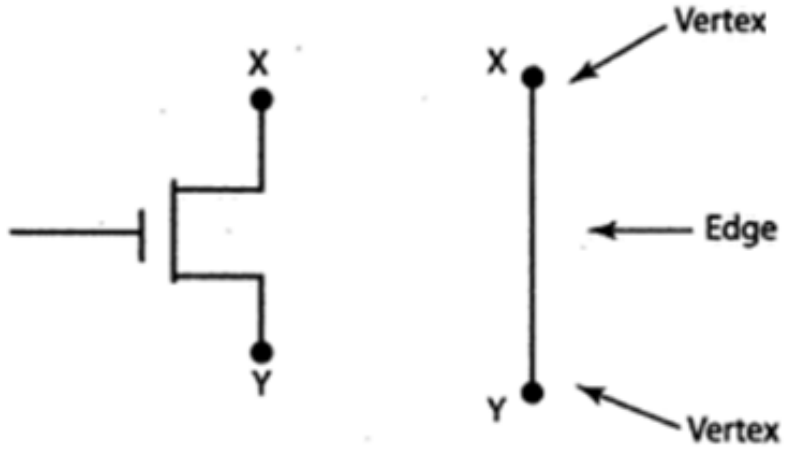
# Euler’s path, Stick diagrams and Layouts



Dr. Arjun Sunil Rao, Department of Electronics and Communication Engineering, MIT Manipal

# Euler's path

- An uninterrupted diffusion strip is possible only if there exists a Euler path in the logic graph




  - Diffusion without break.
  - Common Polysilicon running between P and N diffusion layer.
  - Less Number of Contacts.

Euler path: A path through all nodes in the graph such that each edge is visited once and only once, Nodes can be repeated.

# Stick diagrams

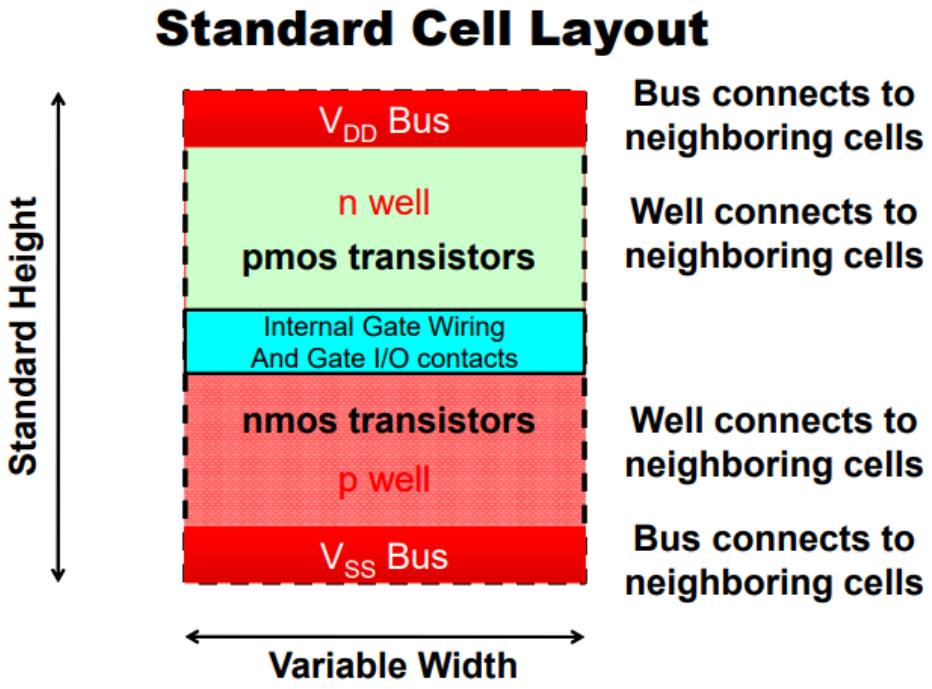
- We have seen that MOS circuits are formed on four basic layers-**n-diffusion**, **p-diffusion**, **polysilicon**, and **metal**, which are isolated from one another by thick or thin (*thinox*) silicon dioxide insulating layers.
- Polysilicon and *thinox* regions interact so that a transistor is formed where they cross the diffusion layer.
- In some processes, there may be a second metal layer and also, in some processes, a second polysilicon layer.
- Layers may deliberately joined together where contacts are formed.

# Stick diagrams

Stick diagrams may be used to convey layer information through the use of a color code

For example, in the case of nMOS design, green for n-diffusion, yellow for p-diffusion, red for polysilicon, blue for metal, yellow for implant, and black for contact areas.

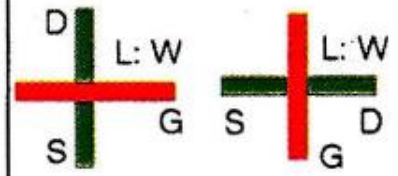
- VLSI design aims to translate circuit concepts onto silicon.
- stick diagrams are a means of capturing topography and layer information using simple diagrams.
- Stick diagrams convey layer information through colour codes (or monochrome encoding).
- Acts as an interface between symbolic circuit and the actual layout.

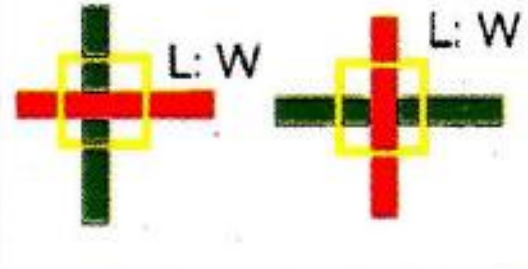


# Stick diagrams

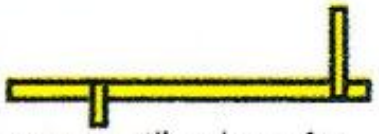
| COLOR | STICK ENCODING | LAYERS |
|---|---|---|
| GREEN |  | n-diffusion (n+ active) Thinox* |
| RED |  | Polysilicon |
| BLUE |  | Metal 1 |
| BLACK |  | Contact cut |
| GRAY | NOT APPLICABLE | Overglass |

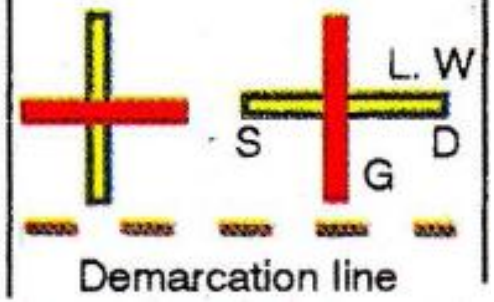
| | | |
|---|---|---|
| nMOS ONLY YELLOW |  | Implant |
| nMOS ONLY BROWN |  | Buried contact |

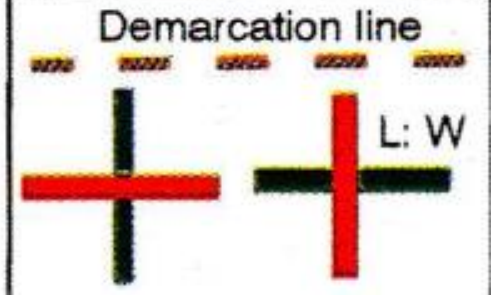
| FEATURE | FEATURE (STICK) |
|---|---|
| n-type enhancement mode transistor |   |

| | |
|---|---|
| n-type depletion mode transistor<br>nMOS only |   |

# Stick diagrams

| | | |
|---|---|---|
| YELLOW (STICK) |  green outline here for clarity | p-diffusion ($p^+$ active) |
| YELLOW | Not shown on diagram | $p^+$ mask |
| DARK BLUE OR PURPLE |  | Metal 2 |
| BLACK |  | VIA |
| BROWN | Demarcation line  p-well edge is shown as a demarcation line in stick diagrams | p-well |
| BLACK |  | $V_{DD}$ or $V_{SS}$ contact |

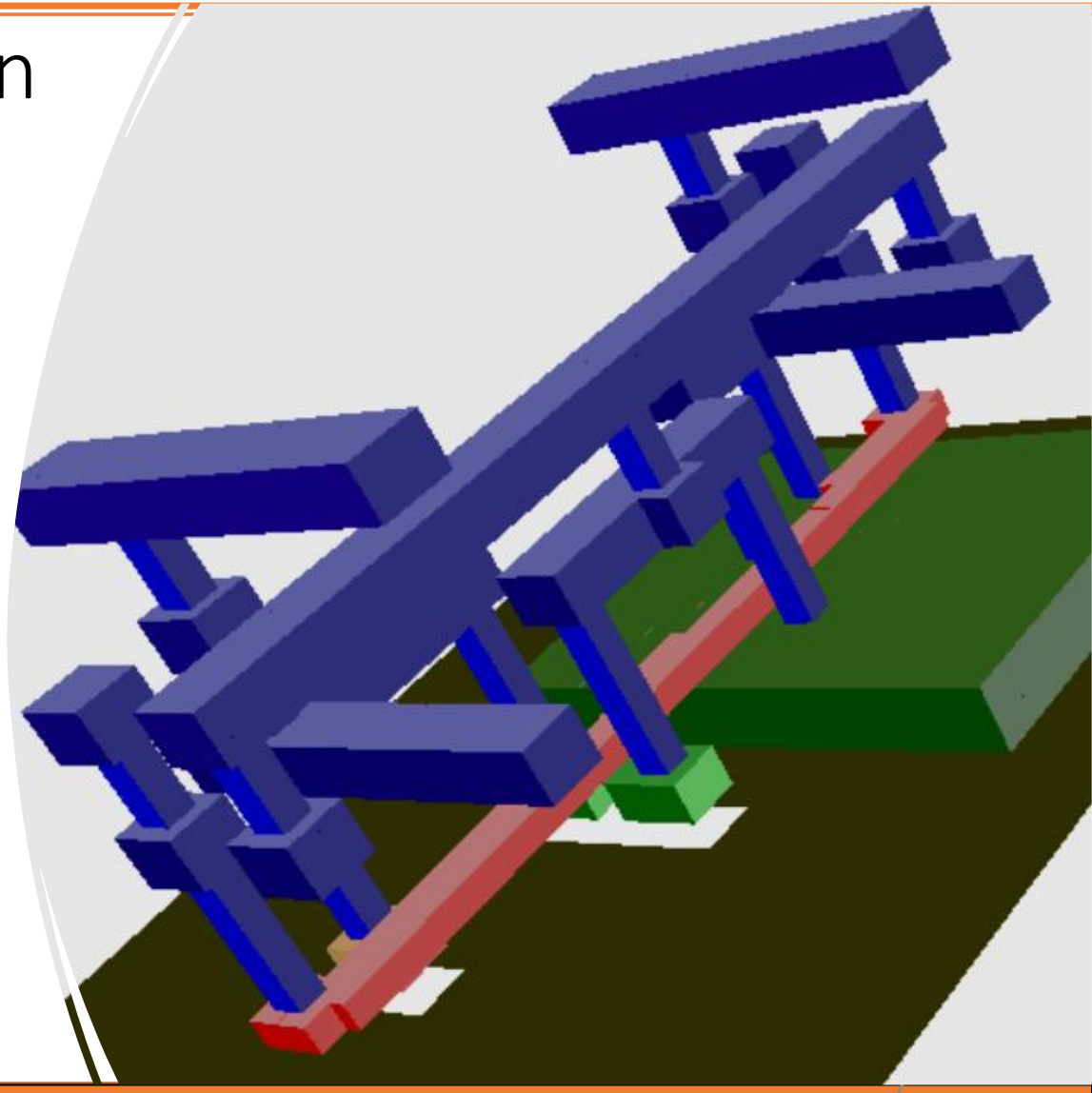
| | |
|---|---|
| *p-type enhancement mode transistor* |   |

| FEATURE | FEATURE (STICK) |
|---|---|
| *n-type enhancement mode transistor (as in Color plate 1(a))* |   |

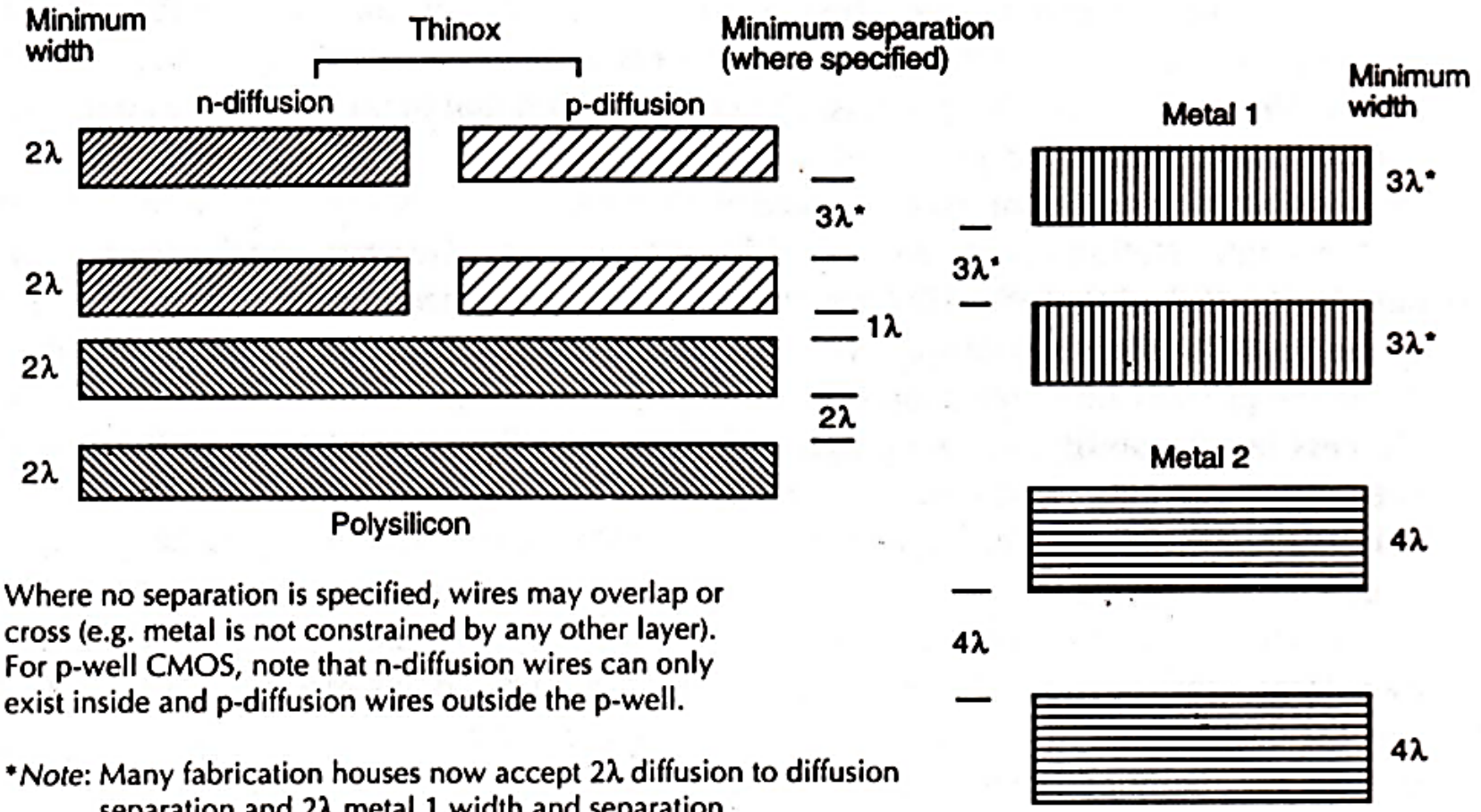
# Layouts and Lambda based design rules

## Importance of Lambda based design rules

- To allow for shape contraction.
- To ensure adequate continuity of the intervening materials.
- To avoid the possibility of metal-metal or polySi-polySi regions overlapping and conducting currents.
- To prevent the lines overlapping to form unwanted capacitor.
- Feature size refers to minimum transistor length, so lambda is half the feature size.



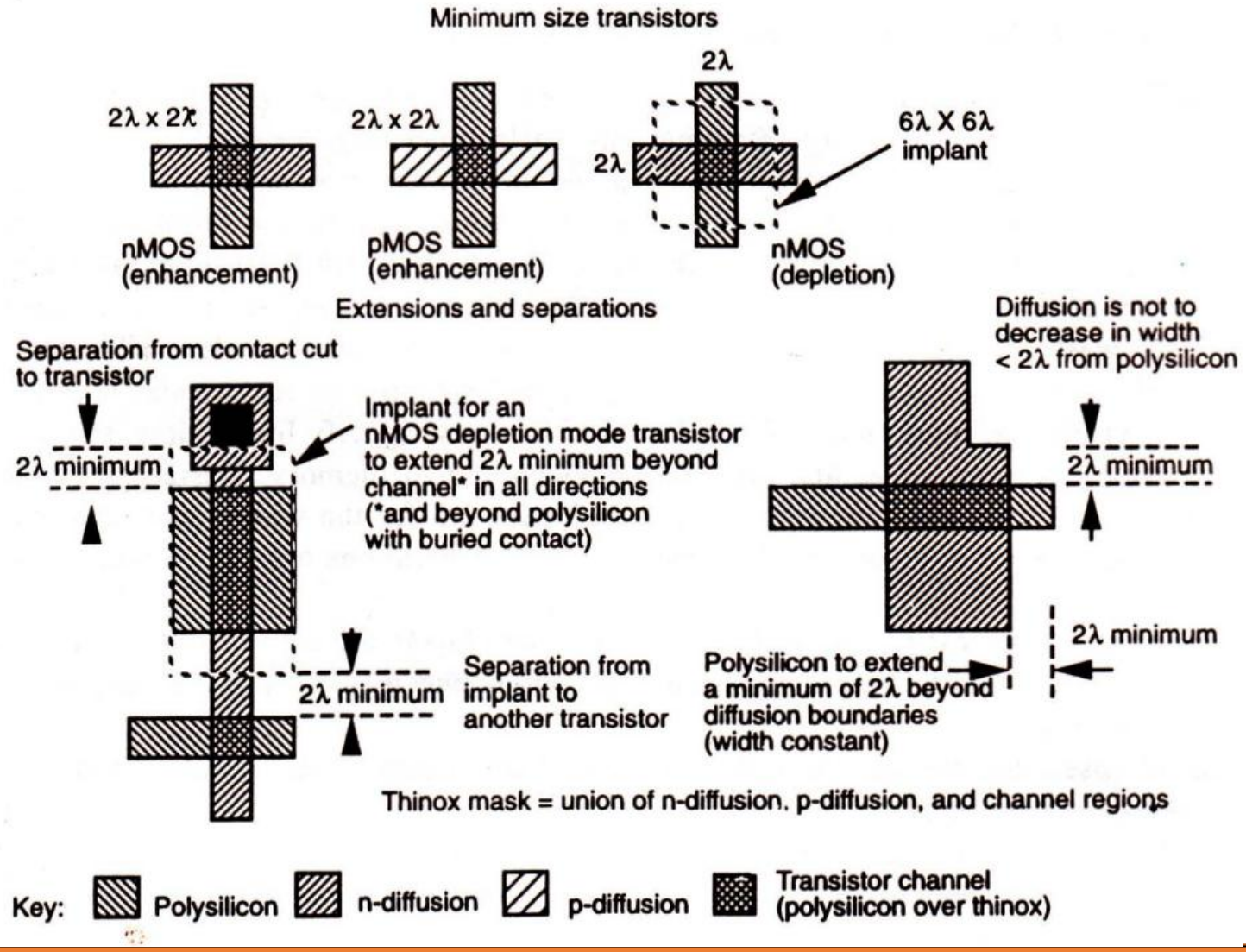
# 1. Design rules for wires (nMOS and CMOS)




Where no separation is specified, wires may overlap or cross (e.g. metal is not constrained by any other layer). For p-well CMOS, note that n-diffusion wires can only exist inside and p-diffusion wires outside the p-well.

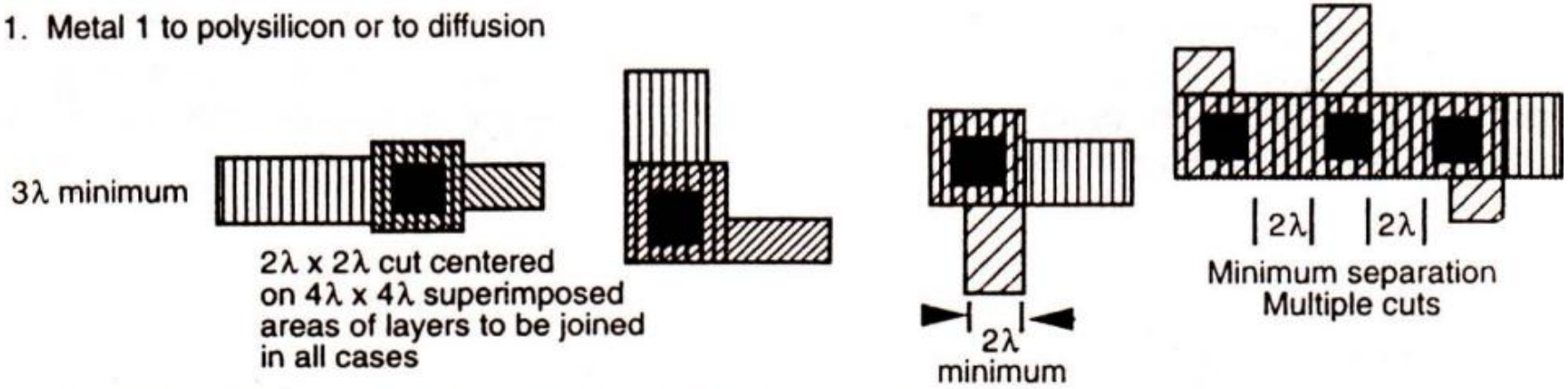
**Note*: Many fabrication houses now accept 2λ diffusion to diffusion separation and 2λ metal 1 width and separation.

# 2. Transistor design rules (nMOS, pMOS and CMOS).

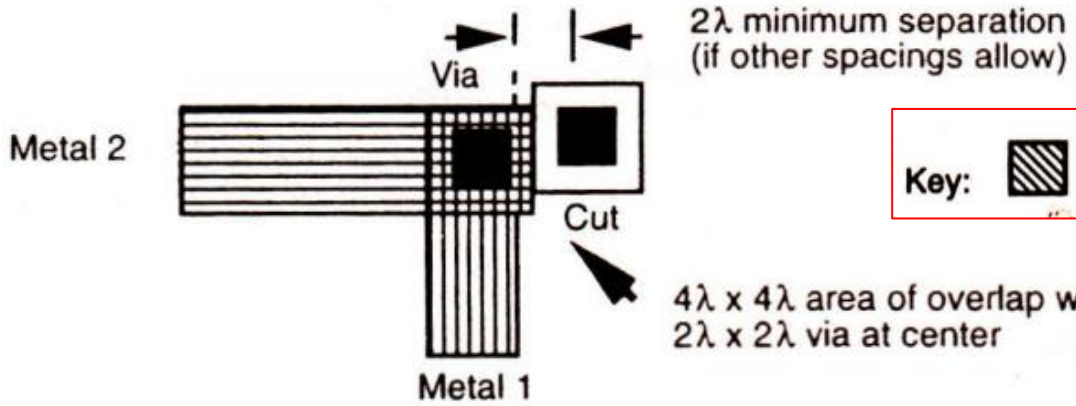


# 3. Contacts (nMOS and CMOS).

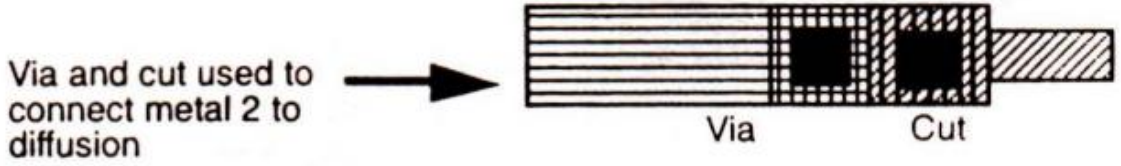
1. Metal 1 to polysilicon or to diffusion




2. Via (contact from metal 2 to metal 1 and thence to other layers)





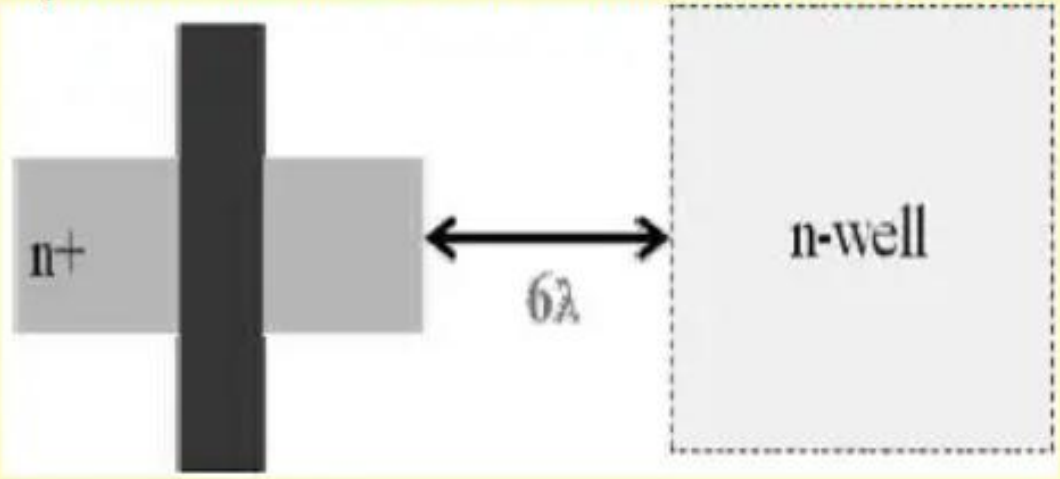


# 4. N-well rule

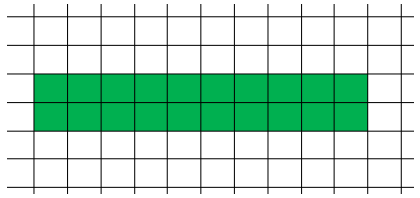
i. To ensure the separation of the PMOS and NMOS devices, n-well supporting PMOS is 6λ away from the active area of NMOS transistor.

ii. N-well must completely surround the PMOS device active area by 2λ

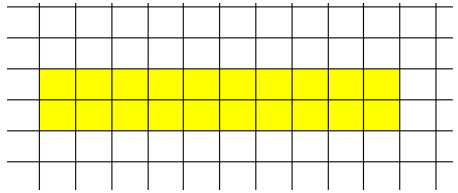
iii. The threshold implant mask covers all n-well and surrounds the n-well by λ



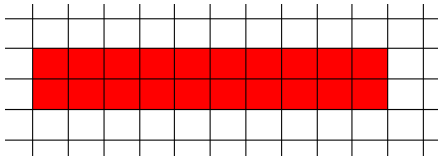
# Recommendation: Use graphical page to draw layout



