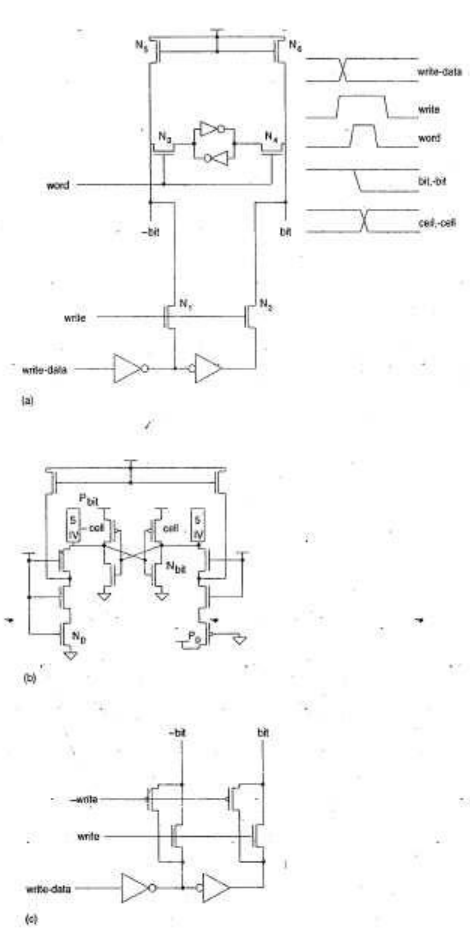
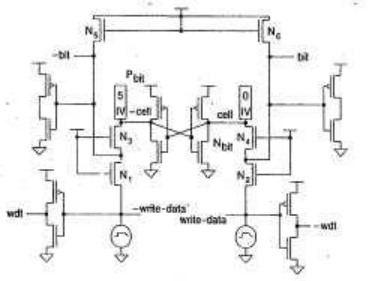
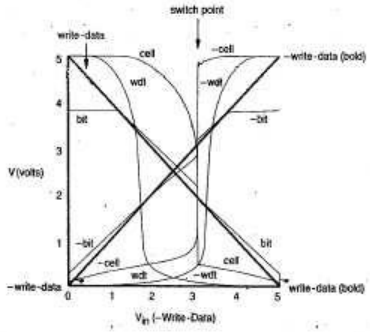


FIGURE Static RAM–write circuits: (a) n-channel pass transistors; (b) circuit model during write; (c) complementary transmission gate version

write enable TrS N_1, N_2 are enabled to allow data & comp to move to bit lines.



(a)



(b)

Figure Static RAM–write waveforms and circuit model

8.3.1.1.3 Row decoders

The simplest row decoder is an AND gate. Figure 8.59 shows two straightforward implementations. The first in Fig. 8.59(a) is a static complementary NAND gate followed by an inverter. This structure is useful for up to 5–6 inputs or more if speed is not critical. The NAND transistors are usually made minimum size to reduce the load on the buffered address lines because there are 2^{n-k} ($N_{load} + P_{load}$)'s on each address line. The second implementation, shown in Fig. 8.59(b), uses a pseudo-nMOS NOR gate buffered with two inverters. The NOR gate transistors can be made minimum size, and the inverters can be scaled appropriately to drive the word line. Large fan-in AND gates can also be constructed from smaller NAND and NOR gates, as shown in Fig. 8.59(c). Figure 8.60 shows two possible layout styles (in sym-

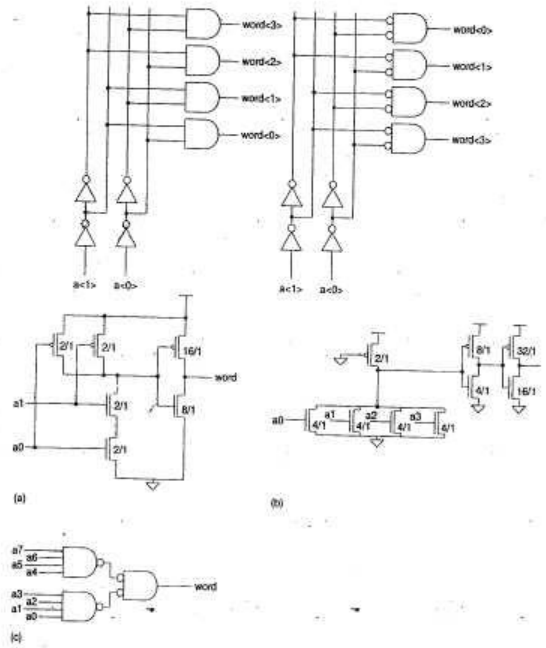


FIGURE Row-decoder circuits: (a) complementary AND gate; (b) pseudo-nMOS gate; (c) cascaded NAND, NOR gates

bolic form) for the row decoders. One passes the address lines over the decode gates, while the other uses a more standard cell style. Choice would depend on the size of the decoder in relation to the size of the RAM cell. Often, speed requirements or size restrict the use of single-level decoding, such as that shown in Fig. 8.59. The alternative is a predecoding scheme, which is illustrated in Fig. 8.61(a). Here the $(n\text{-}k)$ row address lines are split into a p-bit predecode field and a q-bit direct decode field. The q-bit decode field requires a gate per word line, so q is chosen to suit the pitch of the RAM cell. The p-bit predecode field generates 2^p predecode lines (4 in this example), each of which is fed vertically to 2^{n-k}-row decode gates (8 in this example). Figure 8.61(b) shows a possible implementation of a predecode scheme, where the predecode gate is a NAND gate and the word-decode gate is a NOR gate. An additional input ($-clk$) has been included in the NOR gate

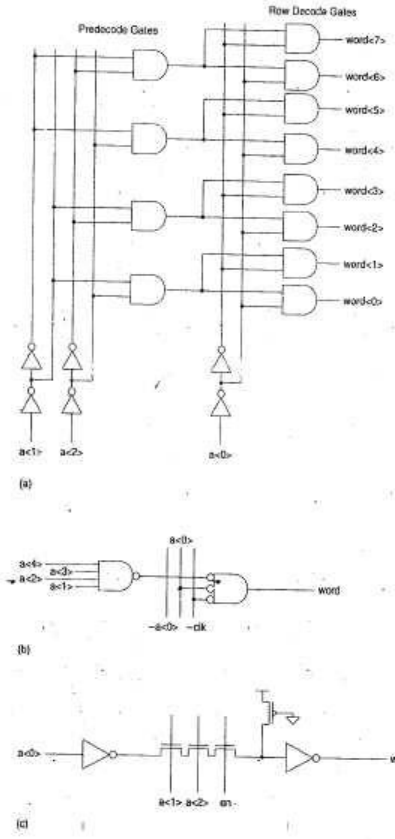


FIGURE Predecode circuits: (a) basic approach; (b) actual implementation; (c) pseudo-nMOS example

via pass gates enabled by the column-address lines. The address decoding is in essence distributed. Decoders for *bit* and *~bit* lines are shown, although one of these may be omitted for single-ended read operations. The read (and, usually of lesser importance, write) operations are somewhat delayed by the series-transmission gates. However, in comparison with gate delays these

COL DECODER.

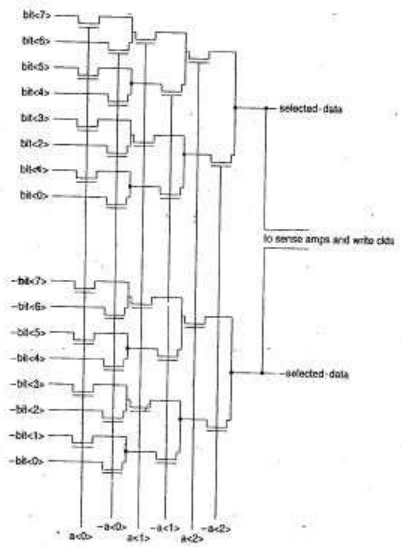


FIGURE Tree-style column decoder

If the delay of the series-pass gates was troublesome, the decoder shown in Fig. 8.64 could be used. Here a NAND decoder is employed on a bit-by-bit basis to enable complementary transmission gates (single transistors may be used where possible) onto a common pair of data lines. These are then routed to a sense amplifier and write circuitry.

8.3.1.1.5 Sense amplifiers

Many sense amplifiers have been invented to provide faster sensing, smaller layouts, and lower power-dissipation sensing.[33] The simple inverter sense amplifier provides for low power sensing at the expense of speed. The differential sense amplifier can consume a significant amount of DC power (Fig. 8.54). Alternatively, one can employ clocked sense amplifiers similar to the SSDL gate shown in Fig. 5.40.

8.3.1.1.6 RAM timing budget

The critical path in a static RAM read cycle includes the clock to address delay time, the row address driver time, row decode time, bit-line sense time, and the setup time to any data register. The column decode is usually not in

Selecting 2^k out of 2^m bits of accessed row.

The CAM portion examines a data word and compares this data with internally "stored" data. If any data word internally matches the I/p data word, the CAM signals that there is a match. These match signals can be passed as word lines to RAM to enable a specific data word to be o/p. This structure may be used as translation look-aside buffer in the virtual memory look up in a microprocessor.

fig : Basic CAM

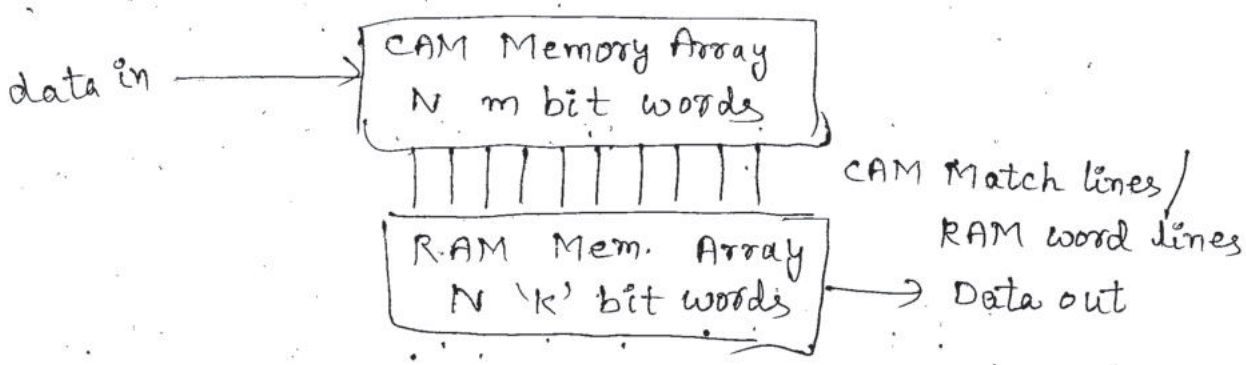


fig: applicⁿ as translation lookaside baffer.

A CMOS CAM cell consists of normal static RAM cell with transistors N_1 and N_2, which form XOR gate, & N_3 which is distributed NOR pull-down. Writes are used to store the match data in the cells, whereas reads are used for testing

