MODULE 5 - PART C

DELAY MODELS OF CMOS GATES AND INTERCONNECTS

In the last part of MODULE 5 (PART B), we had derived the following expression for the PROPAGATION DELAY in terms of the widths W_L (LOAD INVERTER) and W_d (DRIVER INVERTER) :-

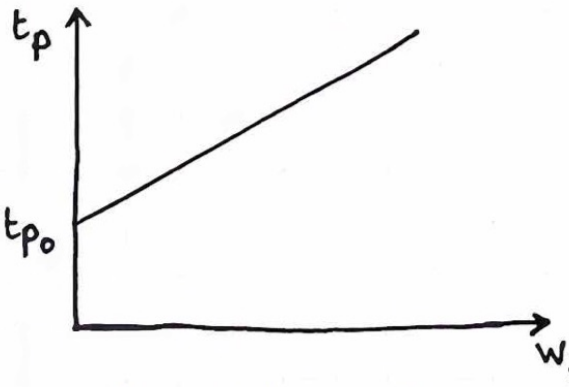
$$t_p = \frac{\tilde{C}_{OUT}}{2V_{DD}\tilde{\beta}} + \frac{C_{WIRE} + W_L\tilde{C}_{IN}}{2V_{DD}W_d\tilde{\beta}} \quad \ldots(i)$$

We will now look at the role of sizing the LOAD INVERTER and the DRIVER INVERTER on the propagation delay.

Let us start by looking at the effect of sizing the LOAD INVERTER, that is increasing W_L on the propagation delay.

(a) EFFECT OF W_L on t_p :-

First let us draw the plot of t_p vs W_L based on equation (i)



Here, the plot of t_p vs W_L is a straight line with $t_p = t_{p_0}$ when $W_L = 0$. Where,

$$t_{p_0} = \frac{\tilde{C}_{OUT}}{2V_{DD}\tilde{\beta}} + \frac{C_{WIRE}}{2V_{DD}W_d\tilde{\beta}} = t_{pi} + t_{pwire}$$

where, $t_{pi} = \frac{\tilde{C}_{OUT}}{2V_{DD}\tilde{\beta}}$

and $t_{pwire} = \frac{C_{WIRE}}{2V_{DD}W_D\tilde{\beta}}$

t_{p_0} represents the case where there is no LOAD INVERTER (hence, $W_L = 0$)

t_{pi} represents the intrinsic delay of an inverter when it is driving its own output capacitance. This is the minimum delay of an inverter and cannot be avoided.

t_{pwire} represents the propagation delay due to the interconnect.

(b) EFFECT OF W_d on t_p :-

First let us draw the plot of t_p vs W_d based on equation (i). Here, we can divide the range of W_d into 2 parts.

(i) W_d is small →

Here, $t_p = \frac{\tilde{C}_{OUT}}{2V_{DD}\tilde{\beta}} + \frac{C_{WIRE} + W_L\tilde{C}_{IN}}{2V_{DD}W_d\tilde{\beta}}$

or, $t_p \simeq \frac{C_{WIRE} + W_L\tilde{C}_{IN}}{2V_{DD}W_d\tilde{\beta}}$

Since, W_d is very small,

$\therefore \frac{C_{WIRE} + W_L\tilde{C}_{IN}}{2V_{DD}W_d\tilde{\beta}} \gg \frac{\tilde{C}_{OUT}}{2V_{DD}\tilde{\beta}}$

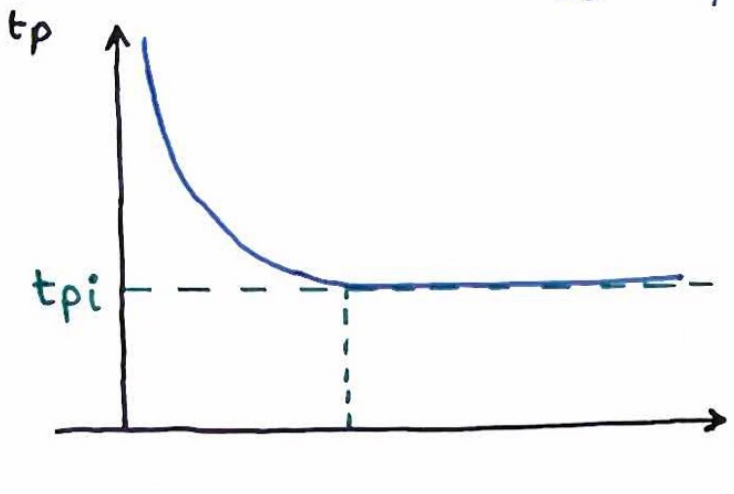
(ii) W_d is large →

Here, $t_p = \frac{\tilde{C}_{OUT}}{2V_{DD}\tilde{\beta}} + \frac{C_{WIRE} + W_L\tilde{C}_{IN}}{2V_{DD}W_d\tilde{\beta}}$

or, $t_p \simeq \frac{\tilde{C}_{OUT}}{2V_{DD}\tilde{\beta}} = t_{pi}$ (intrinsic delay)

since, W_d is very large,

$$\therefore \frac{\tilde{C}_{OUT}}{2V_{DD}\tilde{\beta}} >> \frac{C_{WIRE} + W_L \tilde{C}_{IN}}{2V_{DD} W_d \tilde{\beta}}$$



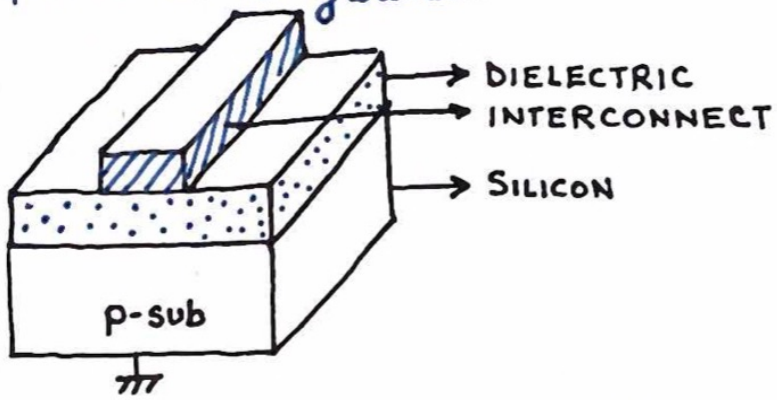
From this plot, we can see that when W_d is small $t_p \propto \frac{1}{W_d}$ and hence upsizing the driver (increasing W_d) decreases t_p and improves performance. This is the region when the load capacitance C_L is dominated by C_{WIRE} and C_{IN}. However, upsizing the driver beyond a point has diminishing returns as $C_{OUT} \propto W_d$

Hence, if the driver is upsized, then the propagation delay will first decrease and then saturate to t_{pi}. So, upsizing the driver beyond a point will not deliver higher performance.