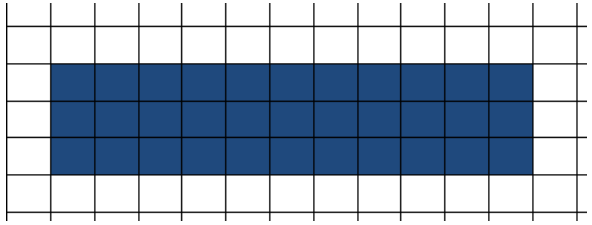
$N^+$ Diffusion



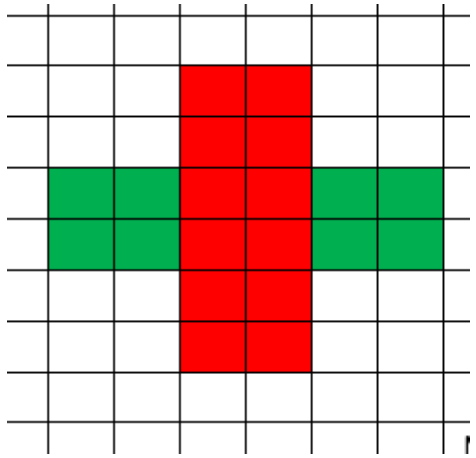
P+ Diffusion



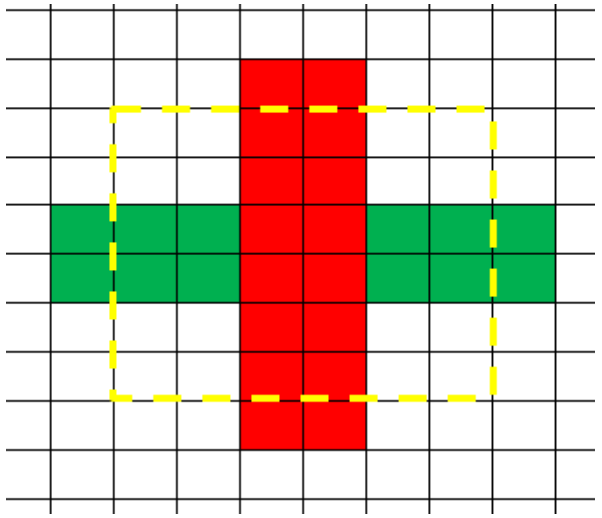
Polysilicon



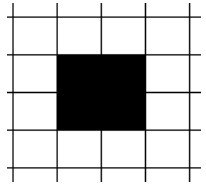
Metal



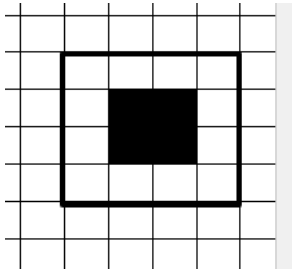
nMOS
(Enhancement) FET



nMOS (Depletion) FET

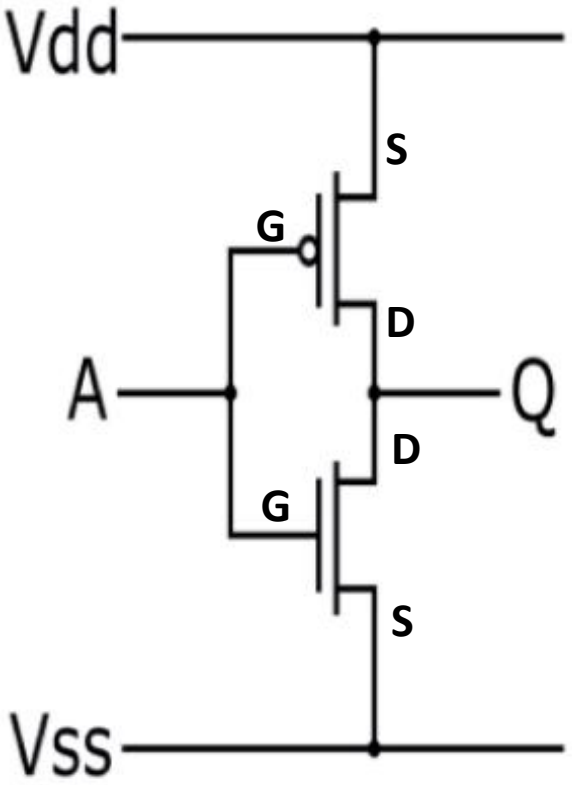


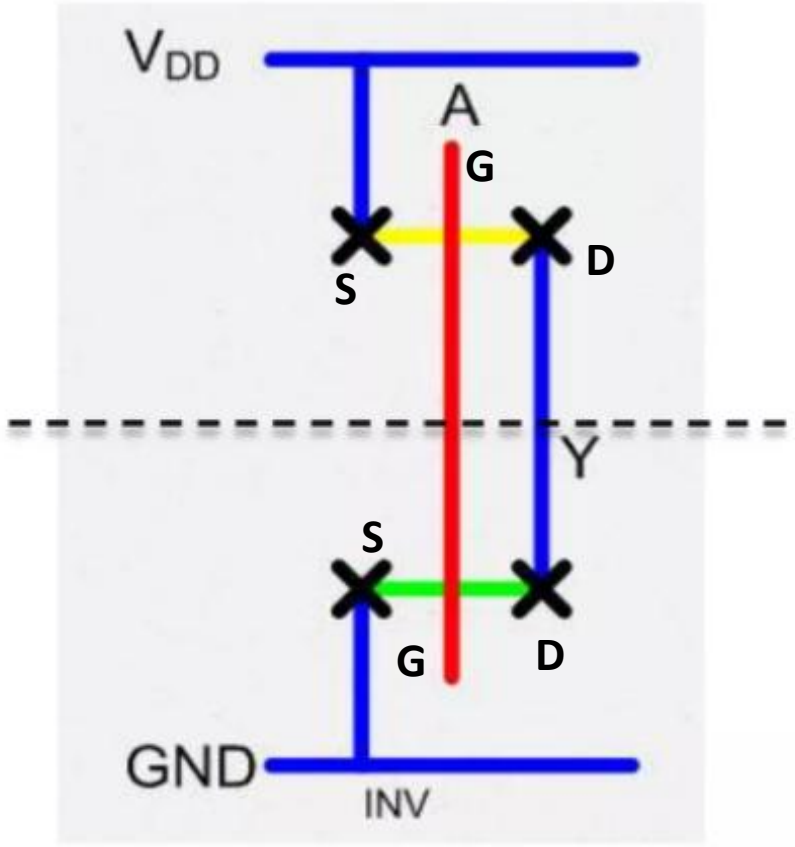
Contact cut

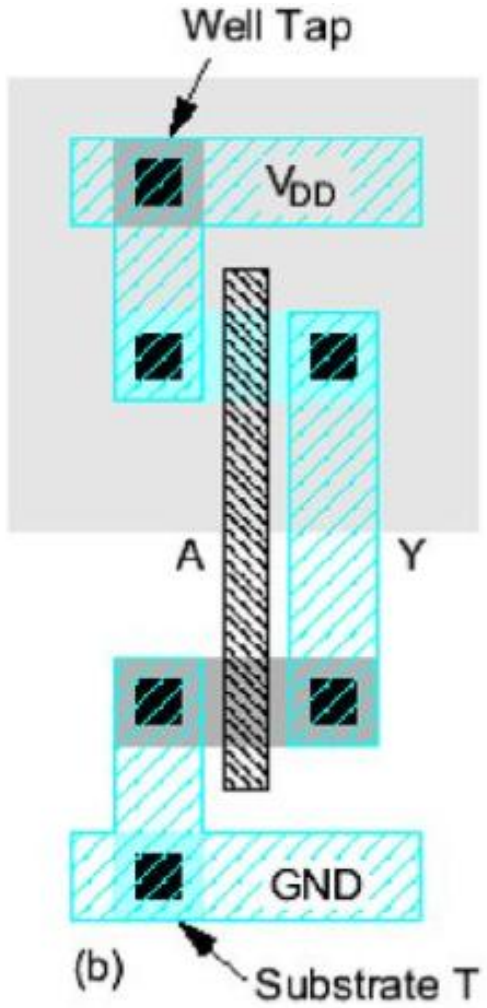


Contact pad

# Draw the stick diagram and layout of CMOS inverter.

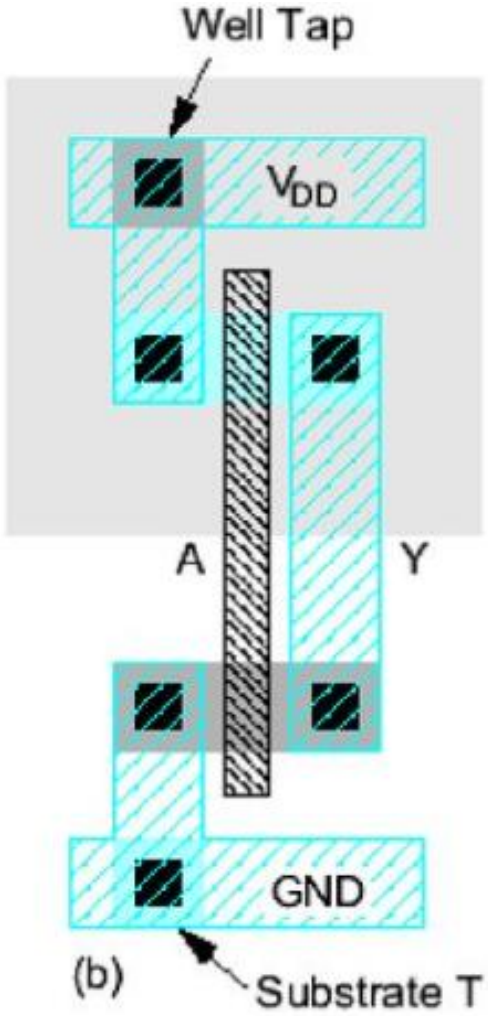


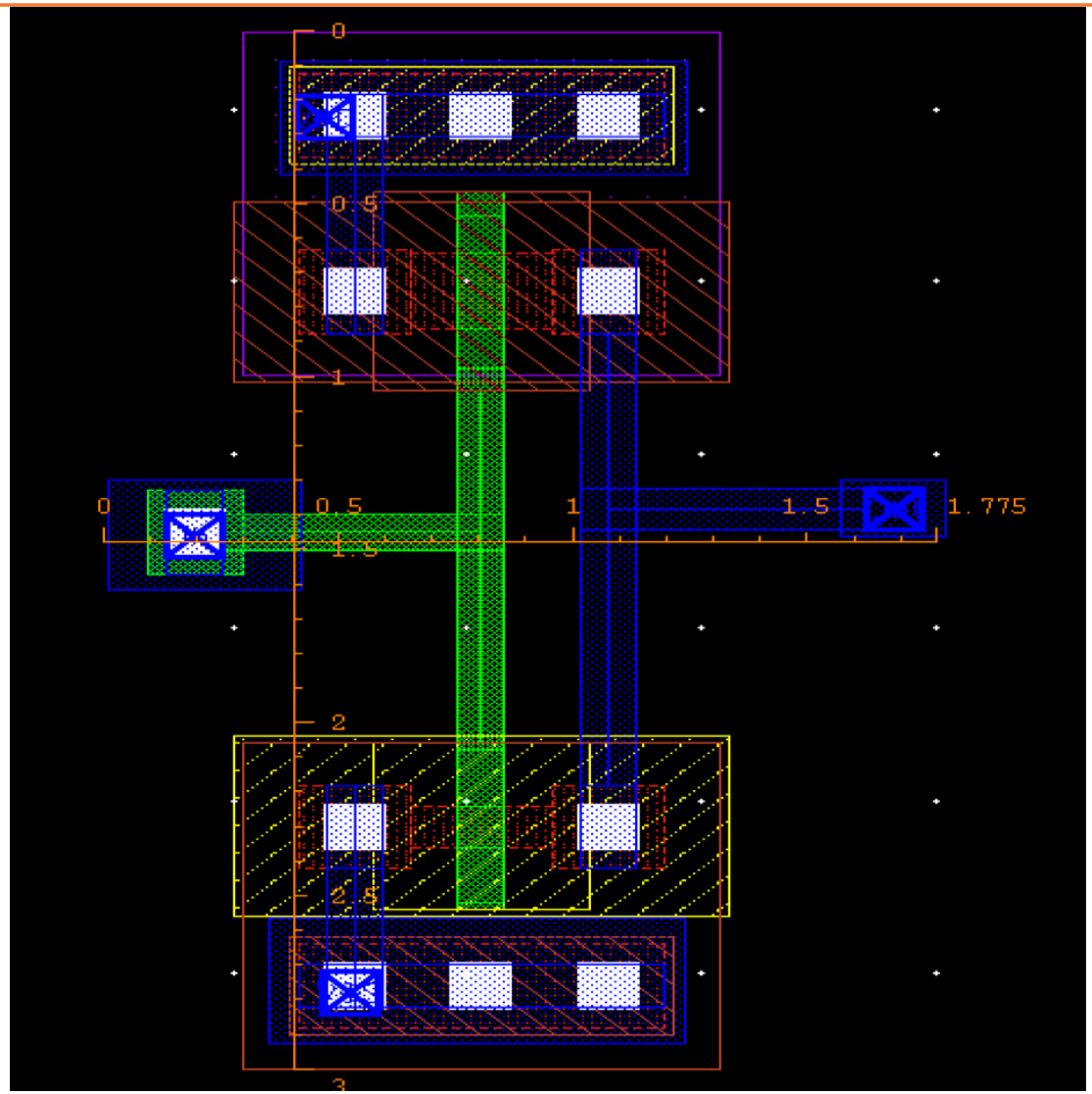



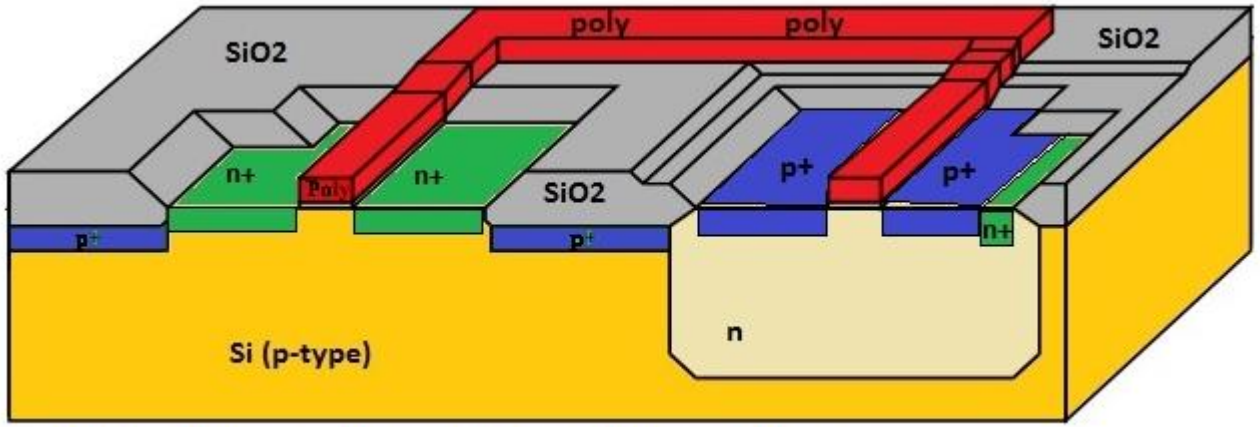




- Metal
- Polysilicon
- Metal Contact

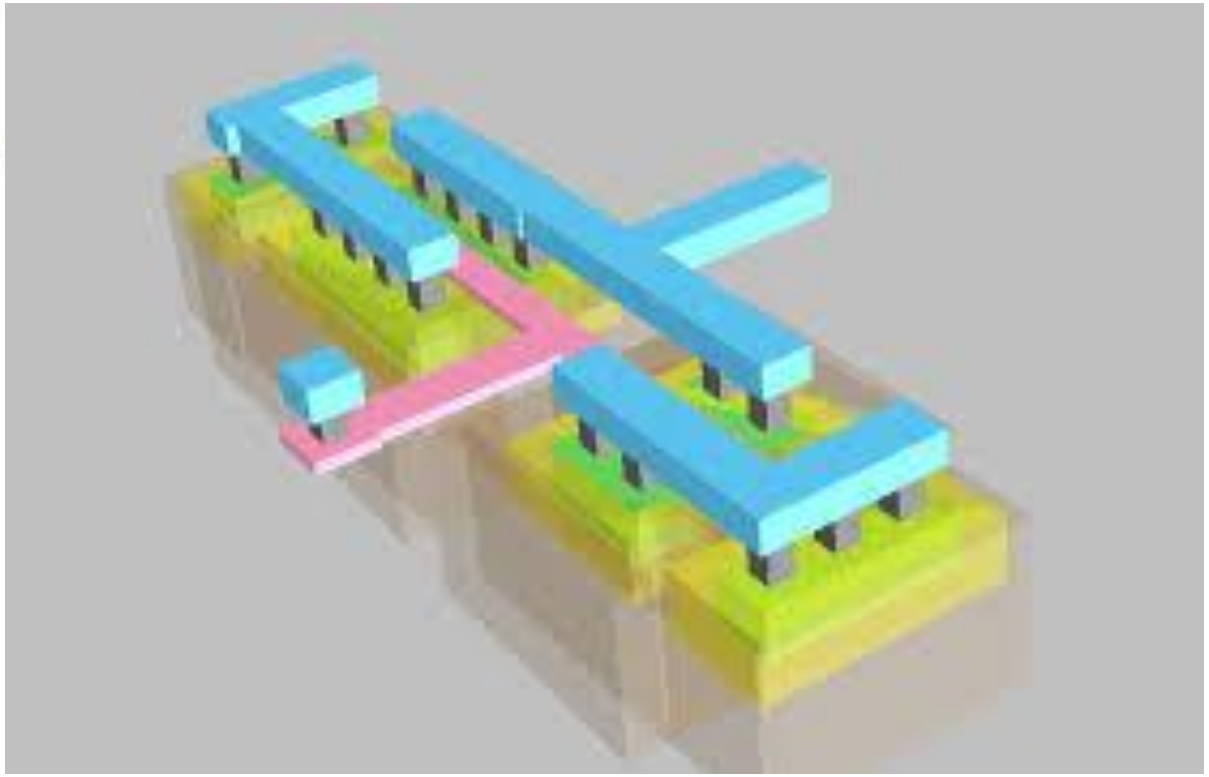
- P-Doping
- N-Doping
- **Demarcation line**


Well Tap
$V_{DD}$
A
Y
GND
(b)
Substrate T

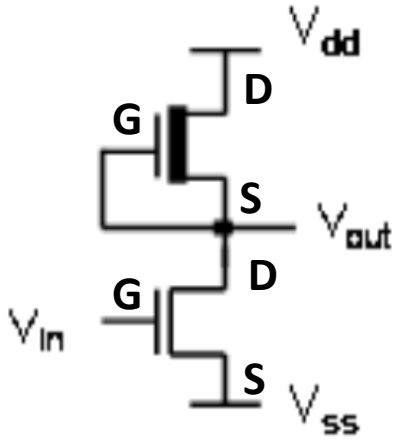

0
0.5
1
0
0.5
1
1.5
1.775
2
2.5
3



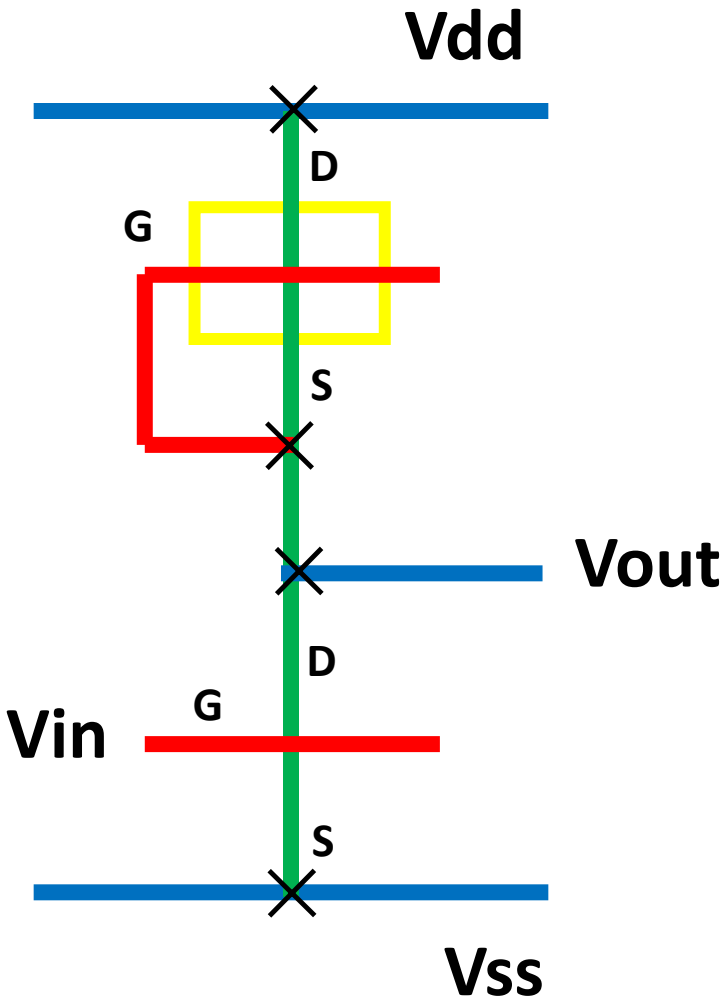

3D view of silicon wafer. Now we have PMOS and NMOS Device.

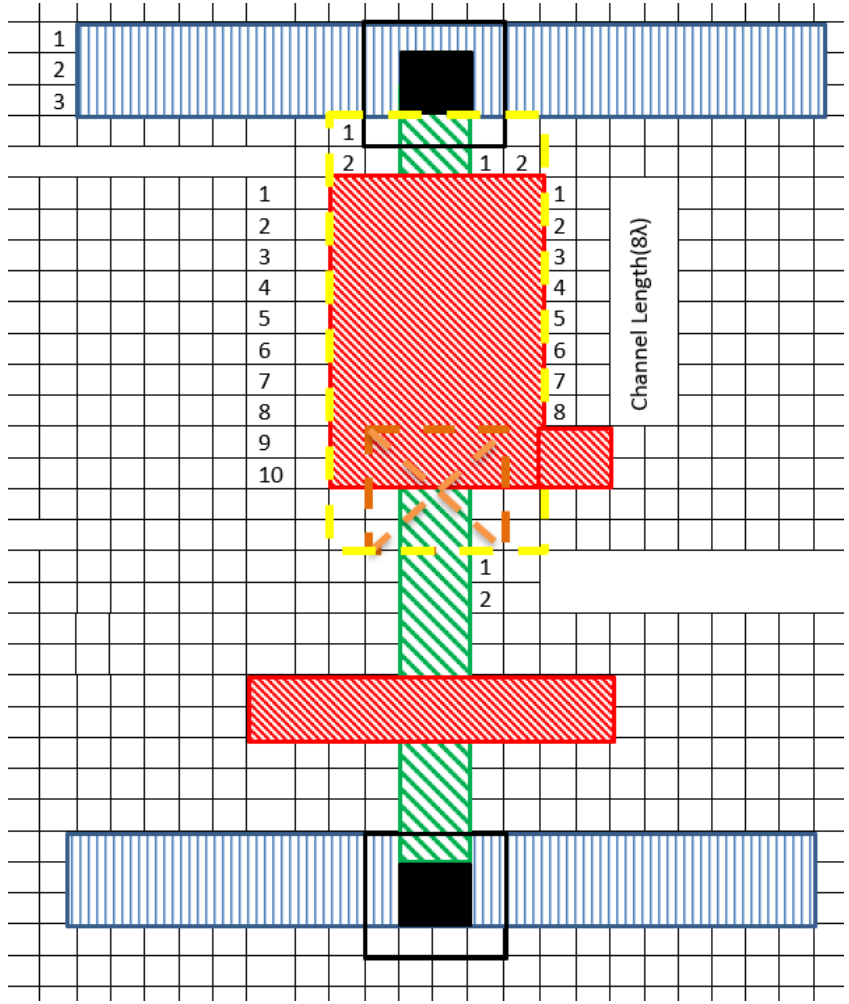


# Draw the stick diagram and layout of nMOS inverter with depletion load.

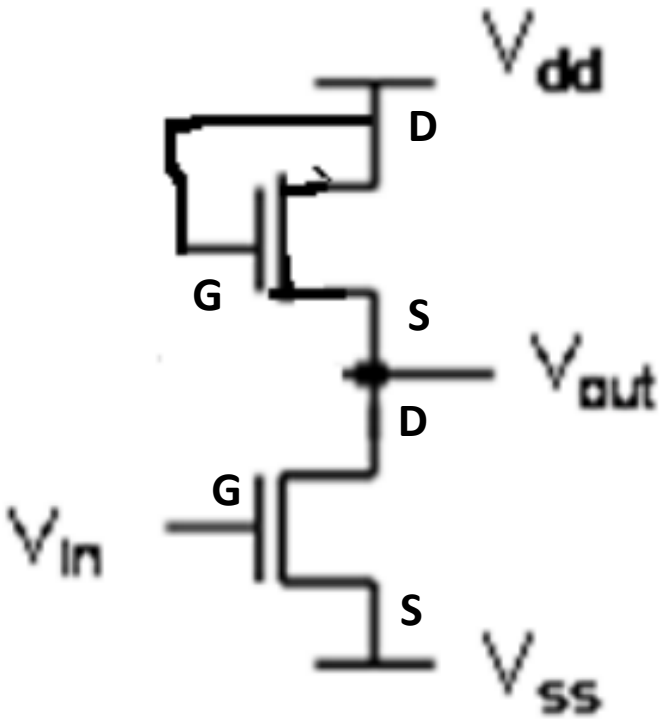



- Metal
- Polysilicon
- Metal Contact
- P-Doping
- N-Doping

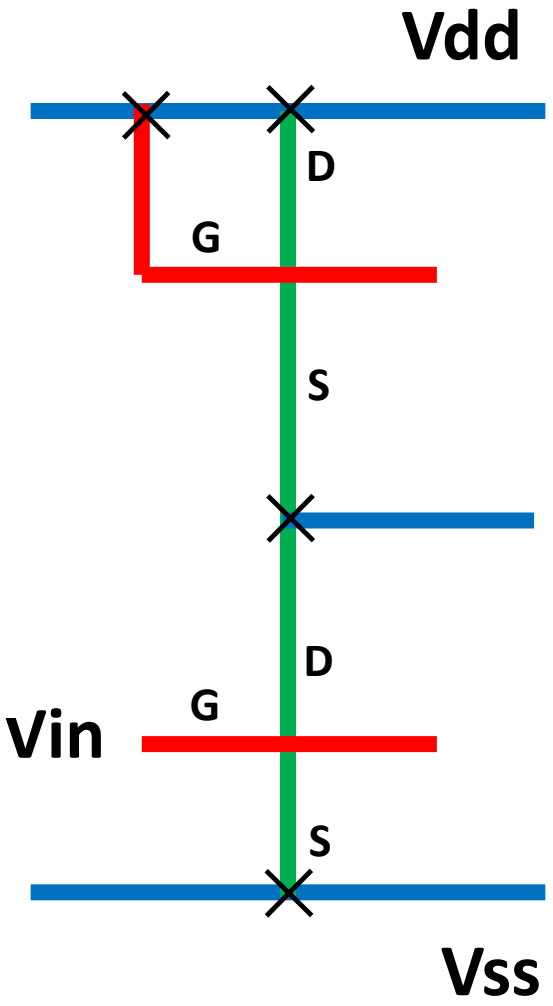


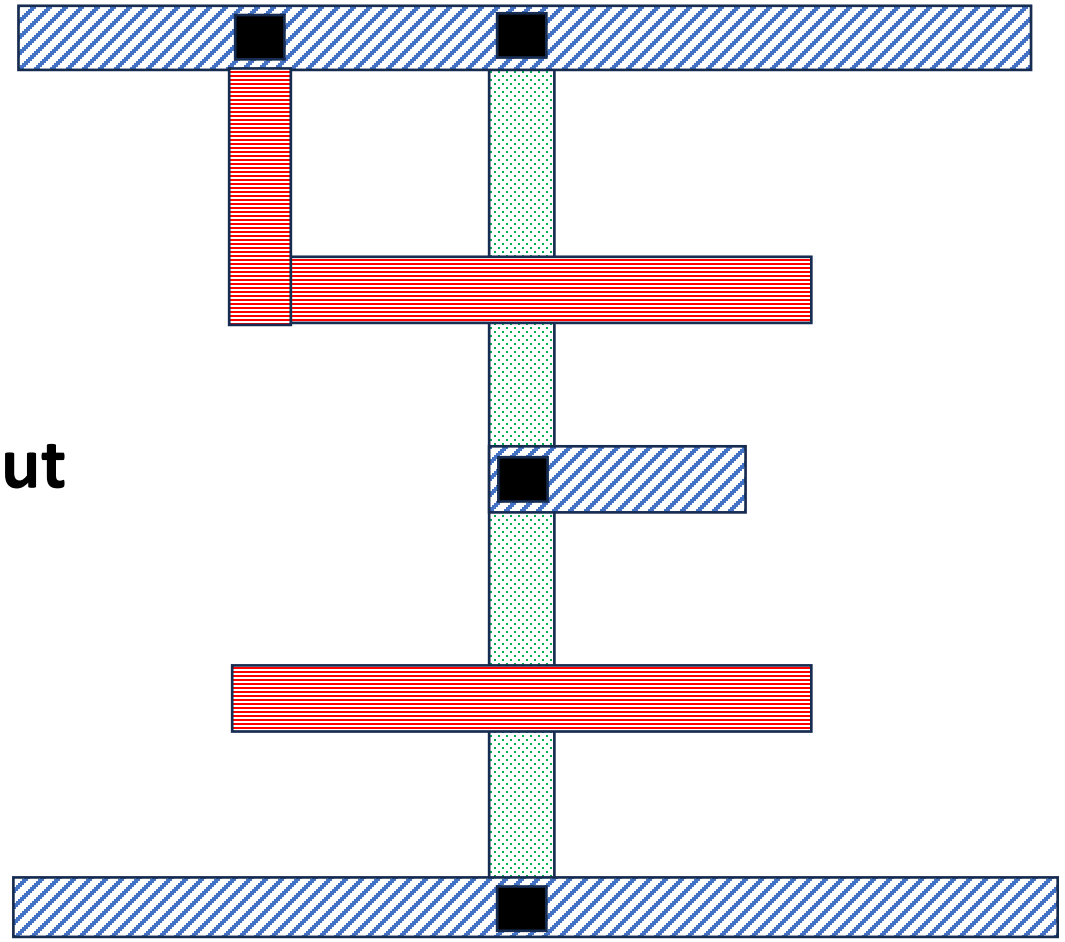



# Draw the stick diagram and layout of nMOS inverter with enhancement load.

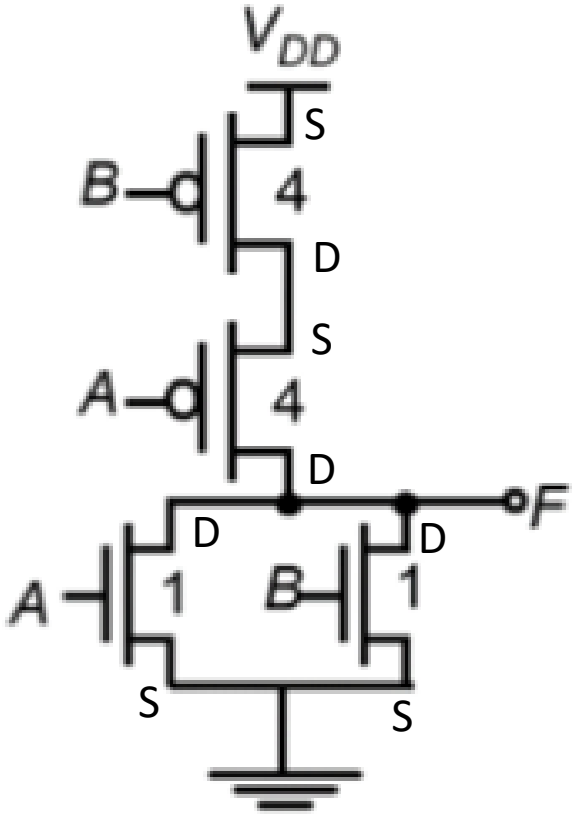


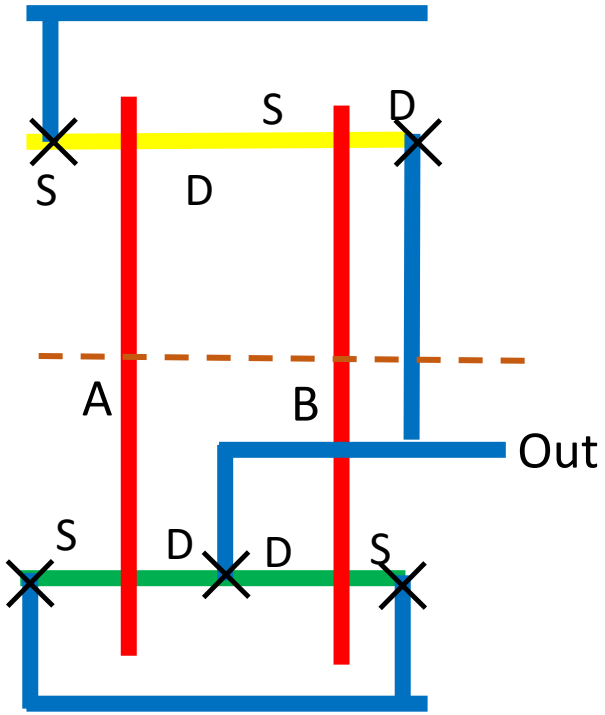

- Metal
- Polysilicon
- Metal Contact
- P-Doping
- N-Doping

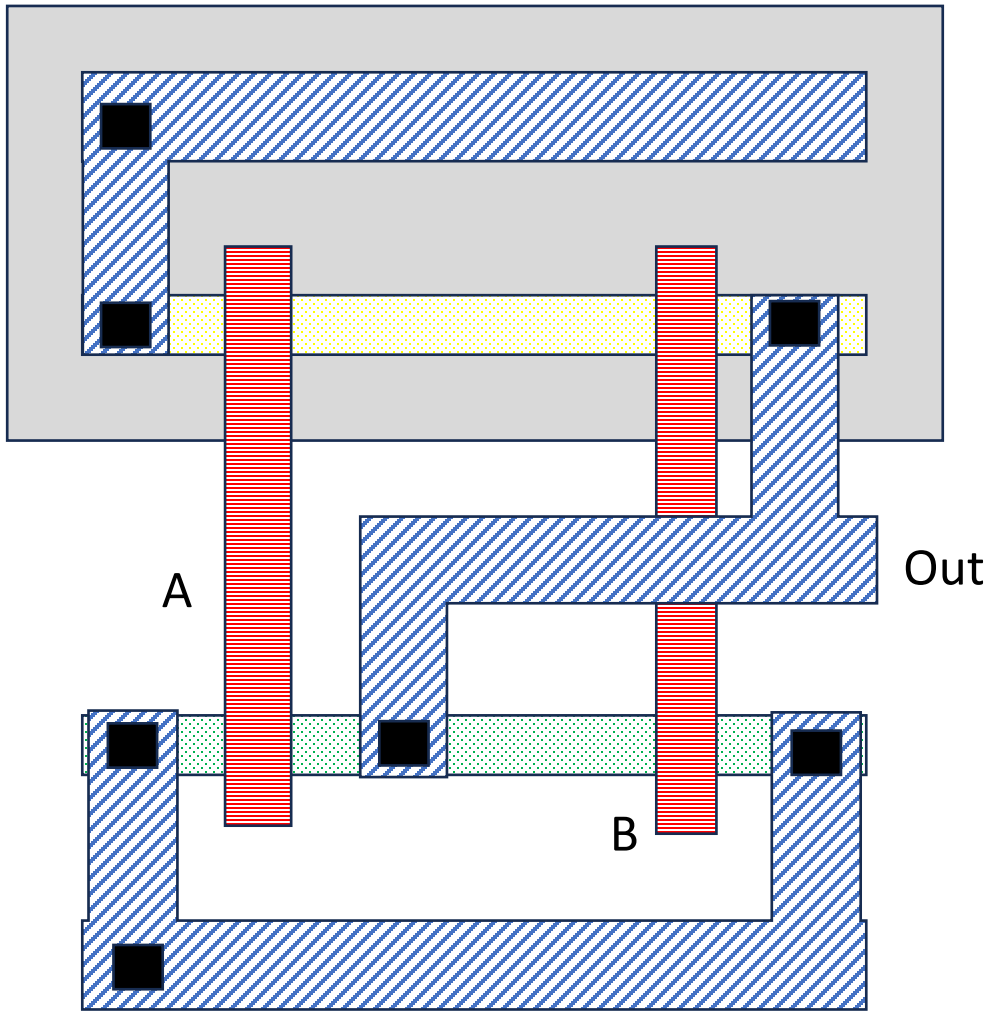


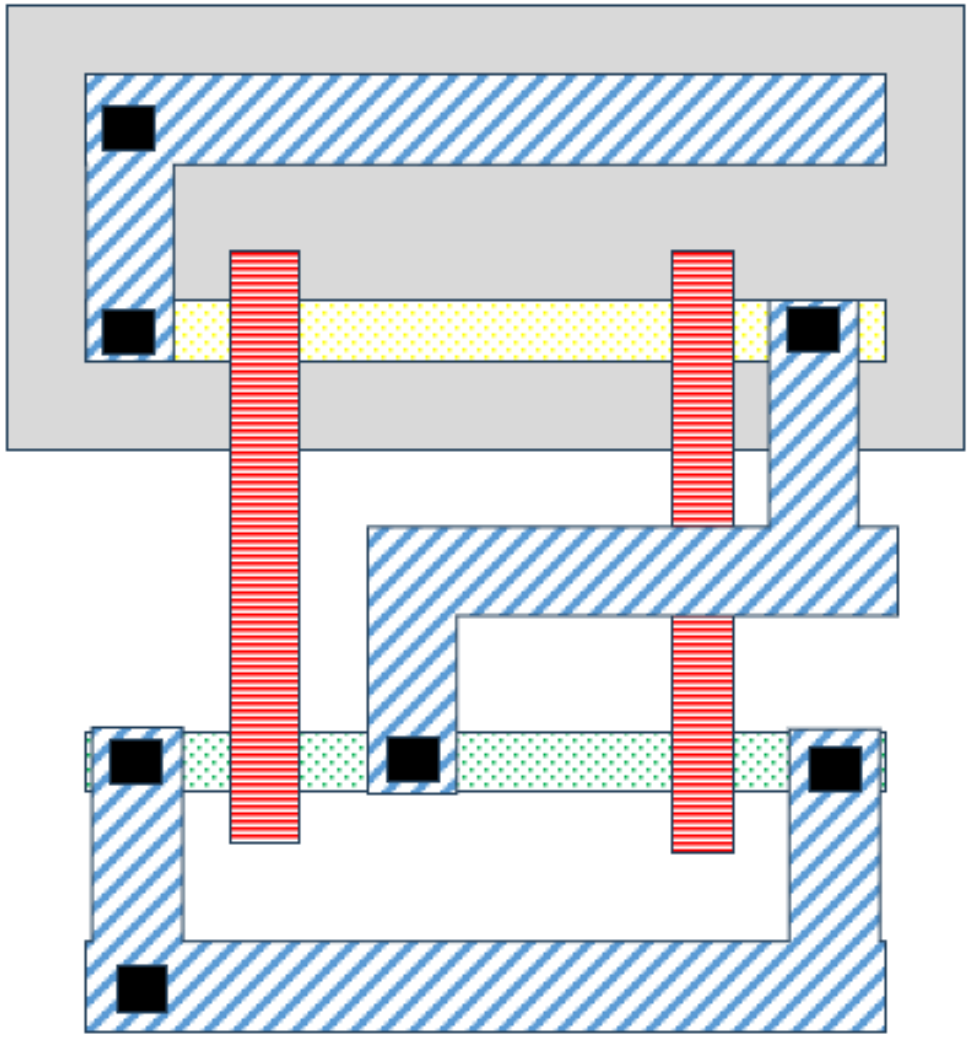



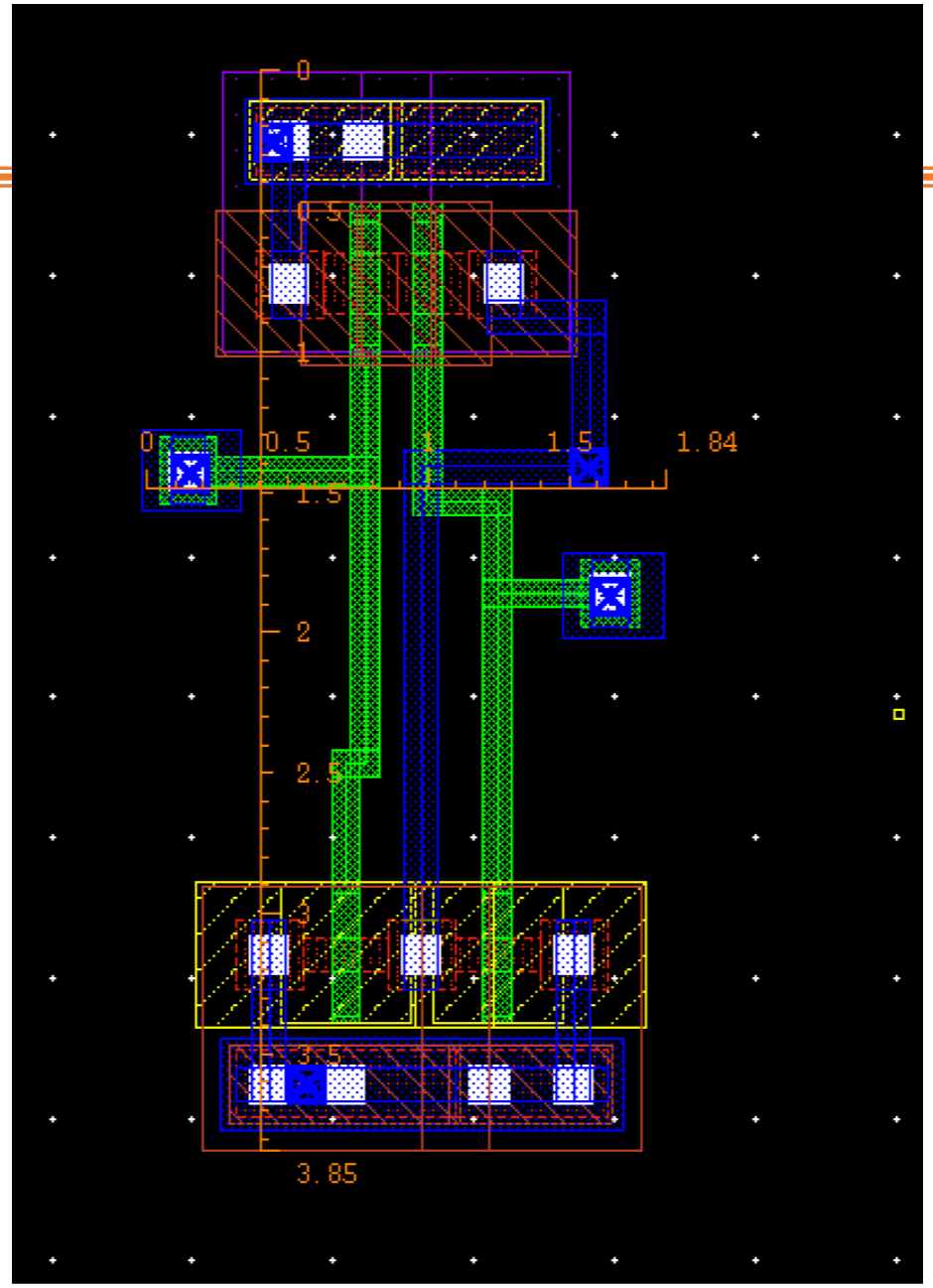
# Draw the stick diagram and layout of CMOS NOR2



