Module 2

MOS and BiCMOS Circuit Design Processes

2.1 MOS Layers:

- MOS circuits are basically formed by 4 layers
  - ✦ Metal
  - ✦ Polysilicon
  - ✦ N diffusion
  - ✦ P diffusion
- Here all the 4 layers are isolated from each other through thick or thin oxide layer (i.e., silicon dioxide layer)
- The thin oxide (thinox) layer includes n-diffusion, p-diffusion and transistor channel.

2.2 Stick diagram:

- Stick diagrams are a means of capturing topography and layer information using simple diagrams.
- They convey layer information through color codes (or monochrome encoding).
- Acts as an interface between symbolic circuit and the actual layout.
- Stick diagrams do show all components/vias(contacts), relative placement of components and helps in planning and routing. It goes one step closer to layout.
- However they do not show exact placement of components, transistor sizes, length and width of wires also the boundaries. Thus we can say that it does not give any low level details.
- The color encodings chosen for different technologies is shown below.
- **Encodings for NMOS process:**

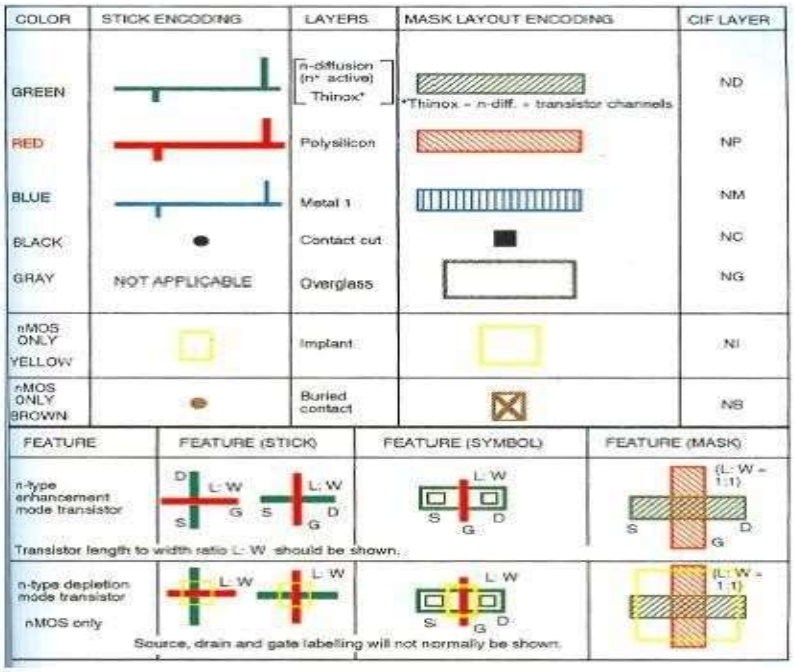


Fig.1 Encodings for NMOS process

Procedure to draw Stick Diagram:

2.3 Nmos Design Process.

1) Draw two metal lines/ power rails providing sufficient space to accommodate all transistors. i.e: Vdd & Vss.
2) Draw common n+ diffusion layer for all the transistors.
3) Provide Vdd and Vss contacts.
4) Draw polysilicon to cross n+ diffusion layer to form transistors.
5) Create buried contact for depletion transistor.
6) Provide input and output connection.

2.4 CMOS Design Process:

- Two type of transistors are used i.e: Nmos and Pmos, thus in stick diagram demarcation line is used to separate them.
- All Pmos transistors are placed above Demarcation line and Nmos transistors below demarcation line.
- While drawing stick Diagram
    1. Diffusion paths must not cross the demarcation line 2. N-diffusion and P-diffusion wires must not join.
    3. Nmos and Pmos transistors are joined by Metal layer when it is required.
    4. Cross must be placed on Vdd and Vss which represent substrate and P-well connection respectively.

Encodings for CMOS process:

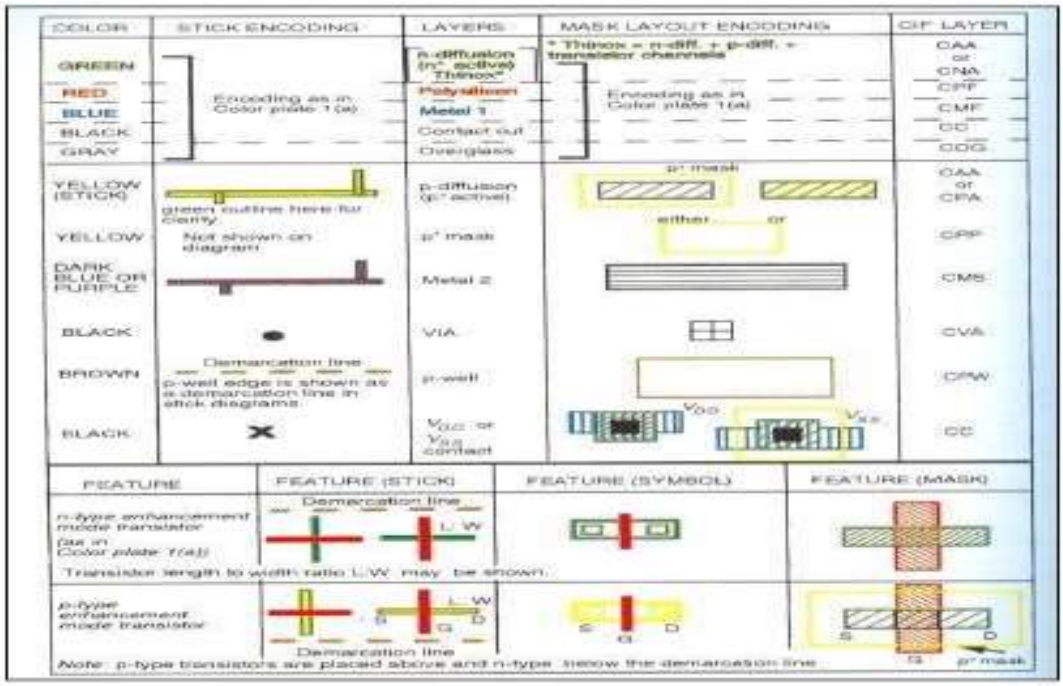


Fig.2 Encodings for CMOS process

Procedure to draw Stick Diagram:

1) Draw two metal lines/ power rails providing sufficient space to accommodate all transistors. i.e: Vdd & Vss.
2) Draw demarcation line in the middle of the two power lines.
3) Draw P+ diffusion above demarcation and N+ diffusion below demarcation
4) Draw polysilicon to represent Pmos and Nmos which represents gates of the transistor.
5) Connect source terminal of transistors to supply.
6) Drain terminals of transistor are connected using metal 1.
7) Place contact cuts wherever necessary.
8) Draw X which represents substrate and P-well contact on power lines.

2.5 DESIGN RULES AND LAYOUTS

Layout: describes actual layers and geometry on silicon substrate to implement a function (Expressions).

[Diffusion region where transistor can be formed is called active region, polysilicon serves as the gate of MOS transistor. L defines channel length and W represents width of channel/active region]

Design rules: are set of guidelines which specify minimum dimension and spacing allowed in layout drawing. Design rules also acts a communication link between circuit designers and process engineers during manufacturing phase.

Goal of design rule: is to achieve optimum yield. Yield = (No. of good chips on wafer)/(Total no. of chips on wafer).

Design rules are also called layout rules. If the circuit performance has to be increased then rules must be more aggressive. Else this leads to non-function of the circuit or yield reduction.
There are two rules.

1. Micron Rule - Absolute Dimension rule, here all sizes and spacing are specified in micron. Here the circuit density is the important goal.
2. Lambda (λ) Based Rules - The Lambda based design rules are proposed by Mead and Conway. Scalable design rules, here this design rule normalizes all geometric design rule by parameter lambda (λ) also called as scaling factor/feature size. In this all mask patterns are expressed as multiples of lambda.

Advantages of lambda based design rules:

1. The mask layout can be scaled down proportionally if the feature size of the fabrication process is reduced.
2. Design rules are conservative.
3. This rule enable technology changes and design reuse and reduced design cost.

Disadvantages:

1. Linear scaling cannot be extended and is limited over range of dimension (1-3 μm)
2. As rules are conservative, results in over dimension and density of design is less.

2.6 Lambda (λ) Based Rules

The Design rules can be conveniently set out in diagrammatic form as shown in fig. 1 for width and separation of conducting path. In fig. 2 shows the design rules associated with extensions and separations with transistor. Fig. 3 and 4 demonstrates the design rules for contacts between layers. Table below also gives the layer and distance of separation dimensions.

Layer	Dimension
n-diffusion	2λ
p-diffusion	2λ
Polysilicon	2λ
Metal 1	3λ
Metal 2	4λ

Layer -Layer	Dimension
n-diffusion – n-diffusion	3λ
p-diffusion – p-diffusion	3λ
n/p diffusion - polysilicon	1λ
Poly-poly	2λ
Metal 1	3λ
Metal 2	4λ

Layer dimension **Distance of Separation**

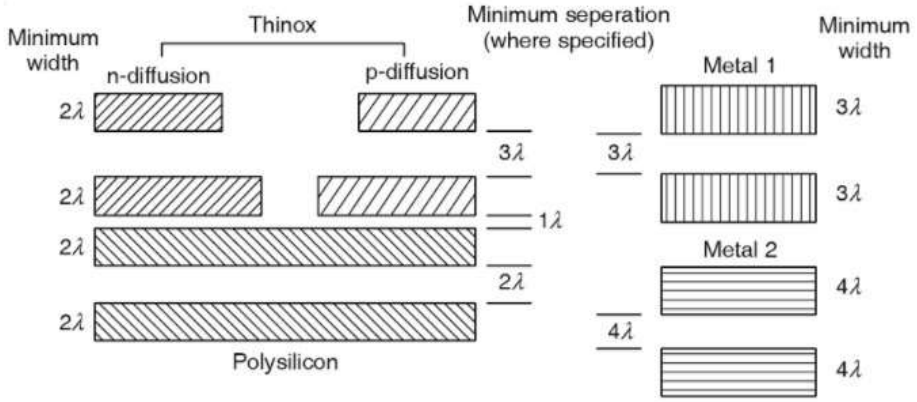


Fig. 3 Design rules for wires and separations (nMOS and CMOS)

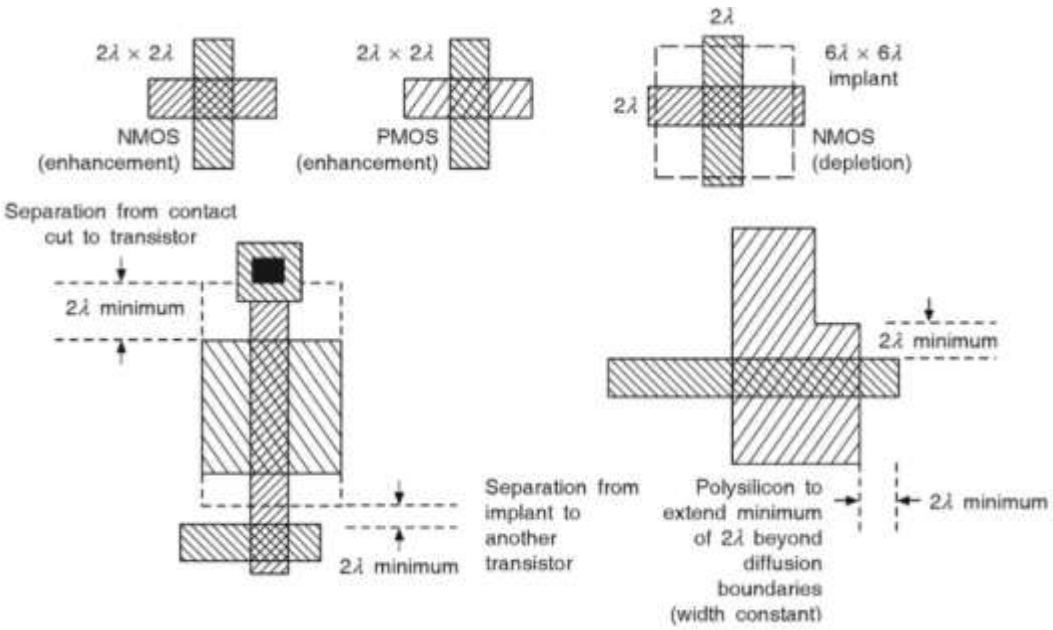


Fig. 4 Design rules for Transistors (nMOS, pMOS and CMOS)