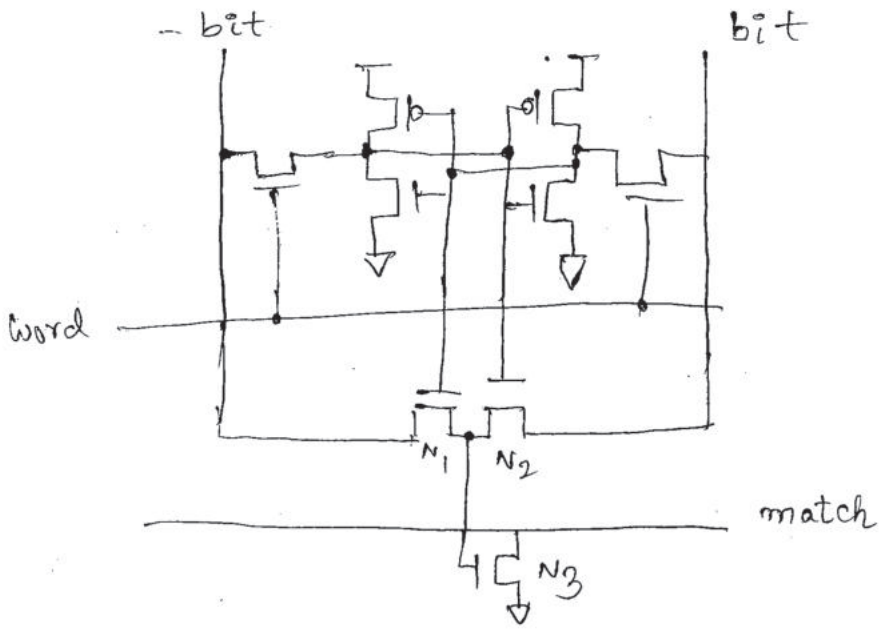


fig: CMOS CAM cell

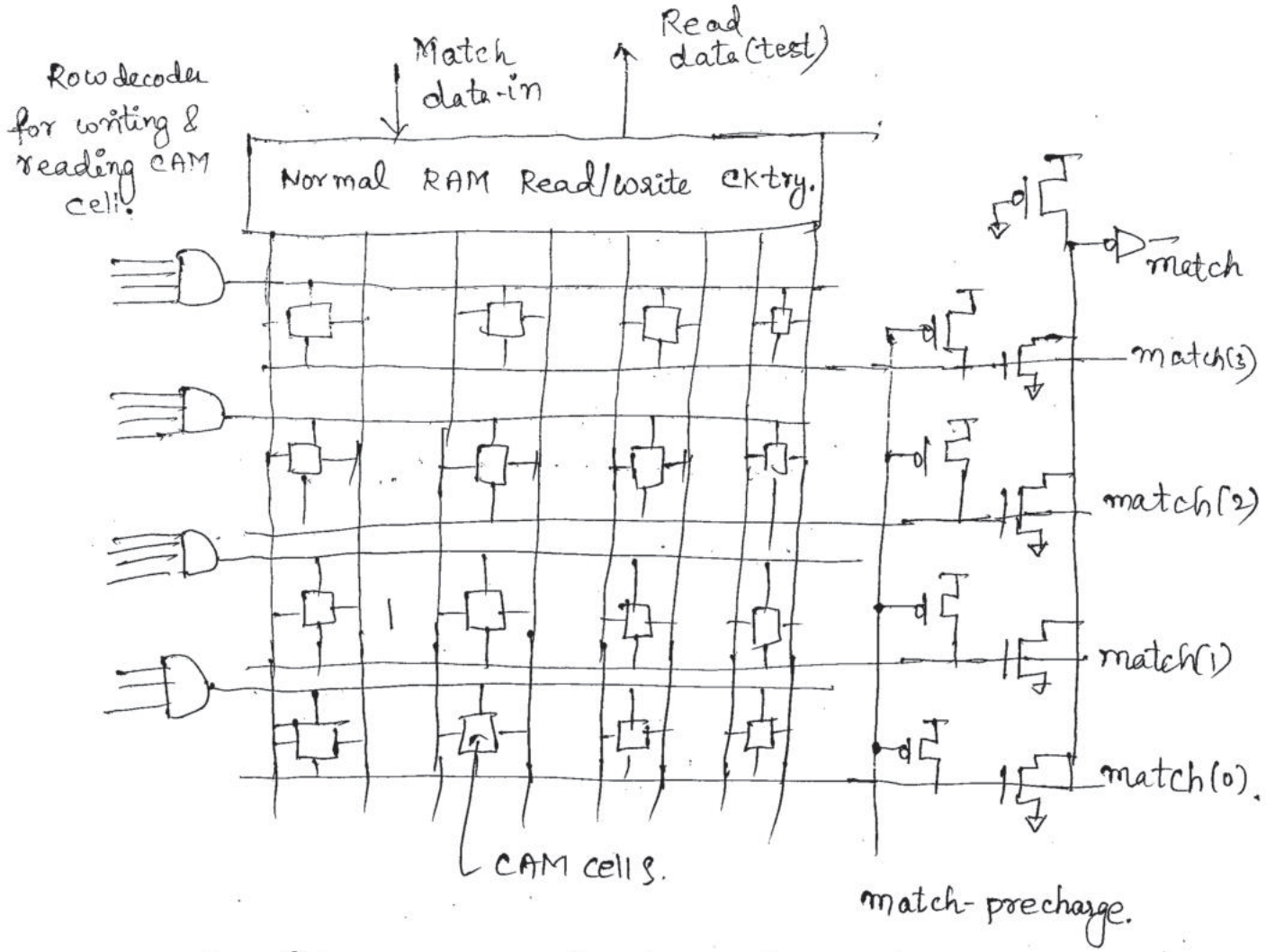


fig: array circuit.

A match operation [illegible] to be matched on the bit lines but not asserting word line. A '1' appears on gate of N3 if the data in the cell is not equal to data on bit lines. The drain of N3 Trs of cells in the same row are commoned. These form distributed NOR gate. Each match line (match<3:0>) remains high if the data in the row matches the data placed on the bit lines. These lines may be used to assert the word lines on a RAM.

SERIAL ACCESS MEMORY

Serial access memories (shift regs) are of use in signal-processing applications for storage and delaying signals. It can be simulated by a RAM & this provides smallest implementation bcoz CMOS static RAM is very area efficient. SAM is surrounded by counters. Some times a dedicated shift register memory may be appropriate from density, speed or floorplanning viewpt.

fig => Blocks of byte shift regs are delayed by 32, 16, 8, 4, 2 & 1 clk cycles & muxes control the pass-around of delay blocks to yield appropriate delay amount. Each mem cell is a

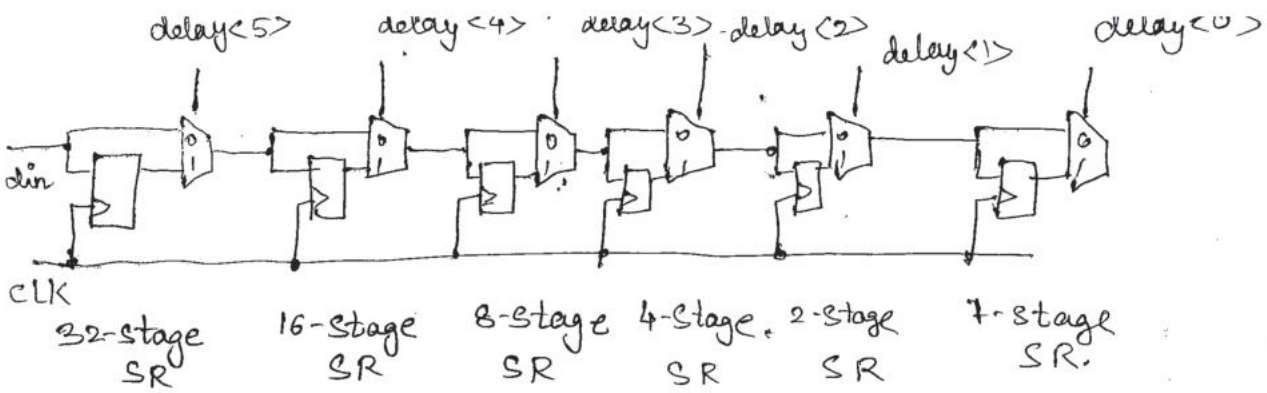


fig: Tapped delay line architecture.

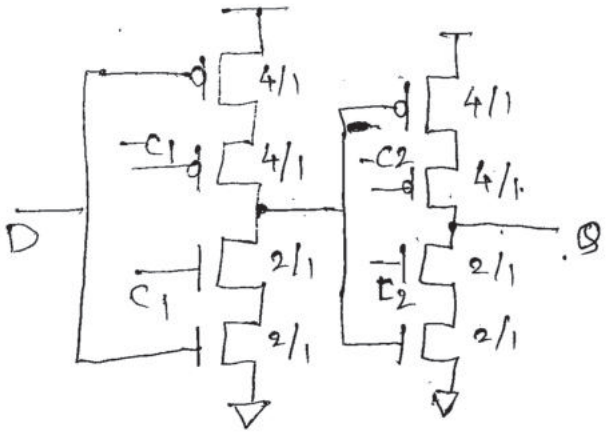


fig: ckt (mem cell).

ROM (READ ONLY MEMORY)

→ It can be implemented with only one Tr per bit of storage. It is a static memory structure in that the state is retained indefinitely without even power. It is generally implemented as a NOR array.

→ Can use NAND array for ultra small ROMs but will be slow.

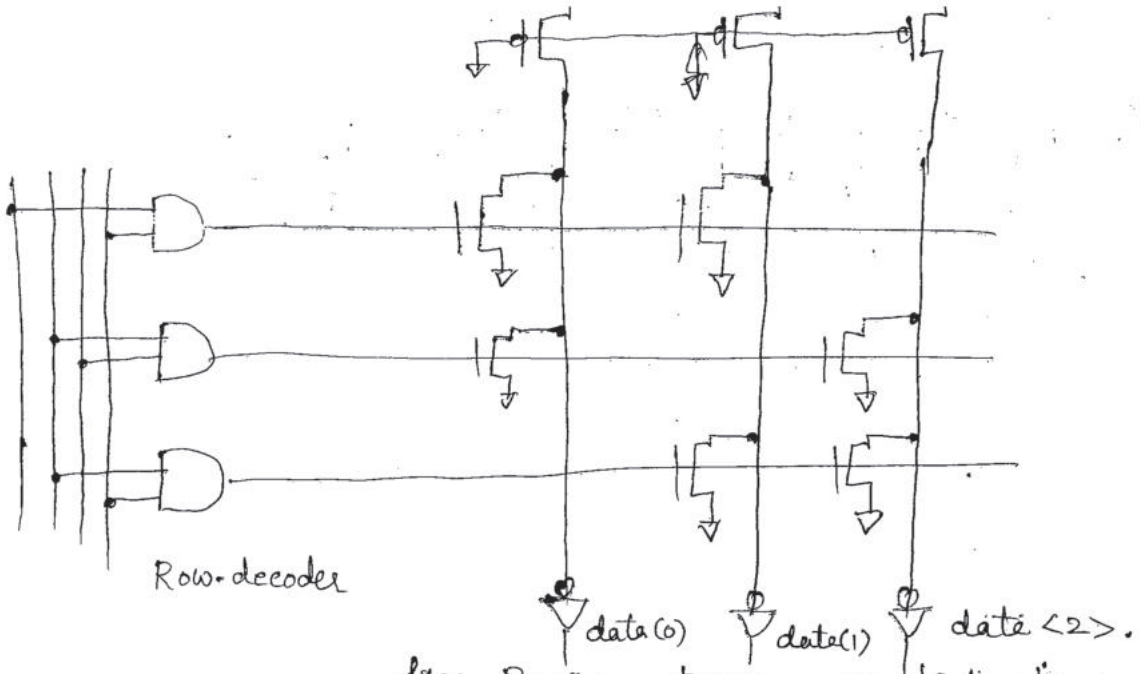


fig: Basic ROM architecture.

Domino logic ROM can slow down bit-line transition for large ROMs. ∴ use dynamic ROM.

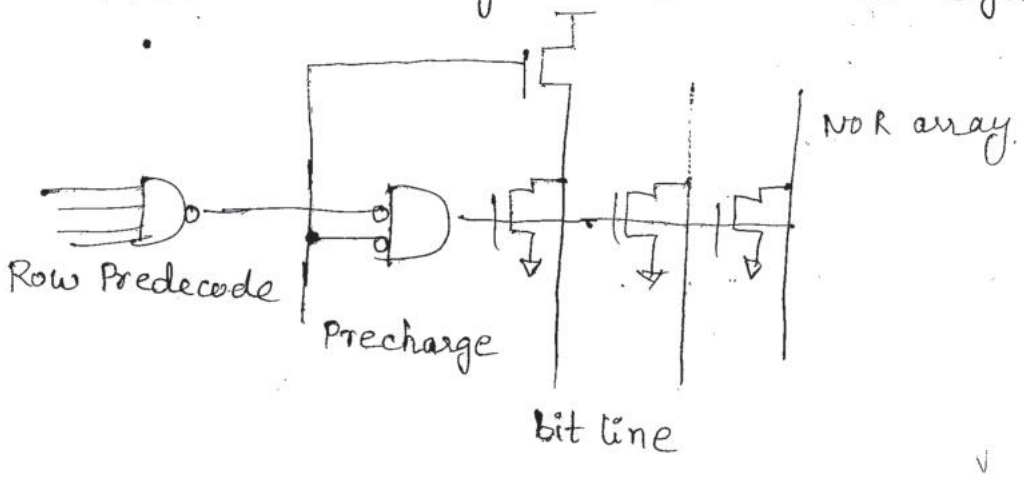


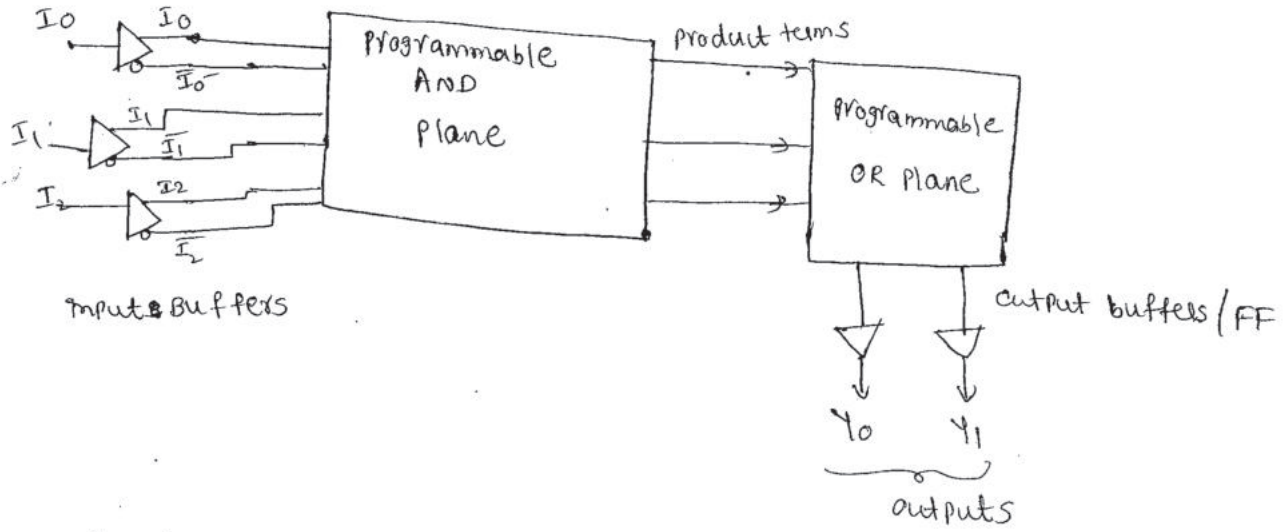
fig: dynamic ROM cktry.