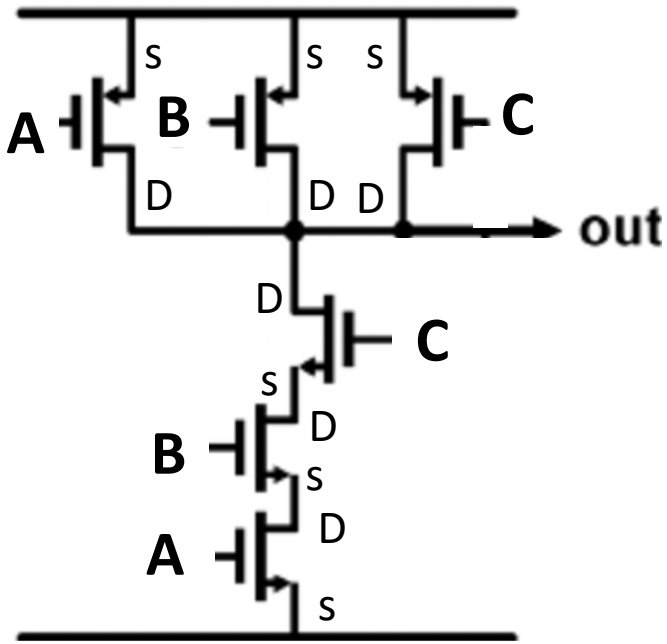



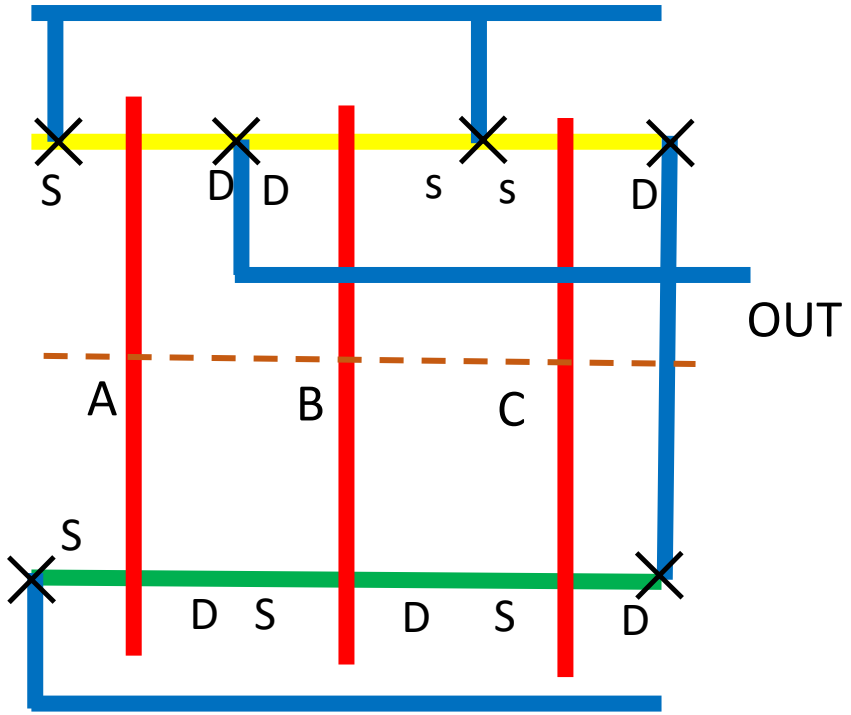



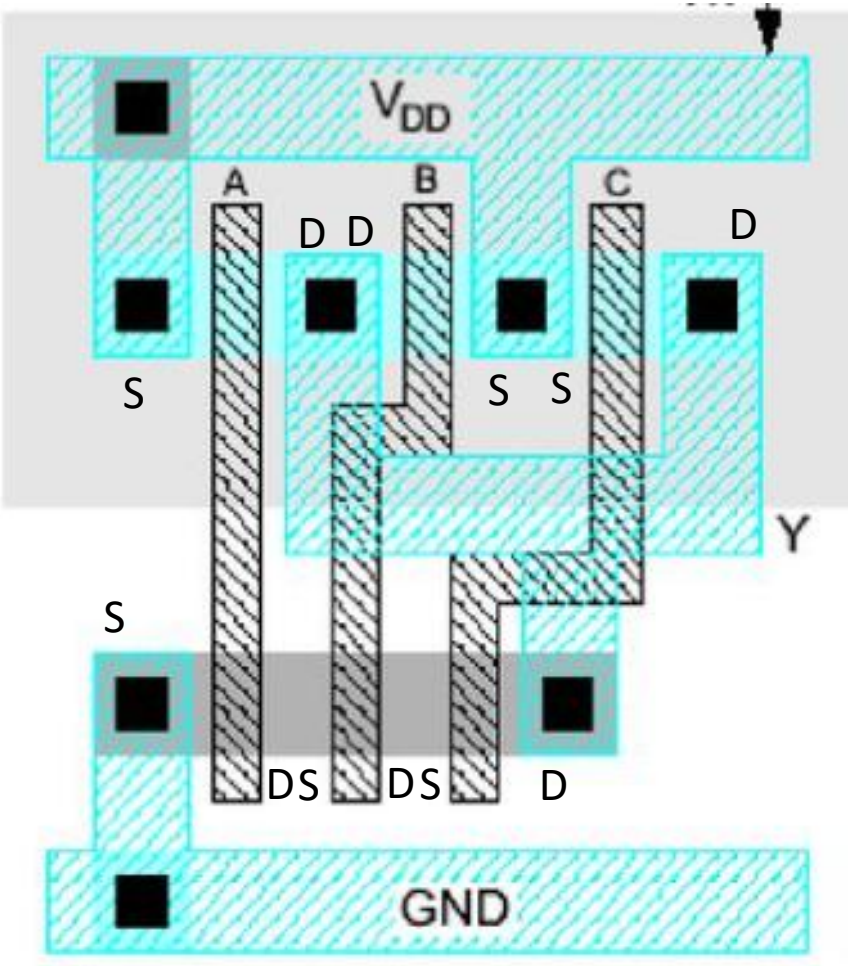



0
0.5
1
1.5
1.84
2
2.5
3.85

# Draw the stick diagram and layout of CMOS NAND3



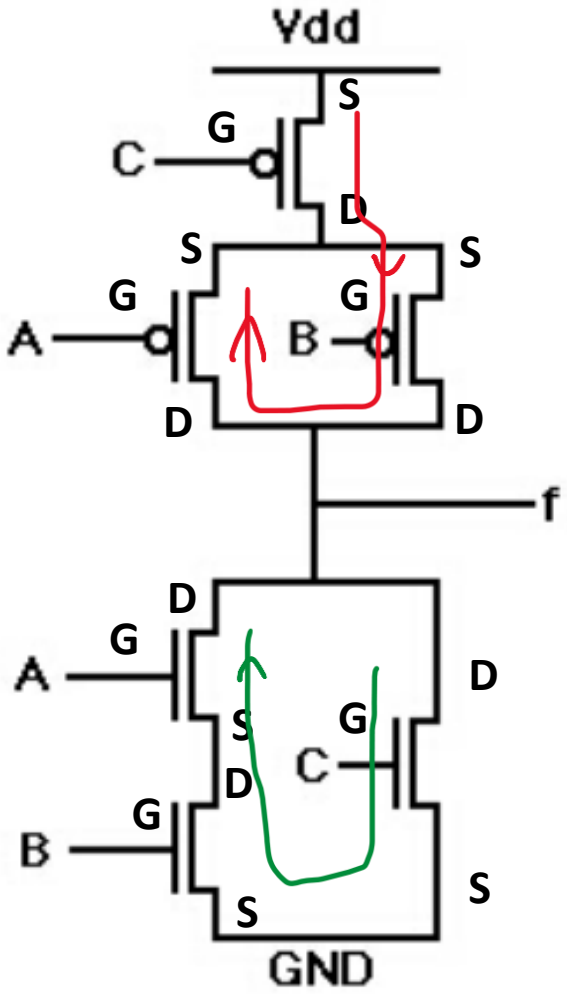



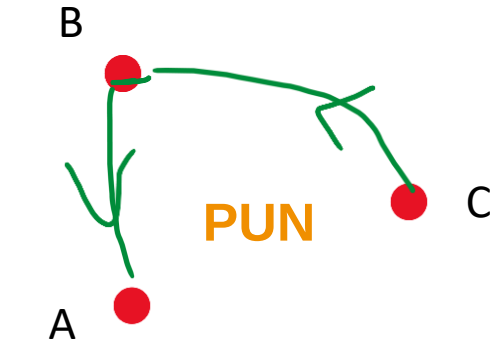
**Draw the stick diagram and layout of CMOS NAND2**

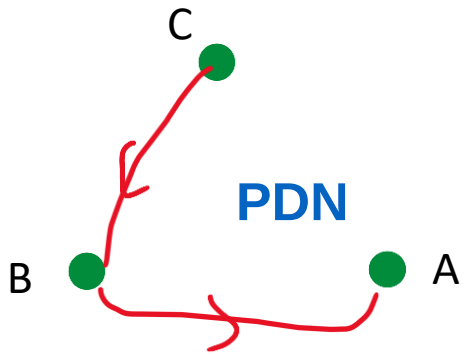
**Draw the stick diagram and layout of CMOS NOR3**

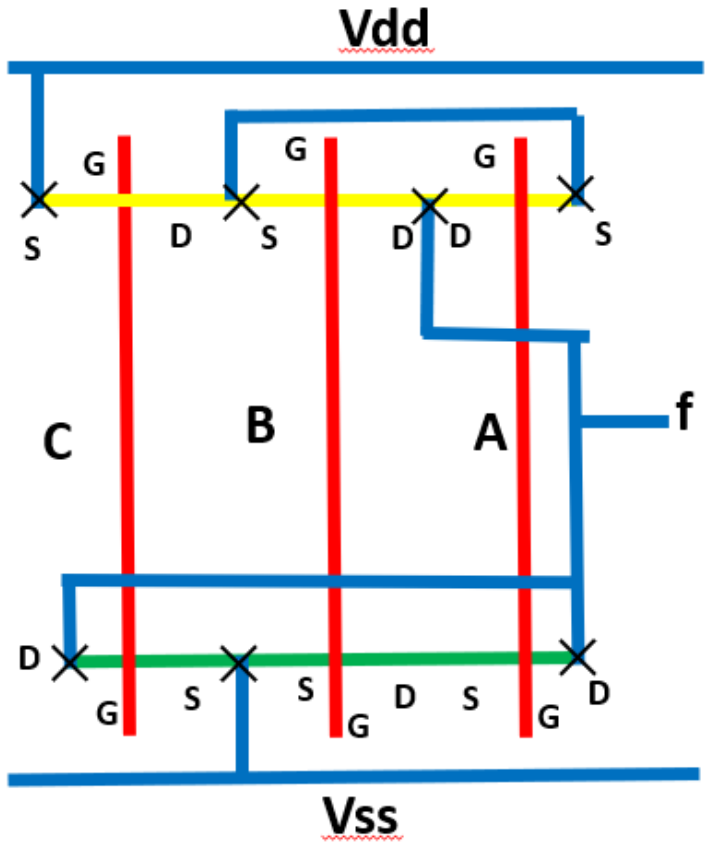
# Draw the stick diagram and layout of graph of $f = \overline{(A \cdot B) + C}$

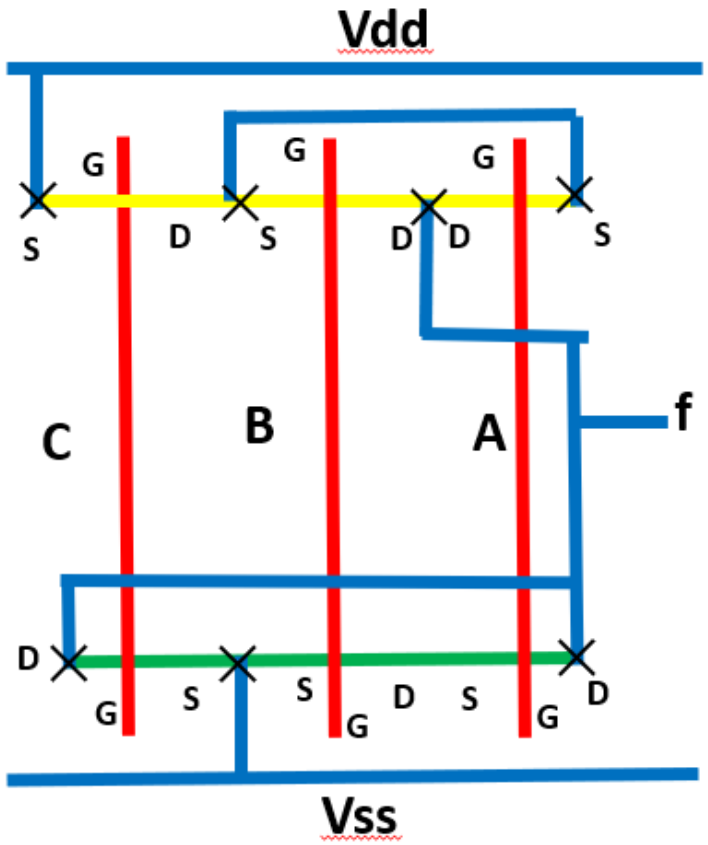


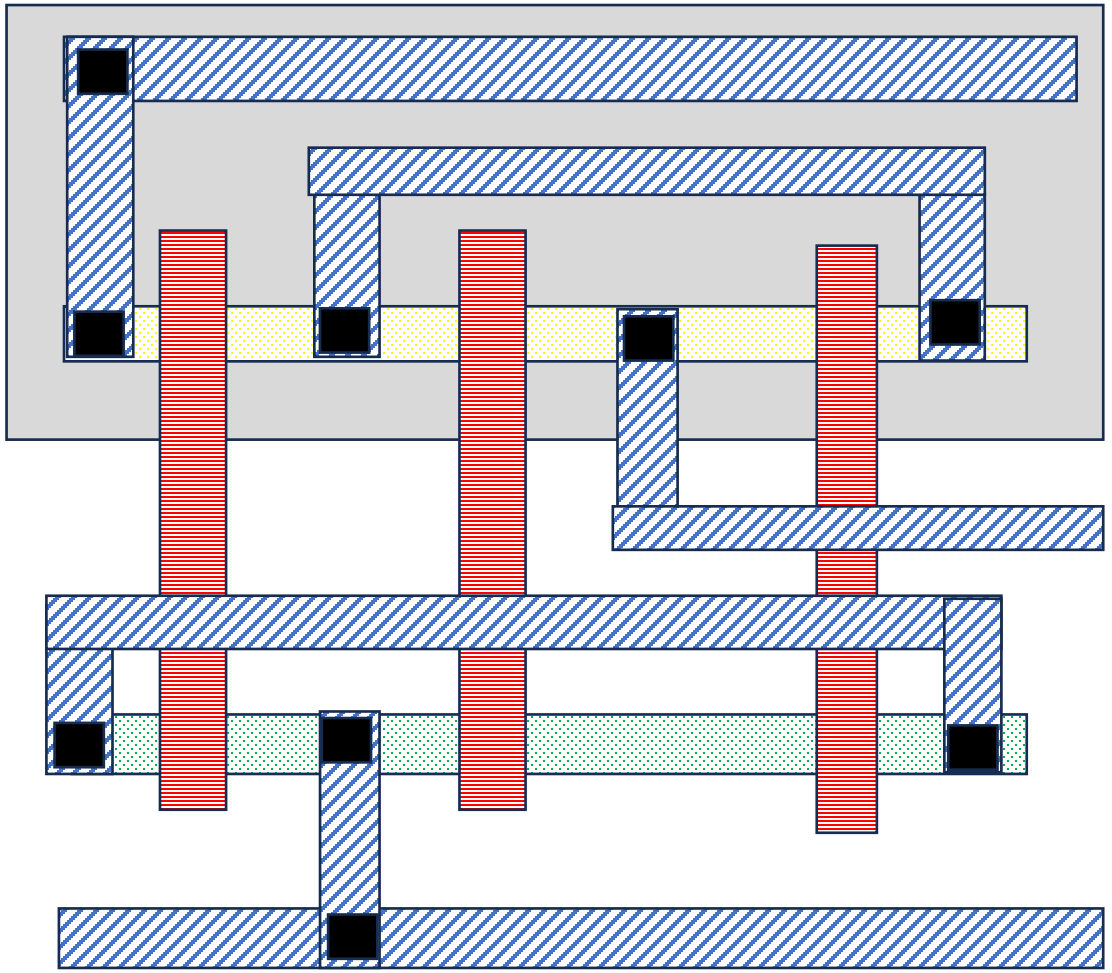

**Euler's path is C → B → A**



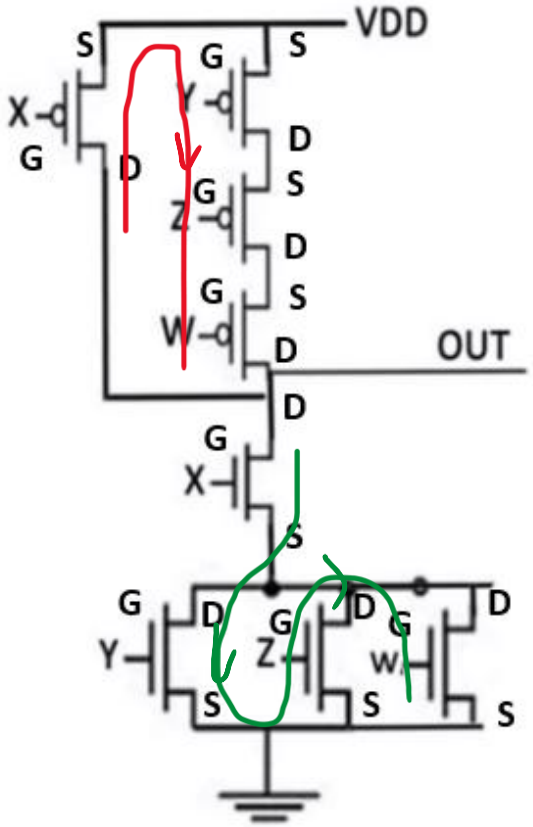



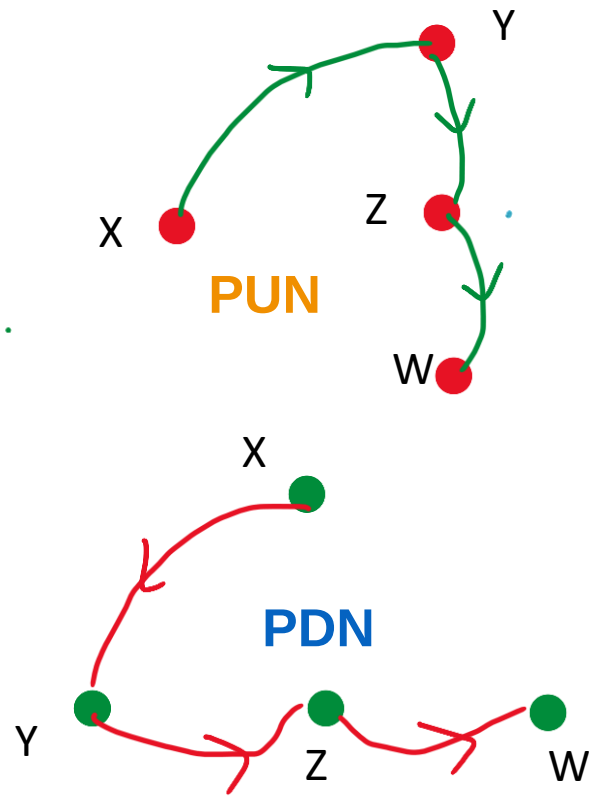

Vdd
G
G
G
S
D
S
D
D
S
C
B
A
f
D
G
S
S
G
D
S
G
D
Vss

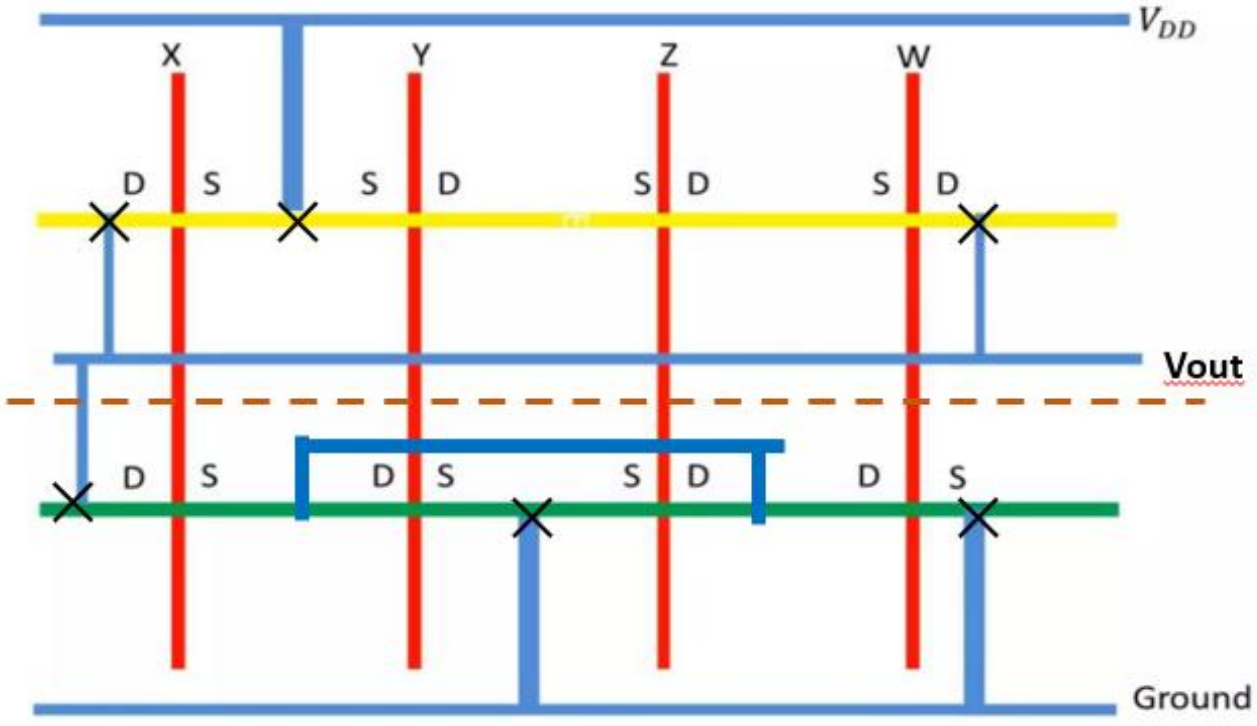


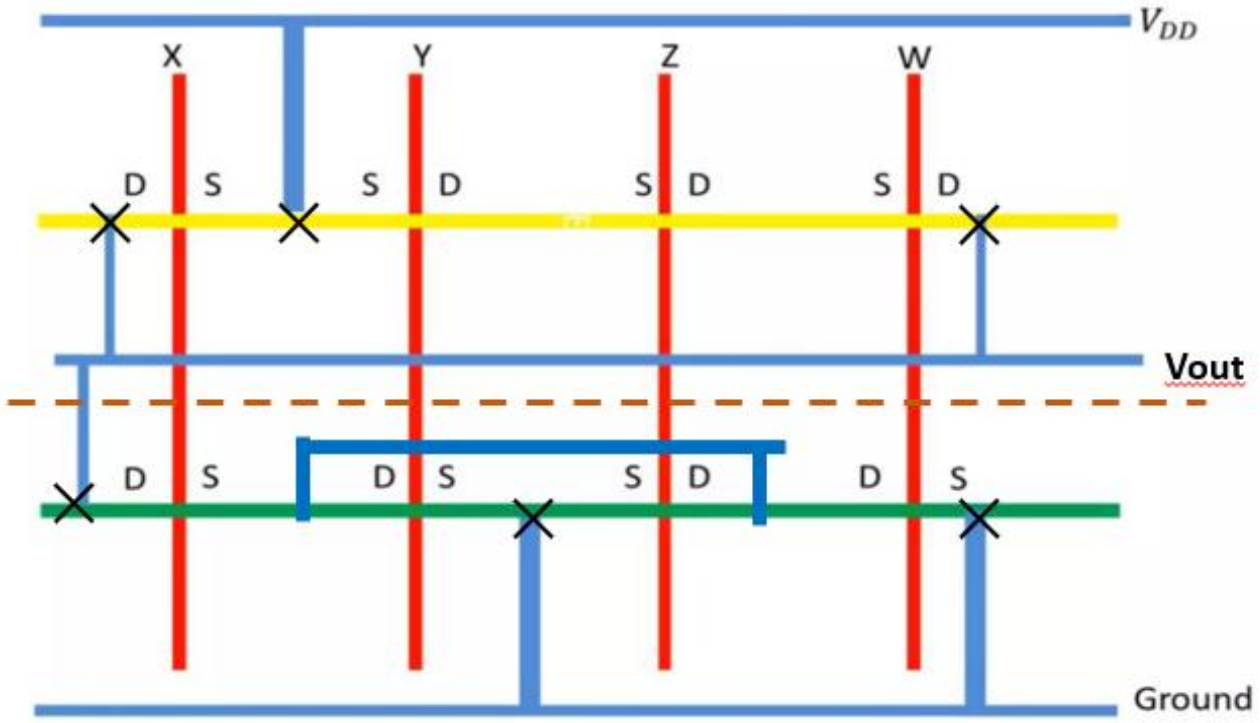
# Draw the stick diagram and layout of graph of F=((y.z.w)+x)'

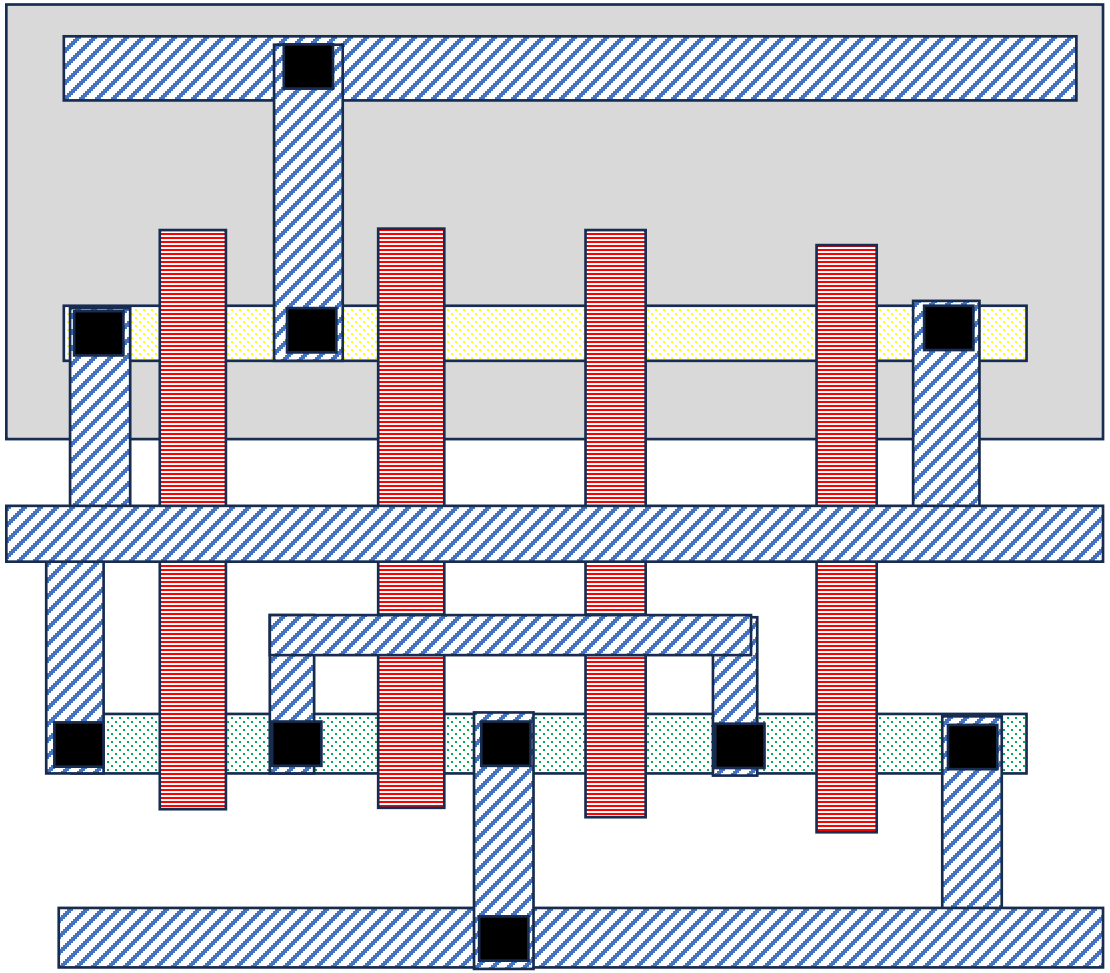
**Euler's path is X → Y → Z → W**



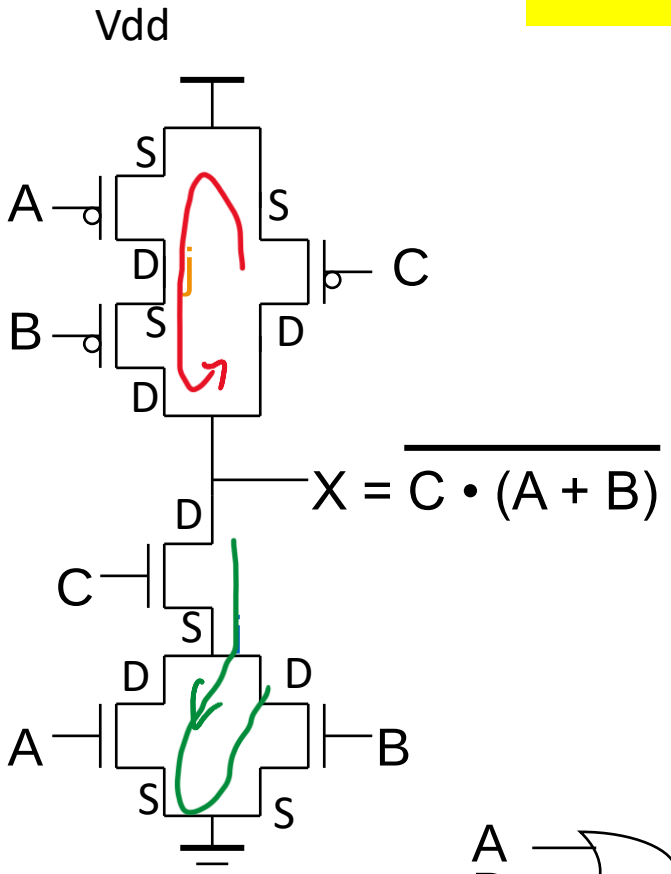



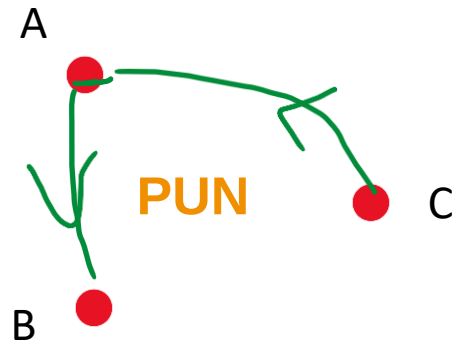

$V_{DD}$
X
Y
Z
W
D
S
S
D
S
D
S
D
Vout
D
S
D
S
S
D
D
S
Ground

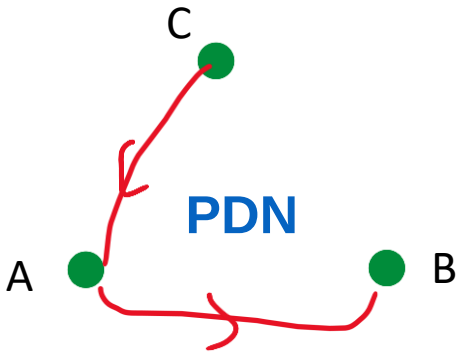


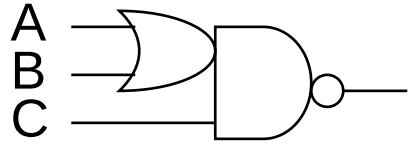
# Draw the stick diagram and layout graph of $X = \overline{C \cdot (A + B)}$

