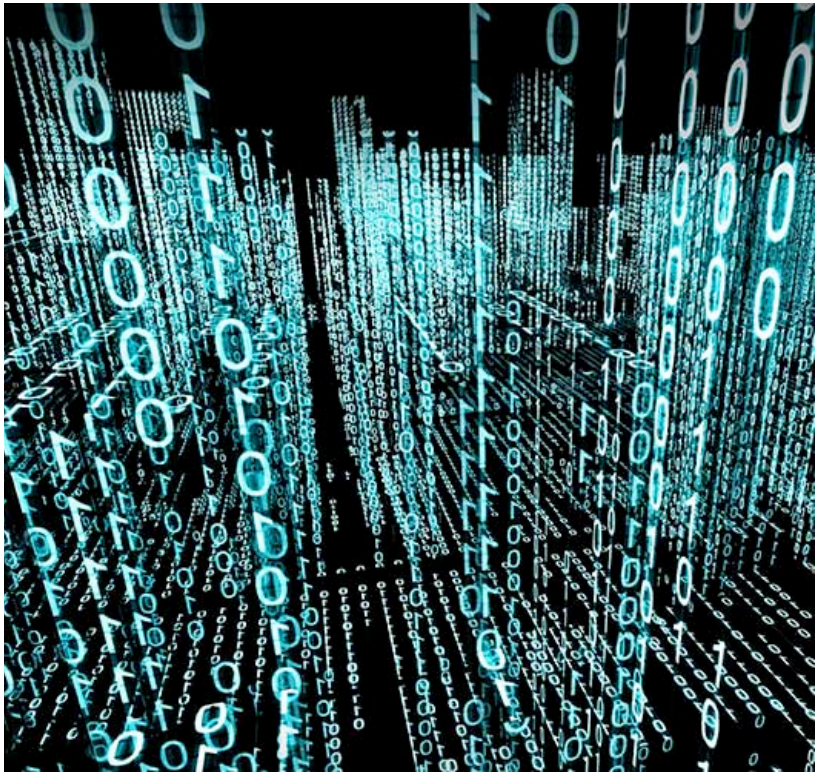
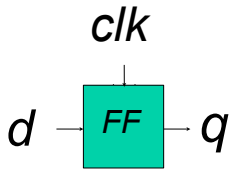


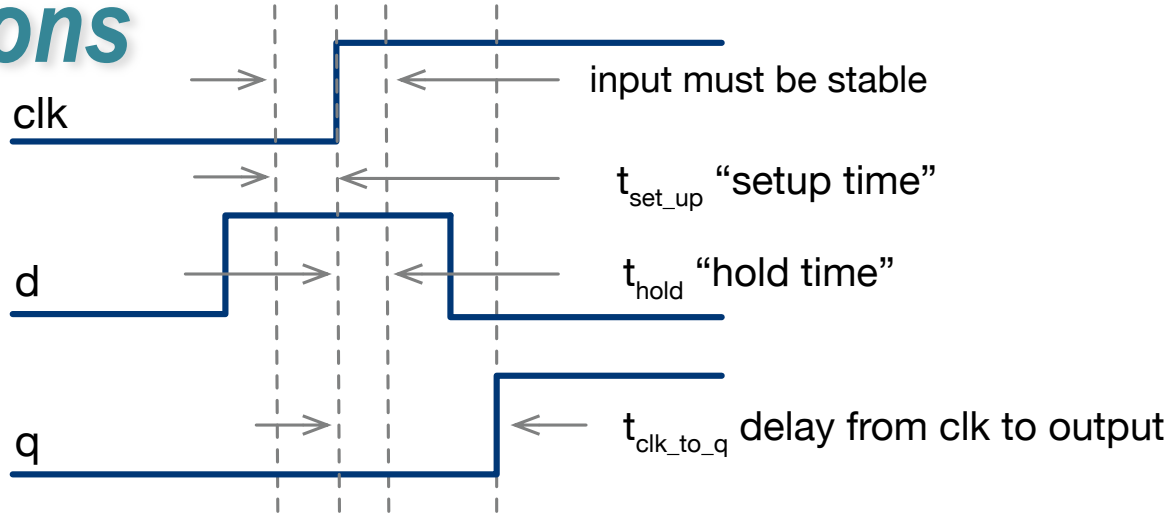
EECS 151/251A
Spring 2024
Digital Design and Integrated Circuits

