CAL POLY
SAN LUIS OBISPO

EE 431: COMPUTER-AIDED DESIGN OF VLSI DEVICES

MOSFET Parasitics & Delay Estimation

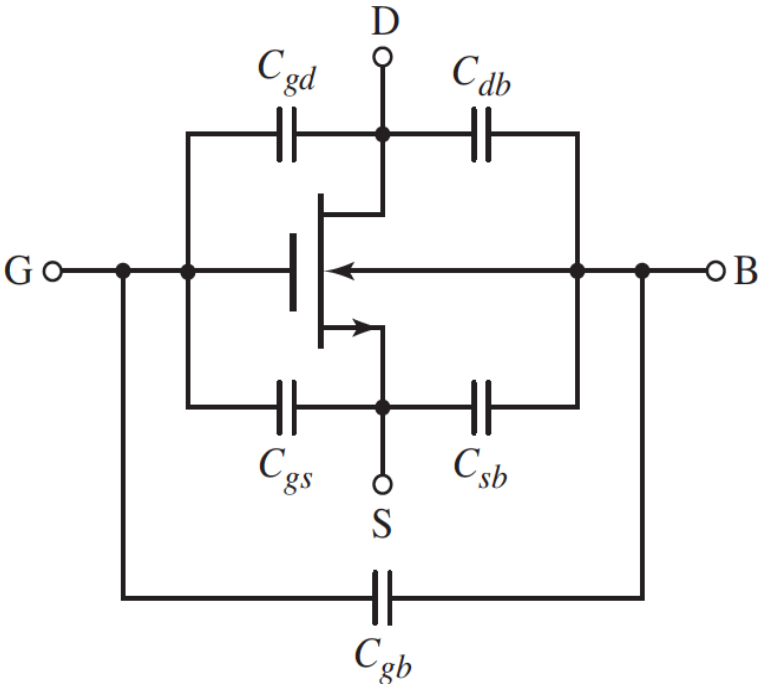
Nishith N. Chakraborty

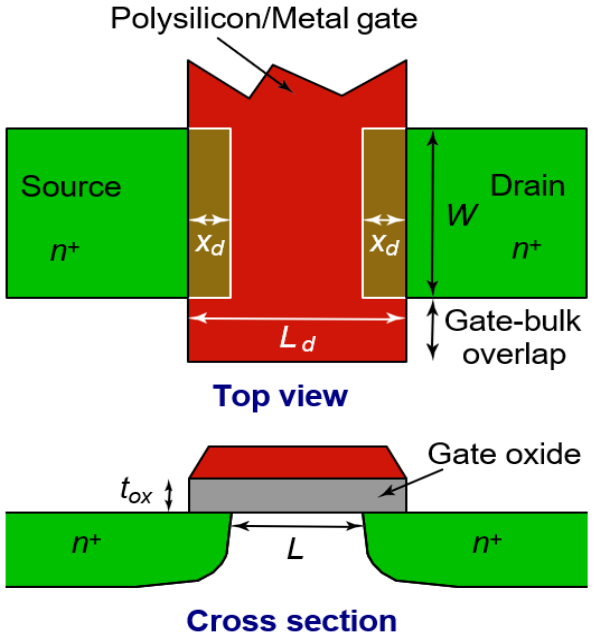
October, 2024

MOSFET CAPACITANCE

- Any two conductors separated by an insulator have capacitance
- Gate to channel capacitor is very important
  - Creates channel charge necessary for operation
- Source and drain have capacitance to body
  - Across reverse-biased diodes
  - Called diffusion capacitance because it is associated with source/drain diffusion

CAPACITANCE COMPONENTS

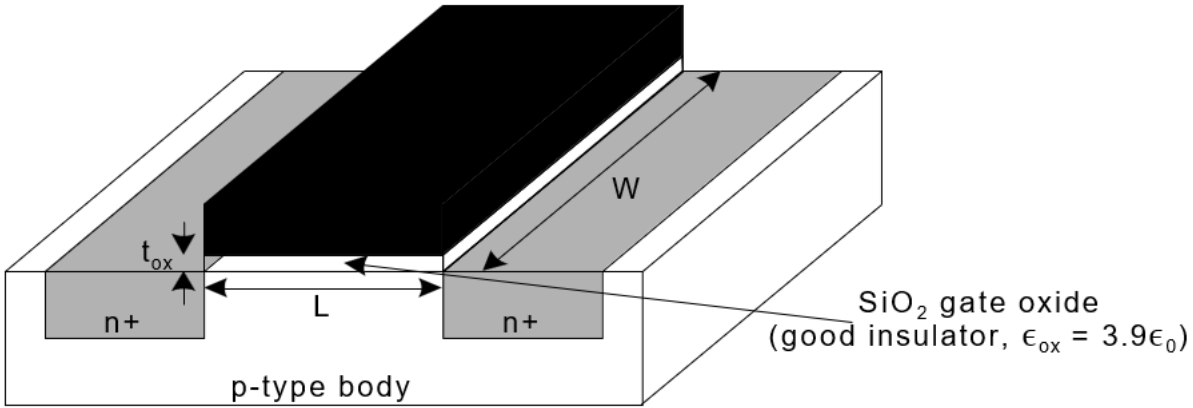




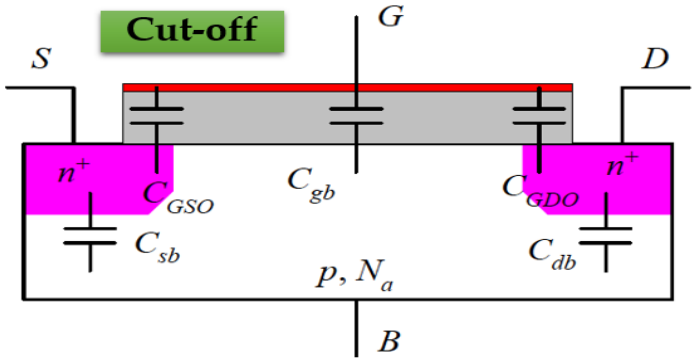
$$C_{gate} = \frac{\varepsilon_{ox}}{t_{ox}} WL$$

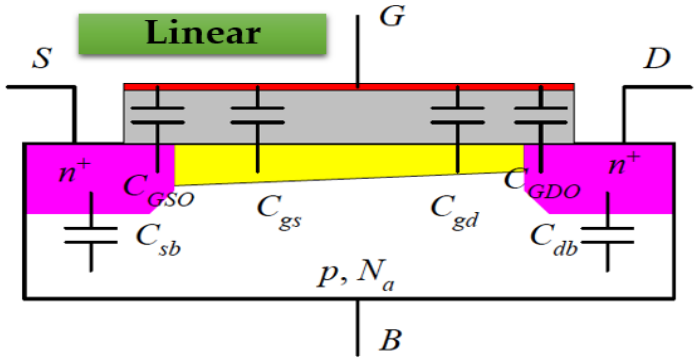
GATE CAPACITANCE

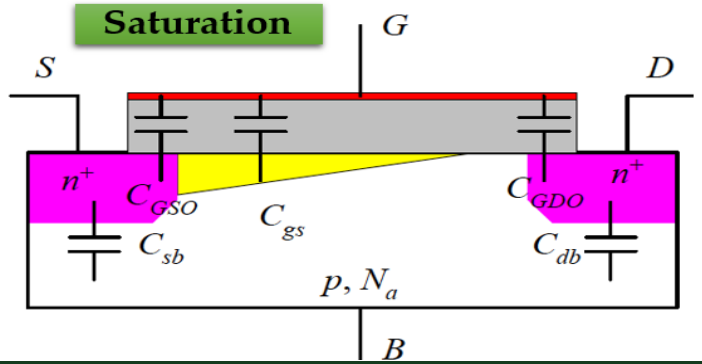
- Approximate channel as connected to source
- $C_g = \varepsilon_{ox}WL/t_{ox} = C_{ox}WL = C_{permicron}W$
- $C_{permicron}$ is typically about 2 fF/µm



