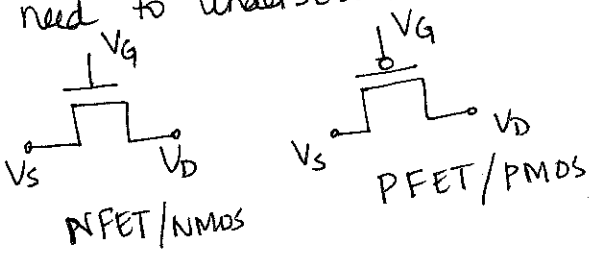
ECE - 6473 (Revision)

1) Technology scaling :- Every generation, shrinking transistor sizes leads to an improvement in their performance and reduces energy dissipation. At least, this was true for classical transistor scaling.

Scaling a technology reduces the gate delay by 30% and the lateral and vertical dimensions are also scaled down by 30%.
The die area decreases by 50%.
The areal & fringing capacitances all decrease by 30%.

2) To understand the design of integrated circuits, you need to understand transistors.

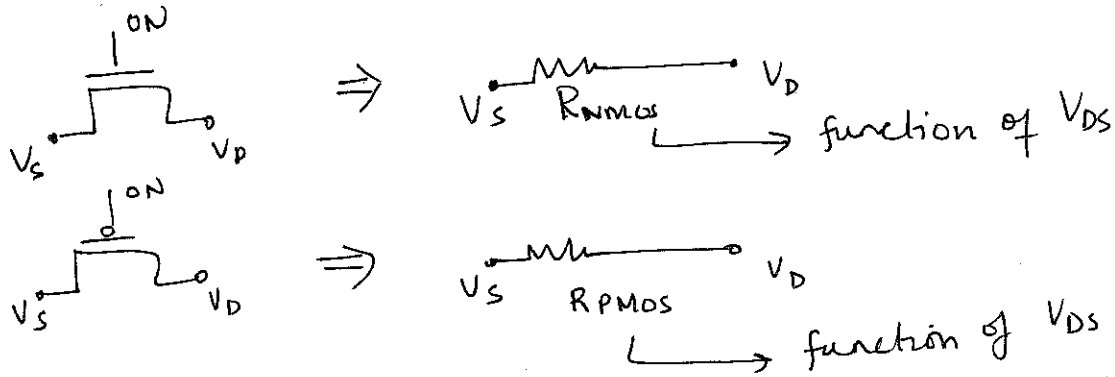


$\Rightarrow$ Essentially a switch that is controlled by the gate voltage V_G

NFET is good for transmitting "0" when it is ON.
PFET is good for transmitting "V_{DD}" when it is ON.

NFET is ON when $V_{GS} > V_{Tn}$
PFET is ON when $V_{SG} < |V_{TP}|$. The actual value of V_{TP} is negative.

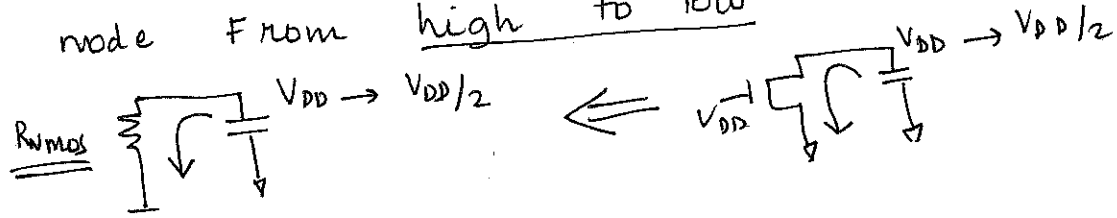
However when the transistor is ON it essentially is not a perfect <u>on</u> switch. That is, it has some finite resistance. Unfortunately, the resistance of the switch is dependent on the drain to source voltage. That is, the transistor is a <u>non-linear resistor</u>.



This makes life a bit harder to work with transistors. But we can often make simplifications to evaluate <u>R_{NMOS}</u> and <u>R_{PMOS}</u>. When we make some simplifications, we call these resistances as "<u>equivalent</u>" resistances. In that case, it really depends on the voltage swings across the transistor terminals.