Transistor design rules

- Minimum dimension of transistor is 2λ × 2λ – overlapping of diffusion and ploy
- Poly and diffusion both must extend beyond the boundary of transistor at least by 2λ
- Implant for depletion mode transistor is 6λ × 6λ i.e., implant must extend boundary of transistor by at least 2λ in all direction.
- From the boundary/ implant of one transistor, the next transistor should maintain min distance of 2λ
- The distance from contact cut to transistor should be at least 2λ

Metal contact – contact between metal 1 to polysilicon OR metal 1 to diffusion (active region) is called metal contact. This is shown in fig. 3

- A 2λ × 2λ cut centered on 4λ × 4λ superimposed area is used to connect layers
- In case of multiple contacts the distance between adjacent contacts should be 2λ

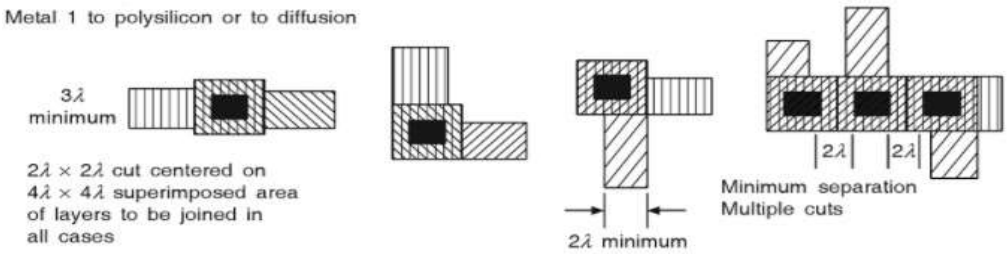


Fig. 5. Contacts (nMOS and CMOS)

Via contact – the contact between metal 1 and metal 2 is called via contact as shown in fig. 4.

- A 2λ × 2λ cut centered on 4λ × 4λ superimposed area is used to connect layers

✦ To connect metal 2 with diffusion via and cut both are used

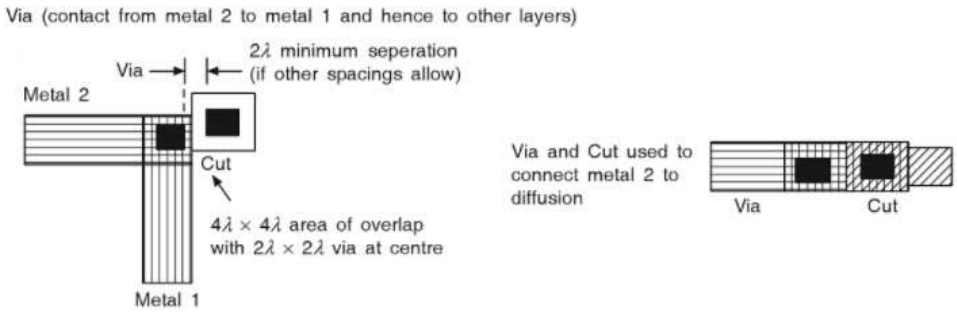


Fig.6 contacts

2.7 Contact Cuts:

- Electrical connection between layers can be done using special structures 'contact cuts'.
- There are 3 approaches for contacts between polysilicon and diffusion in nMOS circuits. They are
    1. Polysilicon to metal and then to diffusion
    2. Buried contact - polysilicon to diffusion
    3. Butting contact - polysilicon to diffusion using metal

✓ among the three buried contact is most used as it gives economy in space and reliable contact.

➢ Buried contact is distinguished feature in nMOS for connection between poly and diffusion and this is most widely used than butting contact.

Buried Contact (nMOS):

✦ Layer is joined over the area of 2λ × 2λ with buried contact cut extending by 1λ in all directions except in the diffusion path. It extends by 2λ in order to avoid formation of unwanted transistors.

✦ The contact cut shown in broken line indicates the region where thinox is removed on the silicon wafer and polysilicon gets deposited on wafer.

✦ When impurities are added, it diffuse into poly and also to diffusion region within the contact area. This provides satisfactory contact between ploy and diffusion as shown in fig 5.

➢ In CMOS poly to diffusion connections are made through metal. The process of making connection between metal and either of 2 layers (poly or diffusion) is by buried contact.

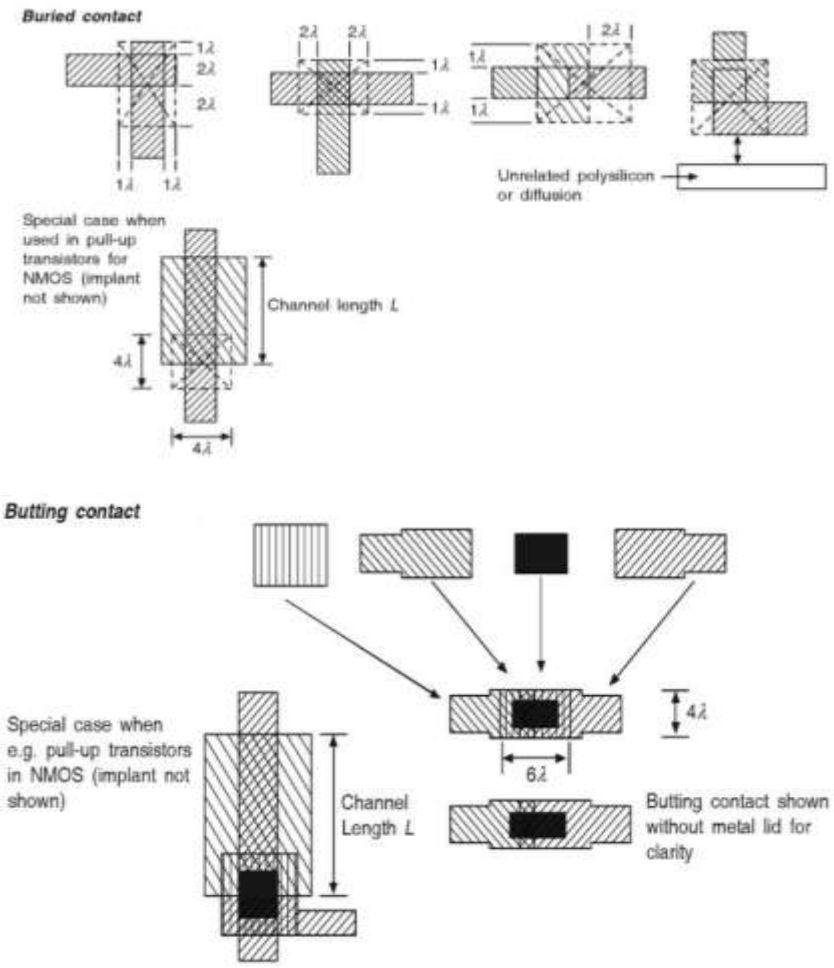


Fig. 7 Buried and butting contacts only for nMOS

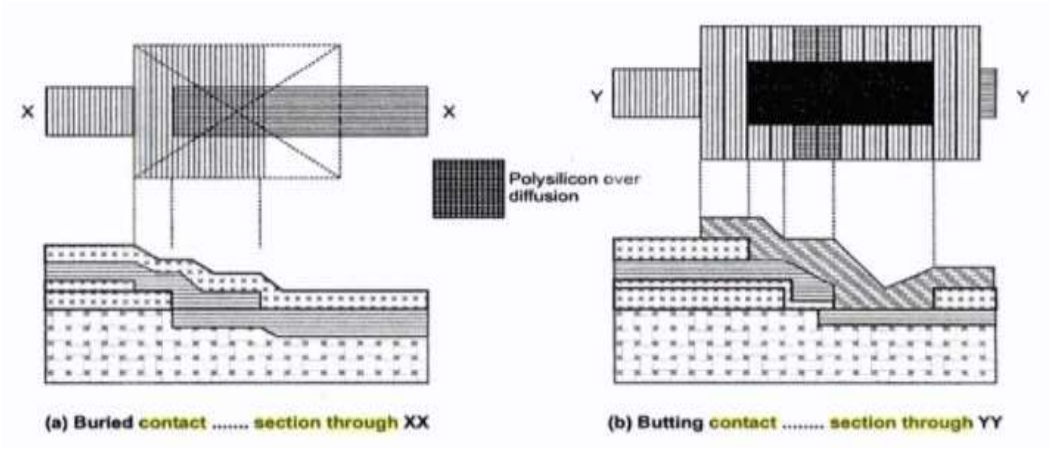


Fig 8. Cross section through contact structures

Butting contact

- Butting contact process is complicated and done when two layers do not overlap. Contact cut of $2\lambda \times 2\lambda$ is made until each of layers is joined. The layers are held in such a way that these two contacts become continuous.

- ✦ The poly and diffusion outlines overlap and thinox under ploy acts as mask during diffusion process. Finally contact between two butting layers is done by a metal. This can be seen in fig. 5 cross-sectional view.

2.8 Double Metal MOS Process Rules:

- If to process considered till now introduction of second metal layer will boosts the design capabilities. It gives more freedom. Ex. this will be helpful for power rail (Vdd and Vss/Gnd) distribution and also for clock.
- This process is called Double Metal MOS Process
- This technique involves connecting metal 1 and metal 2 contacts called 'via'. This is shown in fig. 4 and fig. 6

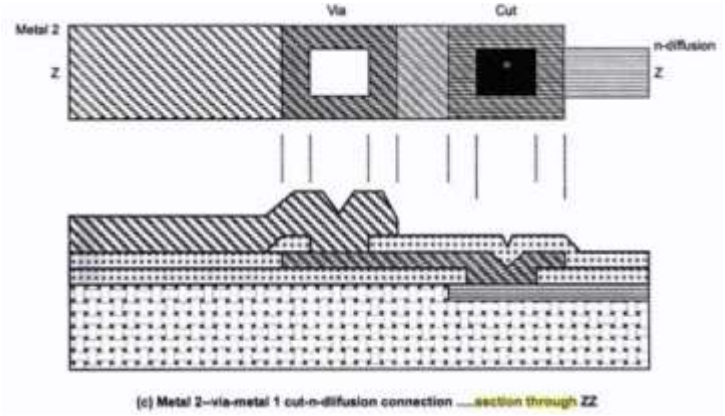


Fig. 9 cross section of via contact structure

- The 2nd metal layer is coarser than 1st metal layer (conventional) and the isolation layer between the 2 is usually thicker than normal.
- To distinguish contacts between 1st and 2nd metal layer they are called as 'vias' rather than contact cut
- In stick diagram representation its color code is dark blue or purple.

