



EECS 151/251A
Spring 2020
Digital Design and
Integrated Circuits

Instructor:
John Wawrzynek

Lecture 12: Timing part 2

Announcements

- ❑ Virtual Front Row for today 2/25:
 - ❑ Naomi Sagan
 - ❑ Peter Trost
 - ❑ Jiefeng Chen
 - ❑ Rajiv Govindjee
 - ❑ James Shi
- ❑ **Questions/comments used in class participation points.**
- ❑ Homework assignment 5 posted - due Monday. *Explanations are important!*

Modeling Gate Delay

Inverter Transient Response

With:

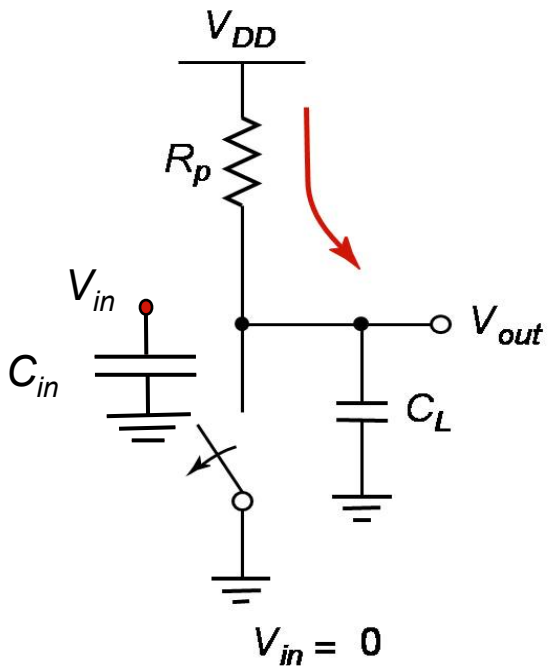
resistive approximation for FETs,
high-to-low (HL)