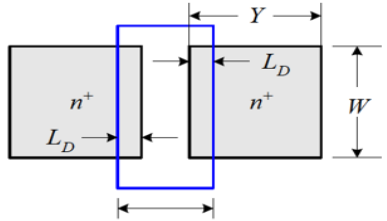
GATE CAPACITANCE







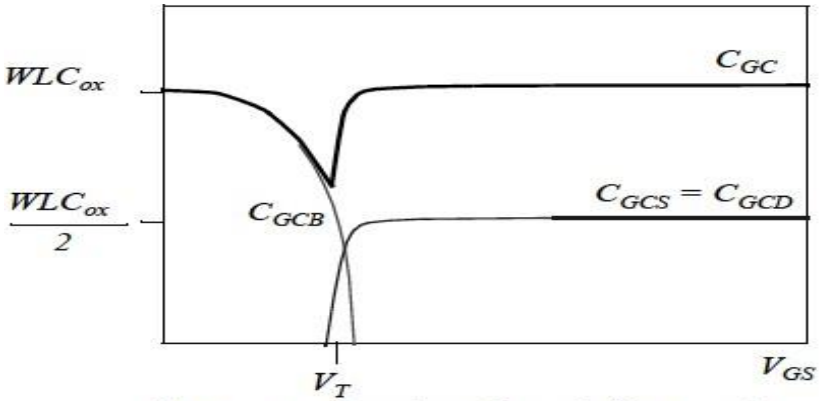
Capacitance	Cutoff	Linear	Saturation
$C_{gb(total)}$	$C_{ox}WL_{actual}$	0	0
$C_{gs(total)}$	C_{GSO}	$\frac{1}{2}C_{ox}WL_{actual} + C_{GSO}$	$\frac{2}{3}C_{ox}WL_{actual} + C_{GSO}$
$C_{gd(total)}$	C_{GDO}	$\frac{1}{2}C_{ox}WL_{actual} + C_{GDO}$	C_{GDO}



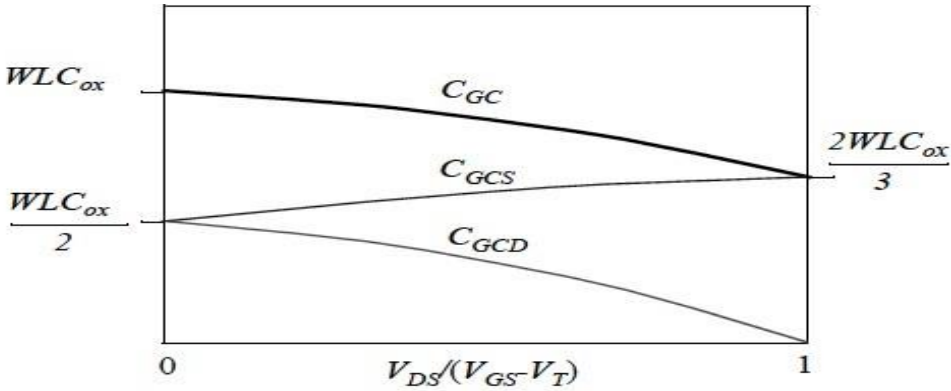
$$L_{actual} = L - 2L_D$$

Most important regions in digital design:
Saturation and Cut-off

GATE CAPACITANCE



(a) C_{GC} as a function of V_{GS} (with V_{DS}=0)

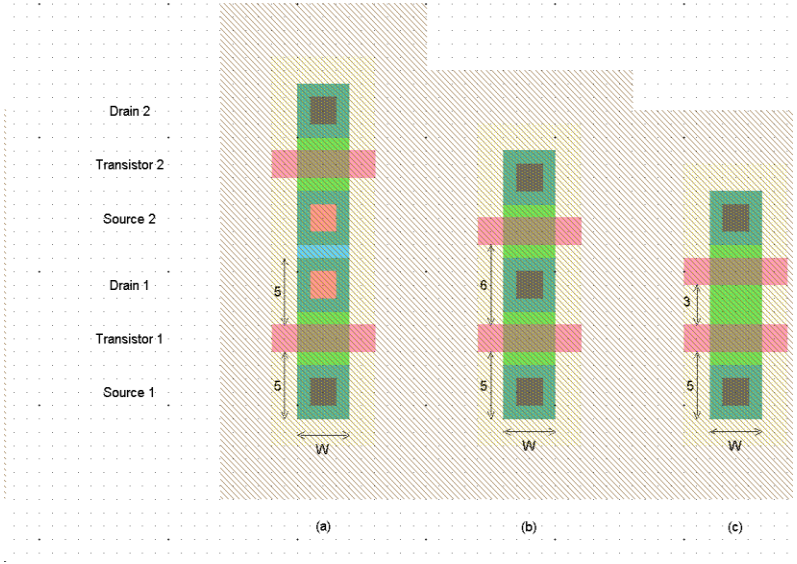


(b) C_{GC} as a function of the degree of saturation

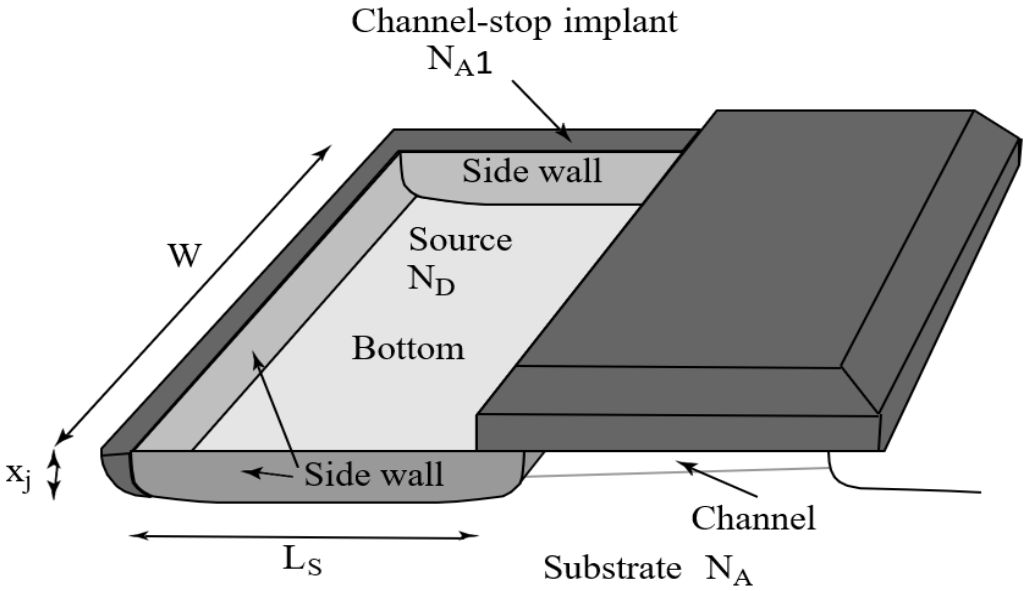
Operation Region	C_{GCB}	C_{GCS}	C_{GCD}	C_{GC}	C_G
Cutoff	$\leq C_{ox}WL$	0	0	$\leq C_{ox}WL$	$\leq C_{ox}WL+2C_oW$
Resistive	0	$C_{ox}WL / 2$	$C_{ox}WL / 2$	$C_{ox}WL$	$C_{ox}WL+2C_oW$
Saturation	0	$(2/3)C_{ox}WL$	0	$(2/3)C_{ox}WL$	$(2/3)C_{ox}WL+2C_oW$