The steps of fabrication process is as follows:

1. Using chemical vapor deposition oxide layer under 1st metal layer is deposited.
2. using same method oxide between 2 metal layers are formed.
3. Selected areas of oxide are removed by using plasma etching. The etching process is done under high vertical ion bombardment to get high and uniform etching.

The layout strategy used with double metal process is summarized as below

1. Second metal layer is usually used for global power railings and clock lines
2. First metal layer is used for local power distribution and signals.
3. The layout of the two metals are such that are mutually orthogonal wherever possible

- ✦ Similar to double metal process, other process allows second poly layer. The process steps are similar to previously described process.

- The first polysilicon (poly 1) layer is deposited and patterned on this a second thinox (thin oxide) layer is grown. On this the second polysilicon (poly 2) layer is deposited and patterned. Thus 2nd thinox isolated the poly layers.
- Presence of poly 2 provides greater flexibility in interconnections and allows transistors to be formed by intersection of poly 2 and diffusion.

2.9 CMOS Lambda-based Design Rules:

- Comparing to Nmos fabrication process, CMOS fabrication is more complex.
- Extending the Nmos design rules, Noting exclusion of butting contact and buried contact rules.
- Additional rules associated with CMOS process concerned with unique feature p-well CMOS, i.e: p-well and P+ Mask and Substrate contact.

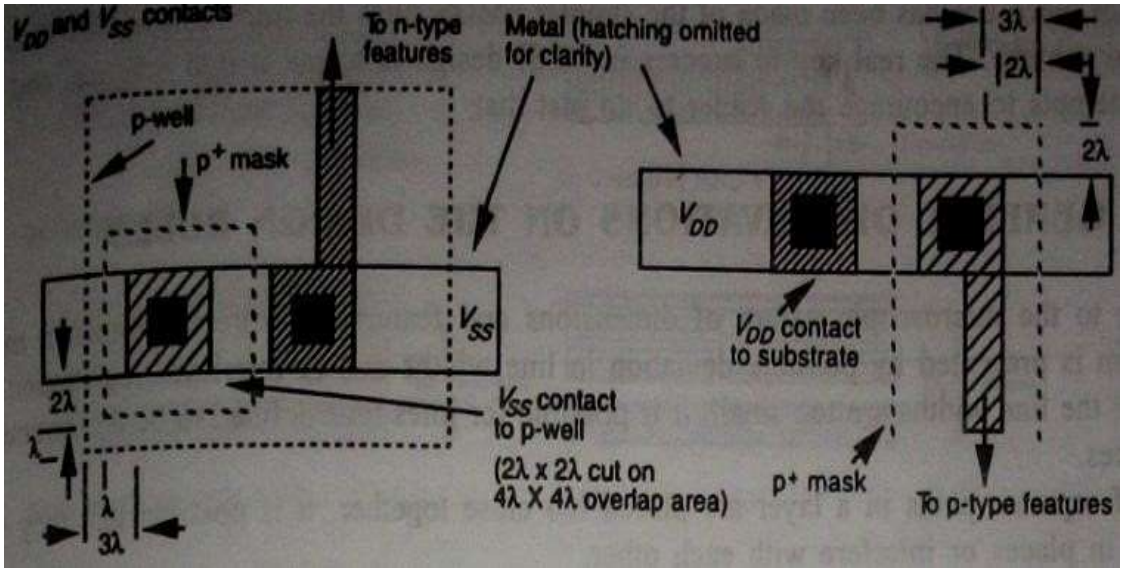
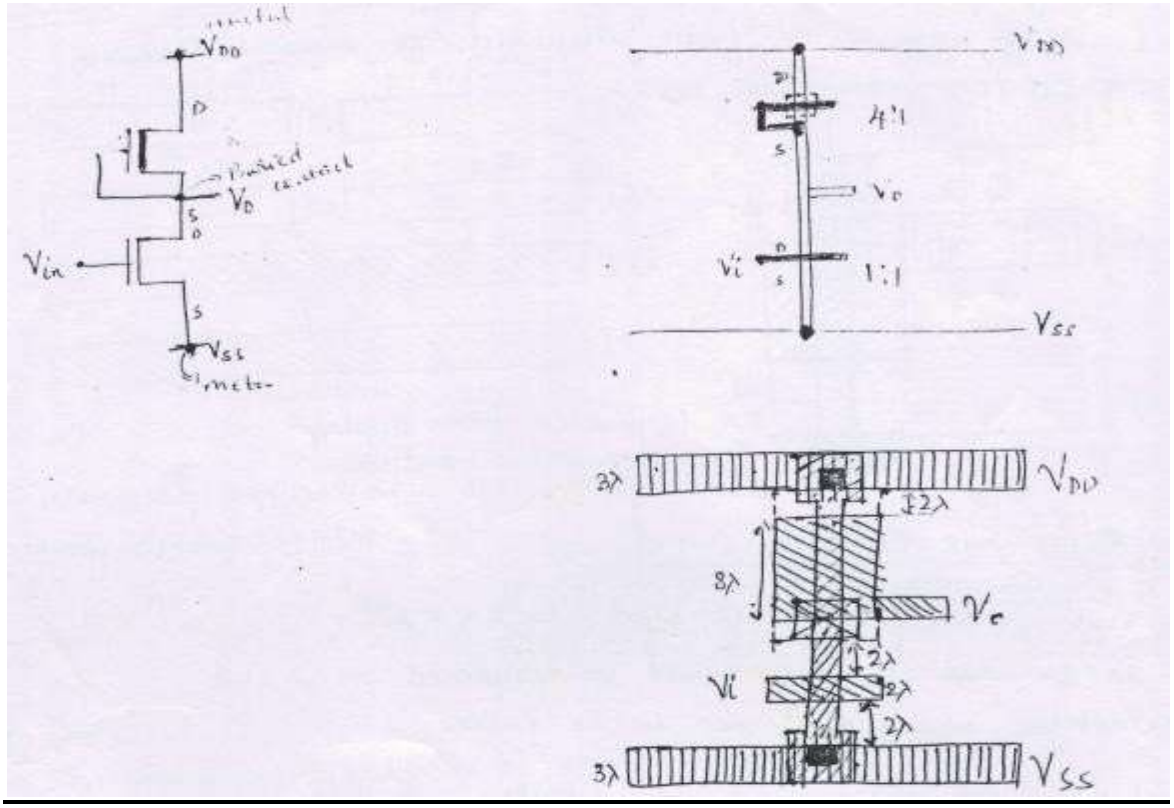
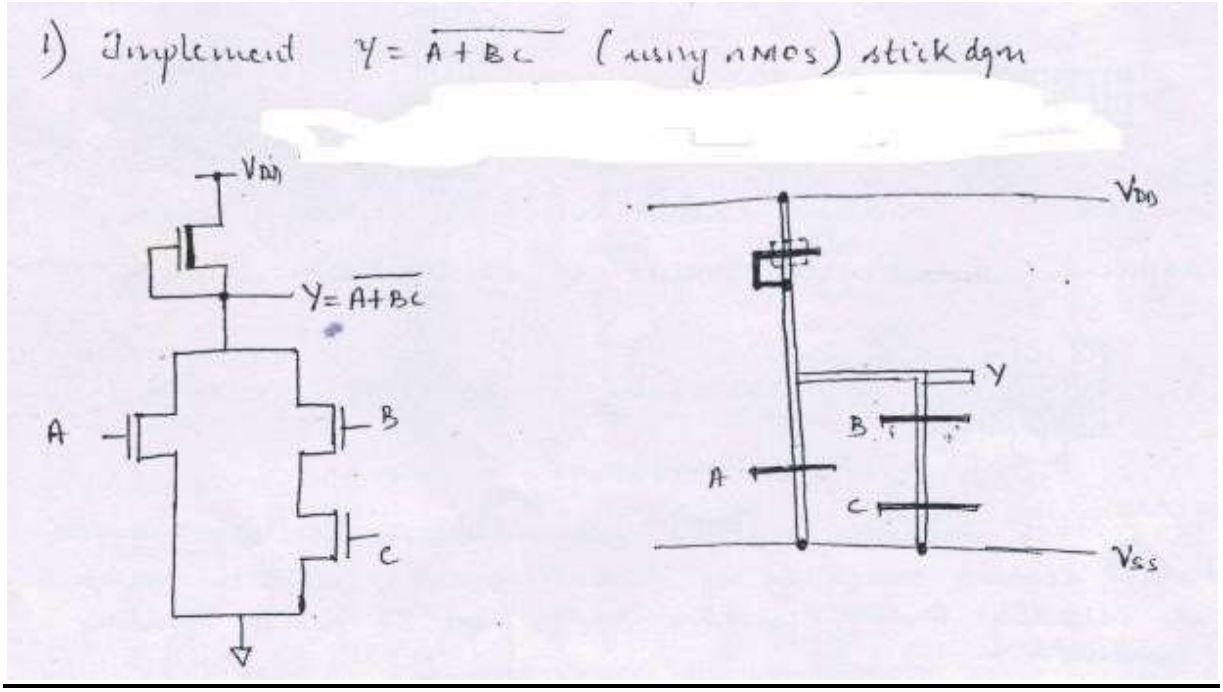


Fig.10 CMOS Lambda-based Design Rules

Problems on stick diagram and layouts.