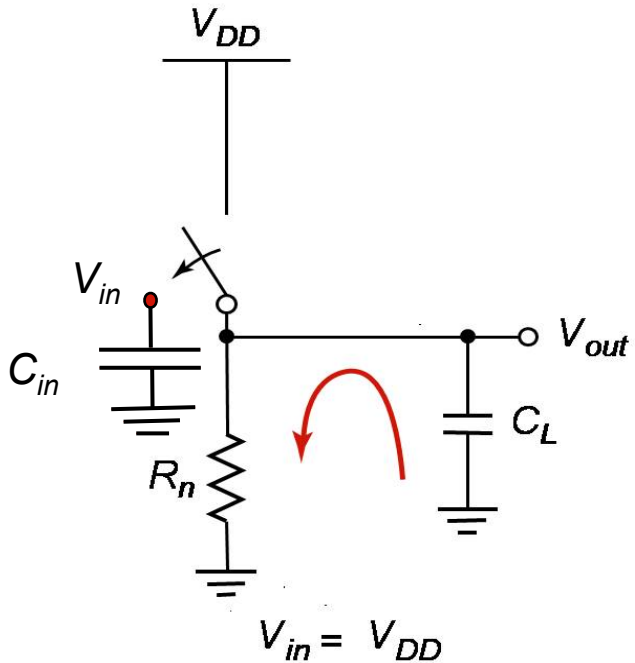
$$V(t) = V_0 e^{-t/RC}$$

$$t_{1/2} = \ln(2) \times RC$$

$$t_{pHL} = f(R_{on} C_L) = 0.69 R_n C_L$$

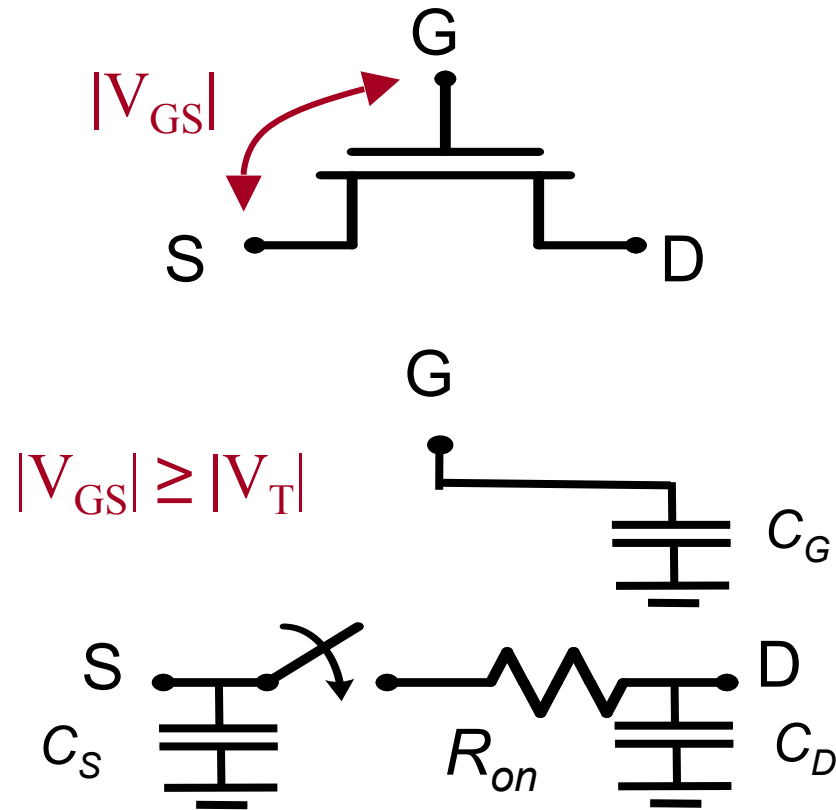


(a) Low-to-high



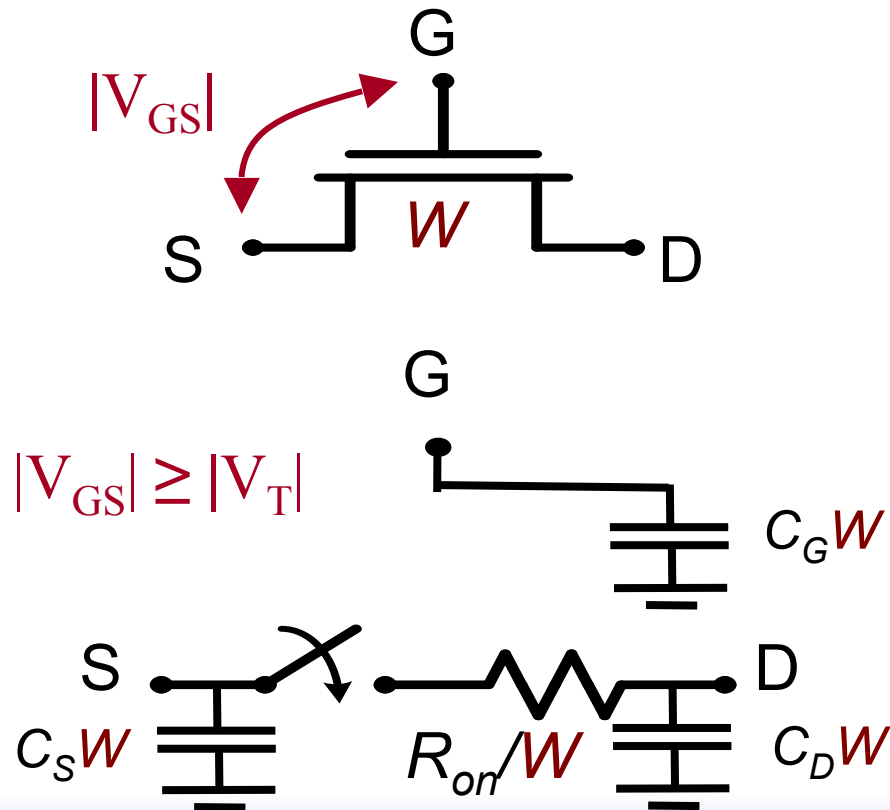
(b) High-to-low

The Switch – Dynamic Model (Simplified)



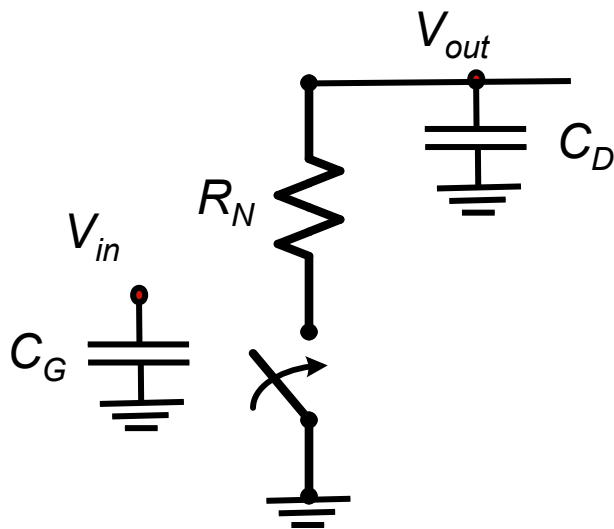
Switch Sizing

What happens if we make a MOSFET W times larger (wider)?

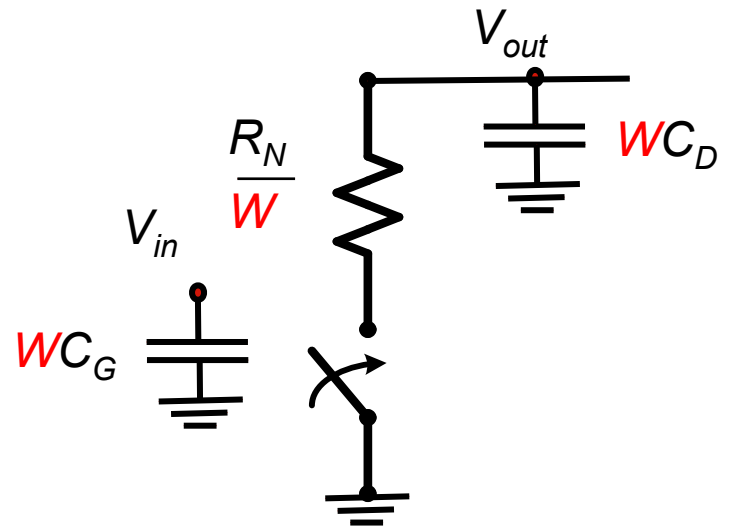


Switch Parasitic Model

The pull-down switch (NMOS)



Minimum-size switch



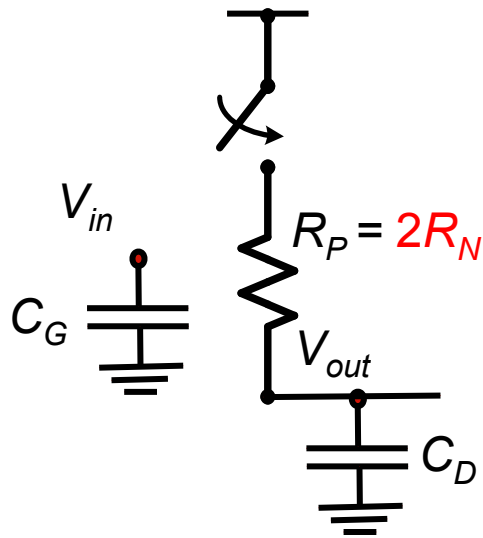
Sizing the transistor (factor W)

We assume transistors of minimal length (or at least constant length). R 's and C 's in units of per unit width.

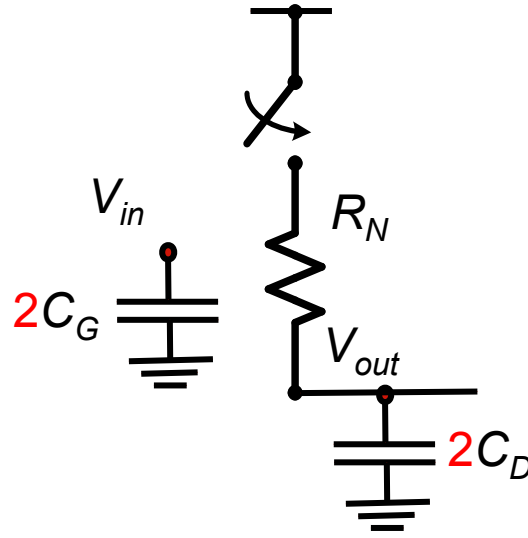
pFET Switch Parasitic Model

For traditional CMOS processes, pFETs are ~twice as resistive as nFETs. (Mobility of holes is 1/2 that of electrons).