We will look at a few specific cases to evaluate equivalent <u>R_{NMOS} and R_{PMOS}</u>

If an NMOS is being used to pull down a node from <u>high to low</u>

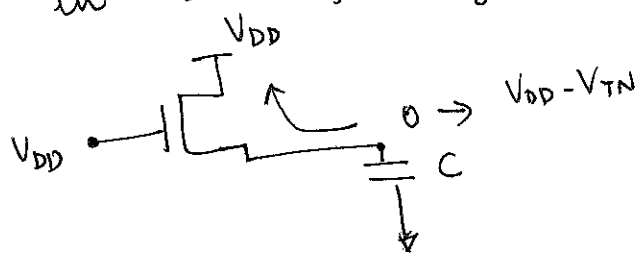


We evaluate R_{NMOS} at the beginning of the transition & then we evaluate R_{NMOS} half-way through the transition and take the average of the two resistances.

$$R_{eq} = \frac{1}{2}\left(R_{NMOS}(V_{DD}) + R_{NMOS}(V_{DD}/2)\right)$$

Also note in this case, NMOS is not under any body effect, since $V_{SB} = 0$.

However, if NMOS were used to pull-up a node in the following configuration.



The problem is slightly harder because now $V_{SB} \neq 0$. At the beginnig of the transition $V_{SB} = 0$ and $\therefore$ $V_{TN} = V_{T0}$.

At half way through the transition, $V_{SB} = \left(\frac{V_{DD} - V_{TN}}{2}\right)$, where V_{TN} must be

obtained self consistently by solving the equation

$$V_{TN} = V_{T0} + \gamma \left[\sqrt{2\phi_f + \frac{V_{DD} - V_{TN}}{2}} - \sqrt{2\phi_f} \right]$$

This is only a quadratic equation in V_{TN} that has two roots. Select the root that is positive and greater than V_{T0}.

<u>NOTE</u>:- whether the NMOS is pulling a node from V_{DD} to 0 or its pulling up a node from 0 to $V_{DD} - V_{TN}$.

$V_{DS} \geq [V_{Dsat} = V_{GS} - V_{TN}]$ So just use the equation for resistance in saturation.

In general (a) $R_{eq} \propto 1/(W/L)$

If the transistor size becomes bigger then R_{eq} becomes smaller.

(b) $V_{DD} \uparrow$ $R_{eq} \downarrow$ (c) $V_T \uparrow$ $R_{eq} \uparrow$ (d) $(\mu C_{ox}) \uparrow$ $R_{eq} \downarrow$

These scaling trends are important to consider.

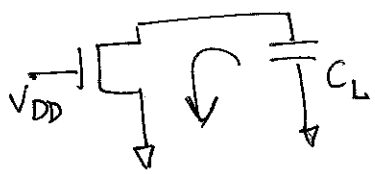
If all process parameters are same for PMOS and NMOS except the mobility such that $\mu_n > \mu_p$ then $R_{eqp} > R_{eqn}$ for $\left(\frac{W}{L}\right)_n = \left(\frac{W}{L}\right)_p$.

$\therefore$ If you want the same equivalent resistance

$$R_{eqn} = R_{eq,p} \Rightarrow \left(\frac{W}{L}\right)_p > \left(\frac{W}{L}\right)_n$$

$$\frac{\left(\frac{W}{L}\right)_p}{(W/L)_n} = \frac{\mu_n}{\mu_p}$$ if all other process parameters are identical. That is, $C_{ox,p} = C_{ox,n}$, $V_{TN} = |V_{TP}|$

Charging and discharging through transistors



$$\tau = 0.69\, R_{eq,NMOS}\, C_L$$

C_L → contains external cap + self loading cap.

Calculate $R_{eq,NMOS}$ as explained earlier.

$$\left\{ \begin{aligned} R_{eq,NMOS} &\cong \frac{3}{4}\frac{V_{DD}}{I_{Dsat}}\left(1 - \frac{5}{6}\lambda V_{DD}\right) \\ I_{Dsat} &= \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_n \left[V_{DD} - V_{TN}\right]^2 \end{aligned} \right\}$$

note, the above is true only when we are discharging a node from V_{DD} to 0 using an NMOS.

Go through HW 2 where we were charging a node using NMOS to calculate $R_{eq,NMOS}$ the proper way.

