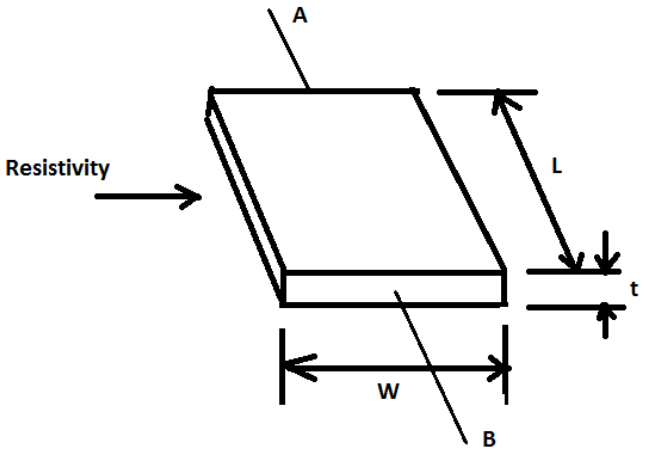
VLSI DESIGN (R 19)
II UNIT MATERIAL

SHEET RESISTANCE



Consider a uniform slab of conducting material of resistivity ρ, width w, length l, and thickness t. The resistance between the terminals A & B is $R_{AB} = \frac{\rho l}{A}$

Where A is cross sectional area=wt

$$\boldsymbol{R_{AB} = \frac{\rho l}{wt}}$$

Let $l = w$, $\therefore R_{AB} = R_s = \frac{\rho}{t}$

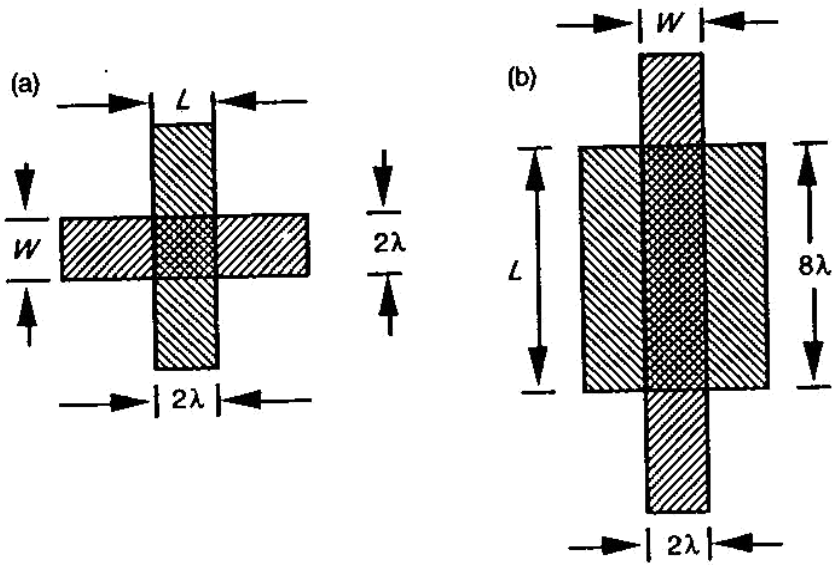
Where R_s is sheet resistance (or) $^{ohms}/_{square}$.

R_s is independent of area of square.

The actual values of R_s associated with different layers depends on thickness of the layer and resistivity of the metal forming the layer.

If $l \neq w$, then resistance $R = ZR_s$ where $Z = \frac{l}{w}$

Consider the following two transistor structures.



For N-MOS transistor in figure(a) has channel length $l = 2\lambda$ and channel width $w = 2\lambda$. The transistor channel becomes square and channel resistance $R = ZR_s = 1.R_s = R_s$

For n-channel transistor, $R_s = 10^4\Omega,\ Z = 1$

The N-MOS transistor has $R = 10^4\Omega = 10\text{k}\Omega$

For N-MOS transistor in figure(b) has channel length $l = 8\lambda$ and channel width $\text{w} = 2\lambda$. The transistor channel has length to width ratio $Z = \frac{l}{w} = 4 \rightarrow R = ZR_s = 4.R_s = 4R_s$

For n-channel transistor, $R_s = 10^4\Omega,\ Z = 4$

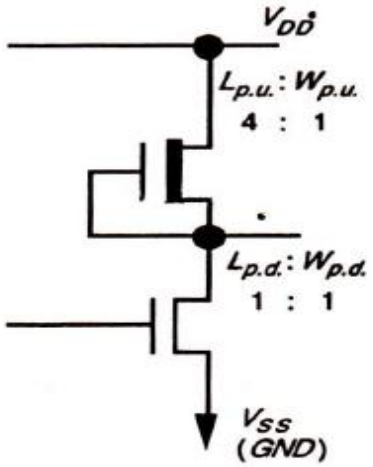
The N-MOS transistor has $R = 4 \times 10^4\Omega = 40k\Omega$

The typical sheet resistance in ohms per square

Layer	5µm	2µm	1.2µm
Metal	0.03	0.04	0.04
Diffusion	10 - 50	20 - 45	20 - 45
Silicide	2 - 4	--	--
Poly-Silicon	15 - 100	15 - 30	15 - 30
n- MOS	10^4	2 X 10^4	2 X 10^4
p-MOS	2.5 X 10^4	4.5 X 10^4	4.5 X 10^4

SHEET RESISTANCE OF INVERTERS

Consider an n-MOS inverter has the channel length 8λ and width 2λ for pull up transistor as shown in figure.



nMOS

For pull-up transistor $Z_{pu} = 4$

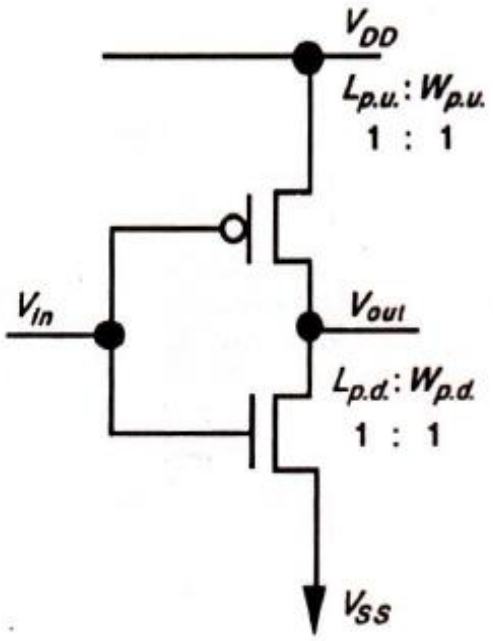
$$\therefore Resistance\ R = 4R_{on} = 4R_{sn} = 4 \times 10^4\Omega = 40k\Omega$$

For pull-down transistor $Z_{pd} = 1$

$$\therefore Resistance\ R = 1.R_{on} = R_{sn} = 10^4\Omega = 10k\Omega$$

The total ON resistance from V_{dd} to ground is $40k + 10k = 50k\Omega$

Consider an C-MOS inverter has the channel length 2λ and width 2λ for pull up & pull-down transistors as shown in figure.



For pull-up transistor $Z_{pu} = 1$

$$\therefore Resistance\ R = 1.R_{on} = R_{sp} = 2.5 \times 10^4\Omega = 25k\Omega$$

For pull-down transistor $Z_{pd} = 1$

$$\therefore Resistance\ R = 1.R_{on} = R_{sn} = 10^4\Omega = 10k\Omega$$

Note:- There is no static resistance between V_{dd} to ground.