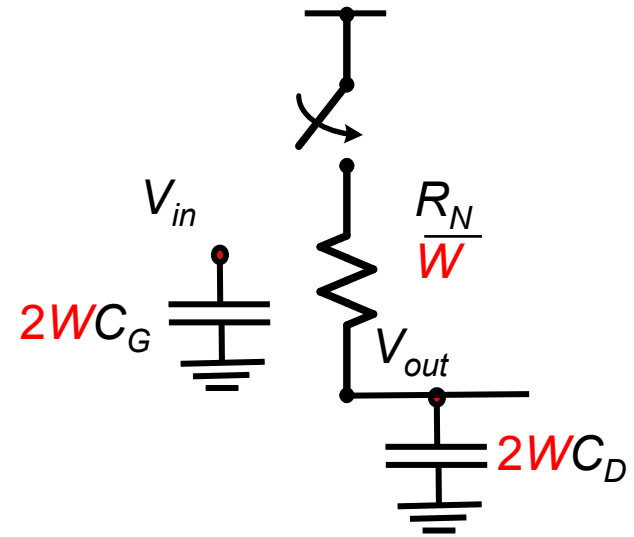
The pull-up switch (PMOS)



Minimum-size switch



Sized for symmetry

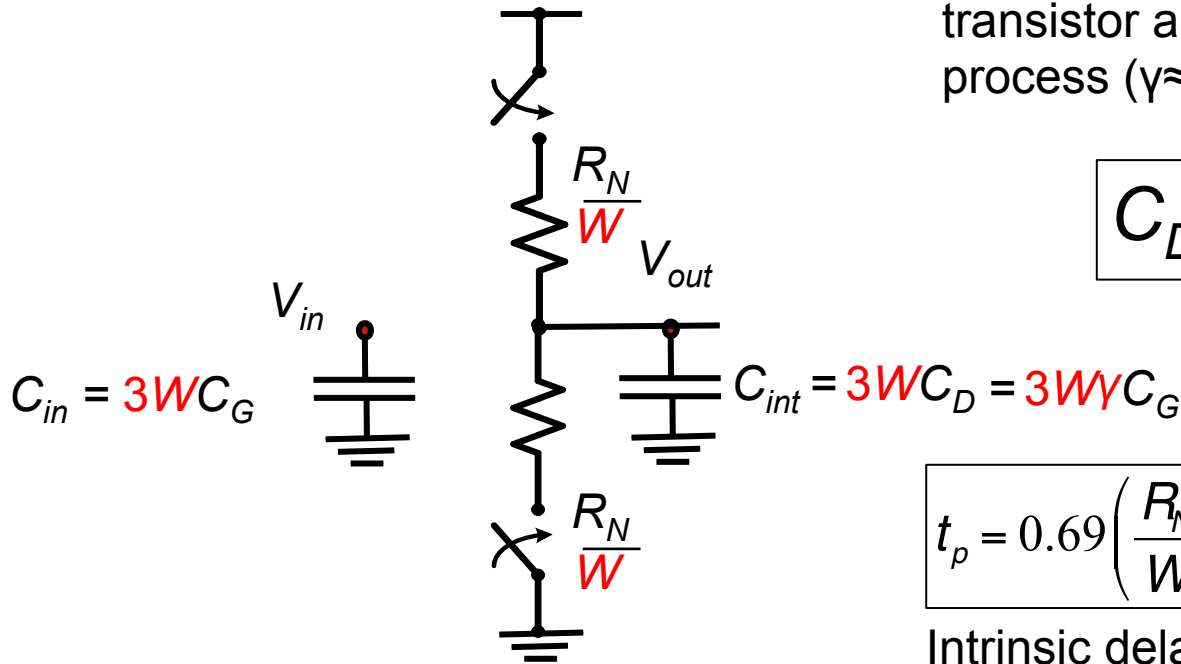


General sizing

“Balanced” Inverter Parasitic Model

Drain and gate capacitance of transistor are **directly** related by process ($\gamma \approx 1$)

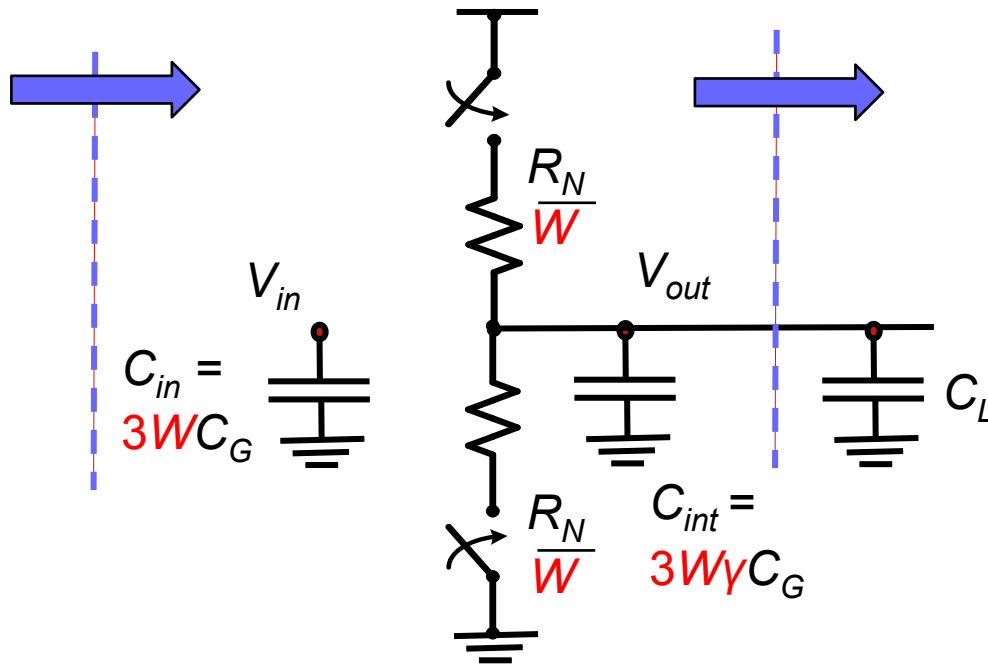
$$C_D = \gamma C_G$$



$$t_p = 0.69 \left(\frac{R_N}{W} \right) (3W\gamma C_G) = 0.69(3\gamma) R_N C_G$$

Intrinsic delay of inverter
independent of size

Inverter with Load Capacitance



$$t_p = 0.69(R_N/W)(C_{int} + C_L)$$

$$= 0.69(R_N/W)(3W\gamma C_G + C_L)$$

factor out $3W\gamma C_G$

replace $C_{int} = 3WC_G$

$$= 0.69(3\gamma R_N C_G) \left(1 + \frac{C_L}{\gamma C_{in}}\right)$$

$$= t_{p0} \left(1 + \frac{C_L}{\gamma C_{in}}\right) = t_{p0} (1 + f/\gamma)$$

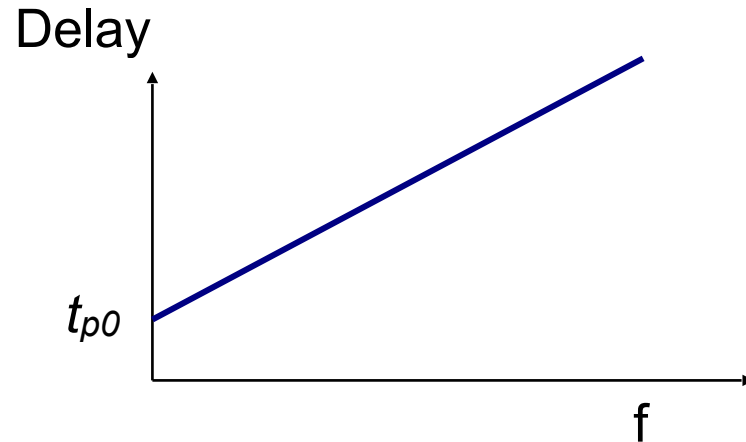
$f = \text{fanout}$ = ratio of load capacitance (C_L) to and input capacitance (C_{in})

Inverter Delay Model

Delay linearly proportional to fanout, f .