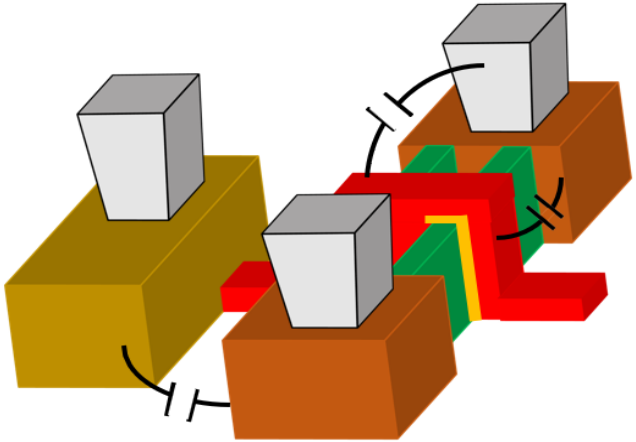
DIFFUSION CAPACITANCE

- C_{sb}, C_{db}
- Undesirable, called parasitic capacitance
- Capacitance depends on area and perimeter
  - Use small diffusion nodes
  - Comparable to C_g for contacted diffusion
  - ½ C_g for uncontacted
  - Varies with process



DIFFUSION CAPACITANCE

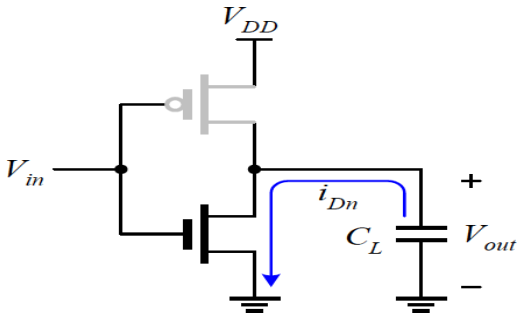


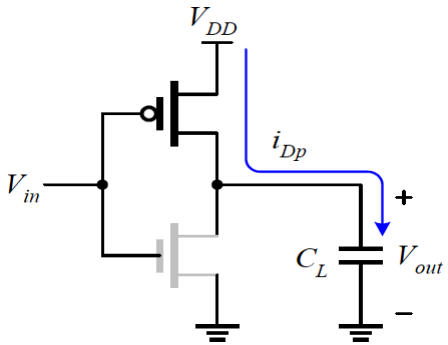


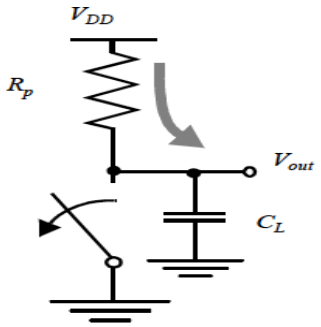
$$C_{diff} = C_{bottom} + C_{sw} = C_j \times AREA + C_{jsw} \times PERIMETER$$
$$= C_j L_S W + C_{jsw}(2L_S + W)$$

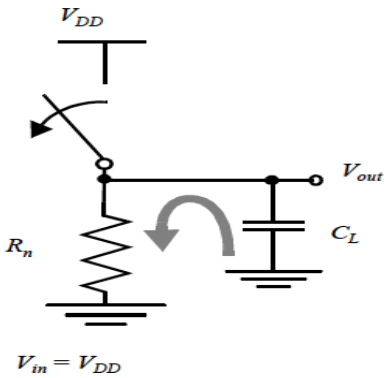
Higher capacitance in FinFETs compared to bulk MOSFETs (due to 3D structure)