Concept of self-loading

Self loading capacitance is the capacitance at the output node due to the transistor structure itself.

Components include :- Overlap + Junction capacitance

Overlap → Source, Drain

Junction capacitance → Source, Drain

Typically all these capacitances are collectively termed as 'Parasitic capacitances'. All of these scale according to $C_{par} \propto W$

$$\begin{cases} C_{ovs} = C_{ovs} W \\ C_{ov,d} = C_{ovd} W \end{cases}$$

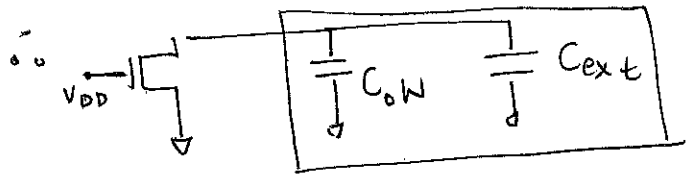
where $C_{ovs} = \frac{\varepsilon}{t_{ox}} X_{ov,s}$, $C_{ovd} = \frac{\varepsilon}{t_{ox}} X_{ov,d}$ } typically same

$$C_{jn} = C_j W L_s + C_{jsw} (W + 2L_s)$$

To remember:- parasitic capacitances arise due to the transistor structure & scale proportionally w/ device width.

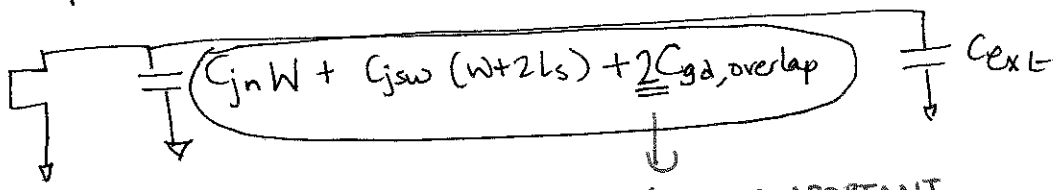
If we lump all of the parasitics into one component $C_{par} = C_0 W$

C_0 → per-unit-width capacitance.



total capacitance
at the output node
of the transistor

C_o comes from overlap and junction of the NMOS.
However, the overlap capacitance must be doubled because it is a miller capacitance. And remember, when capacitance connects input node with the output node then it is essentially a miller capacitance and its effect is doubled.



VERY IMPORTANT.

DO NOT FORGET !!

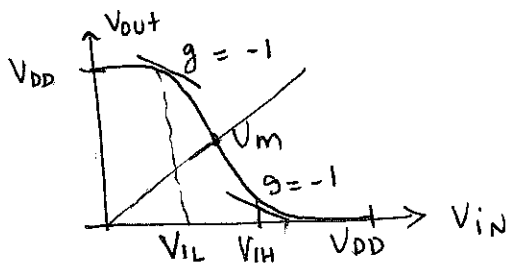
INVERTER ANALYSIS

a) Inverter consists of PMOS in pull-up and NMOS in pull-down network.

b) Figures of merit

(a) Switching threshold
(b) Low to high delay
(c) High to low delay
(d) Driving another inverter
(e) inverter gain → useful concept for noise margins.

Inverter can work even when supply voltage $< V_{TN}$ because it still conducts in sub-threshold. As long as $|gain| > 1$, inverter will continue to work.



$$\left.\begin{aligned} NM_L &= V_{IL} \\ NM_H &= V_{DD} - V_{IH} \end{aligned}\right\} \text{noise margins}$$

$$g = \text{gain} = \left(\frac{dV_{out}}{dV_{in}}\right)$$

→ We want $V_{IL} \uparrow$ and $V_{IH} \downarrow$

Max. value of $V_{IL} = V_m$ and min. value of $V_{IH} = V_M$.

$\underline{V_m = V_{IN} = V_{out}}$ ← switching threshold

V_m is calculated when both NMOS & PMOS are in saturation.

$$V_m = \frac{V_{DD} - |V_{TP}| + \sqrt{\frac{\beta_n}{\beta_p}}\, V_{TN}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}$$

A general rule is that when PMOS becomes stronger then V_m goes toward V_{DD}, while when NMOS becomes stronger then V_M goes toward GND.