For $f=0$, delay is intrinsic inverter delay $t_{inv} = t_{p0}$

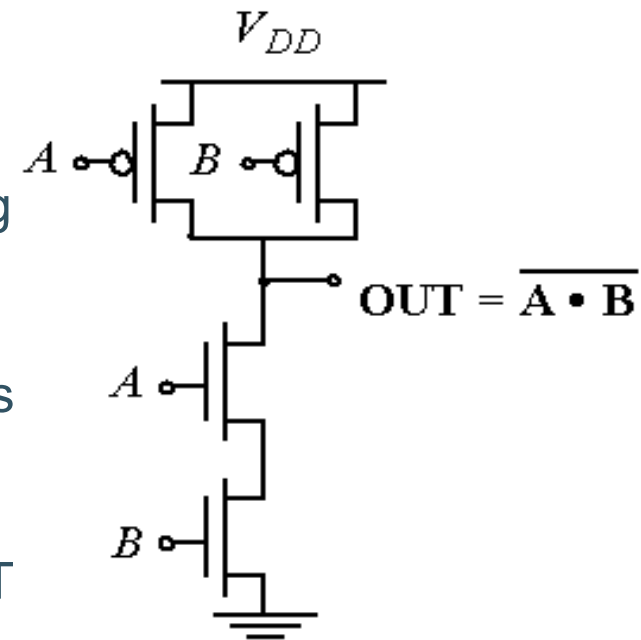
$$t_p = t_{p0}(1 + f/\gamma)$$



Question: how does transistor sizing (W) impact delay?

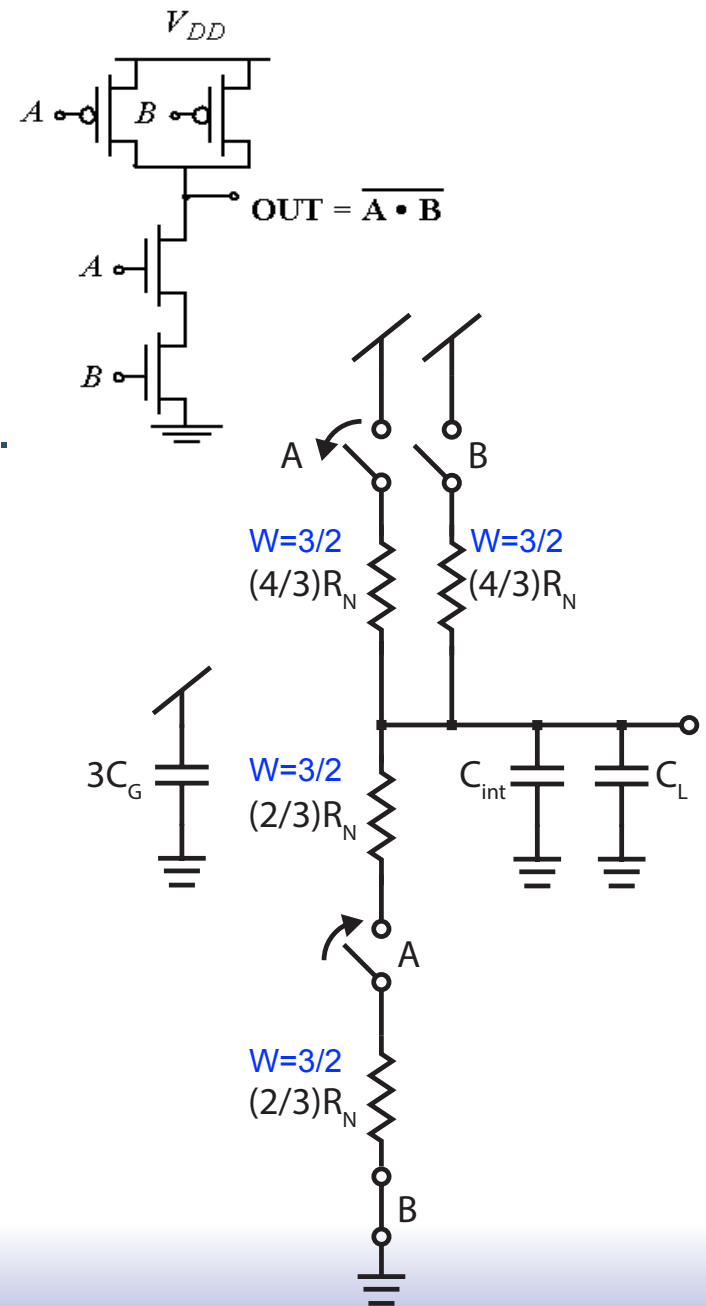
2-input NAND Gate

- ❑ Let's derive the formula for gate propagation as a function of fanout, f (as with inverter)
- ❑ We derive the equations based on the input connected to the transistor closest to the output (A),
- ❑ assuming the B input had been set to 1 (for a long time)
- ❑ So we can fairly compare to the inverter, size the transistors so that the capacitance of each input is equivalent to the input capacitance of the inverter
- ❑ Assume that the resistance of the pFET is twice that of the nFET ($R_p = 2R_n$) if the pFET and nFET have the same width
- ❑ Size the transistors so that the rise time and fall times are equivalent
- ❑ For the 2 transistors in series, ignore the capacitance at their shared node