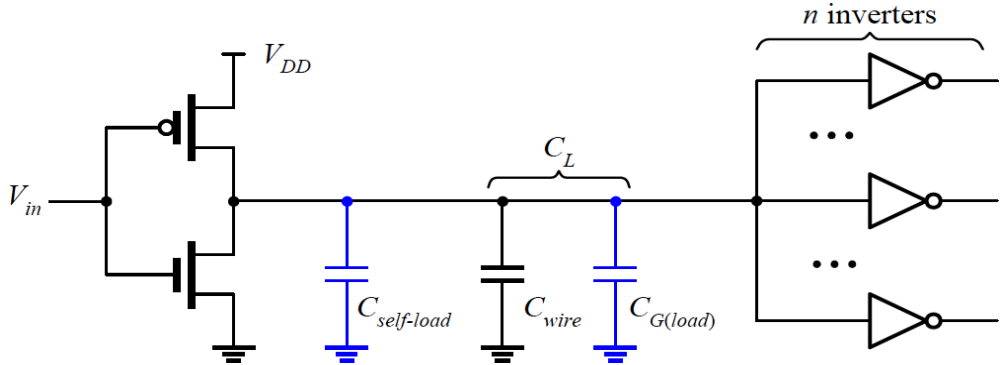
RC EQUIVALENT DELAY MODEL

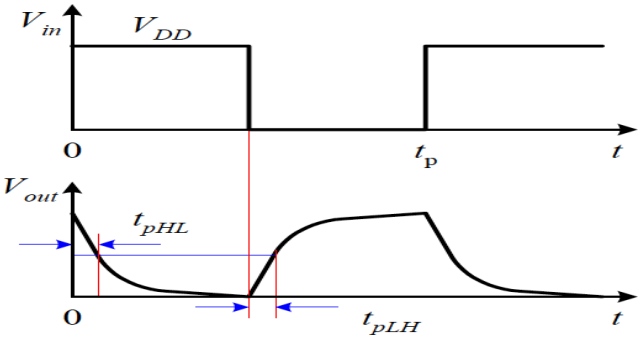




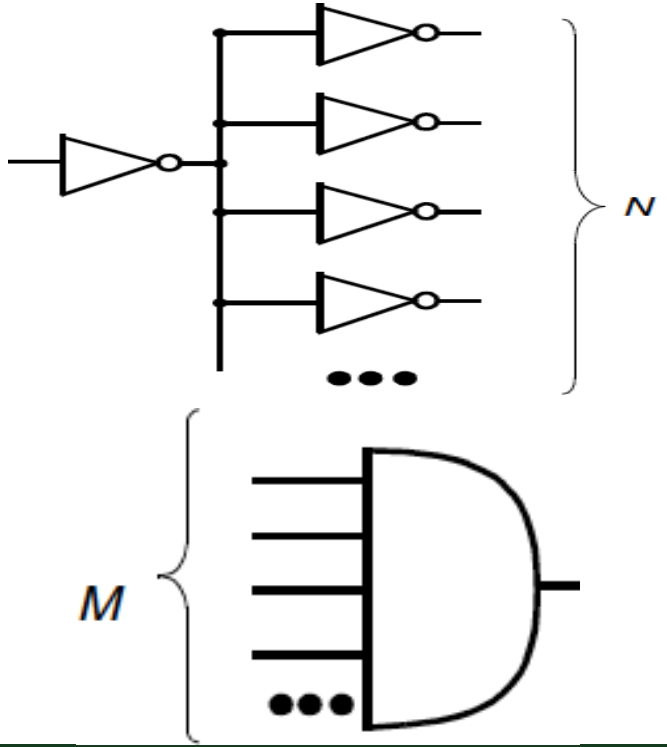


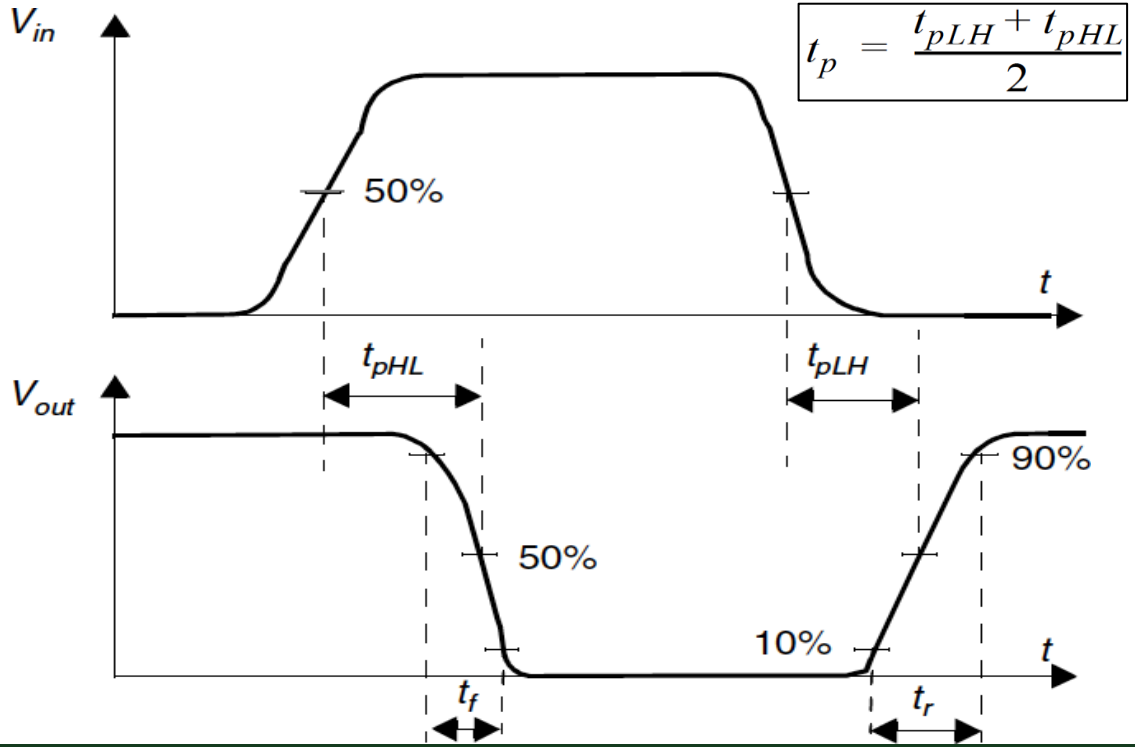






FAN-OUT, FAN-IN AND DELAY



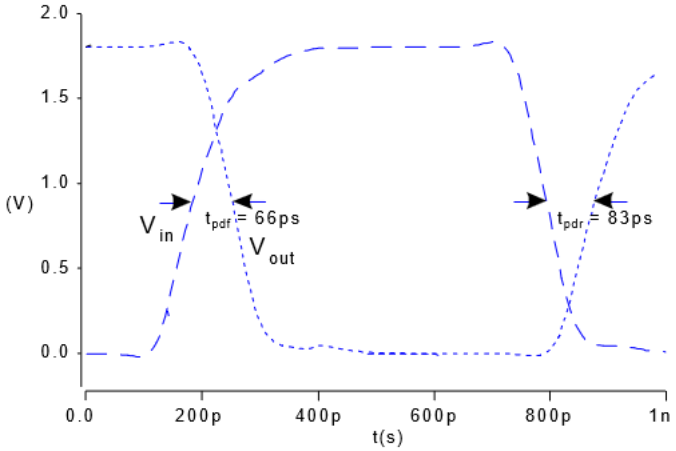


DELAY DEFINITIONS

- t_{pdr}: rising propagation delay
  - From input to rising output crossing $V_{DD}/2$
- t_{pdf}: falling propagation delay
  - From input to falling output crossing $V_{DD}/2$
- t_{pd}: average propagation delay
  - $t_{pd} = (t_{pdr} + t_{pdf})/2$
- t_r: rise time
  - From output crossing 0.1 V_{DD} to 0.9 V_{DD}
- t_f: fall time
  - From output crossing 0.9 V_{DD} to 0.1 V_{DD}

SIMULATED INVERTER DELAY

- Solving differential equations by hand is too hard
- SPICE simulator solves the equations numerically
  - Uses more accurate I-V models too! But simulations take time to write

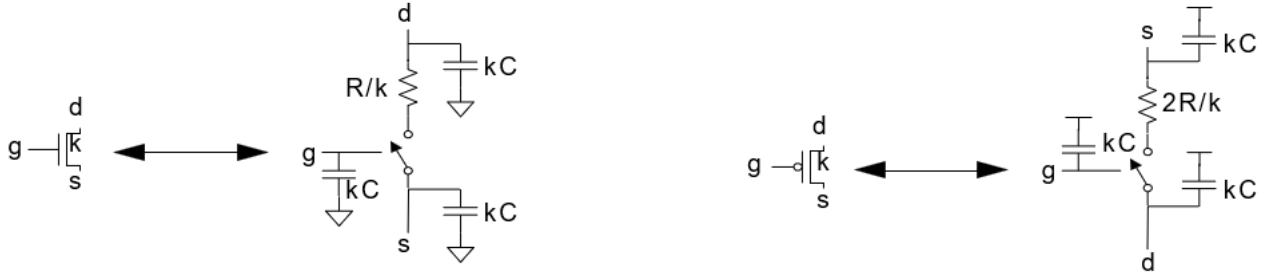


DELAY ESTIMATION

- We would like to be able to easily estimate delay
  - Not as accurate as simulation
  - But easier to ask “What if?”
- The step response usually looks like a 1st order RC response with a decaying exponential.
- Use RC delay models to estimate delay
  - C = total capacitance on output node
  - Use effective resistance R
  - So that t_{pd} = RC ln 2, lets just replace R ln 2 with R, so t_{pd} = RC
- Characterize transistors by finding their effective R
  - Depends on average current as gate switches