Instructor:
John Wawrzynek

Lecture 12: Timing part 2

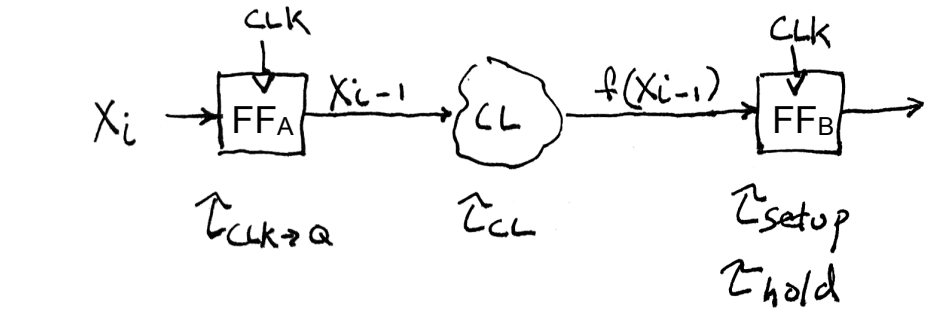
Hold-time Violations

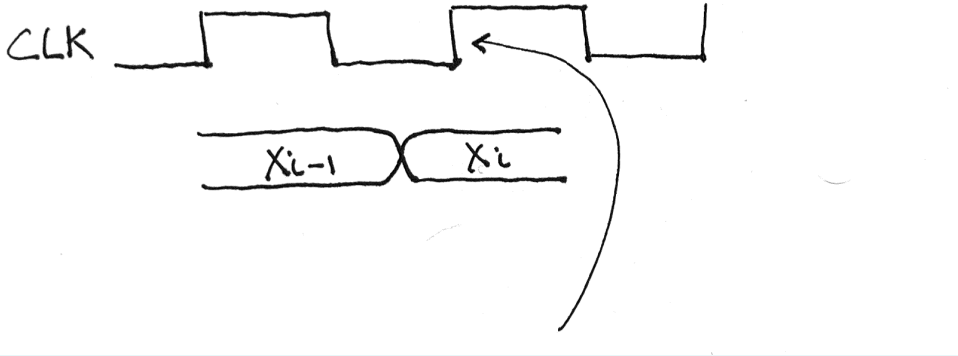




- Some state elements have positive hold time requirements.
  - How can this be?
- Fast paths from one state element to the next can create a violation.
- CAD tools do their best to fix violations by inserting delay (buffers).
  - Of course, if the path is delayed too much, then F_{max} suffers.
  - Difficult because buffer insertion changes layout, which changes path delay.