Nmos Inverter





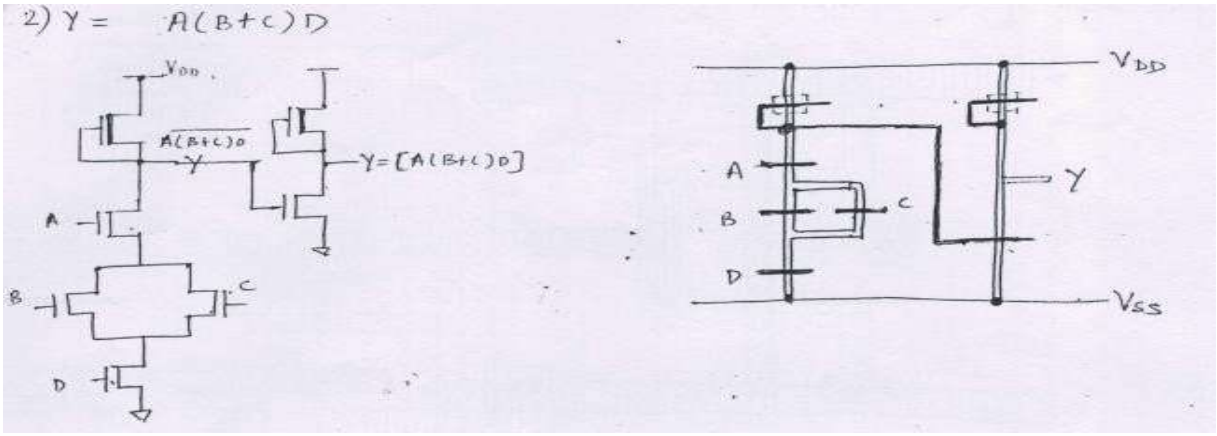


Fig.11 Nmos inverter

CMOS Inverter

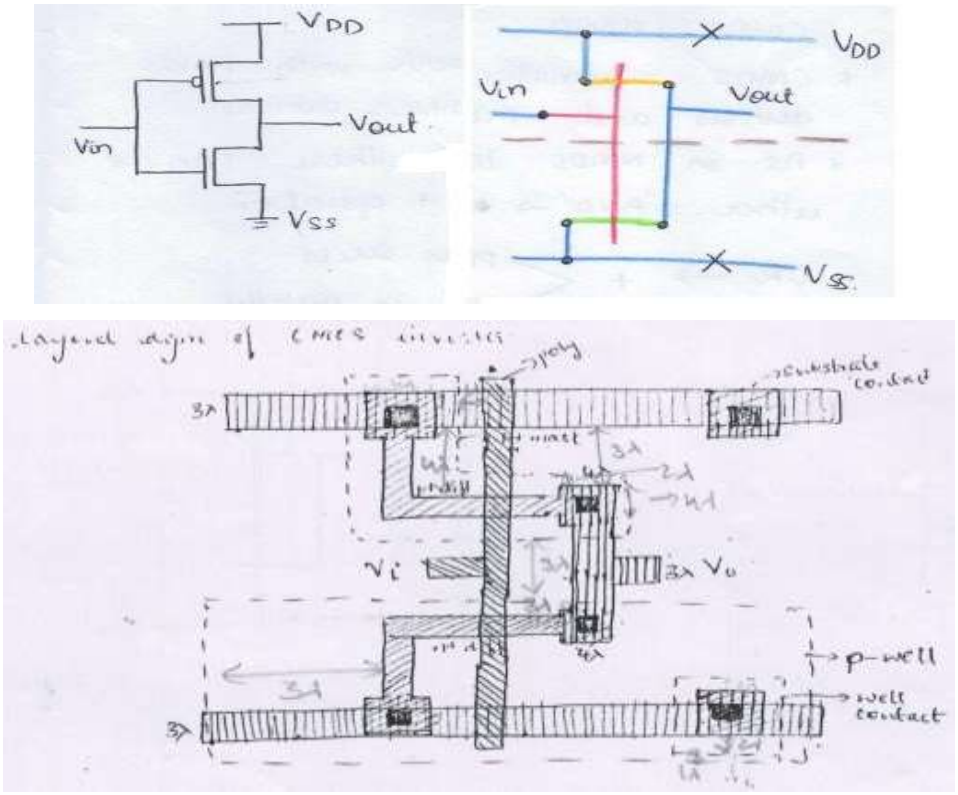
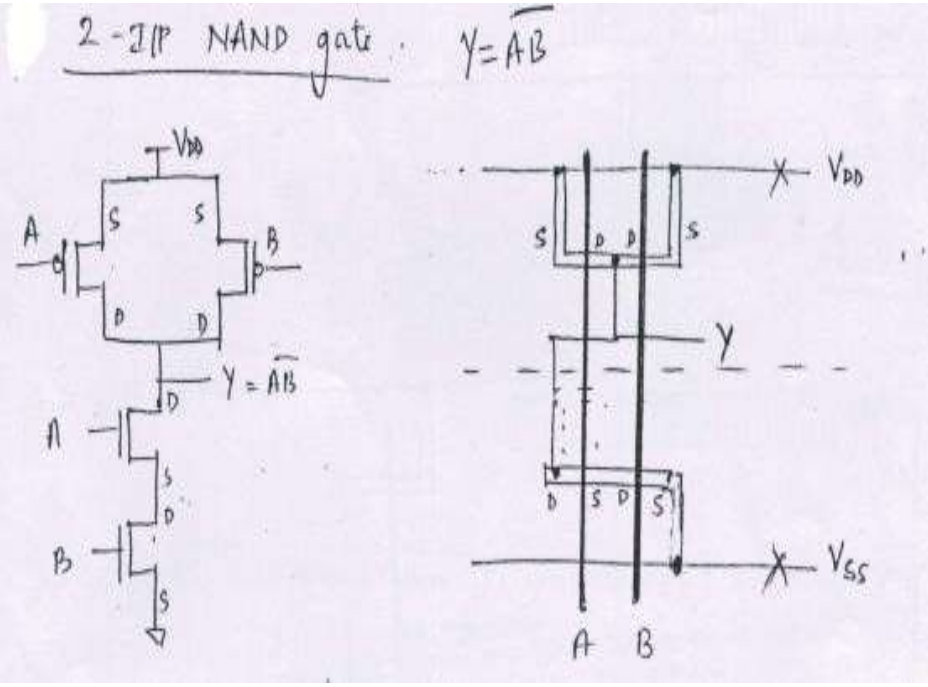


Fig.12 CMOS inverter



Diffusion lines running horizontal and polysilicon lines in vertical direction.

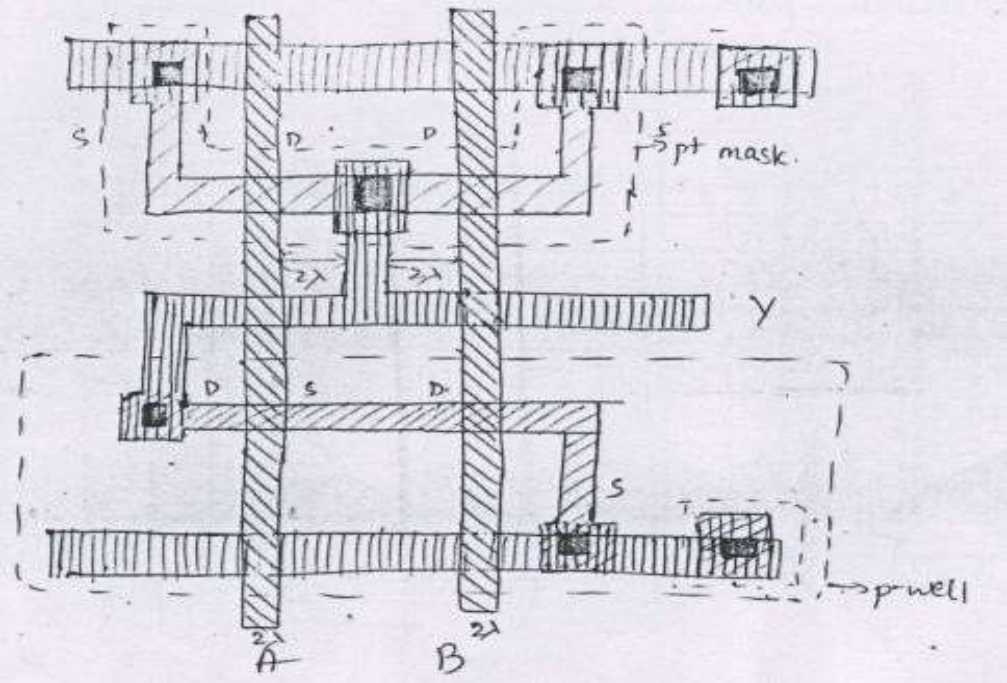


Fig.13 CMOS 2 input NAND gate

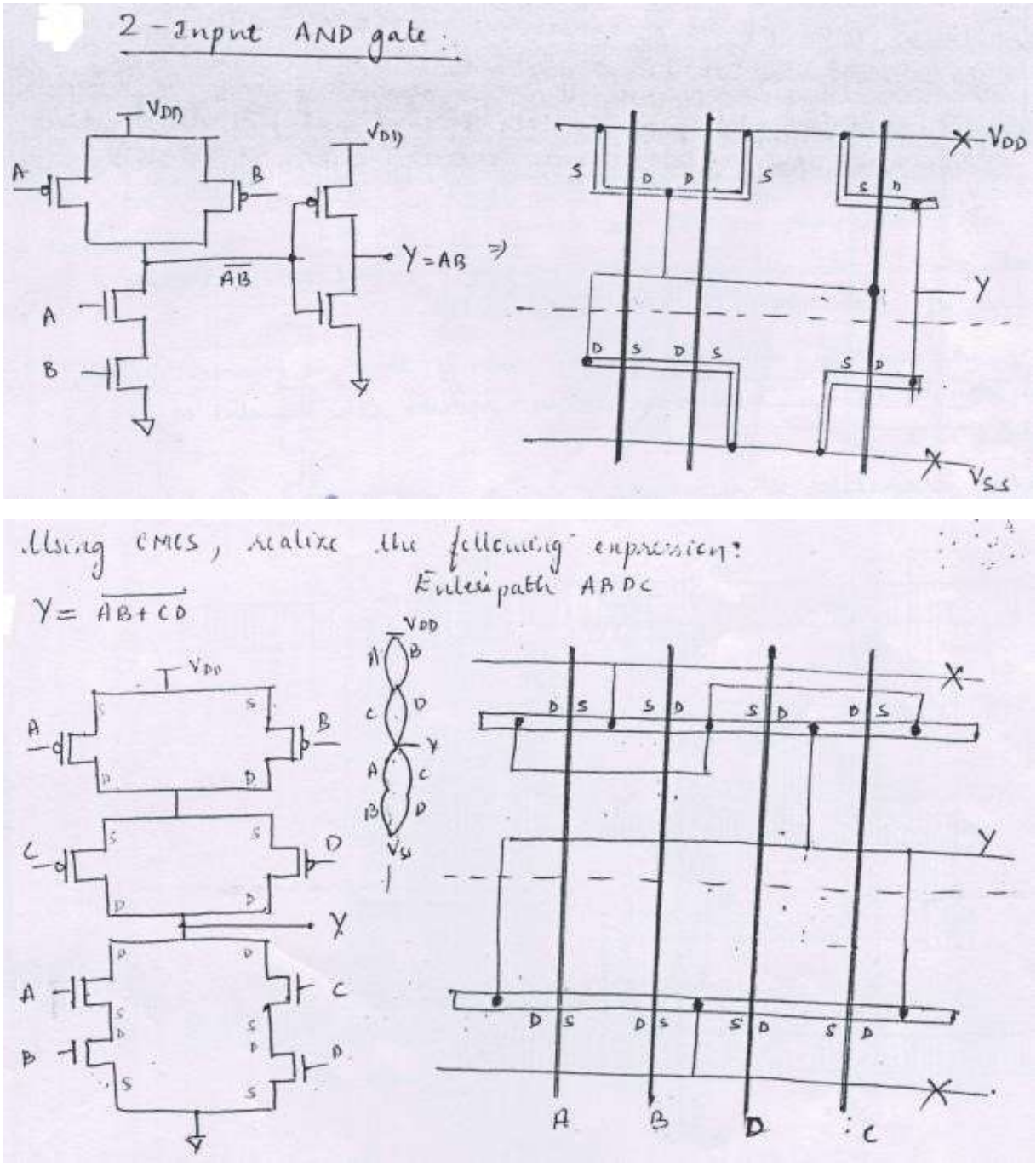
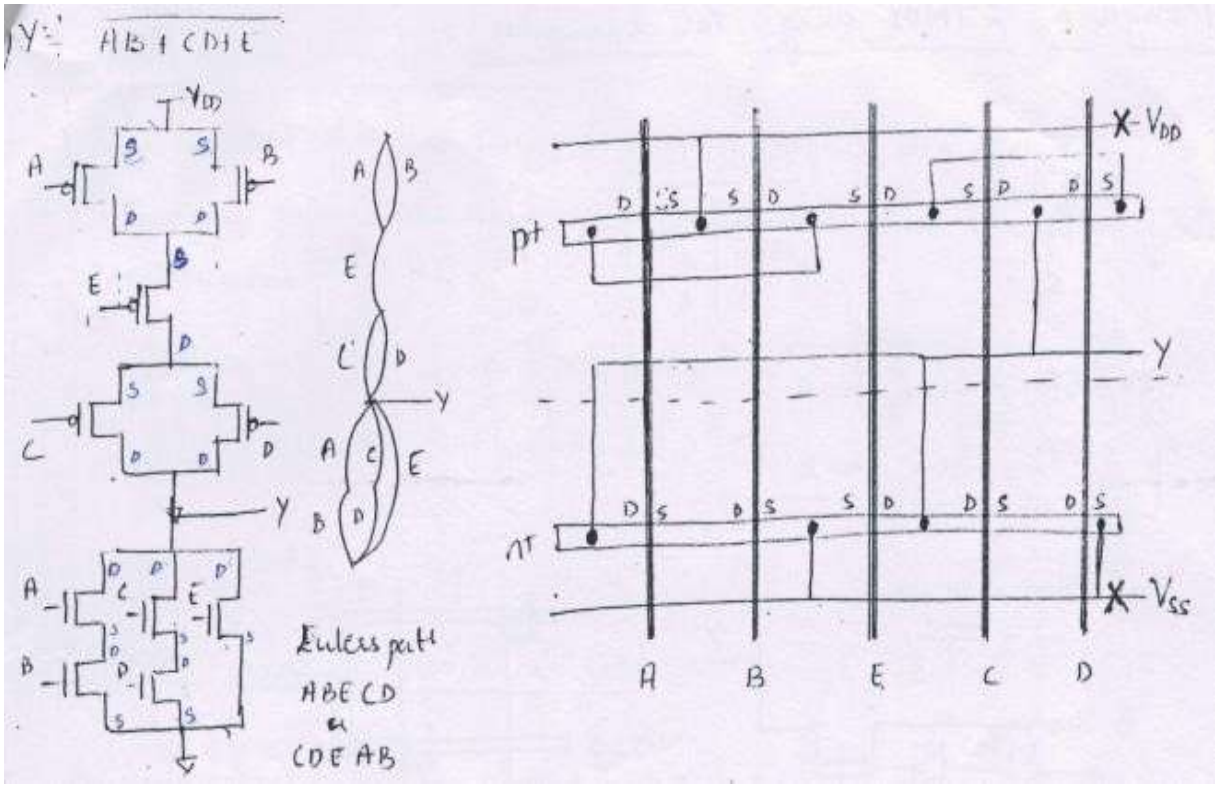