What does a transistor being "strong" mean?

A transistor is said to be strong when its ON current is high. By ON-current, we mean the current in saturation.

A transistor can be made strong by increasing its size or by increasing $(V_{GS} - V_T)$ drop.

For $V_m = \frac{V_{DD}}{2} \Rightarrow \boxed{\sqrt{\frac{\beta_n}{\beta_P}} = \frac{0.5\,V_{DD} - |V_{TP}|}{0.5\,V_{DD} - V_{TN}}}$

General equation for $V_m = V_{DD}/2$

Switching Delay of the inverter

$$t_{pLH} = 0.69\, R_{eq,p}\, C_L$$

$$t_{pHL} = 0.69\, R_{eq,n}\, C_L$$

$$t_p = \tfrac{1}{2}\left(t_{pLH} + t_{pHL}\right)$$

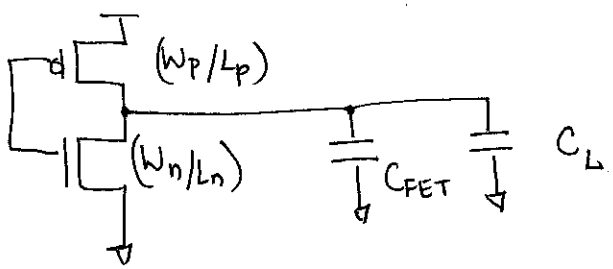
For making $t_{pLH} = t_{pHL}$

$$\boxed{\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p}}$$

$$R_{eq,p} = \frac{3}{4}\,\frac{V_{DD}}{I_{Dsatp}}$$

$$R_{eqn} = \frac{3}{4}\,\frac{V_{DD}}{I_{Dsatn}}$$

What is C_L? (IMPORTANT)



Assume $L_p = L_n = L_{min}$

C_{FET} = PARASITIC

C_L = Fan-out or a constant load + wire.

Lets compute C_{FET}

$$C_{FET} = C_{FET\_nmos} + C_{FET\_pmos}$$

$$C_{FET\,nmos} = C_{jn} W_n + 2 C_{ov} W_n$$

$$C_{FET\,pmos} = C_{jn} W_p + 2 C_{ov} W_p$$

$$\boxed{C_{FET} = C_{FET-nmos}\left(1 + \frac{W_p}{W_n}\right)}$$ V.V. Important.

$$t_{PLH} = 0.693 \frac{V_{DD}}{4 I_{Dsatp}} \left(C_{FET-nmos}\left(1 + \frac{W_p}{W_n}\right) + C_L \right)$$

Assume that there is no extra load at the output from C_L but that the inverter is self loaded.

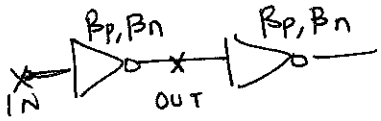
$$t_{PLH} = 0.693 \frac{V_{DD}}{4 I_{Dsatp}} C_{FET-nmos}\left(1 + \frac{W_p}{W_n}\right)$$

$$I_{Dsatp} = \frac{1}{2} \beta_p \left(V_{DD} - |V_{TP}|\right)^2$$

$$t_{PLH} = \frac{V_{DD}}{\beta_p \left(V_{DD} - |V_{TP}|\right)^2} C_{FET-nmos}\left(1 + \frac{W_p}{W_n}\right)$$

As $\left(\frac{W_p}{L_p}\right) \uparrow$ $\beta_p \uparrow$ $R_{eq,p} \downarrow$ but parasitic capacitance increases. Hence, self-loading ~~does~~ means that increasing the Device size does not affect the delay.

When a CMOS inverter is loaded by an identical inverter,



choose $\boxed{\frac{W_p}{W_n} = \alpha}$

How do you minimize the Delay From <u>IN to OUT</u>. ?