AREA CAPACITANCE OF LAYERS

The conducting layers are separated by insulation(SiO_2) from the substrate and each other. So, a parallel plate capacitive effects must be present in the MOS device. For any layer, if dielectric constant and thickness is known the capacitance is given by

$$C = \frac{\epsilon_0 \epsilon_{ins} A}{D}$$

Where D is thickness of oxide layer

ϵ_{ins} is relative permittivity of $SiO_2 \approx 4.0$

ϵ_0 is standard dielectric constant $8.85 \times 10^{-12}\ {}^{F}/_{m}$

The unit of area capacitance is ${}^{pF}/_{\mu m^2}$

The standard unit of area capacitance is denoted by $\square C_g$.

It is defined as gate to channel capacitance of MOS transistor having $w = l$ feature size.

To calculate $\square C_g$

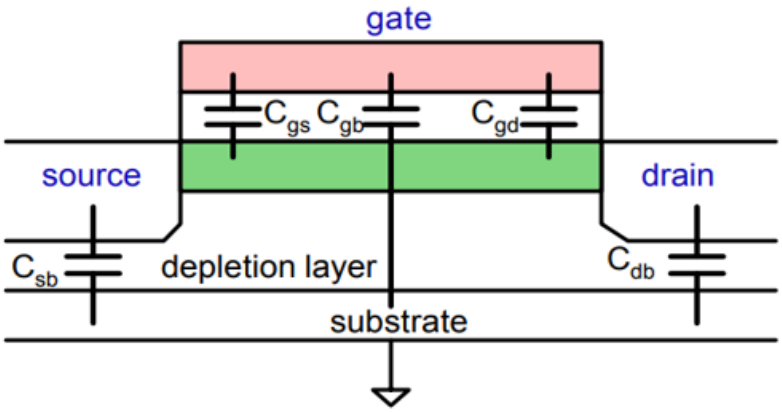
For example, if channel length is $5\mu m$ for a transistor

The area of the channel $= 25\mu m^2$

The gate capacitance C_g for $5\mu m$ technology is $C_g = 4 \times 10^{-4}\, {}^{pF}/_{\mu m^2}$

standard area capacitance $\square C_g = 25\mu m^2 \times 4 \times 10^{-4}\, {}^{pF}/_{\mu m^2} = 0.01pF$

MOS DEVICE CAPACITANCES



$C_g = C_{gb} + C_{gs} + C_{gd}$

For any MOS device following parasitic capacitances must exist. They are

1. $\boldsymbol{C_{gs}}$ & $\boldsymbol{C_{gd}}$**:-** These are gate to channel capacitances which are lumped at source and drain regions of channel respectively.
2. $\boldsymbol{C_{sb}}$ & $\boldsymbol{C_{db}}$**:-** These are source and drain diffusion capacitance to the substrate or bulk.
3. $\boldsymbol{C_{gb}}$**:-** Gate to substrate capacitance.

For any MOS device, the total capacitance $C_g = C_{gs} + C_{gb} + C_{gd}$

These capacitances and their behavior can be explained in terms of following models in three regions of operation.

1. **Cut-Off Region:-** $(V_{gs} < V_t)$

The device is OFF. There is no channel formation. So, C_{gs}&C_{gd} are zero. The gate to substrate capacitance C_{gb} can be modelled as a series combination of C_o & C_{dep}.

2. **Non-Saturation Region:-** $(V_{ds} < V_{gs} - V_t)$

When $V_{ds} < V_{gs} - V_t$, the channel is formed and the capacitances C_{gs} & C_{gd} becomes significant values which depends on gate voltage V_g. In this region $C_{gs} = C_{gd} = \frac{1}{2}\left\{\frac{\varepsilon_0 \varepsilon_{ins} A}{t_{ox}}\right\}$

In this region C_{gb} is reduced to zero.

3. **Saturation Region:-** $(V_{ds} > V_{gs} - V_t)$

When $V_{ds} > V_{gs} - V_t$ the channel is heavily inverted and pinched OFF at the drain region. The capacitance C_{gd} becomes zero and the capacitance C_{gs} increases.

$$C_{gs} = \frac{2}{3}\left\{\frac{\varepsilon_0 \varepsilon_{ins} A}{t_{ox}}\right\}$$

Typical area capacitance values of MOS circuits

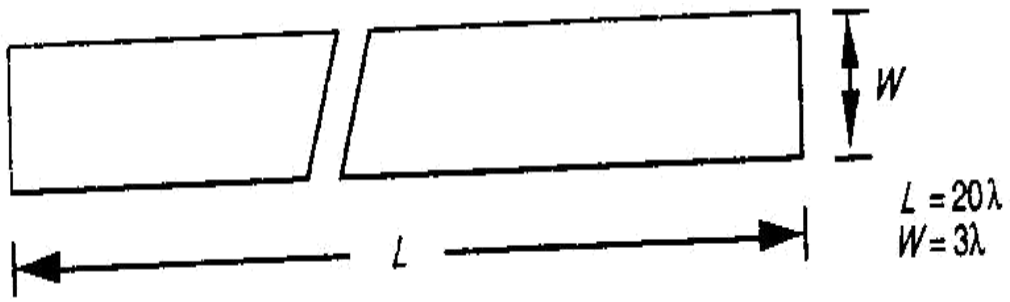
Typical area capacitance values for MOS circuits

Capacitance	*Value in pF × 10^{-4}/μm² (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)

CALCULATION OF AREA CAPACITANCE

The calculation of capacitance value is the ratio between the area of interest and the area of standard gate and multiplying this ratio by the appropriate relative C value from tabular form. The product will give the required capacitance in $\square c_g$ units.

Calculate the capacitance of the area as shown in figure of length 20λ and width 3λ



$$Relative\ Area = \frac{Area(l \times w)}{std\ gate\ area} = \frac{20\lambda \times 3\lambda}{2\lambda \times 2\lambda} = \frac{60\lambda^2}{4\lambda^2} = 15$$

Consider the area as metal1

capacitance to substrate = relative area x relative C value (from table)

= 15 x 0.075 □ c_g = 1.125 □ c_g

Consider the area as polysilicon

capacitance to substrate = relative area x relative C value (from table)

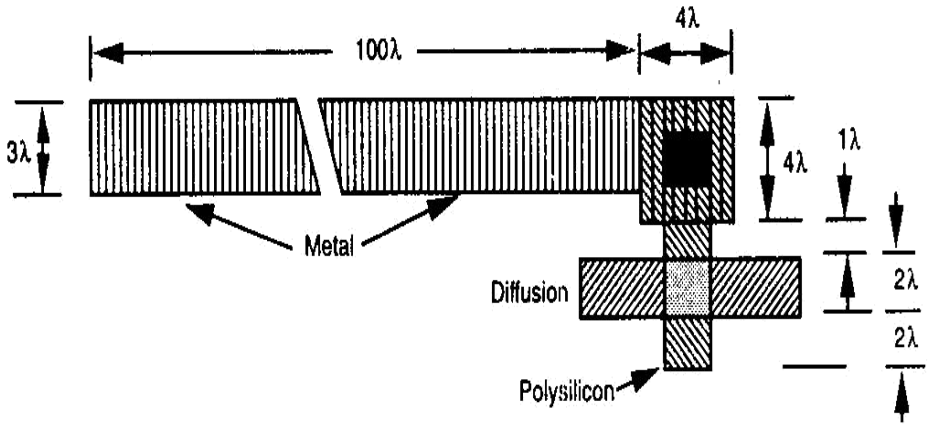
= 15 x 0.1 □ c_g = 1.5 □ c_g

Consider the area as n-type diffusion

capacitance to substrate = relative area x relative C value (from table)

= 15 x 0.25 □ c_g = 3.75 □ c_g

Consider the following structure which occupies more than one layer as shown in figure and calculate the area capacitance value.