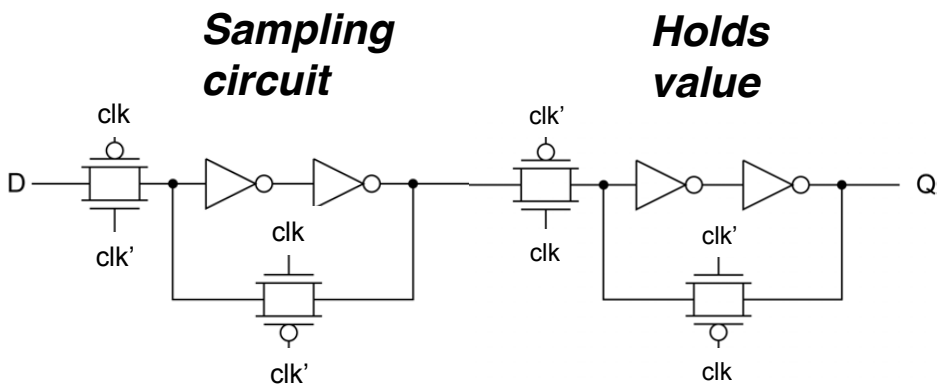
Hold-time Violations



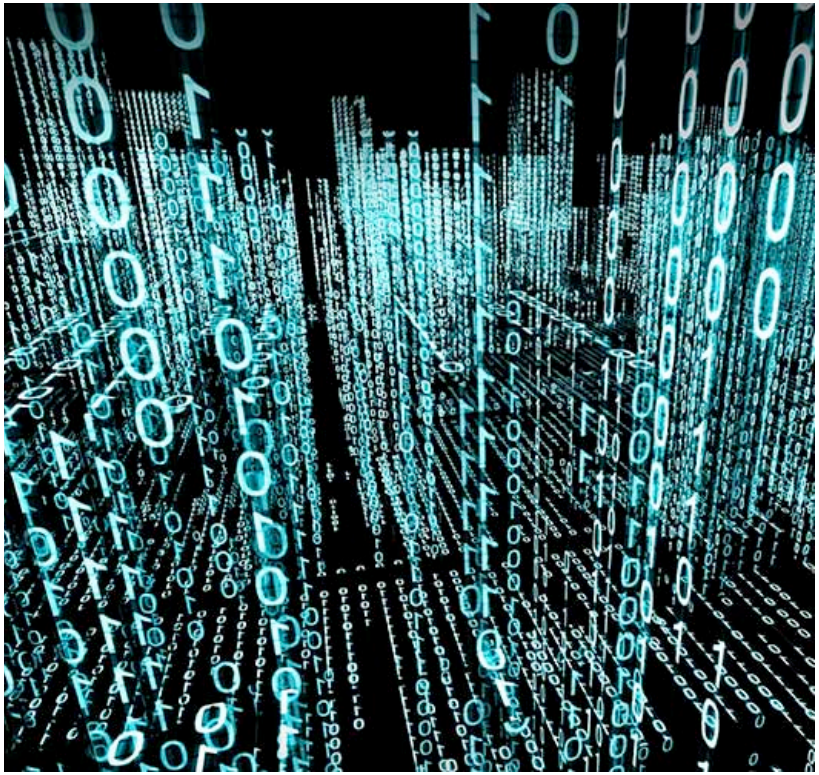


- On this clock edge:
  - FF_A captures x_i (overwrites x_{i-1})
  - FF_B captures $f(x_{i-1})$
- If x_i from FF_A overwrites x_{i-1} and f(xi-1) before FF_B captures it then failure!
- T_{hold} for FF_B specifies how long (after the clock edge) it requires to properly capture $f(x_{i-1})$

Hold-time Violations



- Origin of hold time:
  - Local rebuffering of the clock delays the clock (perhaps relative to the input data)
  - Worse for large state elements with large internal clock load and local rebuffering (such as memory blocks, register files)
- The fix for paths with hold time violations is to increase the delay to the state element
  - If critical path, then Fmax suffers
  - If logic changes (for instance buffer insertion fix), then layout changes which can change timing.



Modeling Gate Delay

Inverter Transient Response

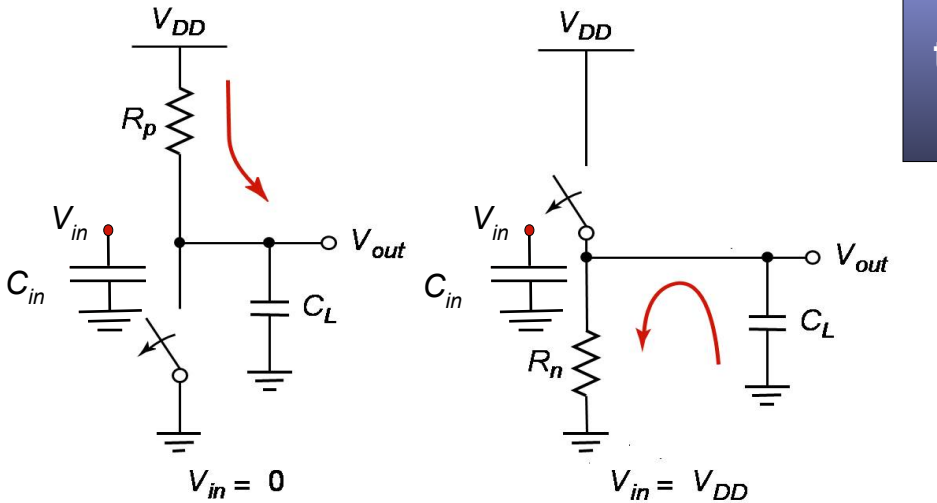
With:

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

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

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

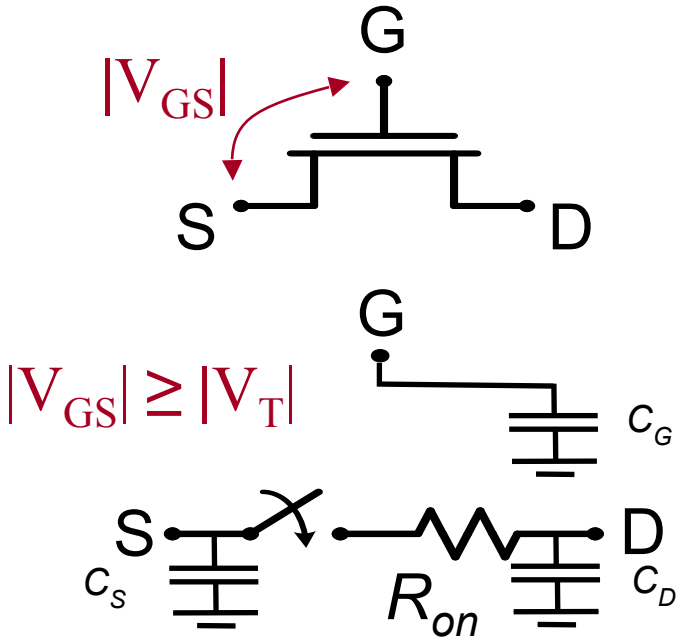
$$t_{pHL} = f(R_{on}C_L) = 0.69\, R_n\, C_L$$