RC DELAY MODEL

- Use equivalent circuits for MOS transistors
  - Ideal switch + capacitance and ON resistance
  - Unit sized NMOS has resistance R, capacitance C
  - Unit sized PMOS has resistance 2R, capacitance C
- Capacitance proportional to width
- Resistance inversely proportional to width

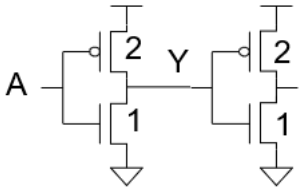


RC VALUES

- Capacitance
  - $C = C_g = C_s = C_d = 2$ fF/µm of gate width
  - Values similar across many processes
- Resistance
  - R ~ 6 KΩ*µm in 0.6um process
  - Improves with shorter channel lengths
- Unit transistors
  - May refer to minimum contacted device
  - Or maybe 1 µm wide device
  - Doesn't matter as long as you are consistent

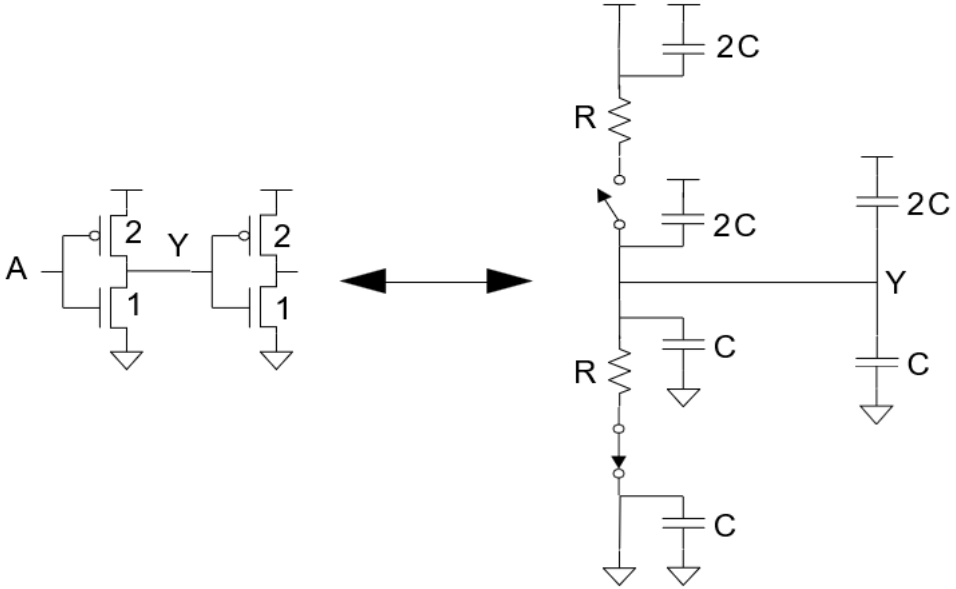
INVERTER DELAY ESTIMATION

- Estimate the delay of a “fanout-of-1” inverter



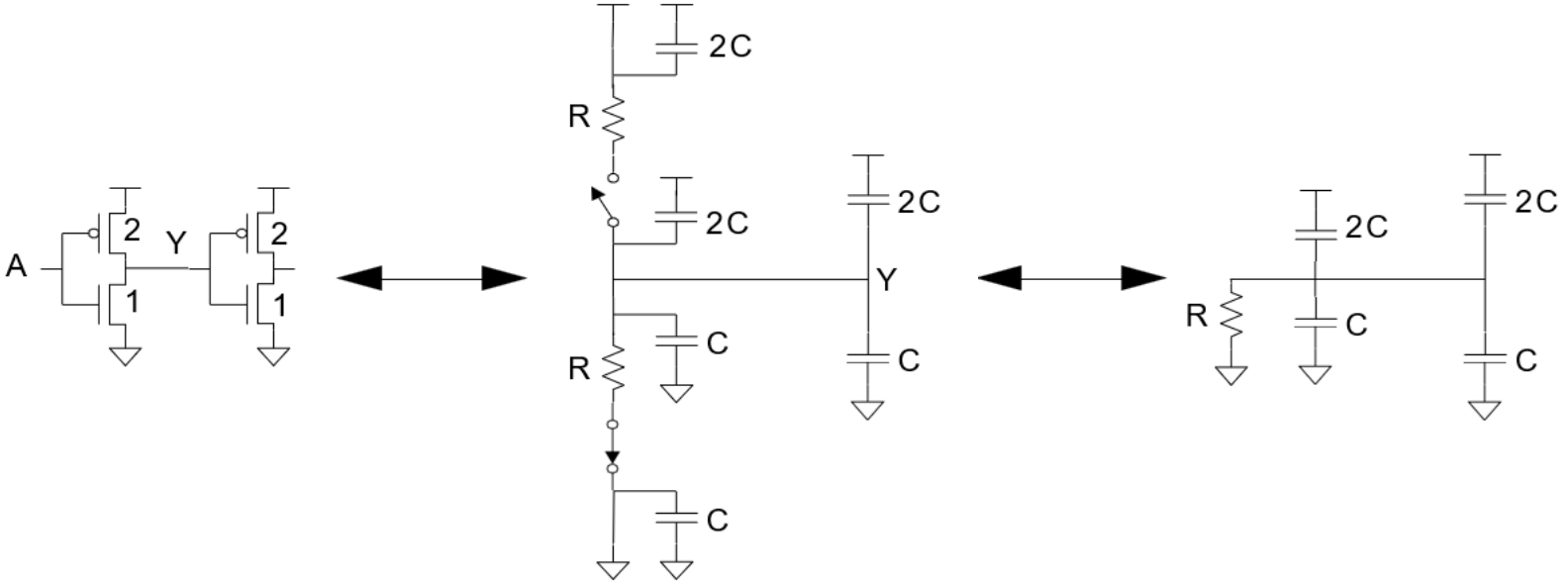
INVERTER DELAY ESTIMATION

- Estimate the delay of a “fanout-of-1” inverter



INVERTER DELAY ESTIMATION

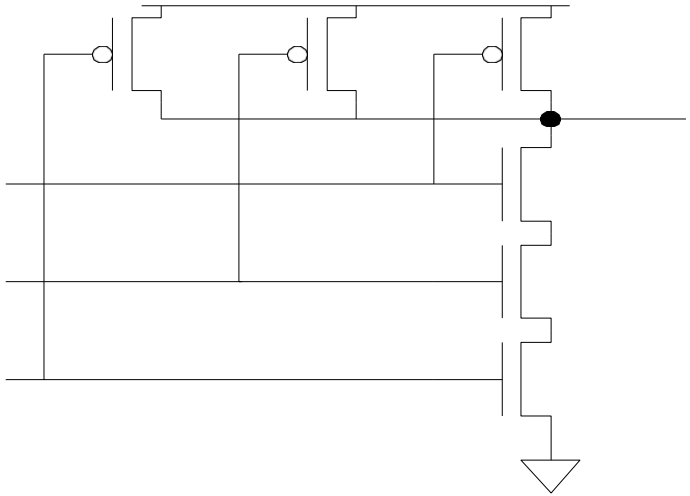
- Estimate the delay of a “fanout-of-1” inverter