Consider the metal area

$$Relative\ Area = \frac{Area(l \times w)}{std\ gate\ area} = \frac{100\lambda \times 3\lambda}{2\lambda \times 2\lambda} = \frac{300\lambda^2}{4\lambda^2} = 75$$

capacitance to substrate C_m= relative area x relative C value (from table)

= 75 x 0.075 □ c_g = 5.625 □ c_g

Consider the polysilicon area

$$Relative\ Area = \frac{Area(l \times w)}{std\ gate\ area} = \frac{(4\lambda \times 4\lambda) + (3\lambda \times 2\lambda)}{2\lambda \times 2\lambda} = \frac{22\lambda^2}{4\lambda^2} = 5.5$$

capacitance to substrate C_p = relative area x relative C value (from table)

= 5.5 x 0.1 □ c_g = 0.55 □ c_g

Consider the transistor (Diffusion)

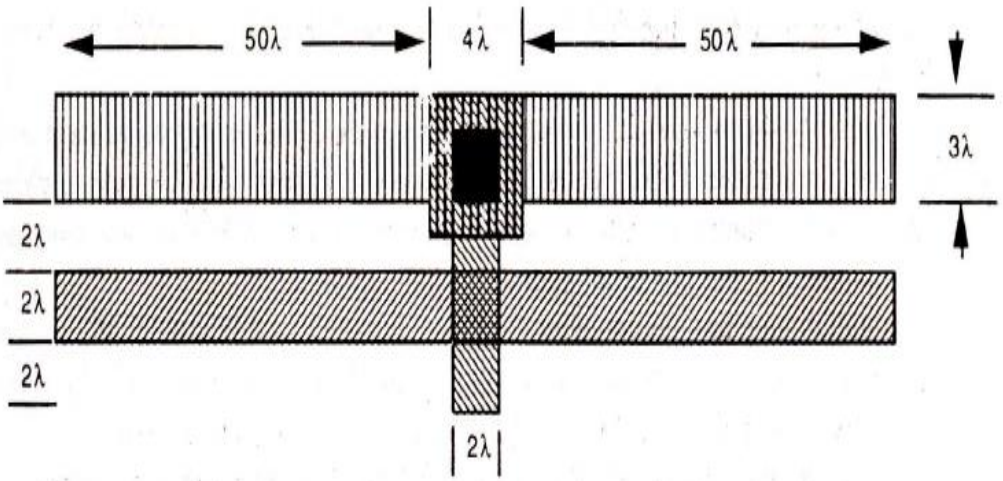
$$Relative\ Area = \frac{Area(l \times w)}{std\ gate\ area} = \frac{(2\lambda \times 2\lambda)}{2\lambda \times 2\lambda} = \frac{4\lambda^2}{4\lambda^2} = 1$$

Gate to channel capacitance C_g = relative area x relative C value (from table)

= 1 x 1 □ c_g = 1 □ c_g

The total capacitance is $C = C_m + C_p + C_g = 7.2$ □ cg

Example 2



Consider the metal area

$$Relative\ Area = \frac{Area(l \times w)}{std\ gate\ area} = \frac{100\lambda \times 3\lambda}{2\lambda \times 2\lambda} = \frac{300\lambda^2}{4\lambda^2} = 75$$

capacitance to substrate C_m= relative area x relative C value (from table)

= 75 x 0.075 □ c_g = 5.625 □ c_g

Consider the polysilicon area

$$Relative\ Area = \frac{Area(l \times w)}{std\ gate\ area} = \frac{(4\lambda \times 4\lambda)+(2\lambda \times 2\lambda)+(1\lambda \times 2\lambda)}{2\lambda \times 2\lambda} = \frac{22\lambda^2}{4\lambda^2} = 5.5$$

capacitance to substrate C_p = relative area x relative C value (from table)

= 5.5 x 0.1 □ c_g = 0.55 □ c_g

Consider the transistor (Diffusion)

$$Relative\ Area = \frac{Area(l \times w)}{std\ gate\ area} = \frac{(2\lambda \times 2\lambda)}{2\lambda \times 2\lambda} = \frac{4\lambda^2}{4\lambda^2} = 1$$

Gate to channel capacitance C_g = relative area x relative C value (from table)

= 1 x 1 □ c_g = 1 □ c_g

The total capacitance is $C = C_m + C_p + C_g = 7.175$ □ cg

ROUTING CAPACITANCE

The area capacitance associated with the layers to substrate are considered to calculate the delay, rise time, fall time etc. There are other significant sources of capacitance which effects the overall performance of the MOS device. It is called as "Routing (or) Wiring capacitance". This capacitance includes

- Fringing field capacitance
- Inter layer capacitance
- Peripheral capacitance (Diffusion)

1) **Fringing Field Capacitance:-**

Capacitance due to fringing fields can be a major component of overall routing capacitance of interconnect wires. The value of fringing field capacitance can be of same order as that of area capacitance. Thus, fringing field capacitance should be taken into account to get actual performance. The fringing field capacitance is $C_{ff} = \varepsilon_{SiO_2}\varepsilon_0 l \left\{ \frac{\pi}{\ln\left(1+\frac{2d}{t}\left(1+\sqrt{1+\frac{t}{d}}\right)\right)} - \frac{t}{4d} \right\}$

where l is the length of the interconnection.

t is thickness of interconnect.

d is separation between interconnect and substrate.

2) **Interlayer Capacitance:-**

This capacitance occurs only when different layers cross each other or when one layer under lies another layer. This capacitance depends on the layout of circuit for accurate circuit modelling & delay calculations.

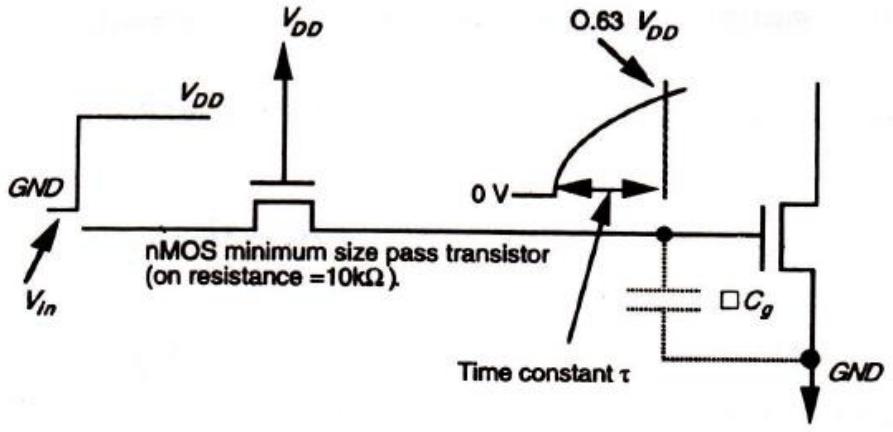
3) **Peripheral capacitance:-**

Source & drain regions forms the junctions with the substrate at a well-defined uniform depth. Each diode thus formed associate a peripheral capacitance in Pico farads (pF).