(a) Low-to-high

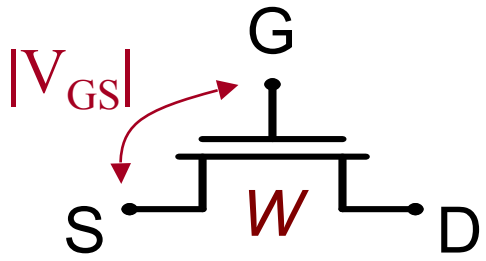
(b) High-to-low

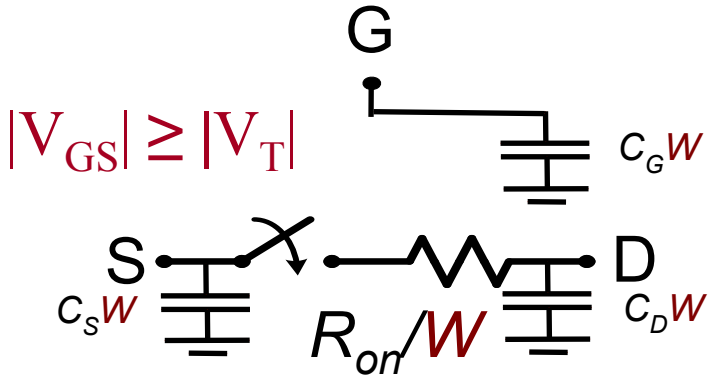
The Switch – Dynamic Model (Simplified)



Transistor Sizing

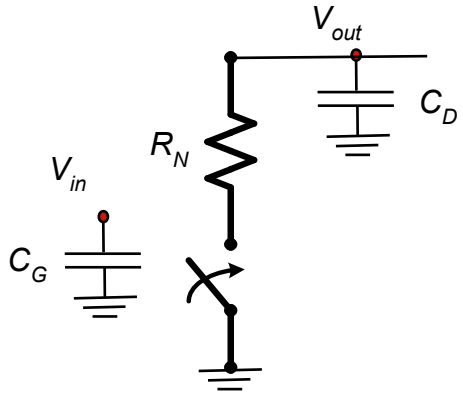
What happens if we make a MOSFET W times larger (wider)?



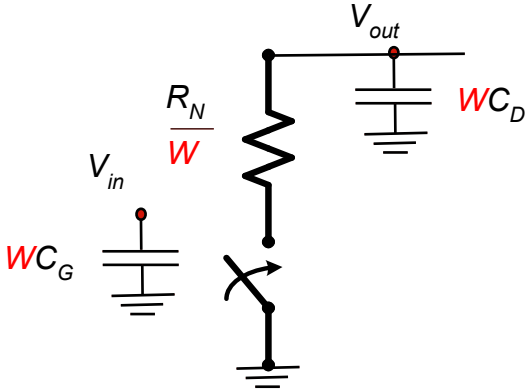


Switch Parasitic Model

The pull-down switch (NMOS)



Minimum-size switch



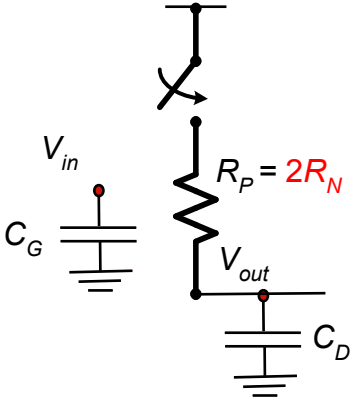
Sizing the transistor (factor *W*)

We assume transistors of minimal length (or at least constant length).

