Layout power optimization:

- Layout design is very important for performance and power optimization in deep submicron technology.

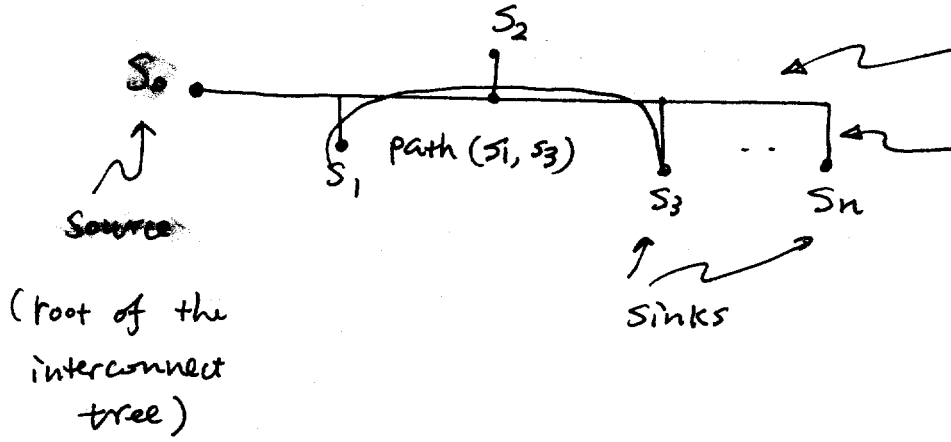
  ∵ it determines the lengths & widths of interconnect.

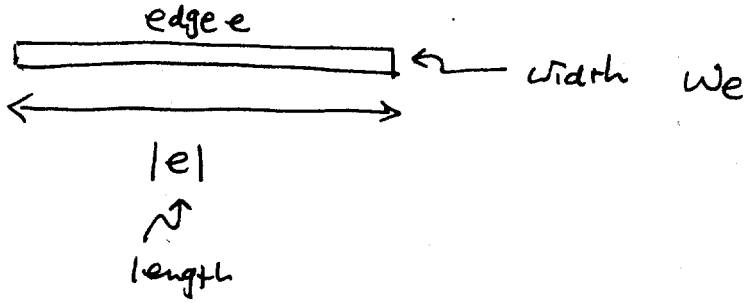
- In deep submicron design, 50% ~ 70% of clock cycle is consumed by interconnect delay.
- Significant amount of power is spent on charging & discharging interconnects.

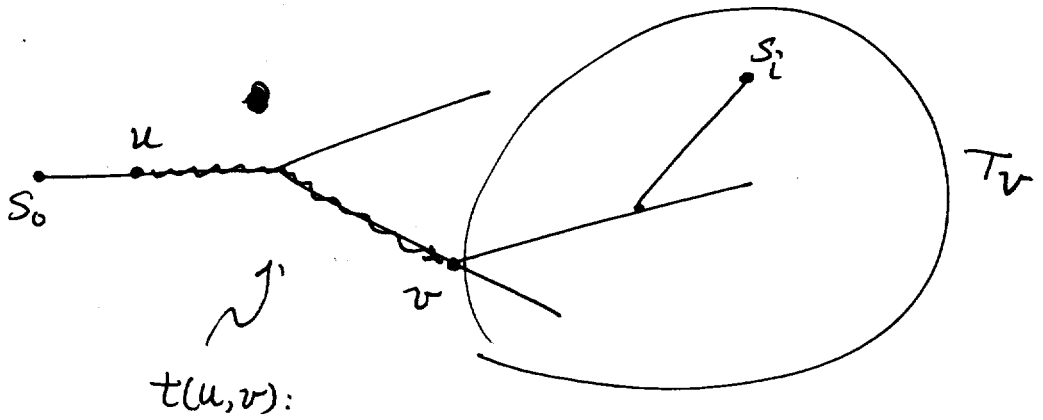
  ∵ Interconnect capacitance ↑

  short circuit power ↓
  static power ↓
  } secondary factor when compared to total capacitive power

- Will cover:
  - power & delay models ⟷ interconnects & gates
    (must study this ∵ delay must be maintained same level while optimizing for power)
  - device layout optimization
  - interconnect layout optimization
  - device & interconnect simultaneous optimization

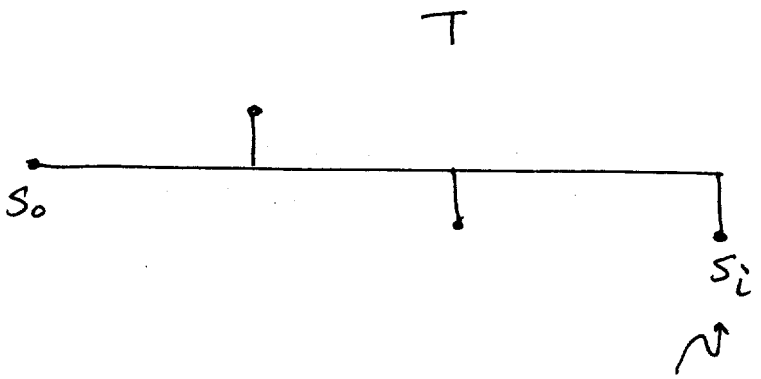






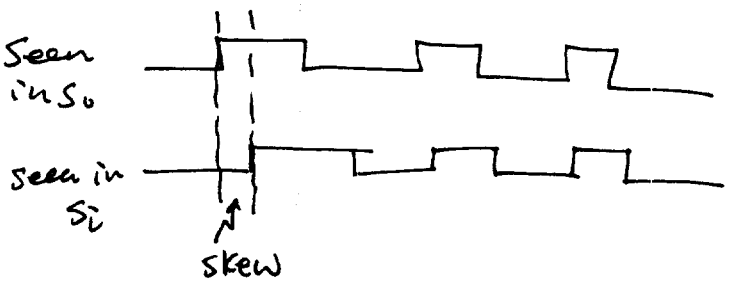
$t(s_0, s_i)$: delay from source to sink s_i

$t(s_0, s_i)$ is shorted by t_i.



Clock skew of s_i:

The difference in the clock signal delay from source to s_i.



$$\text{skew}(T) = \max_{s_i, s_j \in S} |t_i - t_j|$$

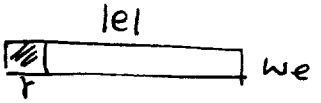
$S = \{$ the set of all sinks $\}$.

r: unit square wire resistance

C_a: unit area capacitance

C_f: unit length fringing capacitance (both sides)

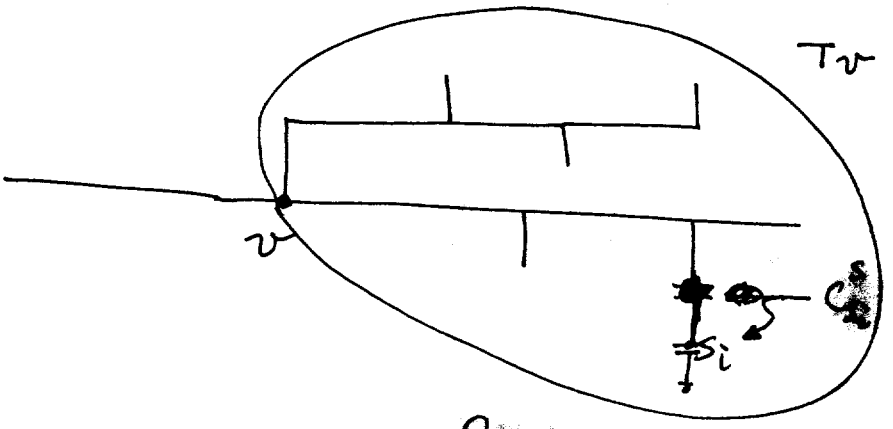
Wire resistance of edge e:



$$r_e = \frac{r \cdot |e|}{w_e}$$

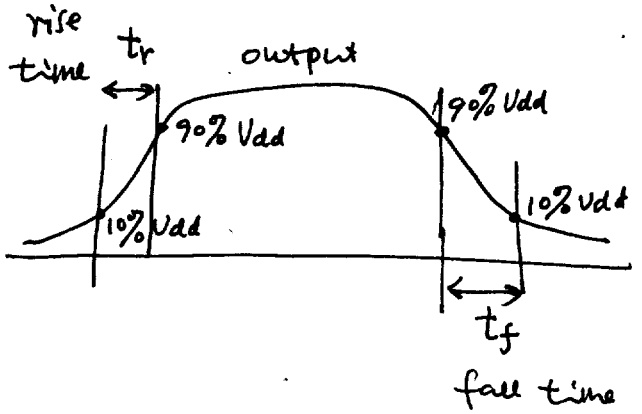
wire capacitance of e

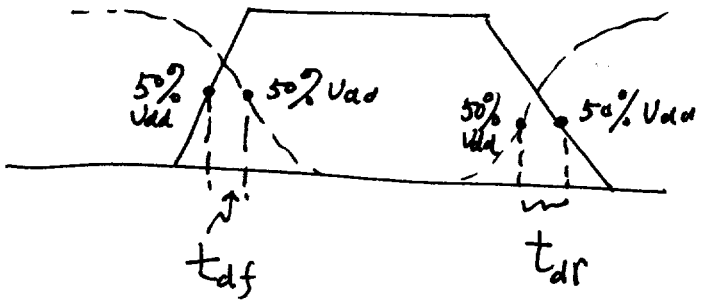
$$C_e = C_a |e| w_e + C_f |e|.$$



C_i : loading capacitance at sink s_i.

Cap(Tv) : total capacitance of Tv.