Fig.14 CMOS 2 input AND gate

- When more number of inputs available, Euler's path is determined to know gate ordering.
- Advantage of using Euler's path is to that a common diffusion line can be used which reduces number of contact cuts.
- Uninterrupted path in both pull-up and pull down network represents optimized gate ordering which helps in drawing layout without breaking the diffusion layer.


Y = $\overline{AB + CD + E}$
V_{DD}
A
B
E
C
D
Y
pt
nt
V_{SS}
Euler path
ABECD
or
CDEAB

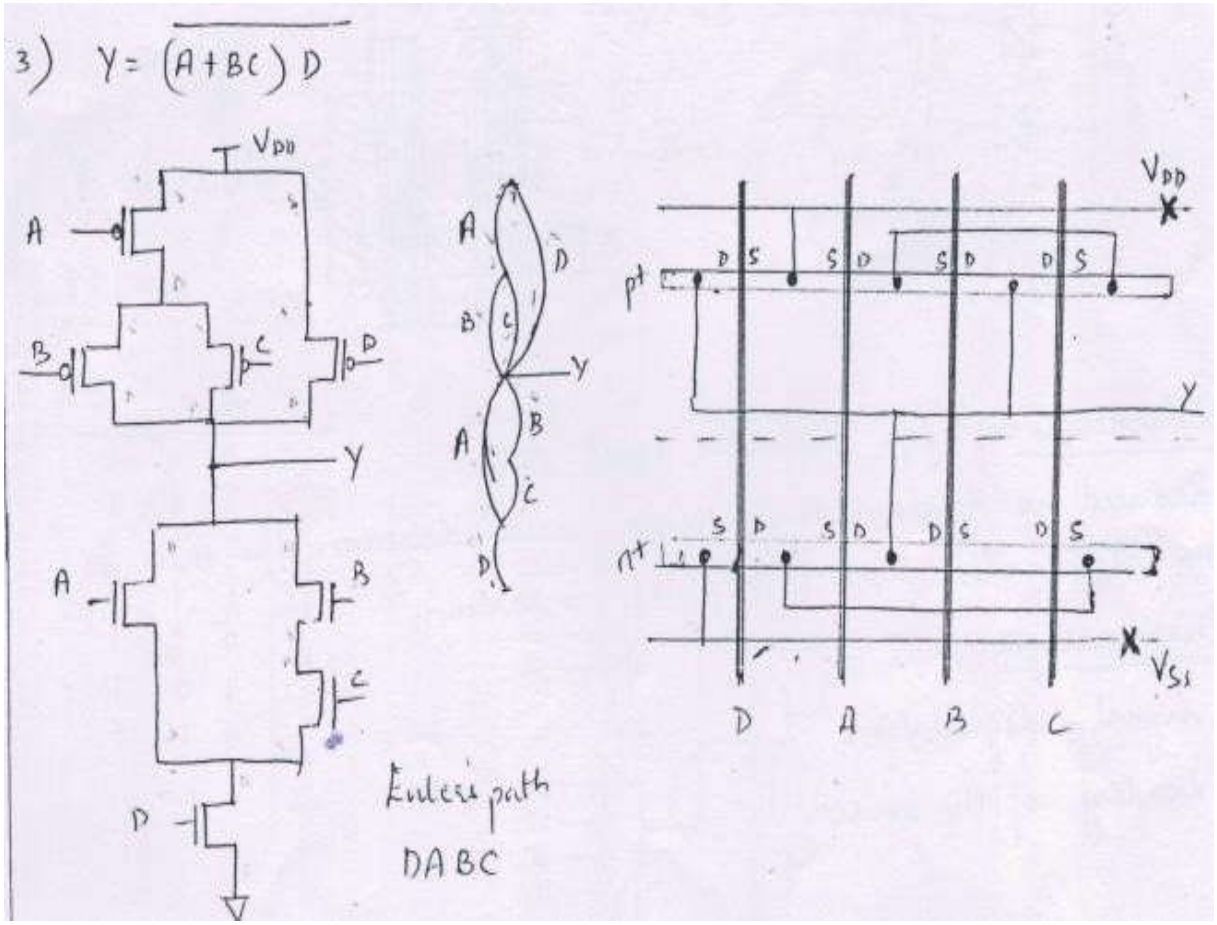
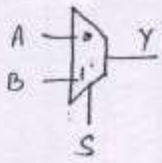

3) Y = $\overline{(A+BC)D}$
V_{DD}
A
B
C
D
Y
pt
nt
V_{SS}
Euler path
DABC

Fig.15 CMOS design

Design a 2:1MUX using Pass transistors.



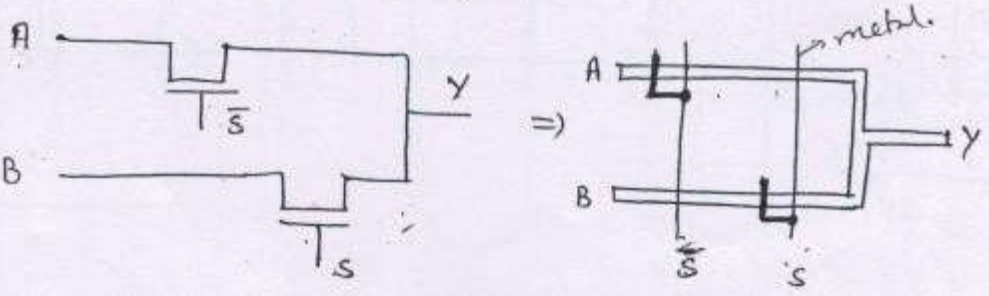
Truth table

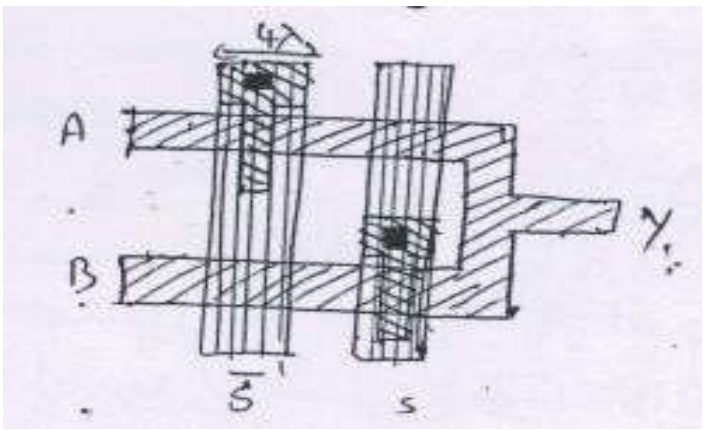
S	Y
0	A
1	B

$Y = \bar{S}A + SB$

if S = 0, then Y = A

if S = 1, then Y = B



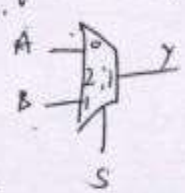


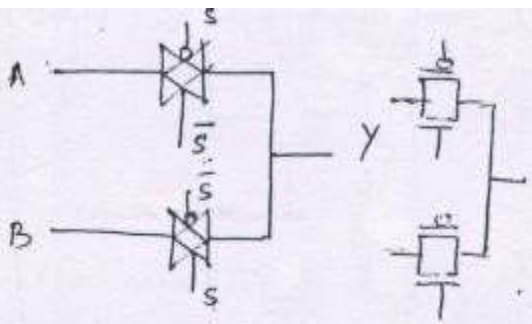
2:1 MUX using transmission gates

S	Y
0	A
1	B

if S = 0, Y = A

if S = 1, Y = B.





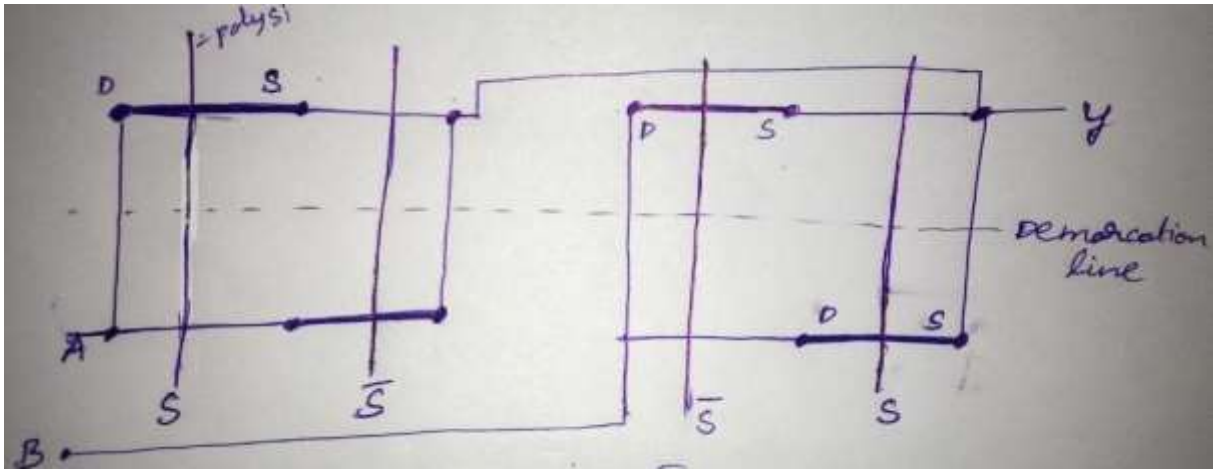


Fig.16 stick diagram of 2:1 MUX using transmission gates.