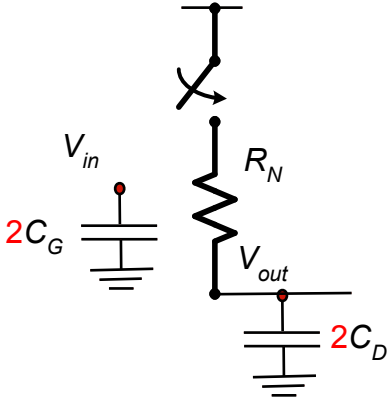
pFET Switch Parasitic Model

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

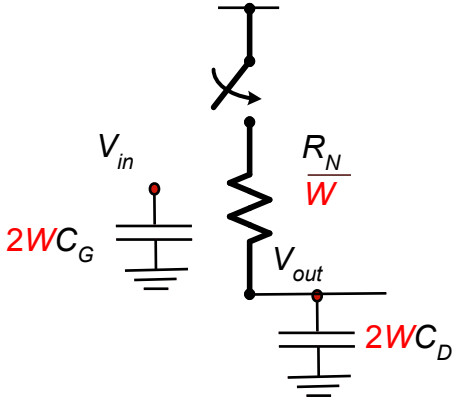
The pull-up switch (PMOS)



Minimum-size pFET

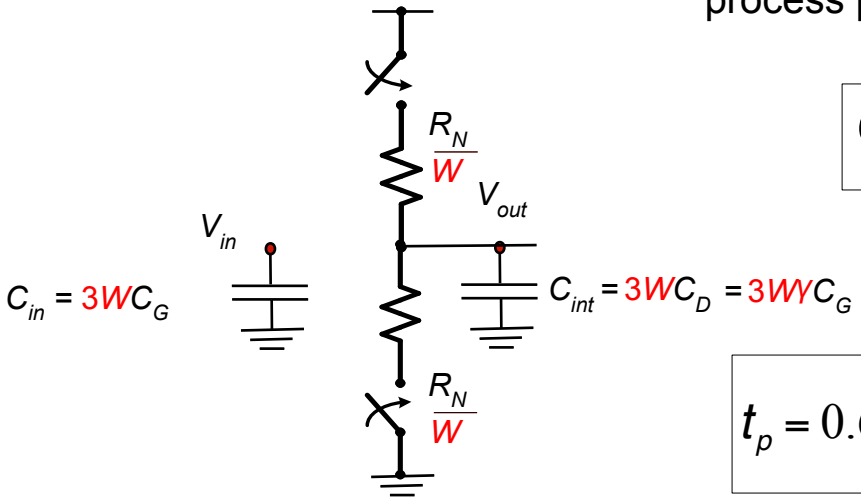


Sized for symmetry



General sizing

“Balanced” Inverter Parasitic Model



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

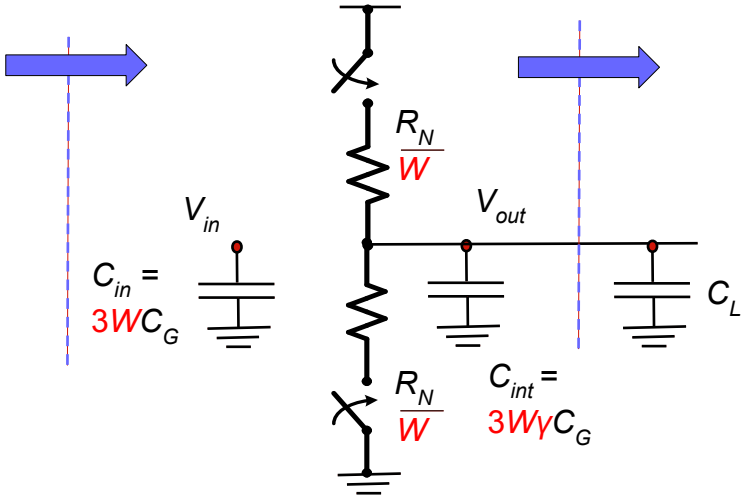
$$C_D = \gamma C_G$$

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

$= t_{p0}$ 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 $C_{in} = 3WC_G$

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

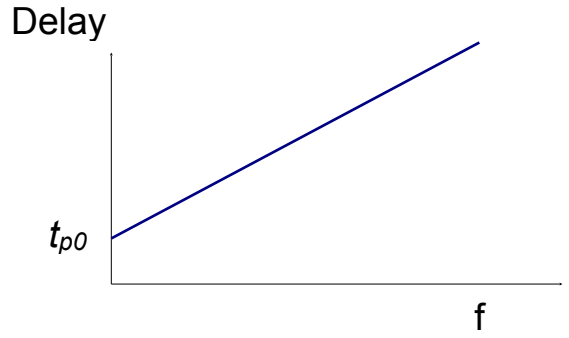
$$= t_{p0}(1 + \frac{C_L}{\gamma C_{in}}) = t_{p0}(1 + f/\gamma)$$

f = **electrical fanout** = ratio of load capacitance (C_L) to and input capacitance (C_{in})

Inverter Delay Model

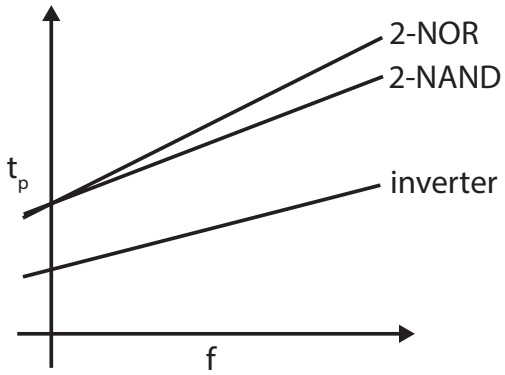
Delay linearly proportional to fanout, f.
For f=0, delay is intrinsic inverter delay $t_{inv} = t_{p0}$

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



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

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).

