Performance of CMOS Circuits

Instructed by Shmuel Wimer
Eng. School, Bar-Ilan University

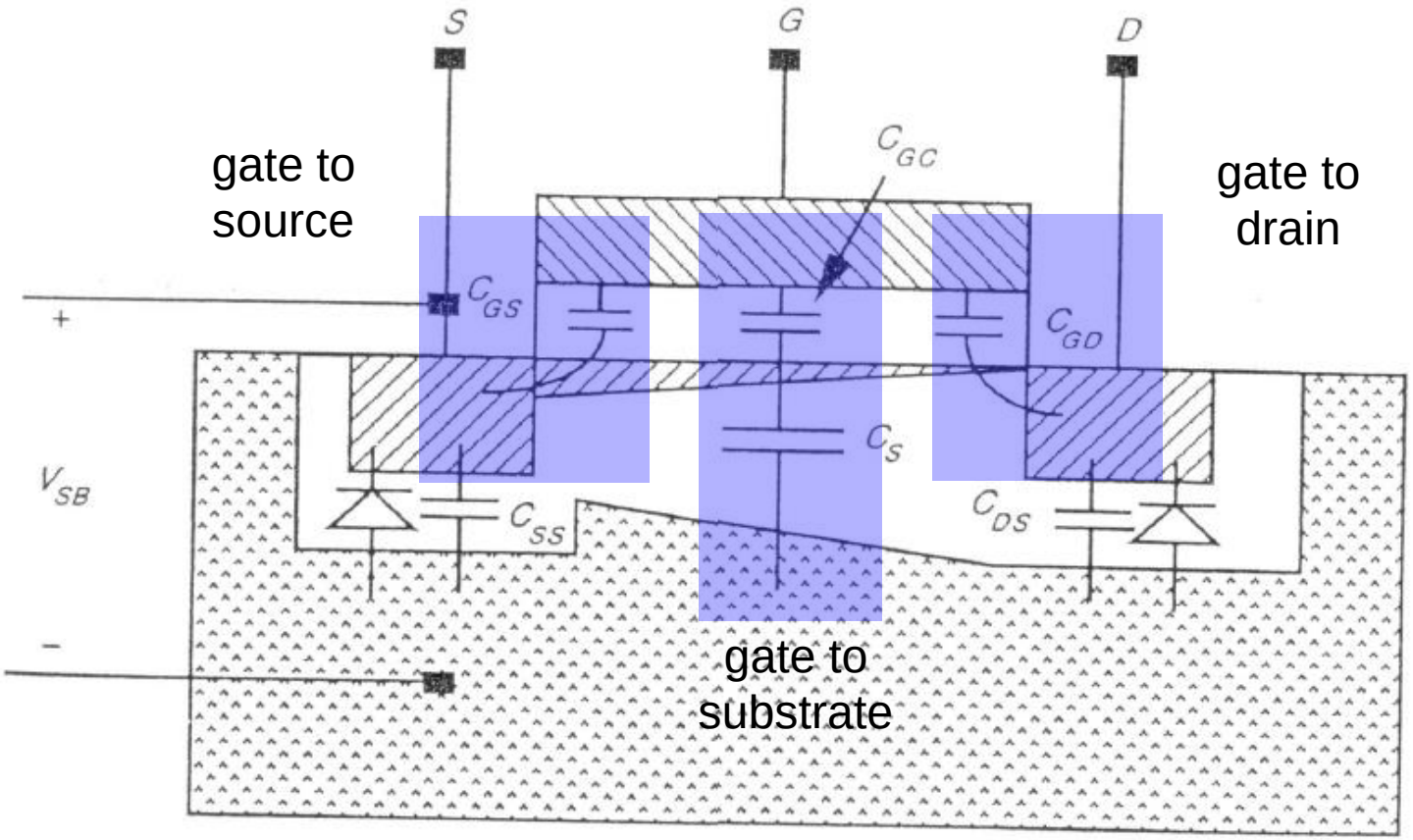
Credits: David Harris
Harvey Mudd College

(Some material copied/taken/adapted from Harris' lecture notes)

Outline

- Gate and Diffusion Capacitance
- RC Delay Models
- Power and Energy
- Dynamic Power
- Static Power
- Low Power Design

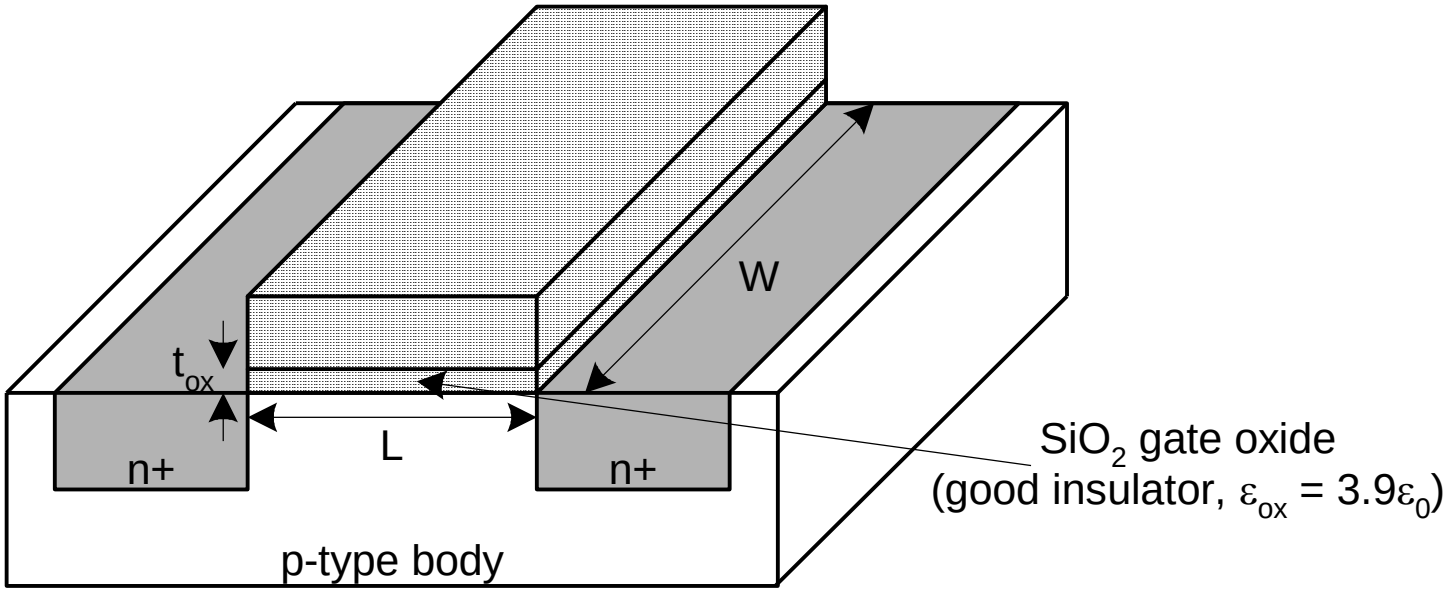
MOSFET Capacitance

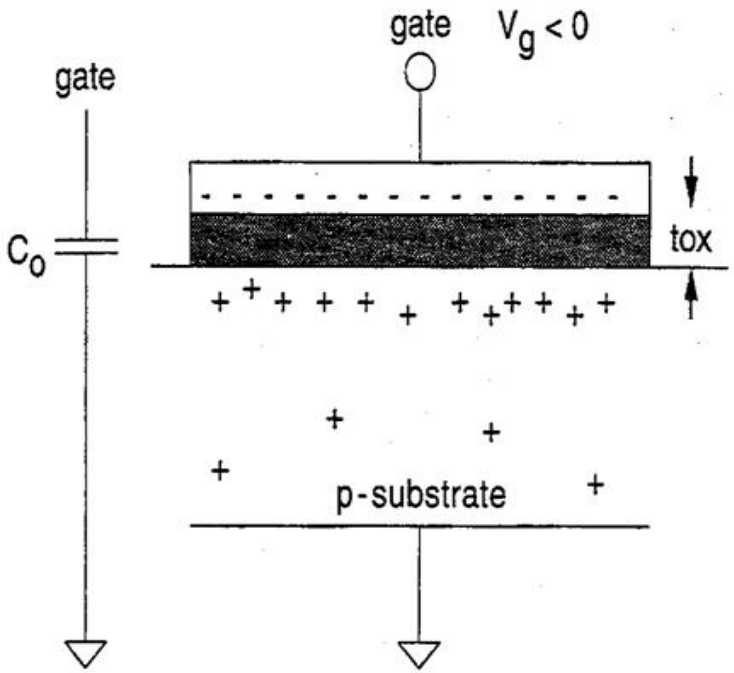


- Any two conductors separated by an insulator have capacitance
- Gate to channel capacitor is very important
  - Creates channel charge necessary for operation
- Source and drain have capacitance to body
  - Across reverse-biased diodes
  - Called diffusion capacitance because it is associated with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

Gate Capacitance

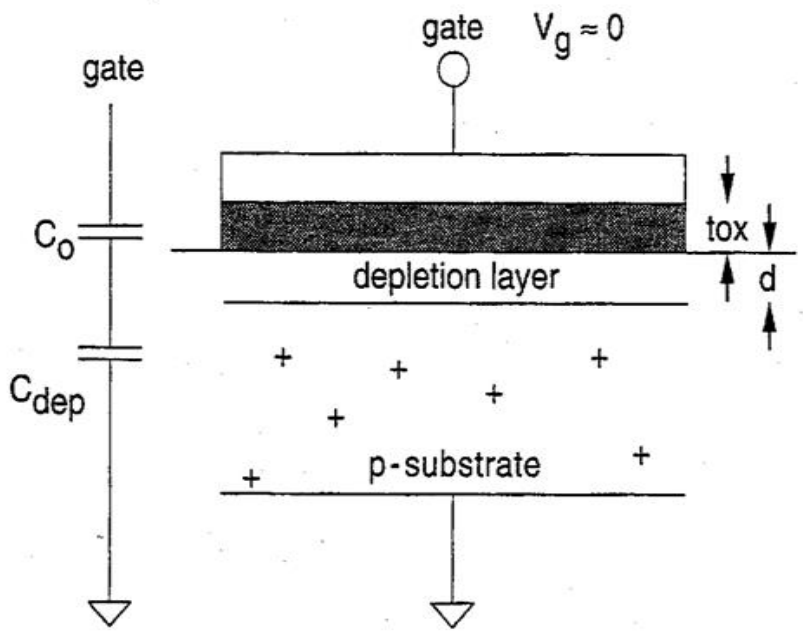
- Approximate channel as connected to source
- $C_{gs} = \varepsilon_{ox}WL/t_{ox} = C_{ox}WL = C_{permicron}W$
- $C_{permicron}$ is typically about 2 fF/μm