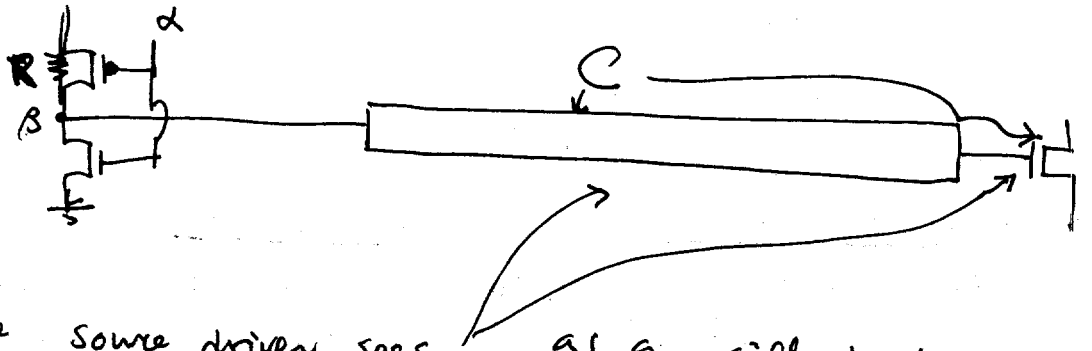
delay time for falling signal

delay time for ~~falling~~ rising signal

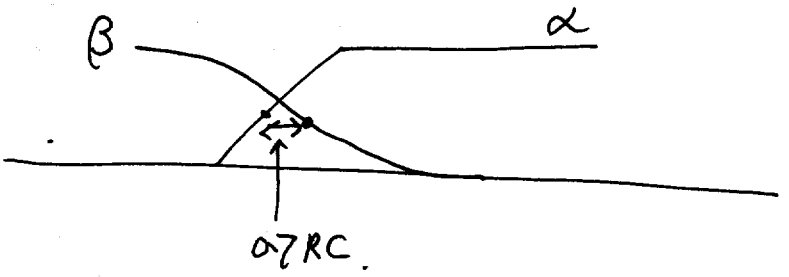
Interconnect Delay Model

- Lumped delay model ⟶ pathlength delay model → Elmore delay model.

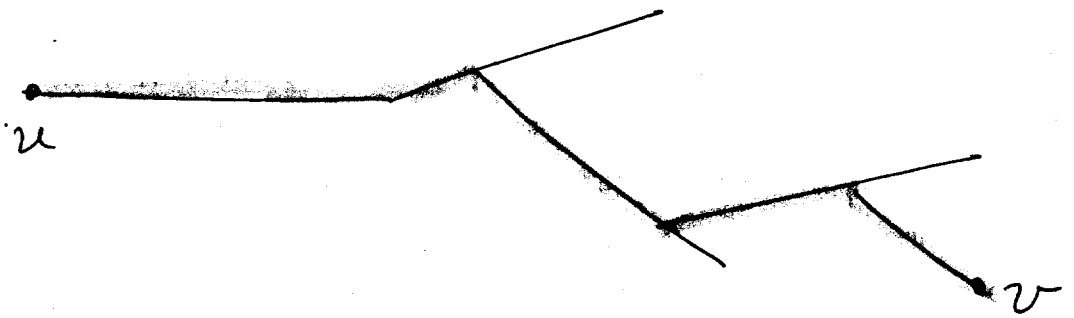
- Lumped capacitance model.



- The sourse driver sees as a sigle load capacitance
- Ignore the resistance at the interconnect (∵ driver r >> interconnect r).
- The delay for a (driver) gate to switch to 50% of its final value under step input ≈ 0.7RC.
- True for feature size ≥ 1.2 μm.



· Pathlength (linear) model:



$$t(u,v) \propto \sum_{e_w \in path(u,v)} |e_w|$$

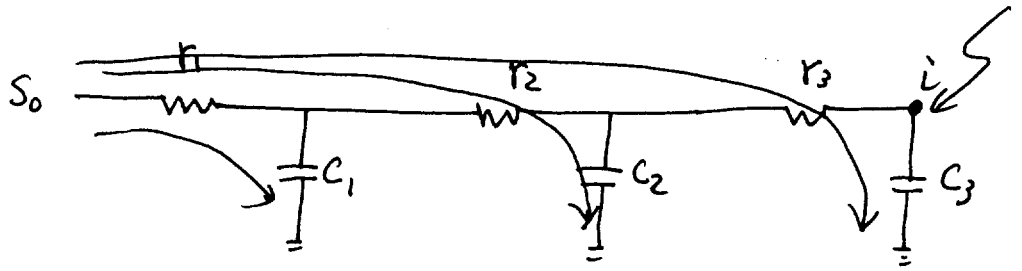
Limitations:

① It ignores the wire resistance, considers only wire capacitance along the path.

② It ignores the effect of edges not along the path.

· Elmore delay model.

· The most commonly used delay model.



$$t(s_0, i) = r_1 C_1 + (r_1 + r_2) C_2 + (r_1 + r_2 + r_3) C_3$$

either charging or discharging

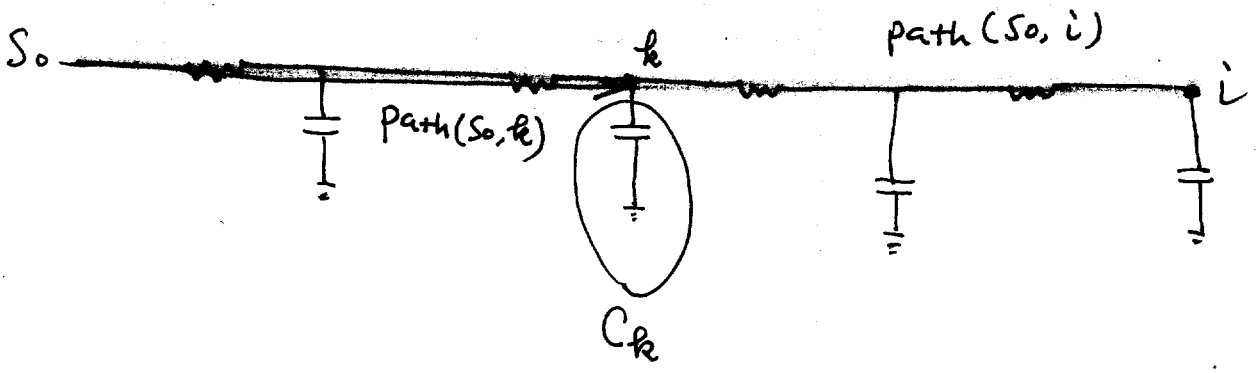
$$= r_1 (C_1 + C_2 + C_3) + r_2 (C_2 + C_3) + r_3 C_3$$

· General form

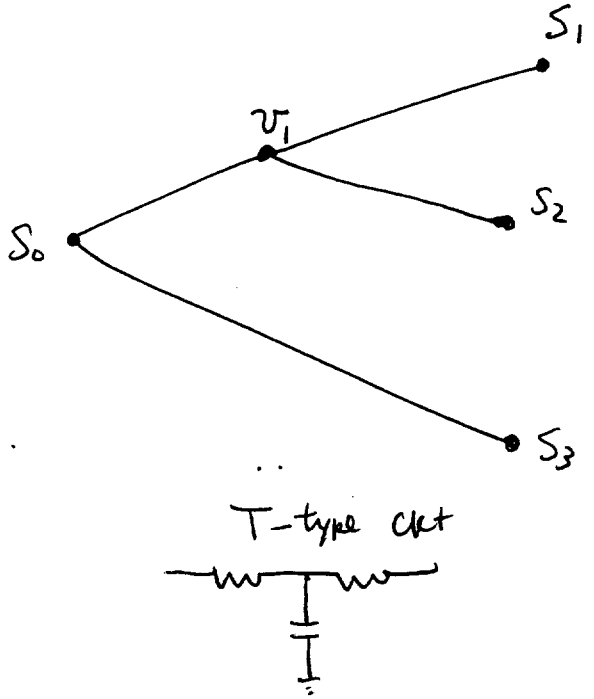
$$t(s_0, i) = \sum_{\text{all nodes } K} R_{Ki} \cdot C_K$$

C_K: lumped capacitance at node K.

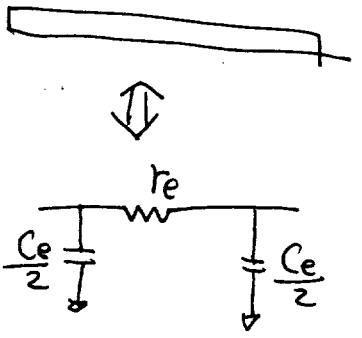
R_{Ki}: the resistance of the portion of path (s_0, i) common with path (s_0, k)