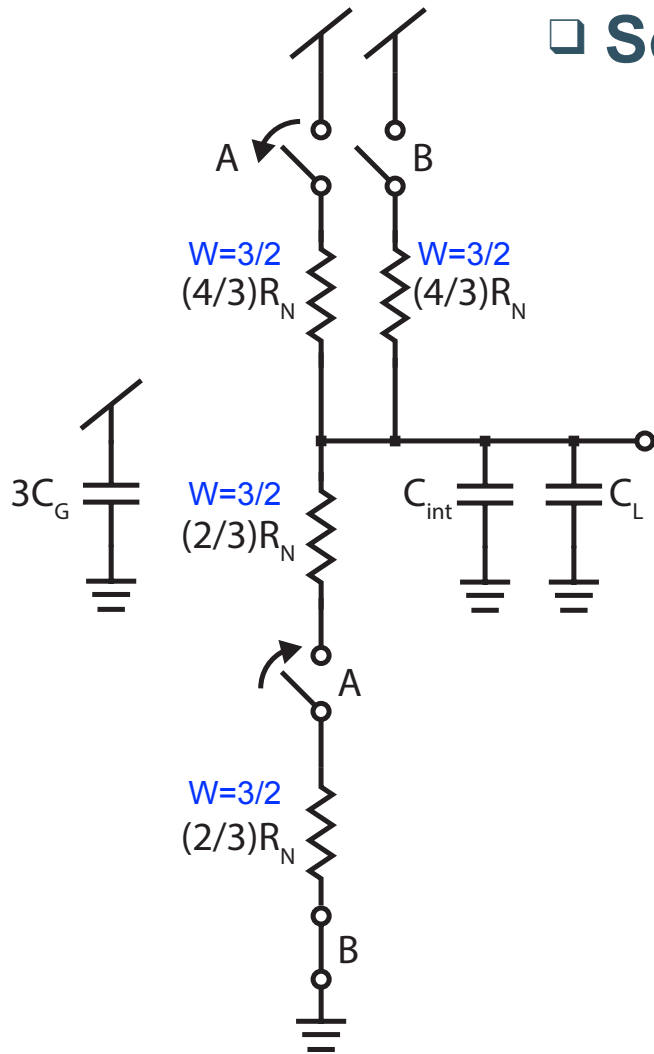


2-input NAND Gate

- **Solution setup:**
- To keep the pullup and pulldown delays the same, $W_p = W_n$, because $R_p = 2R_n$.
- Remember, inverter had input C of $3C_G$, with $W_p = 2W_n$.
- Therefore, here we increase widths by $3/2$ relative to the inverter, so
- R changes by $2/3$ (shown in figure).



2-input NAND Gate



□ Solve:

$$C_{int} = (6/2)WC_D + (3/2)WC_D = (9/2)WC_D$$

$$t_p = 0.69 \cdot 2 \left(\frac{2R_N}{3W} \right) (C_{int} + C_L)$$

$$= 0.69 \left(\frac{4R_N}{3W} \right) \left(\frac{9}{2}\gamma WC_G + C_L \right)$$

$$= 0.69 \left(\frac{R_N}{W} \right) \left(6\gamma WC_G + \frac{4}{3}C_L \right)$$

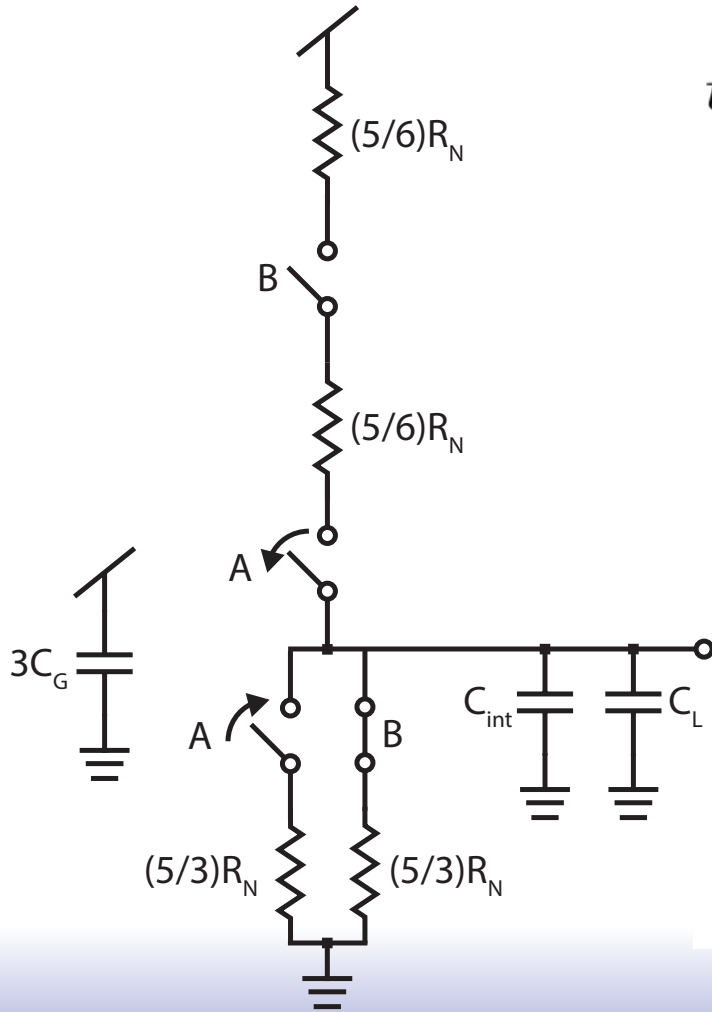
$$= 0.69 \left(\frac{R_N}{W} \right) 3\gamma WC_G \left(2 + \frac{4C_L}{3\gamma(3WC_G)} \right)$$

$$= [0.69 \cdot 3R_N\gamma C_G] \left(2 + \frac{4C_L}{3\gamma C_{IN}} \right)$$

$$= t_{p0} \left(2 + \frac{4f}{3\gamma} \right)$$

2-input NOR Gate

$$W_p = 4W_n$$



$$C_{int} = 2 \times (3/5/C_D + (12/5)C_D) = (18/5)C_D$$

$$t_p = 0.69 \left(\frac{3R_N}{5W} \right) (C_{int} + C_L)$$

$$= 0.69 \left(\frac{5R_N}{3W} \right) \left(\frac{18}{5} \gamma W C_G + C_L \right)$$

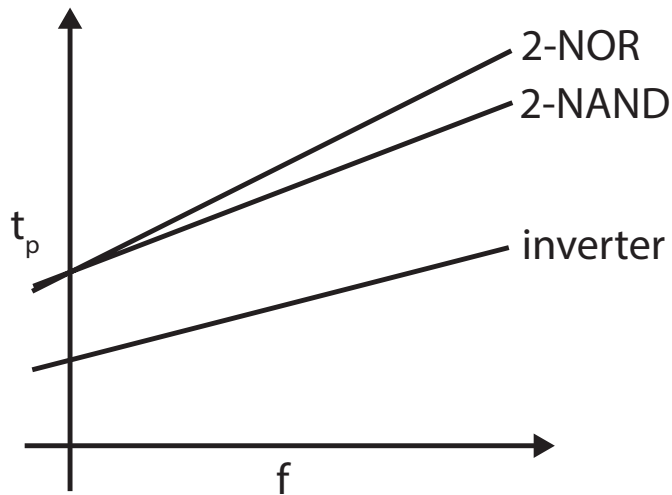
$$= 0.69 \left(\frac{R_N}{W} \right) \left(6\gamma W C_G + \frac{5C_L}{3} \right)$$

$$= 0.69 \left(\frac{R_N}{W} \right) 3\gamma W C_G \left(2 + \frac{5C_L}{3(3\gamma W C_G)} \right)$$

$$= [0.69 \cdot 3R_N \gamma C_G] \left(2 + \frac{5C_L}{3\gamma C_{IN}} \right)$$

$$= t_{p0} \left(2 + \frac{5f}{3\gamma} \right)$$

Gate Delay Summary



The y-intercepts (intrinsic delay) for NAND and NOR are both twice that of the inverter. The NAND line has a gradient $4/3$ that of the inverter (steeper); for NOR it is $5/3$ (steepest).