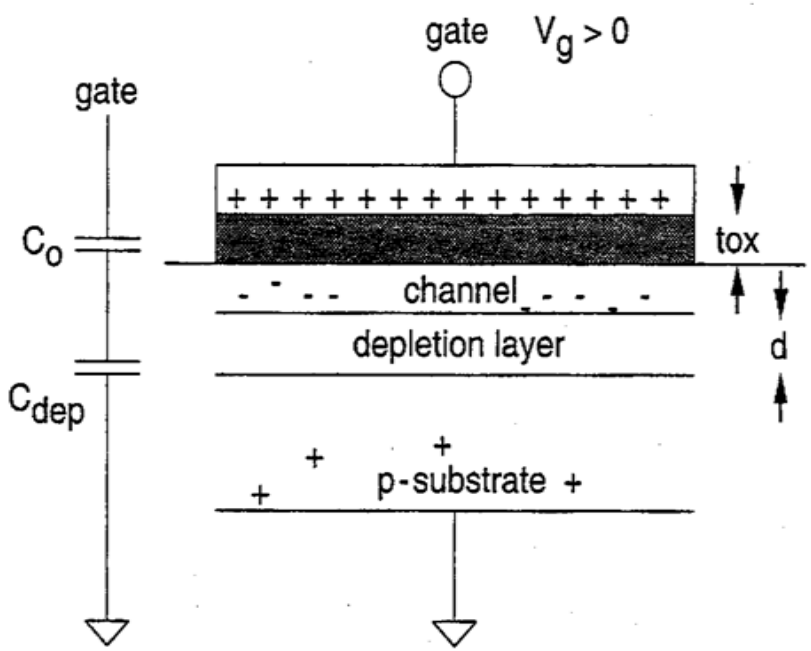


Accumulation occurs when Vg is negative (for P material). Holes are induced under the oxide.

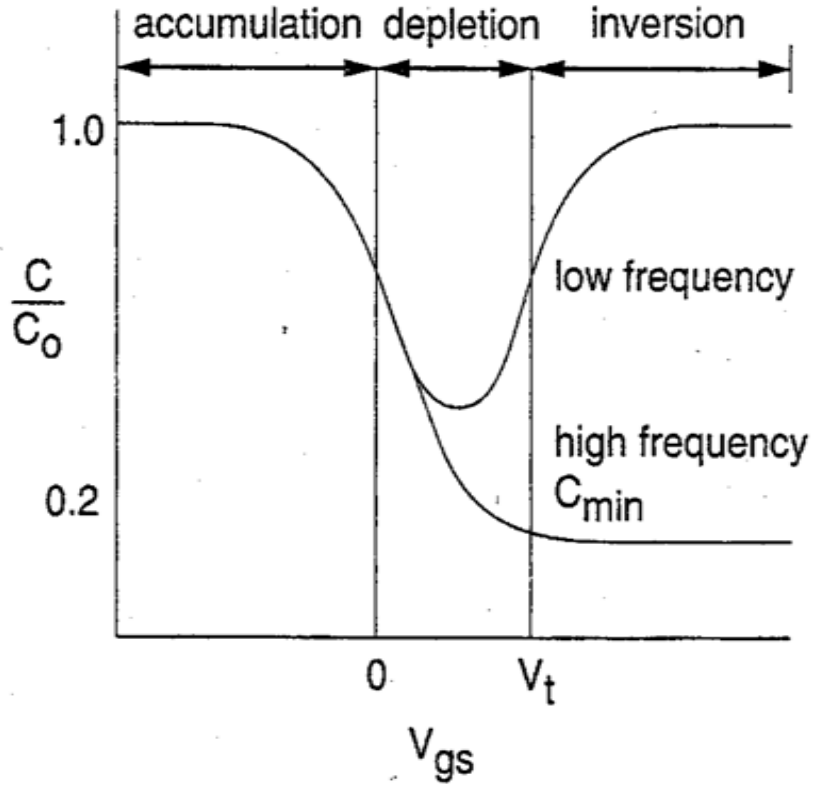
$C_{gate} = C_{ox}A$ where $C_{ox} = \varepsilon SiO2\varepsilon o/t_{ox}$



Depletion occurs when V_g is near zero but < V_{tn}. Here the C_{gate} is given by $C_{ox}A$ in series with depletion layer capacitance C_{dep}



Inversion occurs when V_g is positive and > V_{tn} (for P material). A model for inversion in comprised of C_{ox} A connecting from gate-to-channel and C_{dep} connecting from channel-to-substrate.

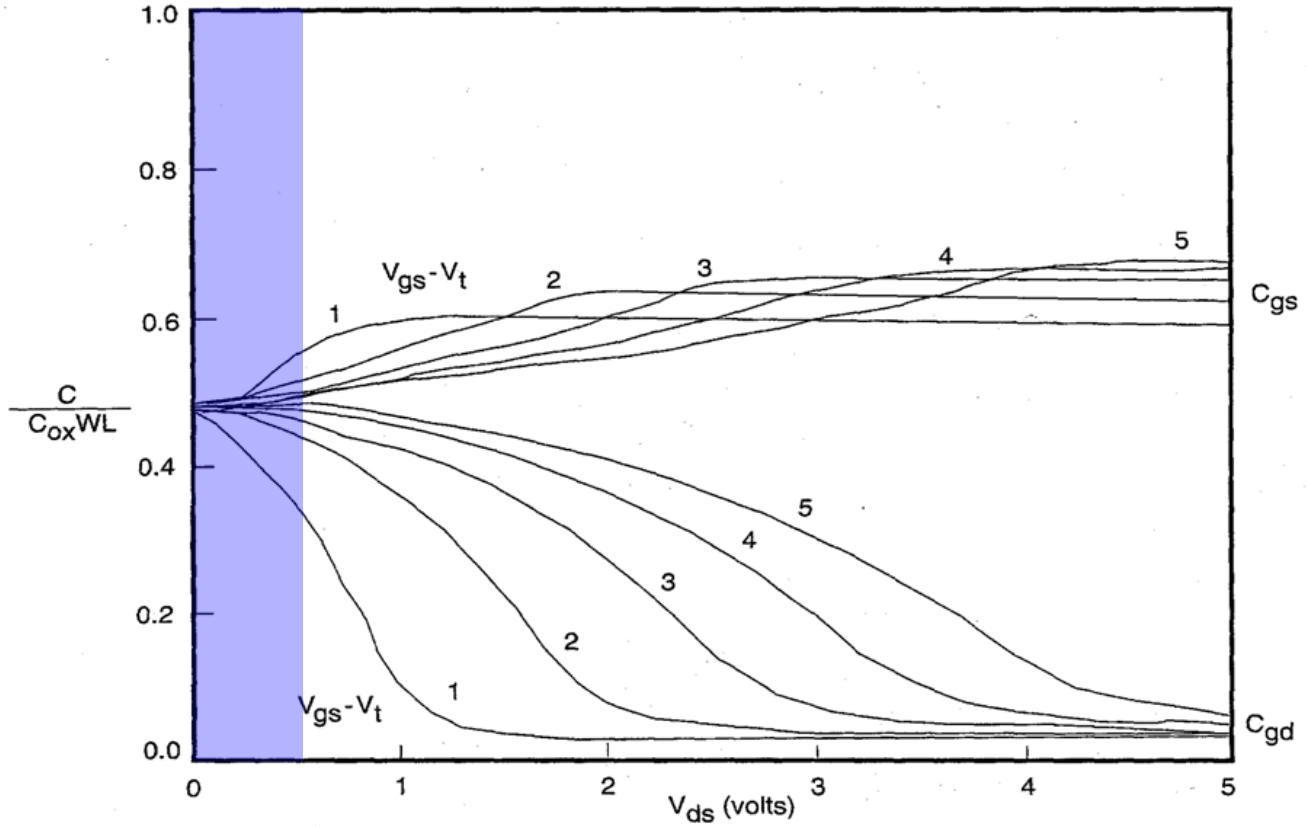


Normalized gate capacitance versus Gate voltage V_{gs}.