Two input XOR gate realization using transmission gates.

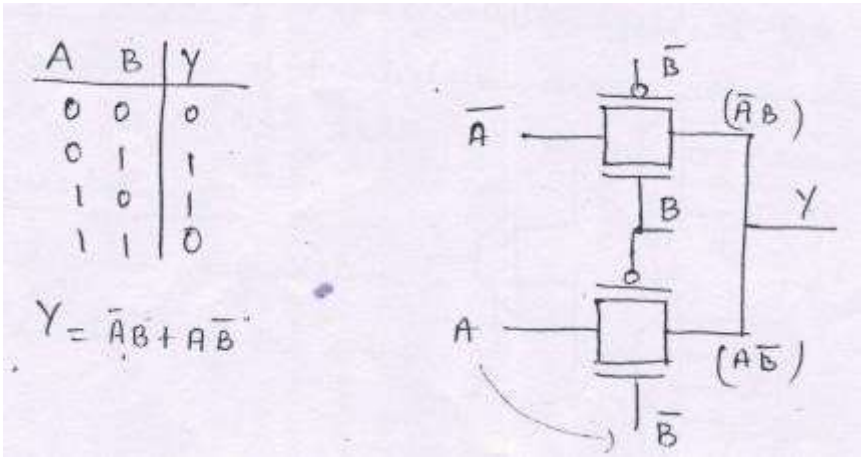


Fig.17 Two input XOR gate realization using transmission gates.

Two way selector with enable

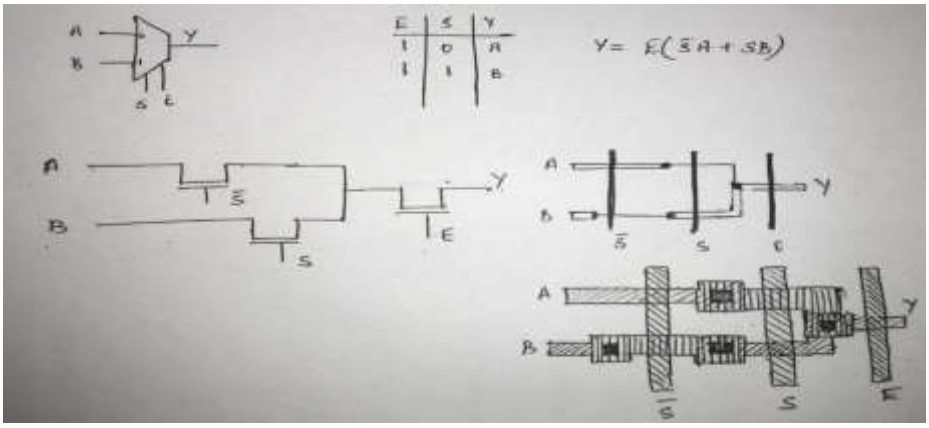


Fig 18. Two way selector with enable

Basic circuit concepts

- In MOS technology, Active devices are dealt with some measurement.
- Wiring up of circuits is done through various conductive layers which is produced by MOS Processing.
- Therefore it is necessary to be aware of resistive and capacitive characteristics of each layers.
- For evaluating the effects of wiring, input and output capacitances, sheet resistance and standard unit capacitances are used.
- Further delay associated with wiring, inverters are evaluated by the term delay unit τ.

Table 1: Typical sheet resistance of MOS layers

Typical sheet resistances R_s of MOS layers for 5 μm*, and Orbit 2 μm* and 1.2 μm* technologies

Layer	*R_s ohm per square*		
	5 μm	*Orbit*	*Orbit 1.2 μm*
Metal	0.03	0.04	0.04
Diffusion (or active)**	10→50	20→45	20→45
Silicide	2→4	—	—
Polysilicon	15→100	15→30	15→30
n-transistor channel	$10^{4\dagger}$	$2 \times 10^{4\dagger}$	$2 \times 10^{4\dagger}$
p-transistor channel	$2.5 \times 10^{4\dagger}$	$4.5 \times 10^{4\dagger}$	$4.5 \times 10^{4\dagger}$

Note: In some processes a silicide layer is used in place of polysilicon.

* 5 micron (μm) technology implies minimum line width (and feature size) of 5 μm and in consequence λ = 2.5 μm. Similarly, 2 μm and 1.2 μm technologies have feature sizes of 2 μm and 1.2 μm respectively.

** The figures given are for n-diffusion regions. The values for p-diffusion are 2.5 times these values.

† These values are approximations only. Resistances may be calculated from a knowledge of V_{ds} and the expressions for I_{ds} given earlier.

2.10 Sheet Resistance R_s

Consider a transistor with a channel having resistivity ϱ, width W, thickness t and length between source and drain is L.

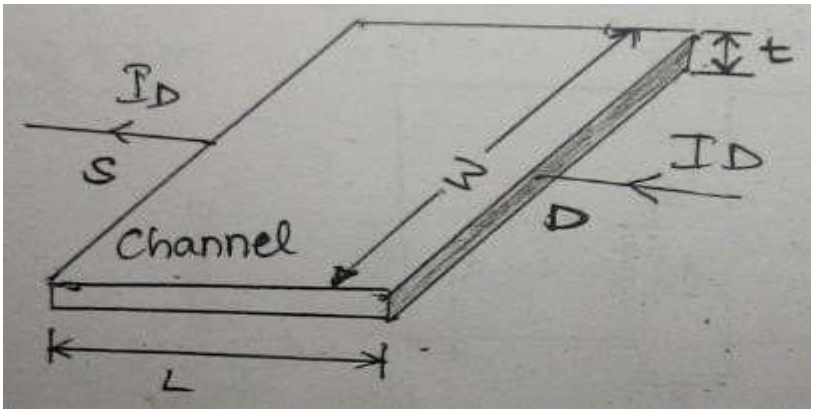


Fig.19 Sheet Resistance

Resistance of the channel between drain and source is expressed as.

$$R_{DS} = \frac{\rho.Length}{Area\ of\ cross\ section} = \frac{\rho.L}{t.W}$$

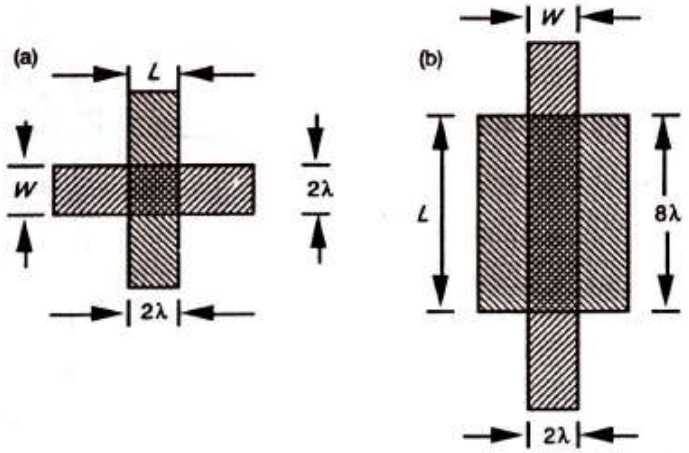
$$R_{DS} = R_s \frac{L}{W}$$

Where $R_s = \frac{\rho}{t}$ is a constant and it is called sheet resistance.

From the above equation, sheet resistance can be defined as resistance of the channel whose length and width are equal.

R_s is completely independent of area square. Ex: 1µm per side square slab of material has exactly same resistance as 1cm per side square slab of same material if thickness is same.

Example



Resistance calculation for transistor channels.

$$R = 1 \text{ square} \times R_s \frac{\text{ohm}}{\text{square}} = R_s = 10^4 \text{ ohm}^*$$

The length to width ratio, denoted Z, is 1:1 in this case. The transistor structure of Figure has a channel length $L = 8\lambda$ and width $W = 2\lambda$. Therefore,

$$Z = \frac{L}{W} = 4$$

Thus, channel resistance

$$R = ZR_s = 4 \times 10^4 \text{ ohm}$$

2.11 Area capacitance

In between gate and channel exists a capacitance and it is called gate capacitance and denoted by C_g.

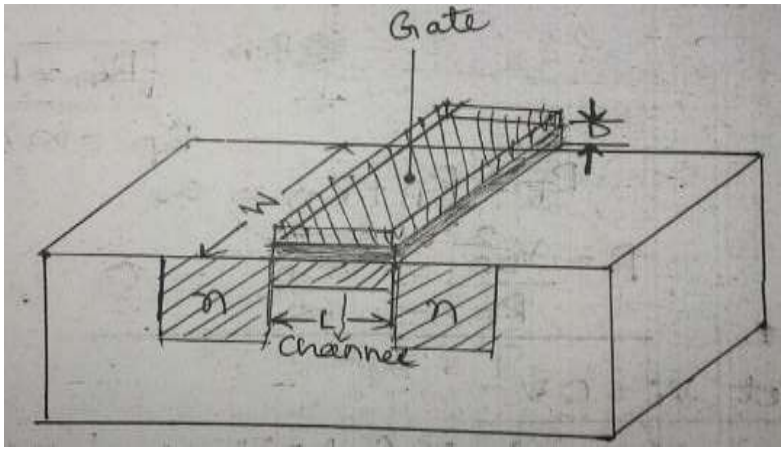


Fig 20. Area capacitance

From the above diagram.

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

A is area of the channel or surface area of the gate

$$\frac{C_g}{A} = \frac{\epsilon_0 \epsilon_r}{D} \, pF/(\mu m)^2$$

$$C_A = \frac{\epsilon_0 \epsilon_r}{D}$$

ϵ_0 = permittivity of free space = 8.854x 10^{-12} F/m.

ϵ_r= relative permittivity of a given material

D = thickness of sio2 constant for a given technology.

Area capacitance is defined as capacitance per unit area at the gate of transistor and denoted by C_A.

Table 2: Typical area capacitance for MOS layers

Typical area capacitance values for MOS circuits

Capacitance	*Value in pF × $10^{-4}/\mu m^2$ (Relative values in brackets)*		
	5 μm	*2 μm*	*1.2 μm*
Gate to channel	4 (1.0)	8 (1.0)	16 (1.0)
Diffusion (active)	1 (0.25)	1.75 (0.22)	3.75 (0.23)
Polysilicon* to substrate	0.4 (0.1)	0.6 (0.075)	0.6 (0.038)
Metal 1 to substrate	0.3 (0.075)	0.33 (0.04)	0.33 (0.02)
Metal 2 to substrate	0.2 (0.05)	0.17 (0.02)	0.17 (0.01)
Metal 2 to metal 1	0.4 (0.1)	0.5 (0.06)	0.5 (0.03)
Metal 2 to polysilicon	0.3 (0.075)	0.3 (0.038)	0.3 (0.018)

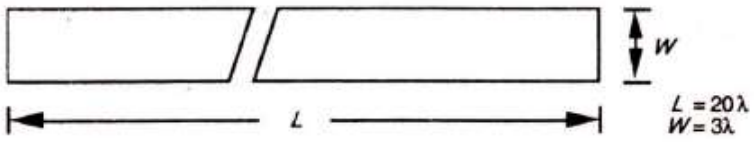
Notes: Relative value = specified value/gate to channel value for that technology.
*Poly. 1 and Poly. 2 are similar (also silicides where used).