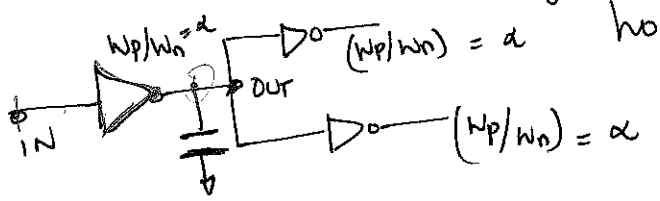
Calculate the net capacitance at the output.

$$C_L = \underbrace{C_{FET\_n}}_{\text{parasitic}} \underbrace{(1+\alpha)}_{(W_p/W_n)} + C_{wire} + (1+\alpha)\underbrace{C_{gn}}_{\text{input gate cap}}$$

$$t_d = 0.5 \frac{C_L V_{DD}}{(V_{DD}-V_T)^2}\left(\frac{1}{\beta_n} + \frac{1}{\beta_p}\right)$$

Substitute C_L in the above equation & minimize the delay $\frac{\partial t_d}{\partial \alpha} = 0 \Rightarrow \alpha_{opt} = \sqrt{\frac{\mu_n}{\mu_p}}$ For neg. C_w.

Now if the inverter is driving two identical inverters. how should you size the First inverter



$$C_L = C_{FETn}(1+\alpha) + C_{wire} + \underbrace{2}_{\text{extra due to two inverters as the Fan-out.}}(1+\alpha)C_{gn}$$

$$t_d = \frac{0.5\, C_L V_{DD}}{(V_{DD}-V_T)^2}\left(\frac{1}{\beta_n} + \frac{1}{\beta_p}\right) \Rightarrow \frac{\partial t_d}{\partial \alpha} = 0$$

So the general principle of obtaining delay is the same. All you need to know is two things: (a) Equivalent resistance (b) Load Capacitance.

For parasitic capacitance of a transistor, I always use the symbol C_{FET}.

For input capacitance of a transistor, I always use the symbol C_g.

$d = p + ghb$ $\quad g = 1$, $b = 1$

$d = p + h$ $\quad h = \frac{C_{out}}{C_n} = 2$

$d = 2$ for $p = 0$

$d \Rightarrow$ normalized delay.

$\tau_{ref} = 0.69 R_{ref} C_{ref}$

$C_{ref} = (1+\alpha) C_{gn}$

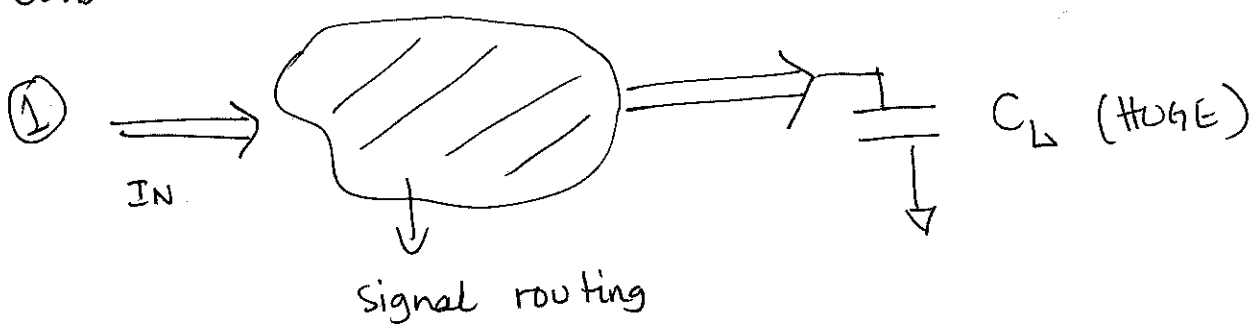
$D = 0.69 \underline{\underline{R_{ref}}} (1+\alpha) C_{gn} \times 2$

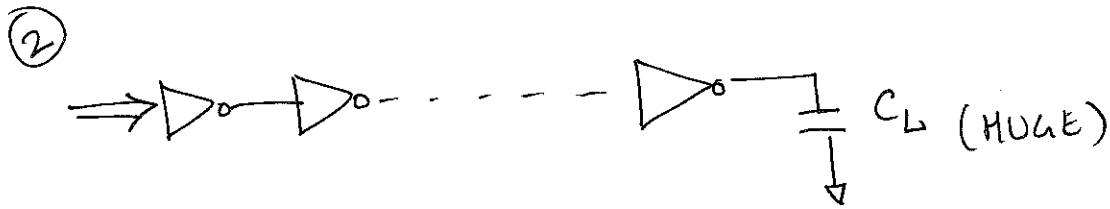
$R_{ref} = \frac{3}{4} \frac{V_{DD}}{I_{Dsat}}$ where $I_{Dsatp} = I_{Dsatn}$ so that pull-up & pull-down are the same.