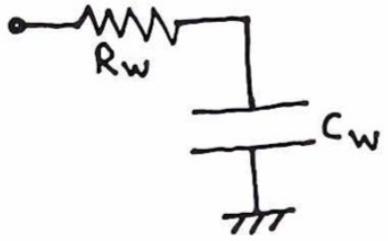
ELMORE DELAY :-

So far we have looked at the delay of inverters. We have also seen that the capacitance of the wire is C_{WIRE}. Now, we will look into this capacitance a little closely.

We know that every interconnect has a capacitance to ground.

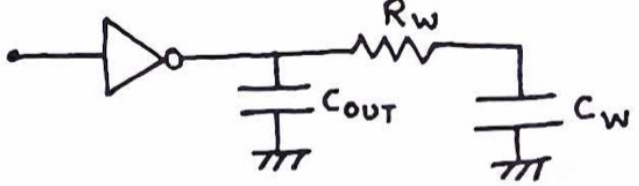


Interconnects also have capacitances to other interconnects, but we will ignore this in our discussions here. Since interconnects are fabricated using metallic material, they also have resistances associated with them. ∴ wire can thus be thought of as a combination of resistance (R_w) and capacitance (C_w)

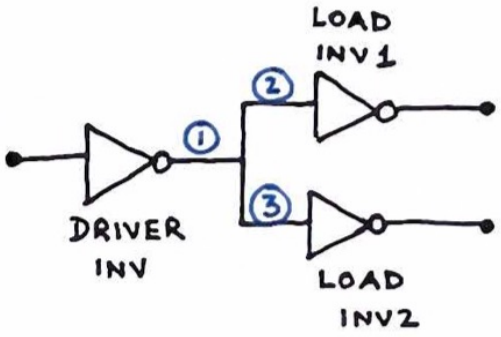


For short interconnects we typically ignore R_w and only consider C_w as we have seen before. Now, let us consider two scenarios :-

(A) Consider the situation where an inverter drives a LONG INTERCONNECT.

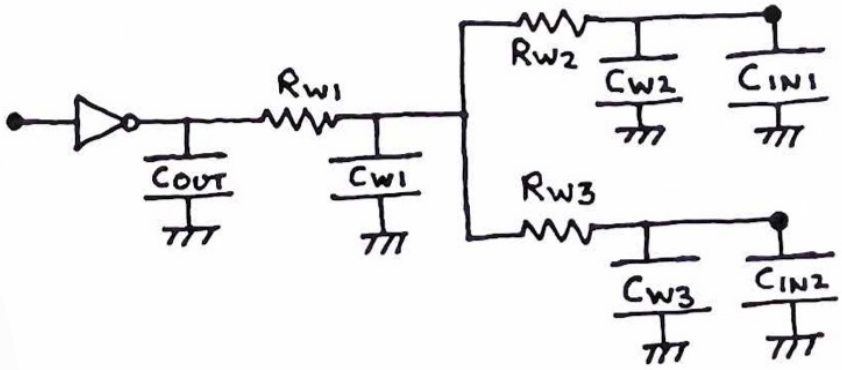


(B) Consider the situation where one interconnect connects to multiple other interconnects that have different loads at their outputs.



① → INTERCONNECT 1
② → INTERCONNECT 2
③ → INTERCONNECT 3

This is equivalent to :-



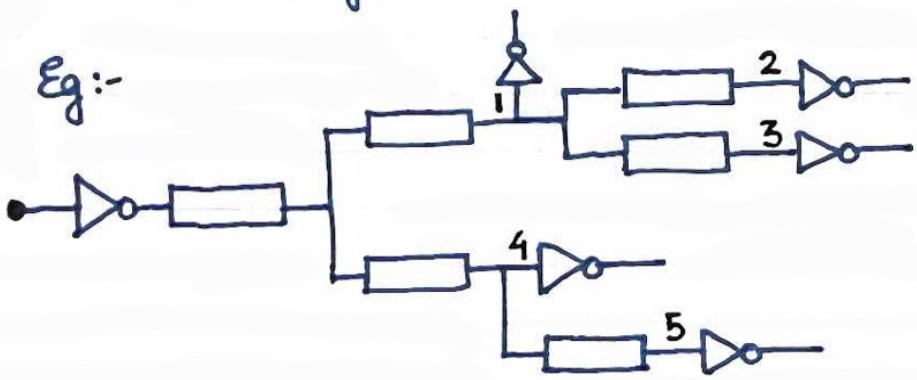
Here, we assume that the interconnect 1 is represented by R_{W1}, C_{W1}, interconnect 2 is represented by R_{W2}, C_{W2} and interconnect 3 is represented by R_{W3}, C_{W3}. The load inverter 1 offers a load capacitance of C_{IN1} and the load inverter 2 offers a load capacitance of C_{IN2}.

In both of these scenarios, we need to calculate the delay of a distributed system of R-C segments. This is done analytically using the ELMORE'S DELAY MODEL.

ELMORE'S DELAY MODEL states that for tree-structured networks, the delay through each segment is R times the downstream C and the total delay is the SUM of the delays of each of these segments from the SOURCE to the SINK.

TREE-STRUCTURED NETWORKS → An interconnect tree structure is a network of interconnect segments where each input (also called SOURCE) can propagate signals to multiple outputs (also called SINK) through the network structure.

Eg:-



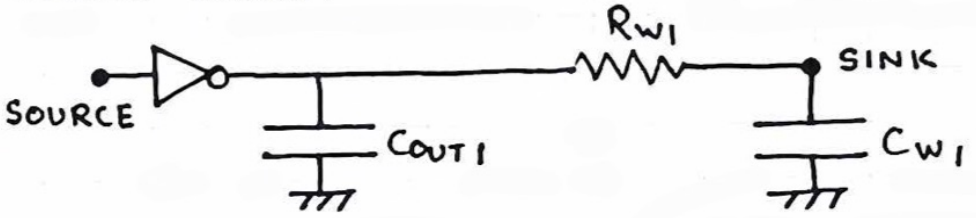
1 SOURCE, 5 SINKS

2 SOURCES, 1 SINK

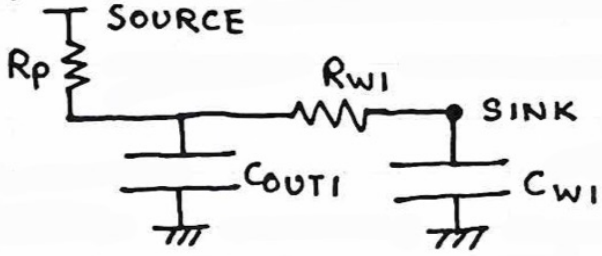
The DOIVNSTREAM C, refers to all the capacitances that appear between the resistance R and the SINK.