d = 6RC

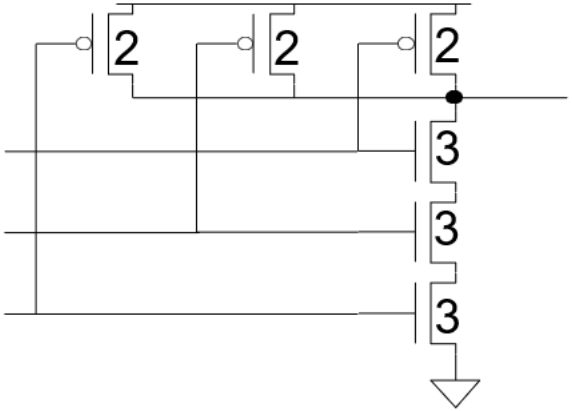
EXAMPLE: 3-INPUT NAND GATE

- Sketch a 3-input NAND with transistor widths chosen to achieve effective rise and fall resistances equal to a unit inverter (R).



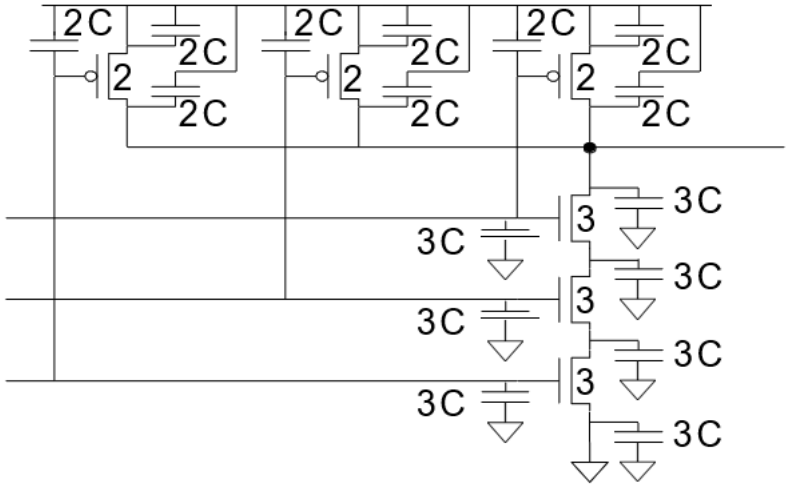
EXAMPLE: 3-INPUT NAND GATE

- Sketch a 3-input NAND with transistor widths chosen to achieve effective rise and fall resistances equal to a unit inverter (R).



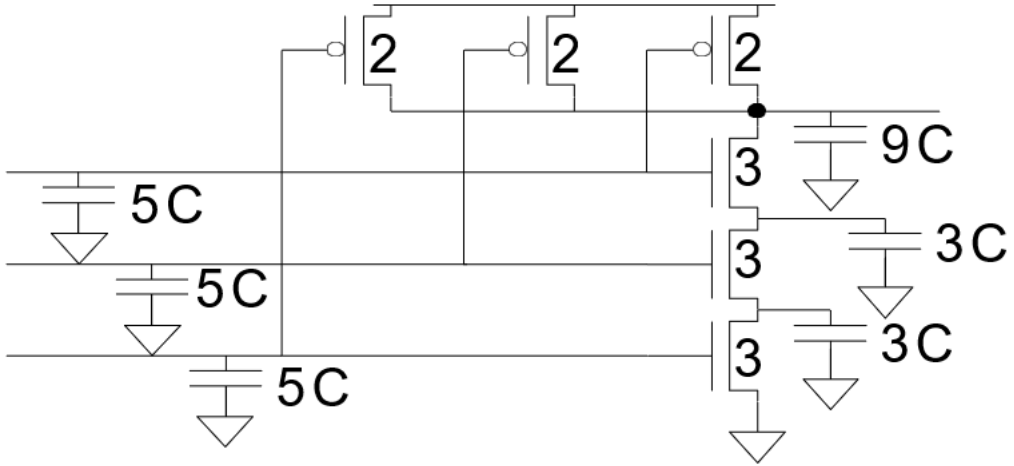
EXAMPLE: 3-INPUT NAND GATE

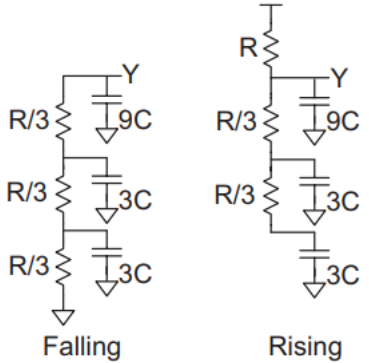
- Annotate the 3-input NAND gate with gate and diffusion capacitance.



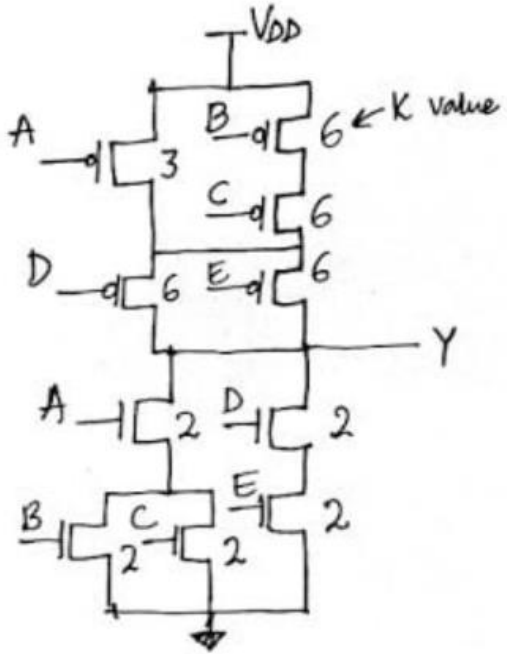
EXAMPLE: 3-INPUT NAND GATE

- Annotate the 3-input NAND gate with gate and diffusion capacitance. Show the rise and fall paths.





EXAMPLE: $Y=\overline{A(B+C) + DE}$



PMOS sizing: For a unit PMOS transistor, the effective resistance with the width k is given by 2R/k.
By looking at the pull-up network in the above circuit, we should find out the worst-case or the longest path to VDD. In the above network, the path E-C-B is the longest path. So we can write the equation (2R/k)+(2R/k)+(2R/k) = R, where R is the effective resistance. The equation gives the value of k = 6. Therefore the k value transistors E, C, and B will be 6.

One more path D-C-B also contributes to the worst-case or longest path, So the k value of the transistor D also becomes 6. The transistor A is equivalent to two transistors B and C (by looking at the circuit). Therefore we can write 2R/k = 2 * 2R/6 Since we know the k values of B and C transistors. So k for A is 3.

NMOS sizing:

For a unit NMOS transistor, the effective resistance with the width k is given by R/k.