Here word lines are forced low while bit lines are being precharged. This ensures that DC current does not flow. After bit-line pull-ups have been turned off, the word-line drivers are asserted and one word line is active. This reqs careful designs of timing chain of sequence of events.

PLAs :- (Programmable Logic Array)

The PLA is one type of Programmable logic device which has a set of Programmable AND planes followed by a set of Programmable OR planes and which can then be Conditionally Complemented to produce an output.

Block diagram of a PLA device:-

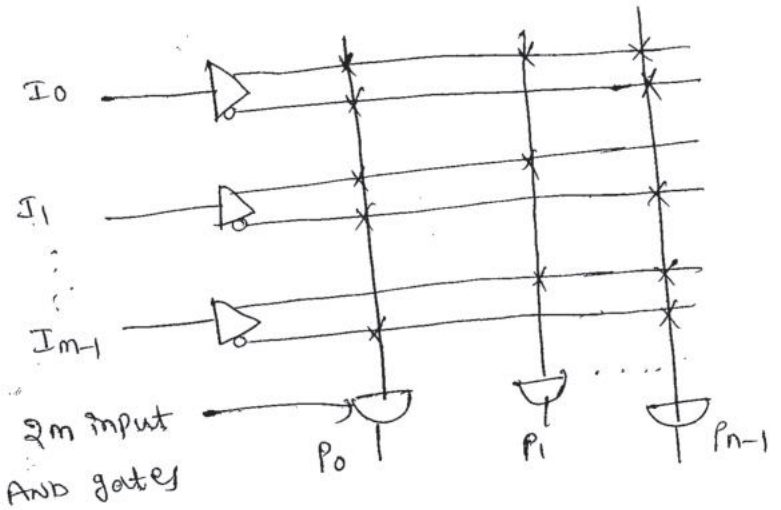


→ In the PLA approach instead of generating all the min terms a seperate logic is implemented which generates only the required product terms.

→ This saves lot of silicon area and also the Common product terms are identified and only one product term is

generated for that particular term.

Simplified representation of AND matrix:-

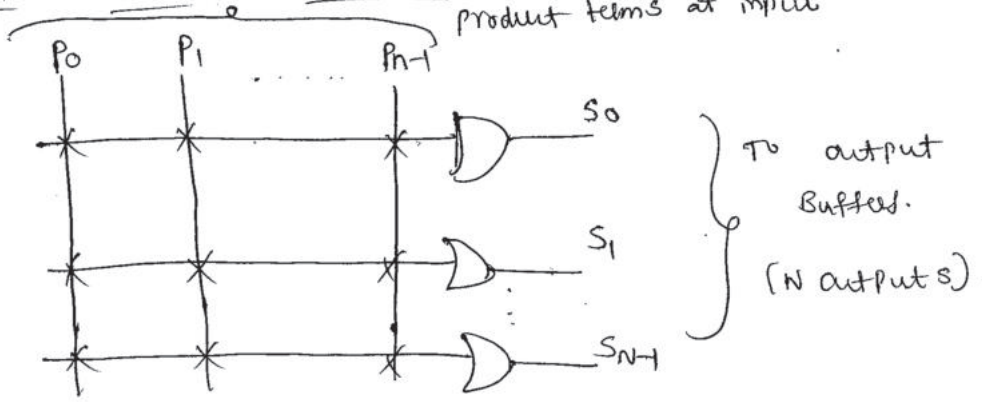


X → represent connection

where M → number of inputs.

n → number of product terms.

Simplified representation of OR matrix:-



Application of PLA:-

(1) we can implement both combinational as well as sequential circuits using PLA.

(2) For combinational circuits, the PLA devices with only

→ To implement the sequential circuits we use the PLA device with Flipflop and buffers included in the output stage.

Designing of Combinational circuits using PLA :-

Procedure :

step1 : prepare the truth table

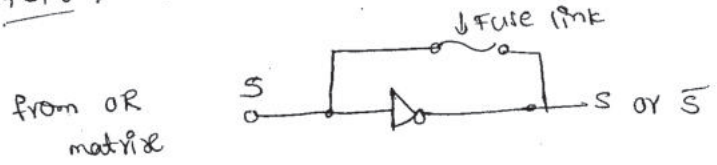
step2 : write boolean expression in sop form

step3 : obtain the minimum SOP form to reduce the number of product terms to minimum.

step4 : Decide the input connections of AND matrix for generating the required product terms.

step5 : Then decide the input connections of OR matrix to generate the required sum terms.

step6 : Decide the connections of invert / non-invert matrix



→ It can invert its input if active low output is required. Its input is passed without any inversion if active high output is required.

→ The output will be same as input if fuse link is closed, where as we get $\bar{S}$ at output if fuse link is open circuit.

step7 :-

Program the PLA .

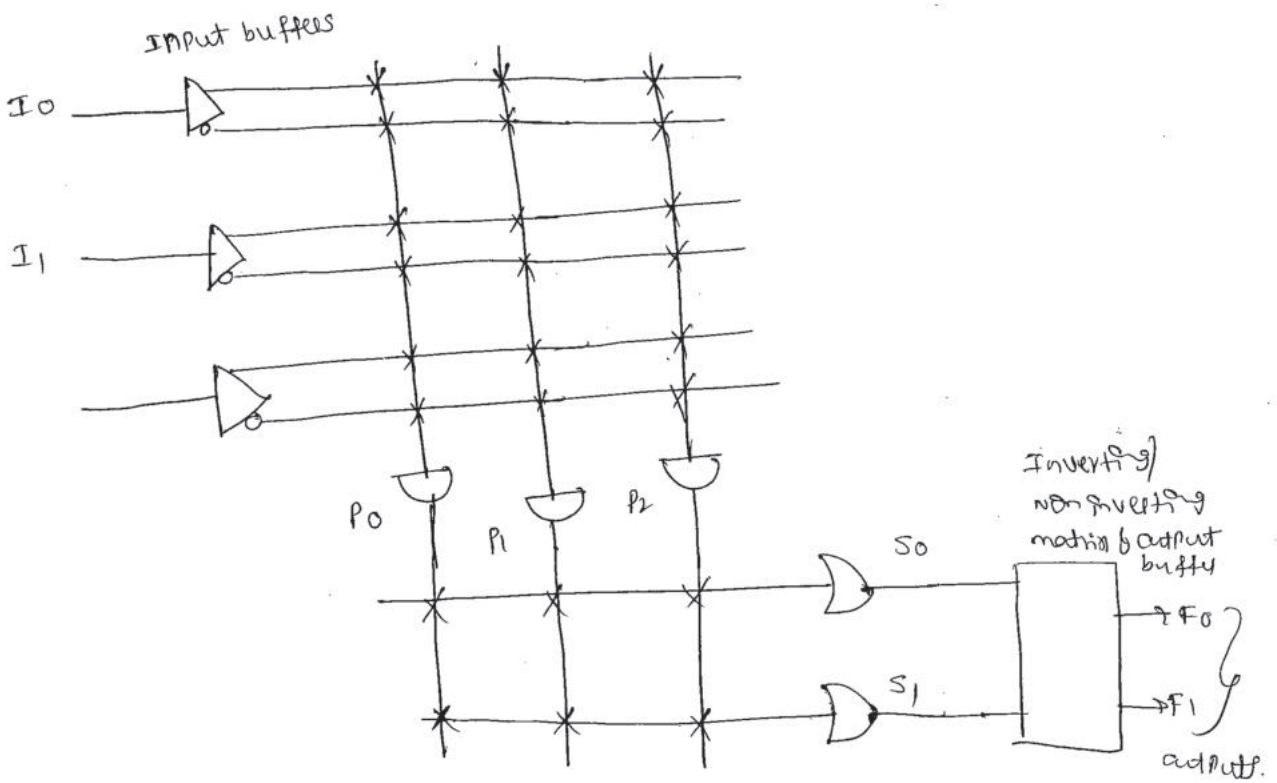
Combinational circuit for a PLA with 3 inputs, three product terms and 2 outputs;

let 3 inputs I_0, I_1, I_2.

number of outputs = number of OR gate = 2.

number of product terms = number of AND gate = 3.

→ Hence this PLA device will have 3 AND gate each of which has 6 inputs & 2 OR gate.



Specifying size of a PLA

The size of a PLA is specified by m x p x n.

m → no. of inputs

p → No. of product terms

n → no. of outputs.

(1) A combinational circuit is defined by the functions

$F_1(A,B,C) = \Sigma m(4,5,7)$

$F_2(A,B,C) = \Sigma m(3,5,7)$.

Implement this circuit with a PLA having 3 inputs, 3 product terms and 2 outputs.

$F_1(A,B,C) = \Sigma m(4,5,7) = A\bar{B}\bar{C} + A\bar{B}C + ABC$

$F_2(A,B,C) = \Sigma m(3,5,7) = \bar{A}BC + A\bar{B}C + ABC$.

Minimization:

$F_1(A,B,C) = A\bar{B}\bar{C} + A\bar{B}C + ABC$

$= A\bar{B}(\bar{C}+C) + ABC$

$= A\bar{B} + ABC$ ($\because C+\bar{C} = 1$ in Boolean algebra).

$= A(\bar{B}+BC)$

~~$F_2 = \bar{A}B\bar{C} + A\bar{B}C + ABC$~~

$= A[\bar{B}(C+1) + BC]$

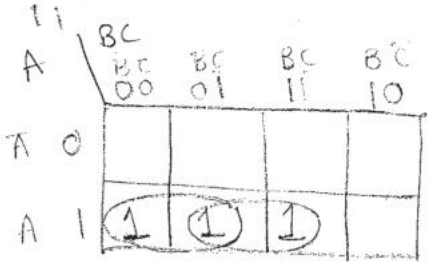
$= A[\bar{B}C + \bar{B} + BC]$

$= A[\bar{B} + C(B+\bar{B})]$

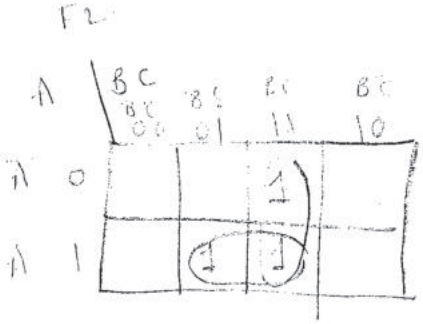
$= A[\bar{B}+C]$

$F_1(A,B,C) = A\bar{B} + AC$ —— (1)

OR minimization using K-maps:



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



$F_2 = BC + AC$.

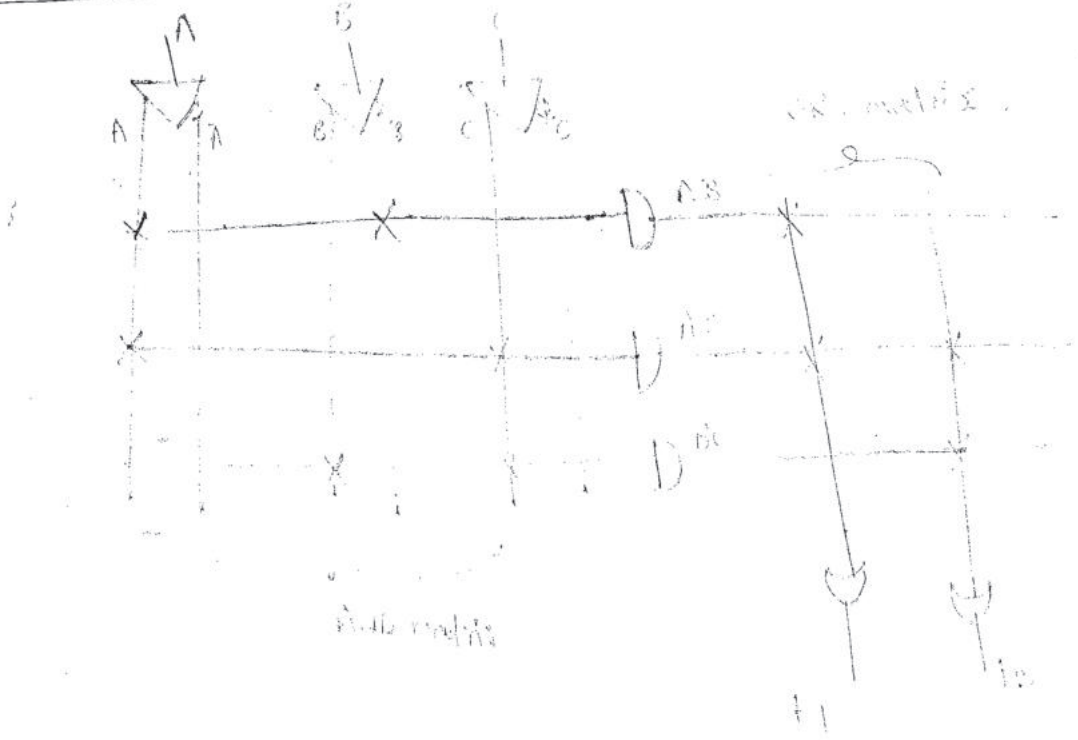
→ Decide Inputs to AND matrix.

From expressions of F_1 & F_2 we have A, $\bar{B}$, B, C Inputs.