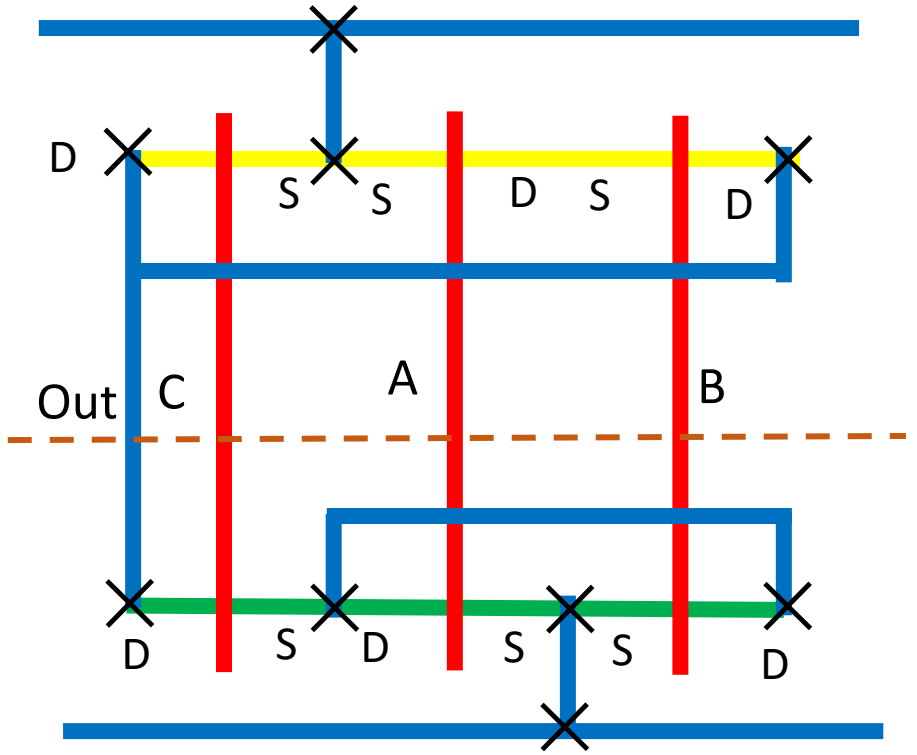
**Euler's path is C → A → B**

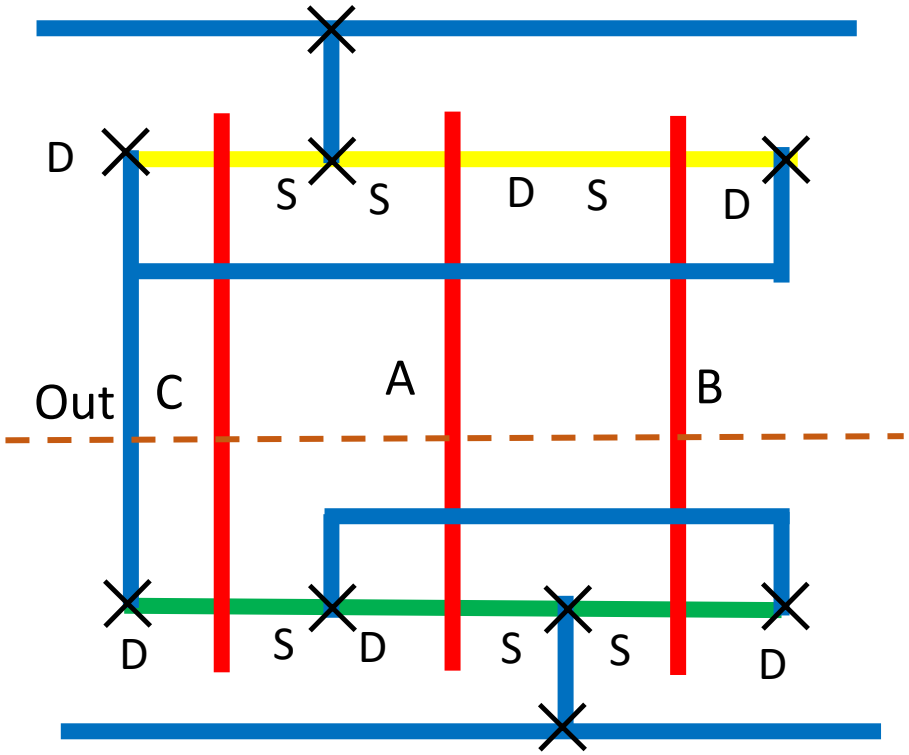


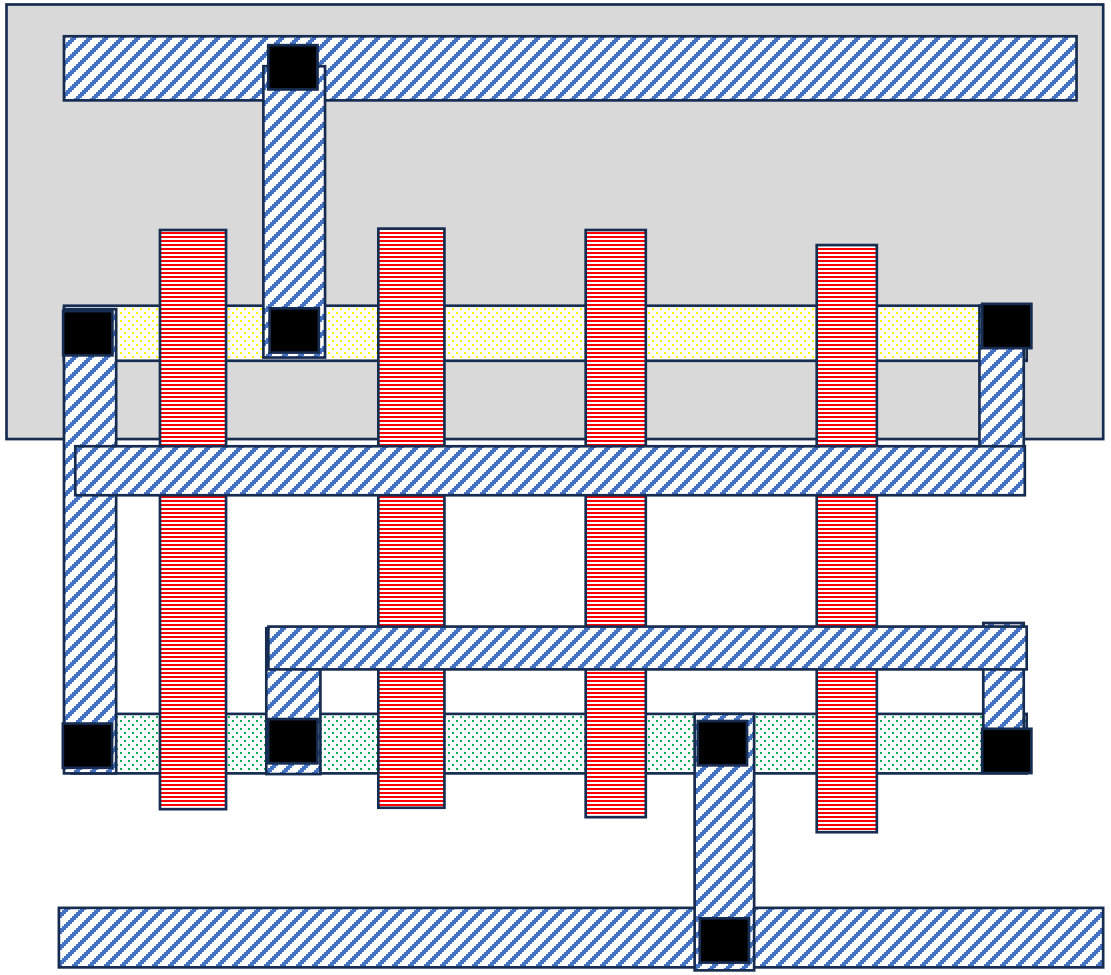



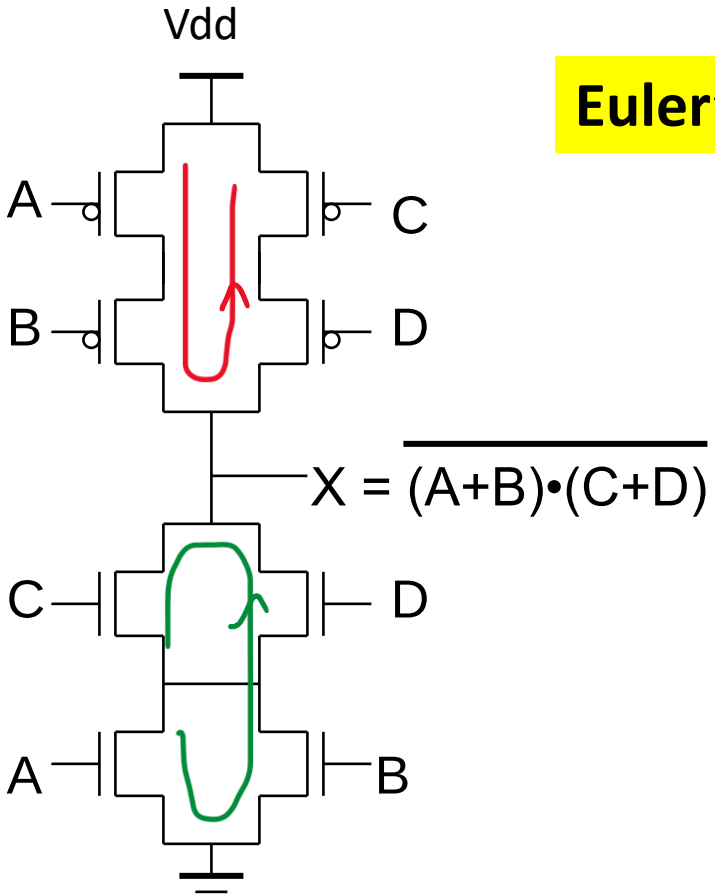




D
S
S
D
S
D
Out
C
A
B
D
S
D
S
S
D

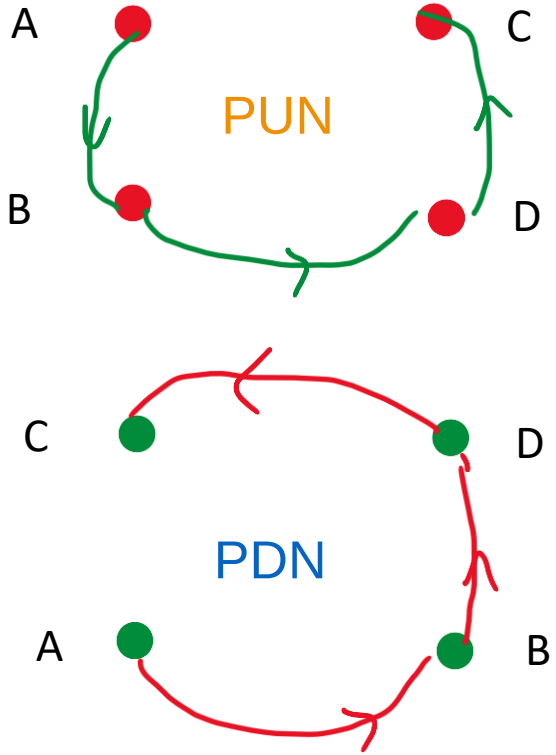


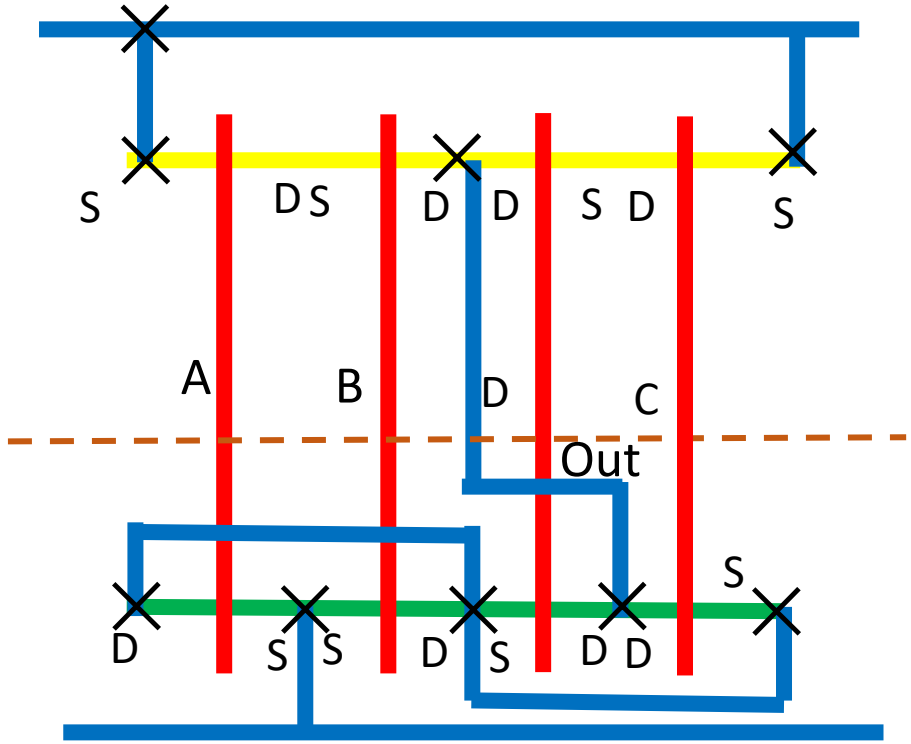
# Sketch the stick diagram and layout of OAI22

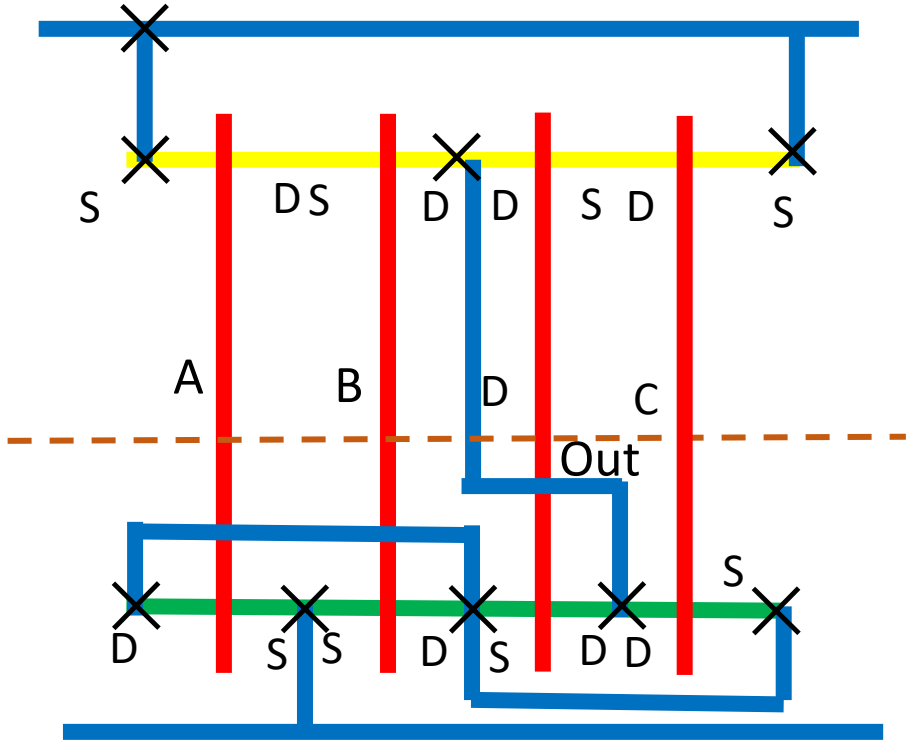


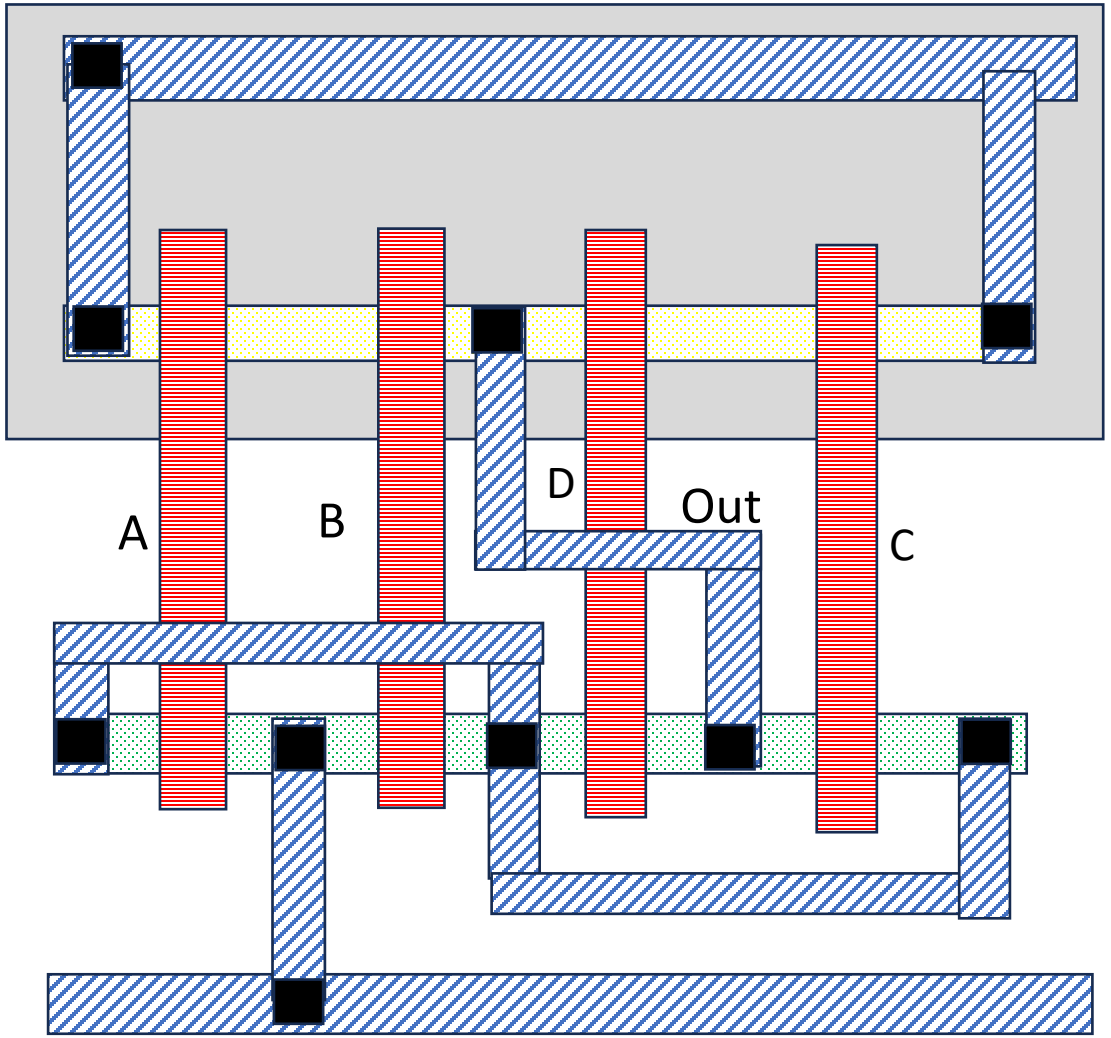

$$X = \overline{(A+B)\bullet(C+D)}$$

**Euler's path is A → B →D →C**

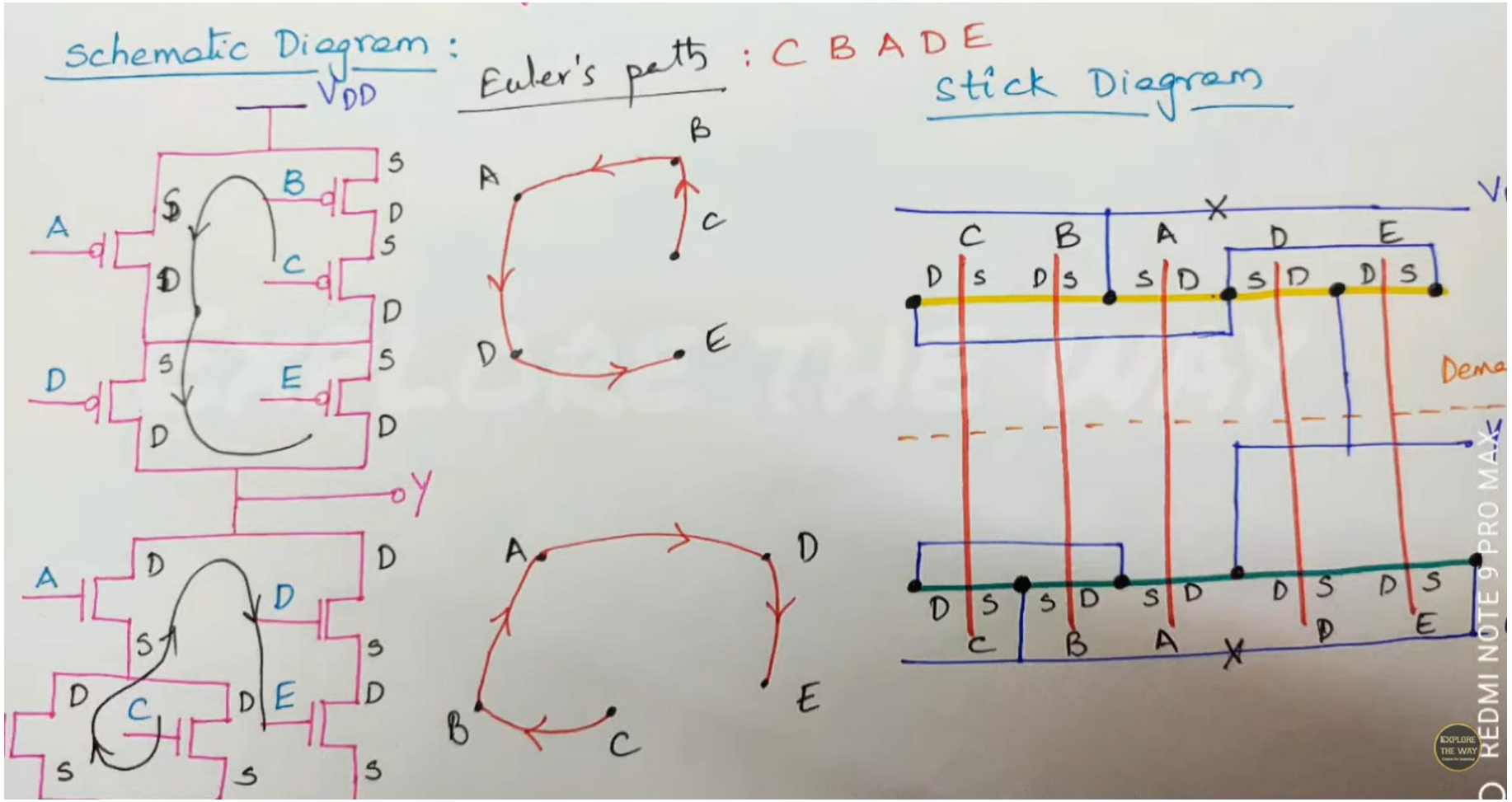


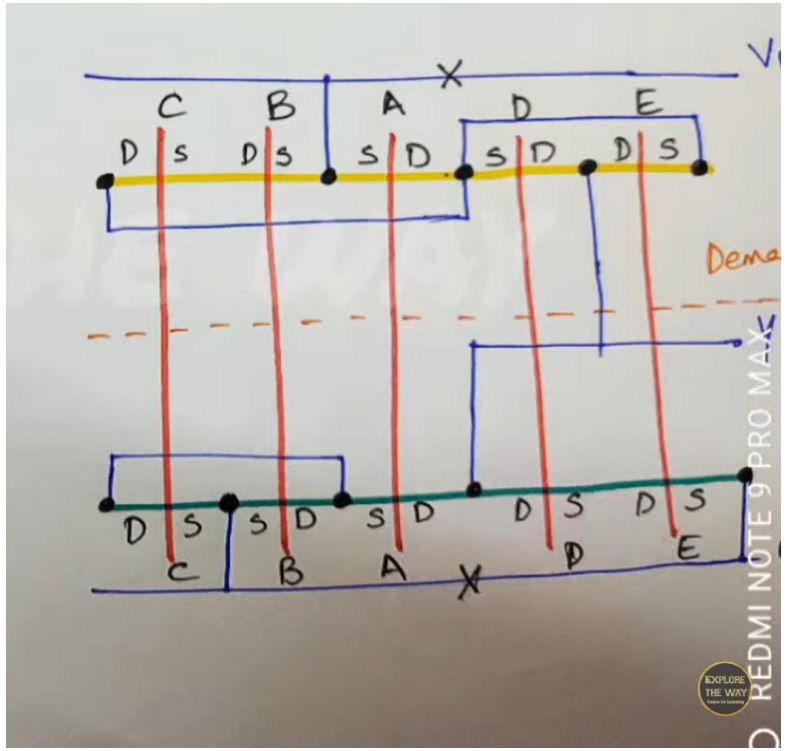



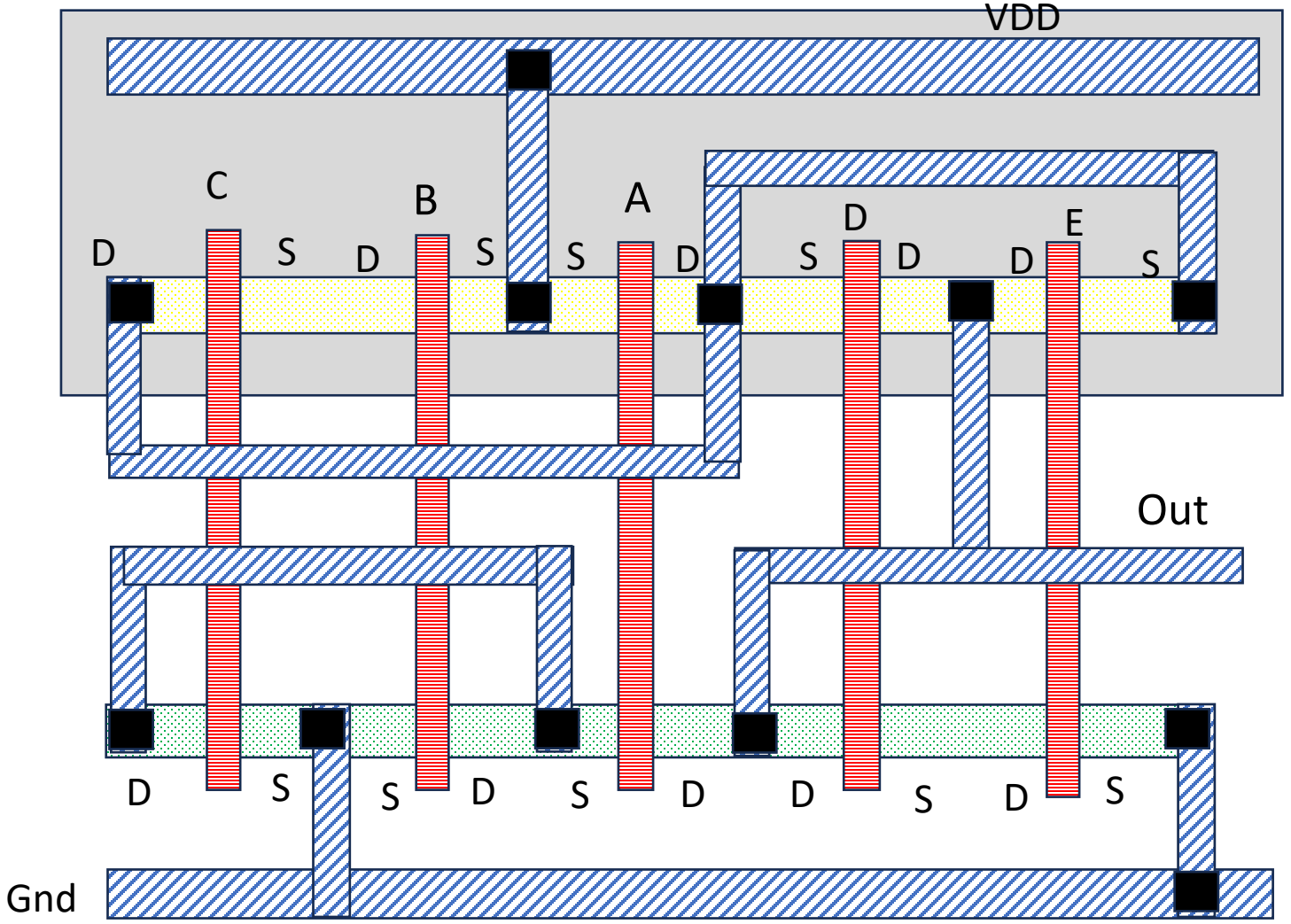

S
D S
D
D
S D
S
A
B
D
C
Out
D
S S
D
S
D D
S


A
B
D
Out
C

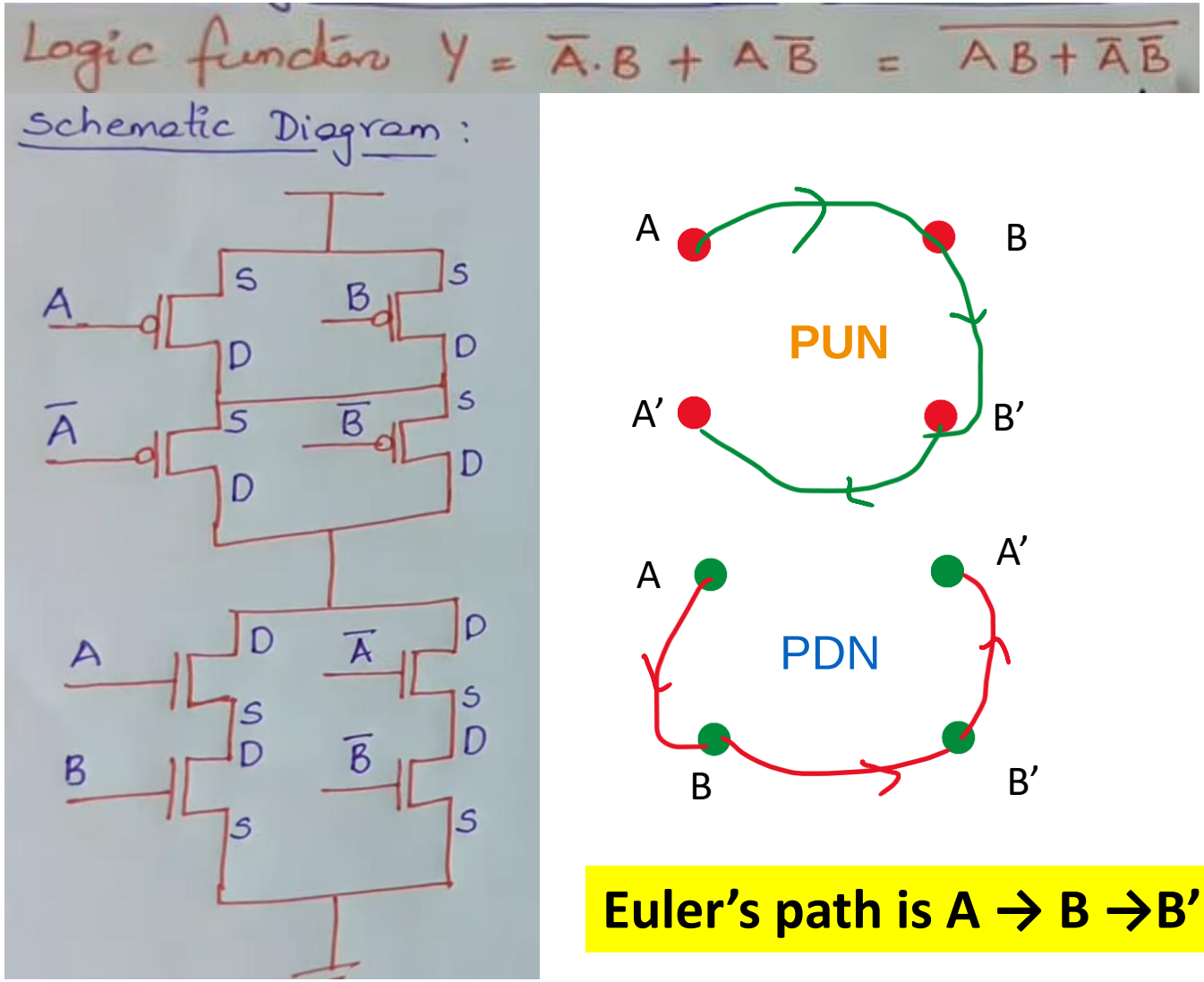
# Draw stick diagram and layout of y=(A(B+C)+DE)'