What about gates with more than 2-inputs?

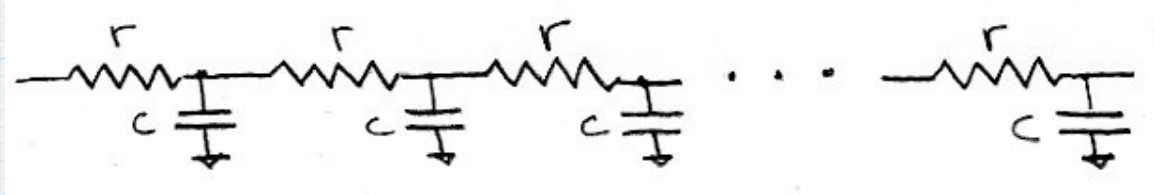
Look at 4-input NAND:

$$t_p = t_{p0} \left(4 + \frac{2f}{\gamma} \right)$$

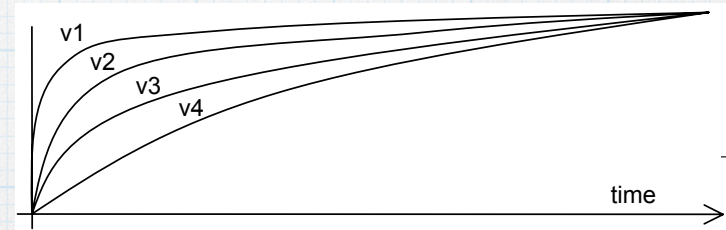
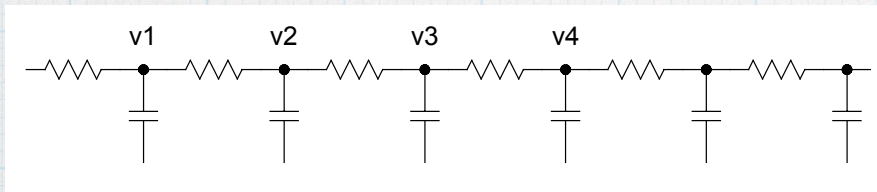
intercept slope

Adding Wires to gate delay

- ▶ Wires have finite resistance, so have distributed R and C:

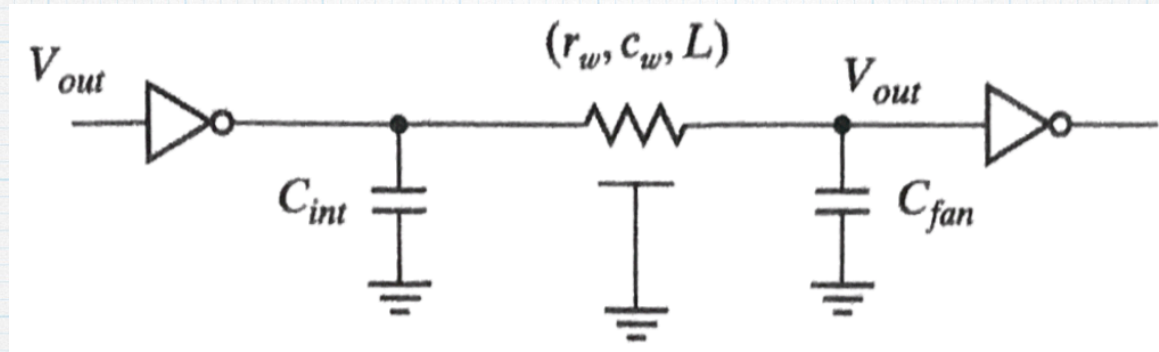


with $r = \text{res/length}$, $c = \text{cap/length}$, $\Delta t \propto r c L^2 \cong rc + 2rc + 3rc + \dots$



- ▶ Wire propagation delay is around half of what it would be if R and C were "lumped": $t_p = 0.38(rL * cL) = 0.38rcL^2$

Gate Driving long wire and other gates

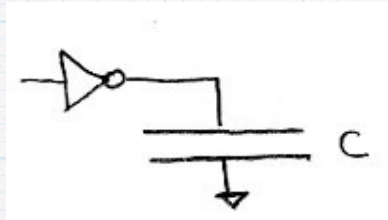


$$R_w = r_w L, \quad C_w = c_w L$$

$$\begin{aligned} t_p &= 0.69R_{dr}C_{int} + 0.69R_{dr}C_w + 0.38R_wC_w + 0.69R_{dr}C_{fan} + 0.69R_wC_{fan} \\ &= 0.69R_{dr}(C_{int} + C_{fan}) + 0.69(R_{dr}c_w + r_wC_{fan})L + 0.38r_wc_wL^2 \end{aligned}$$

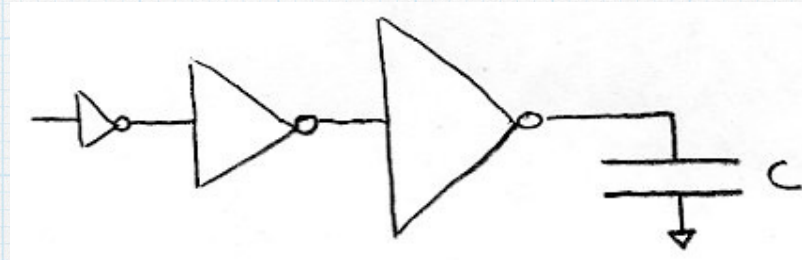
Driving Large Loads

- ▶ Large fanout nets: clocks, resets, memory bit lines, off-chip
- ▶ Relatively small driver results in long rise time (and thus large gate delay)



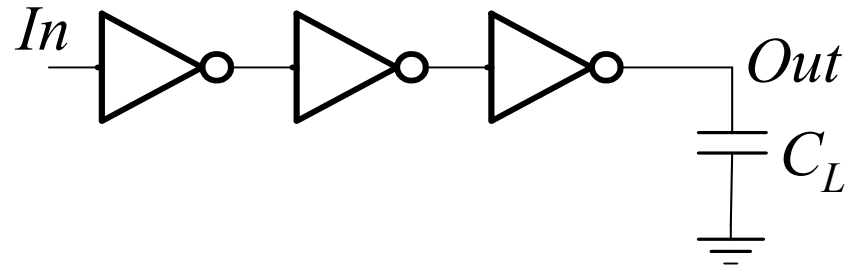
- ▶ Strategy:

Staged Buffers

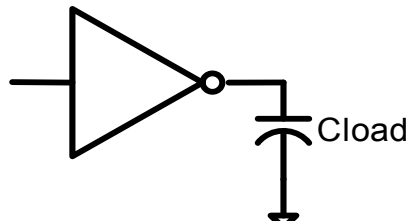


- ▶ How to design for optimal performance (least delay)?
- ▶ Should be obvious that total delay is minimized with equal delay at each stage.