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

2-input NAND

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

2-input NOR

What about gates with more than 2-inputs?

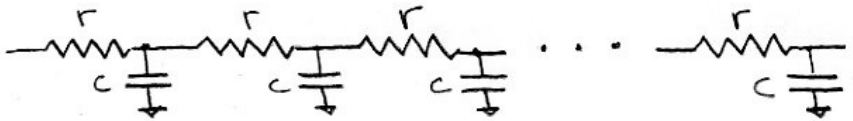
4-input NAND:

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

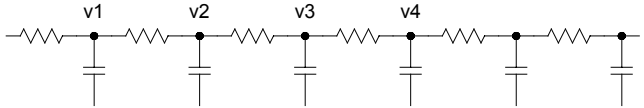
intercept *slope*

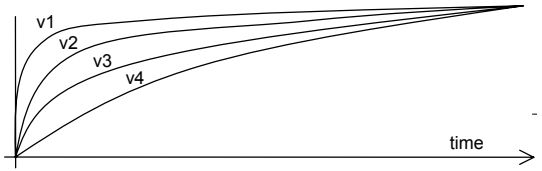
Adding Wires to gate delay

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



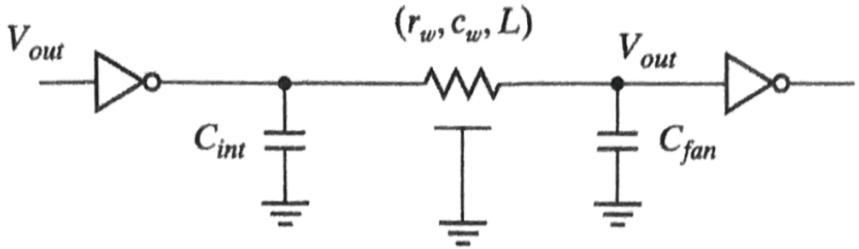
with r = res/length, c = cap/length, $\Delta t \propto rcL^2 \cong rc + 2rc + 3rc + \ldots$





- Wire propagation delay is 0.38 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



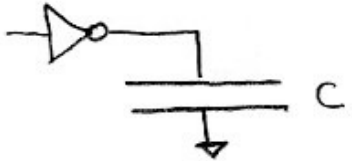
$$R_w = r_w L, \quad C_w = c_w L$$

$$t_p = 0.69R_{dr}C_{int} + 0.69R_{dr}C_w + 0.38R_wC_w + 0.69R_{dr}C_{fan} + 0.69R_wC_{fan}$$

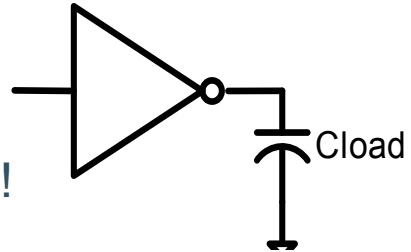
$$= 0.69R_{dr}(C_{int} + C_{fan}) + 0.69(R_{dr}c_w + r_wC_{fan})L + 0.38r_wc_wL^2$$

Driving Large Loads

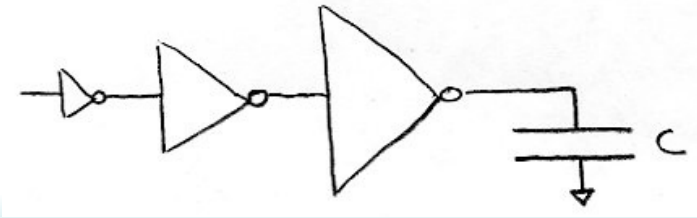
- Large fanout nets: clocks, resets, memory bit lines, off-chip



- Relatively small driver results in long rise time (and thus large gate delay)