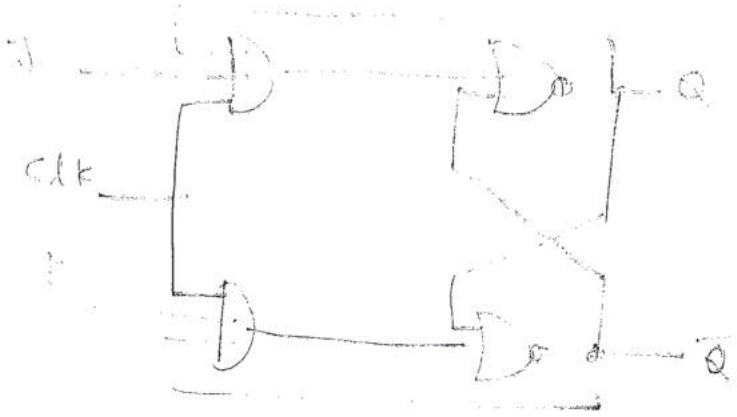
→ Decide inputs to OR matrix.

The Inputs to OR matrix are $A\bar{B}$, AC, $\bar{B}C$.

PLA Implementation :—



(2) Implement JK Flip-flop circuit using PLA.



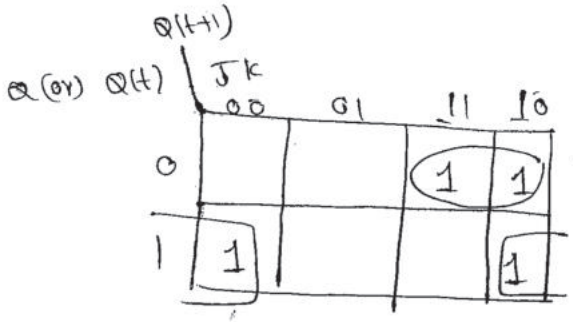
The Characteristic table or truth Table of JK-FF is,

Present state Q(t)	J	K	next state Q(t+1)
0	0	0	0
0	0	1	0
0	1	0	1 ✓
0	1	1	1 ✓
1	0	0	1 ✓
1	0	1	0
1	1	0	1
1	1	1	0

$Q(t) = 0$ then $Q(t+1) = J$

$Q(t) = 1$ then $Q(t+1) = \overline{K}$

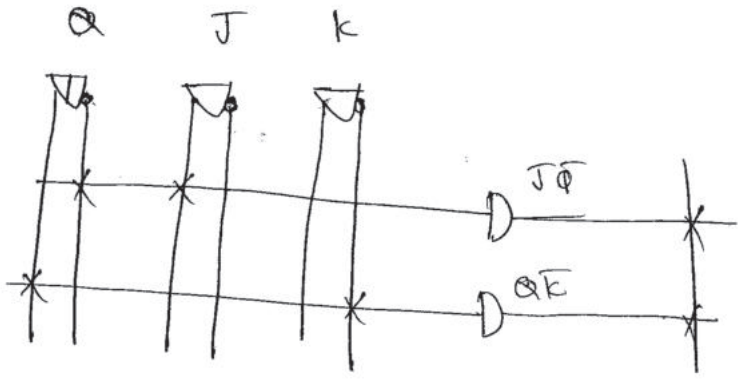
Simplify using k-map;



$$Q(t+1) = \overline{Q}J + Q\overline{K}$$

→ Inputs required for AND gates are, Q, $\overline{Q}$, J, $\overline{K}$

→ 2 AND gates required & one OR gate for 1 output.

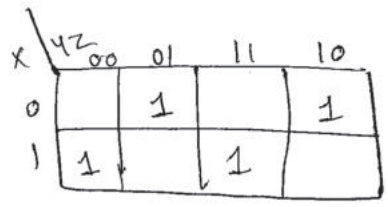


(1) Sketch a diagram for 2 input XOR using PLA.

(2) Implement full adder using PLA.

X	Y	Z	sum	carry
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

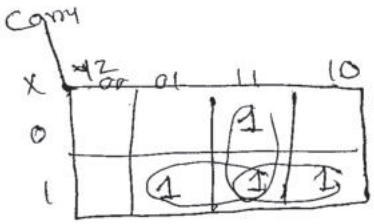
Sum



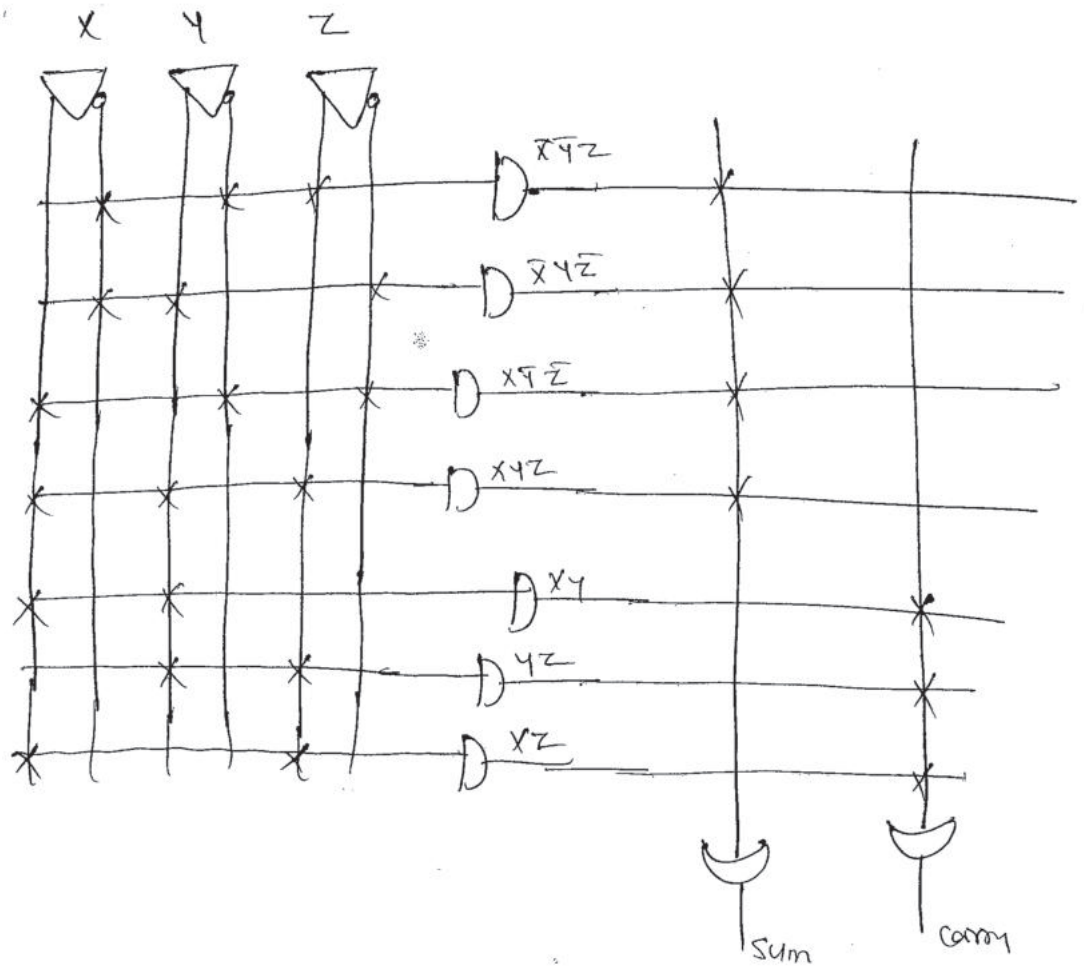
$Sum = \bar{X}\bar{Y}Z + \bar{X}Y\bar{Z} + X\bar{Y}\bar{Z} + XYZ$

Cannot simplified using k-map.

Carry



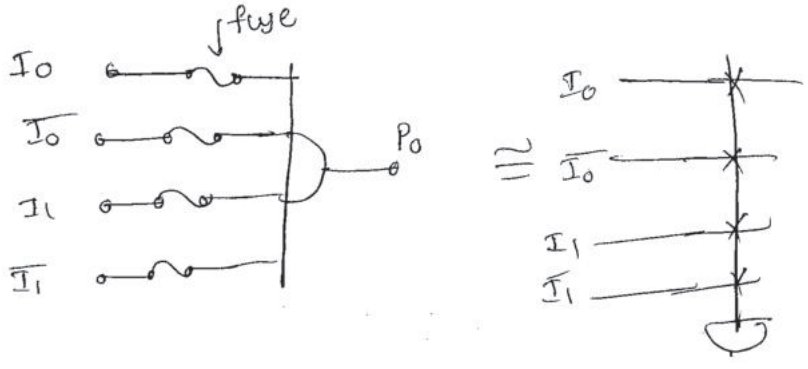
$Carry = XY + XZ + YZ$.



Programmable Array logic (PAL) ;—

The PAL is a special type of PLA where the OR array is not programmable.

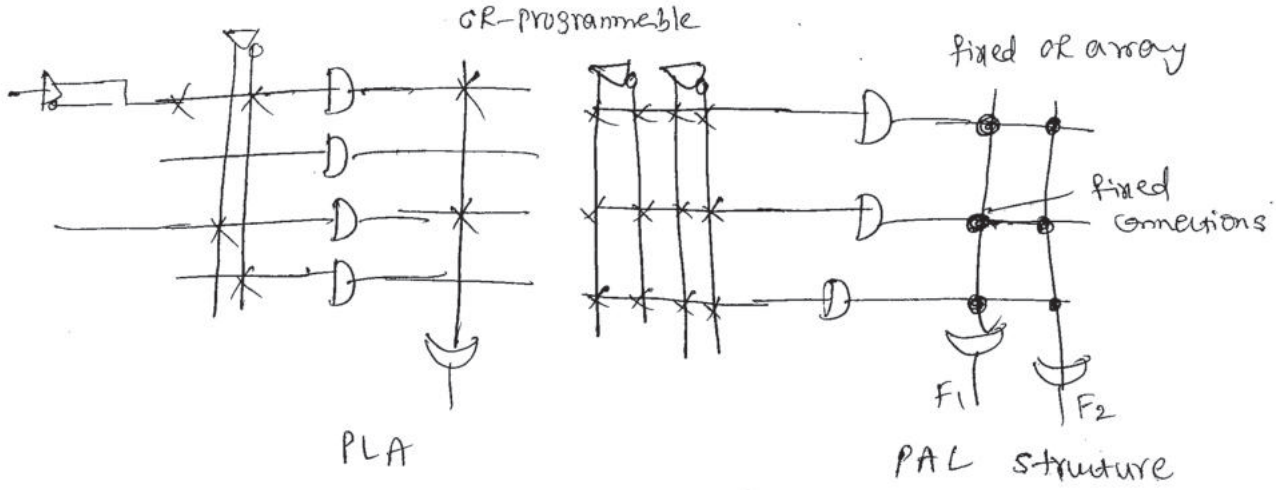
→ means In PAL AND array is programmable but OR array is fixed, whereas in PLA, both arrays are programmable.



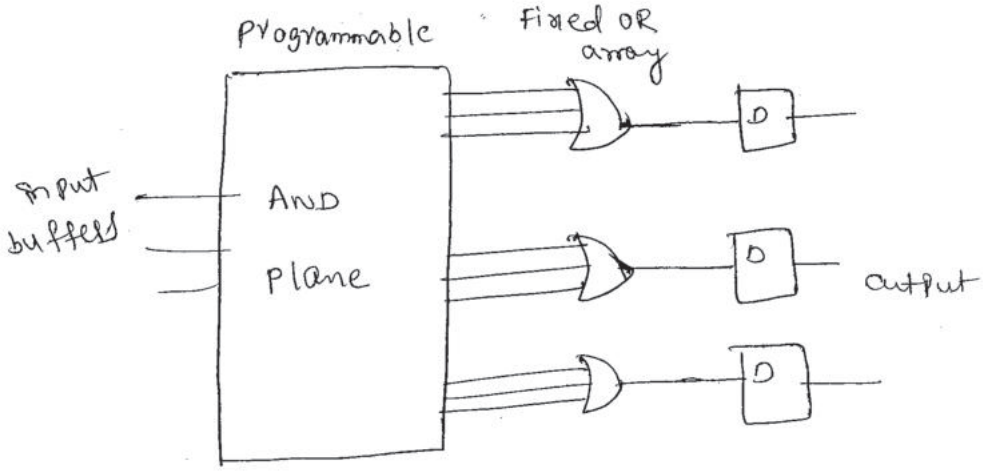
→ The advantage of PALs is the elimination of fuses in the OR array and special electronic circuits to blow these fuses.

→ since these special electronic circuits and programmable OR array occupy a very large area, the area is reduced in PAL.