Driving Large Loads

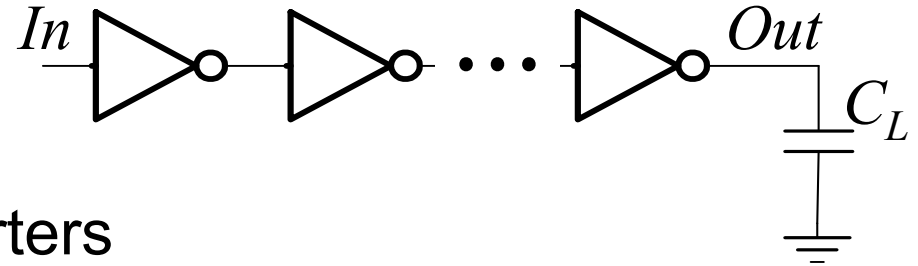


- For some given C_L :
 - How many stages are needed to minimize delay?
 - How to size the inverters?
- Get smallest delay if build one **very** big inverter
 - So big that delay is set only by self-loading



- Not an interesting solution. Why?
 - Something has to drive this inverter (a big inverter has a large input capacitance!) ...

Delay Optimization

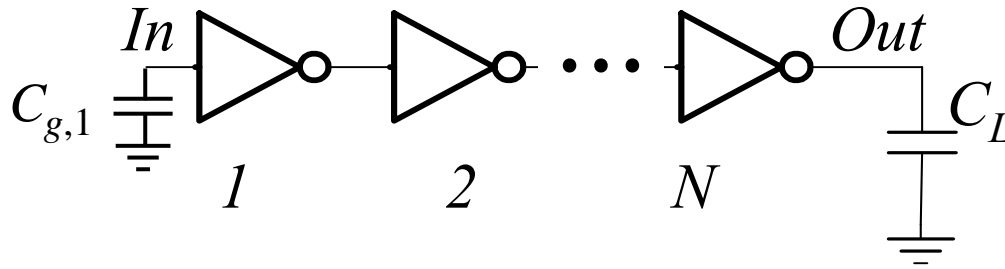


□ First assume given:

- A fixed number of inverters
- The size* of the first inverter
- The size of the load that needs to be driven

□ What is the minimal delay of the inverter chain?

** note: When we talk about inverter (or gate) “size”, we refer to the wide of the transistors making up the circuit.*



- Delay for the j -th inverter stage:

$$t_{p,j} = t_{p0} \left(1 + \frac{C_{g,j+1}}{\gamma C_{g,j}} \right) = t_{p0} (1 + f_j / \gamma)$$

- Total delay of the chain:

$$t_p = \sum_{j=1}^N t_{p,j} = t_{p0} \sum_{j=1}^N \left(1 + \frac{C_{g,j+1}}{\gamma C_{g,j}} \right), \quad C_{g,N+1} = C_L$$

Optimum Delay and Number of Stages

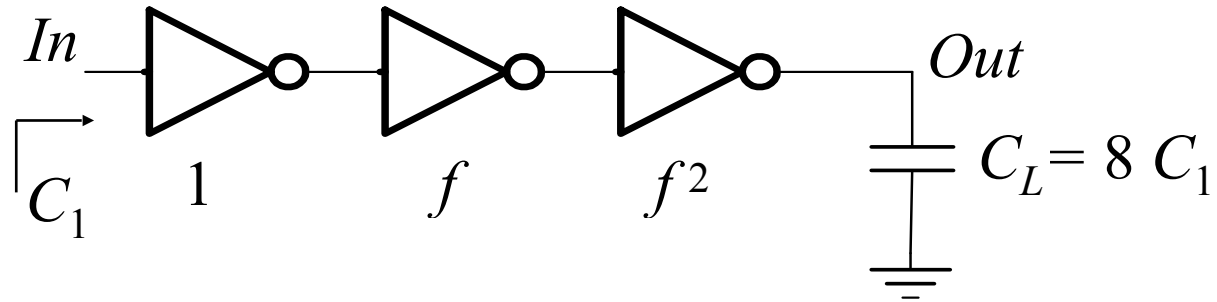
- Each inverter should be sized up by the same factor f with respect to the preceding inverter
- Therefore each stage has the equal delay
- Given $C_{g,1}$ and C_L

$$f = \sqrt[N]{C_L/C_{g,1}} = \sqrt[N]{F}$$

- Where F represents the overall fan-out of the circuit
- Therefore the minimal delay through the chain is:

$$t_p = N \cdot t_{p0}(1 + \sqrt[N]{F}/\gamma)$$

Example



C_L/C_1 has to be evenly distributed across $N = 3$ stages:

$$f = \sqrt[N]{C_L/C_{g,1}} = \sqrt[N]{F} = \sqrt[3]{8} = 2$$

Delay Optimization

- Now assume given:
 - The size of the first inverter
 - The size of the load that needs to be driven
- Minimize delay by finding optimal number and sizes of inverters
- So, need to find N that minimizes:

$$t_p = N \cdot t_{p0} (1 + \sqrt[N]{F/\gamma}), \quad F = C_L / C_{g,1}$$

Finding optimal fanout per stage

$$t_p = N \cdot t_{p0} (1 + \sqrt[N]{F}/\gamma), \quad F = C_L/C_{g,1}$$

- Differentiate w.r.t. N and set = 0:

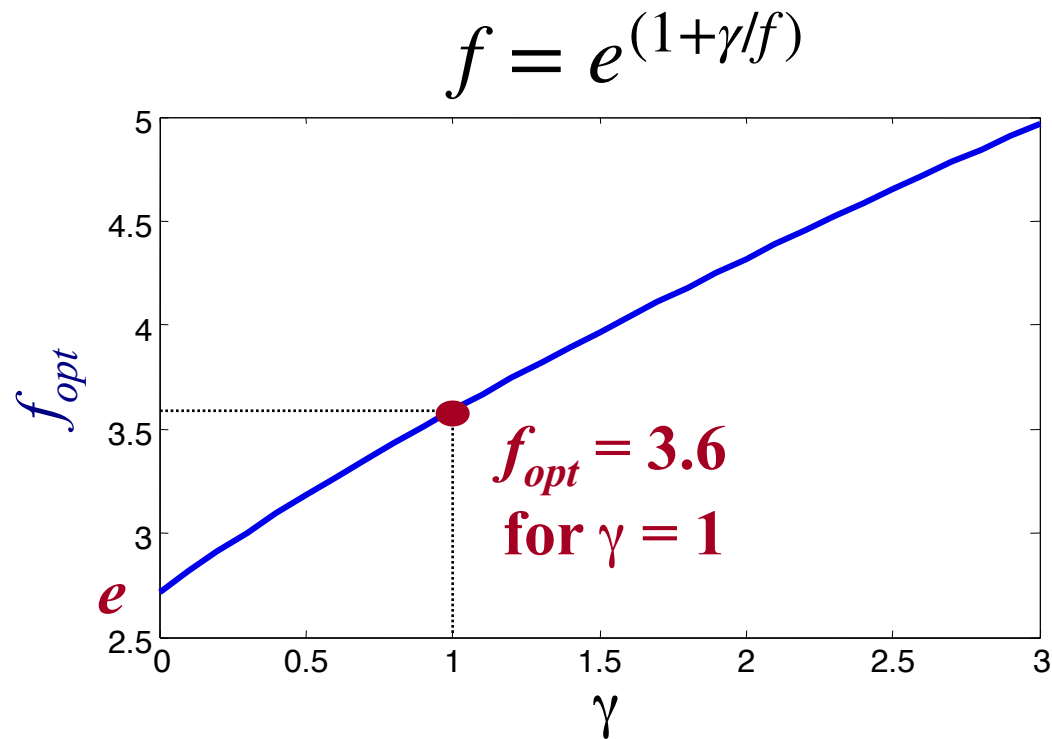
$$\gamma + \sqrt[N]{F} - \frac{\sqrt[N]{F} \ln F}{N} = 0$$

$$\Rightarrow \boxed{f = e^{(1+\gamma/f)}}$$

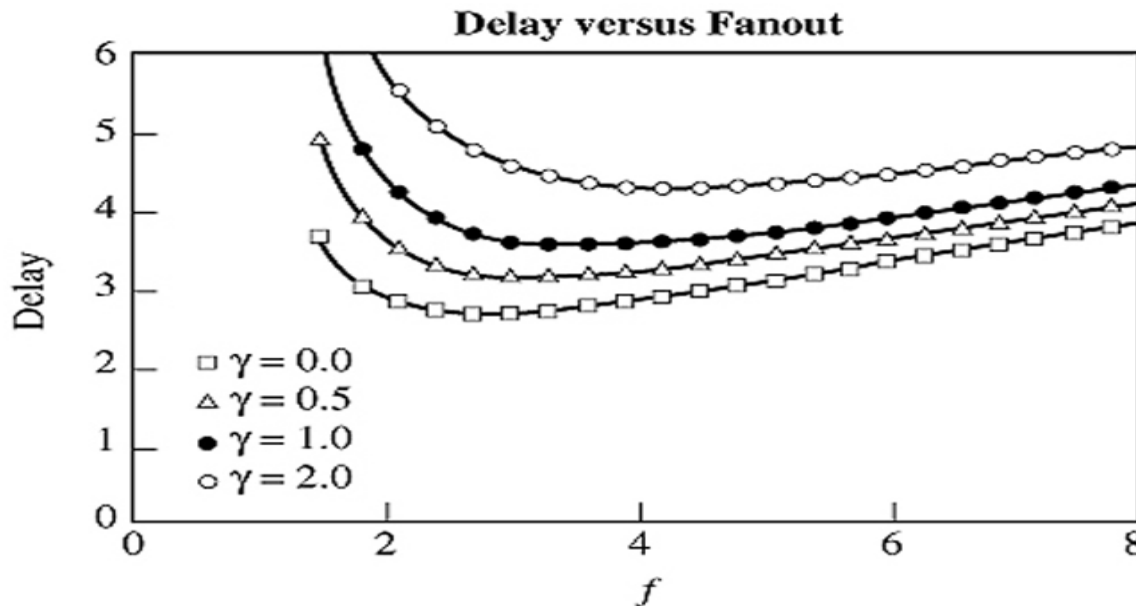
- Closed form only if : $\gamma = 0 \Rightarrow N = \ln(F), f = e$

Optimum Effective Fanout f

- Optimum f for given process defined by γ



In Practice: Plot of Total Delay

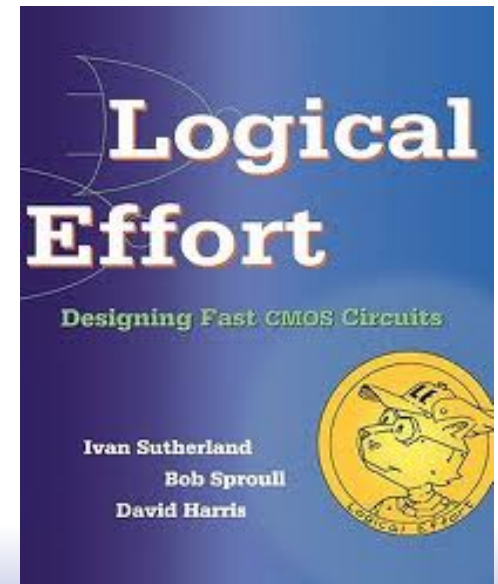
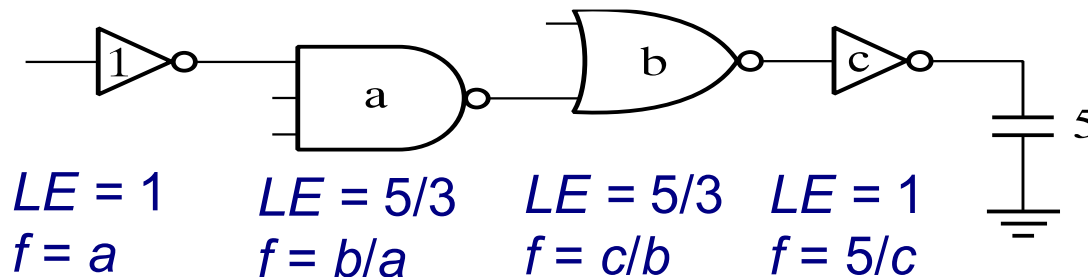


[Hodges, p.281]

- Why the shape?
- Curves very flat for $f > 2$
 - Simplest/most common choice: $f = 4$

Transistor Sizing in Logic Circuits

- ❑ Similar optimization challenges exist within all combinational logic blocks. How do we size transistors to minimize a given circuit?
- ❑ ASIC standard cell libraries include cells with various output drive strength (transistor sizes)
- ❑ Tools will automatically choose the proper size and/or add buffers to minimize critical path logic delay
- ❑ Hand methods exist for minimizing logic path delay:



End of lecture 12