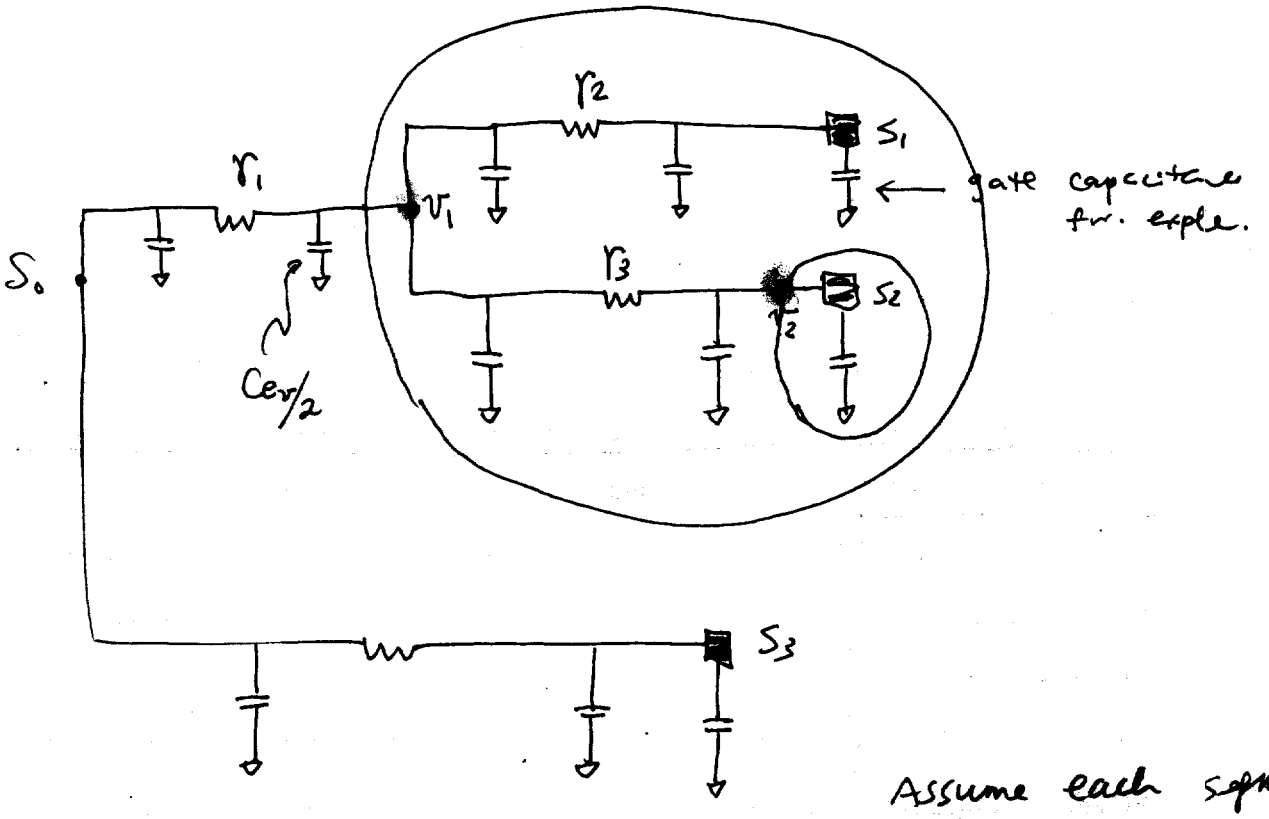


Example:



Each edge can be modeled by π-type circuit





Assume each segment has cap = c_{ev}.

$$t(s_0, s_i) = \sum_{e_v \in path(s_0, s_i)} r_{e_v} \cdot \left(c_{ev}/2 + cap(v) \right)$$

$$t(s_0, s_2) = r_1 \left(c_{ev}/2 + cap(v_1) \right) +$$

$$r_3 \left(c_{ev}/2 + cap(v_2) \right)$$

- Good things for Elmore model:
  - * An optimal or near optimal solution using this model is also nearly optimal using spice for routing, wiresizing, clock skew.

Disadvantages of Elmore model:

① Absolute value of Elmore delay may not be accurate.

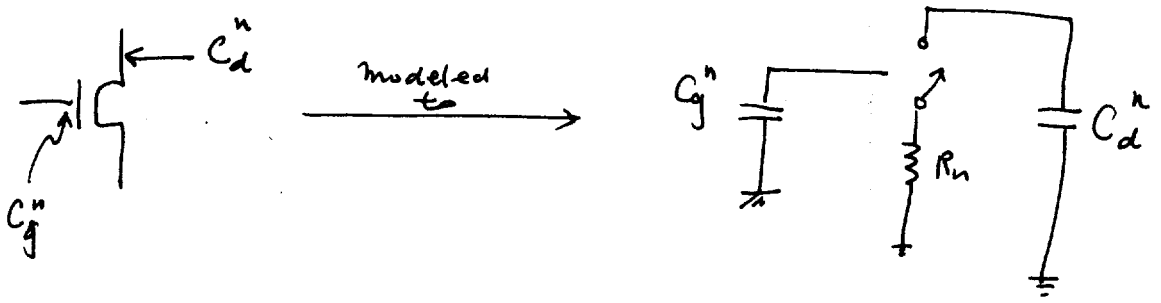
② It can't handle inductive effects

∴ More accurate delay can be modeled by RLC tree for interconnects.

$$\text{Delay} \cong k\, l \sqrt{L(w)\, C}$$

(k: constant, l: length)

Driver delay model



R_n: resistance of minimum transistor n.

C_g^n: gate cap. of min n transistor

C_d^n: output diffussion cap. of min n transistor.

- For an arbitrary transistor with size $d \geq 1$ (Width)

$$R_d = R_n / d$$

gate cap $= C_g^n \cdot d$

diff cap $= C_d^n \cdot d$ ← part of C_L

fixed channel length

$$t_f = K \cdot \frac{C_L}{\beta_{min}^n \cdot d \cdot V_{DD}}$$

for step input.

$$\beta_{min}^n = \frac{W}{L} \cdot \frac{\varepsilon_{ox} \cdot \mu}{t_{ox}}$$

C_L includes diff cap & load capacitance.

$$\beta_{min}^n = \frac{W}{L} \cdot \frac{\varepsilon_{ox} \cdot \mu}{t_{ox}}$$

$t_{df} = t_f/2$ ← delay time for the falling signal.

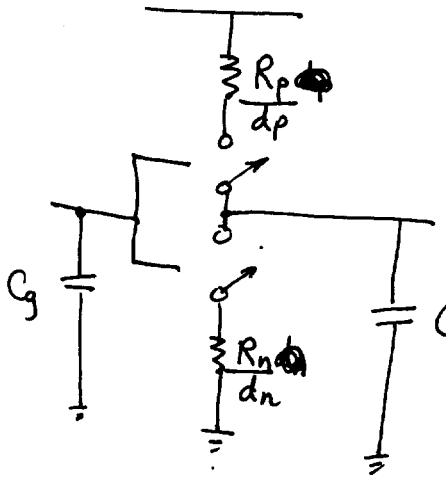
$\therefore R_d \propto \frac{1}{\beta_{min} \cdot d}$

$t_{df} = t_{f/2} = \frac{k}{2 V_{DD}} \left(\frac{1}{\beta_{min}^n \cdot d}\right) C_L$ (constant)

$\therefore t_{df} \propto R_d \cdot C_L$.

t_{df} can be approximated by $R_d \cdot C_L$

• Inverter

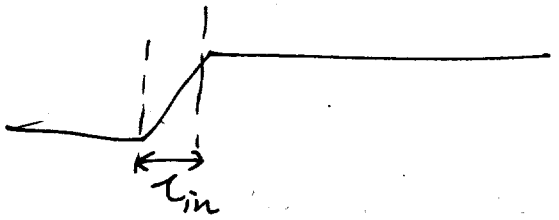


C_d ← sum of diff cap. due to the p-, n- transistors

$$t_f = K \cdot \frac{C_L}{\beta^n_{min} \cdot d_n \cdot V_{DD}}$$

$$t_r = K' \cdot \frac{C_L}{\beta^p_{min} \cdot d_p \cdot V_{DD}}$$

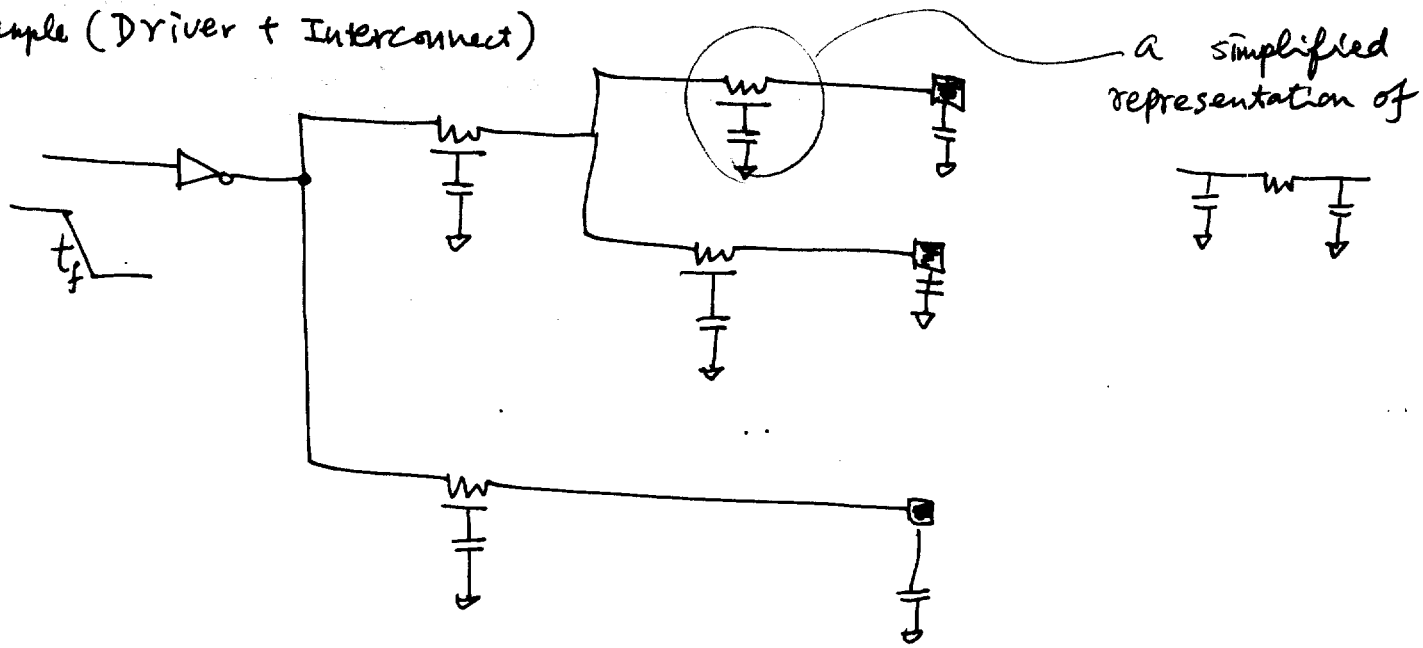
- Shortcoming of the above simple RC model:
  Can't deal with the shape of input waveform.