C
B
A
X
D
E
D S D S S D S D D S
D S S D S D D S D S
C B A X D E

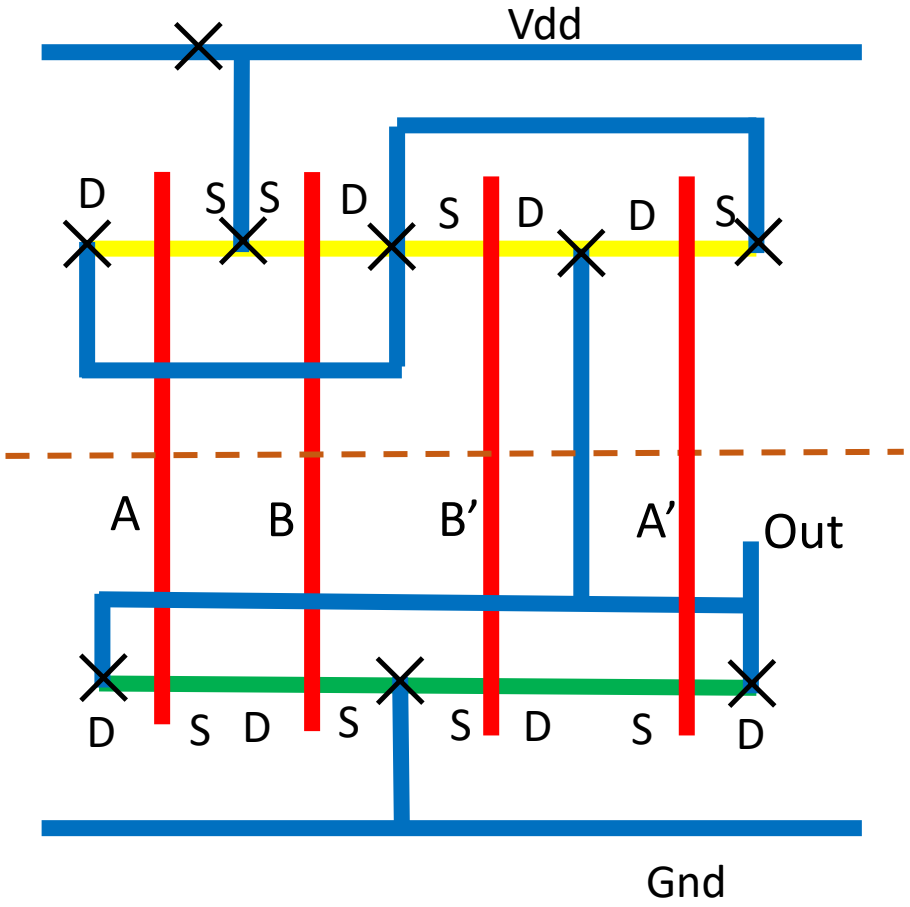

VDD
C
B
A
D
E
D S D S S S D S D D S
D S S D S D D S D S
Out
Gnd

# Draw stick diagram and layout of two input EXOR gate

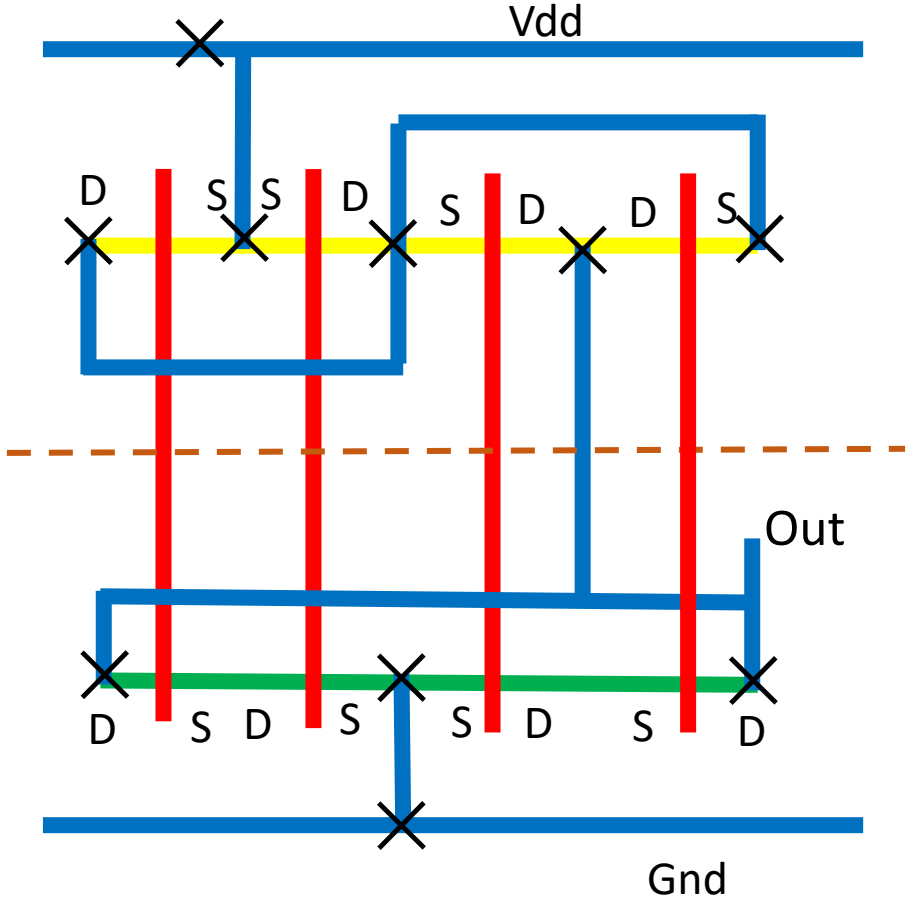


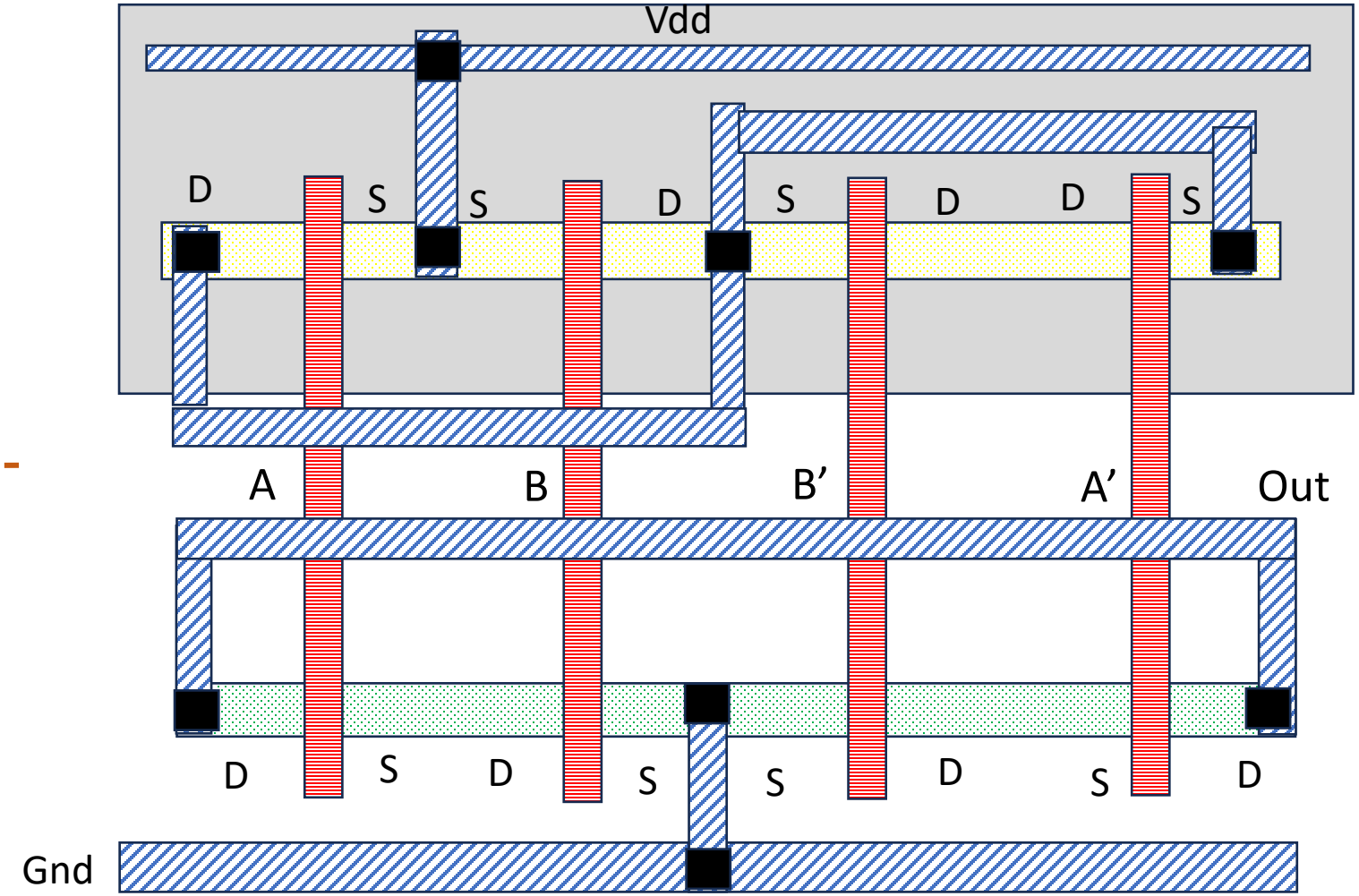




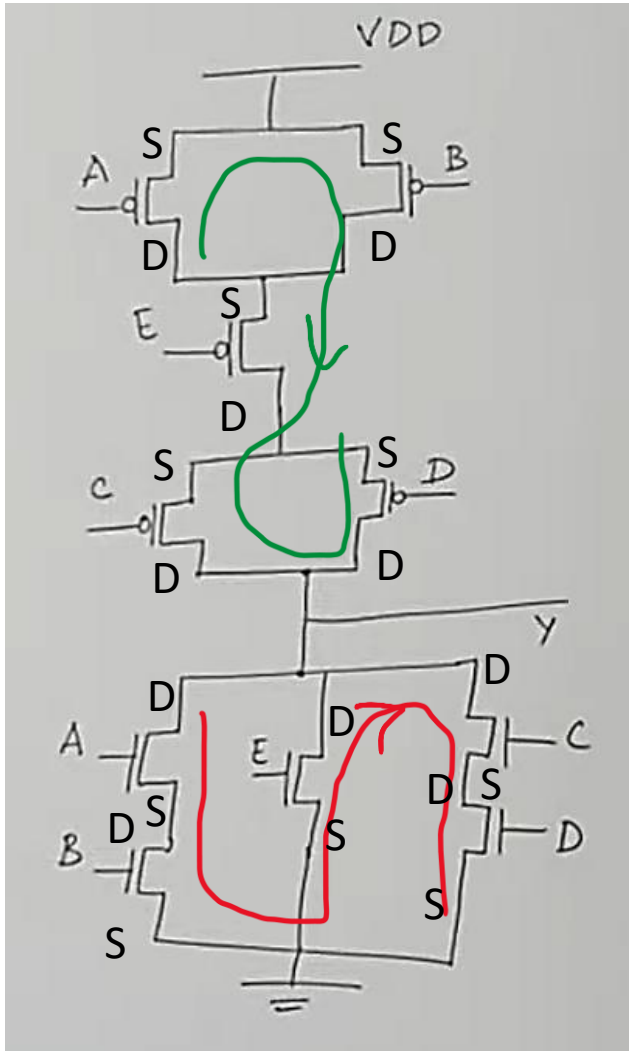


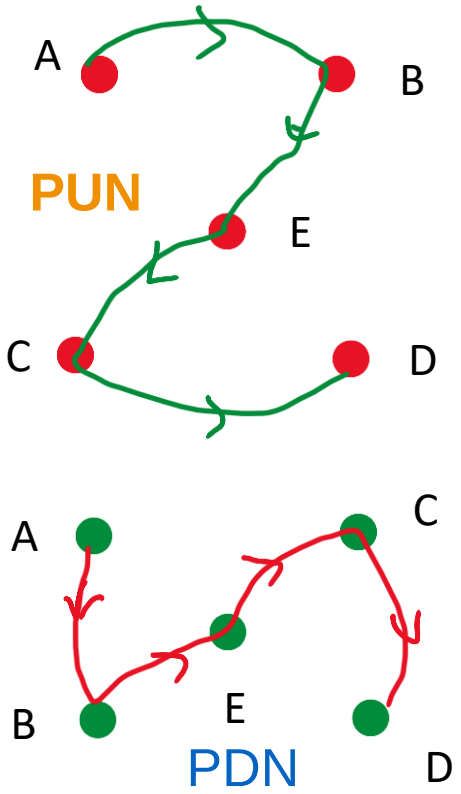

**Euler's path is A → B →B' →A'**

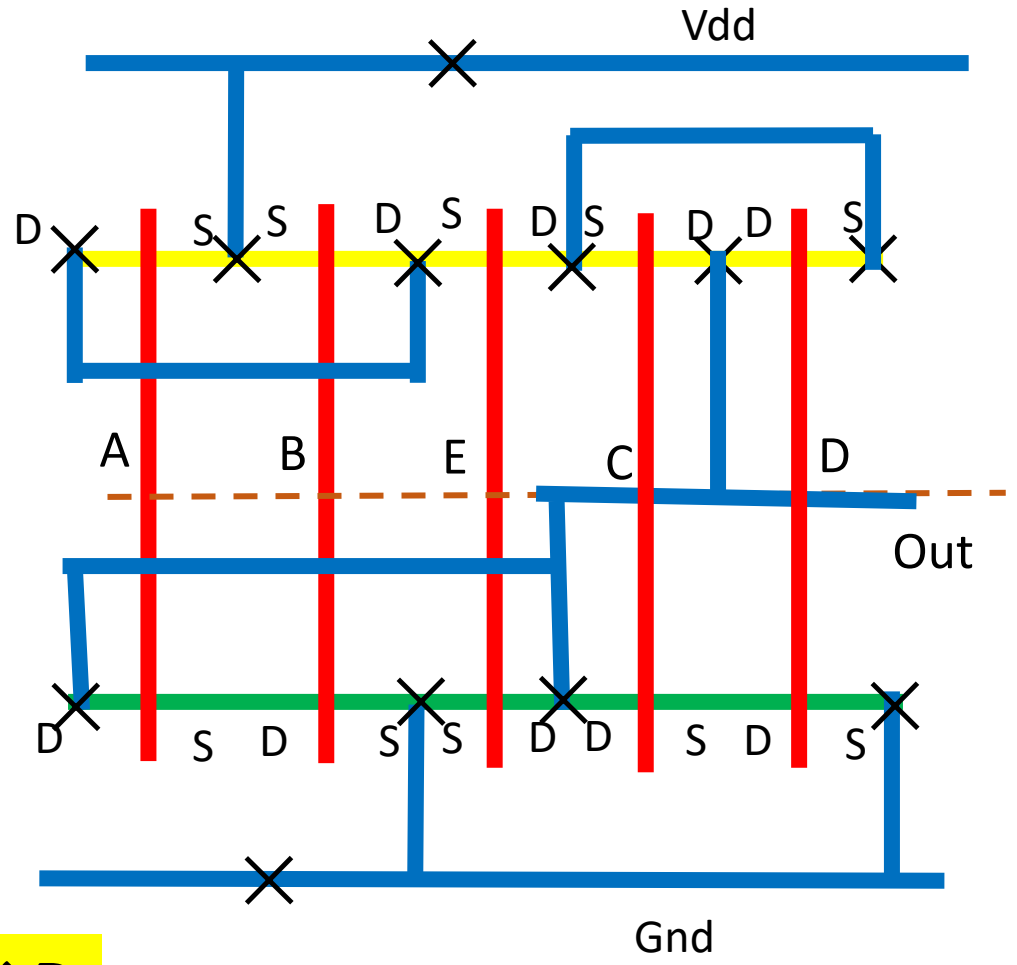

Vdd
D
S
S
D
S
D
D
S
Out
D
S
D
S
S
D
S
D
Gnd


Vdd
D
S
S
D
S
D
D
S
A
B
B’
A’
Out
D
S
D
S
S
D
S
D
Gnd

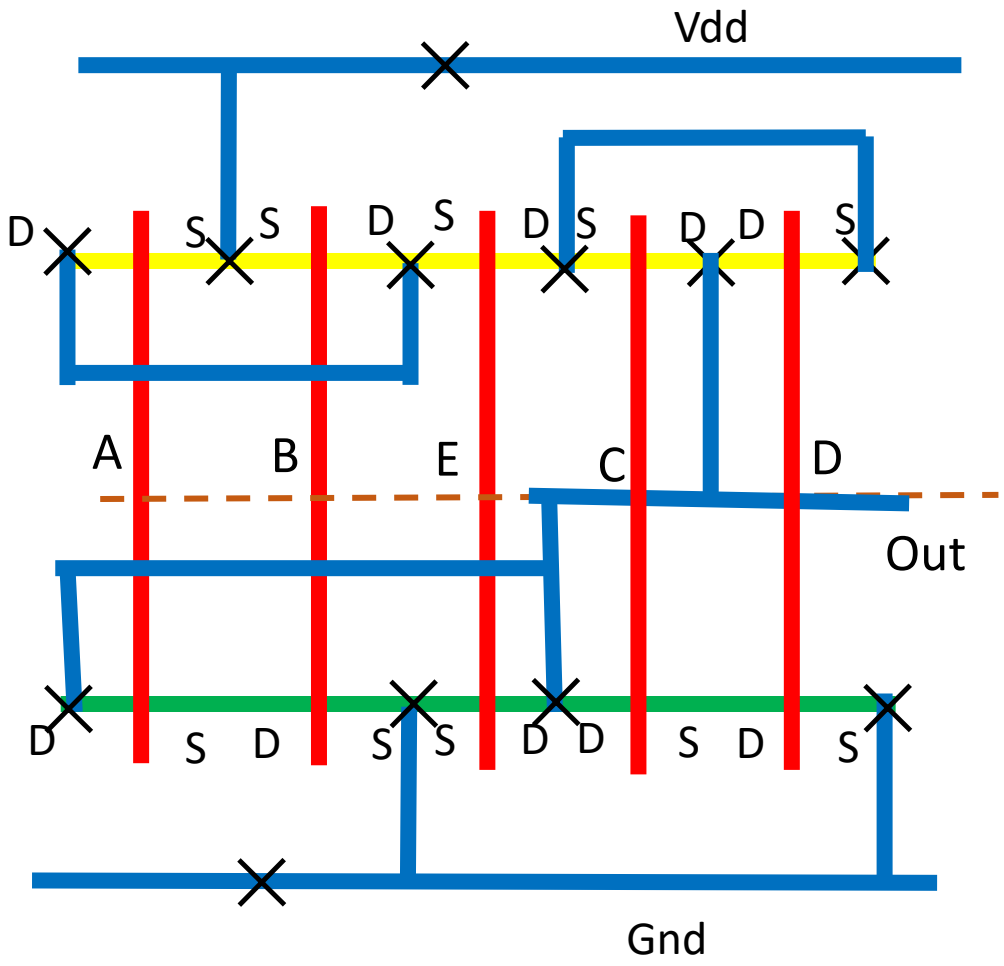
# Draw stick diagram and layout of y=(AB+E+CD)'

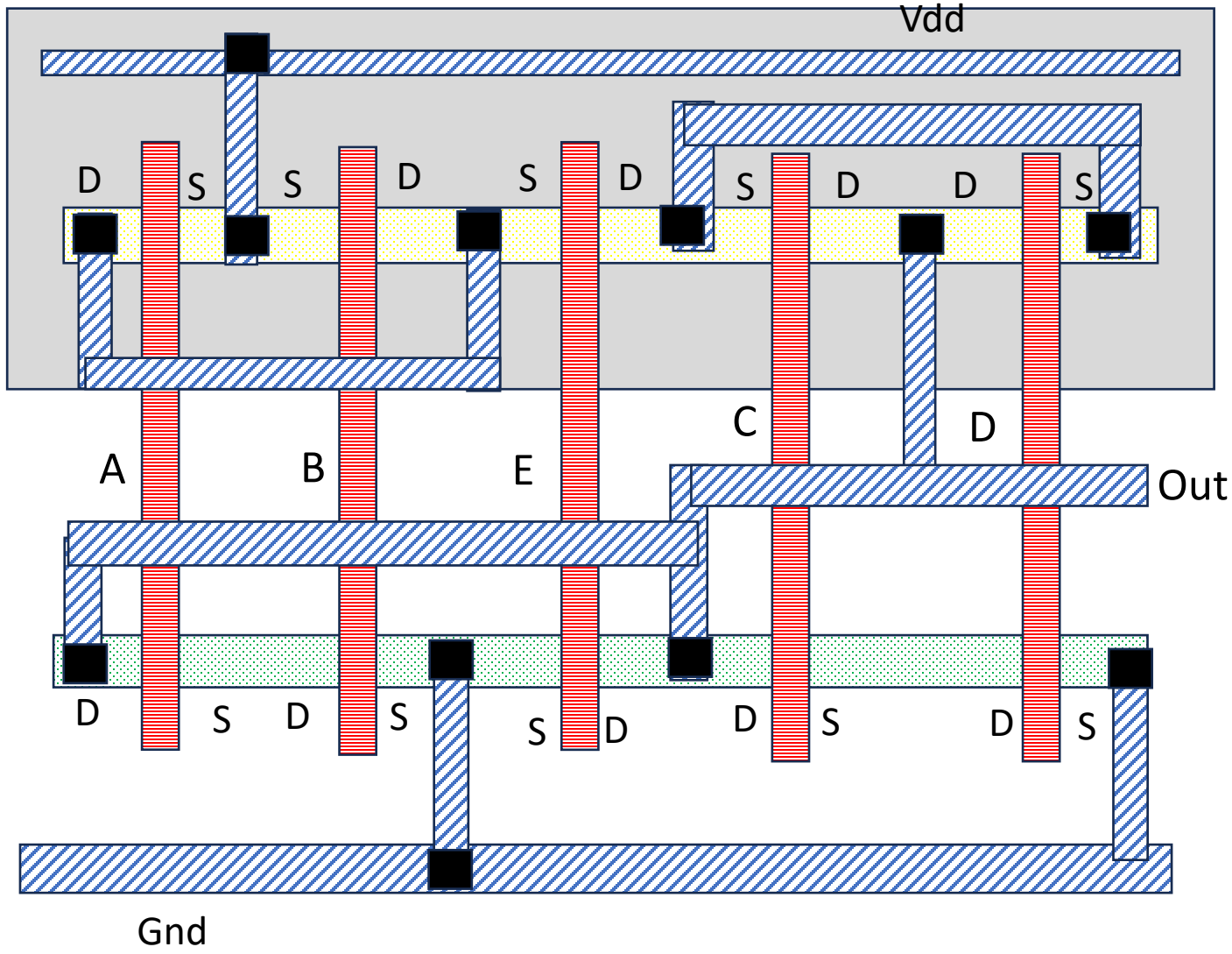


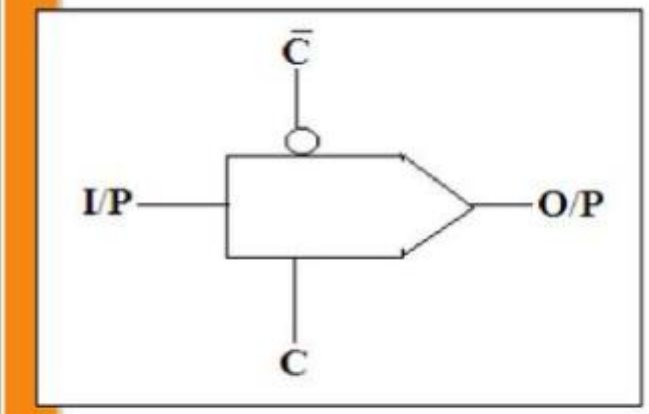




**Euler's path is A → B →E→C →D**

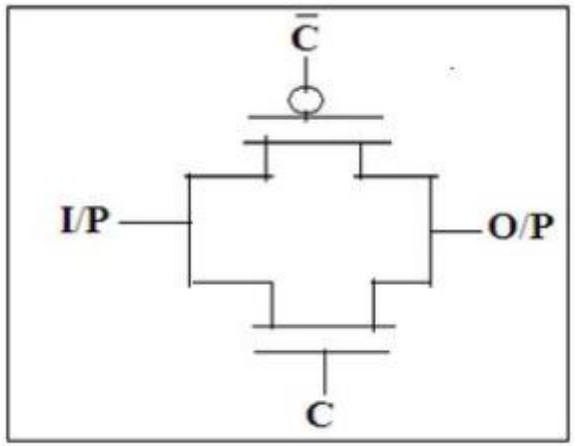

Vdd
D
S
S
D
S
D
S
D
D
S
A
B
E
C
D
Out
D
S
D
S
S
D
D
S
D
S
Gnd


Vdd
D
S
S
D
S
D
S
D
D
S
C
D
A
B
E
Out
D
S
D
S
S
D
D
S
D
S
Gnd

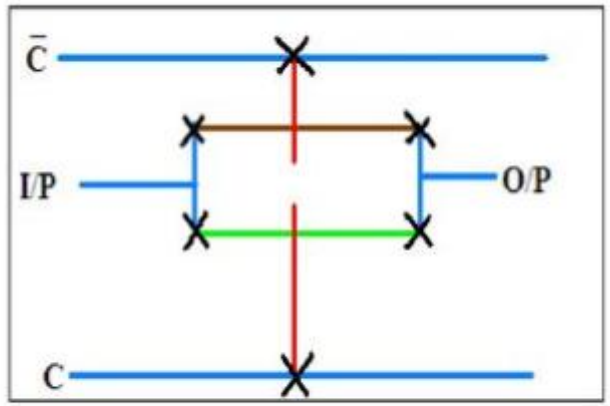
# Draw stick diagram and layout of transmission gate