Thus, the total wiring capacitance is $C_{wiring} = C_{ff} + C_{interlay} + C_{peri}$

DELAY UNIT

The concept of sheet resistance & standard gate capacitance are applied to calculate the delay of MOS devices. Consider a standard gate area capacitance □c_g is to be charged through the minimum feature size of n-channel resistance as shown below:



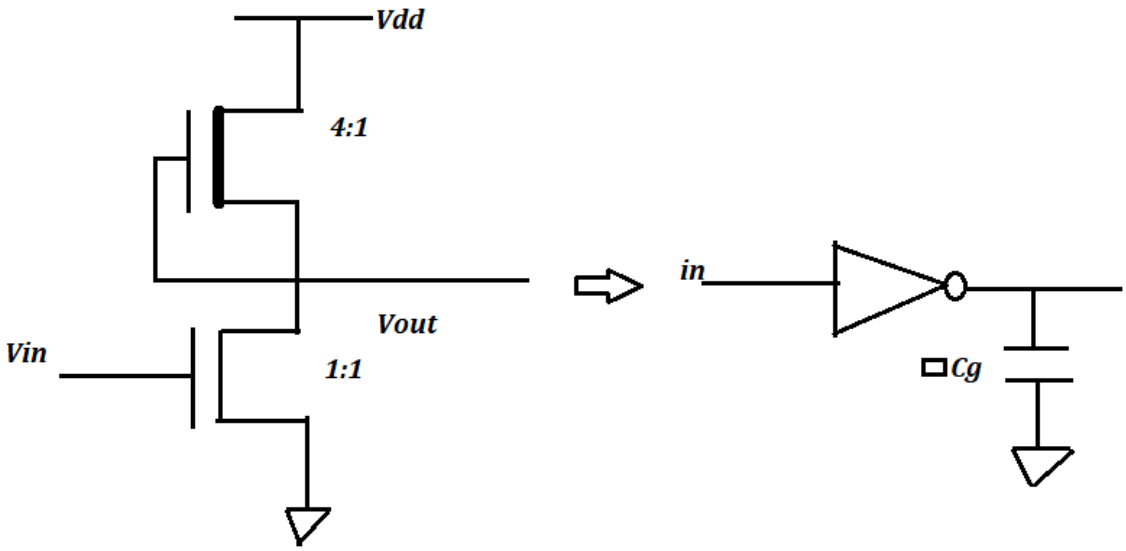
$Delay\ Unit\ \tau = time\ constant\ to\ charge$ □cg through R_s

$$\therefore \tau = R_s \times \square C_g$$

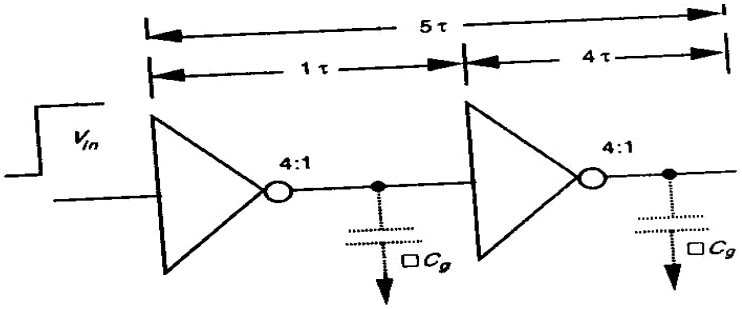
In practical conditions, the wiring and parasitic capacitances are allowed for calculation of delay unit τ. The time constant obtained does not have much difference from the transit time of electrons in the channel τ_{sd}. So the delay unit τ is taken as the fundamental time unit and all the timings in a system are accessed in relation to 'τ'.

Inverter Delays:-

Consider the basic 4:1 n-MOS inverter

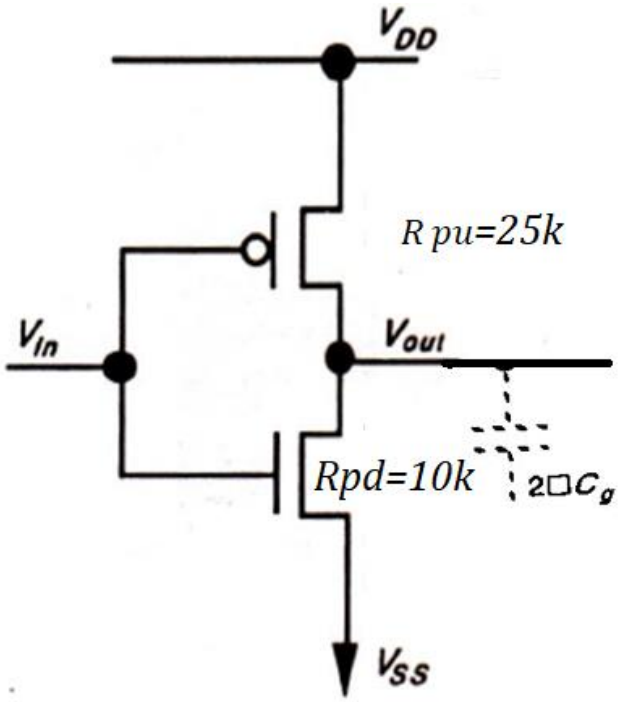


To get 4:1 Z_{pu} to Z_{pd} ratio, the resistance for pull-up transistor R_{pu} should be 4 times greater than pull down transistor R_{pd}. The time constant required to charge $1 \square C_g$ of capacitance through the n-MOS inverter is $4R_s \times \square C_g = 4\tau$. The time constant required to discharge $1 \square C_g$ of capacitance through the n-MOS inverter is $R_s \times \square C_g = 1\tau$. For the n-MOS inverter there is asymmetrical charging and discharging time delay. To get a constant time delay consider a pair of n-MOS inverters as shown below:

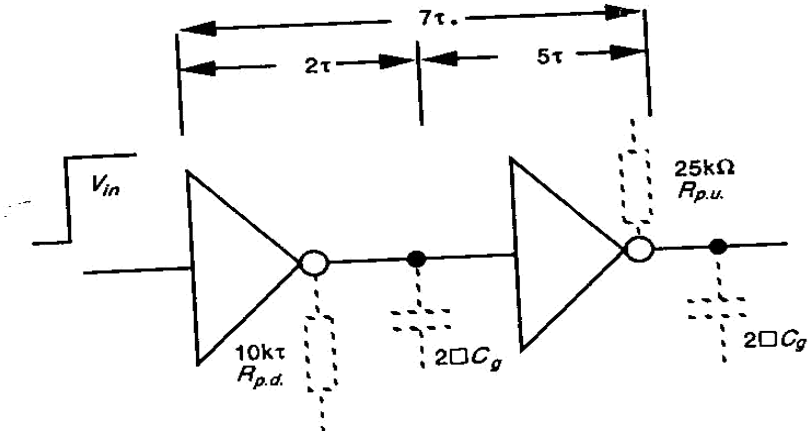


For a pair of cascaded inverters, the delay will be constant irrespective of logic level transition at the input. The overall delay is 5τ. In general, for a pair of n-MOS inverters the total time delay is $T_d = \left(1 + \frac{Z_{pu}}{Z_{pd}}\right)\tau$

Consider a CMOS inverter which is having natural asymmetry at the pull-up p-MOS transistor and the pull-down n-MOS transistor.