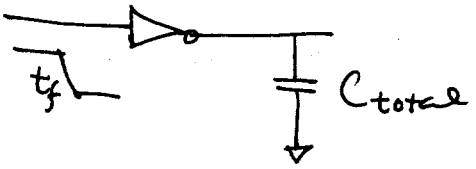
$$t_{df} = t_f/2 + \frac{\tau_{in}}{6} \cdot \left(1 + 2\frac{V_{Tn}}{V_{DD}}\right)$$

V_{Tn}: threshold voltage of n-transistor.

- Example (Driver + Interconnect)



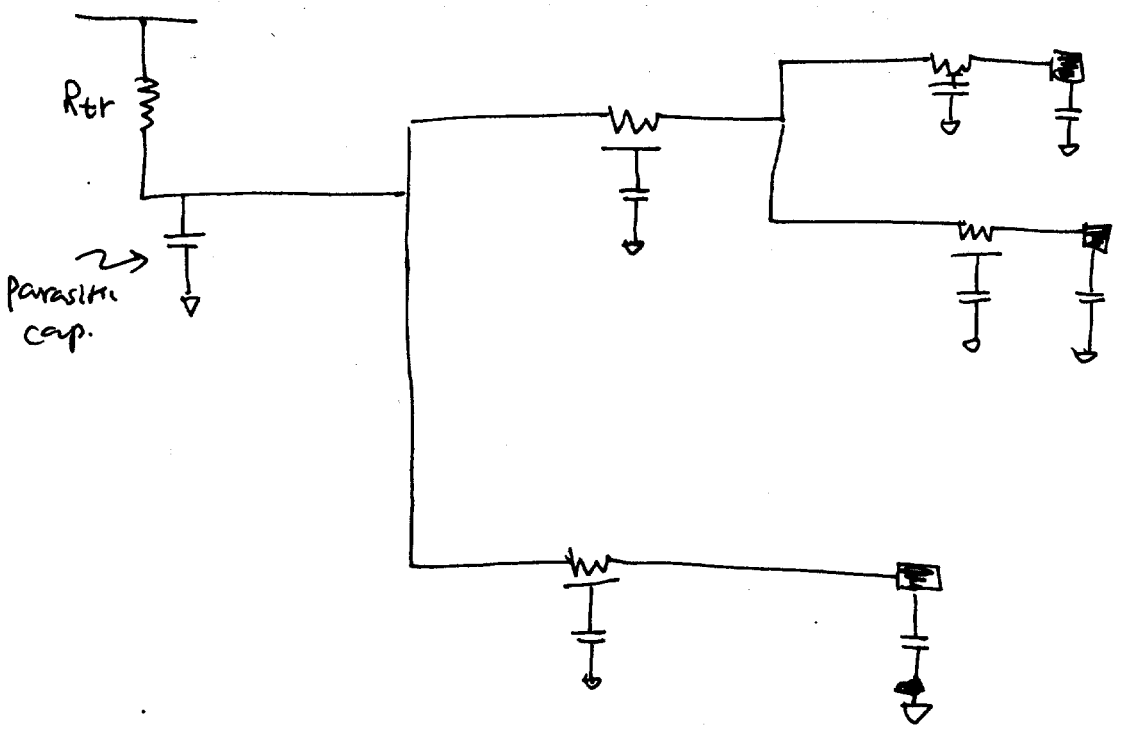
Can be simplified to



$$\therefore t_r = k' \frac{C_{total}}{\beta^p_{min} \cdot d_p \cdot V_{DD}}$$

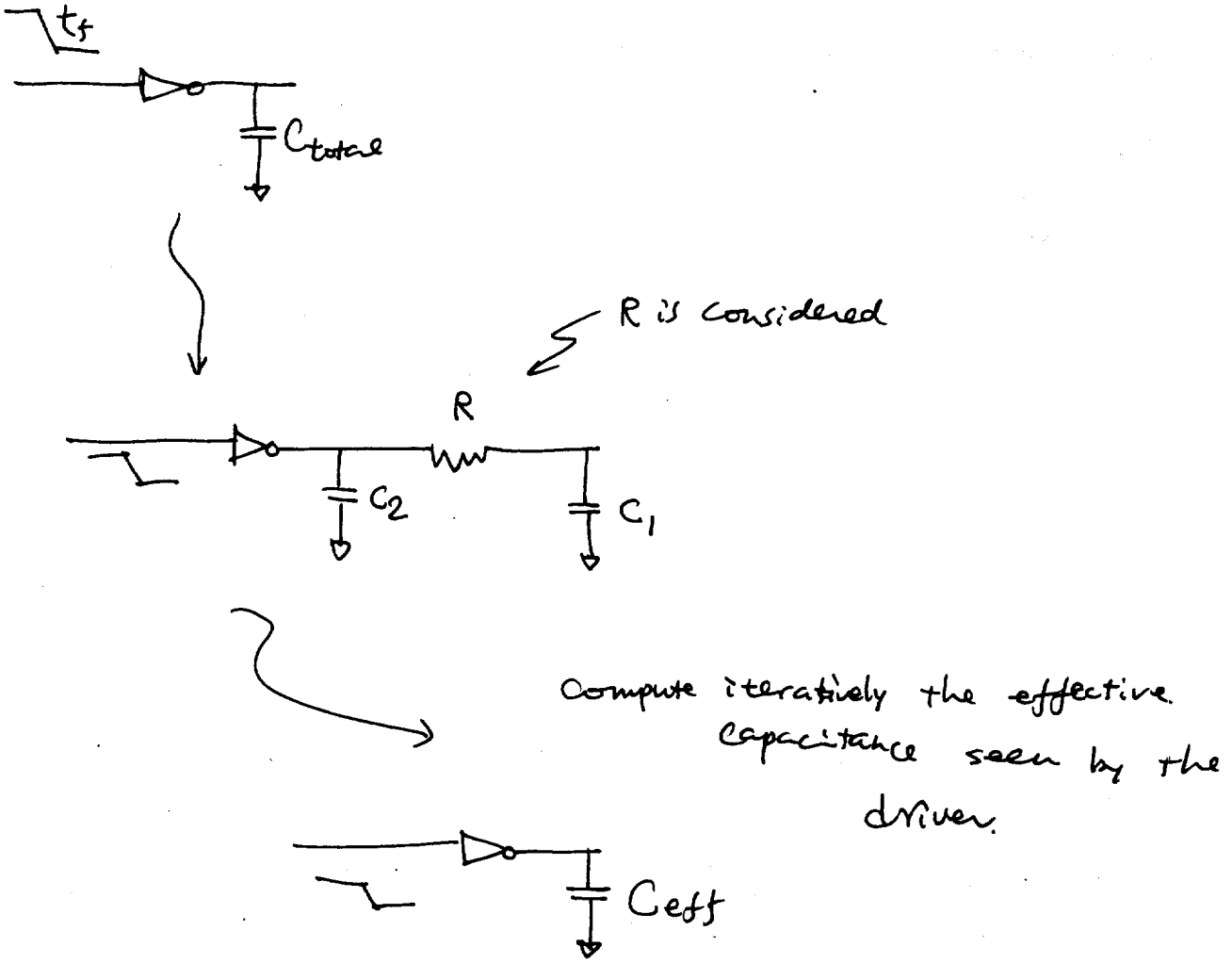
← including $C_{parasitic}$ ($C_{diff(P)} + C_{diff(N)}$ + C_w + $C_{driven\text{-}gate}$

- A better estimation



Then, use Elmore model to estimate the delay.

Another Estimation



Power Model

Short Circuit

input transition time

$$\bar{I}_{SC} = \frac{\beta \cdot \tau_{in}}{12 \cdot V_{DD}} \cdot (V_{DD} - 2V_T)^3 \cdot f$$

$$P_{SC} = \bar{I}_{SC} \cdot V_{DD} = \frac{\beta \cdot \tau_{in}}{12} (V_{DD} - 2V_T)^3 \cdot f$$

$$\beta = \frac{W_{eff}}{L_{eff}} \cdot \frac{\varepsilon_{ox} \cdot M}{t_{ox}}$$

Assume ~~$\beta_n = \beta_p = \beta$~~

The gain factor

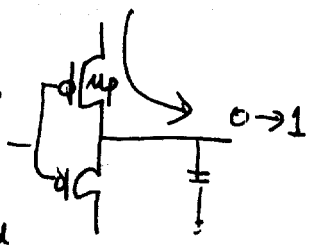
$$\beta = \frac{W}{L} \cdot \frac{\varepsilon_{ox}\, \mu}{t_{ox}}$$

is determined by Width of transistor W &

mobility of carrier responsible for the transition

(μ_p for low-to-high transition)

(μ_n for high-to-low transition)



$\therefore$ P_{sc} for V_{dd} → output

$\therefore$ P_{sc} can be reduce to dominated by pull-up transistor.
∵ pull-down was on

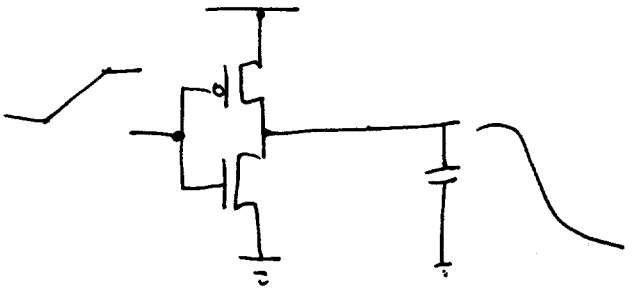
$$P_{sc} = K \cdot \underbrace{\mu W}_{\text{comes from } \beta.} \tau_{in} f$$

P_{sc} for output → V_{gnd}
dominated by pull-down transistor
∵ pull-up was on.

process & voltage dependent.

$$P_{sc} = \frac{W_{eff}}{L_{eff}} \frac{\varepsilon_{ox}\, \mu}{t_{ox}} \frac{\tau_{in}}{12} (V_{DD} - 2V_T)^3 \cdot f$$

$$= \underbrace{\left[\frac{\varepsilon_{ox} \cdot (V_{DD} - 2V_T)^3}{L_{eff} \cdot t_{ox} \cdot 12}\right]}_{K} \mu \cdot W_{eff} \cdot \tau_{in} \cdot f$$



When input transition time >> output transition time

⇒ short ckt path achieved for a longer time

∴ incur more power dissipation

∴ For power minimization, try to achieve equal input and output transition times. &

Keep both as small as possible.