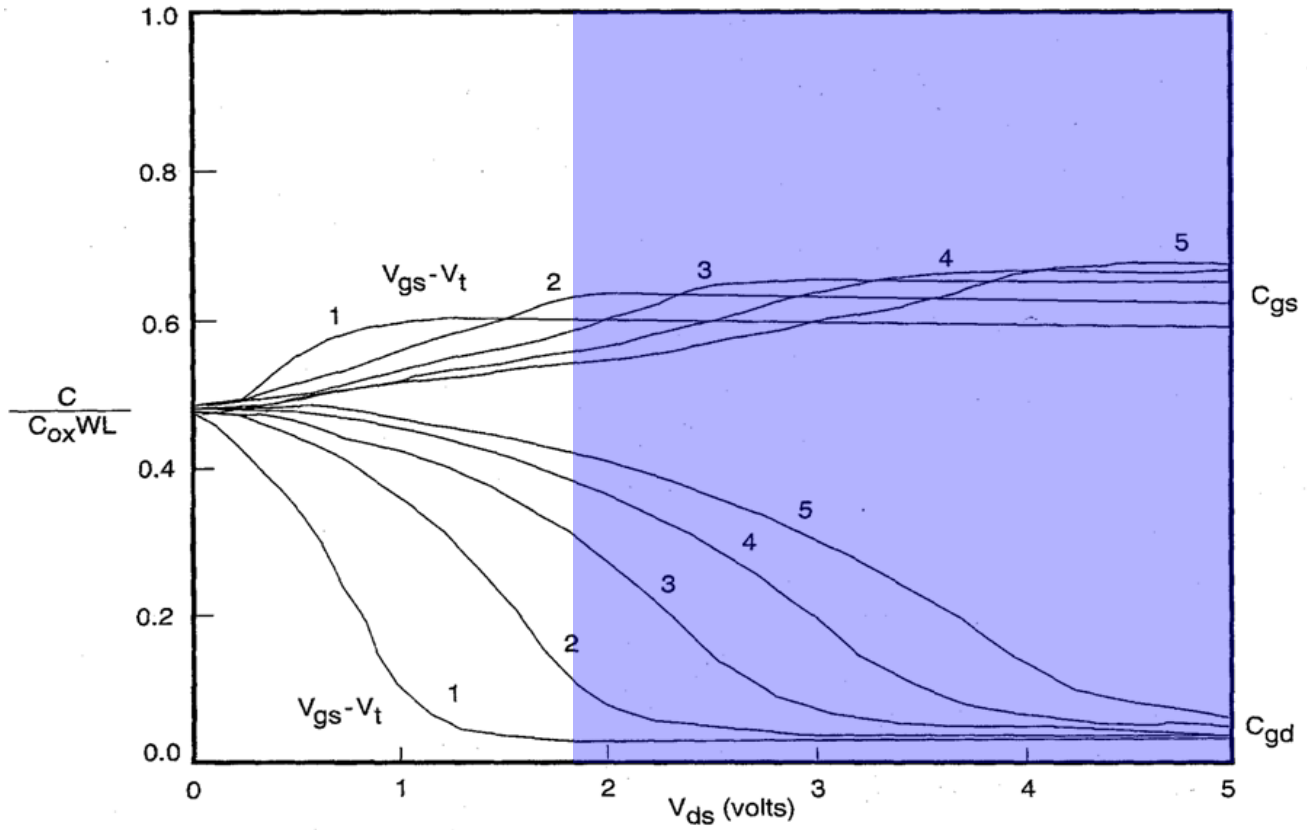
High freq behavior is due to the distributed resistance of channel

Normalized Experimental MOS Gate Capacitance Measurements vs V_{ds}, V_{gs}

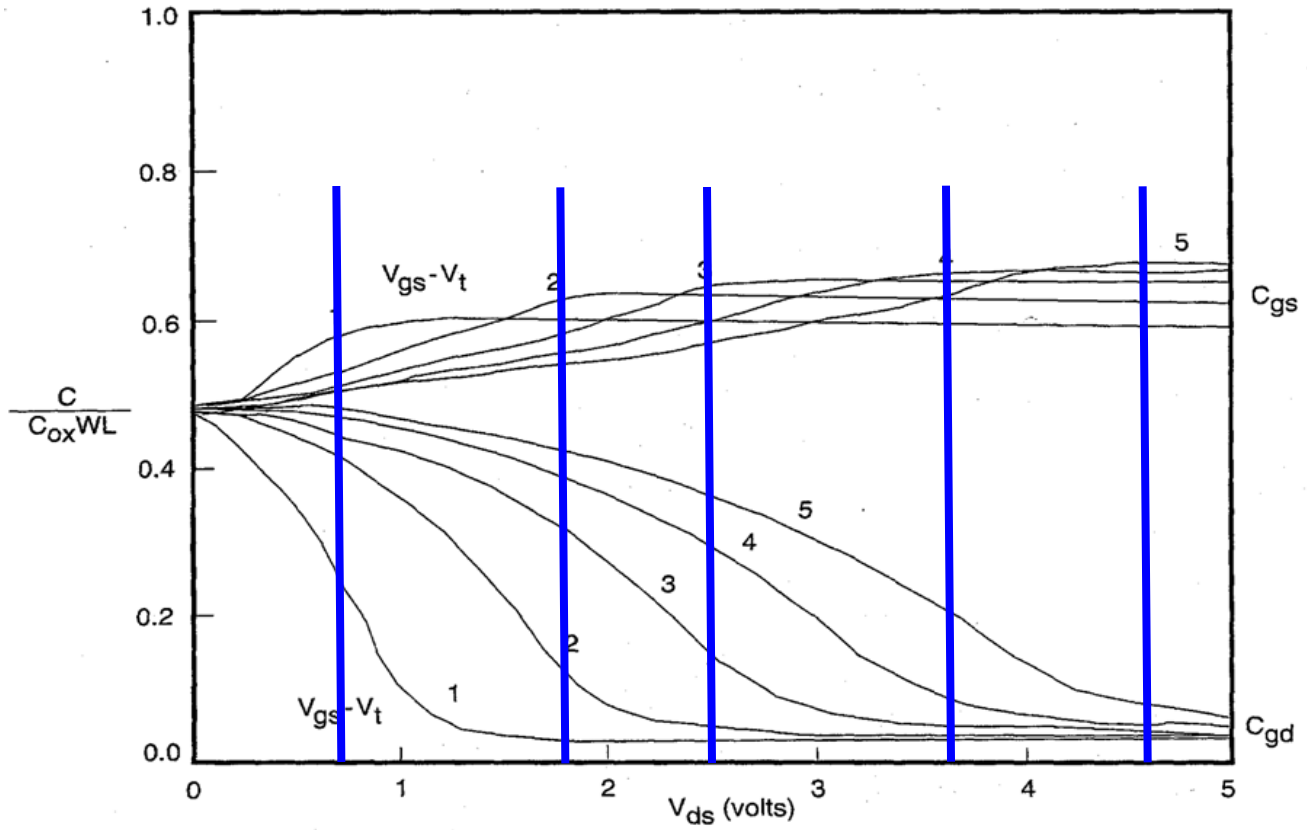
For $V_{ds} = 0$, the total gate capacitance C_{ox} A splits equally to the drain and source of the transistor.



For $V_{ds} > 0$, the gate capacitance tilts more toward the source and becomes roughly 2/3 $C_{ox}A$ to the source and 0 to the drain for high V_{ds}.

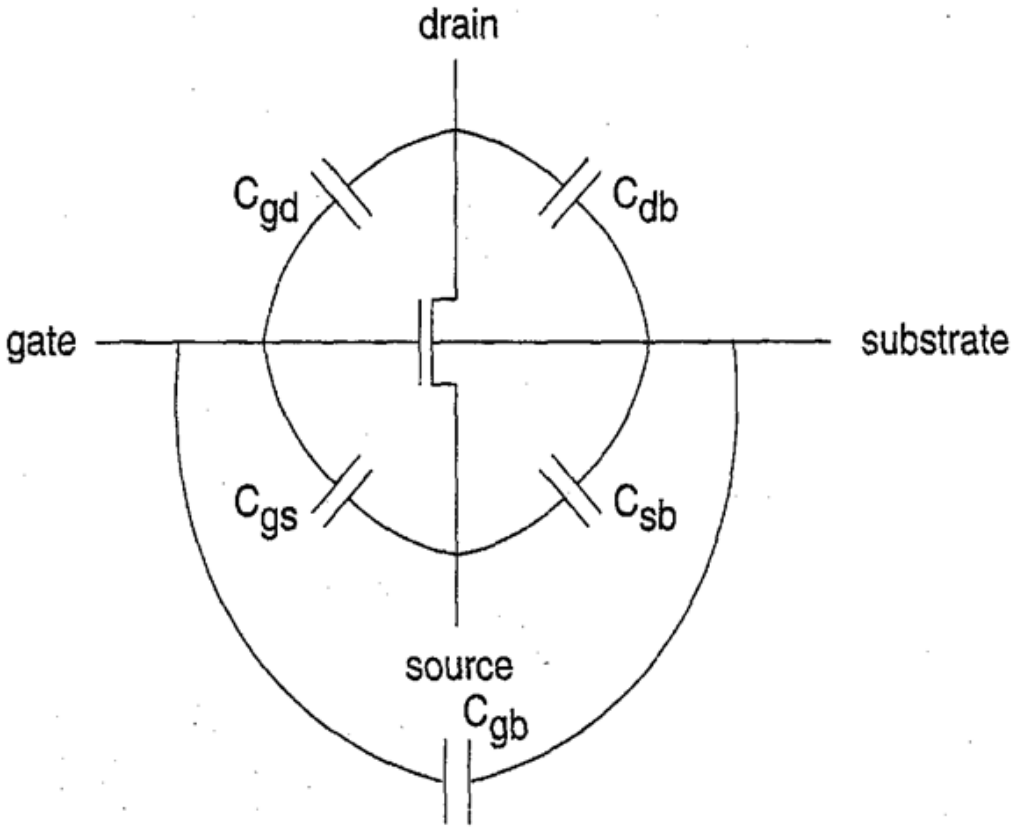


Higher $V_{gs} - V_t$ forces this tilting to occur later, since the device is linear up to $V_{gs} - V_t = V_{ds}$.



MOS Transistor Gate Capacitance Model

Gate capacitance has different components in different modes, but total remains constant.



Gate capacitance has different components in different modes, but total remains constant.