The whole deal about inverter sizing :-

Often times we want to route a signal from one point on the chip to another and the load capacitance is <u>large</u>. It always makes sense to route the signal via a chain of inverters to minimize the propagation delay.

Consider the scenario :-



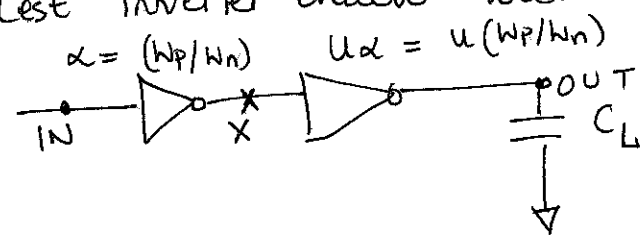


It was shown a long time ago that if you were to use 2 to drive the large load and sized the inverters appropriately, the delay of 2 < delay of 1.

So the question becomes a) how do you size this chain of inverters b) how large should this load capacitance be for us to know that scenario ② makes sense.

First, lets address point (b) above. That is, how large should C_L be for us to consider the inverter chain.

The smallest inverter chain will be



$R_{eq} \rightarrow 1^{st}$ inverter equivalent resistance.
Assume $R_{eqn} = R_{eqp}$ for a given α.

Lets say the first inverter is sized 'α' and second one is sized '$u\alpha$'.

$$t_{delay} = t_{INV1} + t_{INV2}$$

$$t_{INV1} = 0.69\, R_{eq} C_X = 0.69\, R_{eq} \left(C_{INV1}\, u \right)$$

$$t_{INV2} = 0.69 \frac{R_{eq}}{u} C_L$$

$$C_X = \text{input capacitance of inverter 2}$$
$$= C_{INV1} * u$$

Now we have the delay :-

$$t_d = 0.69\, R_{eq}\, C_{inv,1}\, u + 0.69 \frac{R_{eq}}{u} C_L$$

$$\frac{\partial t_d}{\partial u} = 0 \Rightarrow u_{opt} = \sqrt{\frac{C_L}{C_{INV,1}}}$$

$$t_{d,opt} = 0.69\, R_{eq} \sqrt{C_L C_{INV1}} + 0.69\, R_{eq} \sqrt{C_L C_{INV,1}}$$

$$= 2 \times 0.69\, R_{eq} \sqrt{C_L C_{INV1}}$$

Without the second inverter, the delay of driving the load will be :-

$$t_d = 0.69\, R_{eq}\, C_L$$

$\therefore$ Adding the second inverter makes sense only when $2 \times 0.69\, R_{eq} \sqrt{C_L C_{INV}} < 0.69\, R_{eq}\, C_L$

$$\left[\frac{C_L}{C_{INV}} > 4 \right]$$

Makes sense to add only when the load is atleast "4" times bigger than the input capacitance of the 1st inverter.

Now we know when we must add a chain of inverters to drive a large load.

<u>Key idea is :-</u>

Make the 1st inverter in the chain as the "unit" inverter or the reference inverter. Upsize each following inverter by a factor of $\underline{\underline{u}}$.

$$N_{opt} = \ln\left(\frac{C_L}{C_1}\right) \Rightarrow u^{N_{opt}} = (C_L/C_1)$$

$$u_{opt} = e$$

In a general scenario, first calculate N_{opt} and round it off to the nearest integer.

Then calculate $u_{opt} = (C_L/C_1)^{1/[N_{opt}]}$ ~~where [Nopt] is greatest int. less than or equal to.~~

→ This is equivalent to saying lets make the electrical effort of each stage the same.

Electrical effort is (output cap)/(input cap)

ENERGY DISSIPATION

$$E_{VDD} = \int_0^{\infty} i_{VDD}(t) V_{DD} dt$$

$i_{VDD}(t)$ → current drawn from supply

$$E_C = \int_0^{\infty} i_C(t) \, V_{out} \, dt$$

$$E_R = \int_0^{\infty} i_R^2(t) \, R \, dt$$

BASIC DEFINITIONS.