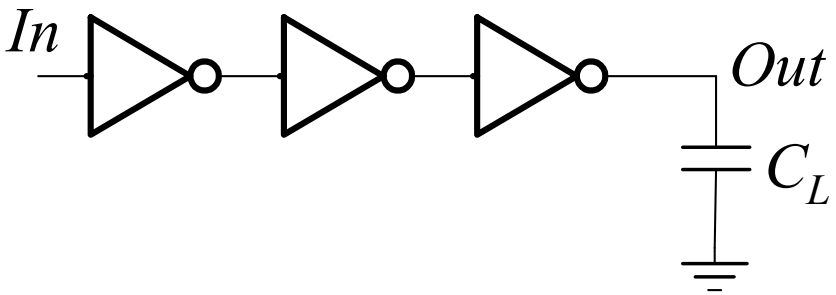


- Could build one very big inverter
- The large inverter has have C_{in} and something needs to drive it!
- Strategy: **Staged buffers**



Driving Large Loads with Staged Buffers

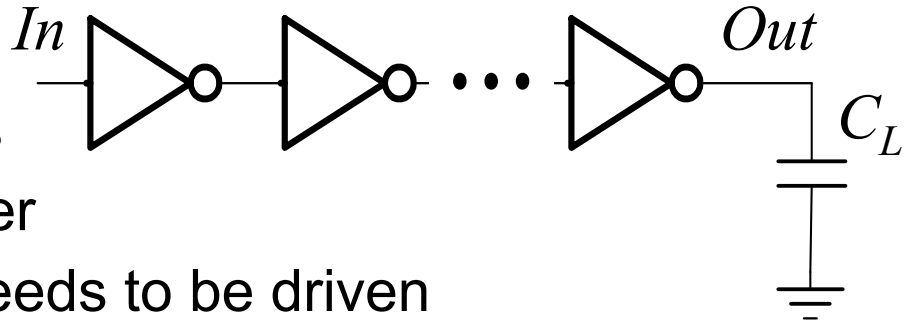
- Should be obvious that total delay is minimized with equal delay at each stage.



- For some given C_L:
  - How many stages are needed to minimize delay?
  - How to size the inverters?

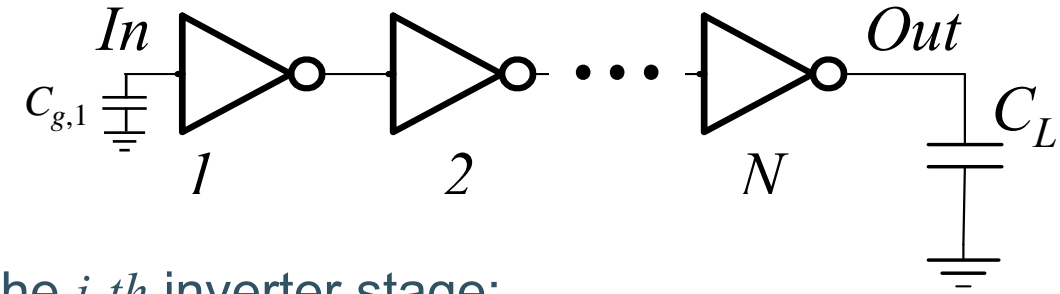
Delay Optimization

- 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:

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

Optimum Delay

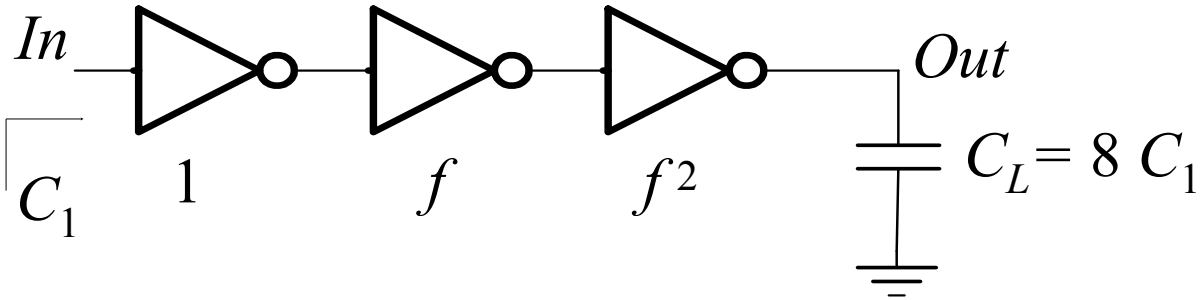
- Each inverter should be sized up by the same factor f with respect to the preceding inverter
- This way each stage has equal delay
- Given $C_{g,1}$ and C_L

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

- 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:

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

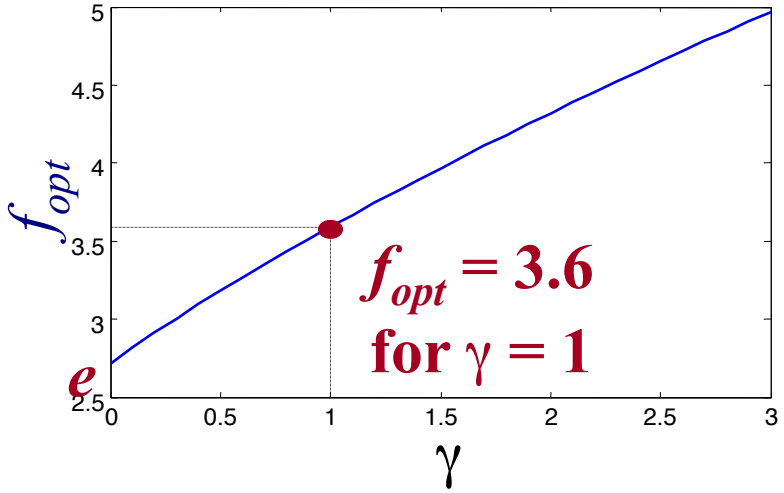
$$\Rightarrow \quad \boxed{f = e^{(1+\gamma/f)}} \qquad f = \sqrt[N]{F}$$

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

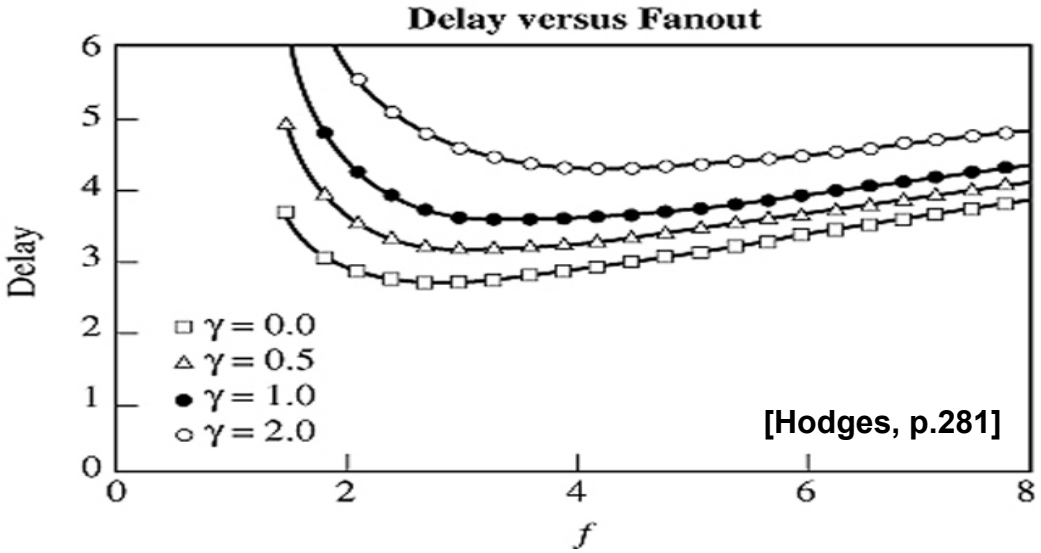
Optimum Effective Fanout f

- Optimum f for given process depends on γ

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



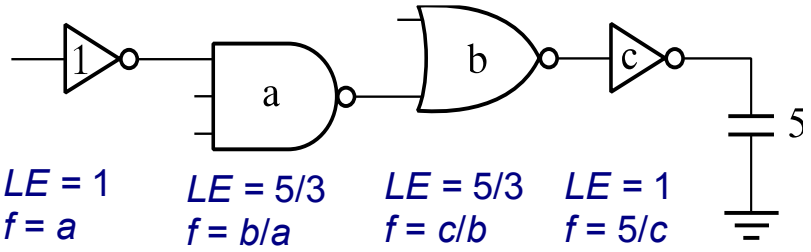
In Practice: Plot of Total Delay

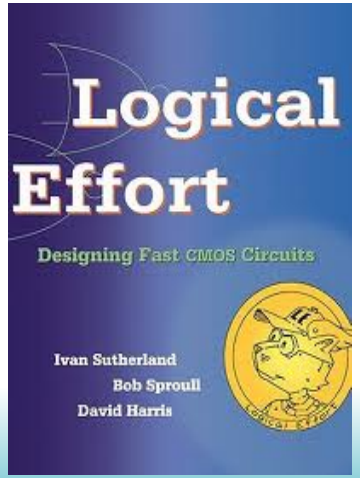


- Why the shape?
- Curves very flat for $f > 2$
  - Simplest/most common choice: $f = 4$

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