TABLE 4.3 Approximation of intrinsic MOS gate capacitance

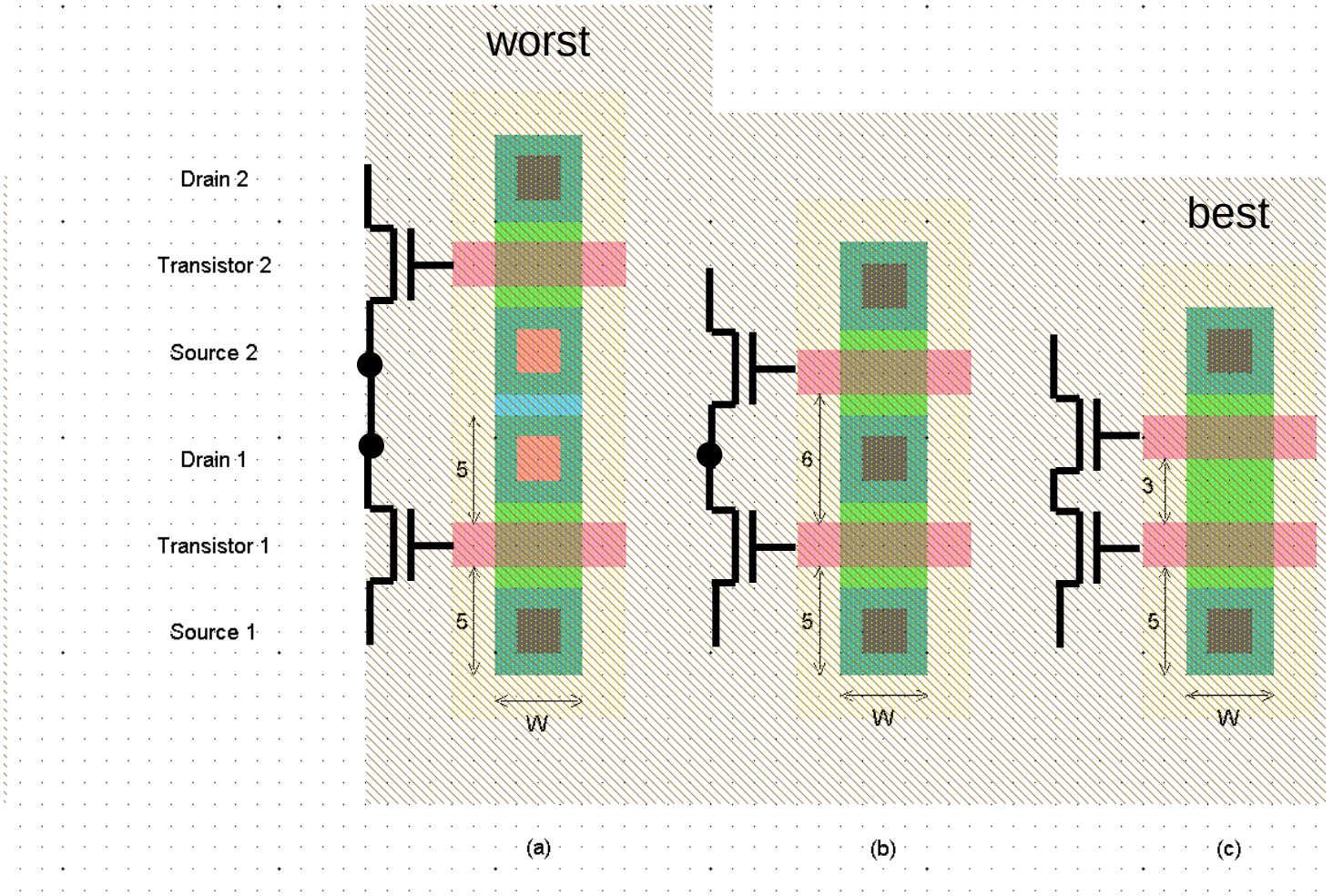
	CAPACITANCE		
Parameter	Off	Non-saturated	Saturated
C_{gb}	$\frac{\varepsilon A}{t_{ox}}$	0	0
C_{gs}	0	$\frac{\varepsilon A}{2t_{ox}}$	$\frac{2\varepsilon A}{3t_{ox}}$
C_{gd}	0	$\frac{\varepsilon A}{2t_{ox}}$	0 (finite for short channel devices)
$C_g = C_{gb} + C_{gs} + C_{gd}$	$\frac{\varepsilon A}{t_{ox}}$	$\frac{\varepsilon A}{t_{ox}}$	$\frac{2\varepsilon A}{3t_{ox}} \rightarrow \frac{.9\,\varepsilon A}{t_{ox}}$ (short channel)

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 diff
  - ½ C_g for uncontacted
  - Varies with process

Diffusion Capacitance (Cont'd)

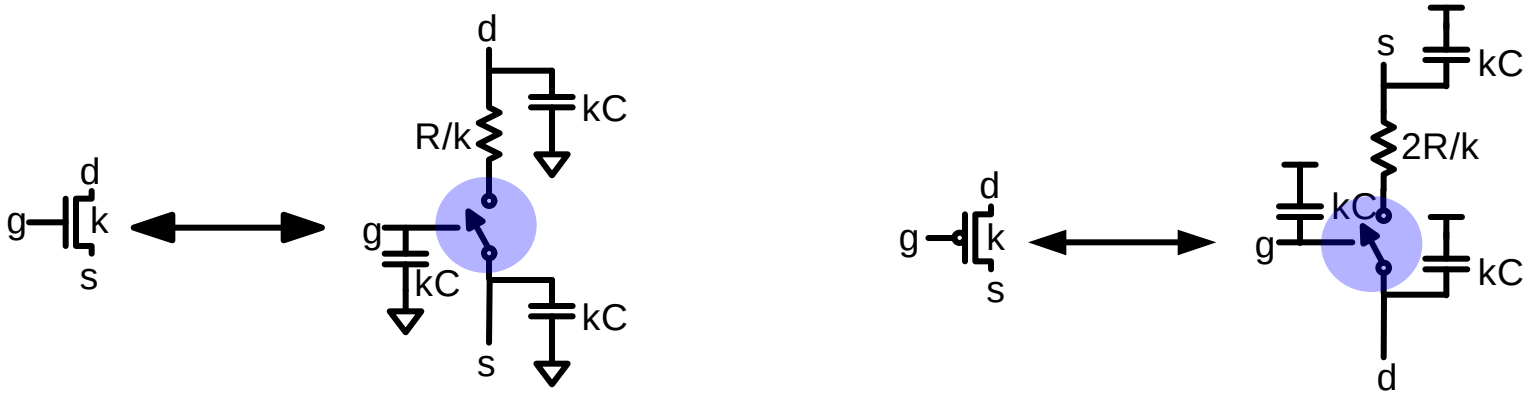


Effective Resistance

- Shockley models have limited value
  - Not accurate enough for modern transistors
  - Too complicated for much hand analysis
- Simplification: treat transistor as resistor
  - Replace $I_{ds}(V_{ds}, V_{gs})$ with effective resistance R
    - $I_{ds} = V_{ds}/R$
  - R averaged across switching of digital gate
- Too inaccurate to predict current at any given time
  - But good enough to predict RC delay

RC Delay Model

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

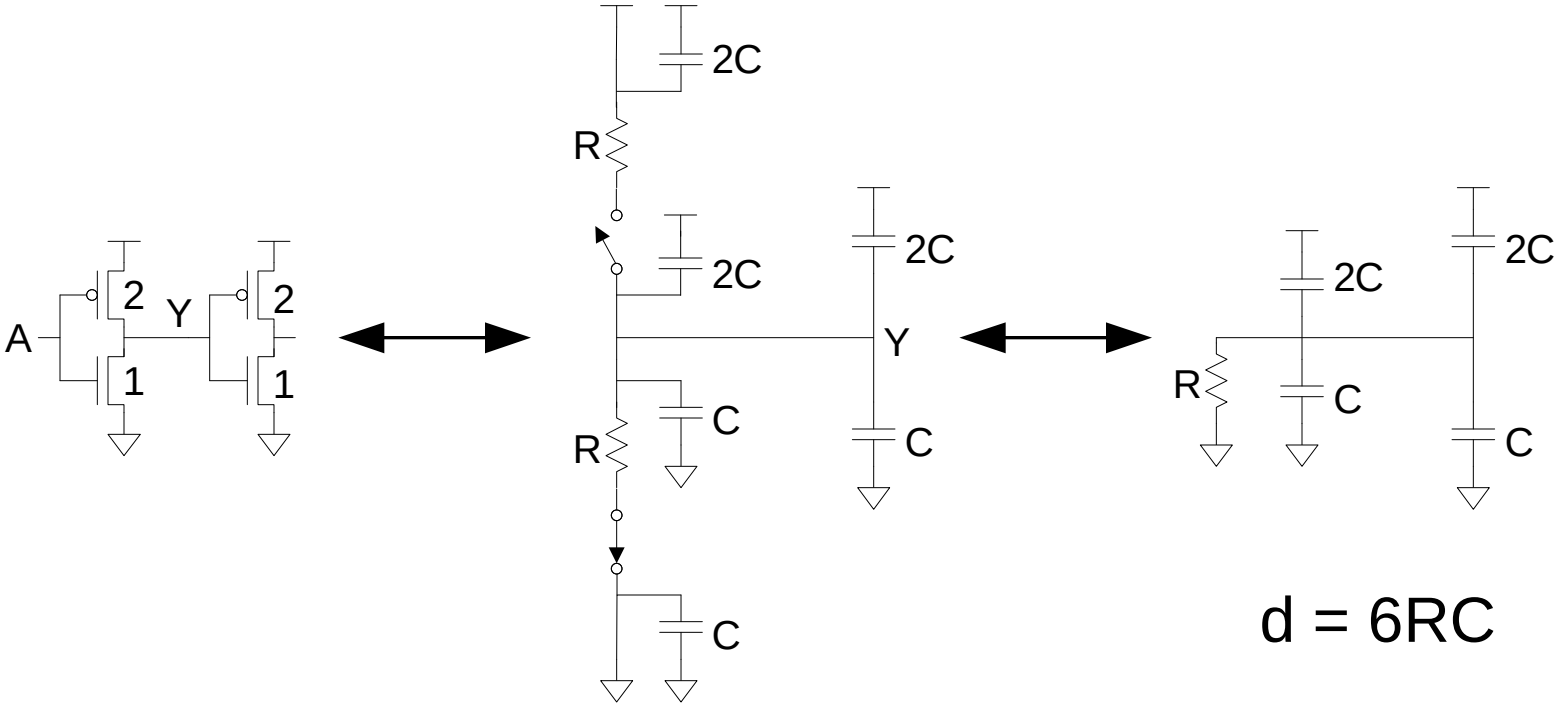


RC Values

- Capacitance
  - $C = C_g = C_s = C_d = 2$ fF/μm of gate width
  - Values similar across many processes
- Resistance
  - $R \approx 6$ KΩ*μm in 0.6um process
  - Improves with shorter channel lengths
- Unit transistors
  - May refer to minimum contacted device (4/2 λ)
  - Or maybe 1 μm wide device
  - Doesn't matter as long as you are consistent

Inverter Delay Estimate

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



Transient Response

- *DC analysis* tells us V_{out} if V_{in} is constant
- *Transient analysis* tells us $V_{out}(t)$ if $V_{in}(t)$ changes
  - Requires solving differential equations
- Input is usually considered to be a step or ramp
  - From 0 to V_{DD} or vice versa