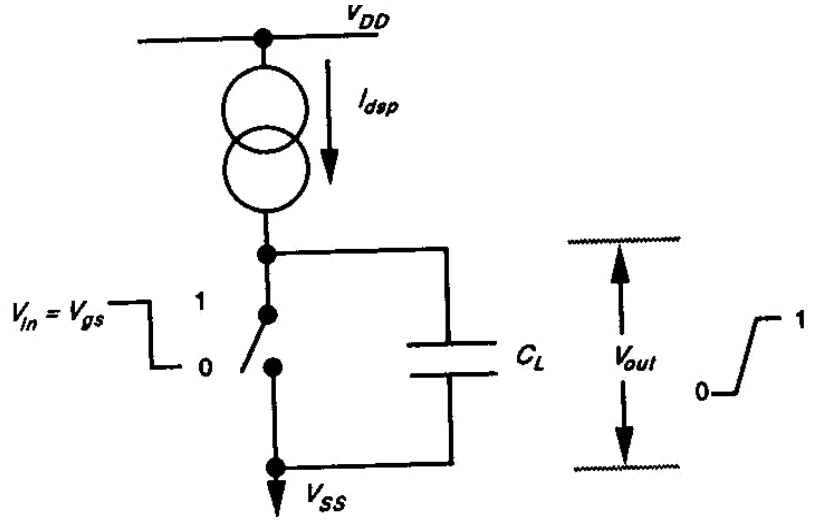


The gate capacitance is doubled that of n-MOS inverter since the input is connected to both p-MOS & n-MOS transistor gates. To charge $2\square C_g$ capacitance the time constant required is 5τ. To discharge $2\square C_g$ capacitance the time constant required is 2τ. To remove the asymmetrical charging & discharging time, a pair of CMOS inverters are considered as shown below:



The total time constant required for any logic level is 7τ.

ESTIMATION OF RISE TIME FOR CMOS INVERTER



For rise time, the output changes from logic 0 to logic 1 when the input changes from logic 1 to logic 0. So, the p-MOS transistor is in saturation for the entire period of rise time to charge C_l. When the p-MOS transistor is in saturation the current is equal to $I_{dsp} = \frac{\beta_p(V_{gs}-V_t)^2}{2}$

These current charges the C_L to V_{out} which is given by $V_{out} = \frac{I_{dsp}t}{C_L}$

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

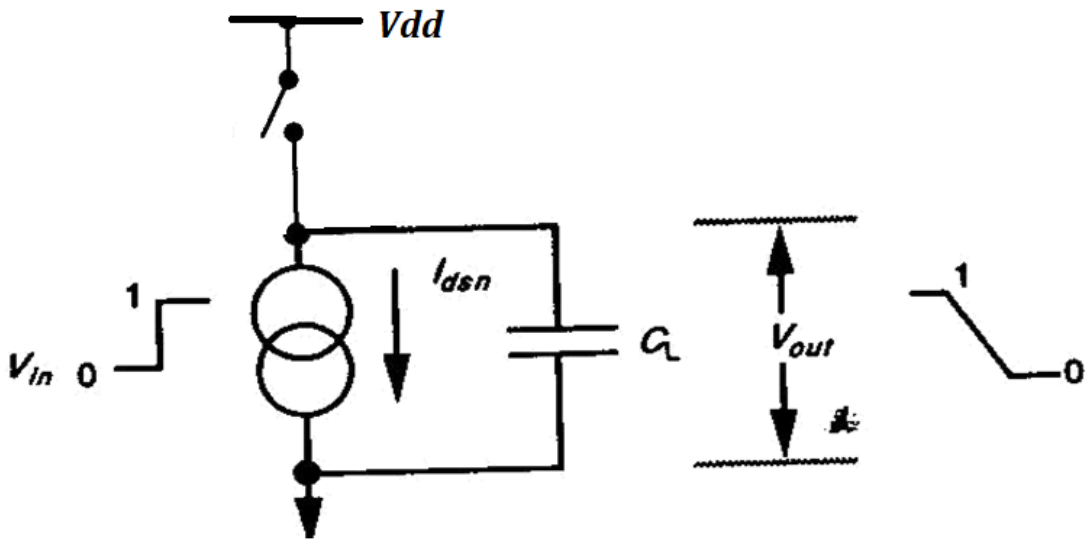
Substituting I_{dsp},

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

For rise time $t = t_r$, $V_{out} = V_{dd}$, $V_t = 0.2V_{dd}$ & $V_{gs} = V_{dd}$

$$t_r = V_{dd}C_L\frac{2}{\beta_p(V_{dd}-0.2V_{dd})^2} = \boldsymbol{\frac{3C_L}{\beta_p V_{dd}}}$$

ESTIMATION OF FALL TIME FOR CMOS INVERTER



For fall time, the output changes from logic 1 to logic 0 when the input changes from logic 0 to logic 1. So the n-MOS transistor is in saturation for the entire period of fall time to discharge C_l. When the n-MOS transistor is in saturation the current is equal to $I_{dsn} = \frac{\beta_n (V_{gs} - V_t)^2}{2}$

Similar to the rise time calculation, the fall time $t_r \approx \frac{3C_L}{\beta_n V_{dd}}$

The ratio of rise time to fall time is $\frac{t_r}{t_p} = \frac{\beta_n}{\beta_p}$

Where $\beta_n = k\frac{W_n}{L_n}$ and $k = \varepsilon_0 \varepsilon_{ins} \mu_n$

Where $\beta_p = k\frac{W_p}{L_p}$ and $k = \varepsilon_0 \varepsilon_{ins} \mu_p$

The mobility of n-type impurities is $\mu_n = 2.5\mu_p \rightarrow \beta_n = 2.5\beta_p$

The ratio $\frac{t_r}{t_p} = 2.5$

From the above equation the rise time is slower than fall time for minimum size n and p type devices. To achieve symmetrical operation the required condition is $W_p = 2.5W_n$

Based on the above condition the total area capacitance is given by

$$1 \square C_g(n - MOS) + 2.5 \square C_g(p - MOS) = 3.5 \square C_g$$

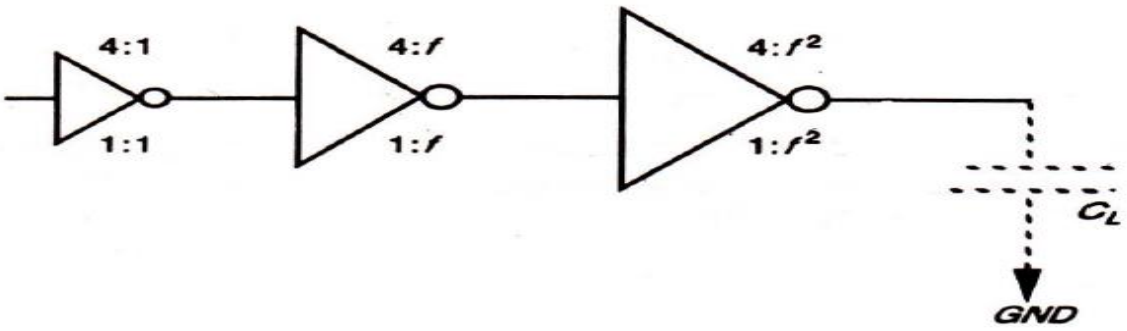
DRIVING LARGE CAPACITIVE LOADS

The problem of driving the large capacitive loads arises when the signals are propagated from chip to off chip destinations. Generally, the off-chip destinations are having high capacitance than $\square C_g$ values (on chip). The load capacitance $C_L \geq 10^4 \square c_g$. So, the capacitance of this order must be driven by low resistance to maintain less delays at the output. Large capacitance is presented at the input, which in turn slows down the rate of change of voltage at input. To achieve this there are 3 methods.

- Cascaded Inverters as drivers.
- Super buffers
- Bi-CMOS drivers.

Cascaded Inverters as Drivers:-

If the inverters are used for driving the large capacitive loads they must have low pull-up and pull-down resistances. For MOS circuits low resistance values requires low length to width ratios. To reduce $\frac{L}{w}$ ratio the channel width must be made very high to reduce the resistance. In this condition the MOS device occupies large area. The length 'L' cannot be reduced below the minimum value. In order to get the less resistance N cascaded inverters are used where width of each one is larger than the preceding stage by a width factor 'f'.