Symbol

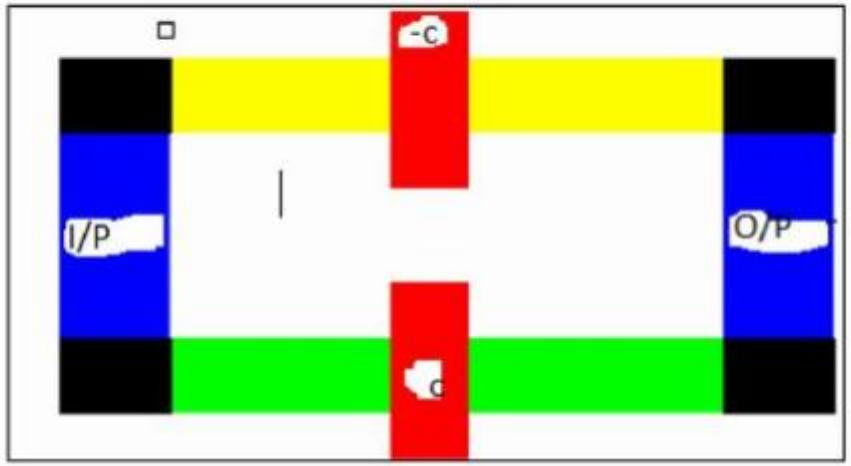



schematic

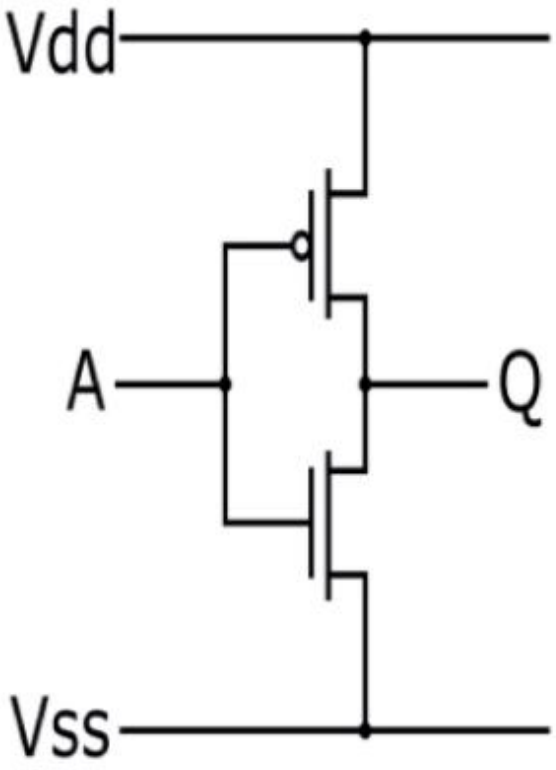



stick diagram

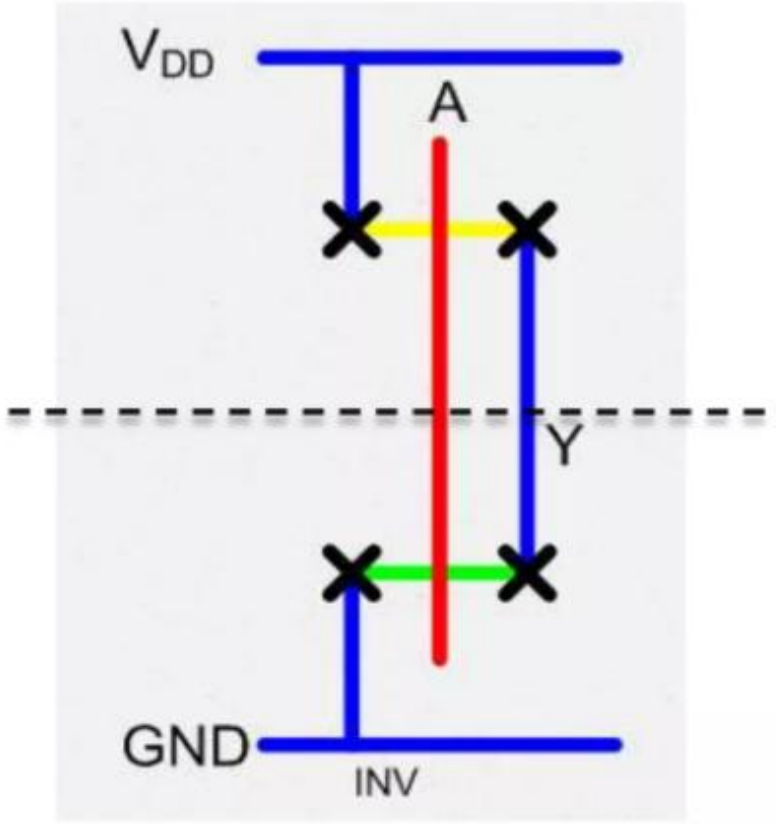
layout



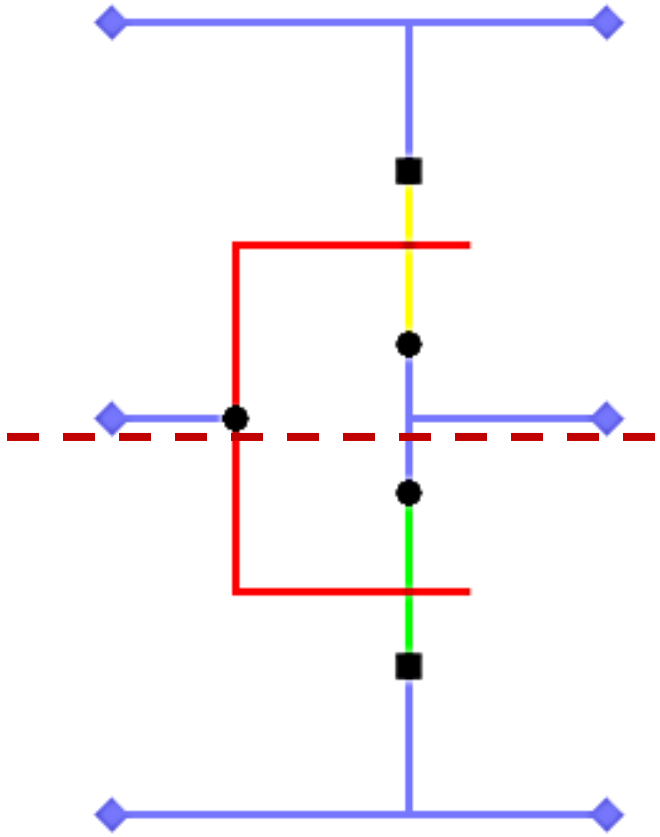
# FYI



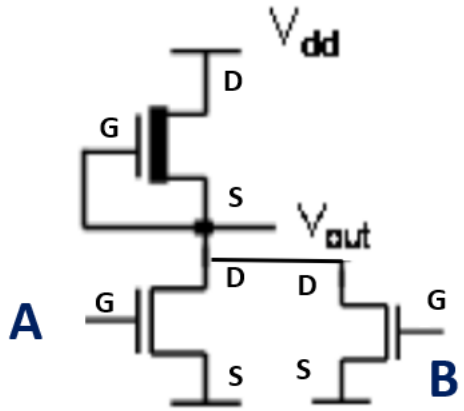

**Horizontal Placement**

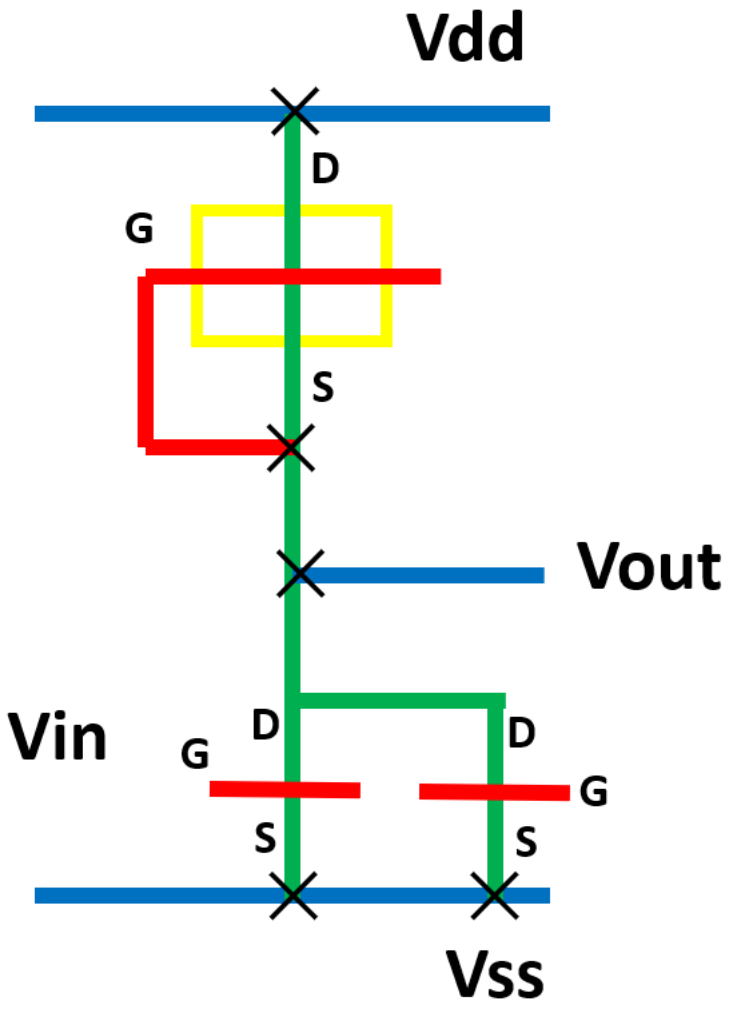



**Vertical Placement**



# Draw stick diagram and layout of nMOS depletion load NOR2



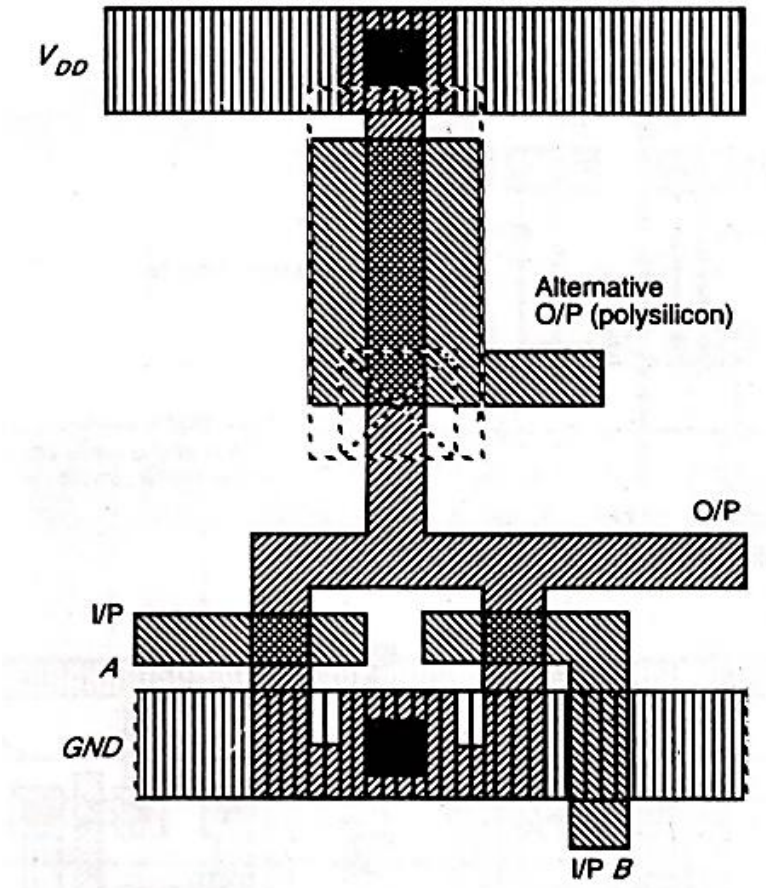
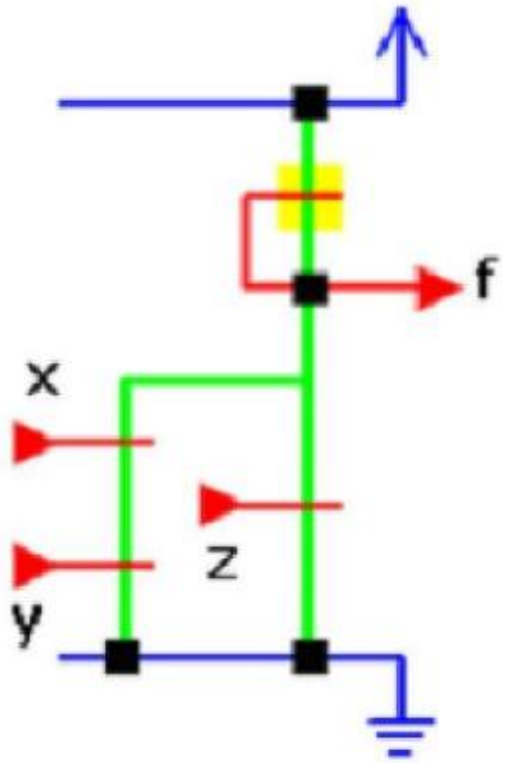




FIGURE 3.16 Two I/P nMOS *Nor* gate.

# Draw stick diagram and layout of nMOS depletion load F=(XY+Z)'




igure 7: stick diagram of a given function f.

**Draw layout**

# Sheet resistance ($R_s$)

Consider a uniform slab of conducting material of resistivity ρ, width W, thickness t, and length between faces L.

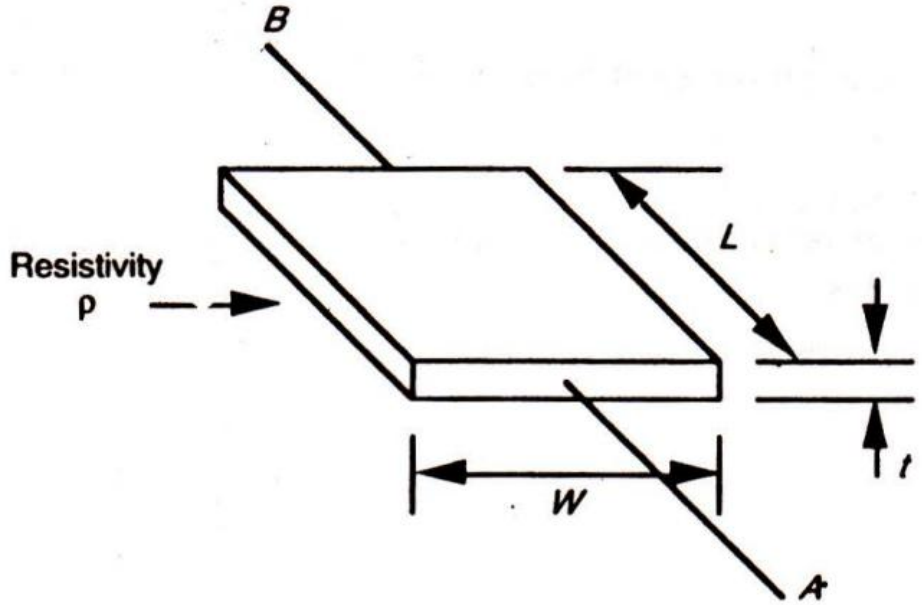



**FIGURE 4.1 Sheet resistance model.**

$$R_{AB} = \frac{\rho L}{A} \text{ ohm}$$

$$R_{AB} = \frac{\rho L}{tW} \text{ ohm}$$

Now, consider the case in which $L = W$, that is, a square of resistive material, then

$$R_{AB} = \frac{\rho}{t} = R_s$$

where

$R_s$ = ohm per square or sheet resistance

Thus

$$R_s = \frac{\rho}{t} \text{ ohm per square}$$

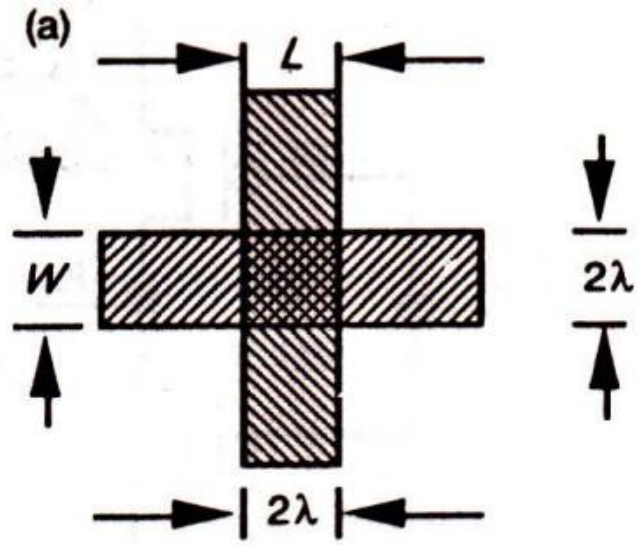
| *Layer* | *$R_s$ ohm per square* | | |
|---|---|---|---|
| | *5 μm* | 2μm *Orbit* | *Orbit 1.2 μm* |
| Metal | 0.03 | 0.04 | 0.04 |
| Diffusion (or active)** | 10→50 | 20→45 | 20→45 |
| Silicide | 2→4 | — | — |
| Polysilicon | 15→100 | 15→30 | 15→30 |
| n-transistor channel | $10^{4\dagger}$ | $2 \times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

# Sheet resistance concept applied to enhancement MOSFETs

The simple n-type pass transistor of Figure has a channel length L = 2 λ and a channel width W = 2λ. The channel is, therefore, square and channel resistance (with or without implant).

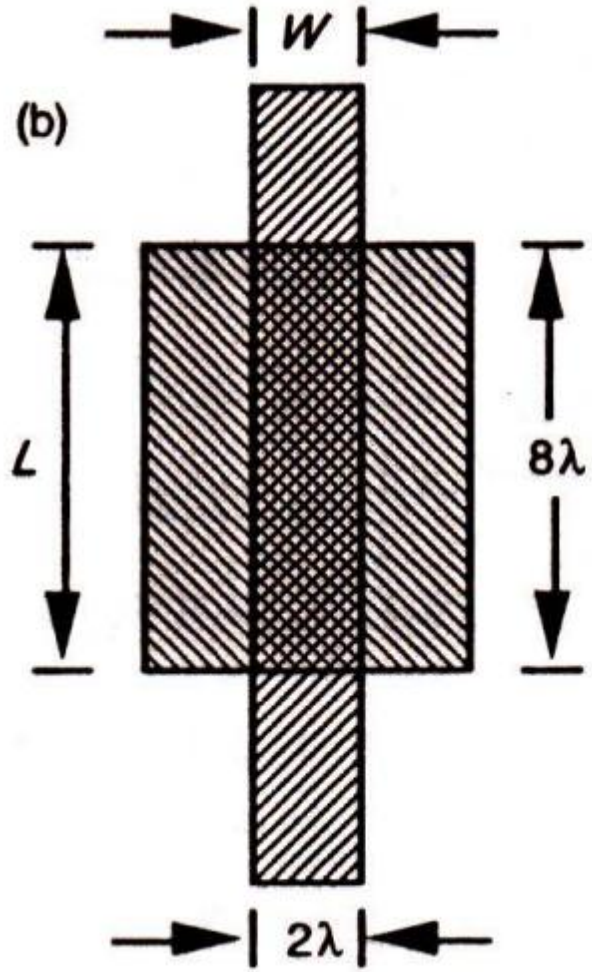
$$R = 1 \text{ square} \times R_s \frac{\text{ohm}}{\text{square}} = R_s = 10^4 \text{ ohm}^*$$

The length to width (L/W) ratio is 1:1 in this case.




| *Layer* | *$R_s$ ohm per square* | | |
|---|---|---|---|
| | *5 μm* | 2μm *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 | $2.5 \times 10^{4\dagger}$ | $4.5 \times 10^{4\dagger}$ | $4.5 \times 10^{4\dagger}$ |

# Sheet resistance concept applied to depletion MOSFETs



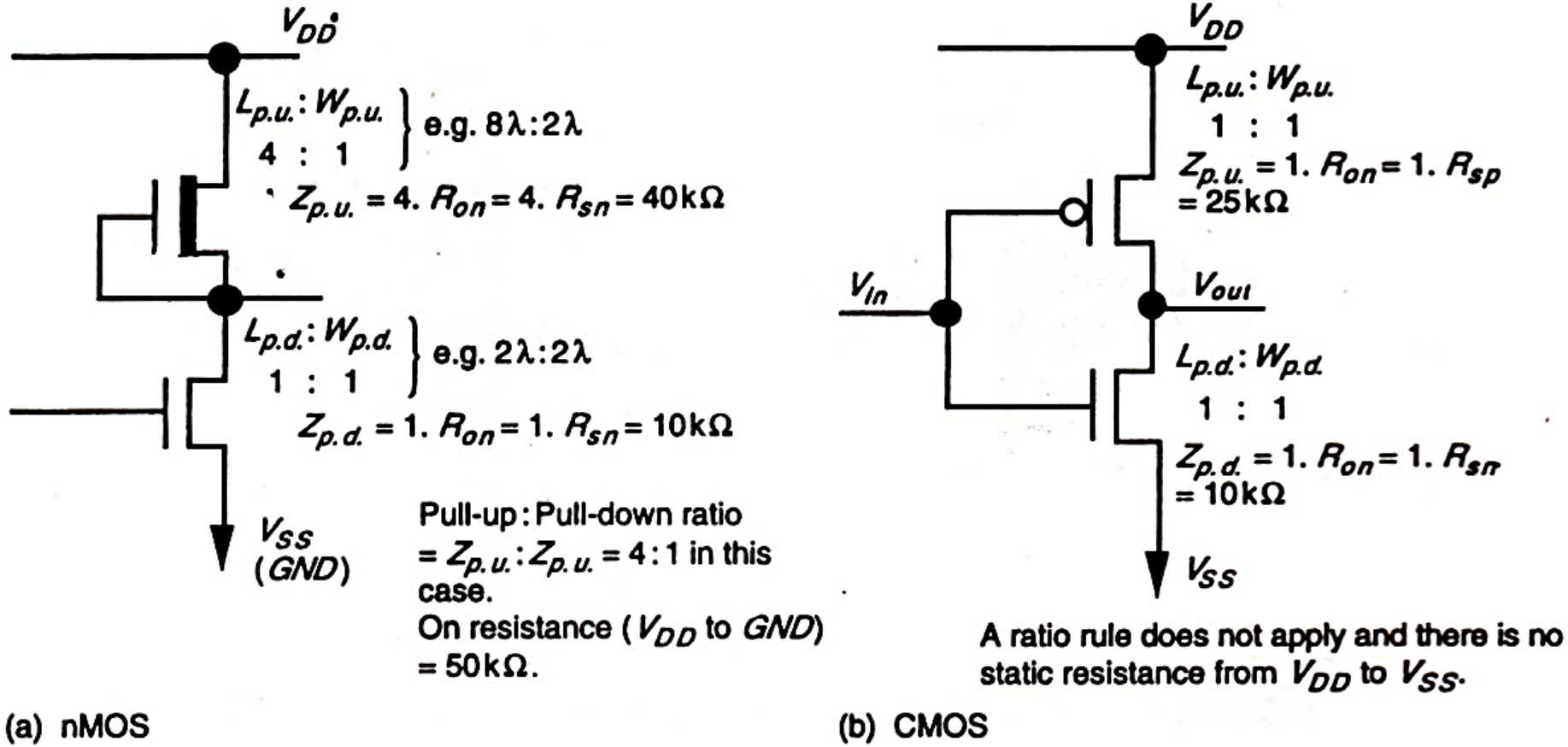

The transistor structure of ·Figure 4.2(b) has a channel length L = 8 λ. and width W = 2 λ. Therefore,

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

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

| *Layer* | *$R_s$ ohm per square* | | |
|---|---|---|---|
| | *5 μm* | 2μm *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 | $2.5 \times 10^{4\dagger}$ | $4.5 \times 10^{4\dagger}$ | $4.5 \times 10^{4\dagger}$ |

# Sheet resistance concept applied to inverters




*Note*: $R_{on}$ = 'on' resistance; $R_{sn}$ = n-channel sheet resistance; $R_{sp}$ = p-channel sheet resistance.

**FIGURE 4.3 Inverter resistance calculation.**

# Area Capacitances of layers

For any layer, knowing the dielectric (silicon dioxide) thickness, we can calculate area capacitance as follows:

$$C = \frac{\varepsilon_0 \varepsilon_{ins} A}{D} \text{ farads}$$

A normal approach is to give layer area capacitances in pF/μm$^2$

**TABLE 4.2** Typical area capacitance values for MOS circuits

| *Capacitance* | *Value in pF × $10^{-4}$/μm$^2$ (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) |

# STANDARD UNIT OF CAPACITANCE $\square C_g$

The unit is denoted □Cg and is defined the gate-to-channel capacitance of a MOS transistor having W = L = feature size, that is, a 'standard' or 'feature size' square

$\square C_g$ may be evaluated for any MOS process. For example, for 5 μm MOS circuits:

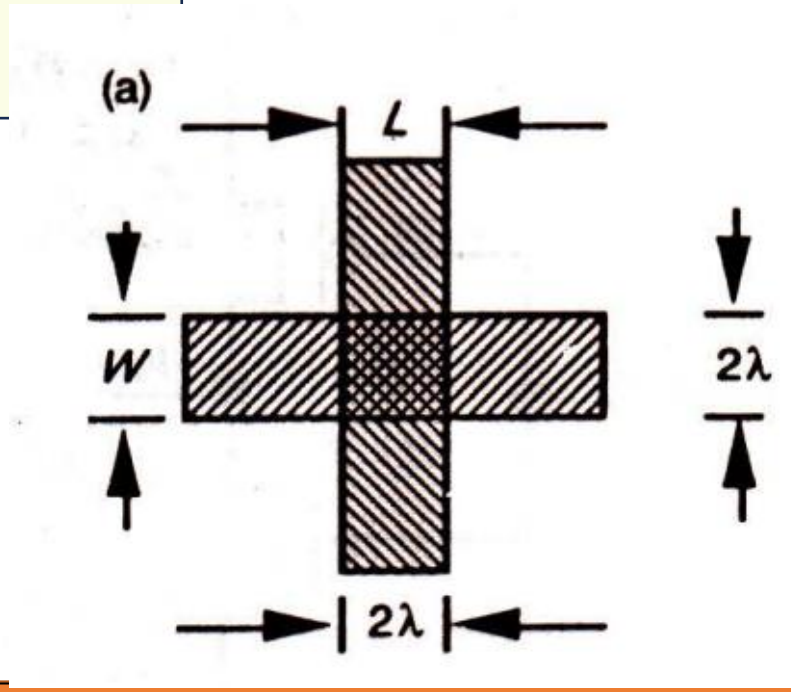
Area/standard square = 5 μm × 5 μm = 25 $\mu m^2$ (= area of minimum size transistor)
Capacitance value (from Table 4.2) = $4 \times 10^{-4}$ $pF/\mu m^2$
Thus, standard value $\square C_g = 25\ \mu m^2 \times 4 \times 10^{-4}\ pF/\mu m^2 = .01$ pF

or, for 2 μm MOS circuits (Orbit):

Area/standard square = 2 μm × 2 μm = 4 $\mu m^2$
Gate capacitance value (from Table 4.2) = $8 \times 10^{-4}$ $pF/\mu m^2$
Thus, standard value $\square C_g = 4\ \mu m^2 \times 8 \times 10^{-4}\ pF/\mu m^2 = .0032$ pF

and, for 1.2 μm MOS circuits (Orbit):

Area/standard square = 1.2 μm × 1.2 μm = 1.44 $\mu m^2$
Gate capacitance value (from Table 4.2) = $16 \times 10^{-4}$ $pF/\mu m^2$
Thus, standard value $\square C_g = 1.44\ \mu m^2 \times 16 \times 10^{-4}\ pF/\mu m^2 = .0023$ pF



# Calculation of Area Capacitance

- 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 x 2λ) and multiplying this ratio by the appropriate relative C value from Table 4.2.
- The product will give the required capacitance in □Cg units.

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

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

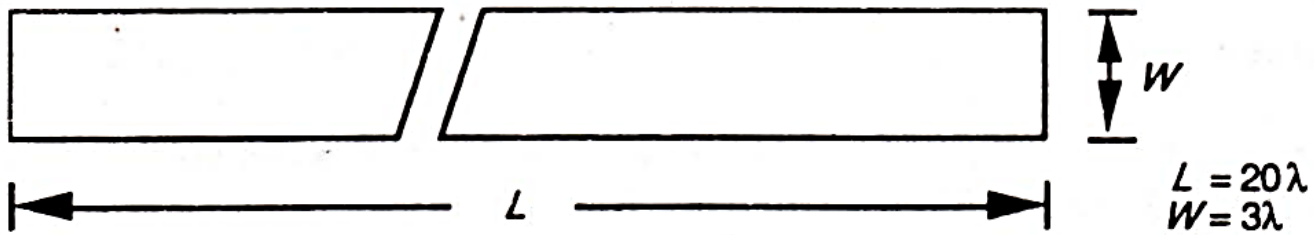



**FIGURE 4.4 Simple area for capacitance calculation.**

**TABLE 4.2** 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) |

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 $= 15 \times 0.25 \square C_g$
   $$= 3.75 \square C_g^*$$

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

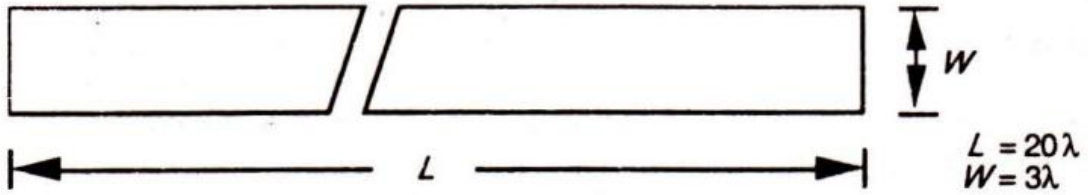



**FIGURE 4.4** Simple area for capacitance calculation.

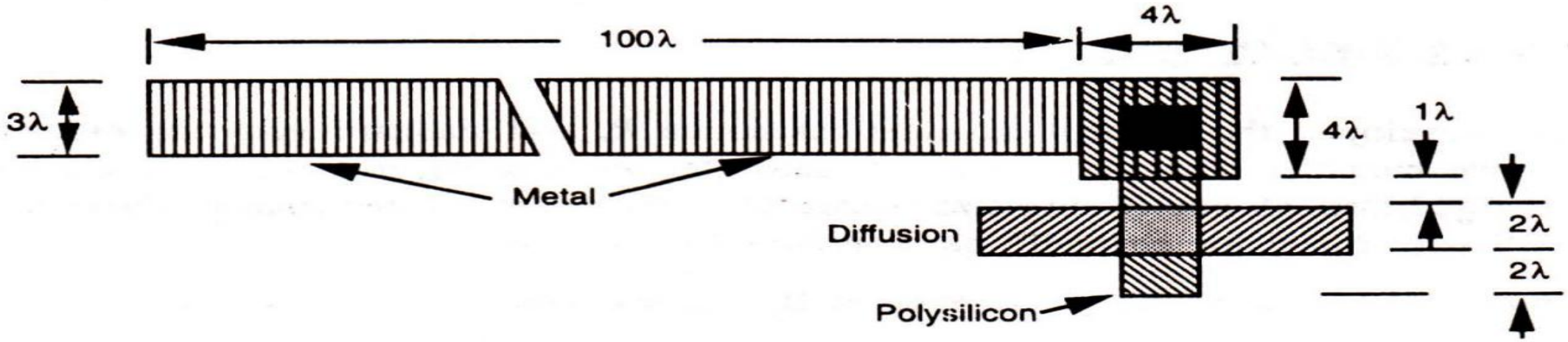



**FIGURE 4.5 Capacitance calculation (multilayer).**

Calculations of area capacitance values associated with structures occupying more than one layer, as in Figure 4.5,

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

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

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

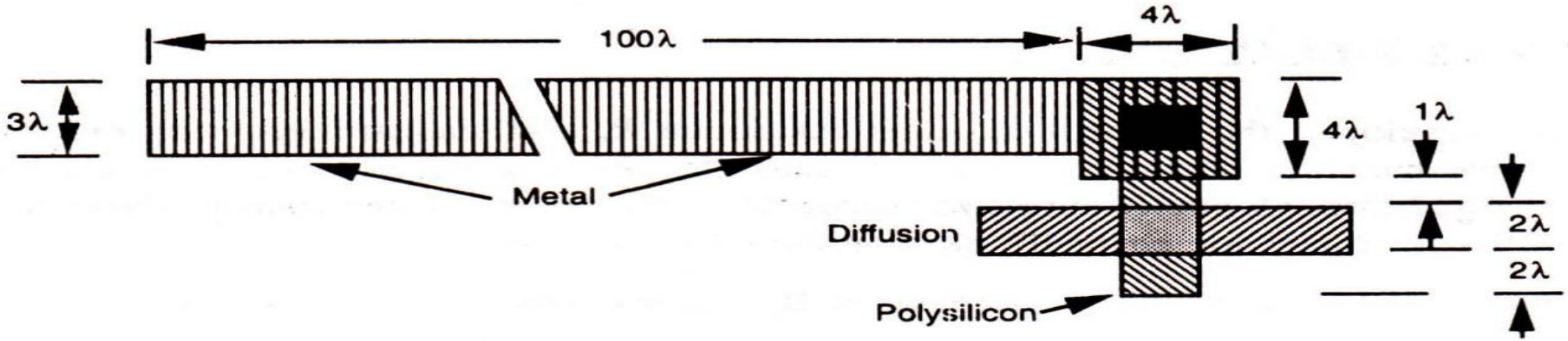



**FIGURE 4.5 Capacitance calculation (multilayer).**

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,

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

$$\textbf{Total capacitance } C_T = C_m + C_p + C_g \doteq 7.20\square C_g$$

# The Delay Unit $\tau$

- We have developed the concept of sheet resistance $R_s$ and standard gate capacitance unit $\square C_g$

- **Time constant**: It is time required to charge one standard (feature size square) gate area capacitance through one feature size square of n channel resistance (that is, through $R_s$ for an nMOS pass transistor channel),

$$\text{Time constant } \tau = (1R_s \text{ (n channel)} \times 1\square C_g) \text{ seconds}$$

If we consider the case of one standard (feature size square) gate area capacitance being charged through one feature size square of n channel resistance (that is, through Rs for an nMOS pass transistor channel), as in Figure 4.6, we have:

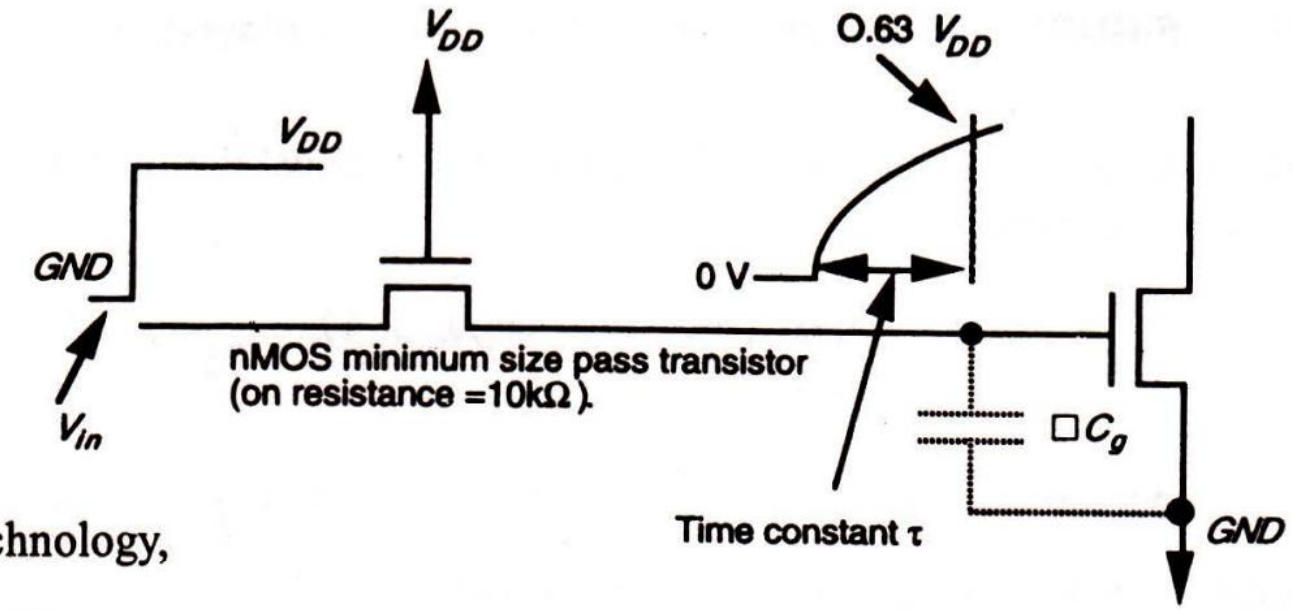



**FIGURE 4.6 Model for derivation of τ.**

This can be evaluated for any technology and for 5 μm technology,

$$\tau = 10^4 \text{ ohm} \times 0.01 \text{ pF} = 0.1 \text{ nsec}$$

and for 2 μm (Orbit) technology,

$$\tau = 2 \times 10^4 \text{ ohm} \times 0.0032 \text{ pF} = 0.064 \text{ nsec}$$

and for 1.2 μm (Orbit) technology,

$$\tau = 2 \times 10^4 \text{ ohm} \times 0.0023 \text{ pF} = 0.046 \text{ nsec}$$

Note that $\tau$ thus obtained is not much different from transit time $\tau_{sd}$ calculated from equation (2.2).

$$\tau_{sd} = \frac{L^2}{\mu_n V_{ds}}$$

Note that $V_{ds}$ varies as $C_g$ charges from 0 volts to 63% of $V_{DD}$ in period $\tau$ in Figure 4.6, so that an appropriate value for $V_{ds}$ is the average value = 3 volts. For 5 μm technology, then,

$$\tau_{sd} = \frac{25\,\mu m^2\ V\ sec}{650\ cm^2\ 3V} \times \frac{10^9\ nsec\ cm^2}{10^8\ \mu m^2}$$

$$= 0.13\ nsec$$

This is very close to the theoretical time constant $\tau$ calculated above.