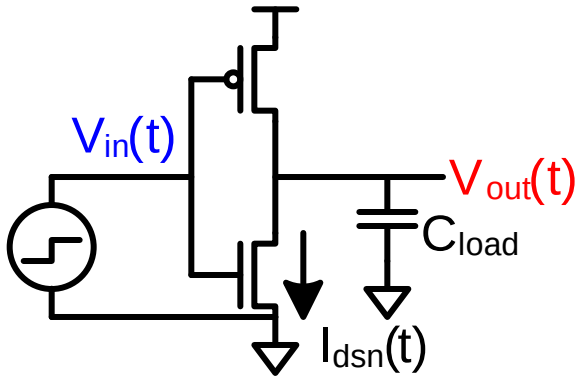
Inverter Step Response

- Ex: find step response of inverter driving load cap

$$V_{in}(t) = u(t - t_0)V_{DD}$$

$$V_{out}(t < t_0) = V_{DD}$$

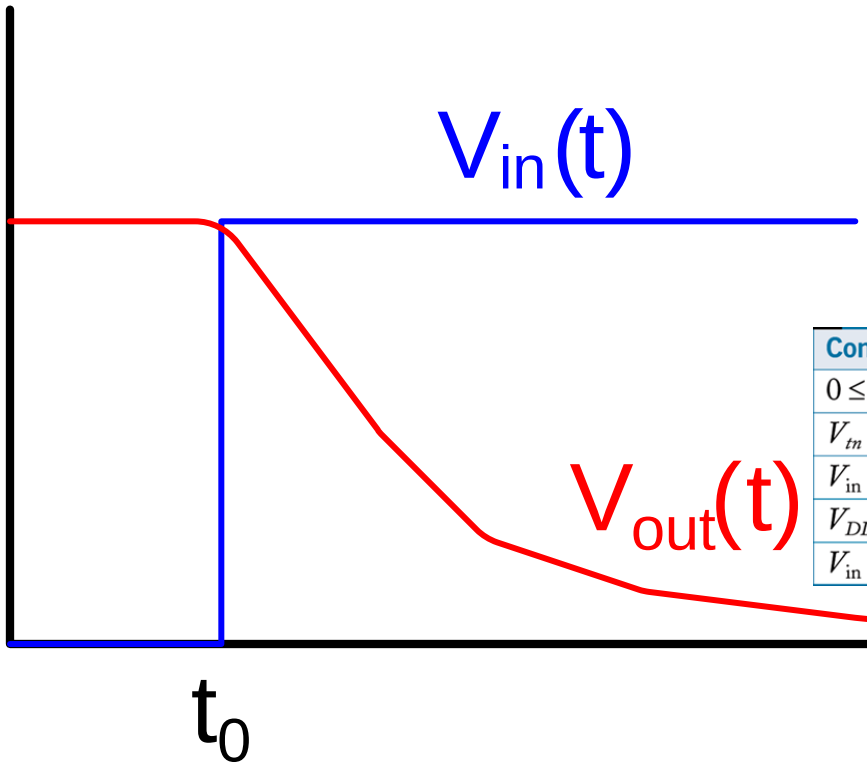
$$\frac{dV_{out}(t)}{dt} = -\frac{I_{dsn}(t)}{C_{load}}$$

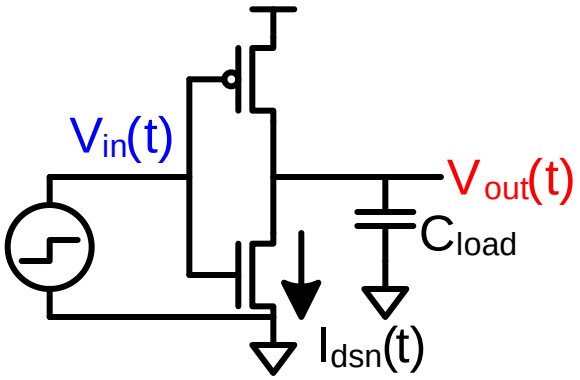


$$I_{dsn}(t) = \begin{cases} 0 & t \le t_0 \\ \frac{\beta}{2}(V_{DD} - V_t)^2 & V_{out} > V_{DD} - V_t \\ \beta\left(V_{DD} - V_t - V_{out}(t)/2\right)V_{out}(t) & V_{out} < V_{DD} - V_t \end{cases}$$

Inverter Step Response

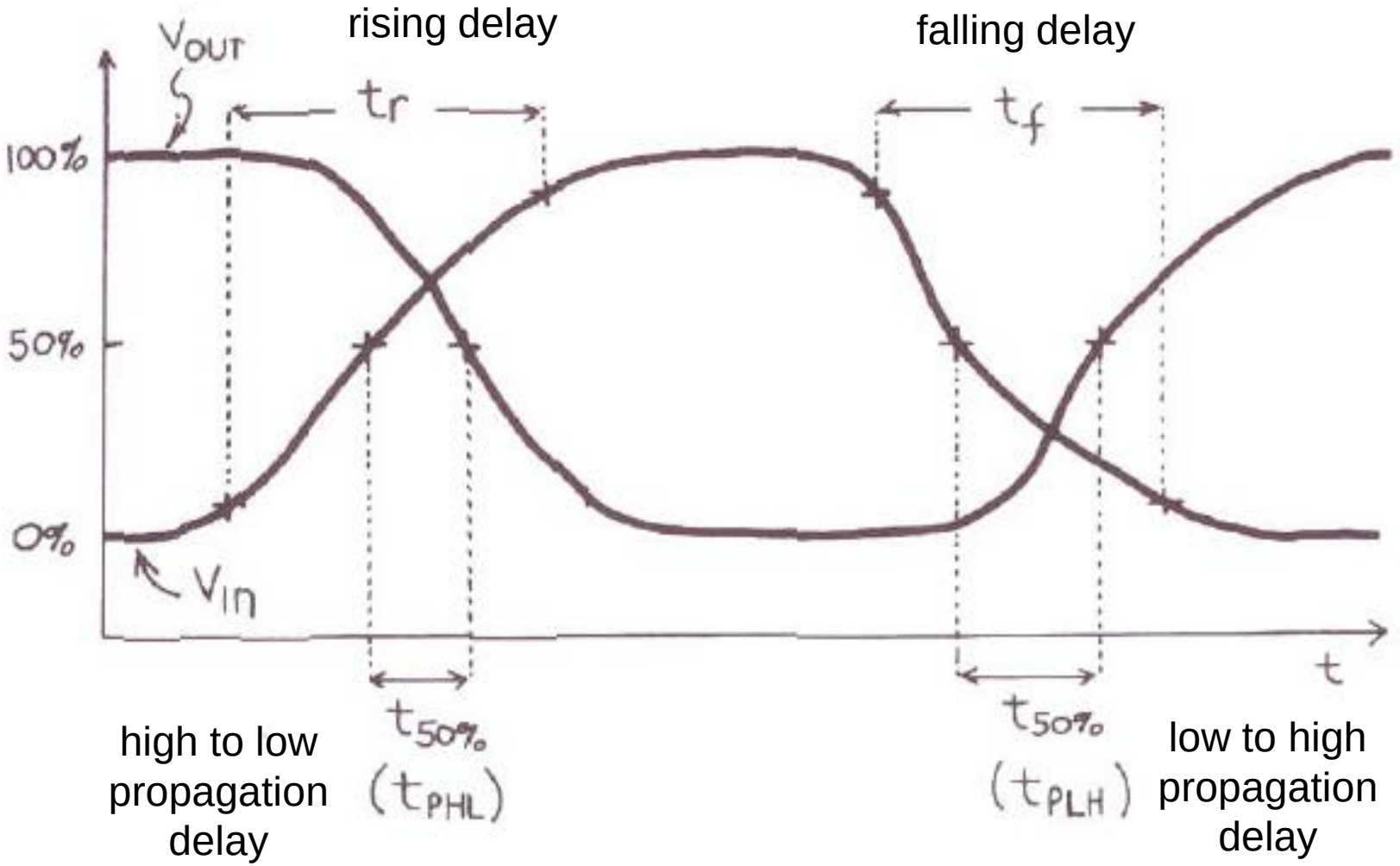
- Ex: find step response of inverter driving load cap