Driving large capacitive loads

As the width increases the capacitance C_g at the inverter input increases. The rate at which the width increases influence the number of stages 'N' which must be cascaded to drive the capacitive loads C_L.

For 4:1 n-MOS inverter $^{delay}/_{stage} = f\tau\ (for\ 1\ to\ 0)$

$$= 4f\tau\ (for\ 0\ to\ 1)$$

$$Total\ delay = 5f\tau\ (For\ nMOS\ inverter)$$

$$= 7f\tau\ (For\ CMOS\ inverter)$$

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

$$y = f^N \rightarrow \ln y = N \ln f$$

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

For N even, $Total\ Delay = \frac{N}{2}(5f\tau) = 2.5Nf\tau(For\ n - MOS\ inverter)$

For N even, $Total\ Delay = \frac{N}{2}(7f\tau) = 3.5Nf\tau(For\ CMOS\ inverter)$

The delay is proportional to $Nf\tau\ delay \cong \frac{\ln y}{\ln f} f\tau$

For delay to be minimum $f = 2.7 = e \rightarrow N = \frac{\ln y}{\ln f} = \frac{\ln y}{1} = \ln y$

i.e Each stage should be 2.7 times wider than its predecessor.

The delay for even number of stages is

$$For\ N - even\ ,\ T_d = 2.5eN\tau(for\ n - MOS)$$

$$T_d = 3.5eN\tau(for\ CMOS)$$

For N-Odd, 1 to 0 transition $T_d = (2.5(N-1)+1)e\tau \quad (for\ n - MOS)$

$$T_d = (3.5(N-1)+2)e\tau \quad (for\ CMOS)$$

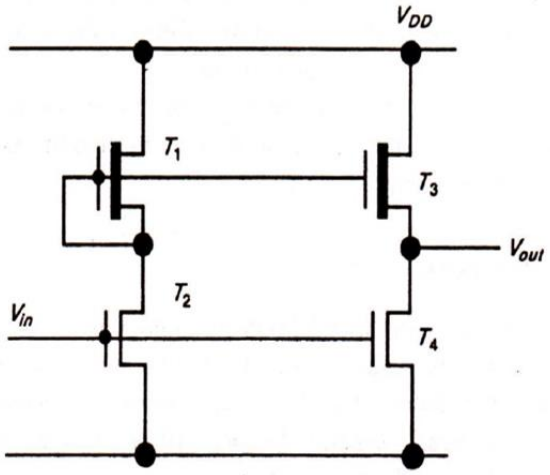
For N-Odd, 0 to 1 transition $T_d = (2.5(N-1)+4)e\tau \quad (for\ n - MOS)$

$$T_d = (3.5(N-1)+5)e\tau \quad (for\ CMOS)$$

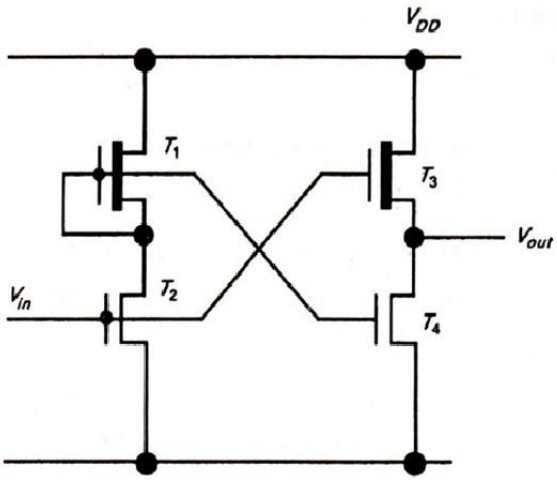
Super Buffers:-

The asymmetry of inverters rises and fall time becomes a significant problem when an inverter is used to drive large capacitive loads. The common approach to reduce this problem in n-MOS technology is to use super buffers. These are two types

- Inverting super buffer.
- Non-Inverting super buffer.



Inverting type nMOS super buffer



Non-inverting type nMOS super buffer.

Let a positive growing logic transition at the input V_{in}. The inverter formed by the transistors T_1 and T_2 is ON. Thus, the gate of transistor T_3 is pull down to logic zero with a small delay while the transistor T_4 is turned ON and the output is pull-down quickly. Let a negative going transition at the input V_{in} so the gate of transistor T_3 is allowed to rise quickly to V_{dd}. The transistor T_4 is turned OFF by the input V_{in}. Here the transistor T_3 is made to conduct with V_{dd} on its gate with twice the average voltage that would apply if the gate is connected to source. For MOS transistors, I_{ds} is proportional to V_{gs}, as V_{gs} increases the current increases which reduces the delay in charging the load capacitance. Here the symmetrical transitions are obtained.

Bi-CMOS Drivers:-

The ability of Bi-polar transistors to get a low resistance is used to drive the large capacitive loads. The availability of Bi-polar transistors in Bi-CMOS technology presents the possibility of using BJT as drivers at the output stage of inverter and logic gates. The BJT has exponential dependance of the output current on the input voltage and provides large output currents.

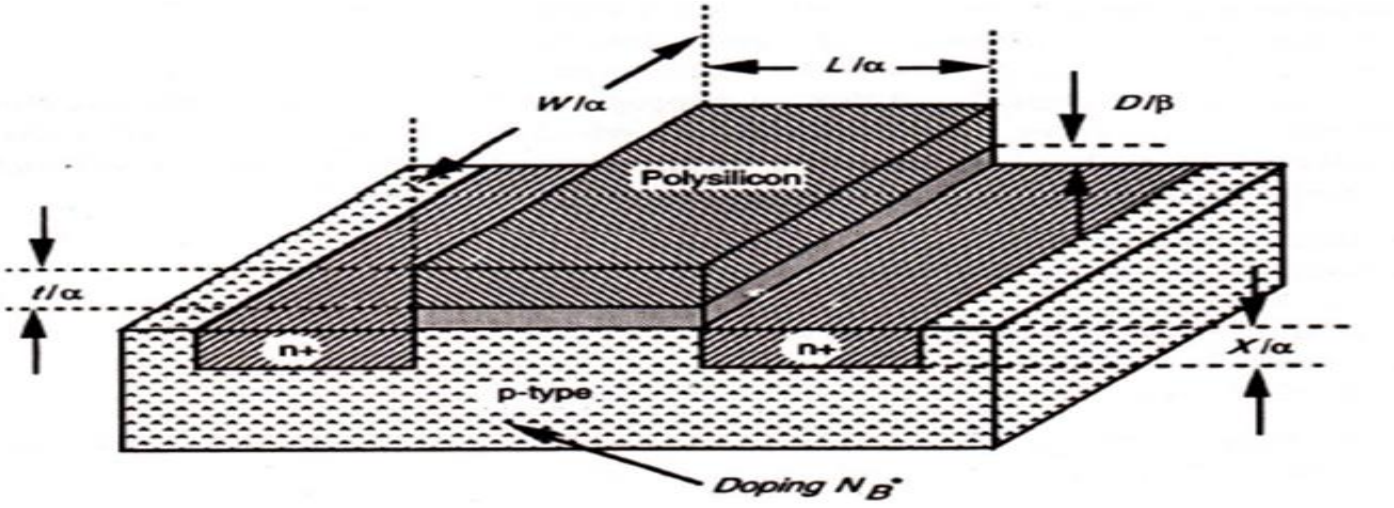
The time necessary to change the output voltage is given by $\Delta t = \frac{C_L}{g_m}$

The above time is a combination of two main components.