Let us use the ELMORE MODEL for situation (A) where an inverter drives a LONG INTERCONNECT.

Let us consider a case where the inverter is charging the total output capacitances as shown below :-



This is equivalent to the following fegure since charging is done through the pMOS of the inverter. (Rp is the lumped resistance of the pMOS)



The downstream capacitance for Rp is $C_{OUT1} + C_{W1}$.

∴ The delay associated with the resistance Rp is :- $T_1 = 0.69\, R_p (C_{OUT1} + C_{W1})$

(We know the low to high propagation delay is given by $t_{PLH} = 0.69\, R_p C_L$ and here $C_L = C_{OUT1} + C_{W1}$)

The downstream capacitance associated with R_{W1} is C_{W1}.

∴ The delay associated with the resistance R_{W1} is similarly given by :-

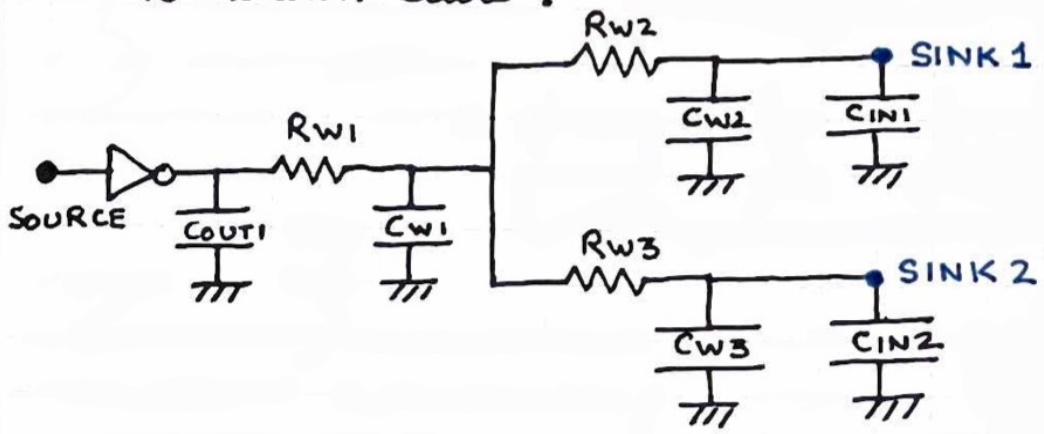
$$T_2 = 0.69\, R_{W1} C_{W1}$$

∴ The total charging delay here, is

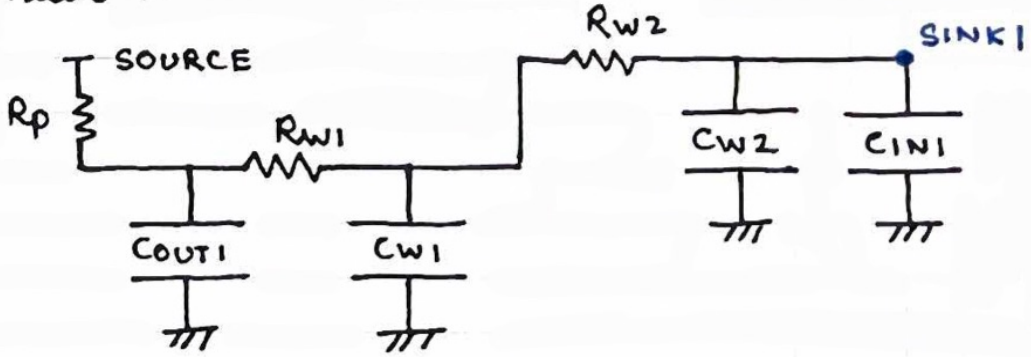
$$T = T_1 + T_2$$

or, $$T = 0.69 \left[R_p (C_{OUT1} + C_{W1}) + R_{W1} C_{W1} \right]$$

Now, let us look at the ELMORE MODEL for scenario (B) where one interconnect connects to multiple other interconnects that have different loads at their outputs. This is shown below :-



If we look at the charging delay from the DRIVER INVERTER to SINK 1, we have :-



Here, the total delay from the source to the SINK1 will have 3 components. The number of these delay components will be equal to the number of resistors between the SOURCE and the SINK.

$$T_1 = 0.69\left[R_p\left(C_{OUT1} + C_{W1} + C_{W2} + C_{IN1}\right)\right]$$

$$T_2 = 0.69\left[R_{W_1}\left(C_{W1} + C_{W2} + C_{IN1}\right)\right]$$

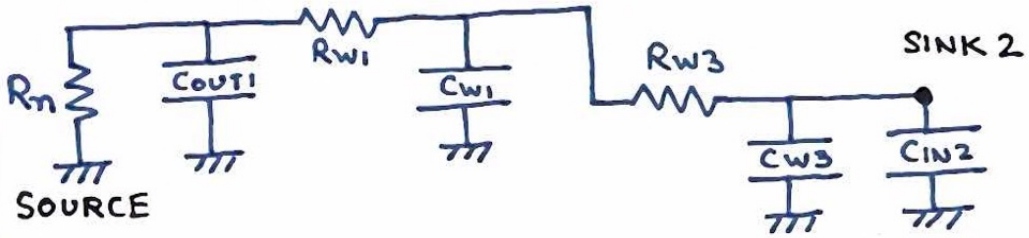
$$T_3 = 0.69\left[R_{W_2}\left(C_{W2} + C_{IN1}\right)\right]$$

Hence, the TOTAL DELAY :-

$$T = T_1 + T_2 + T_3$$

Similarly, we can find the total charging delay from SOURCE to SINK 2.

Now, let us look at the DISCHARGING delay from SOURCE to SINK 2. Here, the nMOS of the DRIVER inverter will discharge all the capacitors between the SOURCE and SINK 2.



Since, there are 3 resistors between the SOURCE and SINK 2, we will have 3 components here too.