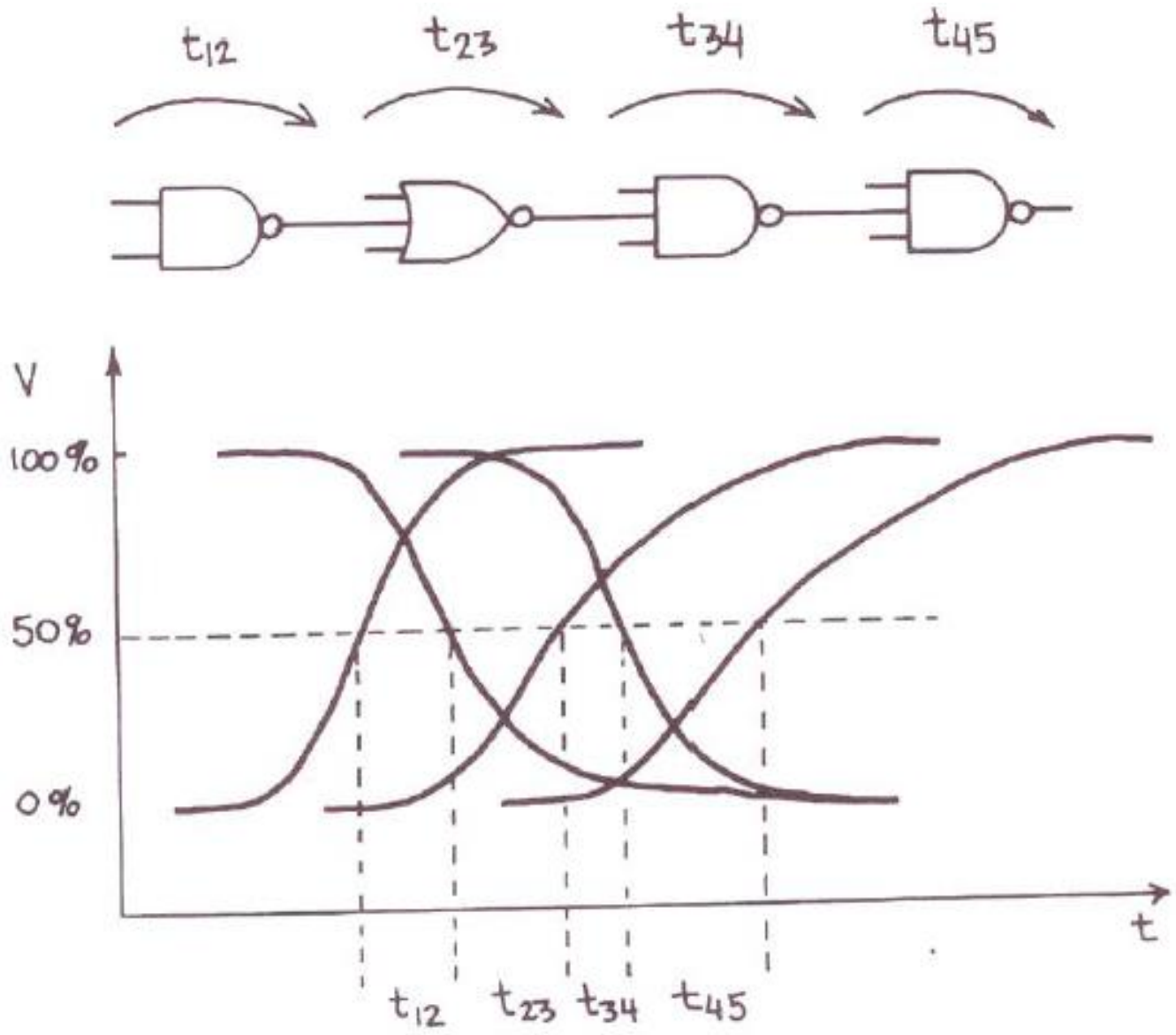




Condition	p-device	n-device	Output
$0 \le V_{in} < V_{tn}$	linear	cutoff	$V_{out} = V_{DD}$
$V_{tn} \le V_{in} < V_{DD}/2$	linear	saturated	$V_{out} > V_{DD}/2$
$V_{in} = V_{DD}/2$	saturated	saturated	V_{out} drops sharply
$V_{DD}/2 < V_{in} \le V_{DD} - \lvert V_{tp} \rvert$	saturated	linear	$V_{out} < V_{DD}/2$
$V_{in} > V_{DD} - \lvert V_{tp} \rvert$	cutoff	linear	$V_{out} = 0$

Delay Definitions




t12
t23
t34
t45
V
100%
50%
0%
t
t12
t23
t34
t45

Delay Definitions (Cont'd)

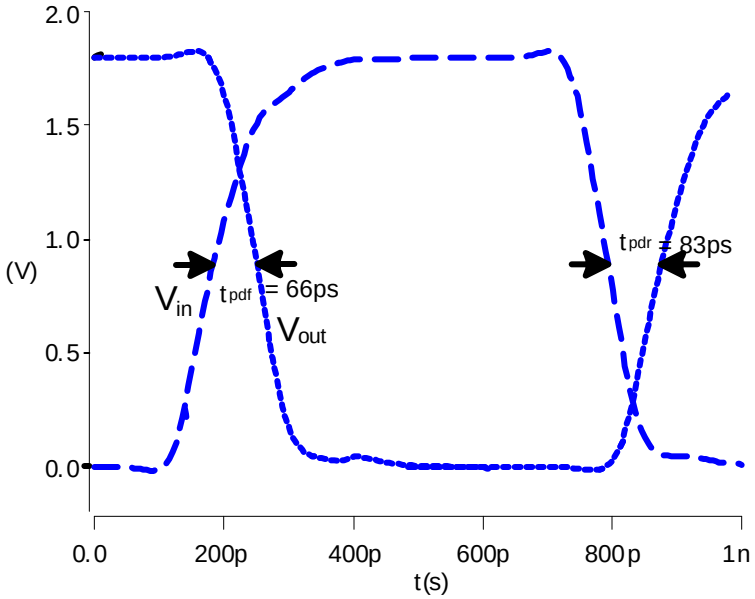
- **t_{pdr}**: *rising propagation delay*
  - **Maximum** time from input crossing 50% to rising output crossing 50%
- **t_{pdf}**: *falling propagation delay*
  - **Maximum** time from input crossing 50% to falling output crossing 50%
- **t_{pd}**: *average propagation delay*
  - $t_{pd} = (t_{pdr} + t_{pdf})/2$
- **t_r**: *rise time*
  - From output crossing 0.2 V_{DD} to 0.8 V_{DD}

Delay Definitions (Cont'd)

- **t_f**: *fall time*
  - From output crossing 0.8 V_{DD} to 0.2 V_{DD}
- **t_{cdr}**: *rising contamination delay*
  - **Minimum** time from input crossing 50% to rising output crossing 50%
- **t_{cdf}**: *falling contamination delay*
  - **Minimum** time from input crossing 50% to falling output crossing 50%
- **t_{cd}**: *average contamination delay*
  - $t_{cd} = (t_{cdr} + t_{cdf})/2$

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

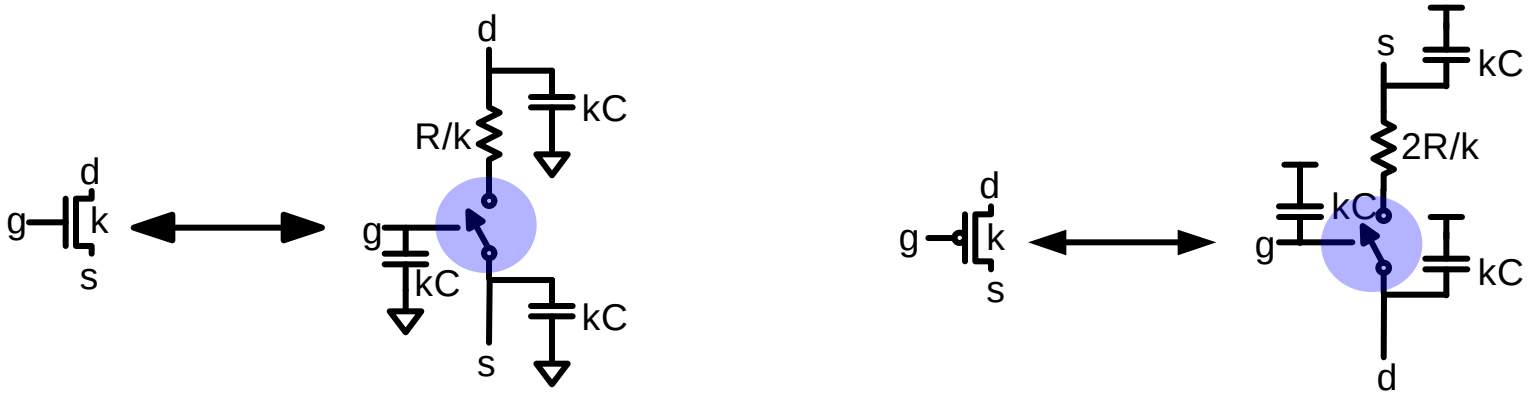


Delay Estimation

- 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
- Characterize transistors by finding their effective R
  - Depends on average current as gate switches

RC Delay Model

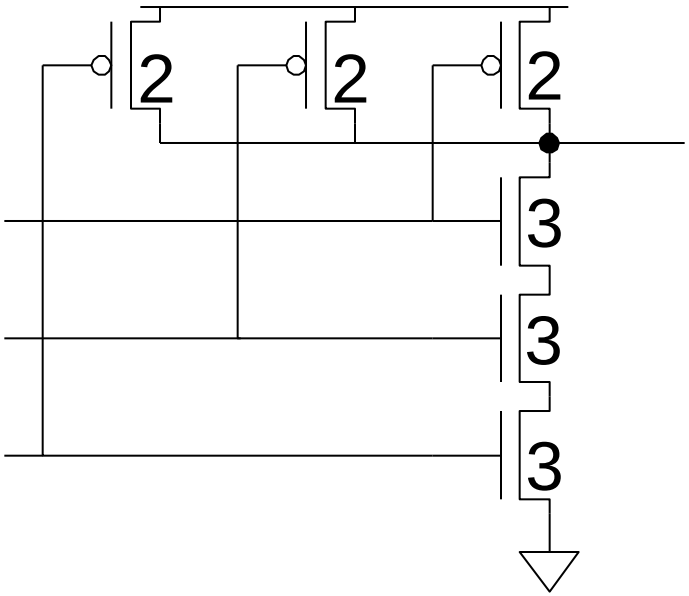
- Use equivalent circuits for MOS transistors
  - Ideal switch + capacitance and ON resistance
  - Unit nMOS has resistance R, capacitance C
  - Unit pMOS has resistance 2R, capacitance C
- Capacitance proportional to width
- Resistance inversely proportional to width



Example: 3-input NAND

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

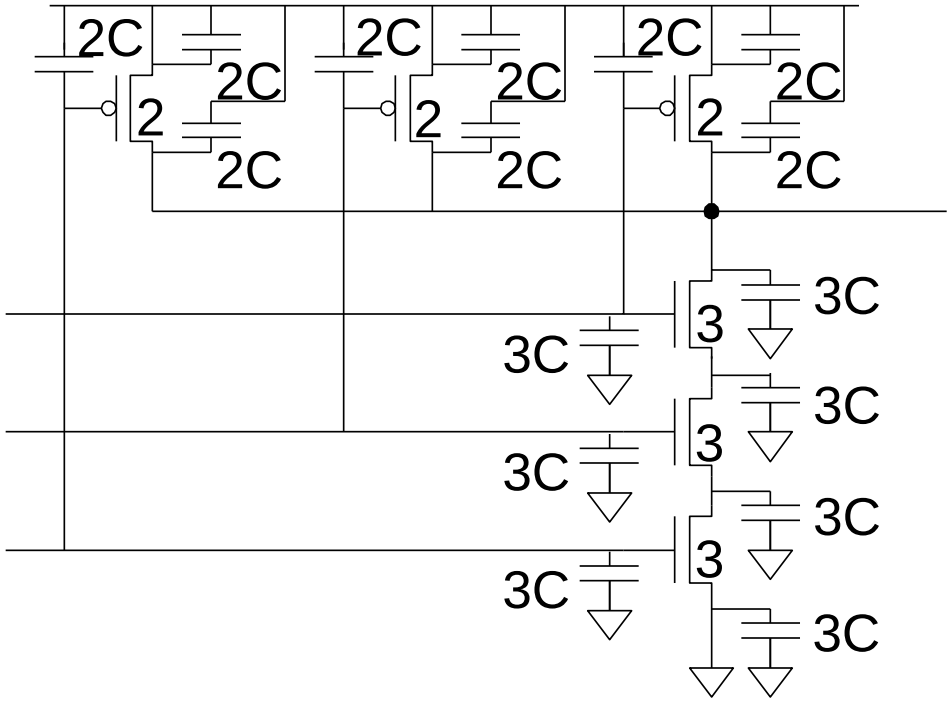
In worst case of P only one device is opened.



all N devices must be opened.