AVERAGE POWER DISSIPATION

$$P_{avg} = \alpha C_L V_{DD}^2 f_{clk} + \alpha E_{sc} + V_{DD} I_{static}$$

$\alpha C_L V_{DD}^2 f_{clk}$ → Dynamic power

αE_{sc} → Short circuit

$V_{DD} I_{static}$ → static power Leakage

α = Activity Factor

Fraction of the times, the output node transitions from "0" to "1"

STATIC CMOS GATES

Consist of PULL-UP network w/ PMOS

Pull-down network w/ NMOS.

a) While sizing the gates, Figure out the worst case path for low to high & high to low transition & make it equal to the reference inverter.

b) When considering dependence of input pattern on the delay, pay special attention to the internal node capacitance.

Large Fan-in and Large FO effect on delay can be represented as :-

$$t_p = a_1 \underbrace{FI}_{\text{Fan in}} + a_2 FI^2 + a_3 \underbrace{FO}_{\text{Fan-out}}$$

Remember techniques to size gates to minimize the delay.

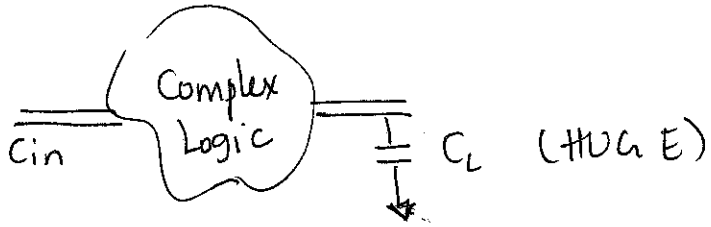
a) progressive sizing.

b) Input re-ordering.

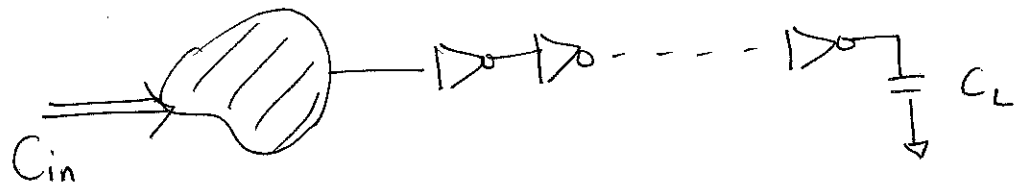
c) alternative logic structures

d) Isolating Fan-in From Fan out.

Lets look at the last point in more detail



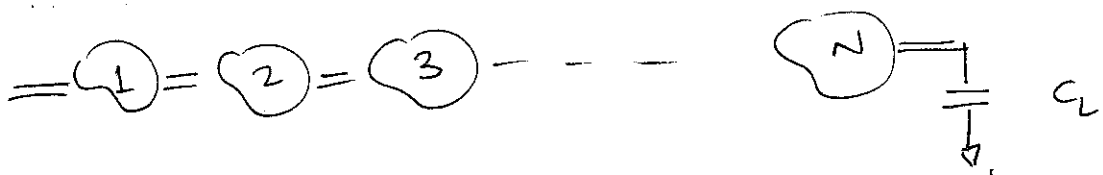
If we have a large capacitance to drive then we can do something like



Remember this technique of buffering makes sense only when $F = \frac{C_L}{C_{in}} > 4$.

Also correlate this with logical effort discussion.

→ Size each one of these $p = e$ For $\underline{p = 0}$ } parasitic
$p = 3.6$ For $\underline{p = 1}$



How should we size each gate in this logic path.

The idea is to make the stage delays equal

That is, $\underline{f_1 = f_2 = \cdots f_N} = F^{1/N} = (BGH)^{1/N}$

$$\left[\begin{array}{l} f_i = b_i g_i h_i \\ b_i = \text{branching at } i^{th} \text{ stage} \\ g_i = \text{logical effort of } i^{th} \text{ stage} \\ h_i = \text{electrical effort of } i^{th} \text{ stage} \end{array}\right]$$

$$B = \prod_i b_i$$

$$G = \prod_i g_i .$$

$$H = \prod_i h_i = (C_L/C_1)$$

$$D_{opt} = N F^{1/N} + P$$

total parasitic delay