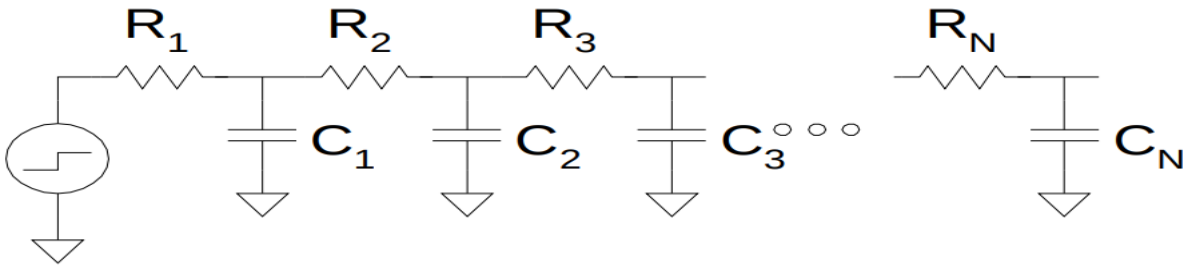
In the above network, the worst-case or the longest path can be seen is with two transistors. (The paths A-B, A-C, and D-E). So we can write the relation 2 * R/k = R, So the value of k of all the NMOS transistors will be 2 since all are in the longest path.

ELMORE DELAY

- ON transistors look like resistors
- Pullup or pulldown network modeled as RC ladder
- Elmore delay of RC ladder

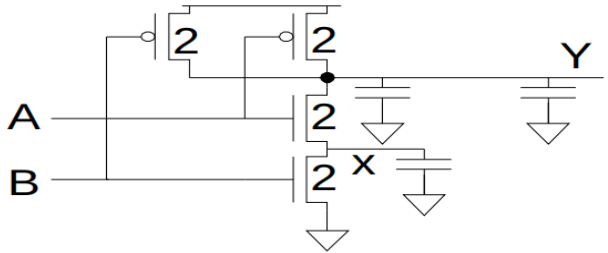
$$t_{pd} \approx \sum_{\text{nodes } i} R_{i\text{- to- source}} C_i$$

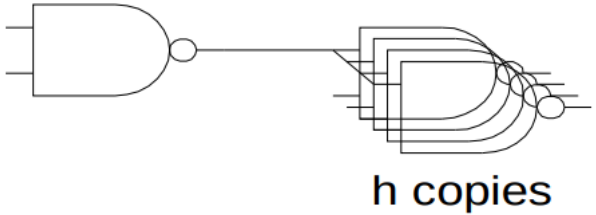
$$= R_1 C_1 + (R_1 + R_2) C_2 + \ldots + (R_1 + R_2 + \ldots + R_N) C_N$$



EXAMPLE: 2-INPUT NAND GATE

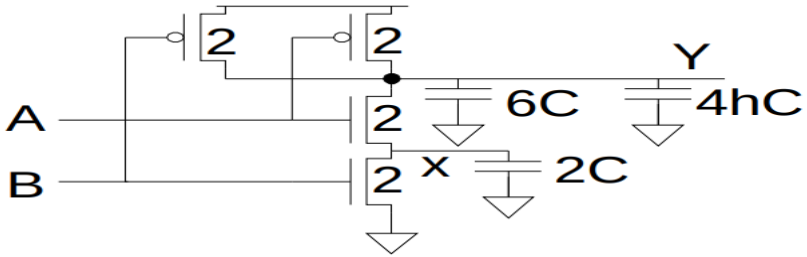
- Estimate worst-case rising and falling delay of 2-input NAND driving *h* identical gates.

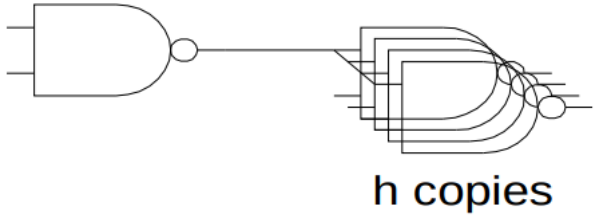




EXAMPLE: 2-INPUT NAND GATE

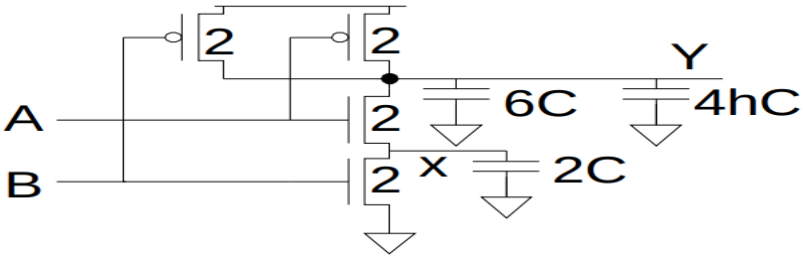
- Estimate worst-case rising and falling delay of 2-input NAND driving *h* identical gates.

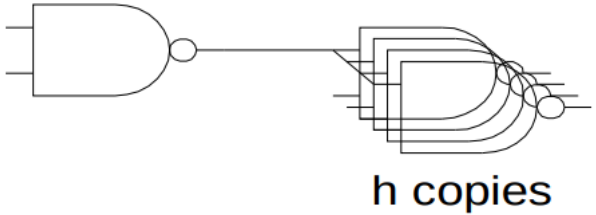


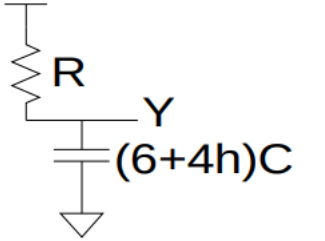


EXAMPLE: 2-INPUT NAND GATE

- Estimate worst-case **rising** and falling delay of 2-input NAND driving *h* identical gates.



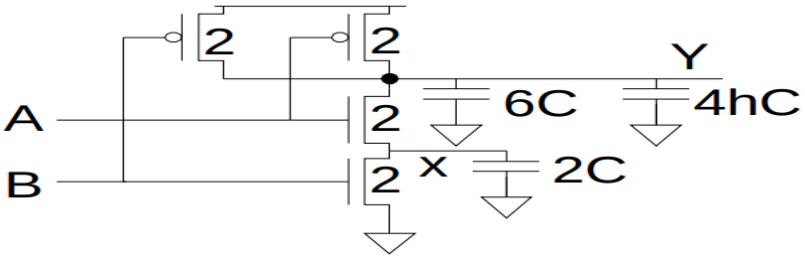


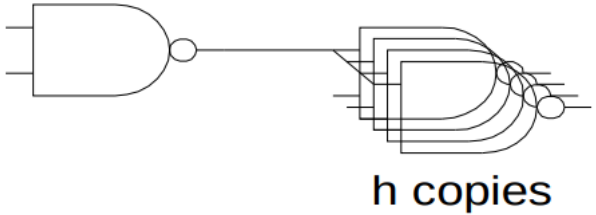


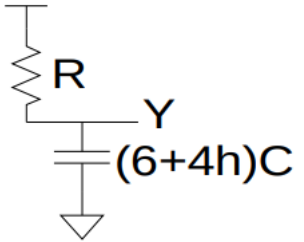
$$t_{pdr} =$$

EXAMPLE: 2-INPUT NAND GATE

- Estimate worst-case **rising** and falling delay of 2-input NAND driving *h* identical gates.