→ many AND gates in first level and one OR gate of the network output



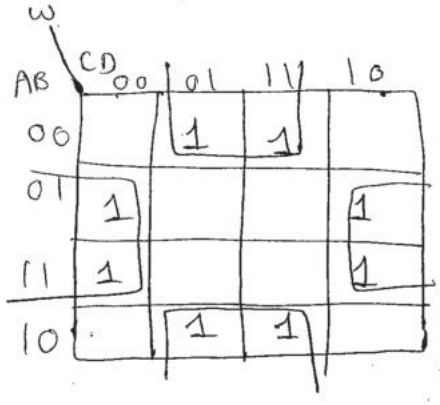
PAL structure:



(1) Implement following Boolean function using PAL.

$$W(A, B, C, D) = \Sigma m(1, 3, 4, 6, 9, 11, 12, 14)$$

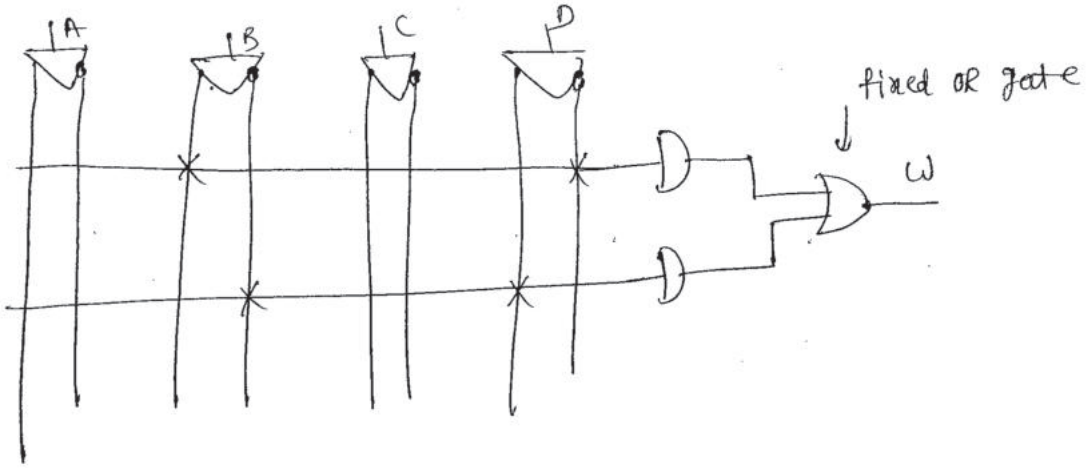
→ Simplify using K-map.



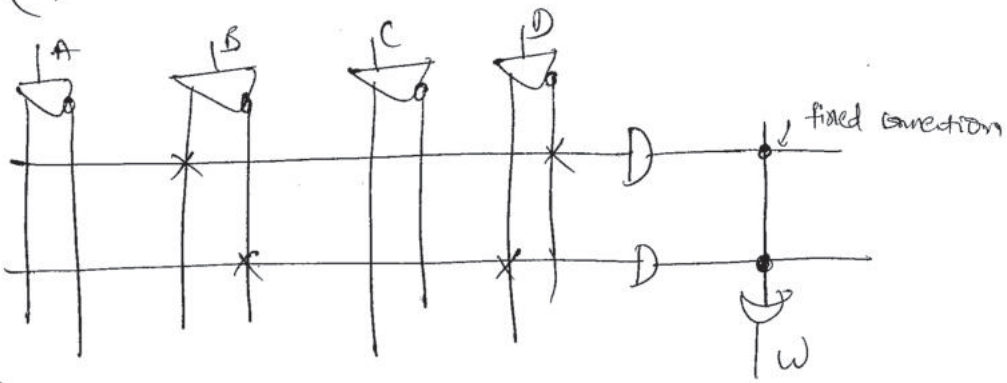
8421

1 → 0001
3 → 0011
4 → 0100
6 → 0110
9 → 1001
11 → 1011
12 → 1100
14 → 1110

$$W = B\bar{D} + \bar{B}D$$



(or)

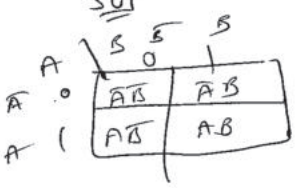


PAL 16L8

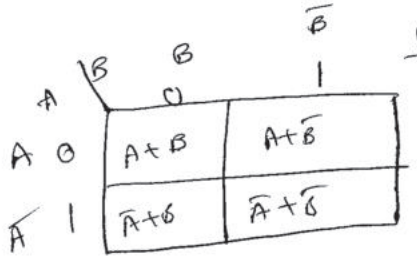
Seven AND gates to each OR gate.

Product of Sum (POS) simplification using k-map:

SOP

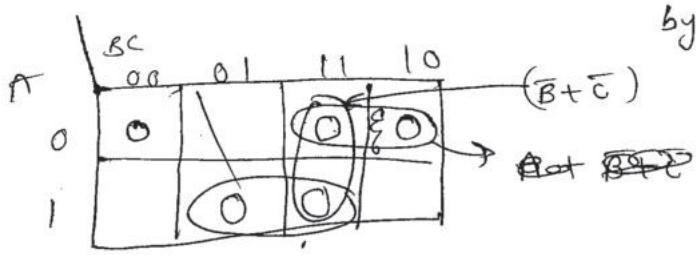


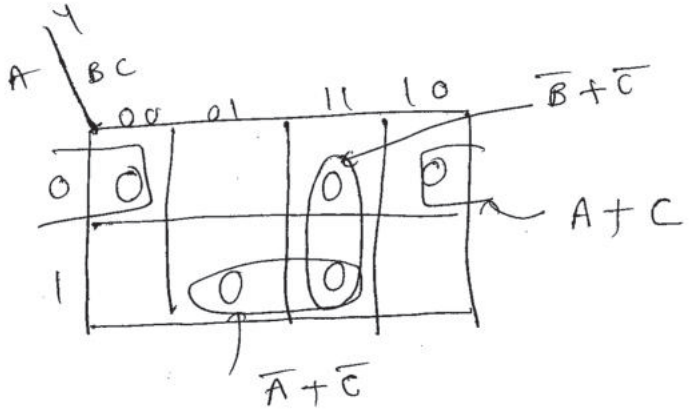
POS



Ex:-

$Y = \pi m(0, 2, 3, 5, 7)$ represents pos terms by '0'.

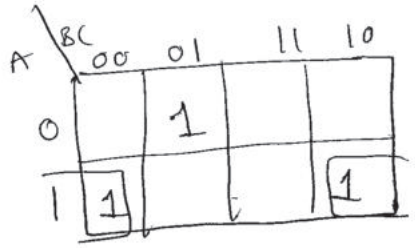




$Y = (\bar{A}+\bar{C}) \cdot (\bar{B}+\bar{C}) \cdot (A+C)$

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

(OR)



1, 4, 6.

$$Y = (\bar{A}+\bar{C}).(\bar{B}+\bar{C}).(A+C)$$

$$= (\bar{A}A + \bar{A}C + \overset{A\bar{C}+}{\bar{C}}C)(\bar{B}+\bar{C})$$

$$= (\bar{A}C + A\bar{C})(\bar{B}+\bar{C})$$

$$= (\bar{A}C\bar{B} + \bar{A}C\bar{C} + AB\bar{C} + A\bar{C}\bar{C})$$

$$= \bar{A}C\bar{B} + 0 + AB\bar{C} + A\bar{C}$$

$$= \bar{A}\bar{B}C + AB\bar{C} + A\bar{C}$$

Binary to Gray encoder :

B_3	B_2	B_1	B_0	G_3	G_2	G_1	G_0
0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	1
0	0	1	0	0	0	1	1
0	0	1	1	0	0	1	0
0	1	0	0	0	1	1	0
0	1	0	1	0	1	1	1
0	1	1	0	0	1	0	1
0	1	1	1	0	1	0	0
1	0	0	0	1	1	0	0
1	0	0	1	1	1	0	1
1	0	1	0	1	1	1	1
1	0	1	1	1	1	1	0
1	1	0	0	1	0	1	0
1	1	0	1	1	0	1	1
1	1	1	0	1	0	0	1

$G_3 = B_3$

$G_2 = B_3 \oplus B_2$

$G_1 = B_2 \oplus B_1$

$G_0 = B_1 \oplus B_0$

$G_{13} =$

Standard cells :-

→ Standard cells are pre-defined logic elements used in the circuit.

→ The design methodology that uses standard cells is known as cell-based design methodology.

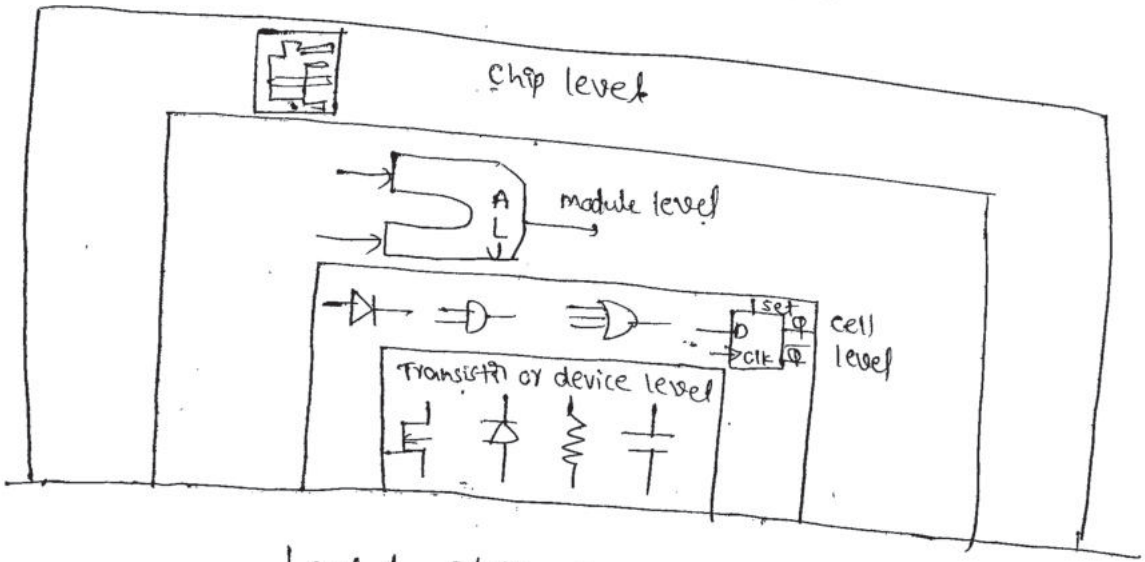
→ Hence, standard cells are the basic building blocks of cell-based IC design methodology.

→ A standard-cell library is one of the foundations upon which the VLSI design approach is built.

→ A standard cell is designed either to store information or perform a specific logic function (Inverter, AND, OR etc).

→ The type of standard cell created to store data is referred to as a sequential cell. (FF & latch).

→ Standard cells are built on transistors. They are one abstraction level higher than Transistor.



Level of abstraction.

→ Hardware block can be represented in four different abstraction levels during the chip implementation process.

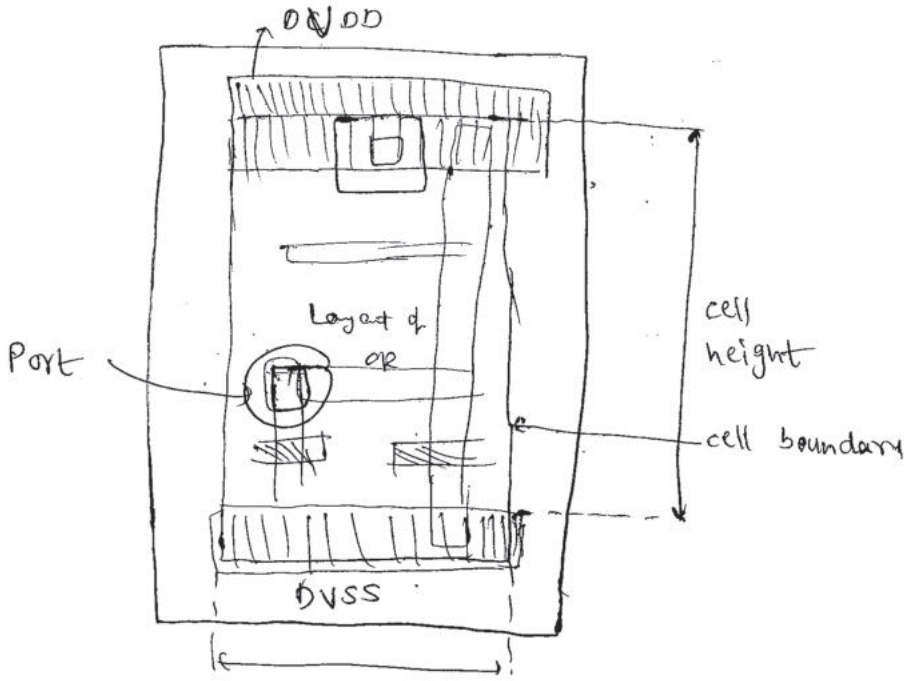
Lowest level :- The lowest level is the transistor or device level.

→ At this level, entire block is described directly by very basic building elements of transistors, diodes, capacitors and resistors.

Cell level :- In this, designs are composed of standard cells.

module level:- Designs are represented by modules such as adders, multiplier, ALU & shifter.

Chip level :- The highest level is chip level, At this level, designs are partitioned into subsystems, such as DSP, micro controller, MPEG decoder, UART, USB, DMA, ADC, DAC & PLL.



→ During chip construction process, the designer's HDL code is transformed to a netlist using synthesis tool. The netlist is composed of a certain number of standard cells, each one having its specific logic function.

→ The intended system functions, described by HDL are realized by standard cells in this netlist.

→ These standard cells are placed within the chip's floorplan by special place and route tool.

Cell's physical size defined by

1) cell height
2) cell width.

→ cell boundary attribute used by place and route tool to place cells during placement stage.

→ Only metal1 layer is used inside the cell layout since higher level metals are reserved for signal routing.

→ It is important to complete each standard cell's layout with least amount of silicon area.

→ Physically, the standard cells within an ASIC library have a fixed size in one dimension (usually height) so that they can be placed and aligned along rows of the chip.

⇒ The other name of cell-based design methodology is cell-based ASIC or CBIC in simple.