∴ The discharging delays will be given by $t_{PHL} = 0.69 R_n C_L$.

$$\therefore T_1' = 0.69 R_n (C_{OUT1} + C_{W1} + C_{W3} + C_{IN2})$$

$$T_2' = 0.69 R_n (C_{W1} + C_{W3} + C_{IN2})$$

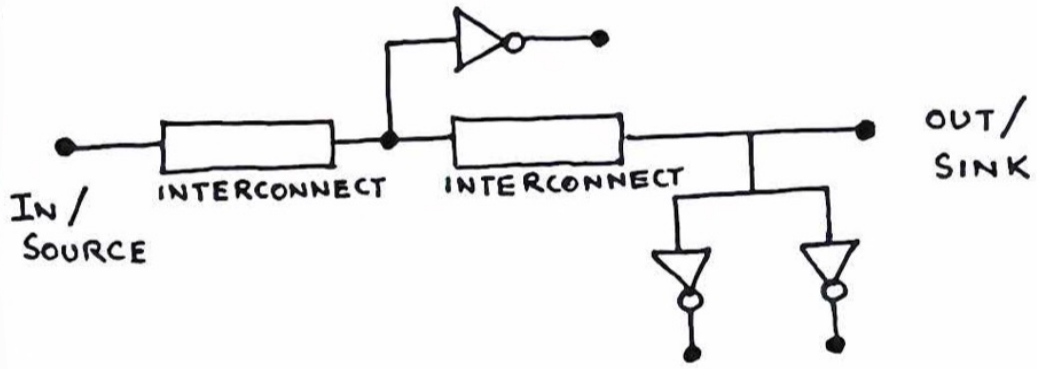
$$T_3' = 0.69 R_n (C_{W3} + C_{IN2})$$

Hence, the total discharging delay from the SOURCE to the SINK 2 is given by :-

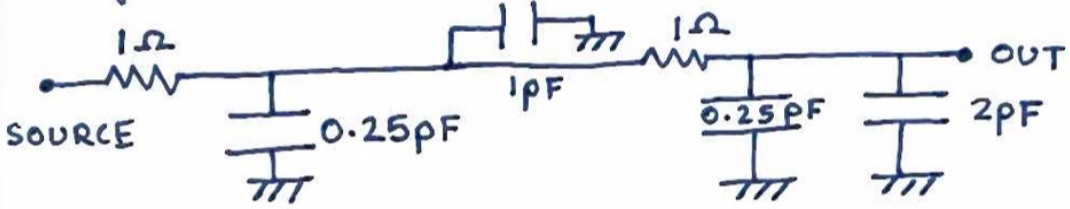
$$T' = T_1' + T_2' + T_3'$$

Let us now look at an example.

EXAMPLE :- Assume that the capacitance at the gate of an inverter is 1 pF. Find the total delay using ELMORE'S DELAY MODEL if each interconnect section is ($R_W = 1\Omega$ and $C_W = 0.25$ pF)



SOLUTION :- The equivalent RC model of the given network is :-



Using ELMORE'S DELAY MODEL, we can write,

$$T_1 = 0.69\left[1(0.25+1+0.25+2)\right]$$

$$= 0.69(3.5)\ ps$$

$$T_2 = 0.69\left[1(0.25+2)\right]$$

$$= 0.69(2.25)\ ps$$

$\therefore$ The total delay $T = 0.69(3.5+2.25)ps$

or, $T = 0.69(5.75)ps$

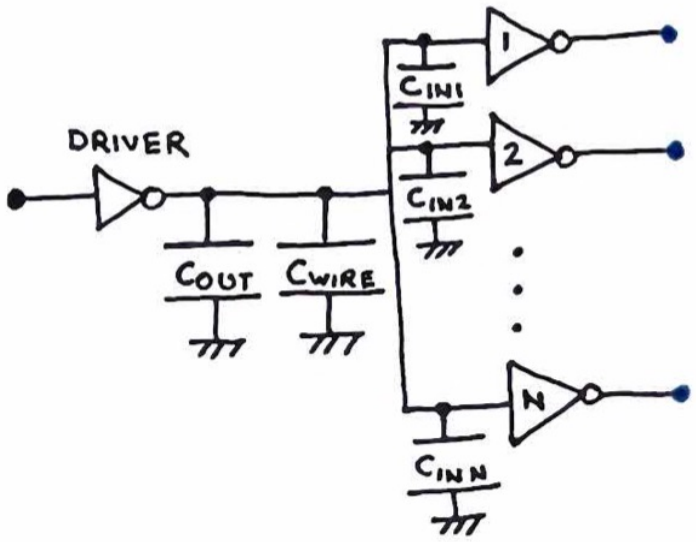
or, $\boxed{T = 3.97\ ps}$

EFFECT of FAN-IN and FAN-OUT on DELAY

So far we have looked at the delay characteristics of inverters and interconnects Now, we will look at the effect of fan-in and fan-out on delay.

First let us consider the effect of fan-out.

Let us look at the case where one inverter drives N inverters (FAN OUT = N)



Here, assuming the N inverters are identical,

$$\therefore C_{IN1} = C_{IN2} = \ldots = C_{INN} = C_{IN}$$

$$\therefore C_L = C_{OUT} + C_{WIRE} + NC_{IN}$$

Hence, the propagation delay is given by :-

$$t_p = \frac{C_L}{2V_{DD}} \left[\frac{1}{\beta_p} + \frac{1}{\beta_n}\right]$$

or,

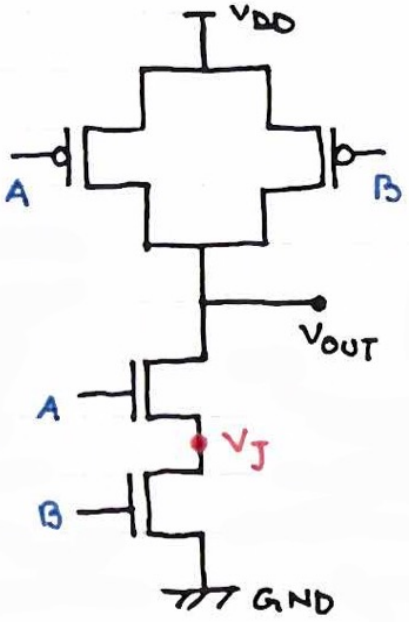
$$t_p = \left[\frac{C_{OUT} + C_{WIRE}}{2V_{DD}} \left(\frac{1}{\beta_p} + \frac{1}{\beta_n}\right)\right] + N\left[\frac{C_{IN}}{2V_{DD}} \left(\frac{1}{\beta_p} + \frac{1}{\beta_n}\right)\right]$$

Thus, we see that the delay of an inverter increases linearly with N.