Dynamic power

$$P_{cap} = \frac{1}{2} E(sw)\, C_L\, V_{DD}^2\, f$$

C_L includes device capacitance & interconnect capacitance

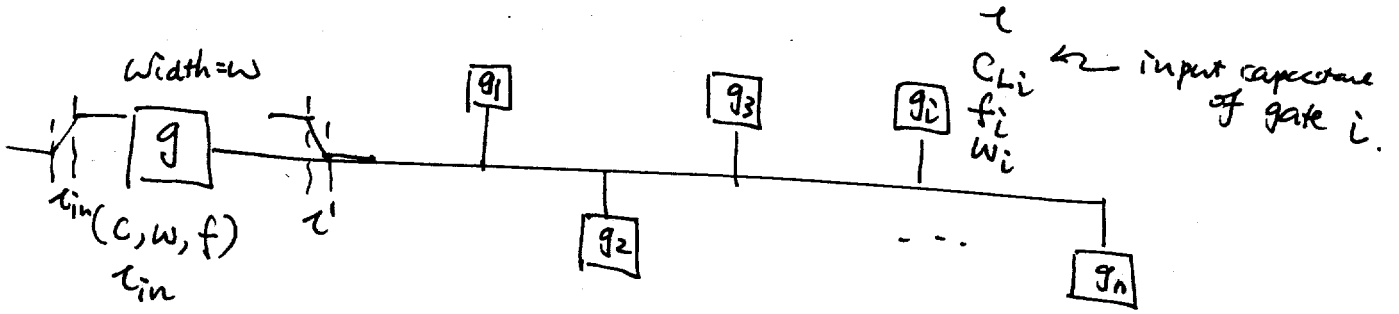
interconnect capacitance critical for deep submicron technology.

Transistor and Gate sizing

- Transistor sizing: Determine the optimal width for each transistor

  ⇒ minimize delay or power under the delay constraint.

- Gate sizing: All transistors in a gate are scaled by a common factor to perform power/delay optimization.



C: input capacitance for a minimum-width transistor

W: size of switching transistor (width)

f: switching frequency of g.

t_{in}: input transition time of g.

τ: output transition time of g, i.e. input transition time of $g_1, g_2, \cdots g_n$.

pull-up w pull-down depends on output ⌐ w ⌐

- If output ⌐ w is for pull-down
- If output ⌐ w is for pull-up.

~~ASST~~ Project III assigned

Due 3/17/03

Interconnect Layout optimizato.

- ckt partitig
- placement
- Routing
- wiresizg

definitns

properties: Monotone.
Separability
Dominance

Sizg method

* Gate or Transistn sizg
* Transistn reordrg

Capacitive power fw g (gate)

· capacitive power fw $g_1, g_2, \cdots g_n$

$$P = V_{dd}^2 \left(C \cdot w \cdot f + \sum_{i=1}^{n} C_{Li} \cdot f_i \right) + \underbrace{\quad}_{= P_1}$$

power of gate cap of driver

contains each C_{Li} C_{di} C_{gi} ↑ can be ignored.

Let $P_1 = V_{dd}^2 \sum_{i=1}^{n} C_{Li} \cdot f_i$

$$k \left(\mu \cdot \tau_{in} \cdot w \cdot f + \sum_{i=1}^{n} \mu' \cdot \tau \cdot w_i \cdot f_i \right)$$

mobility of carrier fw g

mobility of carrier fw g_i

Comes from β of P_{sc}

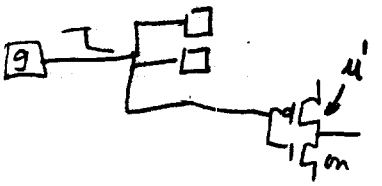
Constant based on technology & voltage (comes from β & V_{DD}) of P_{sc}.

For g with

w: pull-down width

μ: pull-down mobility

μ': pull-up mobility



· The ~~delay~~ output transition time of g:

load cap. of g.

$$\tau = \phi \frac{C_L}{\mu w_o} + \frac{\tau_{in}}{6} \left(1 + \frac{2V_T}{V} \right)$$

process dependent constant

$\Rightarrow$ g short ckt $\quad$ g switch $\quad k_1 = V_{dd}^2 \cdot C$.

$$P = P_1 + (k \mu \tau_{in} + k_1) w f + k \mu' \sum_{i=1}^{n} w_i f_i \left(\phi \frac{C_L}{\mu} + k_2 \tau_{in} \right)$$

P_1: $g_1 \cdots g_n$ switching power

$g_1 \cdots g_n$ short-ckt power

$\Rightarrow P = \alpha * w + \frac{\delta}{w} + \sigma$.

← Increasing w → increase power of g

reduce power of $g_1 \cdots g_n$

∴ trade-off

$\frac{\partial P}{\partial W} = 0$ gives the condition where P is minimum on W.

~~for sp~~

$$W = \frac{\sqrt{\phi \frac{u'}{u} C_L \left(\sum_{i=1}^{n} W_i f_i\right)}}{\sqrt{\left(\frac{k_1}{k} + u\tau_{in}\right) f}}$$

the power optimal p-transistor size:

$u \leftarrow u_p$

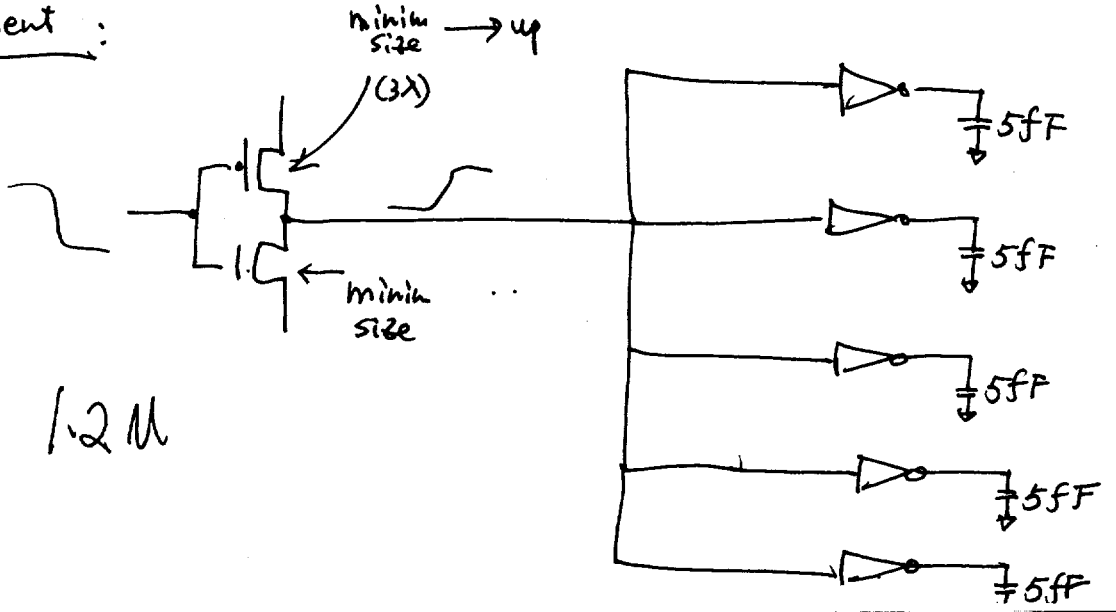
$u' \leftarrow u_n$

the power optimal n-transistor size:

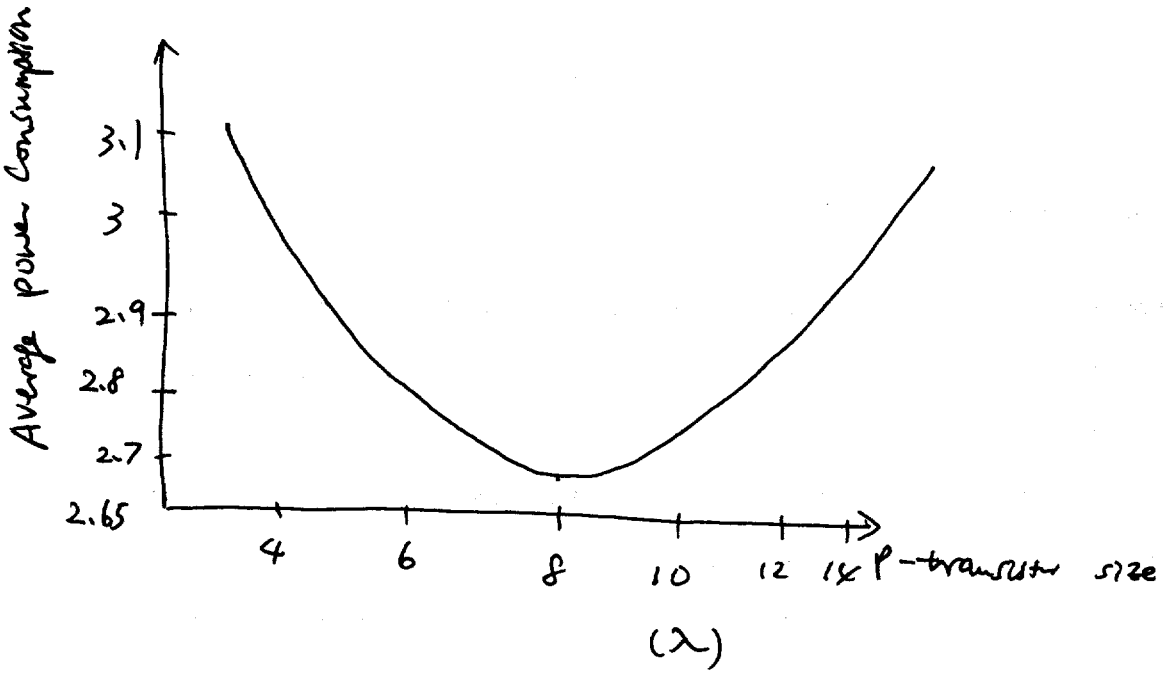
$u \leftarrow u_n$

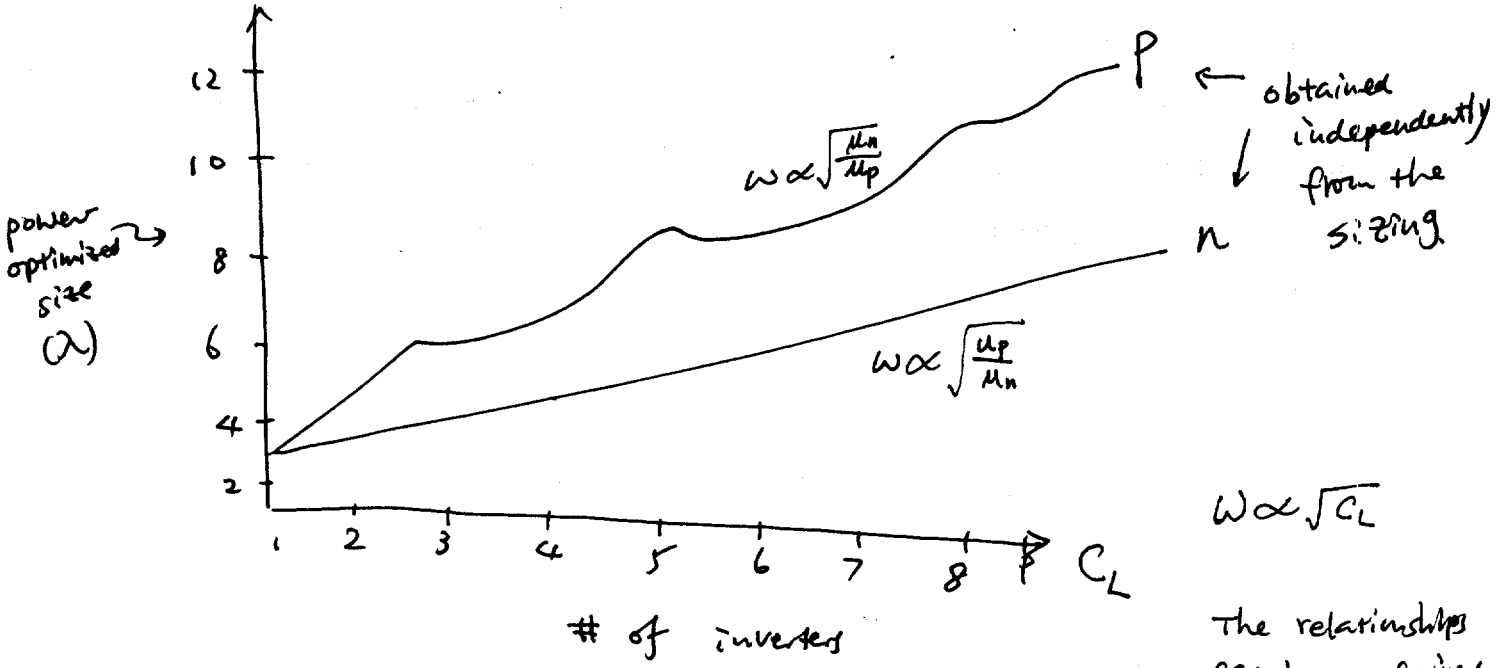
$u' \leftarrow u_p$

Expriement:



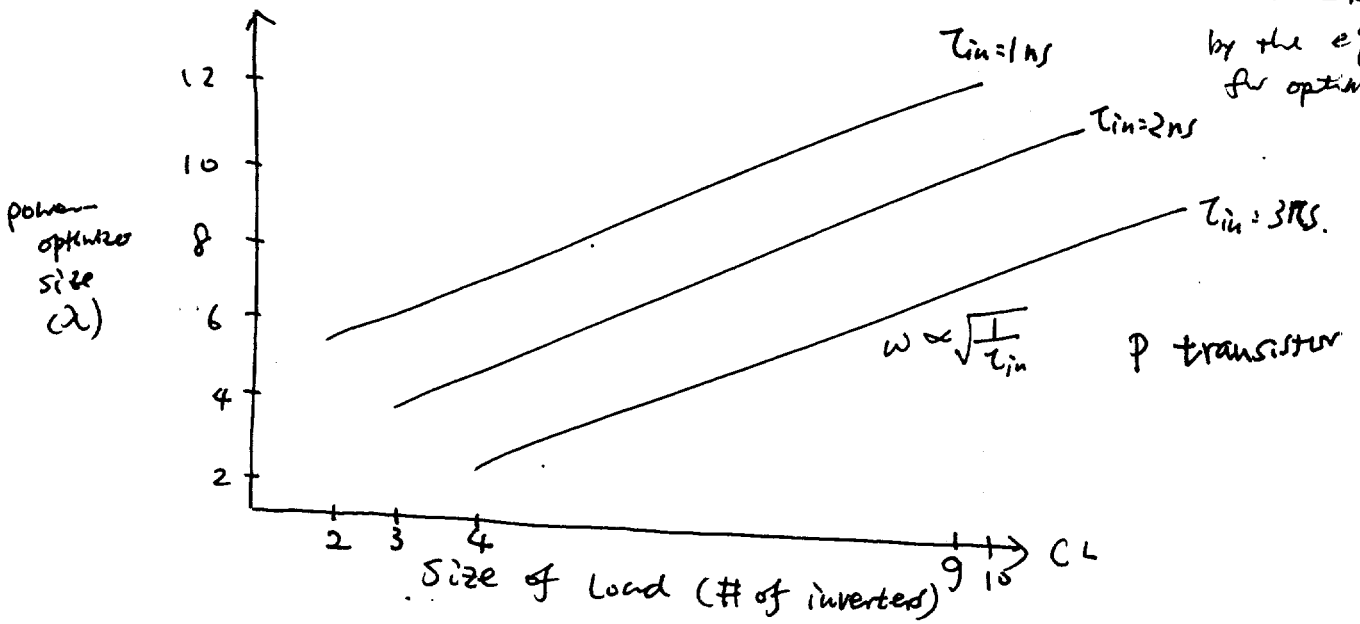
1.2 M

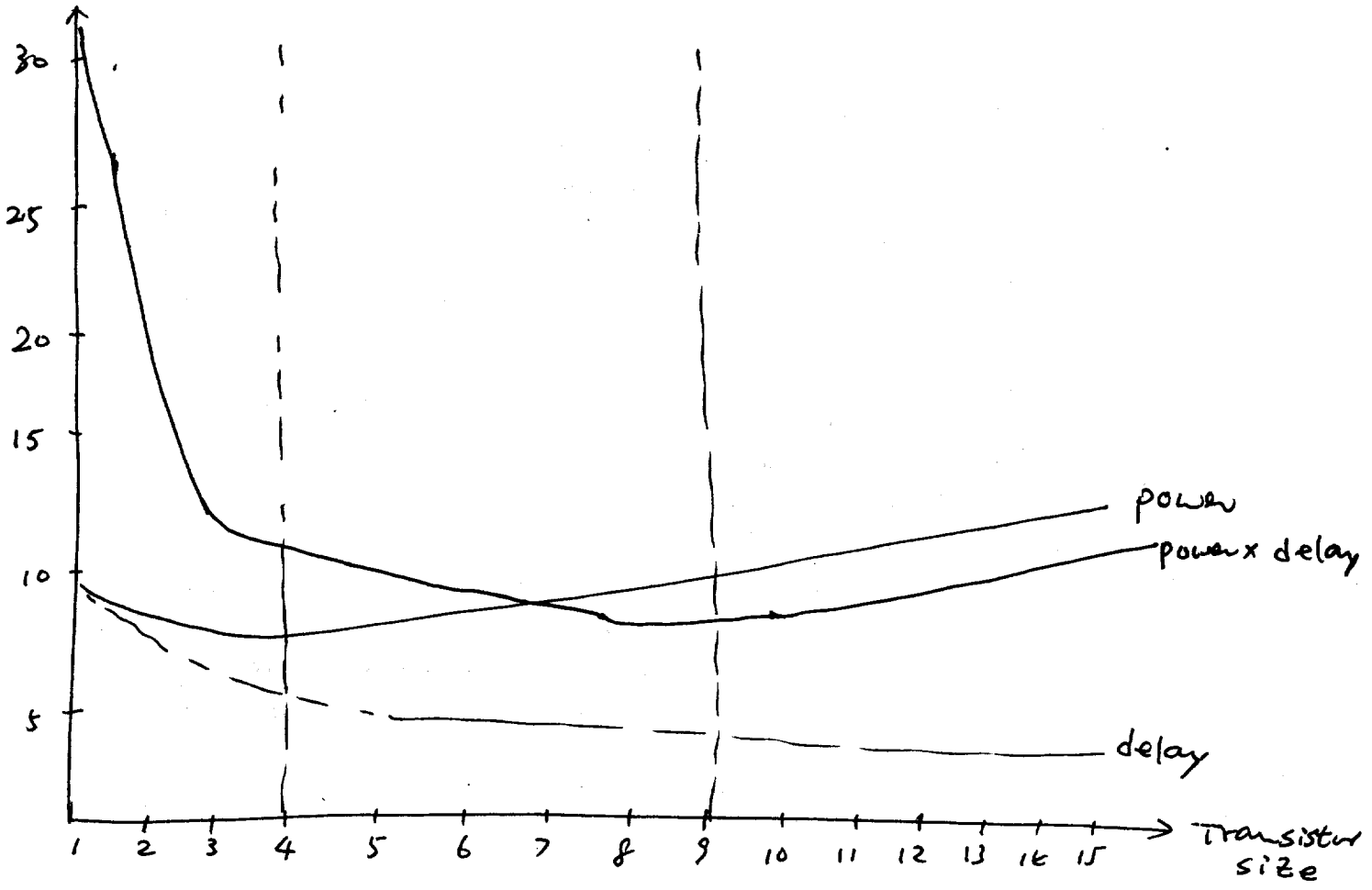




$W \propto \sqrt{C_L}$

The relationships can be explained by the equation for optimal W.