Factors Influencing Low Power VLSI Design :

1. Reduce V_{DD}.
2. Power supply Reduction
3. Variation of the threshold voltage
4. Optimal Power voltage
5. Compensating for lower speed.
6. Voltage switch
7. Reduce C
8. Partition Blocks
9. Locality of Reference
10. Clocks and control.
11. Logic design.
12. Buffer design
13. Reduce A
14. Glitch Avoidance
15. Point-to-Point Buses.

Reduce V_{DD} :-

→ Development in fabrication are already moving from the existing standard 5V towards a new level of 3.3 V and experimental processes are looking at even lower voltages.

Power Supply Reduction :—

→ One of the motivation in technology development has been to increase the levels of Integration by Reducing feature sizes.

→ However as gate lengths are reduced the electric field strength increases in the gate region. This leads to realiability problems as the high electric field strengths accelerate the Conducting es to such speeds cause substrate current and then cause latch-up Problem.

→ There are 3 approaches to enabling further feature size reduction

(1) lightly doped drain (LDD) technique allows the smallest gate length.

(2) new circuit techniques which avoids high electric fields across individual transistors.

(3) Reduce the supply voltage.

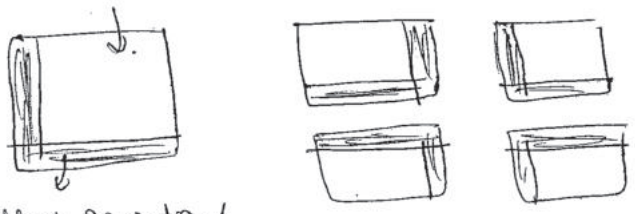
$$\text{Power delay} = \frac{P}{f_{max}} = V_{DD}^2 \, \Sigma (C_i P(r_i))$$

→ The variation in V_{DD} actually leads to a quadratic change in the power-delay product.

f_{max} → max possible frequency of a circuit represents the fastest throughput.

Partition Blocks

In general it is best to partition large blocks into smaller ones. But, the power calculation for each memory access is based upon the capacitances of the bit and word lines which run vertically & horizontally across array.



address generation & bit detectors circuit.

→ Instead of array is broken down into 4 sub-circuits (each with its own support circuitry) and only one unit is addressed with each access, then the product of activity & capacitance is reduced by a half.

Reduce C:

The best strategy is to Reduce capacitance.

Buffer design:-

One problem is the design of circuitry to drive large capacitance. The basic solution is a sequence of buffers with increasing gate widths;

Reduce A:

Reduce A, the average activity on each gate. Power is only expended when a node is switched, If switching is restricted to when Information Changes then Power is minimised.

Glitch Avoidance:-

With some digital logic, there are spurious transitions known as glitches which occurs due to partially Resolved functions.

Testing

→ Tests fall into 3 main categories :-

① Functionality tests or Logic verification :

The first set of tests verify that the chip perfoms its intended function. These tests are run before tapeout to verify the functionality of the circuit.

② Silicon debug :-

The second set of tests are run on the first batch of chips that return from fabrication.

→ They can be much more extensive than the logic verification tests because the chip can be tested at full speed in a system.

For example, a new microprocessor can be placed in a prototype motherboard to try to boot the operating system.

→ The silicon debug requires creative detective work to locate the cause of failures because the designer has much less visibility into

chip compared to during design verification.

(3) Manufacturing Tests :—

These set of tests verify that every transistor, gate and storage element in the chip functions correctly.

→ These tests are conducted on each manufactured chip before shipping to the customer to verify that the silicon is completely intact.

$$\text{Yield of an IC} = \frac{\text{no. of good die}}{\text{total number of die per wafer}}$$

→ Because of the complexity of the manufacturing process, not all die on the wafer function correctly.

→ Dust particles and small imperfections in starting material or photomasking can result in bridged connections or missing features.

→ These imperfections result in what is termed as a Fault.

→ The goal of the manufacturing test procedure is ~~to~~ determine which die are good and should be supplied to customers.

Testing a chip(die) can occur at different levels

1. wafer level
2. Packaged chip level
3. board level
4. system level
5. Field level.

→ By detecting a malfunctioning chip early, the manufacturing cost can be kept low.

→ For instance, the approximate cost to a company of detecting a fault at various levels is,

wafer	$0.01 – $0.10
Packaged chip	$0.10 – $1
board	$1 – $10
system	$10 – $100
field	$100 – $1000

~~Need For Testing~~ :—

From above example, it is clear that the manufacturing cost is kept low, if the faults are detected at wafer level (bad devices mine).

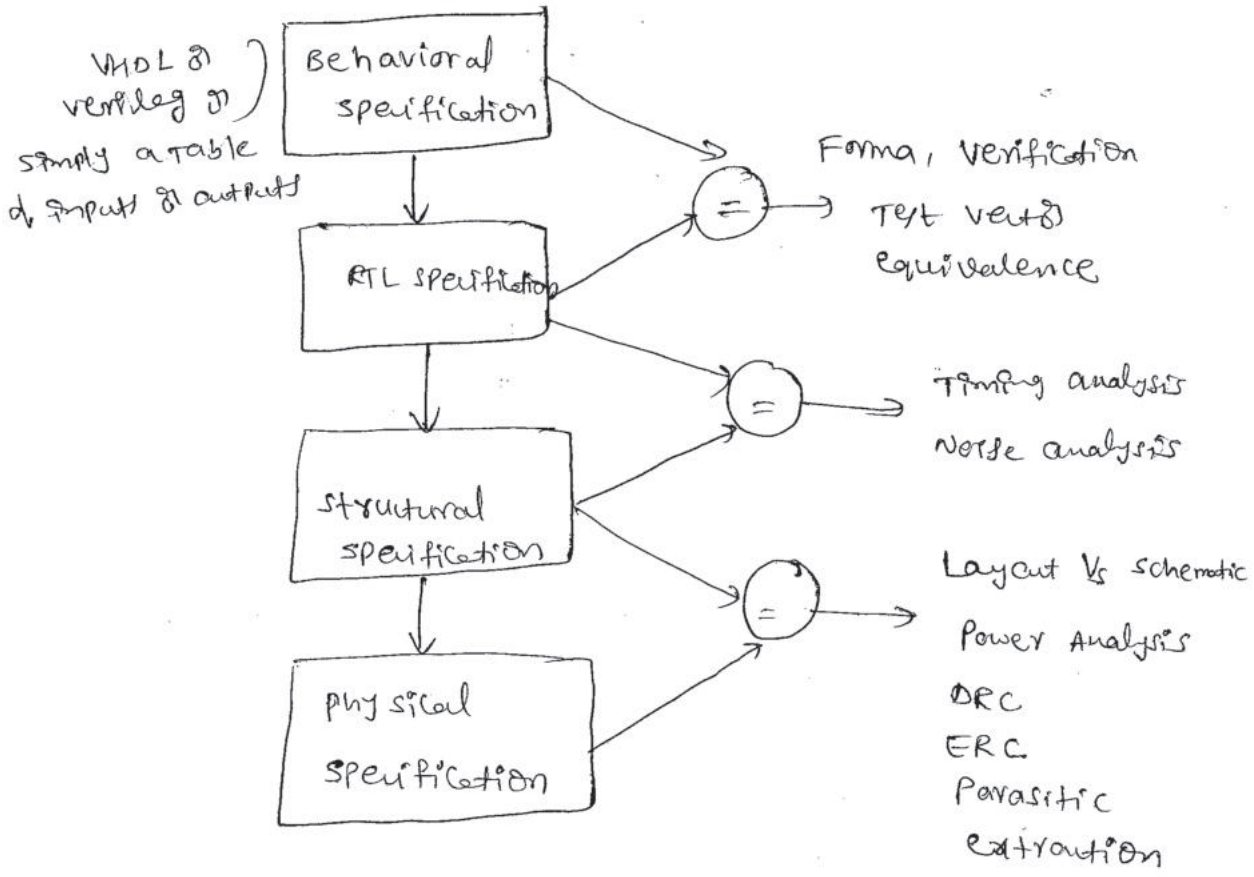
→ If Yield high & Package cost low (ie plastic Package) then Part can be tested only after packaging.

→ If wafer Yield is lower and package cost high (ceramic package) it is better to screen bad dice at wafer level.

Logic & Functional Verification:-

→ verification tests are usually the first one a designer might construct as a part of the design process. like adder adding or not, counter counting or not.

→ In Functional Tests it is required to prove that a synthesized gate description was functionally equivalent to the source RTL.



→ Functional equivalence involves running a simulator at some level on the 2 descriptions of the chip (one at the gate level and one at functional level) and ensure that the outputs are equivalent at some convenient check points in time for all inputs applied. This is mostly convenient in HDL by employing a Test bench.

→ We have to simulate as closely as possible the way in which the chip or system will be used in the real world.

Manufacturing Tests :—

These tests are used to verify that every gate operates as expected. The need to do this arises from a number of manufacturing defects that might occur during either chip fabrication or accelerated life testing (where the chip is stressed by over voltage or over temperature).

Typical defects include:

- Layer to Layer shorts (e.g. metal-to-metal)
- discontinuous wire (eg: metal thins when crossing vertical topology jumps)
- missing or damaged via
- shorts through the thin gate oxide to the substrate or well.

These convey circuit maladies like,

① nodes shorted to power or ground

② nodes shorted to each other

③ inputs floating / outputs disconnected.

Apart from the verification of internal gates, I/o integrity is also tested with the following Tests being completed.

① I/o levels (ie. checking for noise margin for TTL, ECL & CMOS I/o pads).

② speed Test.

→ In general, manufacturing test generation assumes the function of the circuit/chip is correct. It requires ways of exercising all gate inputs and monitoring all gate outputs.

Functional verification principles:—

① Test Benches and Harnesses.

② Regression Testing

③ Version control

④ Bug Tracking.

① Test Benches and Harnesses:—

a verification test bench or harness is a piece of HDL code that is placed as a wrapper around a core piece of HDL.

→ In Test bench, inputs are applied to the module under test and at each cycle, the outputs are examined to determine

wheather they comply with a predefined expected data set.

→ The data set can be derived from another model and available as a file or the value can be computed on the fly.

Regression Testing :—

High level language (C, FORTRAN, Pascal & LISP) scripts are frequently used when running large test benches, especially for "regression testing".

→ Regression Testing involves performing a suite of simulations to automatically verify that no functionality has inadvertently changed in a module or set of modules.

Version Control :—

Combined with regression testing is the use of versioning, ie, the orderly management of different design iterations. Unix / Linux Tools such as CVS (Circuit Vs Schematic) are useful for this.

Bug Tracking :—

Important tool to use during verification is a "bug-Tracking" system.

→ Bug Tracking systems such as Unix/Linux based GNATS allow the management of a wide variety of bugs.

→ Each bug is entered & noted the location, nature and severity of the bug.

Manufacturing Test principle :-

(1) observability :-

→ The observability of a particular circuit node is the degree to which you can observe that node at the outputs of an Integrated circuit (ie, the pins).

→ This metric is relevant when you want to measure the output of a gate within a larger circuit to check that it operates correctly.

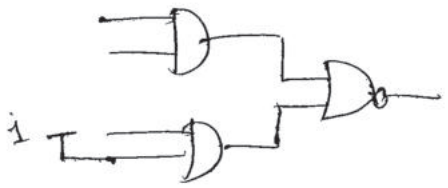
(2) Fault model :-

This is a model for how faults occur and their impact on circuits. The most popular model is called "Stuck-At" model.

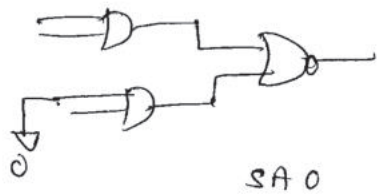
→ The short circuit / open circuit model can be closer fit to reality, but is harder to incorporate into logic simulation tools.

⇒ Stuck-At faults :-

In this model, a faulty gate input is modeled as a stuck at zero (stuck-At-0, S-A-0) or stuck At one (S-A-1).



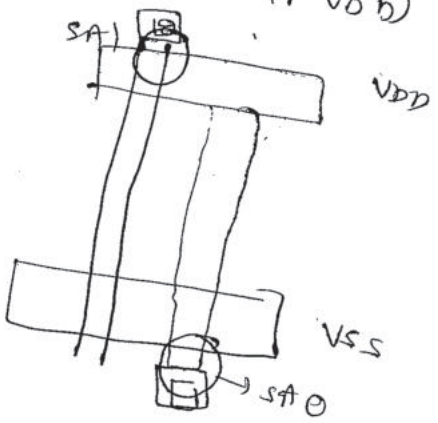
Stuck At 1



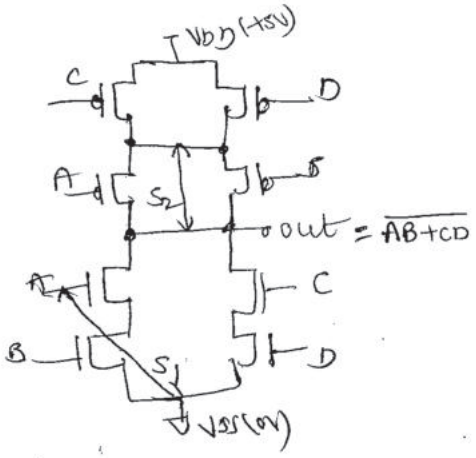
SA 0

Stuck At 0

These faults most frequently occur due to gate oxide shorts (nmos gate to GND, Pmos gate to VDD) & metal–metal shorts.