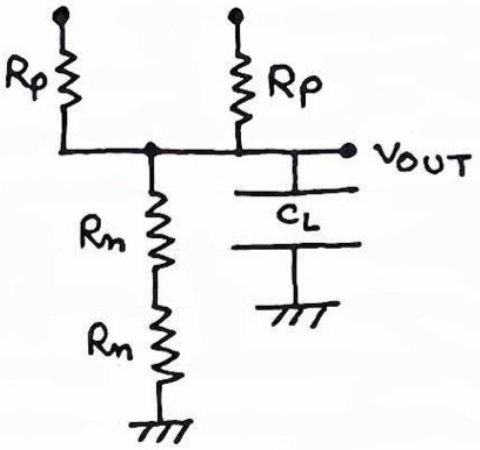
To reduce the delay as N increases, we need to decrease β_p and β_n. This can be done by increasing the width W_d i.e. by upsizing the DRIVER INVERTER.

To understand the effect of fan-in, we have to look at a complex gate.

Let us look at a 2-input NAND gate (NAND2) :-



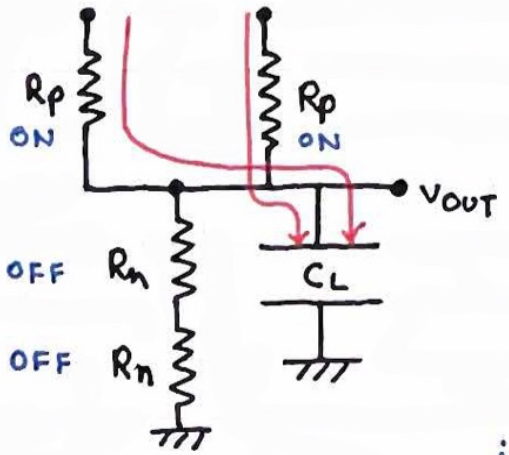
We can draw the resistance diagram for this NAND2 as below :-



Here, C_L is the output capacitor.

In a 2-input NAND gate, when the output capacitor C_L is charging, we can have one of two scenarios :-

(i) BOTH PMOS DEVICES ARE ON :-

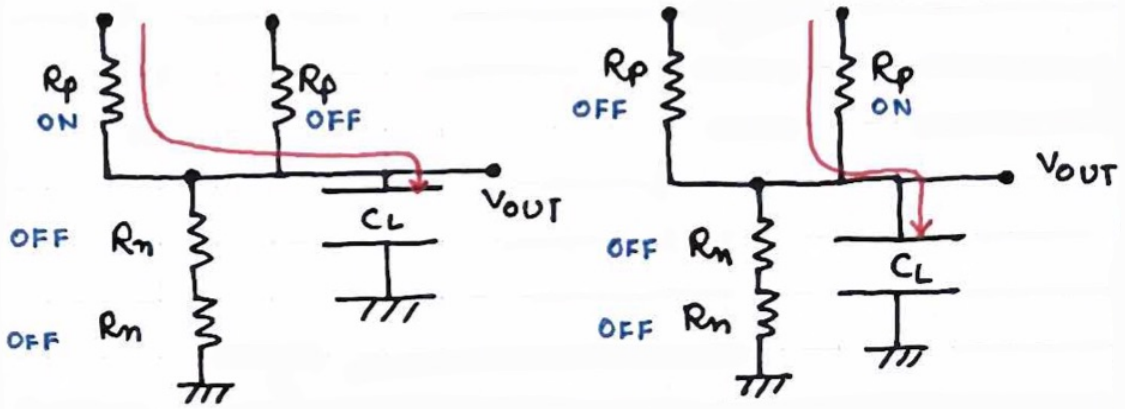


Here, the effective resistance,

$$\tilde{R_p} = R_p \,||\, R_p = \frac{R_p}{2}$$

$\therefore$ The charging delay is given by $0.69 \frac{R_p}{2} C_L$

(ii) ONLY ONE PMOS IS ON :-



In this case, the effective resistance,

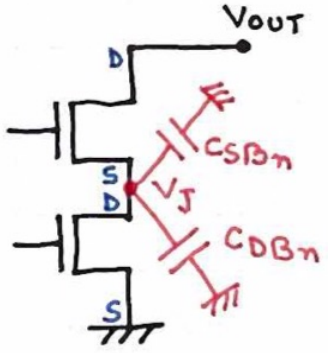
$$\tilde{R_p} = R_p$$

∴ The charging delay is given by,

$$0.69\, R_p C_L$$

Hence, the charging delay here is input dependant.

The discharging process here is a little more complex.
Apart from the total capacitor C_L at V_{OUT}, there will be an additional capacitance at the node V_J.

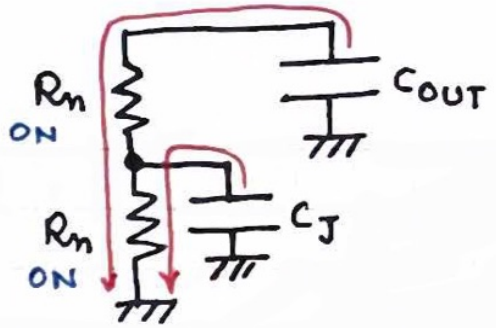


This capacitance is the total JUNCTION capacitance at V_J.

$$\therefore C_J = C_{SBn} + C_{DBn}$$

When V_{OUT} is discharged (that is both of the nMOSes are ON), the junction capacitance C_J also needs to be discharged

∴ The equivalent RC MODEL will be :-



Here, C_{OUT} is discharged through the 2 Rn resistors and C_J is discharged through 1 Rn resistor.

Using ELMORE'S DELAY MODEL, the total discharging delay is :-

$T_1 = 0.69\, R_n\, (C_J + C_{OUT})$

$T_2 = 0.69\, R_n\, C_{OUT}$

or, $T = T_1 + T_2$

or, $T = 0.69\, R_n\, (C_J + 2C_{OUT})$

Here, the discharging delay is input independant.

If we do a similar analysis for a NOR2 gate, we will find that the discharging delay is input dependant and the charging delay is input independant.
(Try to do it yourself)

We can thus, analyze the total delay of a complex gate using the ELMORE'S DELAY MODEL.