① Decrease in power & delay

② Rapid decrease in power-delay Curve

① Decrease in delay

② increase in power

③ Slow decrease in power-delay Curve

① increase in power-delay Curve

② Δ (increase in power) rate $>$ Δ (decrease in delay)

P transistor of driver

Transistor sizing method for power optimization

① All transistors are assigned power-optimal sizes. (region 1)

② Timing analysis is performed.

③ While (∃ path with delay > constraint)
power-delay optimal sizes are used for all transistors in the critical path.

④ If (∃ path with delay > constraint)
the transistor with least slop on the power-delay curve is sized up

⑤ If (∃ path with delay << constraint)
the transistor with steepest power-delay curve is sized down.

⑥ Repeat ④ and ⑤ until delay constraint is satisfied.

Circuit	minim area power	minim-power power	power saving
C432	306.4 MW	298.3 MW	2.64%
C499	227.12 MW	220.10 MW	3.09%
			0.74%
			1.57%
			6.4%
			4.35%
			5.05%
			9.51%
			4.71%
			10.91%

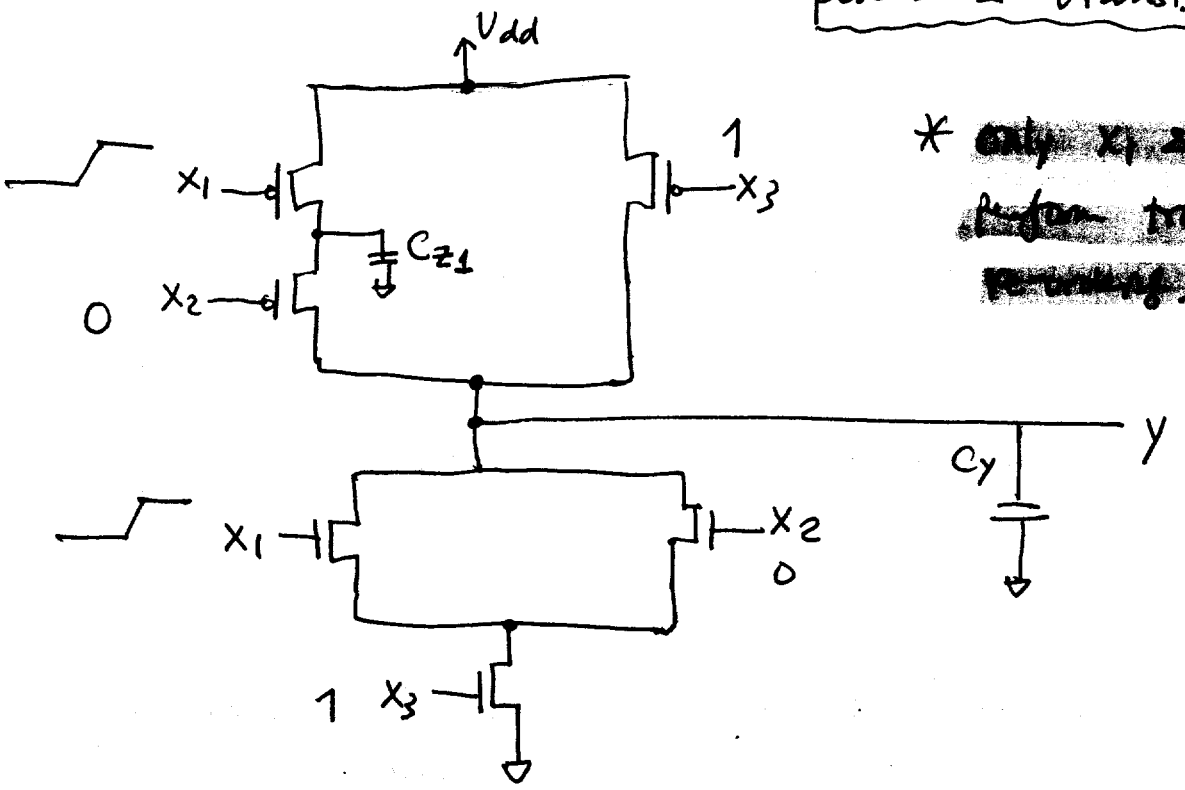
→ It is hard to optimize power dissipation in layout level.

Transistor Re-ordering

- Find the best ordering of series connected transistors in each gate to minimize the delay or power
- No area penalty when compared with device sizing techniques.

/F3_84
STACK:
OFFENDING COMMAND: F3
ERROR: undefined

power & transistor order

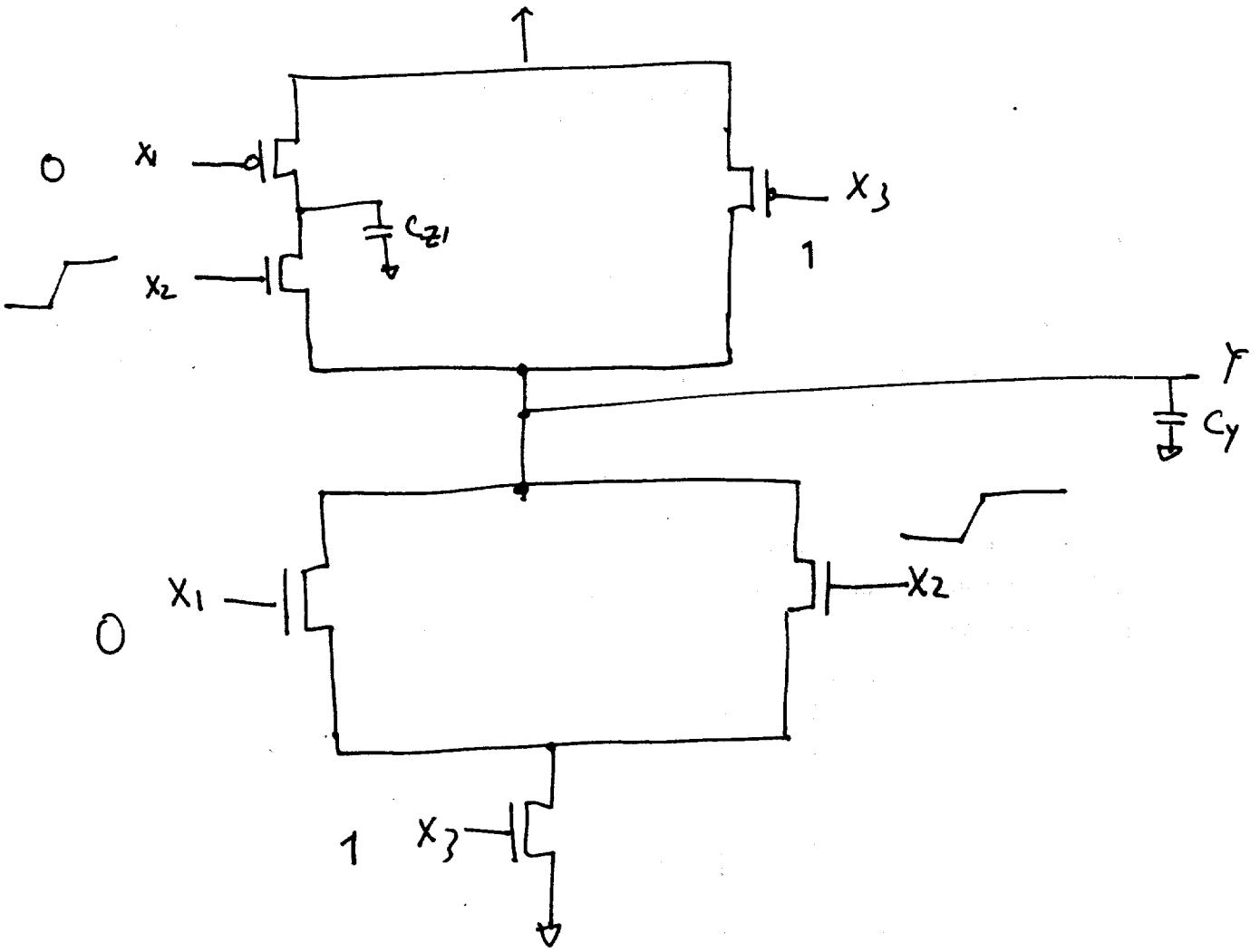


* [redacted]

- When x_1 ↑, C_y and C_{z1} are both discharged.
- With $x_1 = 0$ and x_2 ↑, only C_y is discharged.

∴ If x_2 has a higher switching prob. than x_1, the transistor driven by x_2 in the series-connected P-transistor chain should be placed closer to the output, to reduce the power dissipation.