- T_{in}: The initial time necessary to charge the base to emitter junction.
- T_L: Time taken to charge the output load capacitance.

The total delay $T = T_{initial} + \frac{V}{I_d}\left(\frac{1}{h_{fe}}\right) C_L$

SCALING OF MOS CIRCUITS



Scaled nMOS transistor (pMOS similar).

The VLSI fabrication technology is in continuous process of evolution which leads to smaller line width and feature sizes and higher packing densities. This reduction in size is called as scaling which leads to improved performance. The performance parameters can be improved by shrinking of dimensions of interconnections, separation between layers, supply voltages and size of the transistors.

The scaling factors are:

1. **Gate Area (A_g):-** $A_g = w \times L$

   The length and width are scaled by a factor of $\frac{1}{\alpha}$. So $A_g$is scaled by a factor of $\frac{1}{\alpha^2}$.

2. **Gate capacitance per unit area $(C_0 \& C_{ox})$:-**

   $C_0 = \frac{\varepsilon_{ox}}{D} \rightarrow \varepsilon_{ox} = \varepsilon_o\ \varepsilon_{ins} = C_0 = \frac{\varepsilon_o\ \varepsilon_{ins}}{D}$

   D is thickness of gate oxide which is scaled by a factor of $\frac{1}{\beta}$

   $C_0 = \frac{1}{1/\beta} = \beta$ (C_0 is scaled by β)

3. **Gate Capacitance (C_g):-** $C_g = \frac{C_0 A}{D} = C_{0x}(WL)$ C_g is scaled by β/α^2

4. **Parasitic Capacitance:-** (C_x)

   The parasitic capacitance C_x is proportional to $\frac{A_x}{D}$ where d is depletion width around source and drain which is scaled by $1/\alpha$. A_x is area of depletion region which is scaled by $1/\alpha^2$. So C_x is scaled by $1/\alpha$.

5. **Charge density in the channel:-** (Q_{on})

   Charge Density $Q_{on} = C_{ox}V_{gs}$

   Capacitance C_{ox} is scaled by β and $V_{gs}$is scaled by width of SiO_2 layer

$$Q_{on} = \beta\frac{1}{\beta} = 1$$

   Q_{on} is scaled by 1.

6. **Channel Resistance (R_{on}):-**

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

   Here mobility μ does not change. The channel resistance R_{on} is scaled by $\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}}.\frac{1}{1} = 1$

7. **Gate Delay (τ_d):-**

   $\tau_d \propto R_{on}C_g$ τ_d is scaled by $1.{}^{\beta}/_{\alpha^2}$

8. **Operating Frequency (f_0):-**

   The frequency $f_0 = \frac{W}{L}\frac{\mu C_0 V_{dd}}{C_g}$. Since operating frequency $f_0 \propto \frac{1}{\tau_d}$ So f_0 is scaled by ${}^{\alpha^2}/_{\beta}$

9. **Saturation Current (I_{dss}):-**

   For a MOS device saturation current is assessed to $I_{dss} = \frac{\mu C_{ox}}{2}\frac{W}{L}\left(V_{gs} - V_t\right)^2$

   Since V_{gs} & V_t are scaled by ${}^{1}/_{\beta}$

   I_{dss} is scaled by $\beta \times \frac{1}{\beta^2} = \frac{1}{\beta}$

10. **Current Density(J):-**

    $J = \frac{I_{dss}}{A}$ where A is area of channel.

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

11. **Switching energy per gate(E_g):-**

    The energy per gate is $E_g = \frac{1}{2}C_g{V_{dd}}^2$

    $E_g$is scaled by ${}^{\beta}/_{\alpha^2} \times \frac{1}{\beta^2} = \frac{1}{\alpha^2\beta}$

12. **Power dissipation per gate(P_g):-**

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

    Static component $P_{gs} = \frac{{V_{dd}}^2}{R_{on}}$

    Dynamic Component $=E_g f_0$

    P_g is scaled by $\frac{1}{\beta^2}$ (or) $\frac{1}{\alpha^2\beta} \times \frac{\alpha^2}{\beta} = \frac{1}{\beta^2} + \frac{1}{\beta^2} = \frac{1}{\beta^2}$

13. **Power dissipation per unit area(P_a):-**

$$P_a = \frac{P_g}{A} = \frac{1}{\beta^2} \times \frac{1}{\frac{1}{\alpha^2}} = \frac{\alpha^2}{\beta^2}$$

14. **Power speed product (P_T):-**

$$P_T = P_g \times \tau_d$$

P_T is scaled by $\frac{1}{\beta^2} \times \frac{\beta}{\alpha^2} = \frac{1}{\alpha^2 \beta}$

CHOICE OF DIFFERENT LAYERS

To design an arrangement of layout to meet given specifications there are several possible ways of choosing different layers on which to group certain data & control signal. To choose the layers there are different conditions.

1. V_{dd} and V_{ss} should be distributed on metal layers only.
2. Long lengths of poly silicon should be used only after careful considerations because of its high R_s value on the polysilicon layer.
3. The capacitive effects are to be considered where fast signal lines are required in relation to the signals on wiring which are having relative values of R_s.
4. The diffusion layers have relatively high values of capacitance to the substrate within each layer. The delay associated with signal propagation is small in comparison with gate delays in systems connected by the wires.

LIMITATIONS OF SCALING

1. Limitations on Substrate doping: –

The effect of barrier potential is neglected when barrier potential is much smaller than V_{dd}. When this condition is not satisfied as a result of scaling the effect of V_b is to be considered. The V_b shows its effect on the scaling factor corresponding to the substrate doping. As the channel length is reduced the depletion region width is also scaled down. The depletion width between the junction is given by

$$d = \sqrt{\frac{2\varepsilon_0 \varepsilon_{si} V}{qN_B}}$$

Where N_B is doping level of substrate

V is effective voltage across the junction

$V = V_a + V_b$

V_a is applied voltage

V_b is junction potential$= \frac{KT}{q} \ln\left(\frac{N_D N_B}{n^2}\right)$

If V_{dd} is scaled by $\frac{1}{\beta}$ and depletion width is scaled by $\frac{1}{\alpha}$ then the substrate doping is scaled by $\frac{\alpha^2}{\beta}$

$$d = \sqrt{\frac{2\varepsilon_0 \varepsilon_{si} V}{q N_B}} \rightarrow \frac{1}{\alpha^2} = \frac{1/\beta}{N_B} \rightarrow N_B = \frac{\alpha^2}{\beta}$$

If N_B is increased to reduce 'd' and V_B is also increased. If V_{dd} is scaled down then V_B is not much less than V_{dd} because of scaling. So, a careful scaling of the substrate doping is done to increase the substrate doping by a factor of $\frac{\alpha^2}{\beta}$.

2. **Limitations of Miniaturization:-**

The minimum size of transistor is determined by the process technology of the device. The reduction size of device geometry depends on alignment accuracy of layout. The size of the transistor is defined in terms of channel length 'L'. As the channel length is scaled down the edges of depletion region of source becomes closer around the drain. To maintain the perfect action of the transistor channel length 'L' must be at least 2d. i.e $L \approx 2d$. The minimum transit time for an electron to travel from source to drain is given by $\tau = \frac{2d}{v_{drift}}$.

3. **Limits of interconnections & Contact resistance:-**

Since the width, thickness & spacing of interconnections are scaled by $1/\alpha$ and their cross-sectional area is scaled by $1/\alpha^2$, the conductor length is also scaled by $1/\alpha$. So, the resistance is increased by a factor of α which reduces the driving capability and noise margins. By increasing the interconnections both resistance and capacitance of interconnects increases which produces much larger time constant. This problem is reduced by making use of multi-layer interconnection with wider conductors & thin separating layers.