SOME AREA CAPACITANCE CALCULATIONS

The approach will be demonstrated using λ-based geometry. The calculation of capacitance values may now be undertaken by establishing the ratio between the area of interest and the area of standard (feature size square) gate (2λ × 2λ for λ-based rules) and multiplying this ratio by the appropriate relative *C* value from Table The product will give the required capacitance in $\square C_g$ units.

Consider the area defined in Figure 4.4. First, we must calculate the area relative to that of a standard gate.

$$\text{Relative area} = \frac{20\lambda \times 3\lambda}{2\lambda \times 2\lambda} = 15$$



Simple area for capacitance calculation.

Now:

1. Consider the area in metal 1.
   Capacitance to substrate = relative area × relative *C* value
   $= 15 \times 0.0750 \square C_g$
   $= 1.125 \square C_g$

   That is, the defined area in metal has a capacitance to substrate 1.125 times that of a feature size square gate area.

2. Consider the same area in polysilicon.

   Capacitance to substrate = $15 \times 0.1 \square C_g$

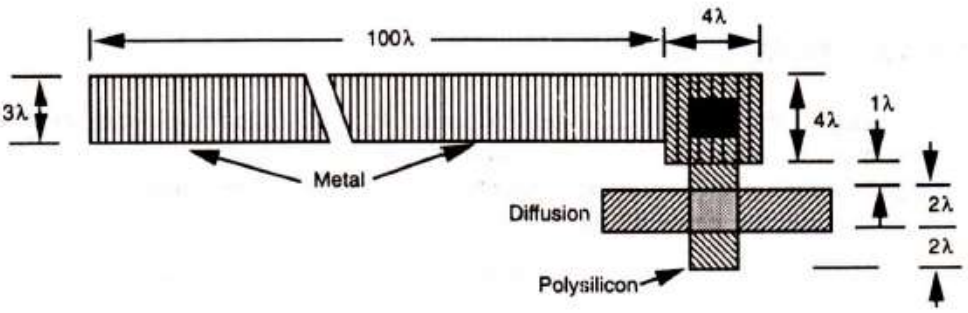
   $= 1.5 \square C_g$

3. Consider the same area in n-type diffusion.

   Capacitance to substrate = $15 \times 0.25 \square C_g$

   $= 3.75 \square C_g^*$

Calculations of area capacitance values associated with structures occupying more than one layer, as in Figure are equally straightforward.



Capacitance calculation (multilayer).

Consider the metal area (less the contact region where the metal is connected to polysilicon and shielded from the substrate)

$$\text{Ratio} = \frac{\text{Metal area}}{\text{Standard gate area}} = \frac{100\lambda \times 3\lambda}{4\lambda^2} = 75$$

$$\text{Metal capacitance } C_m = 75 \times 0.075 = 5.625 \square C_g$$

Consider the polysilicon area (excluding the gate region)

$$\text{Polysilicon area} = 4\lambda \times 4\lambda + 3\lambda \times 2\lambda = 22\lambda^2$$

Therefore

$$\text{Polysilicon capacitance } C_p = \frac{22}{4} \times 0.1 = .55 \square C_g$$

For the transistor,

$$\text{Gate capacitance } C_g = 1 \square C_g$$

2.12 Standard unit of capacitance ($\square C_g$) and calculation

The standard unit of capacitance is defined as the capacitance at the gate of 1:1 transistor.

Ex: consider a 1:1 transistor where L = 2λ and W = 2λ.

Gate area of transistor = L x W

$$A = 2\lambda \times 2\lambda = (2\lambda)^2$$

Actual capacitance at the gate of transistor $C = C_A . A = C_A . (2\lambda)^2$

$$C = 4x10^{-4}\ PF/(\mu m)^2 . (2\lambda)^2$$

Consider 5 μm technology, i.e: $2\lambda = 5\ \mu m$

$C = 4x10^{-4}\ pF/(\mu m)^2 . (5\ \mu m)^2 = 4x10^{-4} \times 25 = 0.01$ pF $C = 1\square C_g$

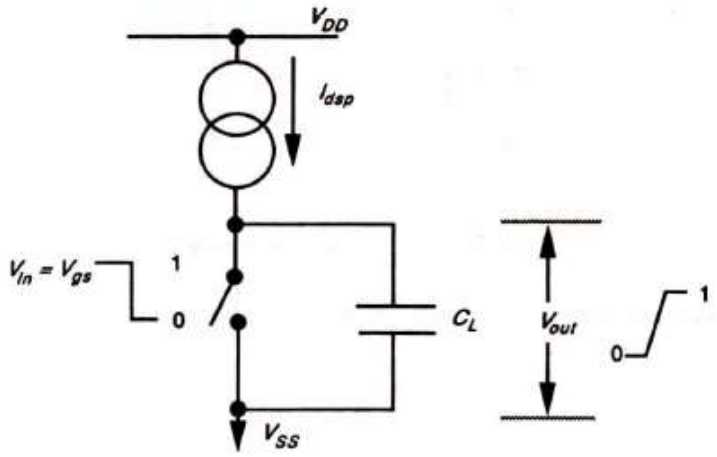
2.13 RISE TIME AND FALL TIME ESTIMATION

Rise-time estimation

In this analysis we assume that the p-device stays in saturation for the entire charging period of the load capacitor C_L. The circuit may then be modeled as in Figure 4.9.

The saturation current for the p-transistor is given by

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



Rise-time model.

This current charges C_L and, since its magnitude is approximately constant, we have

$$V_{out} = \frac{I_{dsp} t}{C_L}$$

Substituting for I_{dsp} and rearranging we

$$t = \frac{2 C_L V_{out}}{\beta_p (V_{gs} - |V_{tp}|)^2}$$

We now assume that $t = \tau_r$ when $V_{out} = +V_{DD}$, so that

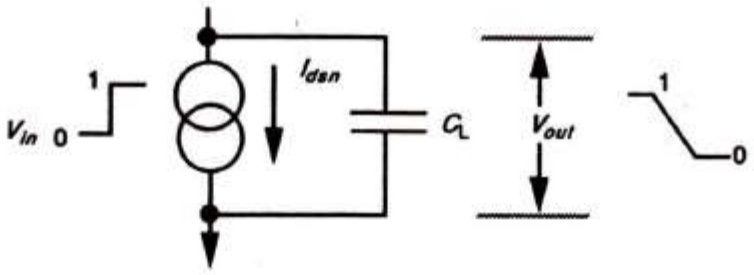
$$\tau_r = \frac{2V_{DD}C_L}{\beta_p(V_{DD} - |V_{tp}|)^2}$$

with $|V_{tp}| = 0.2V_{DD}$, then

$$\tau_r \doteq \frac{3C_L}{\beta_p V_{DD}}$$

Fall-time estimation

Similar reasoning can be applied to the discharge of C_L through the n-transistor.



Fall-time model.

Making similar assumptions we may write for fall-time:

$$\tau_f \doteq \frac{3C_L}{\beta_n V_{DD}}$$

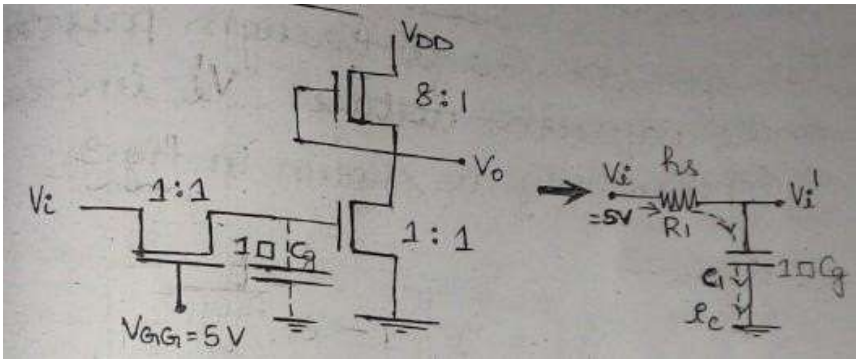
2.14 Standard Delay Unit (τ)

Time delay is measured in terms of standard unit τ. It is

defined as product of and C_g. i.e: $\tau = R_s . \square C_g$

Measurement of τ.

Consider nmos driven by pass transistor shown in below figure and the dimensions are indicated. Pass transistor is ON for given gate voltage V_{GG}. Pass transistor is represented by R_s due to its equal length and width. Pull down transistor of inverter is represented by capacitance $\square C_g$. Since pull down has minimum dimensions.



τ is defined as time taken by capacitor to charge from 0 to 63.2% of maximum value as shown in below figure.

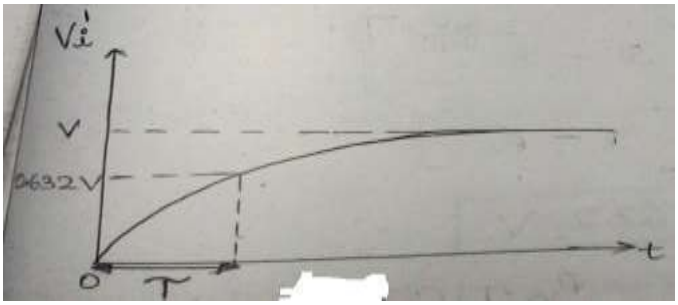


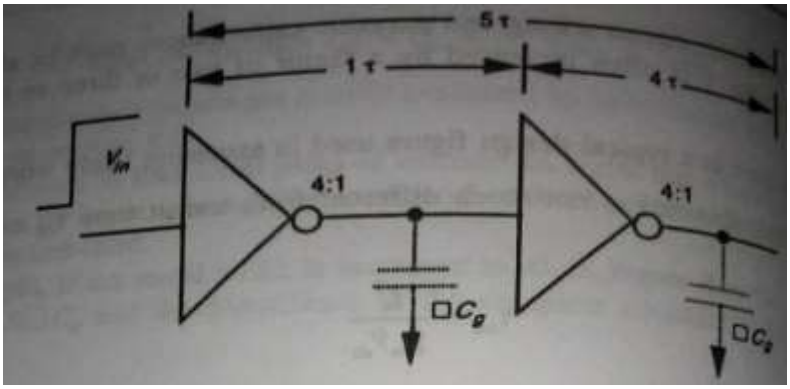
Fig.21 Measurement of τ.

2.15 Inverter Delays

Consider basic 4:1 nmos inverter. To achieve 4:1 Z_{pu} to Z_{pd} ratio, R_{pu} will be $4R_{pd}$. Clearly resistance R_{pu} value is $R_{pu} = 4\ R_s = 40\text{K}\Omega$. Meanwhile R_{pd} value is $10\text{K}\Omega$.

Delay associated with inverter depends on ON and OFF condition of transistors.

Consider a pair of cascaded inverter, delay in this pair will be constant irrespective of sense of logic level transition. The overall delay of nmos inverter is τ + 4τ = 5τ. Shown in below figure.



In general term delay through nmos inverter pair is given by $T_d = (1 + Z_{pu}/Z_{pd})\ \tau$

So single 4:1 inverter exhibits asymmetric delays, delay in turning on τ (capacitor discharging condition) and delay in turning off is 4 τ (capacitor charging condition). Asymmetry becomes worse for inverter with 8:1 ratio.

For CMOS inverter, nmos rules no longer applies, but we need to consider natural asymmetry of equal size pull up and pull down transistors.

Gate capacitance is double compare to nmos inverter since input is connected to both transistors and delay associated with pair of minimum size inverters is shown in below figure.