⇒ short circuit and open-circuit Faults:-

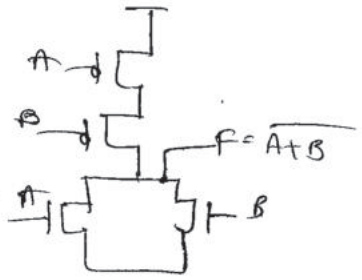


S_1 → modeled by an SA0 fault at input A, while short S_2 modifie the function of the gate.

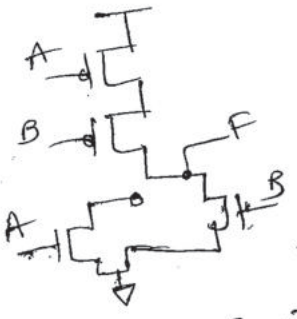
S_2 modifie function of gate

→ A problem that arises with CMOS is that it is possible for a fault to convert a combinational circuit into a sequential circuit.

Consider 2-input NOR gate



Normal circuit.



0	0	1
0	1	0
1	1	0
1	0	Q (Fn)

$$F = \overline{(A+B)} + A\bar{B}.Fn$$

→ Fn is the previous state of the gate.

If B n-Transistor drain connection is missing

the function is,

$$F = (\overline{(A+B)}) + \bar{A}.B.Fn.$$

Controllability :-

The Controllability of an internal circuit node within a chip is a measure of the ease of setting the node to a 0 & 1 state.

→ An easily controllable node would be directly settable via an input pad.

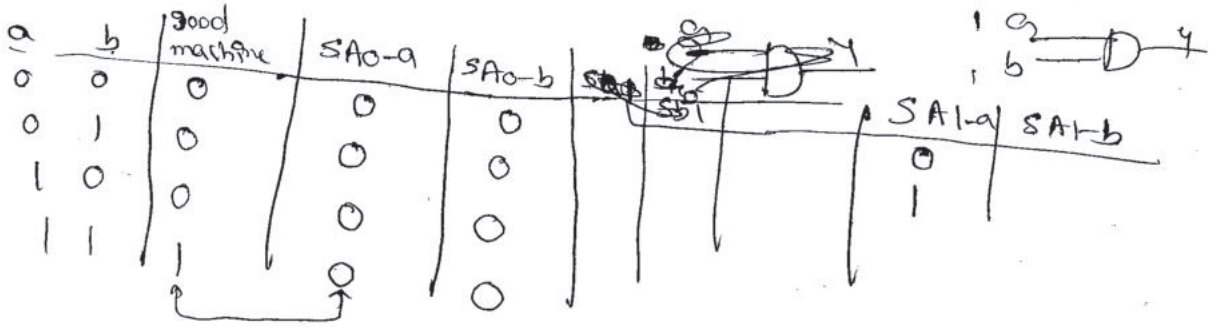
EX: Global reset to get initial positions of all inputs.

Fault coverage :-

A measure of goodness of a Test Program is the amount of fault coverage it achieves.

ie, for the vectors applied, what percentage of the chip's internal nodes were checked.

→ Each circuit node is taken in sequence and held to SA0, and the circuit is simulated, comparing the chip outputs with a known "good machine".



a	b	y	SA0-a	SA0-b	SA1-a	SA1-b
0	0	0	0	0	0	0
0	1	0	0	0	1	0
1	0	0	0	0	0	1
1	1	1	0	0	1	1

11, 01, 10 Test vectors.

→ The total number of nodes that, when set to 0 or 1, do result in the detection of fault, divided by the total number of nodes in the circuit is called "percentage fault coverage).

Fault-sampling :-

An Approach to Fault analysis is "Fault-sampling".

This is used in circuits where it is impossible to fault every node in the circuit.

→ nodes are randomly selected and faulted.

→ The resulting fault-detection rate may be statistically inferred from the no. of faults that are detected in the fault set and the size of the set.

Statistical Fault Analysis:-

→ This method of fault analysis relies on estimating the probability that a fault will be detected.

→ Extra statistics are gathered by a modified simulation on a per-input vector basis.

(1) Zero counter— The 0 count on each gate input when a 1→0 change of output is detected.

(2) One-counter— The 1 count on each gate input when a 0→1 change of output is detected.

(3) sensitization counter— incremented if the input change causes the output to be sensitized.

(4) Loop-counter— used to detect & deal with feedback.

The one-controllability of line l is given by,

$$C1(l) = \text{one-count}/N,$$

N → no. of vectors.

zero-controllability

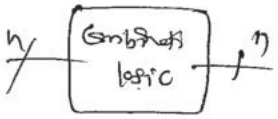
$$C0(l) = \text{zero-count}/N,$$

One-level sensitization probability is.

manufacturing Test principles:

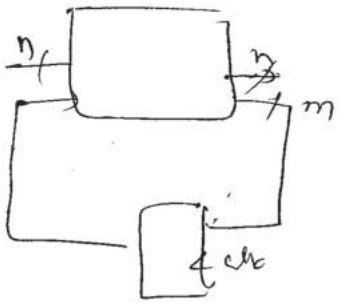
Consider a combinational circuit with n-inputs.

→ To test this circuit exhaustively, a sequence of 2^n inputs (Test vectors) must be applied & observed to fully exercise the circuit.



$n = 10 \Rightarrow 2^{10} = 1024$ Test vectors.

→ The combinational circuit is converted to sequential circuit with additional of m-storage registers.



→ The state of the circuit is determined by the inputs & previous state.

A minimum of 2^{n+m} test vectors must be applied to exhaustively test the circuit.

Ex: A n/w with $n = 25$, $m = 50$

$2^{25+50} = 2^{75}$ Test vectors

$\cong 3.8 \times 10^{22}$.

If 1 μs per pattern, then total test time would be over a billion years (10^9).

```
library IEEE;
use IEEE.STD_LOGIC_1164.ALL;

entity and1 is
    port ( a : in std_logic;
           b : in std_logic;
           y : out std_logic);
end and1;

architecture and12 of and1 is
begin
    y <= a and b;
```

```
process (a,b)
begin
    c <= a and b;
end process;
end and12;
```

Test bench:

```
library IEEE;
use -

entity and1_tb is
end and1_tb;

architecture and1_tb of and1_tb is
    signal a_s, b_s;
begin
    component and1 is
        port ( a : in std_logic;
               b : in std_logic;
               y : out std_logic);
    end component;
begin
U1 : component and1 portmap ( a => a_s,
```

```
b <= 0;
a <= 0;
wait for 1 ns;
a <= 1;
b <= 0;
wait for 1 ns;
a <= 0
b <= 1
wait for 1ns;
a <= 1
b <= 1
wait;
end and1_tb;
```

Chip-level Test Techniques :-

In this we will examine practical methods of incorporating test requirements into a design. This discussion is structured around the main types of circuit structure that will be encountered in a digital cmos chip.

1. Regular Logic Arrays
2. memories
3. Random Logic

Regular Logic Arrays :-

Partial serial scan & parallel scan is probably the best approach for structures such as datapaths.

One approach that has been used in a Lisp microprocessor is,

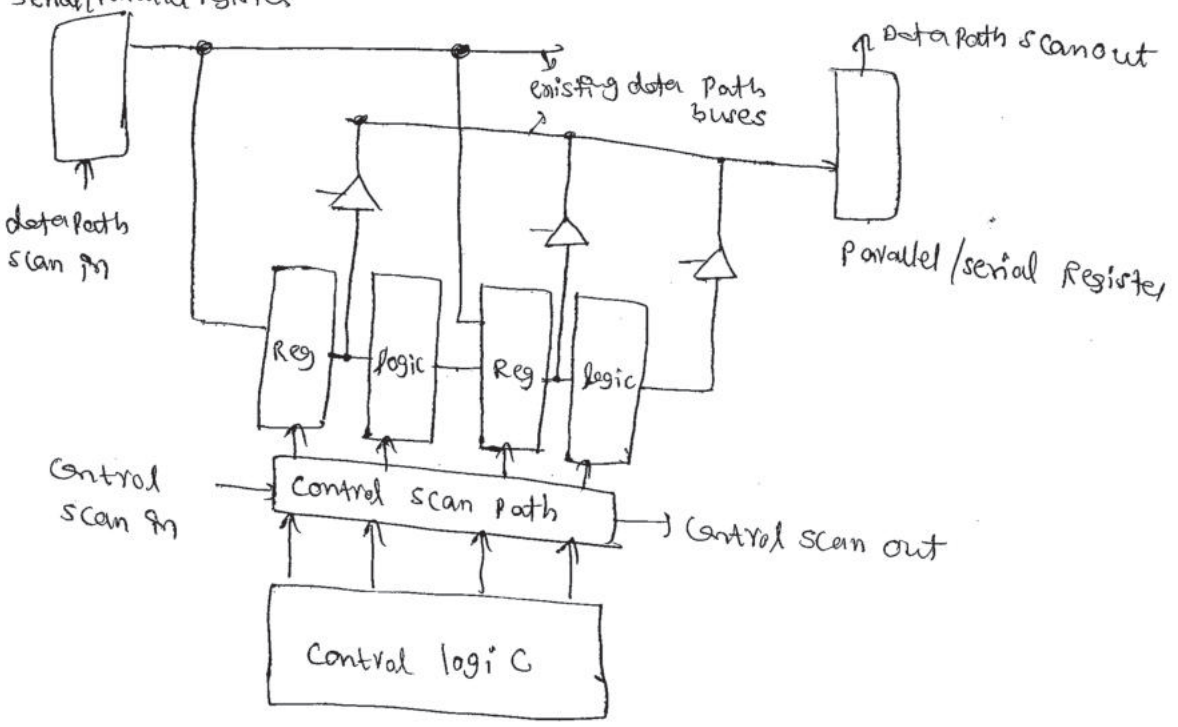


Fig :

→ The input buses may be driven by a serially loaded register.

→ They e in turn may be used to load internal data path Registers.

→ The data path Registers may be sourced onto a bus & thy bus may be loaded into a Register that may be serially accessed. All of the control signals to the Datapath are also made scannable.

Memories:—

memory may use the self-testing techniques by embedding self Test circuits for memories in higher-speed circuits.

→ Another way, the provision of multiplexers on data inputs and addresses and convenient external access to data outputs enables the testing of embedded memories.

→ It is a mistake to have memories indirectly accessible ie, data is written by passing through logic.
data is observed after passing through logic.
address cannot be conveniently sequenced.

→ Because memories have to be tested exhaustively, any overhead on writing and reading the memories can substantially increase the test time.

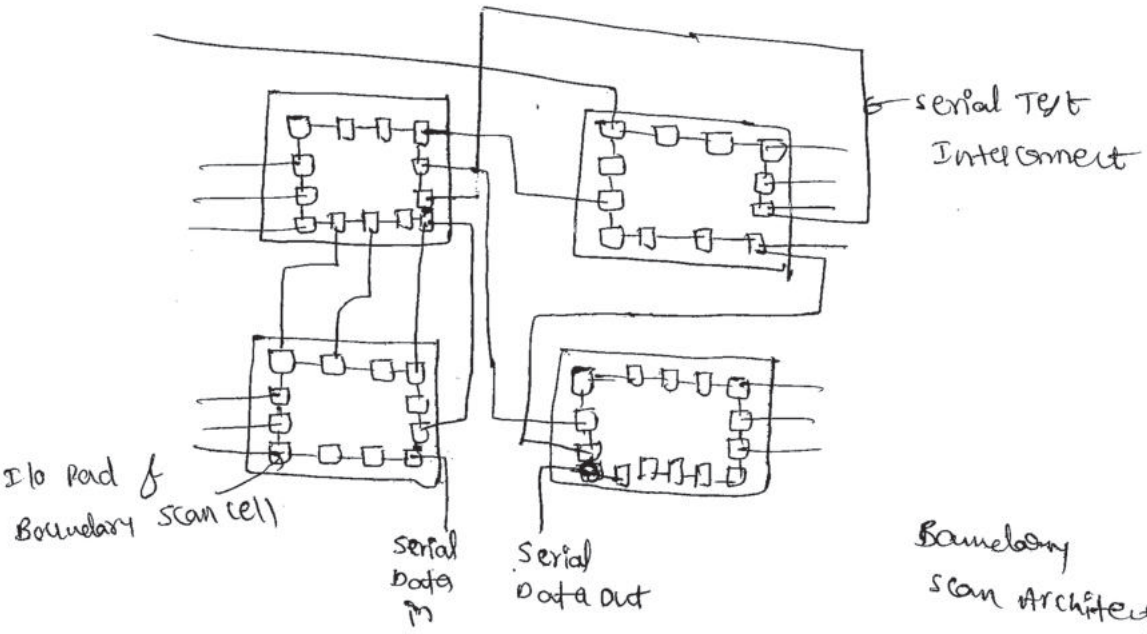
Random logic:—

Random logic if probably best tested via full serial scan &

System level Test Techniques:-

The increasing complexity of boards and the movement to different technologies like MCM (multi chip modules) resulted in system designers agreeing on a unified scan-based methodology for testing chips at the board and system level. This is called "Boundary Scan".

Boundary scan:-



Boundary scan Architecture

→ It provides a standardized serial scan path through the I/o pins of an IC.

TAP: (Test Access Port)

The TAP is a definition of the interface that need to be included in an IC to make it capable of being scan.

in a Boundary-scan architecture.