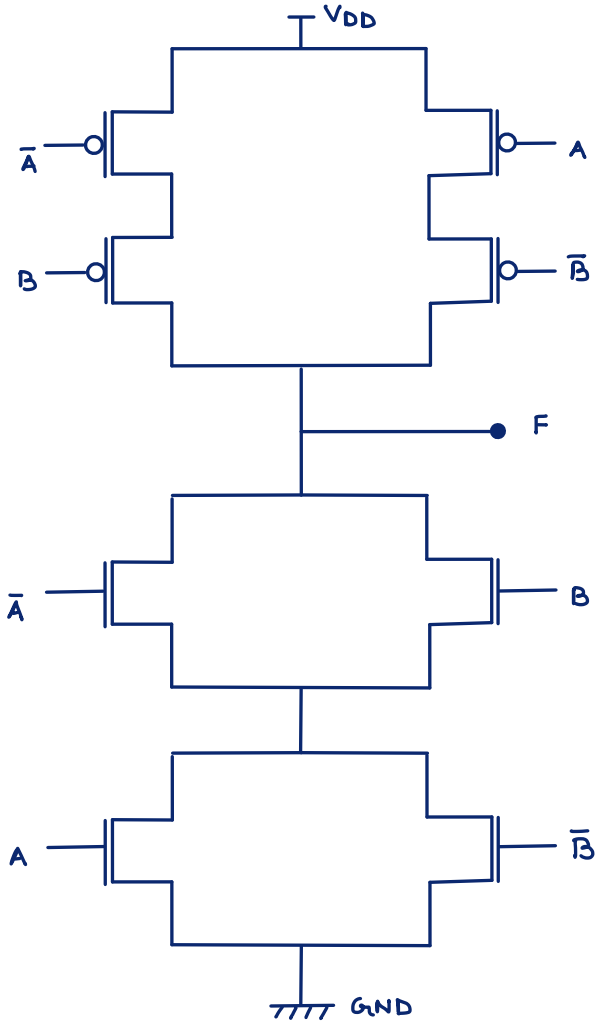
EXAMPLE :- Find the charging and discharging delay of an XOR gate.

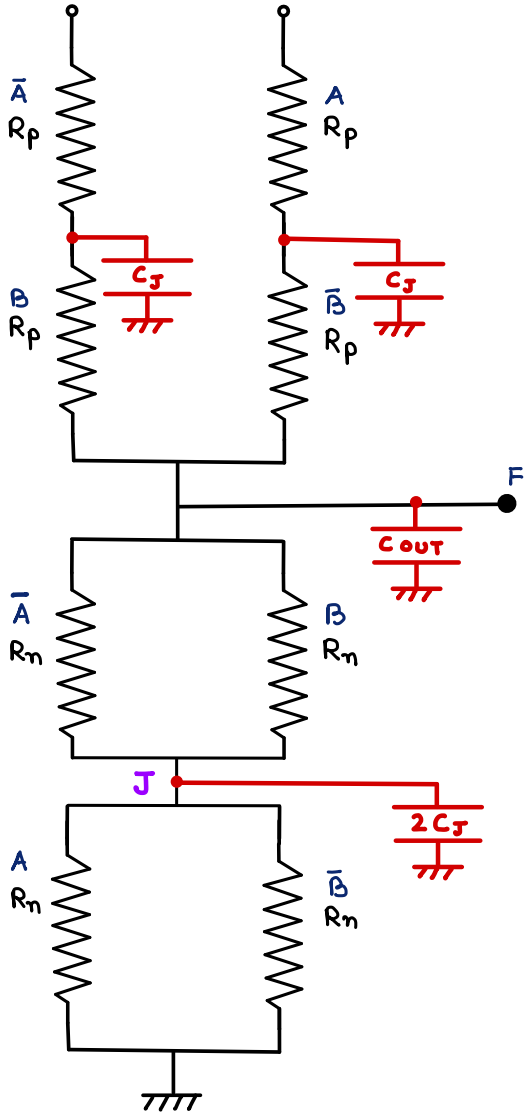
Let us first look at the traditional CMOS implementation of an XOR gate.

The BOOLEAN expression of an XOR gate is given by :-

$$F = A\bar{B} + \bar{A}B$$

The traditional CMOS implementation and its corresponding resistance diagram with the capacitances is shown below :-





Here we have assumed that the transistors are all symmetric i.e. all pMOSes have the same lumped resistance Rp and, all nMOSes have the same lumped resistance Rn

The junction capacitance $C_{Jp} = C_{DBp} + C_{SBp}$

is equal to $C_{Jn} = C_{DBn} + C_{SBn}$

i.e. $C_{Jp} = C_{Jn} = C_J$

Node 'J' is the junction of 4 transistors (instead of 2).

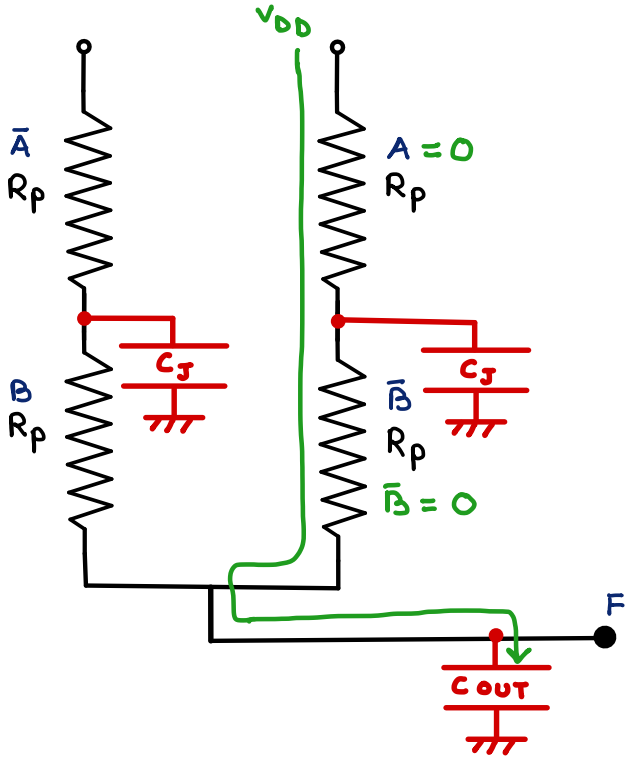
∴ The total junction capacitance here will be $2C_J$.

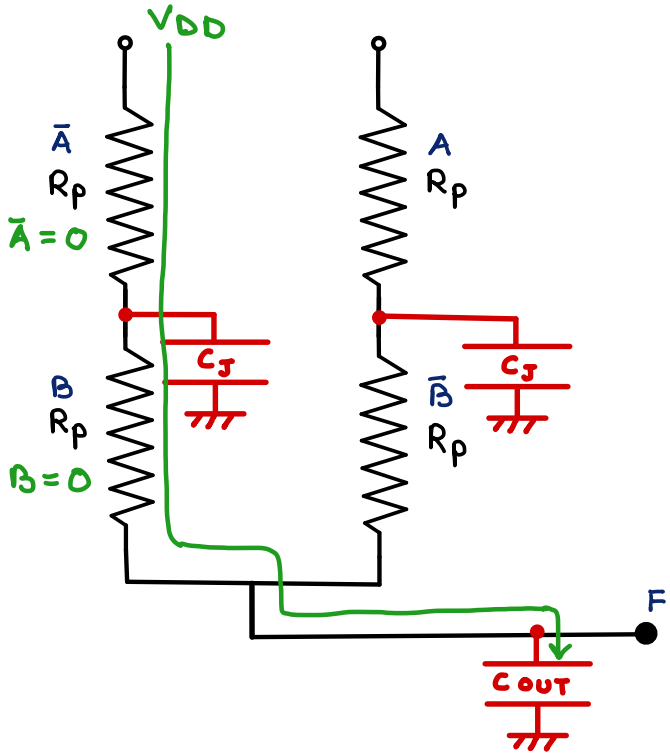
CHARGING DELAY :-

There are two different charging paths for this circuit :-

When A = 0 ; B = 1 and when A = 1 ; B = 0.