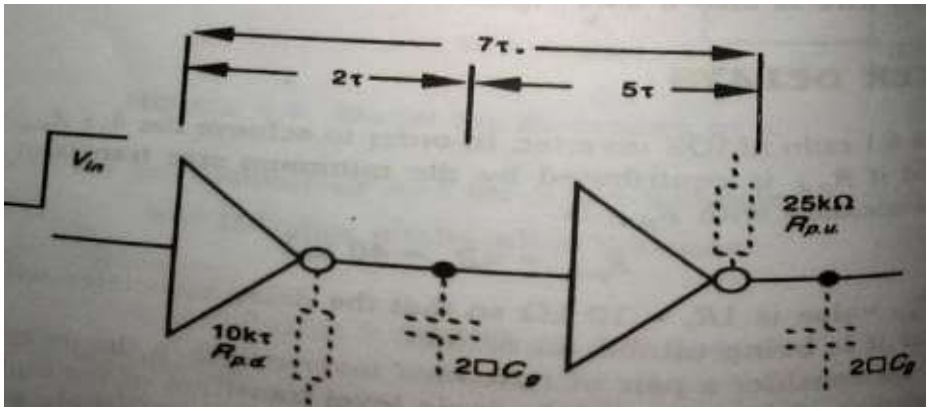


Fig 22. Inverter Delays

Asymmetry of resistance is eliminated by increasing the width of p- device channel by factor of two or three, but gate capacitance increases by the same factor.

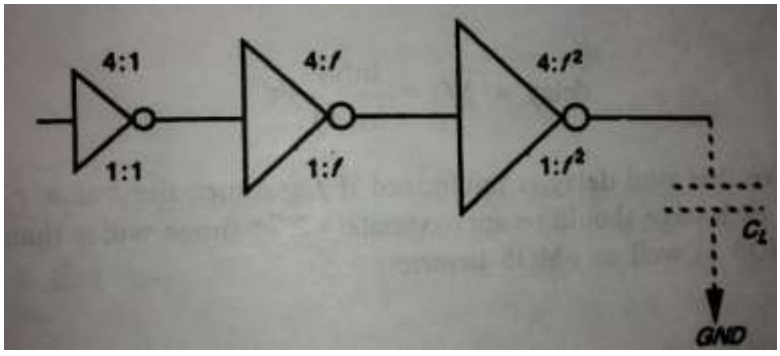
2.16 Driving large capacitive loads

- A large capacitive loads problem arises when a signal to be transmitted from On chip to Off chip destinations.
- Off chip capacitance is is generally higher than On chip $\square C_g$.And it is denoted by .

  $C_L \geq 10^4 \square C_g$
- A capacitance of this order to be driven through low resistance otherwise long delays will occur.

2.17 Cascaded Inverters as drivers

- Inverters to drive large capacitive loads resistance associated with pull up and pull down transistors to be low.
- Low resistance values of Z_{pu} and Z_{pd} implies low L:W ratio or channel width must be made wider to reduce channel resistance but consequently inverter occupies large area.
- Gate area LxW is more significant and large capacitance present at input which slows down rate of change of voltage at input.
- Remedy to use N cascade inverter is by maintaining L to a minimum feature size and width of each successive stage is increased by factor f as shown in below figure.



- With increase in width factor increases capacitive load at input side and area occupied by the inverter also increases.
- The rate of width increase influence on number of stages to be cascaded to drive particular C_L value.
- Total delay associated with nmos pair is 5 τ and cmos pair is 7 τ.

Let y = $C_L/\square C_g$ = , f and N are interdependent.

To determine value of f to minimize overall delay for given y ln(y) = N ln(f)

$$N = \frac{\ln(y)}{\ln(f)}$$

For N even, total delay = $\frac{N}{2} 5 f \tau = 2.5 f \tau$ (nmos) or
$-\frac{N}{2} 7 f \tau = 3.5 f \tau$ (cmos)

In all cases, delay $\dot{\alpha}$ N$^{f \tau = \frac{\ln(y)}{\ln(f)} f \tau}$

- Total delay is minimized If f assumes the value e. i.e: each stage is approximately 2.7 times wider than its predecessor and it is applicable for both cmos and nmos inverters. Thus assuming $f = e$, we have

  Number of stages N= ln(y)

And overall delay t_d

N even: t_d= 2.5 N τ (NMOS) or t_d= 3.5 N τ (CMOS)

N odd: t_d= [2.5 (N-1)+1]e τ (NMOS) or t_d= [3.5 (N-1)+1]e τ (CMOS) For ΔVin which indicates logic 0 to 1 transistion of Vin.

t_d= [2.5 (N-1)+4]e τ (NMOS) or t_d= [3.5 (N-1)+5]e τ (CMOS) For ΔVin which indicates logic 1 to 0 transistion of Vin.

2.18 Super buffers

- Asymmetry of conventional inverter gives rise to significant delay problems when used to drive large capacitive loads.

Common approach used in nmos inverter is to use super buffers an inverting type nmos super buffer is shown in figure.

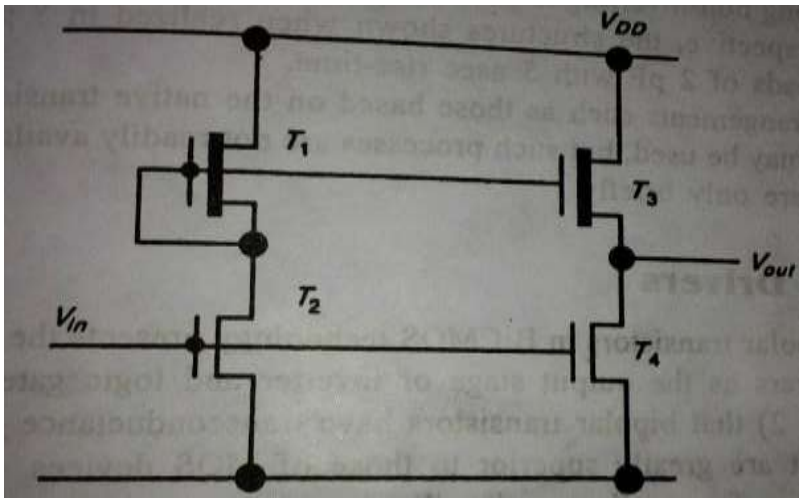


Fig 23. nmos super buffer

- Consider input Vin = 1, the inverter formed by T1 and T2 is turned On and thus gate of T3 is pulled down to zero volts with small delay. So T3 is in cut off and T4 is turned On and output is pulled down.

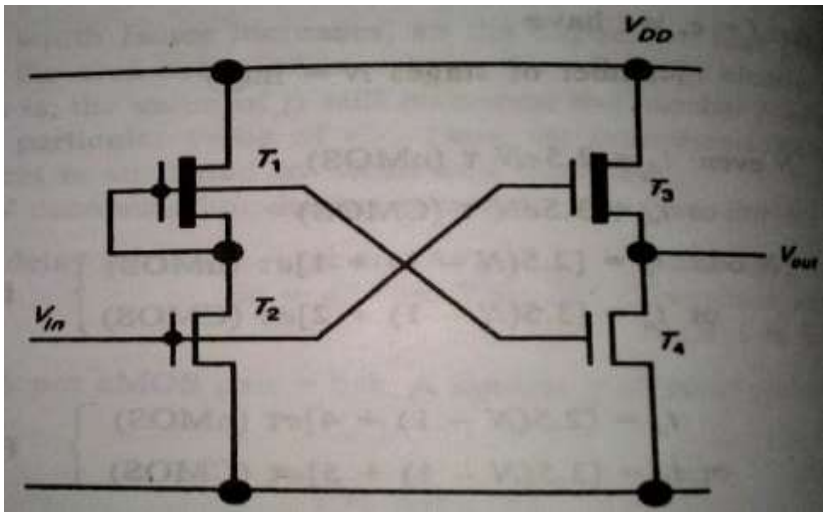


Fig 24 Super buffers

- When Vin = 0, gate of T3 is allowed to rise to Vdd. Thus T4 turned Off, T3 is made to conduct with Vdd on its gate. The voltage applied to gate is twice the average voltage of conventional nmos inverter.
- Doubling effective Vgs will increase current, thus reduces the delay in charging capacitor at output, so symmetry is achieved.
- The Non-inverting type nmos inverter is shown in below figure.

2.19 BICMOS Inverter

- Bipolar transistor availability in Bicmos technology presents possibility of using bipolar transistors as drivers at the output stage of the inverter.
- Transconductance and current/area characteristics are superior than MOS devices. so it has high current driving capability.
- Bipolar transistor has exponential dependence of output current on base emitter voltage which means transistor can operate with small input voltage swing compared to MOS transistors and switches large current.
- So the bipolar transistors have better switching performance results in small input voltage swing and switch large current.
- Switching performance of transistor driving capacitive load can be seen from simple model.
- The time required to change output voltage Vout by an amount equal to input voltage is given by

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

Where is trans conductance of bipolar transistor. As increases Δt decreases.

- Bipolar transistor delay has 2 main components Tin and T_L.
- Tin is the initial time required to charge the B-E junction of the transistor. It is time taken to charge the input gate capacitance.
- T_L is time taken to charge the output load capacitance C_L. This value is less for bipolar by factor of h_{fe}.
- As BJT has higher Tin, T_L is small and because of this faster charging takes place and helps in reducing the delay.

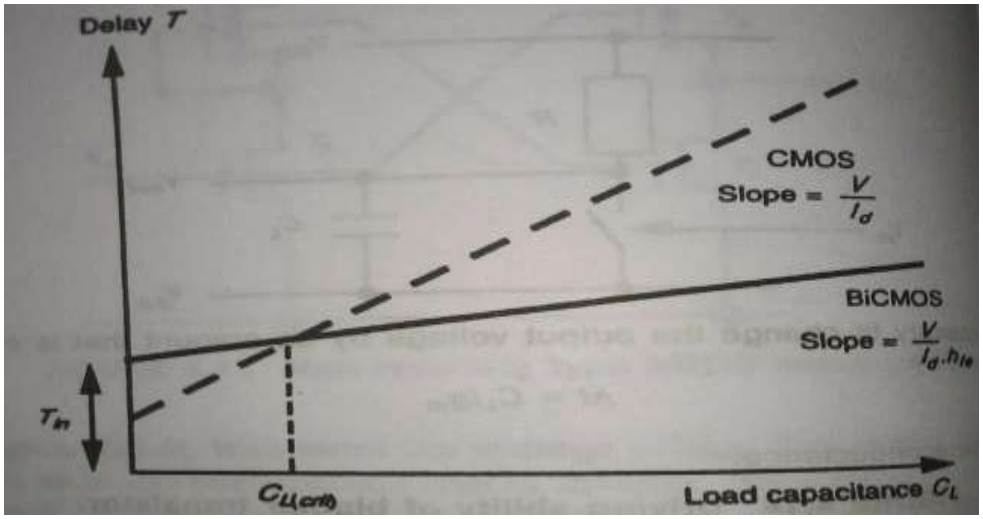


Fig 24 BICMOS driver

Combined effect of Tin & T_L is in graph. There is C_L critical load capacitance below which BICMOS driver is shown than CMOS driver.

- Delay of BICMOS is described by T = Tin+ $(V/I_d)(1/ h_{fe})C_L$.
- Delay for BICMOS inverter is reduced by a factor of h_{fe} when compared with CMOS inverter.