→ The port has 4 or 5 single bit connections.

① TCK (Test clock input)– used to clock tests into & out of chips.

② TMS (Test mode select)– used to control test operations.

③ TDI (The test Data input)– Used to input test data to a chip

④ TDO (Test Data output)– Used to output Test Data from a chip.

optional pin:

* TRST (Test Reset signal) used to asynchronously reset the TAP controller.

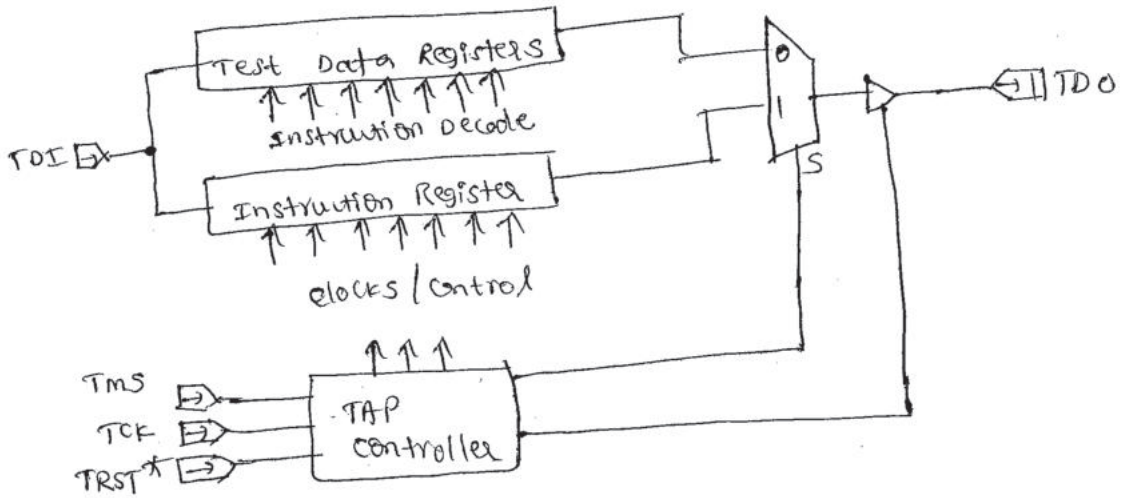
The Test Architecture:–

It consists of ① The TAP Interface pins

② Set of test-data Registers to collect data from chip.

③ an Instruction Register to enable test inputs to be applied to the chip.

④ A TAP controller, which interprets test instructions & control the flow of data into & out of the TAP.



Controller :-

The TAP controller is a 16-state FSM that proceeds from state to state based on the TCK and TMS signals. It provides signals that control the test data Registers, and Instruction Register. These include serial-shift clocks & update clocks.

The Instruction Register (IR):

The IR has to be atleast 2 bits long & logic detecting the state of the IR has to decode at least 3 Instructions. They are,

① BYPASS — This Instruction is represented by an IR having all zero's in it.

→ It is used to bypass any serial-data Registers in a chip with a 1-bit Register.

② EXTEST — This allows for the testing of off-chip circuitry and is represented by all ones in the IR.

③ SAMPLE/PRELOAD — This places the boundary scan registers (ie at chip's I/O pins) in the DR chain & samples or preloads the chips I/o's.

Test Data Registers (DRs):

They are used to set the inputs of module to be tested and to collect the results of running tests.

→ The simplest data-Register configuration would be a boundary scan register and a bypass Register (1-bit long).

→ A multiplexer under the control of TAP controller selects which particular data register is routed to the TDO pin

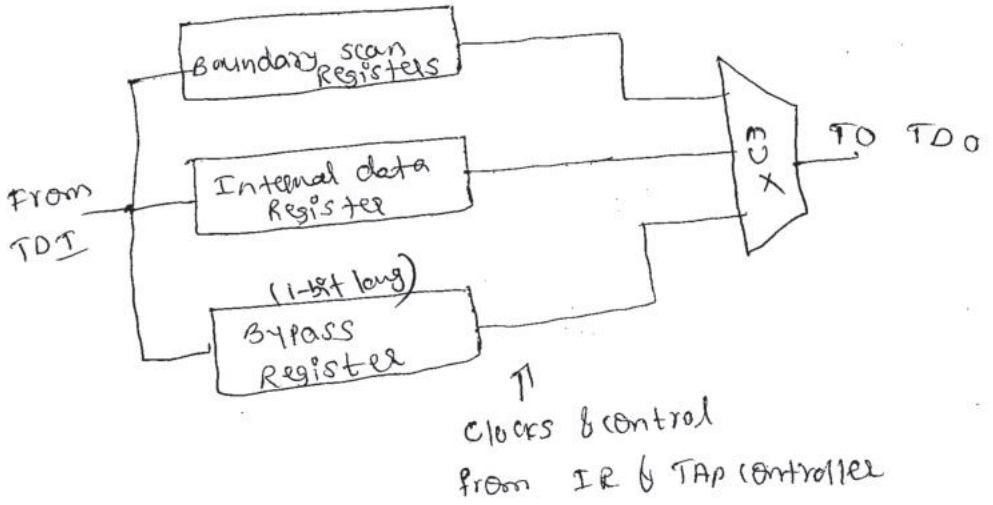


Fig: TAP data Register.

Boundary scan Registers :-

This is a special case of data Register. It allows circuit-board interconnections to be tested, external components tested & the state of the chip digital I/O to be sampled.

→ It consists of 2 multiplexers & 2 edge-triggered Registers.

Design strategies for Test :-

Design For Testability :-

The 3 main approaches to what is commonly called "Design for Testability". These may be categorized as

(1) Ad-hoc Testing
(2) Scan-based approaches
(3) self-test & Built-in Testing

Ad-Hoc Testing :-

Ad-hoc Test Techniques are collections of ideas aimed at Reducing the combinational explosion of Testing. Common Techniques involve:

1. Partitioning large sequential circuits.
2. adding Test points.
3. adding multiplexers
4. Providing for easy state reset.

Ex: Long Counters.

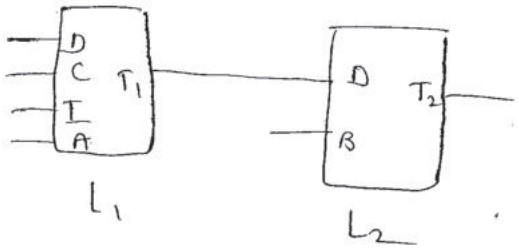
Scan-based Test Techniques :-

1. Level sensitive scan Design (LSSD)

LSSD is based on 2 tenets.

1. The circuit is Level sensitive
2. Each register may be converted to a serial shift Register.

The basic building block of LSSD is shift Register Latch (SRL).



SRL

SRL consists of 2 latches L_1 & L_2.

L_1 has a serial data Port I,
enable A
Data Port D
Enable C.

when A is high ⇒ value of $L_1(T_1)$ is set by value of I.

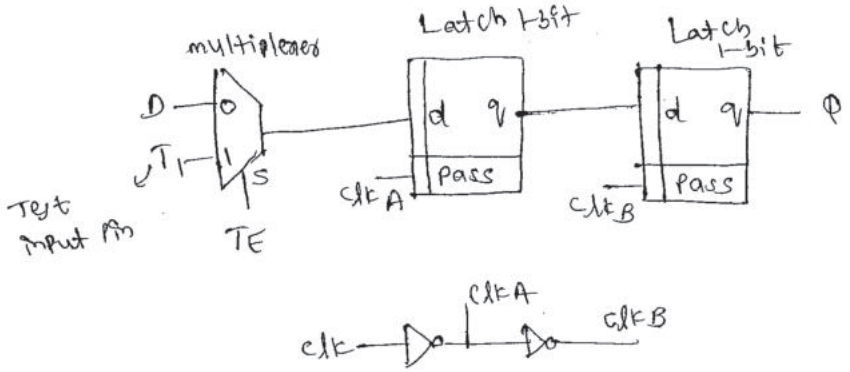
when C is high ⇒ value of $L_1(T_1)$ is set by D.

A & C can not be simultaneously high.

→ when signal B in L_2 is high ⇒ $T_2 \Leftarrow T_1$ (T_1 passed to T_2).

Serial scan:—

Schematic for a cmos edge-sensitive scan-Register is



→ A mux is added before the master latch in a conventional D-Register.

→ TE is Test enable pin, T_1 is Test Input pin

When TE enabled ⇒ T_1 is clocked into Register by Rising edge of Clk.

Partial serial scan :-

In Dataflow section of the chip, pipeline Registers removed. In this case only input & output Registers need be made scannable. This technique of Testing is known as "Partial Scan", & depends on the designer making decisions about which registers need to be made scannable.

→ In full scan Test, all Registers would have to be scannable.

→ The part of the circuit that is being tested and monitored by the scan Registers known as "kernel".

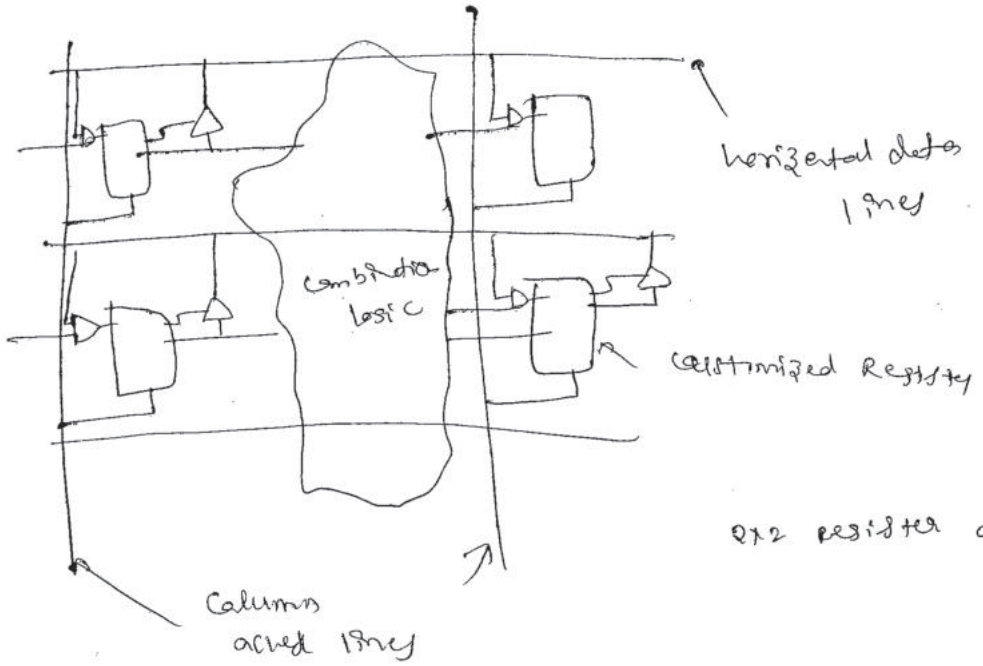
Parallel scan :-

→ serial scan Chains can become quite long & loading, unloading sequence can dominate Test time. An extension of serial scan is called "Random access & Parallel scan".

→ In Parallel scan,

each Register in the design is arranged on an imaginary or real grid where Registers on common rows receive common data lines

Registers in common columns Receive common Read & write control signals.



8x2 register array.

fig: Parallel scan- basic structure.

→ Any registed output may be observed by enabling the appropriate column Read line & setting the appropriate address on an output data multiplexer.

→ similarly, data may be written to any Register.

Layout Design for Improved Testability :—

→ In order to predict layout styles that improve testability, a designer has to have some idea of nature and frequency of defects for a particular process.

→ The types of defects divided into (1) shorts circuits
(2) open circuits.

→ shorts occurs in all layers used for connections ie, diffusion, polysilicon, metal1, metal2. & oxide short to substrate & source or drain.

→ All conducting layers might have open circuits. In addition [illegible] lead to Intralayer

Memory self-Test:-

Embedding self-Test circuits for memories in higher speed circuits not only may be the way of testing the structures at speed but can save on the number of external Test vectors that have to be run.

EX: 4-Mbit RAM with self-Test.

The self-Test consists of 256k cycles that input a checkboard pattern to test for cell to cell interference.

→ This is followed by 256k cycles in which the data is readout. Then a complemented checkboard is written and read.

→ A total of 1 million cycles provide a test sufficient for system maintenance.

→ The advantage of self-Test methods is that "Testing may be completed. when the part is in the field."

→ we can perform self-Test even during normal system operation with care.

Iterative Logic Array Testing: (ILA Testing)

→ Arrays of logic present is the problem to the Test architect because the replication can be used to advantage in reducing number of Tests.

→ An ILA is a collection of identical logic modules (Ex: n-bit adder).

→ An ILA is C-Testable if it can be tested with a constant

An ILA is I-Testable if a particular fault that occur in any module of a vector of an applied input vector is identical for all modules in ILA.

IDDQ Testing :—

Popular method of testing for bridging faults is called "IDD Q" (VDD supply current Quiescent) or current-supply monitoring.

→ When cmos logic gates not switching it draws no DC current (except for leakage).

→ When bridging fault occur, for some combination of inputs a measurable DC IDD will flow.

→ Applying normal vectors, allowing signals to settle & then measuring IDD. This is IDDQ testing.

→ Because current measuring is slow, tests must be run slower than normal ; so increasing Test Time.

↳ But gives a form of indirect native observability at little circuit overhead.