4. **Limitations due to Subthreshold currents:-**

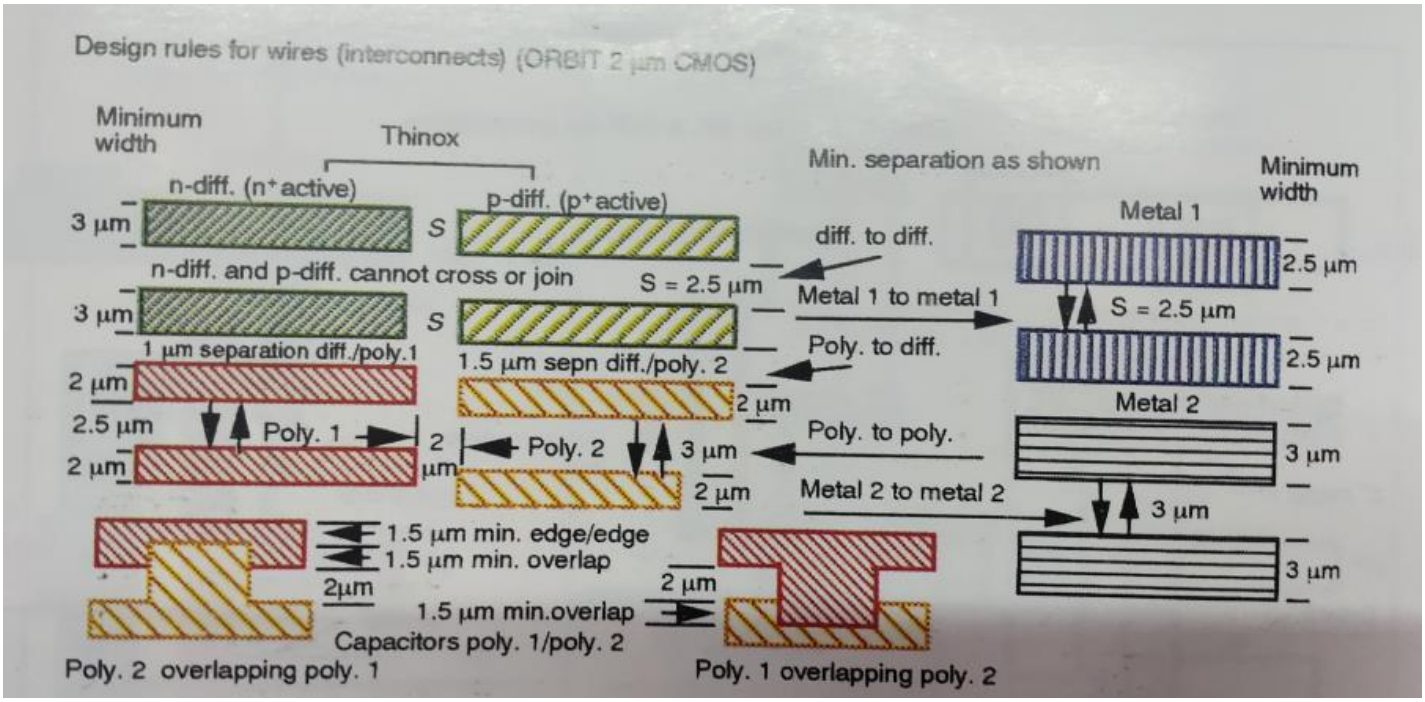
We know the subthreshold current I_{sub} is directly proportional to $e^{\frac{(V_{gs}-V_t)^2}{kT}}$. As voltages are scaled down, the ratio of $(V_{gs} - V_t)$ to kT will reduce so that subthreshold current increases. So V_{gs} and V_t are scaled down together with V_{dd}.

5. **Limitations due to Current Density:-**

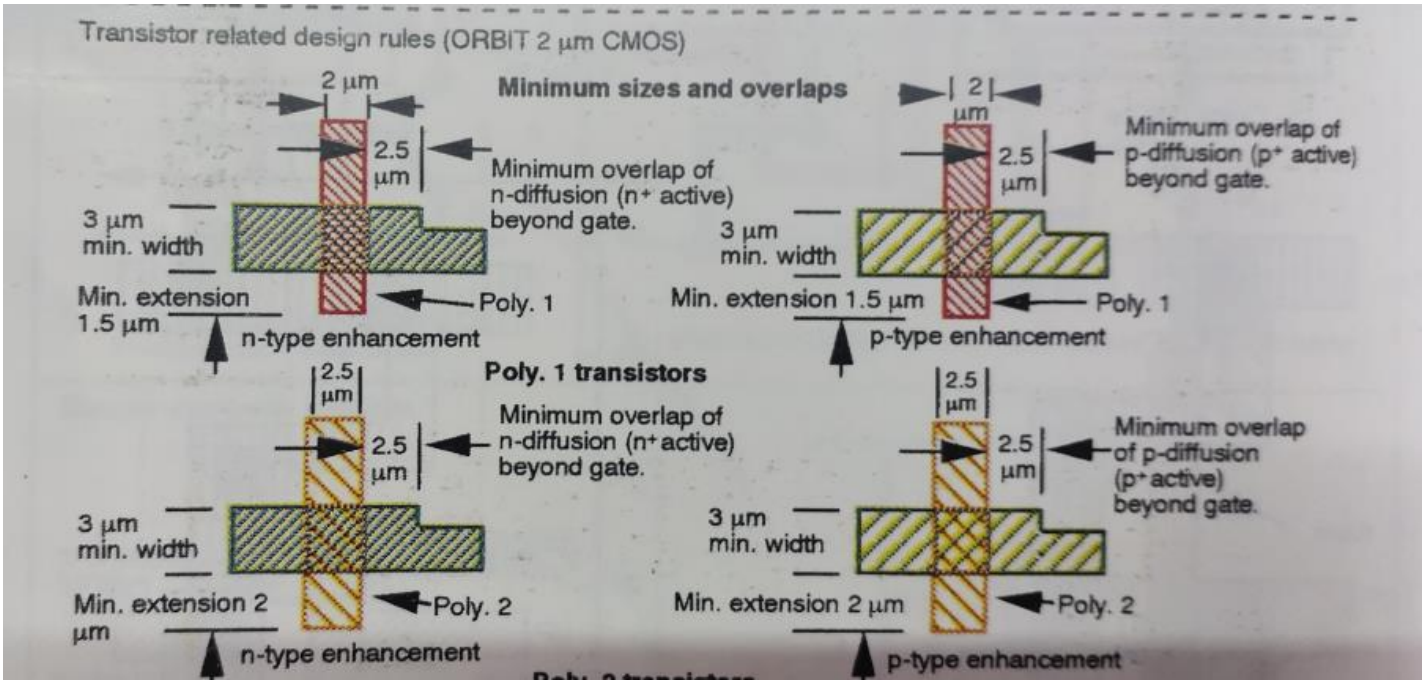
When the current density in aluminum approaches $10^6\ A/cm^2$ $\left(10\ mA/\mu m^2\right)$ the interconnects are likely to be burned off owing to metal migration. Thus, allowable current densities are set well below this limit and order of $J = 1\ to\ 2\ mA/\mu m^2$ are commonly used.

2µm CMOS Design rules for wires, Contacts and Transistors

Design rules for wires & interconnects:-



Design rules for Transistors:-



Design rules for Contact Cuts:-