A	B	F	
0	0	0	
0	1	1	→ CHARGING PATH 1
1	0	1	→ CHARGING PATH 2
1	1	0	





The charging delays for both of these paths are the same and is given by :-

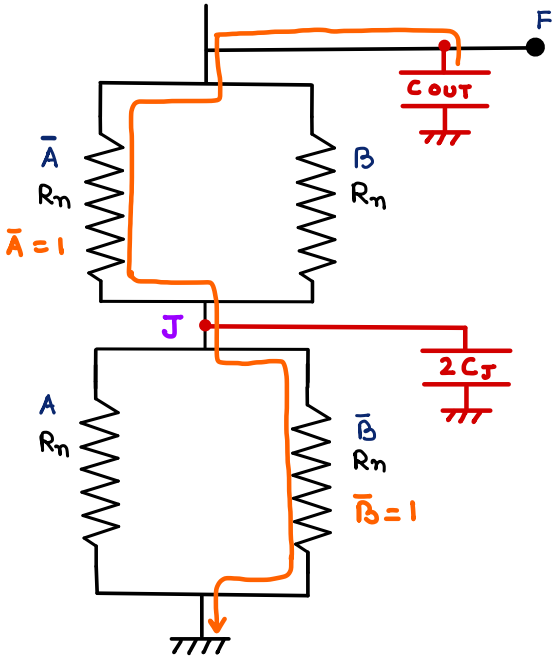
$$T = 0.69\left[R_p\left(C_J + C_{OUT}\right) + R_p C_{OUT}\right]$$

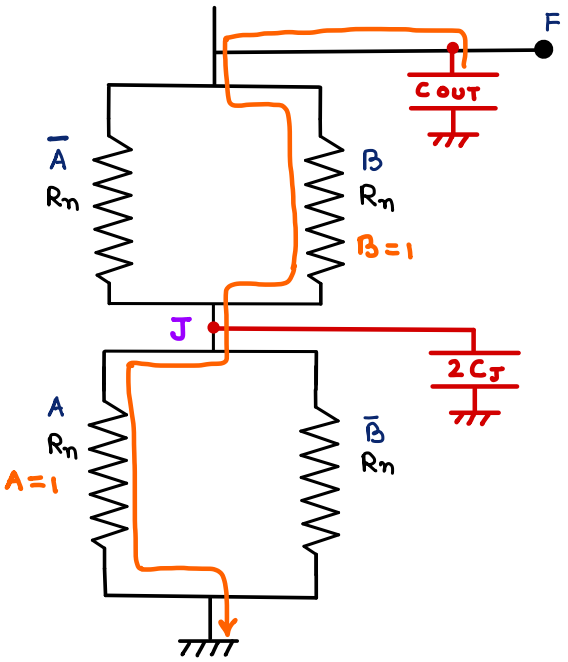
DISCHARGING DELAY :-

There are two different discharging paths for this circuit :-

When A = 0 ; B = 0 and when A = 1 ; B = 1

A	B	F	
0	0	0	→ DISCHARGING PATH 1
0	1	1	
1	0	1	
1	1	0	→ DISCHARGING PATH 2





The discharging delays for both of these paths are the same and is given by :-

$$T = 0.69\left[R_n(2C_J + C_{OUT}) + R_n C_{OUT}\right] \quad \ldots(i)$$

Now let us find the charging and the discharging delays of the MIRROR CIRCUIT implementation of the XOR gate. For the Pull-up network (PUN) the circuit will be exactly the same as the traditional CMOS implementation. Hence, the charging delays for the 2 charging paths will

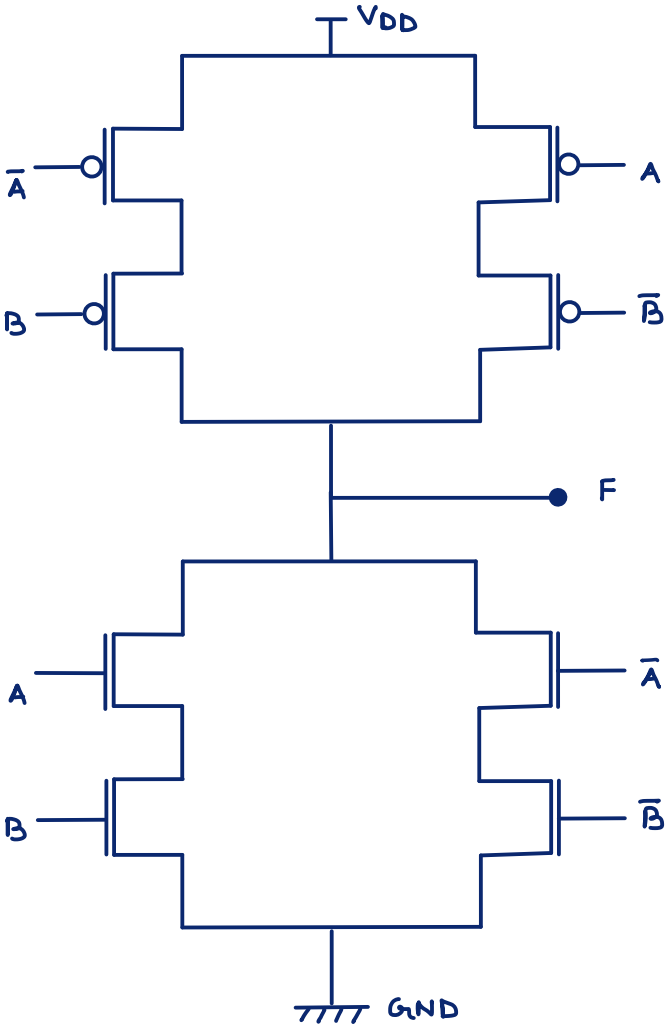
also be the same.