# Estimation of delay in NMOS and CMOS inverters

# nMOS inverter delays

In general, the delay through a pair of similar nMOS inverters is

$$T_d = (1 + Z_{p.u.}/Z_{p.d.})\tau$$

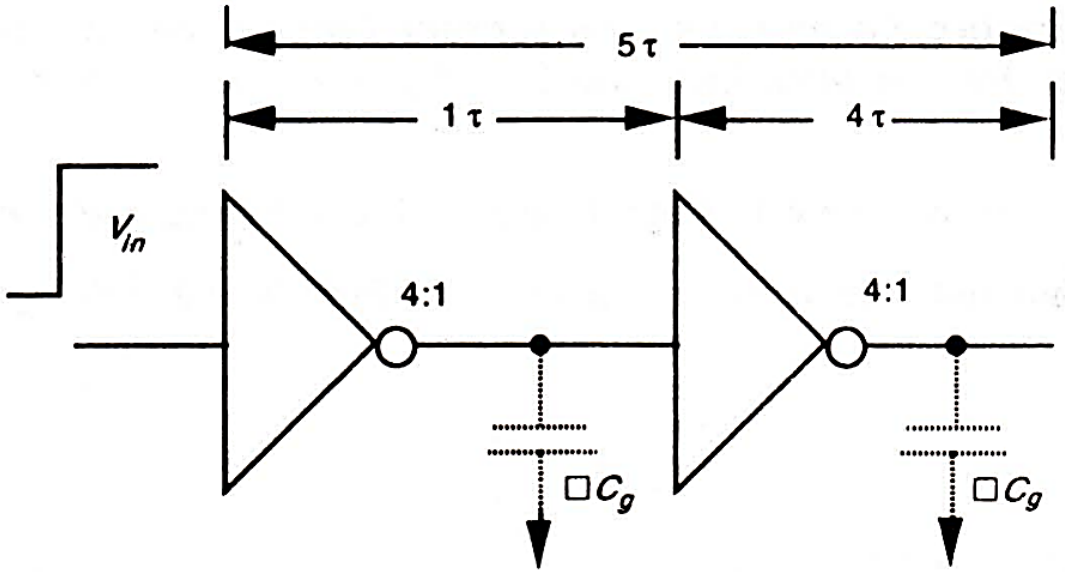



**FIGURE 4.7 nMOS inverter pair delay.**

INV-1

When $V_{in} = 1$,

enh – ON

$\tau_1 = 1 R_s \times 1 \square C_g$

$z = 1$

$\tau_1 = R_s \square C_g$

$\tau_1 = 1\tau$

INV-2

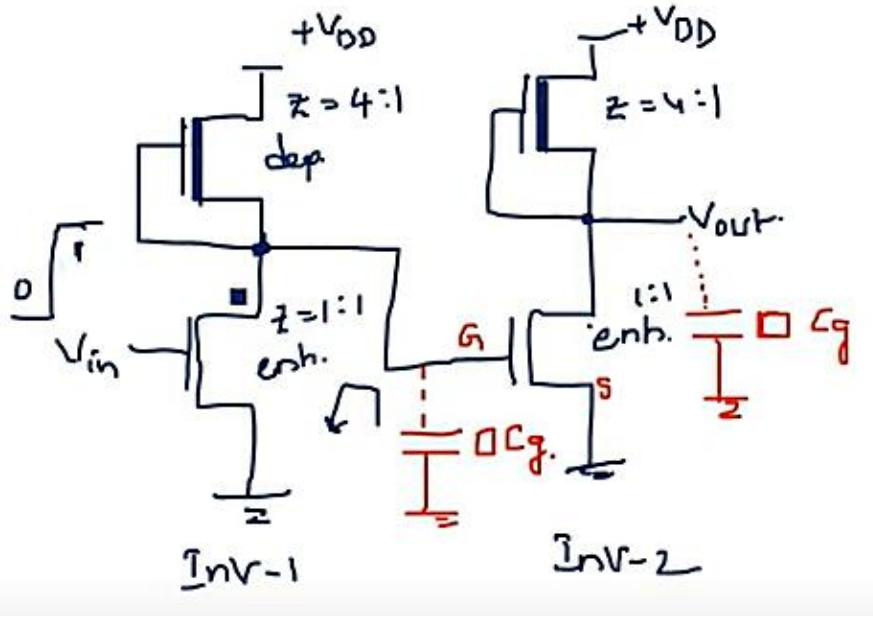
When $V_{in} = 0$ V. then enhancement trans. is in OFF state. then depletion mode comes into action.

$\tau_2 = z \cdot R_s \times 1 \square C_g$

$\simeq 4 R_s * 1 \square C_g$

$\simeq 4 \cdot R_s \square C_g$

$\blacksquare = 4\tau$



# CMOS inverter delays

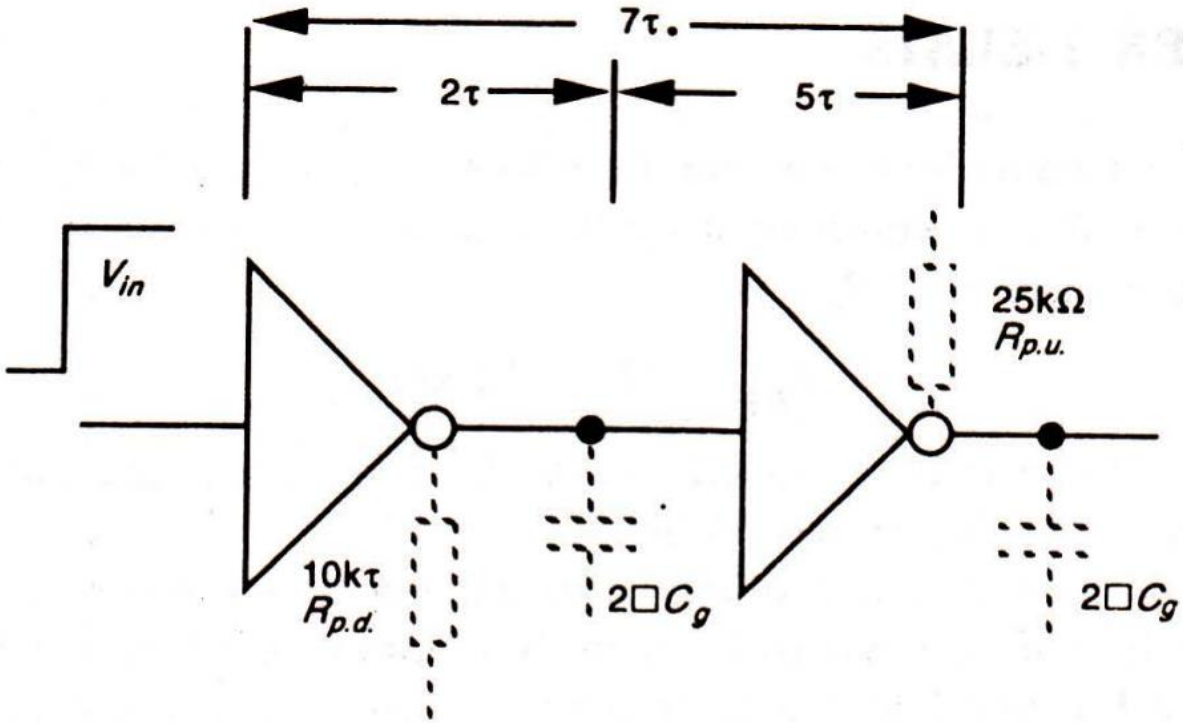



**FIGURE 4.8 Minimum size CMOS inverter pair delay.**

1. For asymmetric CMOS inverter, assuming the width of nMOS and pMOS to be equal, the input gate capacitance will be $2\square C_g$

2. For symmetric CMOS inverter, i.e. for $\beta_n = \beta_p$, the input capacitance is,

$$1\square C_g \text{ (n-device)} + 2.5\square C_g \text{(p-device)} = 3.5\square C_g$$

# A more formal estimation of CMOS Inverter delay

A CMOS inverter, in general, either charges or discharges a capacitive load $C_L$ and rise-time $\tau_r$ or fall-time $\tau_f$ can be estimated from the following simple analysis.

$$\beta = \frac{\mu C_{ox} W}{L}$$

## 1. 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

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

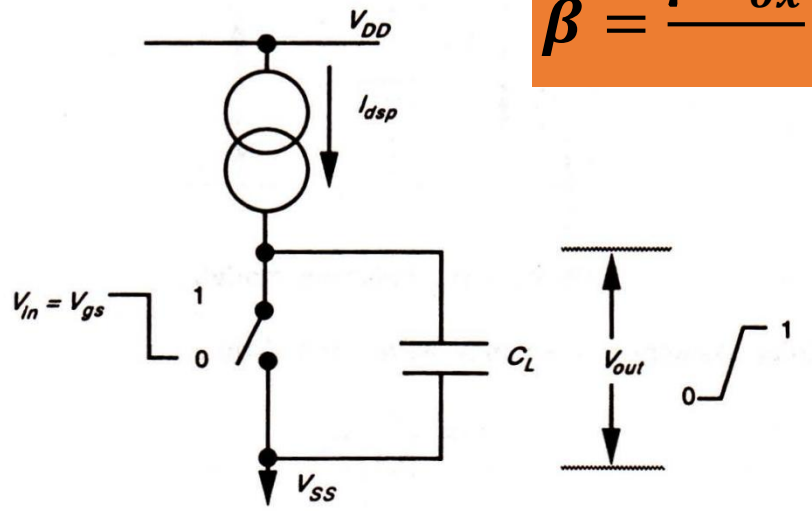
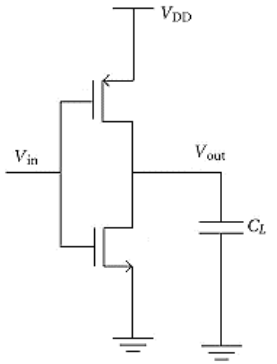



FIGURE 4.9 Rise-time model.

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

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



$$\beta = \frac{\mu C_{ox} W}{L}$$

Substituting for $I_{dsp}$ and rearranging we have

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

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

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

# 2. Fall time estimation

Similar reasoning can be applied to the discharge of $C_L$ through the n-transistor. The circuit model in this case is given as Figure 4.10.

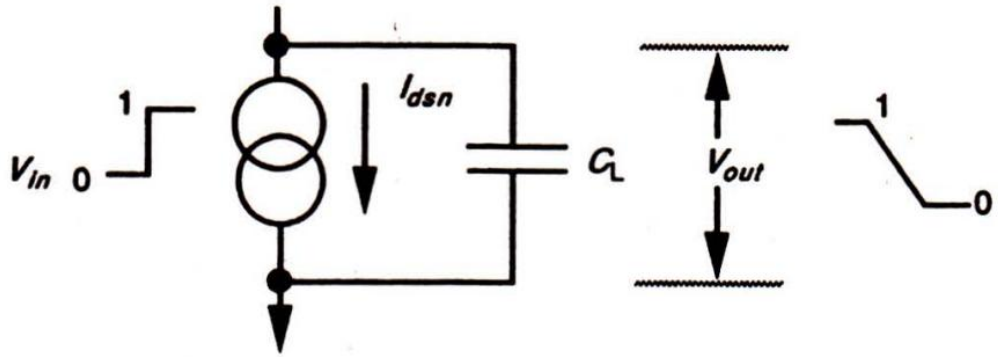



**FIGURE 4.10 Fall-time model.**

Making similar assumptions we may write for fall-time:

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

### 4.7.1.3 Summary of CMOS rise and fall factors

$$\beta = \frac{\mu C_{ox} W}{L}$$

Using these expressions we may deduce that:

$$\frac{\tau_r}{\tau_f} = \frac{\beta_n}{\beta_p}$$

But $\mu_n = 2.5\ \mu_p$ and hence $\beta_n \doteqdot 2.5\beta_p$, so that the rise-time is slower by a factor of 2.5 when using minimum size devices for both 'n' and 'p'.

In order to achieve symmetrical operation using minimum channel length, we would need to make $W_p = 2.5W_n$ and for minimum size lambda-based geometries this would result in the inverter having an input capacitance of $1\square C_g$ (n-device) + $2.5\square C_g$(p-device) = $3.5\square C_g$ in total.

# Driving Large Capacitive loads

- The problem of driving comparatively large capacitive loads arises when signals must be propagated from the on-chip to off-chip destinations.
- Generally, typical off chip capacitances may be several orders higher than on-chip $\square C_g$ values.
- If the off-chip load is denoted $C_L$ then, $C_L \geq 10^4 \square C_g$ (typically)
- Clearly capacitances of this order must be driven through low resistances, otherwise excessively long delays will occur.
- Three methods to derive large capacitance,

1. Cascaded Inverters as Drivers
2. Super buffers
3. Bi-CMOS drivers

# 1. Cascaded Inverters as Drivers

- For MOS circuits, low resistance values for $Z_{p.d.}$ and $Z_{p.u.}$ imply low L: W ratios; in other words, channels must be made very wide to reduce resistance value and, in consequence, an inverter to meet this need occupies a large area.
- Clearly, as the width factor increases, so the capacitive load presented at the inverter input increases, and the area occupied increases also.
- Equally clearly, the rate at which the width increases (that is, the value of f) will influence the number N of stages which must be cascaded to drive a particular value of $C_L$.

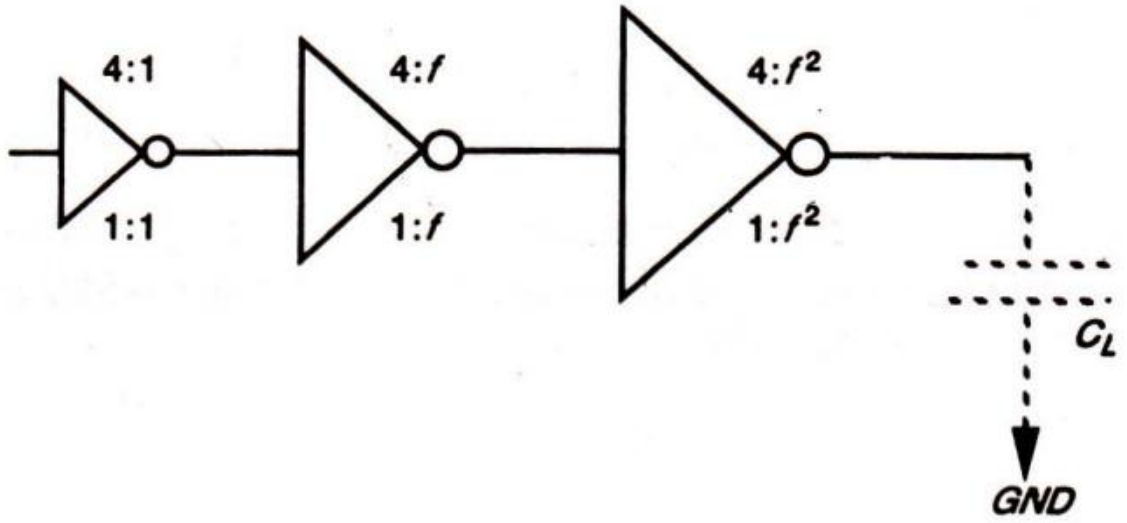



**FIGURE 4.11** Driving large capacitive loads.

With large $f$, $N$ decreases but delay per stage increases. For 4:1 nMOS inverters

delay per stage = $f\tau$ for $\Delta V_{in}$
or = $4f\tau$ for $\nabla V_{in}$

where $\Delta V_{in}$ indicates logic 0 to 1 transition and $\nabla V_{in}$ indicates logic 1 to 0 transition of $V_{in}$

Therefore, total delay per nMOS pair = $5f\tau$. A similar treatment yields delay per CMOS pair = $7f\tau$. Now let

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

so that the choice of $f$ and $N$ are interdependent.

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

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

Thus, for $N$ even

$$\text{total delay} = \frac{N}{2} 5f\tau = 2.5\, Nf\tau \text{ (nMOS)}$$

$$\text{or} = \frac{N}{2} 7f\tau = 3.5\, Nf\tau \text{ (CMOS)}$$

Thus, in all cases

$$\text{delay} \propto Nf\tau = \frac{\ln(y)}{\ln(f)} f\tau$$

Thus, assuming that $f = e$, we have

$$\text{Number of stages } N = \ln(y)$$

and overall delay $t_d$

$$N \text{ even: } t_d = 2.5eN\,\tau \text{ (nMOS)}$$

$$\text{or } t_d = 3.5eN\,\tau \text{ (CMOS)}$$

$$\left.\begin{aligned} N \text{ odd: } t_d &= [2.5(N-1)+1]e\tau \text{ (nMOS)} \\ \text{or } t_d &= [3.5(N-1)+2]e\tau \text{ (CMOS)} \end{aligned}\right\} \text{ for } \Delta V_{in}$$

or

$$\left.\begin{aligned} t_d &= [2.5(N-1)+4]e\tau \text{ (nMOS)} \\ \text{or } t_d &= [3.5(N-1)+5]e\tau \text{ (CMOS)} \end{aligned}\right\} \text{ for } \nabla V_{in}$$

It can be shown that total delay is minimized if f assumes the value e (base of natural logarithms); that is, each stage should be approximately 2.7 times wider than its predecessor. This applies to CMOS as well as nMOS inverters.

# 2. Super buffers

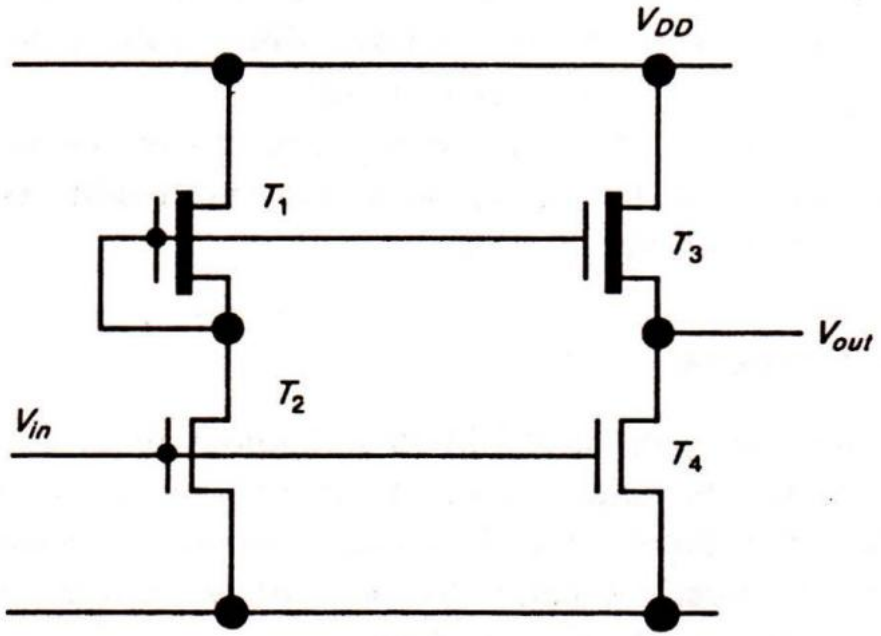



FIGURE 4.12 Inverting type nMOS super buffer.

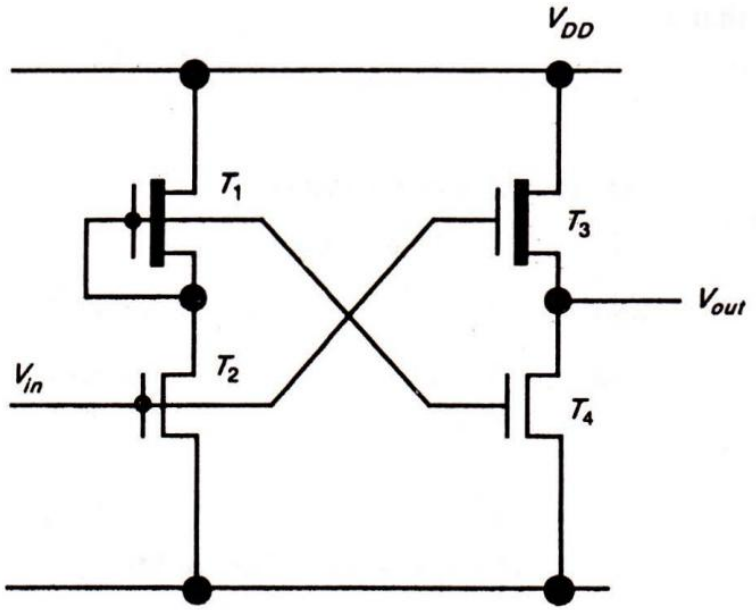



FIGURE 4.13 Non-inverting type nMOS super buffer.

- Since $I_{ds} \alpha V_{gs}$ then, doubling the effective $V_{gs}$ will increase the current and thus reduce the delay in charging any capacitance on the output, so that more symmetrical transitions are achieved.
- The structures shown when realized in 5 µm technology are capable of driving loads of 2 pF with 5 nsec rise-time.

# 3. Bi-CMOS drivers

- BJTs have better transconductance $g_m$ and current/area (I/A) than MOSFETs. This indicates high current drive capabilities for small areas in silicon.
- BJTs have better swinging performance as compared to the MOSFETs.

It may be shown that the time $\Delta t$ necessary to change the output voltage $V_{out}$ by an amount equal to the input voltage $V_{in}$ is given by

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

where $g_m$ is the transconductance of the bipolar transistor.

Clearly, since the bipolar transistor has a relatively high transconductance, the value of $\Delta t$ is small.

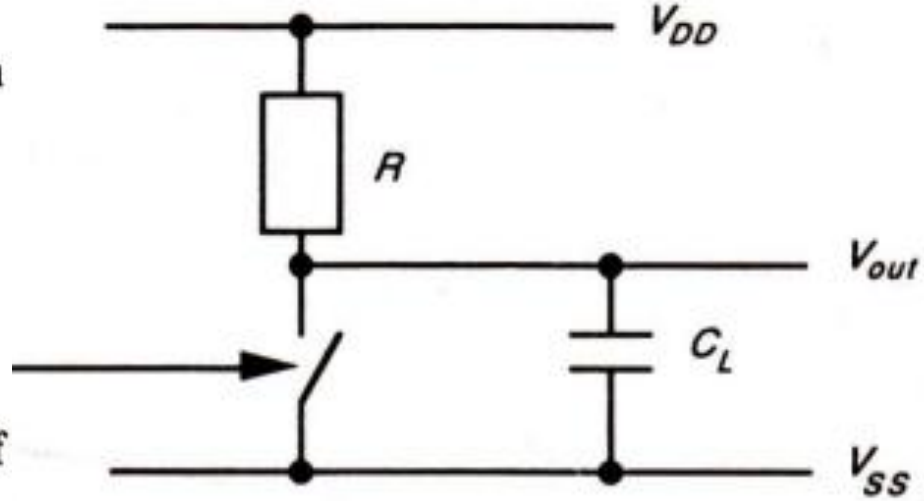



**FIGURE 4.14 Driving ability of bipolar transistor.**

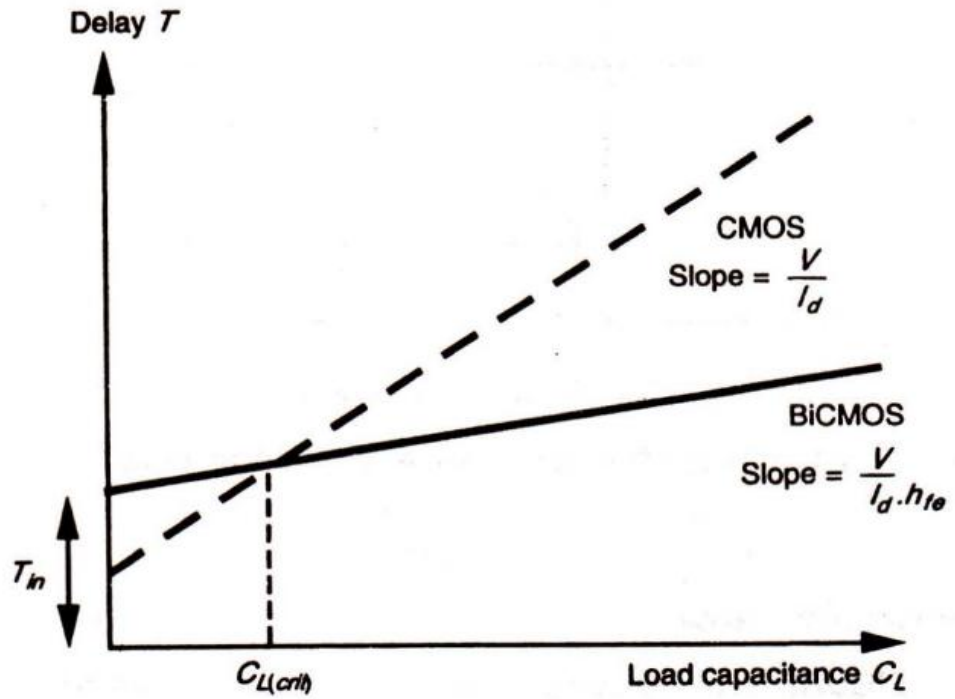
$T_{in}$ - an initial time necessary to charge the base emitter junction of the bipolar (npn) transistor.

For BJT, $T_{in}$=2ns

For MOSFET $T_{in}$=1ns

$T_L$ - the time taken to charge the output load capacitance $C_L$

- It will be seen that there is a critical value of load capacitance $C_{L(cric)}$ below which the BiCMOS driver is slower than a comparable CMOS driver.




- Delay of BiCMOS inverter can be described by

$$T = T_{in} + (V/I_d)\,(1/h_{fe})\,C_L$$

where

$T_{in}$ = time to charge up base/emitter junction
$h_{fe}$ = transistor current gain (common emitter)

- Delay for BiCMOS inverter is reduced by a factor of $h_{fe}$ compared with a CMOS inverter.

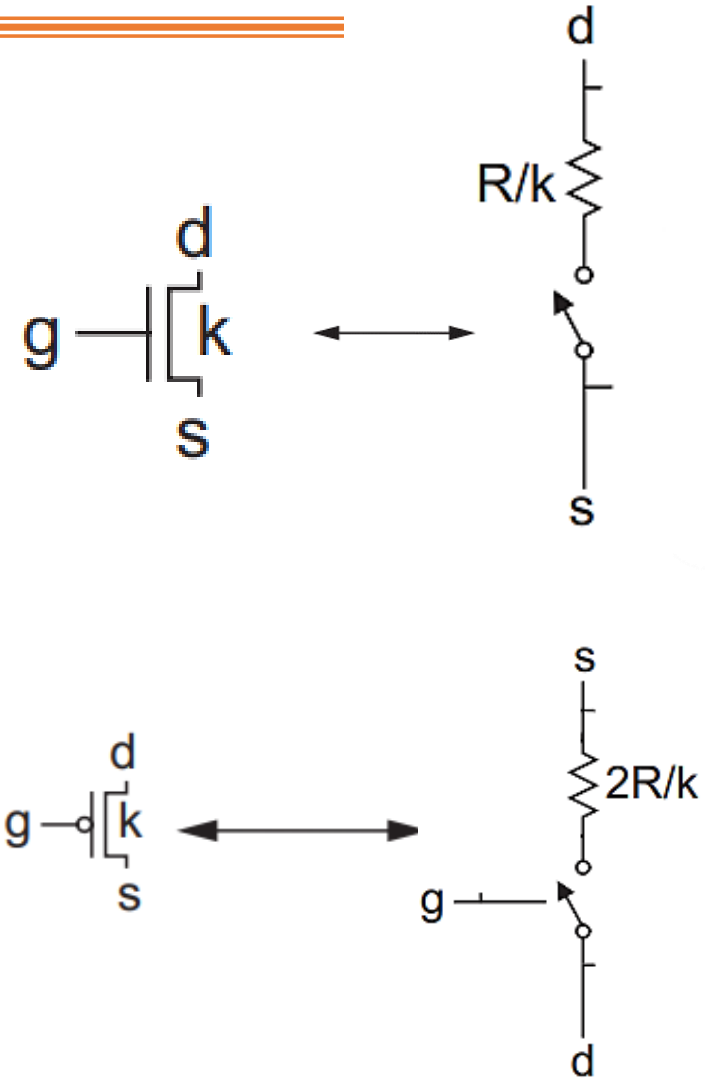
**FIGURE 4.15 Delay estimation.**

# RC delay model

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

1. Effective resistance:

- The RC delay model treats a transistor as a switch in series with a resistor.
- A unit nMOS transistor is defined to have effective resistance R.
- An nMOS transistor of k times unit width has resistance R/k.
- A unit pMOS transistor has greater resistance, generally in the range of 2R.
- A pMOS transistor of k times unit width has resistance 2R/k.



## 2. Gate and diffusion capacitances

- Each transistor has gate and diffusion capacitance.
- We define C to be the gate capacitance of a unit transistor of either flavor.
- A transistor of k times unit width has capacitance kC.
- Wider transistors have proportionally greater diffusion capacitance.
- Increasing channel length increases gate capacitance proportionally but does not affect diffusion capacitance.

| | Resistance | Capacitance (both gate and diffusion capacitances) |
|---|---|---|
| **nMOS** | **R/k** | **kC** |
| **pMOS** | **2R/k** | **kC** |

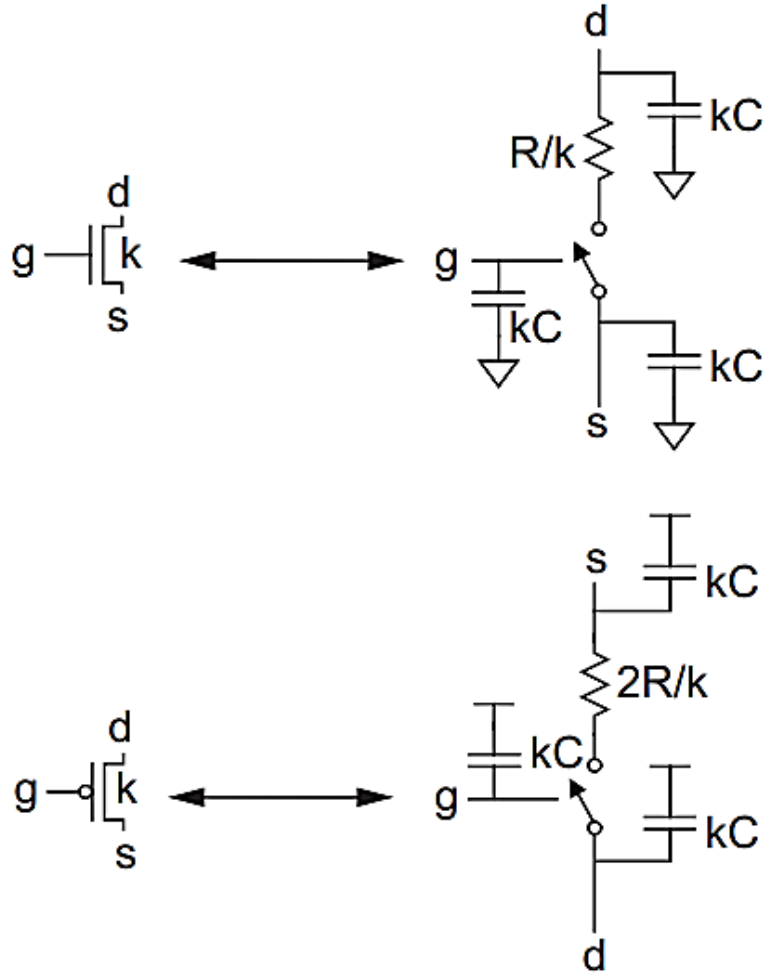



**FIGURE 4.5**
Equivalent circuits for transistors

## 3. Equivalent RC circuits

- Figure 4.5 shows the RC equivalent circuit of nMOS and pMOS.
- Figure 4.6 shows the equivalent circuit for a fanout-of-1 inverter with negligible wire capacitance