∴ Based on switching activity, we can find the best ordering for power dissipation.

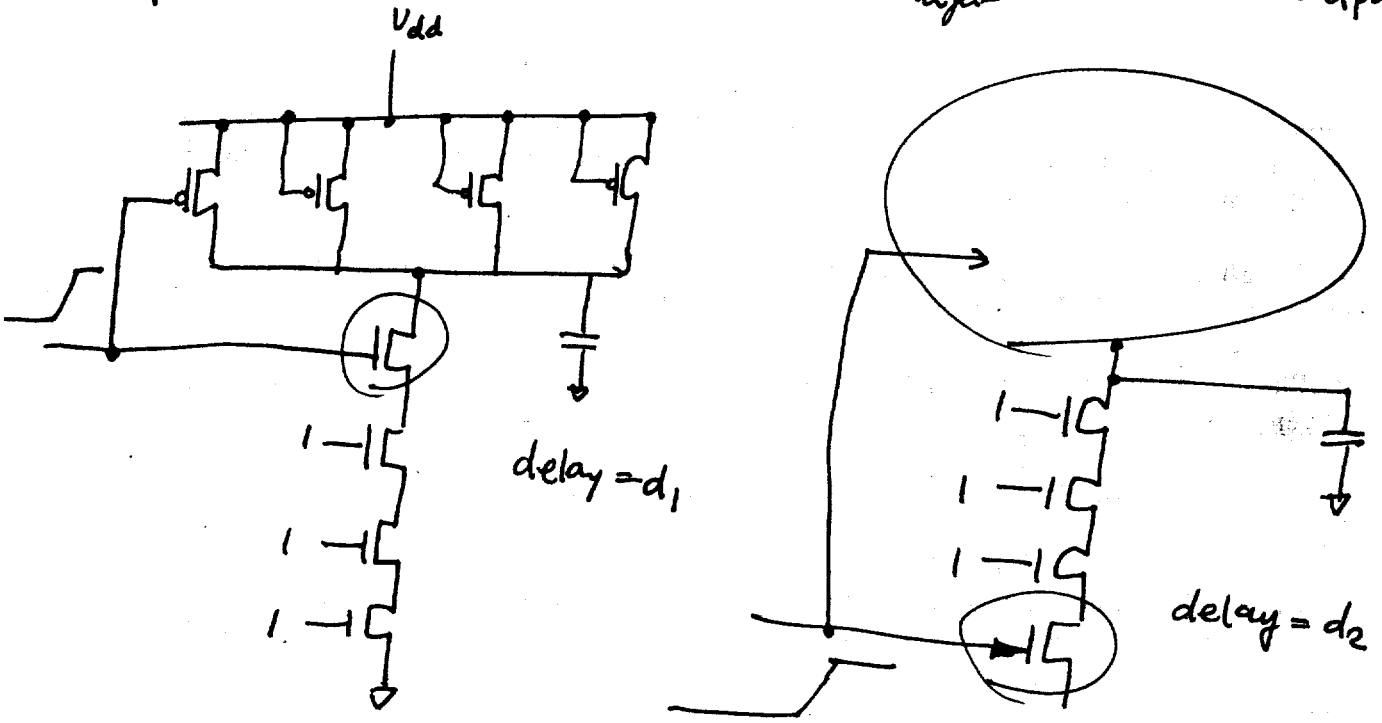


- C_{Z1} will not be discharged

Delay and Transistor ordering

- Difficult to determine the order for delay minimization.
- Depends on input arrivals time and transition times.
- The example at page 150 suggests to move transition signal close to output.

Example:



If rise time of input = 5ns, $d_1/d_2 = 1.23$

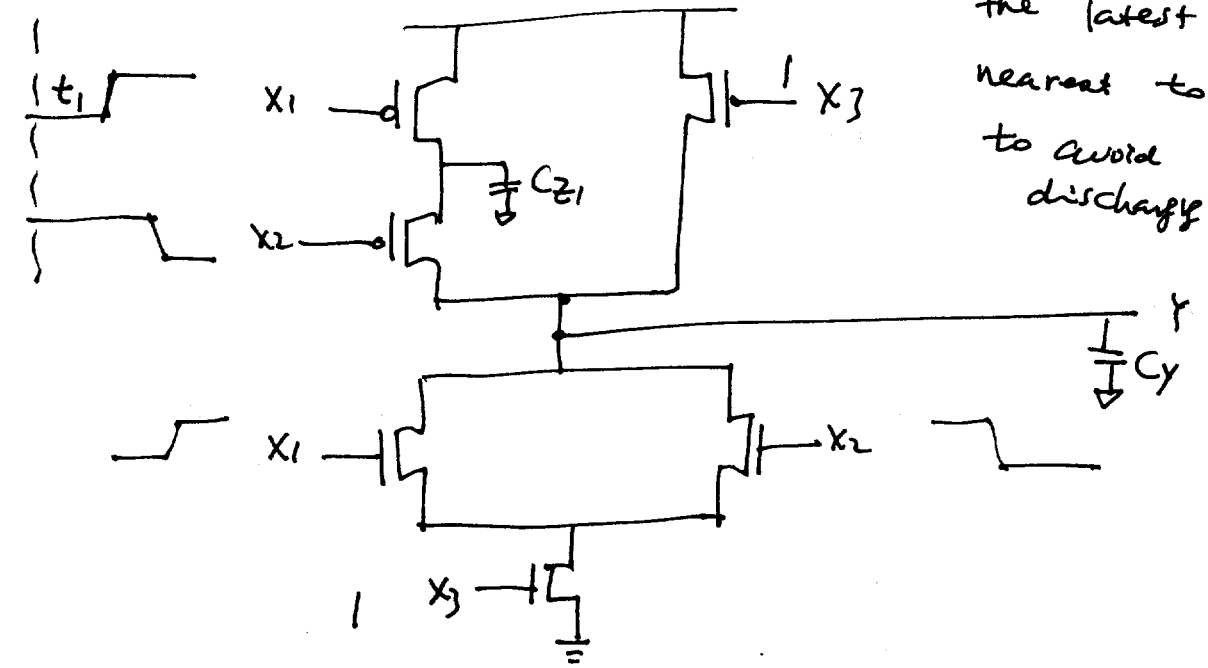
If rise time of input = 1ns, $d_1/d_2 = 0.92$

∴ Spice simulation must be used to determine the optimal ordering for delay minimization.

Rule of thumb.

* place transistors with the latest arrive signal nearest to the output. to avoid charge or discharge the internal capacitances to reduce gate delay.

True or False.



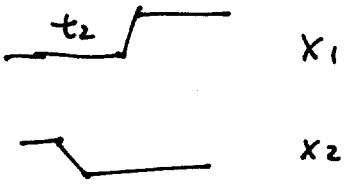
Case I.

X_1 charge C_{z1}

X_2 discharge $C_{z1} + C_y$

$t(X_2 \text{ transition}) > t(X_1 \text{ transition})$

Case II. $t(X_2 \text{ transition}) < t(X_1 \text{ transition})$



$\therefore$ $t_2 > t_1$

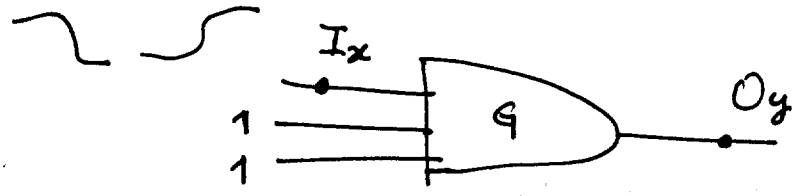
$\therefore$

X_1 charge C_{z1} more than above

than X_2 discharge $C_{z1} + C_y$.

• A possible solution

• Delay characterization for all gates in the library.



It appears that t_f and t_r are not considered.

$T^i_{xy}(G)$: intrisic delay (without output load)

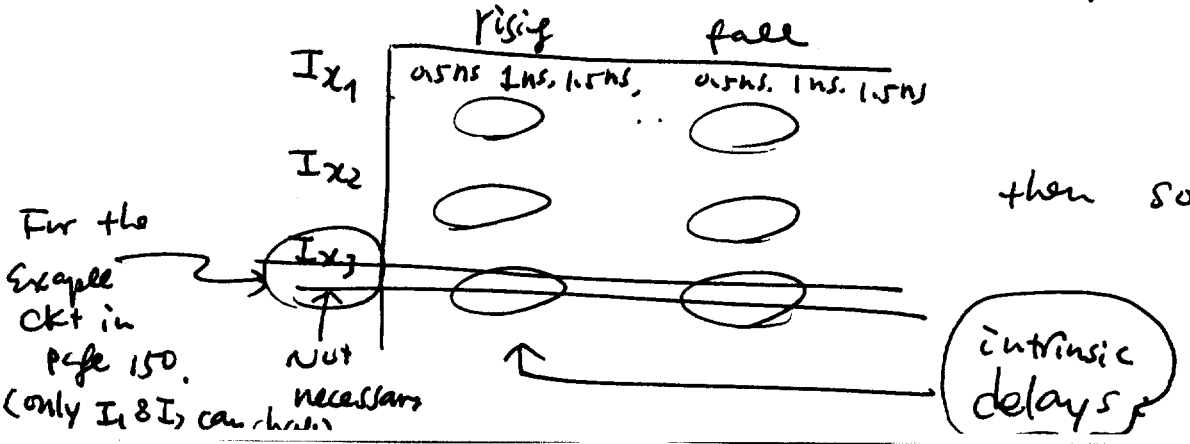
$R_{xy}(G)$: additional delay per unit fanout load

⇒ Total propagation delay through a gate G from I_x to O_y is

$$T^i_{xy}(G) + R_{xy}(G) \cdot C_y(G)$$

(total output load)

- Create characterization for each input
- Sort the delay according to their ordering



then sorted.