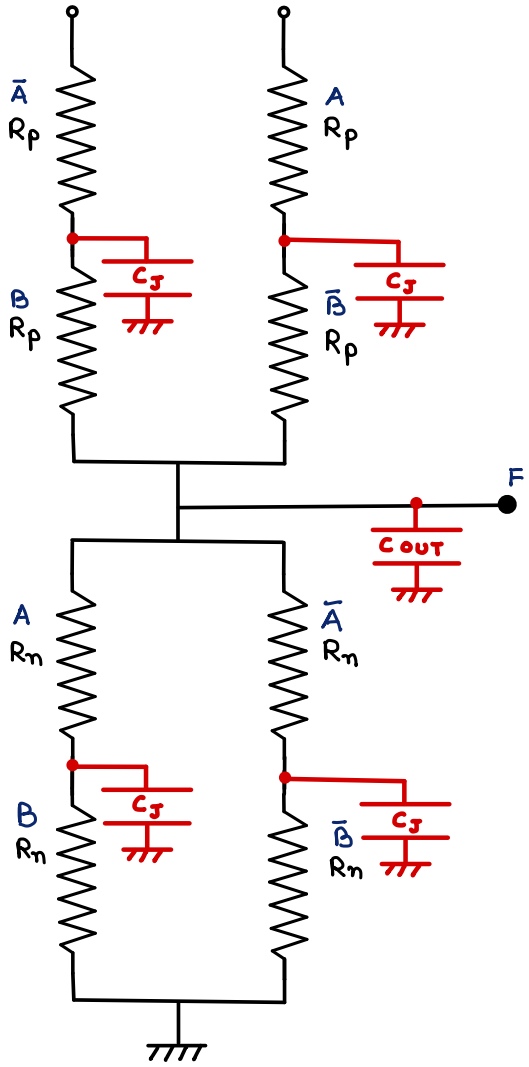
$$\text{i.e.} \quad T = 0.69\left[R_p\left(C_J + C_{OUT}\right) + R_p C_{OUT}\right]$$

For the Pull-down network (PDN) in a mirror circuit, we will implement $\bar{F} = \overline{(A\bar{B} + \bar{A}B)} = AB + \bar{A}\bar{B}$ (XNOR)

To review MIRROR CIRCUITS, you can read the Class Notes of MODULE 2 - PART B of ECE 2020.

The MIRROR CIRCUIT implementation and its corresponding resistance diagram with the capacitances is shown below :-



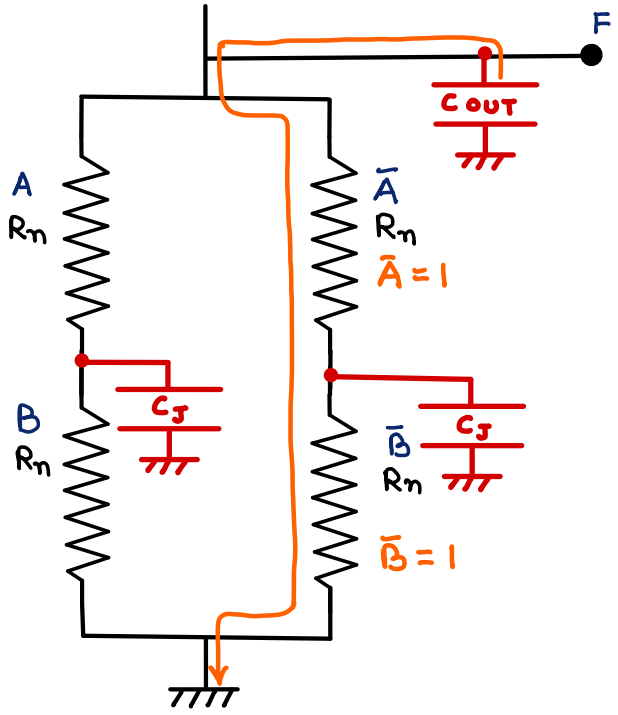


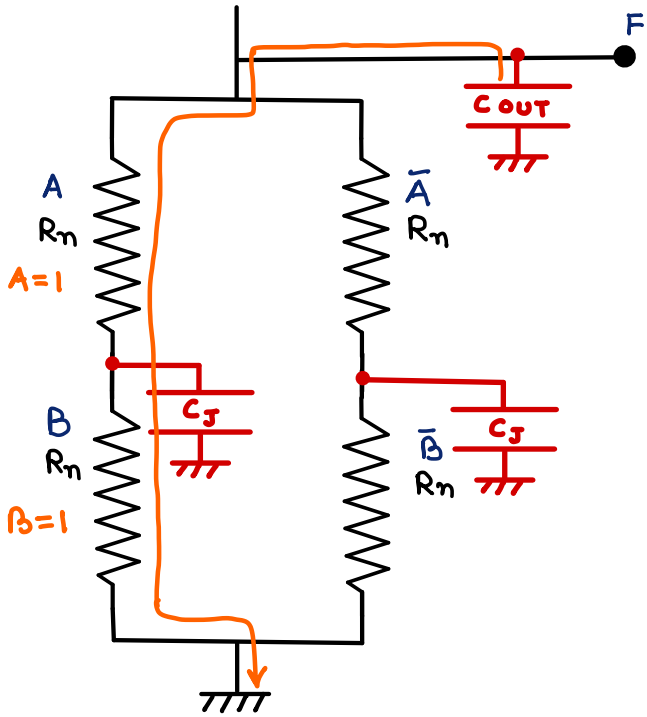
DISCHARGING DELAY :-

There are two different discharging paths for this circuit :-

When A = 0 ; B = 0 and when A = 1 ; B = 1

A	B	F	
0	0	0	→ DISCHARGING PATH 1
0	1	1	
1	0	1	
1	1	0	→ DISCHARGING PATH 2





The discharging delays for both of these paths are the same and is given by :-

$$T' = 0.69\left[R_n\,(C_J + C_{OUT}) + R_n C_{OUT}\right] \quad \ldots(ii)$$

Now if we compare T and T' from (i) and (ii),

$$T = 0.69\,(2R_n C_J + 2R_n C_{OUT})$$

or, $$T = 1.38\,R_n\,(C_J + C_{OUT})$$

and, $$T' = 0.69\,R_n\,(C_J + 2C_{OUT})$$

We can thus, see that $T' < T$.

$\therefore$ While the charging delays of the traditional CMOS implementation and the MIRROR CIRCUIT implementation of the XOR gates are the same, the discharging delay of the traditional CMOS implementation is higher because of the node J which is the junction of the 4 NMOS transistors.

$\therefore$ We prefer the MIRROR CIRCUIT implementation over the traditional CMOS implementation wherever applicable.