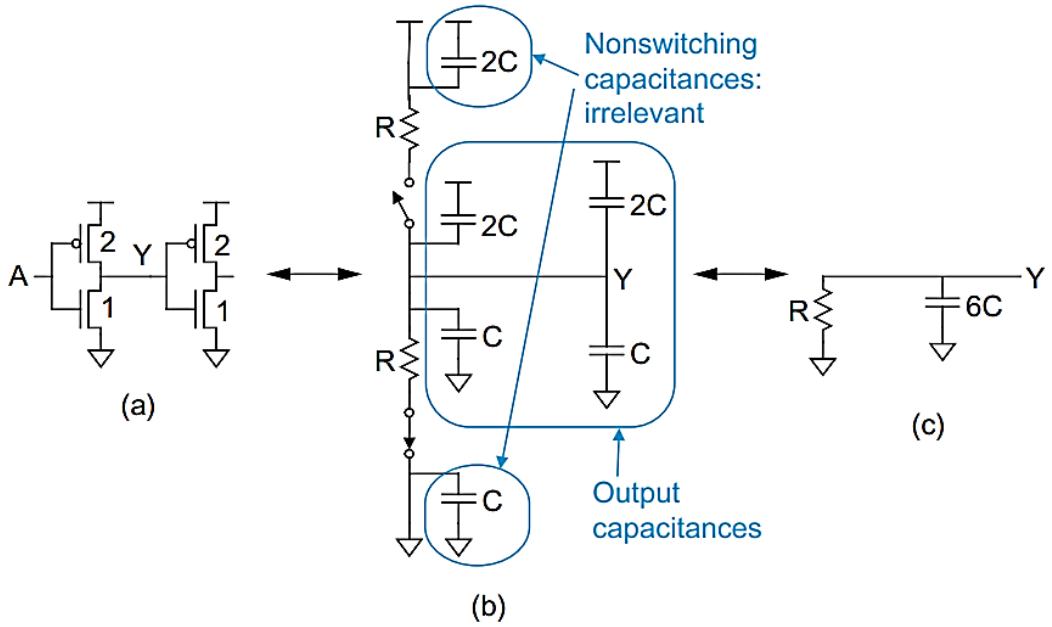



**FIGURE 4.6** Equivalent circuit for an inverter

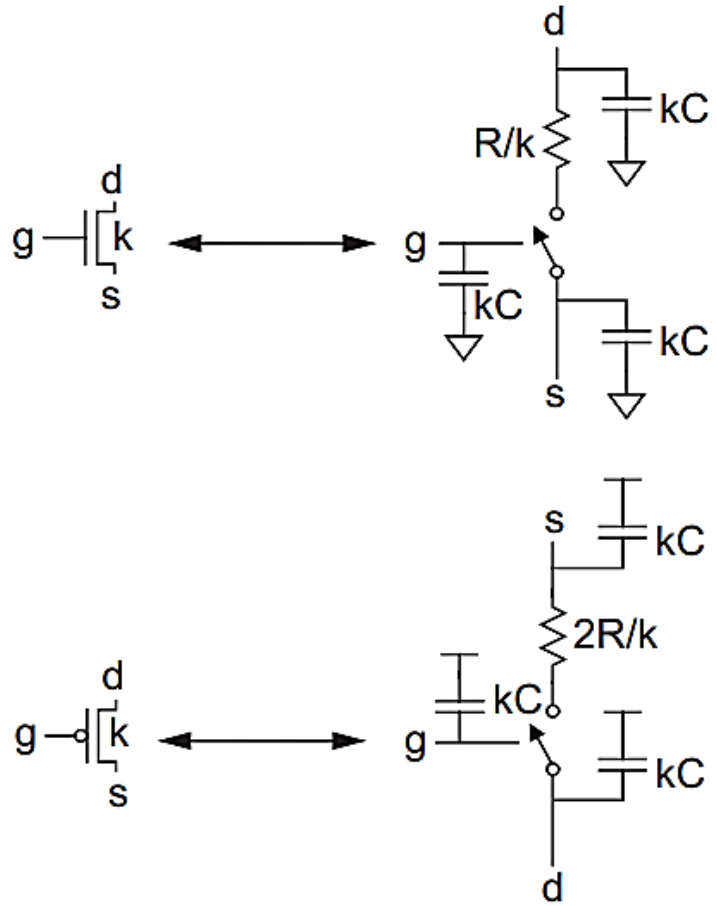
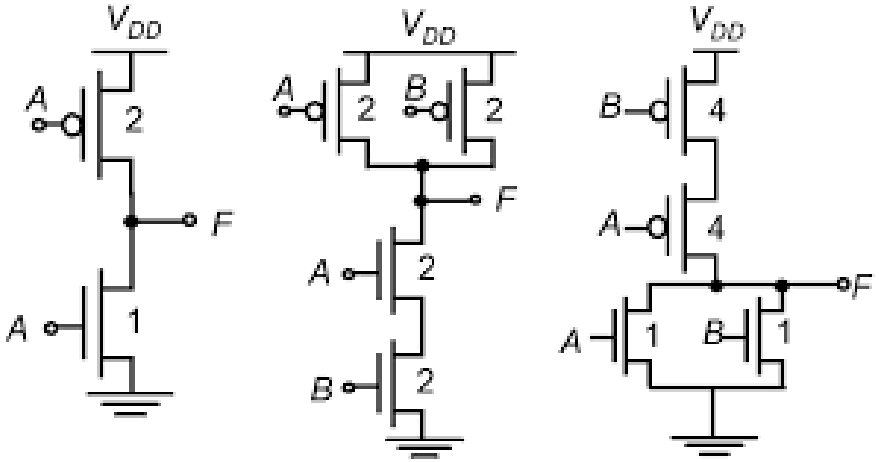
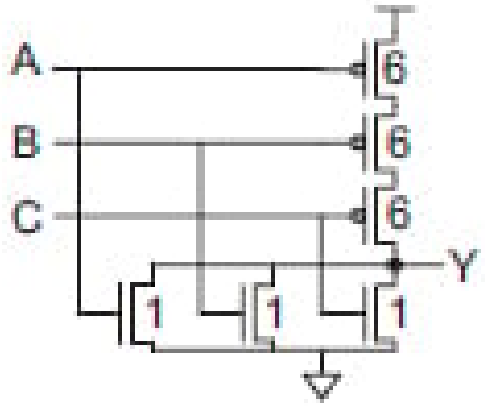



**FIGURE 4.5**
Equivalent circuits for transistors

# Transistor sizing



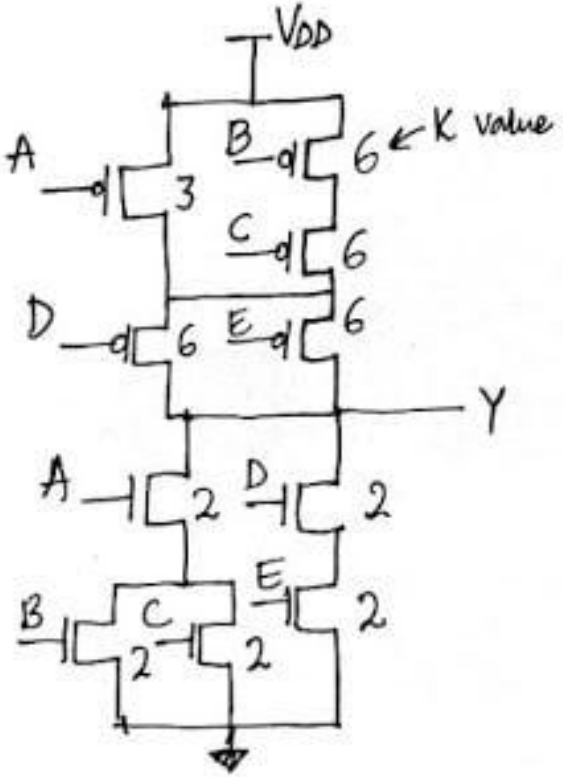




- for symmetrical response (dc, ac)
- for performance

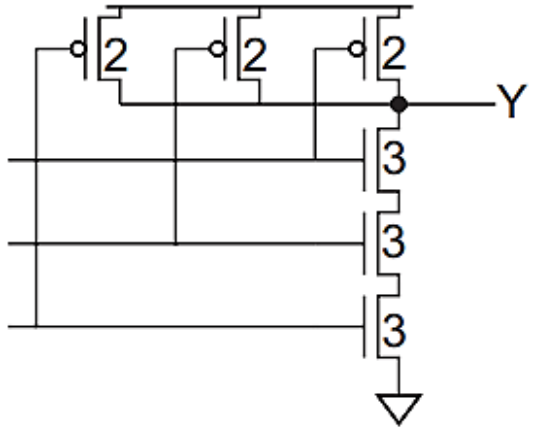
**Input Dependent**

**Focus on worst-case**

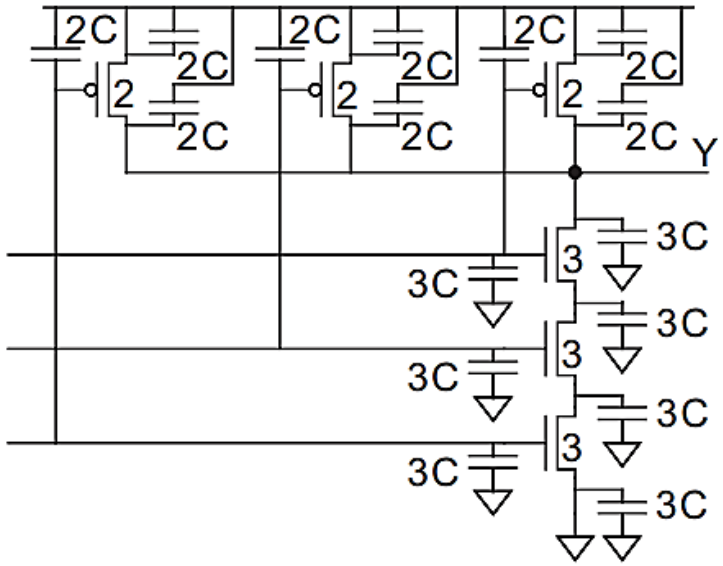
Numbers indicate transistor sizing with minimum size equal to 1



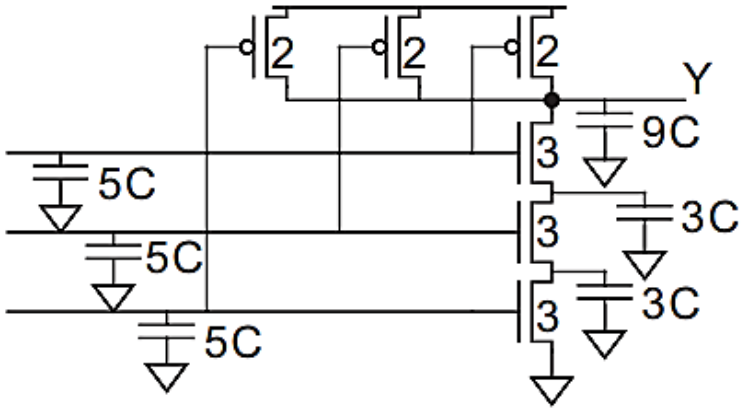
**Sketch a 3-input NAND gate with transistor widths chosen to achieve effective rise and fall resistance equal to that of a unit inverter (R). Annotate the gate with its gate and diffusion capacitances. Assume all diffusion nodes are contacted. Then sketch equivalent circuits for the falling output transition and for the worst-case rising output transition.**



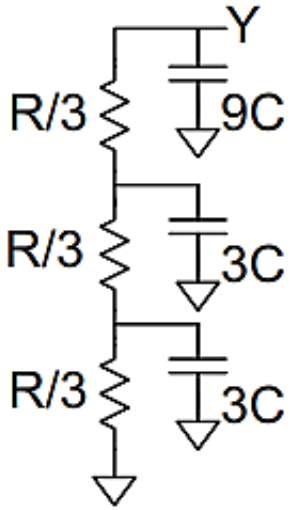

(a)




(b)

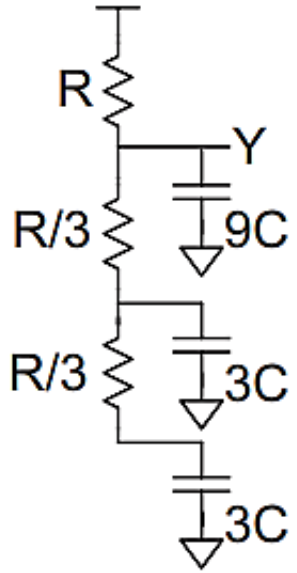



(c)




Falling

(d)




Rising

(e)

## 4. Elmore delay

- In general, most circuits of interest can be represented as an RC tree, i.e., an RC circuit with no loops.
- The root of the tree is the voltage source, and the leaves are the capacitors at the ends of the branches.
- The Elmore delay model estimates the delay from a source switching to one of the leaf nodes changing as the sum over each node ***i*** of the capacitance $\boldsymbol{C_i}$ on the node, multiplied by the effective resistance $\boldsymbol{R_{is}}$ on the shared path from the source to the node and the leaf.

$$t_{pd} = \sum_i R_{is} C_i$$

## 1. Compute the Elmore delay for $V_{out}$ in the 2nd order RC system from Figure 4.10.

- The circuit has a source and two nodes.
- At node $n_1$, the capacitance is $C_1$ and the resistance to the source is $R_1$.
- At node $V_{out}$, the capacitance is $C_2$ and the resistance to the source is $(R_1 + R_2)$.
- Hence, the Elmore delay is $t_{pd} = R_1C_1 + (R_1 + R_2)C_2$

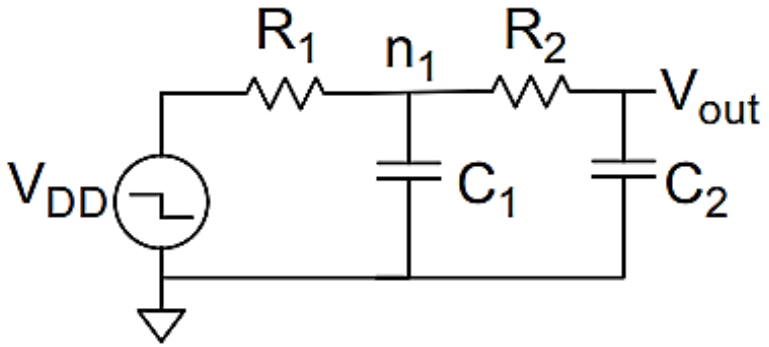



**FIGURE 4.10** Second-order RC system

## 2. Estimate $t_{pd}$ for a unit inverter driving m identical unit inverters.

- Figure 4.12 shows an equivalent circuit for the falling transition.
- Each load inverter presents 3C units of gate capacitance, for a total of 3mC.
- Elmore delay is $t_{pd} = (3 + 3m)RC$

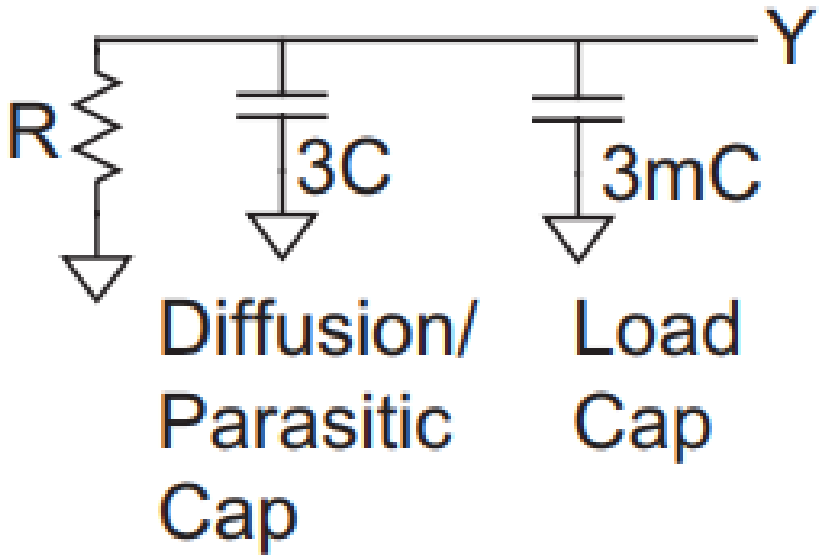



**FIGURE 4.12** Equivalent circuit for inverter

## 3. Repeat Example 2 if the driver is *w* times unit size.

- The driver transistors are w times as wide, so the effective resistance decreases by a factor of w.
- The diffusion capacitance increases by a factor of w.
- The Elmore delay is $t_{pd} = ((3w + 3m)C)(R/w)$

$$t_{pd} = (3 + 3m/w)RC.$$

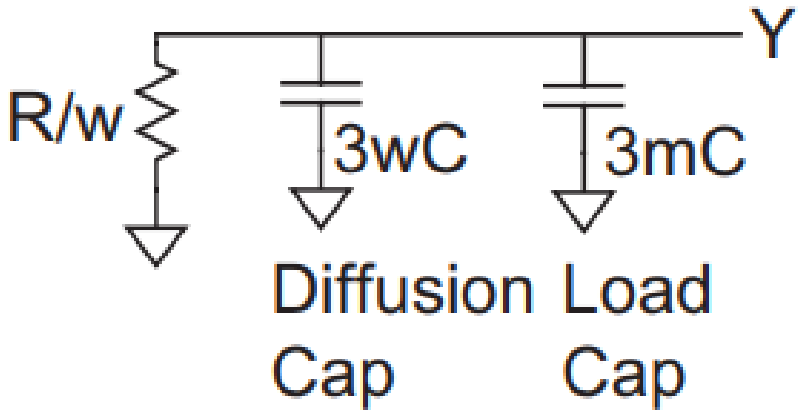



**FIGURE 4.13** Equivalent circuit for wider inverter

**4. If a unit transistor has R = 10 k ohm and C = 0.1 fF in a 65 nm process, compute the delay, in picoseconds, of the inverter in Figure 4.14 with a fanout of m = 4.**

- The RC product in the 65 nm process is (10 k ohm)(0.1 fF) = 1 ps.
- For m=4, the delay is (3+3m)(1 ps) = 15 ps
- The inverter can switch about 66 billion times per second.

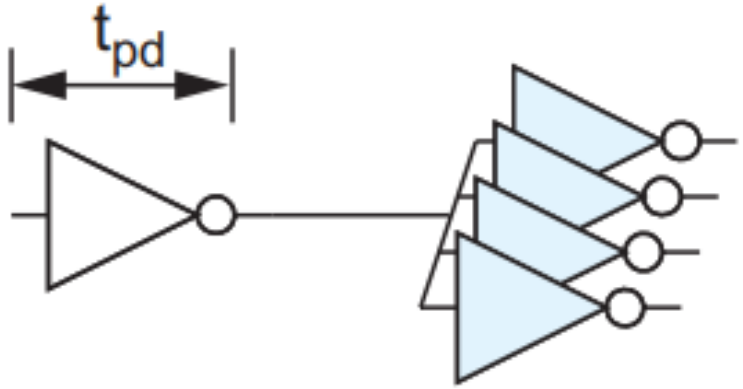



**FIGURE 4.14** Fanout-of-4 (FO4) inverter

# Linear delay model

- The RC delay model showed that delay is a linear function of the fanout of a gate.
- The normalized delay of a gate can be expressed in units of $\tau$ as,

$$d = f + p$$

- ***p*** is the ***parasitic delay*** inherent to the gate when no load is attached. ***f*** is the ***effort delay*** or ***stage effort*** that depends on the complexity and fanout of the gate:

$$f = gh$$

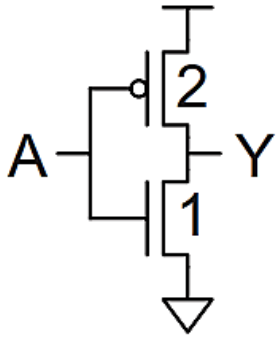
- ***g*** is the logical effort. A gate driving ***h*** identical copies of itself is said to have a fanout or electrical effort of ***h***.

$$h = \frac{C_{\text{out}}}{C_{\text{in}}}$$

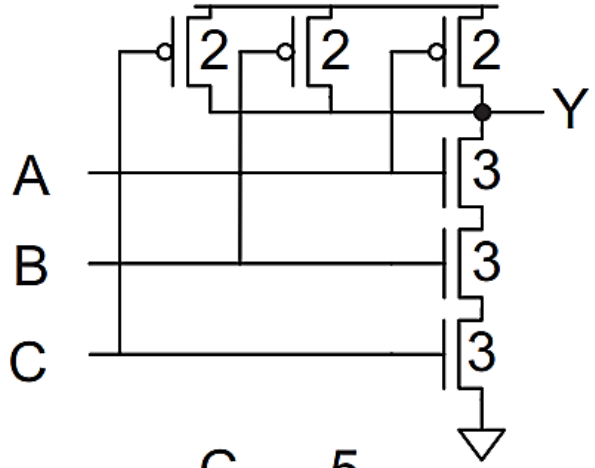
- An inverter is defined to have a logical effort of 1.

1. Logical effort (*g*)

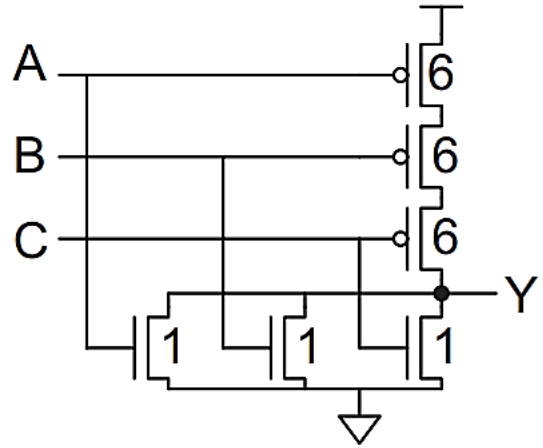
- Logical effort of a gate is defined as ***the ratio of the input capacitance of the gate to the input capacitance of an inverter*** that can deliver the same output current.
- Equivalently, logical effort indicates how much worse a gate is at producing output current as compared to an inverter, given that each input of the gate may only present as much input capacitance as the inverter.




(a) $C_{in} = 3$, $g = 3/3$




(b) $C_{in} = 5$, $g = 5/3$




(c) $C_{in} = 7$, $g = 7/3$

**TABLE 4.2** Logical effort of common gates

| Gate Type | Number of Inputs | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | *n* |
| inverter | 1 | | | | |
| NAND | | 4/3 | 5/3 | 6/3 | $(n+2)/3$ |
| NOR | | 5/3 | 7/3 | 9/3 | $(2n+1)/3$ |
| tristate, multiplexer | 2 | 2 | 2 | 2 | 2 |

## 2. Parasitic delay (*P*)

- The parasitic delay of a gate is the delay of the gate when it drives zero load.
- It can be estimated with RC delay models.
- A crude method good for hand calculations is to count only diffusion capacitance on the output node.
- For example, consider the gates in Figure 4.22, transistor widths were chosen to give a resistance of R in each gate.
- The inverter has three units of diffusion capacitance on the output, so the parasitic delay is 3RC = $\tau$
- This is then normalized parasitic delay equal to 1. This is generally denoted as $P_{inv}$
- **$P_{inv}$ is the ratio of diffusion capacitance to gate capacitance.**

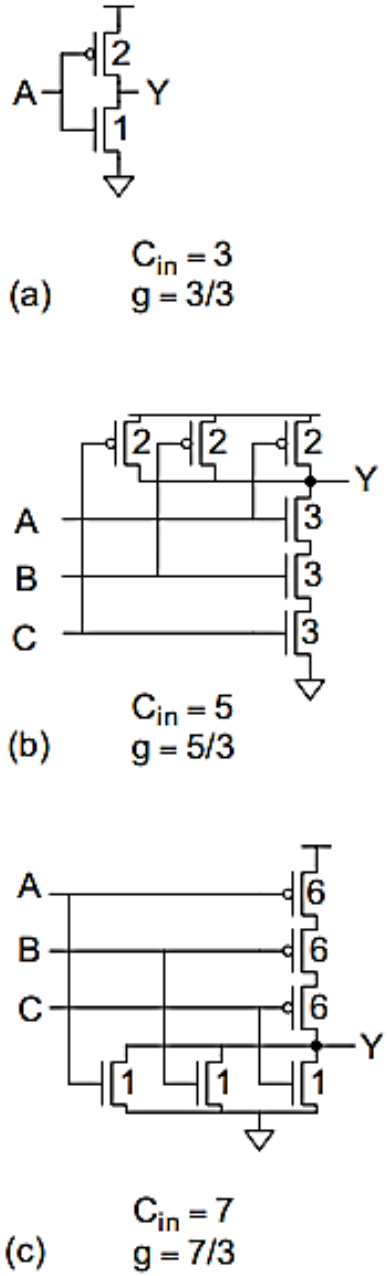



**FIGURE 4.22** Logic gates sized for unit resistance

- The 3-input NAND and NOR each have 9 units of diffusion capacitance on the output, so the parasitic delay is three times as great ($3P_{inv}$, or simply 3).

**TABLE 4.3** Parasitic delay of common gates

| Gate Type | Number of Inputs | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | *n* |
| inverter | 1 | | | | |
| NAND | | 2 | 3 | 4 | *n* |
| NOR | | 2 | 3 | 4 | *n* |
| tristate, multiplexer | 2 | 4 | 6 | 8 | 2*n* |

## 3. Delay in logic gates

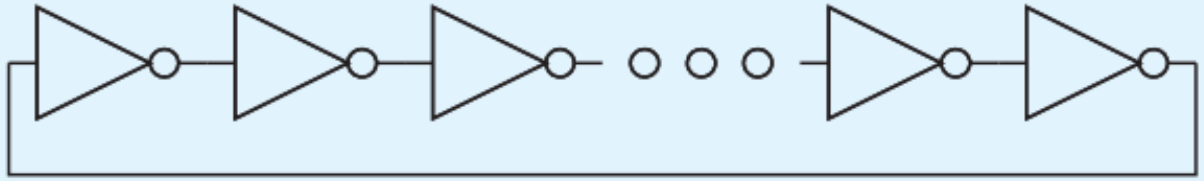
### Example 4.10

Use the linear delay model to estimate the delay of the fanout-of-4 (FO4) inverter from Example 4.6. Assume the inverter is constructed in a 65 nm process with $\tau = 3$ ps.

- The logical effort of the inverter is g = 1, by definition.
- The electrical effort, h = 4 because the load is four gates of equal size.
- The parasitic delay of an inverter is $p_{inv} = 1$.
- The total delay is $\boldsymbol{d = gh + p = 1 \times 4 + 1 = 5}$ in normalized terms, or $\boldsymbol{t_{pd} = 15\ ps}$ in absolute terms.

## Example 4.11

A ring oscillator is constructed from an odd number of inverters, as shown in Figure 4.24. Estimate the frequency of an *N*-stage ring oscillator.



- The logical effort of the inverter is ***g = 1***, by definition.
- The electrical effort ***h*** of each inverter is also 1 because it drives a single identical load.
- The parasitic delay ***p*** is also 1.
- The delay of each stage is ***d = gh + p = 1 × 1 + 1 = 2.***
- An N-stage ring oscillator has a period of 2N stage delays because a value must propagate twice around the ring to regain the original polarity. Therefore, the period is T = 2 × 2N. The frequency is the reciprocal of the period, 1/4N.
- A 31-stage ring oscillator in a 65 nm process has a frequency of 1/(4 × 31 × 3 ps) = 2.7 GHz

# Scaling of MOS circuits

- VLSI fabrication technology is still in the process of evolution which is leading to smaller line widths and feature size and to higher packing density of circuitry on a chip.
- The scaling down of feature size generally leads to improved performance.
- VLSI technology may be characterized in terms of several indicators like:
    1. Minimum feature size
    2. Number of gates on one chip
    3. Power dissipation
    4. Maximum operational frequency
    5. Die size
    6. Production cost
- It is essential for the designer to understand the implementation and the effects of scaling.

# Scaling models and scaling factors

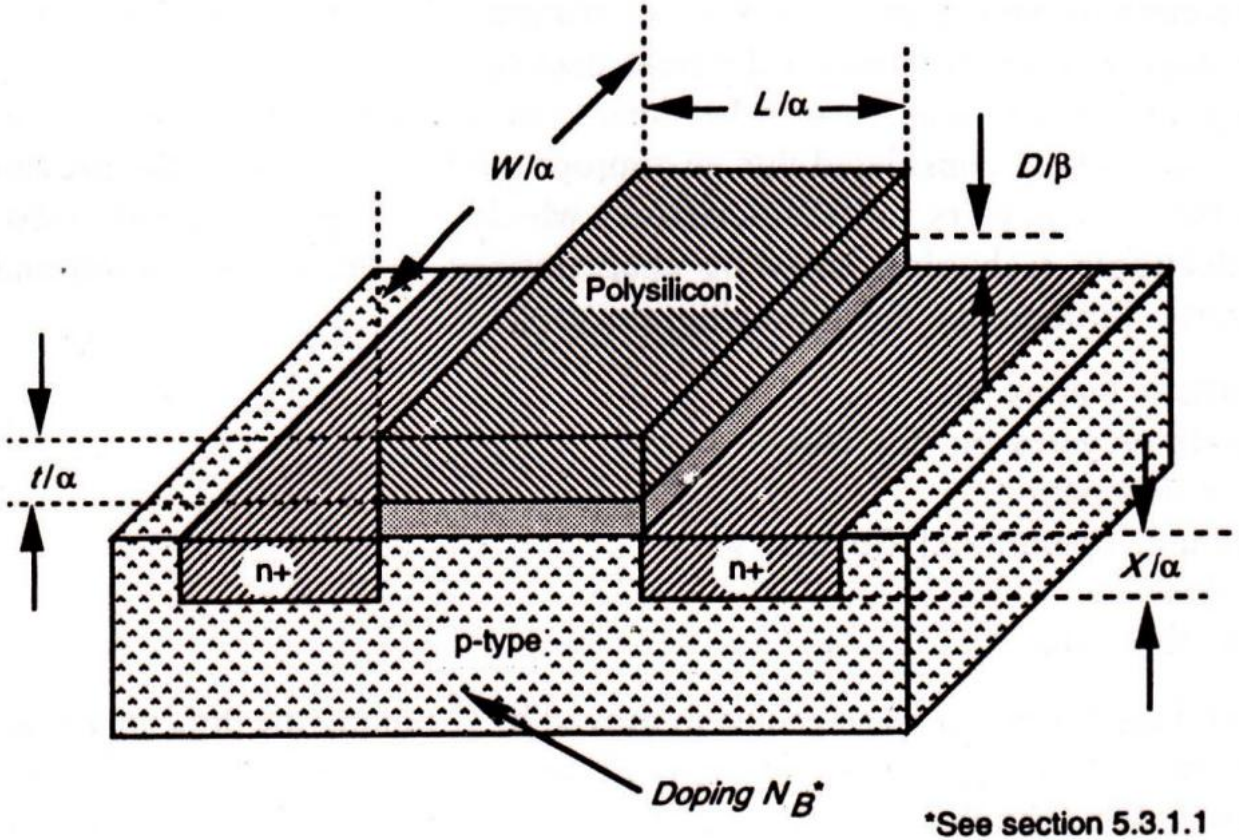



**FIGURE 5.1 Scaled nMOS transistor (pMOS similar).**

Scaling can be done using 2 factors, i.e., **α and β.**

**β** is for scaling oxide thickness D and voltages.

**α** is of all other parameters like L, A, W, t etc.

# Scaling factors for device parameters

**1. Gate Area $A_g$:** $A_g = L.W.$

where L and W are the channel length and width, respectively. Both are scaled by 1/α. Thus $A_g$ is scaled by $1/\alpha^2$

**2. Gate capacitance Per Unit Area $C_o$ or $C_{ox}$** $C_0 = \frac{\varepsilon_{ox}}{D}$

where $\varepsilon_{ox}$ is the permittivity of the gate oxide (thinox) $[= \varepsilon_{ins}.\varepsilon_0]$

D is the gate oxide thickness which is scaled by 1/β

Thus $C_0$ is scaled by $\frac{1}{1/\beta} = \beta$

**3. Gate capacitance $C_g$**

$$C_g = C_0 L.W.$$

Thus $C_g$ is scaled by $\beta\frac{1}{\alpha^2} = \frac{\beta}{\alpha^2}$

## 4. Parasitic Capacitance $C_x$

$C_x$ is proportional to $\frac{A_x}{d}$

- where d is the depletion width around source or drain which is scaled by 1/α,
- $A_x$ is the area of the depletion region around source or drain which is scaled by $1/\alpha^2$.

Thus $C_x$ is scaled by $\frac{1}{\alpha^2} \cdot \frac{1}{1/\alpha} = \frac{1}{\alpha}$

## 5. Carrier Density In Channel $Q_{on}$

$$Q_{on} = C_0 \cdot V_{gs}$$

- where $Q_{on}$ is the average charge per unit area in the channel in the 'on' state.
- Note that $C_0$ is scaled by β and $V_{gs}$ is scaled by 1/β.
- Thus-$Q_{on}$ is scaled by 1

## 6. Channel Resistance $R_{on}$

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

where μ is the carrier mobility in the channel and is assumed constant.

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

## 7. Gate Delay $T_d$

$T_d$ is proportional to $R_{on} \cdot C_g$

Thus $T_d$ is scaled by $\frac{1.\beta}{\alpha^2} \frac{\beta}{\alpha^2}$

## 8. Maximum Operating Frequency $f_o$

$$f_0 = \frac{W}{L} \frac{\mu C_0 V_{DD}}{C_g}$$

or, $f_0$ is inversely proportional to delay $T_d$.

Thus $f_0$ is scaled by $\frac{1}{\beta/\alpha^2} = \frac{\alpha^2}{\beta}$

## 9. Saturation Current $I_{dss}$

$$I_{dss} = \frac{C_0 \mu}{2} \frac{W}{L} (V_{gs} - V_t)^2$$

noting that both $V_{gs}$ and $V_t$ are scaled by $1/\beta$, we have

$I_{dss}$ is scaled by $\beta(1/\beta)^2 = 1/\beta$

## 10. Current Density J

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

where $A$ is the cross-sectional area of the channel in the 'on' state which is scaled by $1/\alpha^2$

So, $J$ is scaled by $\frac{1/\beta}{1/\alpha^2} = \frac{\alpha^2}{\beta}$

## 11. Switching Energy Per Gate $E_0$

$$E_g = \frac{1C_g}{2}(V_{DD})^2$$

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

## 12. Power Dissipation Per Gate $P_g$

$P_g$ comprises two components such that

$$P_g = P_{gs} + P_{gd}$$

where the static component

$$P_{gs} = \frac{(V_{DD})^2}{R_{on}}$$

and the dynamic component

$$P_{gd} = E_g f_0$$

It will be seen that both $P_{gs}$ and $P_{gd}$ are scaled by $1/\beta^2$

So, $P_g$ is scaled by $1/\beta^2$

## 13. Power Dissipation Per Unit Area $P_a$

$$P_a = \frac{P_g}{A_g}$$

So, $P_a$ is scaled by $\frac{1/\beta^2}{1/\alpha^2} = \alpha^2/\beta^2$

## 14. Power-speed Product $P_T$

$$P_T = P_g \cdot T_d$$

So, $P_T$ is scaled by $\frac{1}{\beta^2} \cdot \frac{\beta}{\alpha^2} = \frac{1}{\alpha^2\beta}$