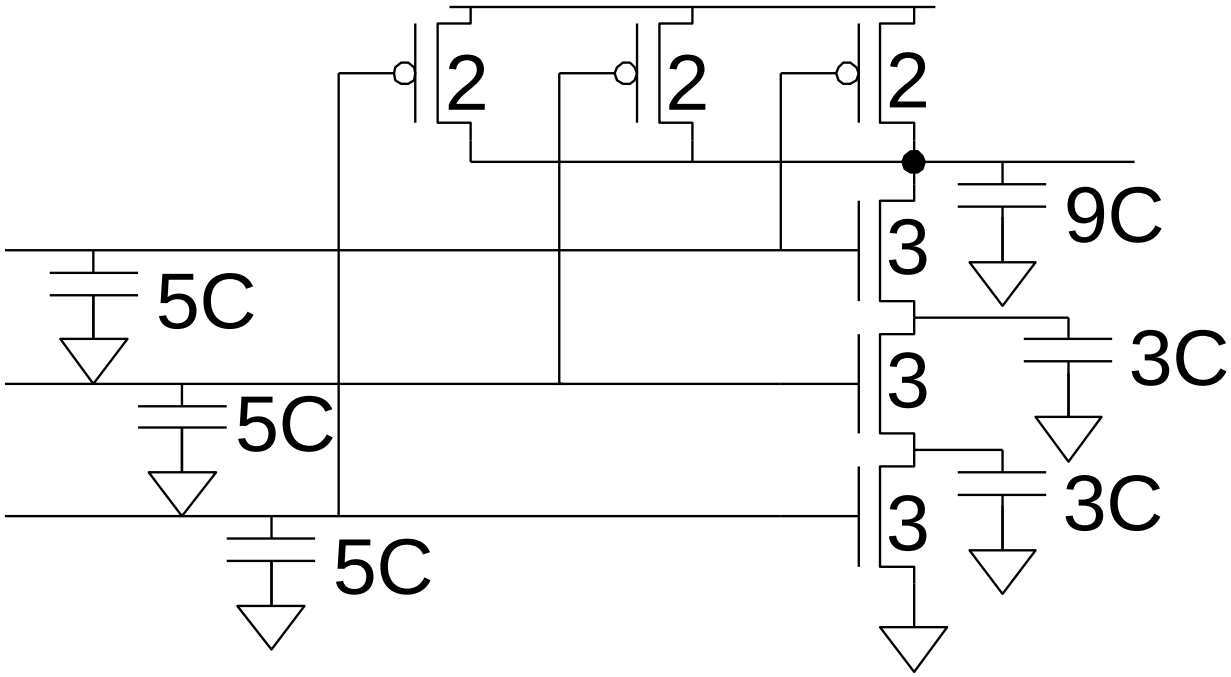
3-input NAND Caps

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



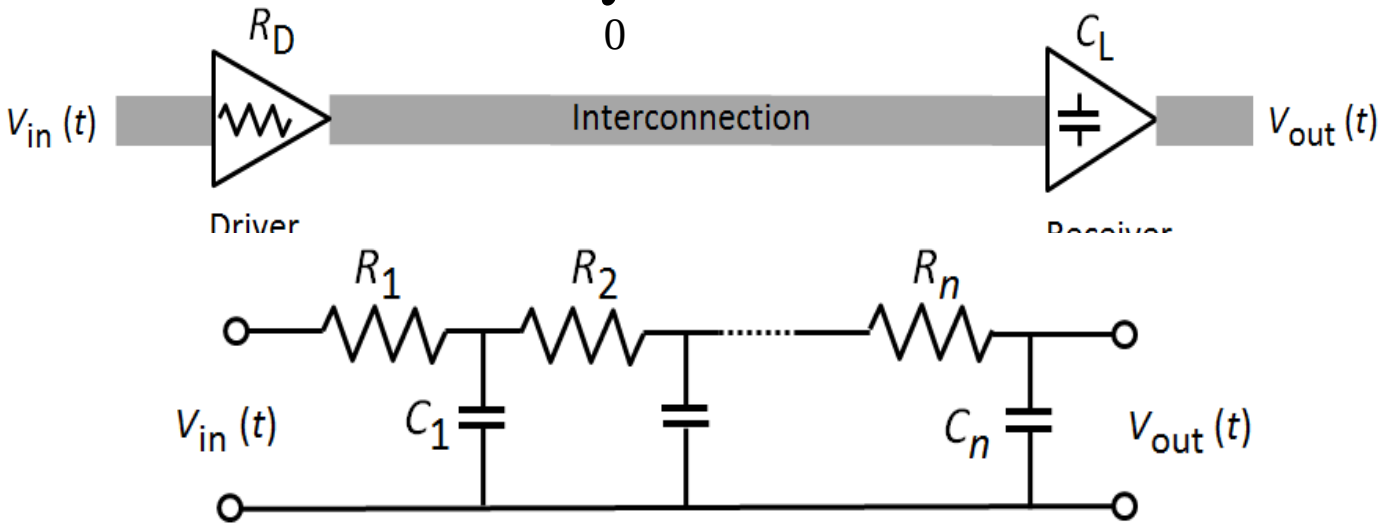
3-input NAND Caps (Cont'd)

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



Elmore Delay

$$\delta \triangleq \int_0^\infty t V_{\text{out}}(t)\,dt$$



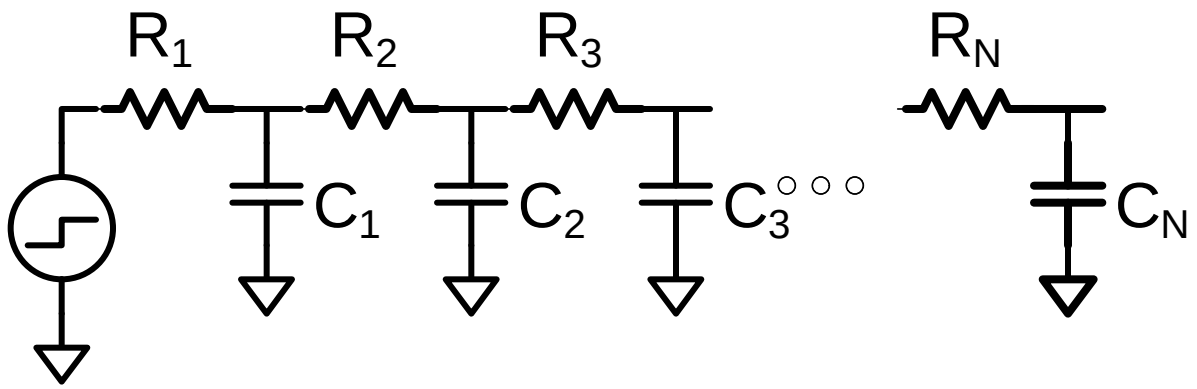
$$\delta \triangleq \int_0^\infty t V_{\text{out}}(t)\,dt \cong \sum_{i=1}^{n} R_i \sum_{j=i}^{n} C_j = \sum_{j=1}^{n} C_j \sum_{i=1}^{j} R_i$$

Elmore Delay

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

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

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



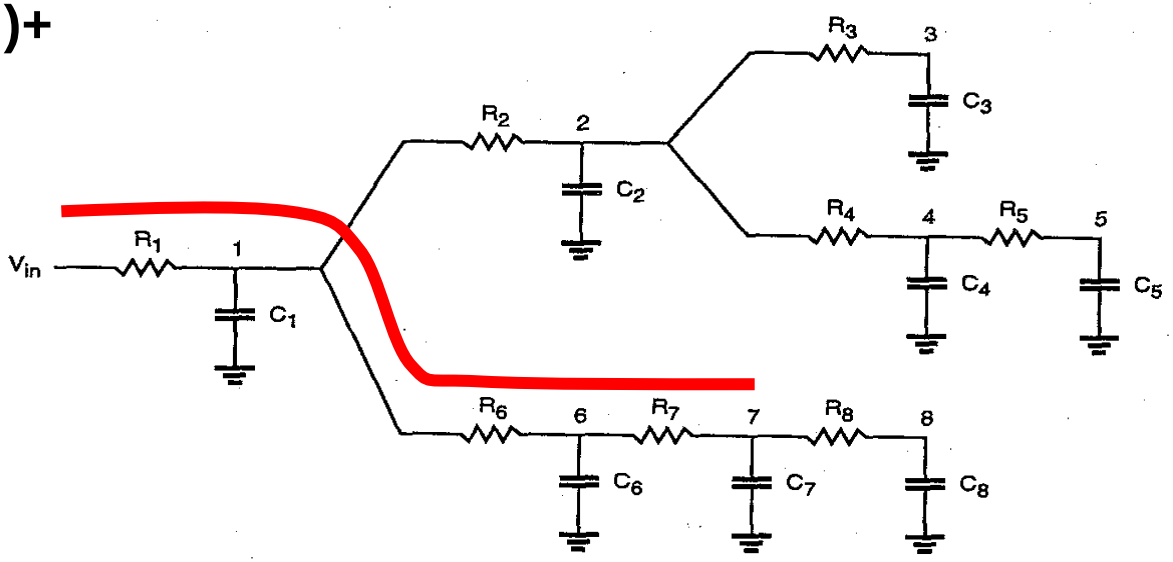
For a step input Vin, the delay at any node can be estimated with the Elmore delay equation $t_{Di} = \Sigma C_j \ \Sigma R_k$