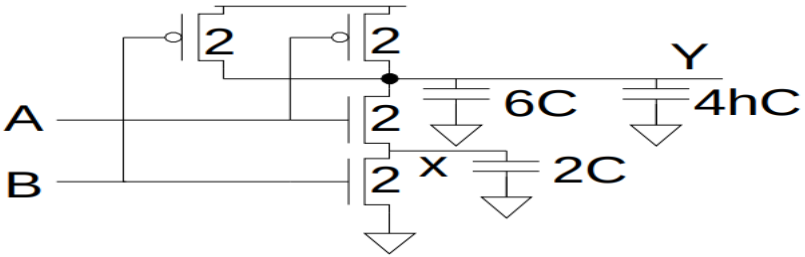


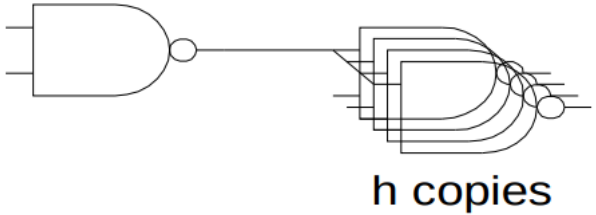


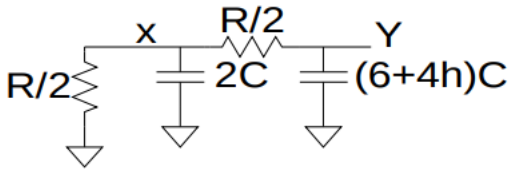
$$t_{pdr} = (6 + 4h)\,RC$$

EXAMPLE: 2-INPUT NAND GATE

- Estimate worst-case rising and **falling** delay of 2-input NAND driving *h* identical gates.



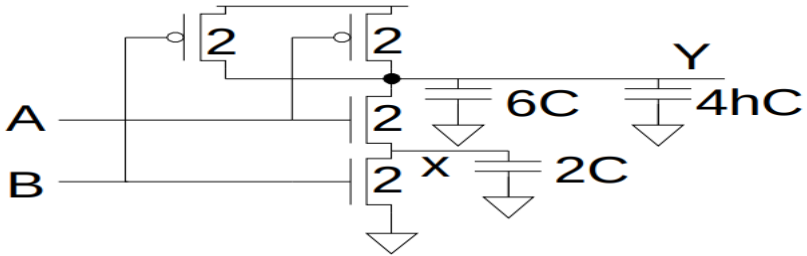


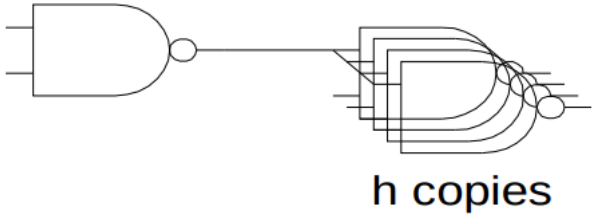


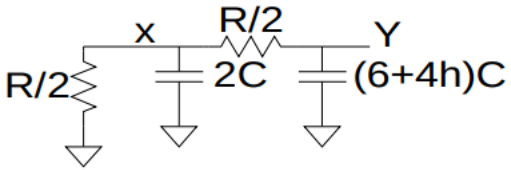
$$t_{pdf} =$$

EXAMPLE: 2-INPUT NAND GATE

- Estimate worst-case rising and **falling** delay of 2-input NAND driving *h* identical gates.







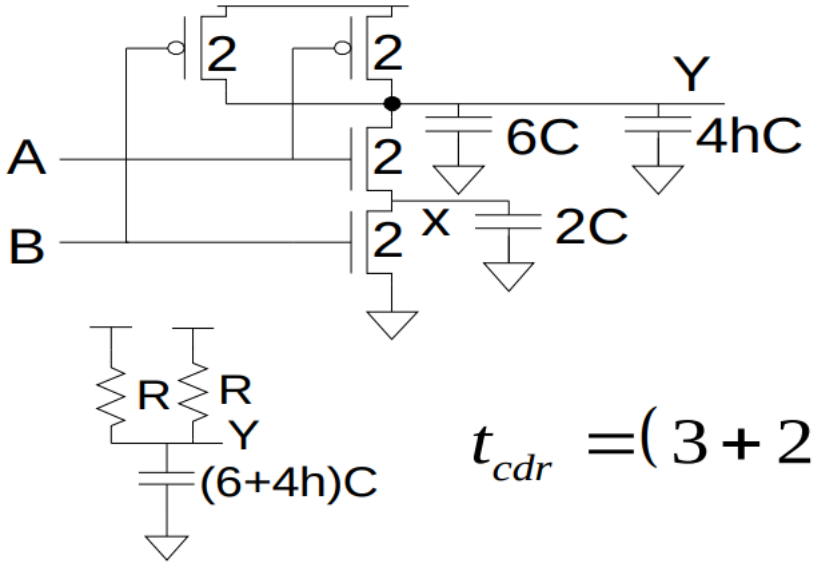
$$t_{pdf} = (2C)\left(\frac{R}{2}\right) + \left|(6+4h)C\right|\left(\frac{R}{2}+\frac{R}{2}\right)$$
$$= (7+4h)RC$$

DELAY COMPONENTS

- Delay has two parts
  - Parasitic delay
    - 6 or 7 RC
    - Independent of load
  - Effort delay
    - 4h RC
    - Proportional to load capacitance

CONTAMINATION DELAY

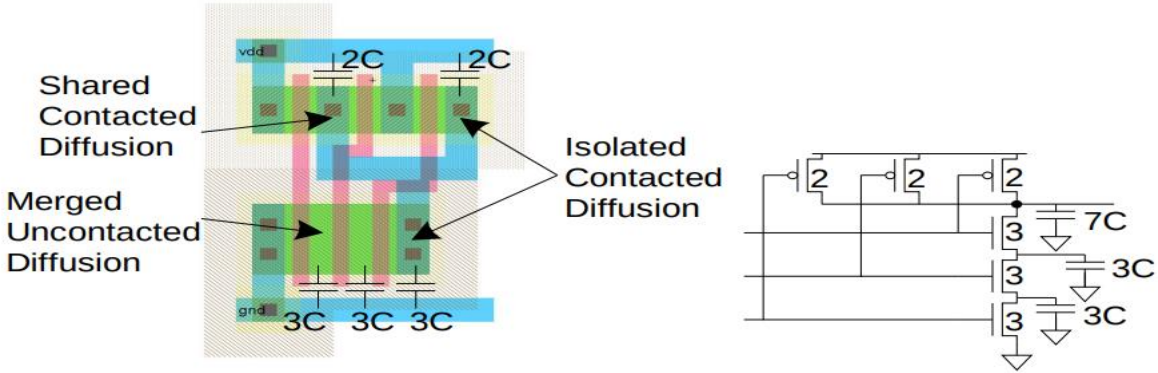
- Best-case (contamination) delay can be substantially less than propagation delay.
- Ex: If both inputs fall simultaneously



$$t_{cdr} = (3 + 2h)\,RC$$

DIFFUSION CAPACITANCE

- We assumed contacted diffusion on every s / d.
- Good layout minimizes diffusion area
- Ex: NAND3 layout shares one diffusion contact
  - Reduces output capacitance by 2C
  - Merged uncontacted diffusion might help too



Thank you!