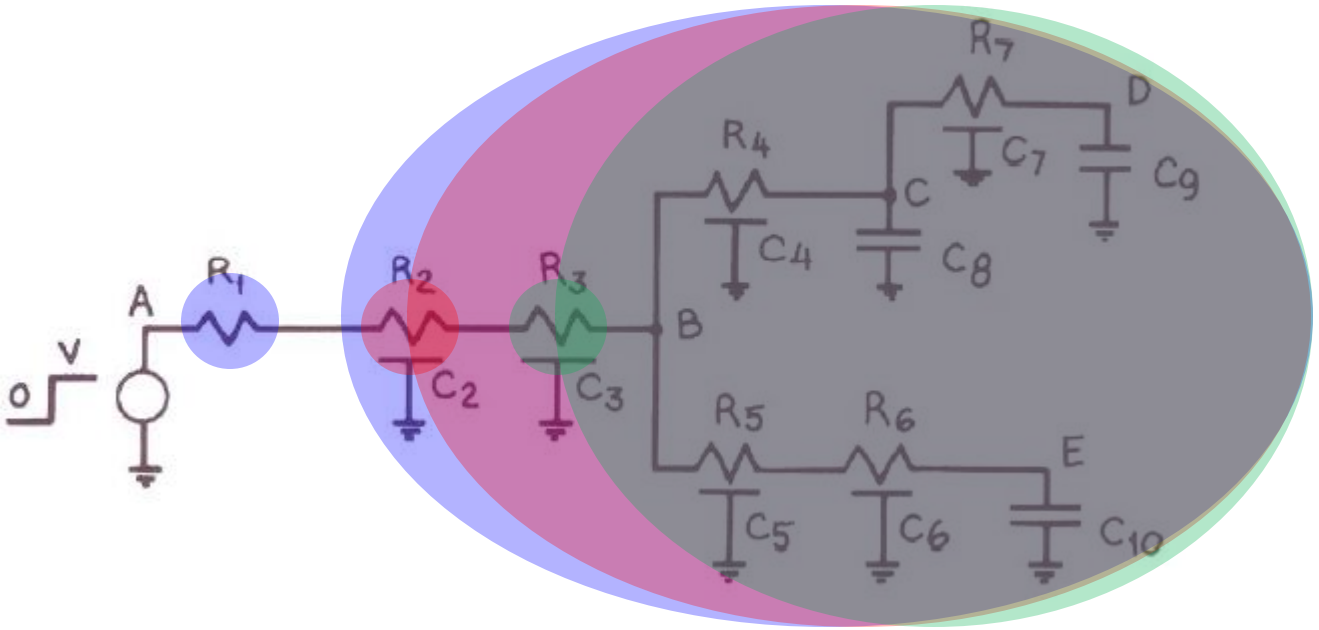
For example, the Elmore delay at node 7 is give by:

R1 (C1 + C2 + C3 + C4 + C5+ C6+ C7 + C8) +

R6 (C6+ C7+ C8)+

R7 (C7 + C8)



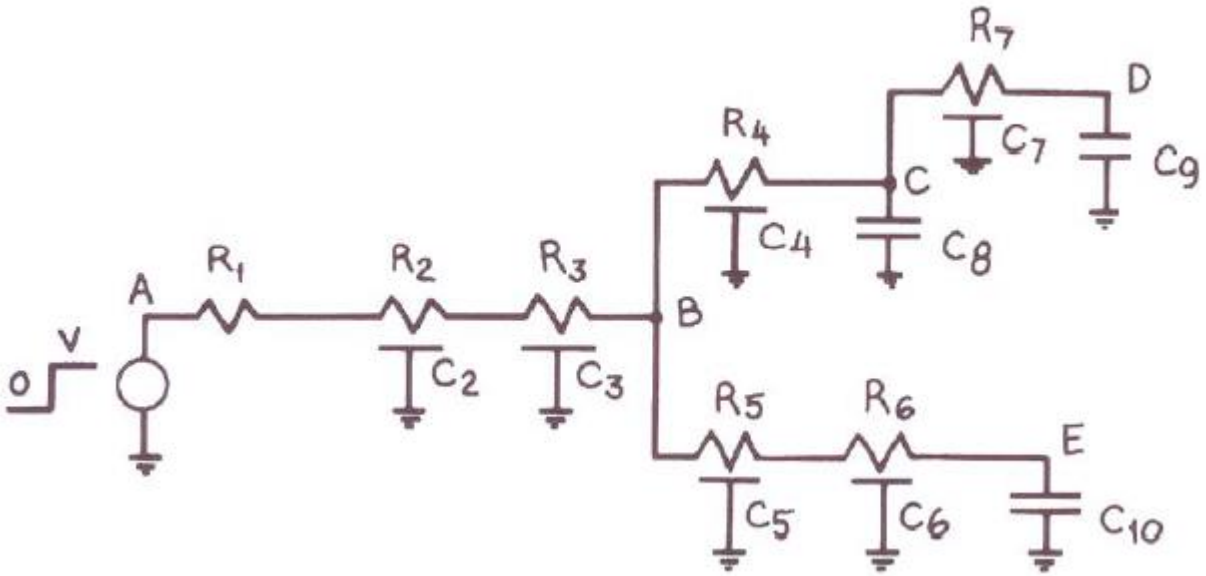


The 50 percent delay from A to B is

$$T_{AB} = R_1(C_2 + C_3 + C_4 + C_5 + C_6 + C_7 + C_8 + C_9 + C_{10})$$

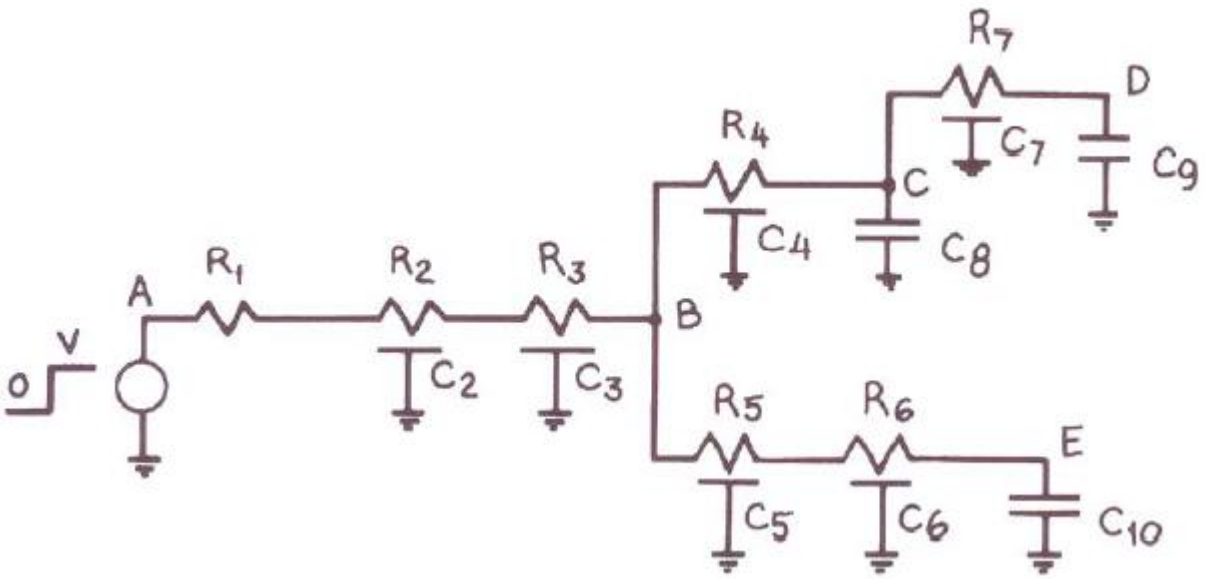
$$+R_2\left(\frac{C_2}{2} + C_3 + C_4 + C_5 + C_6 + C_7 + C_8 + C_9 + C_{10}\right)$$

$$+R_3\left(\frac{C_3}{2} + C_4 + C_5 + C_6 + C_7 + C_8 + C_9 + C_{10}\right).$$



The 50 percent delay from B to D is

$$T_{BD} = R_4\left(\frac{C_4}{2} + C_7 + C_8 + C_9\right) + R_7\left(\frac{C_7}{2} + C_9\right).$$

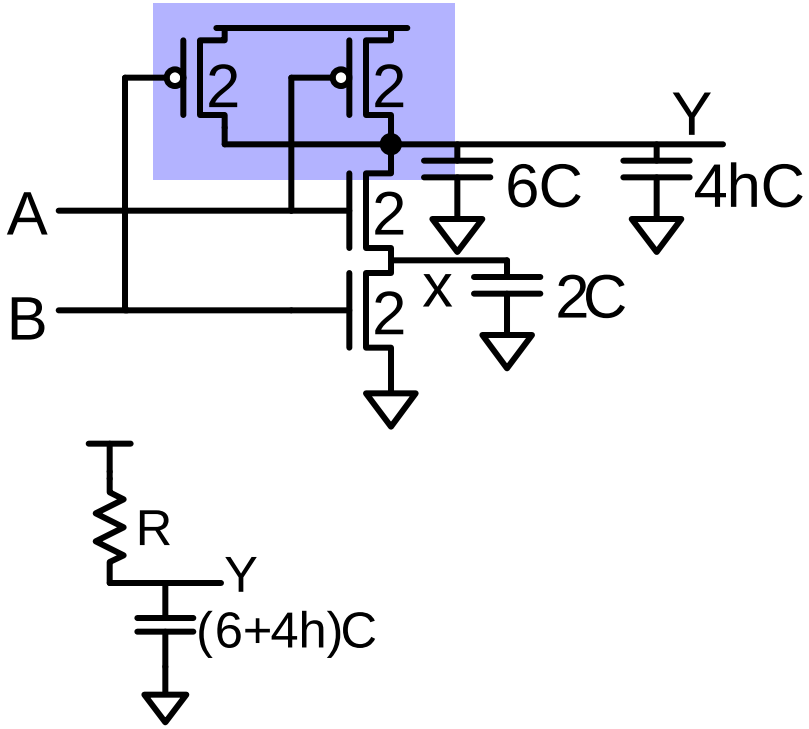


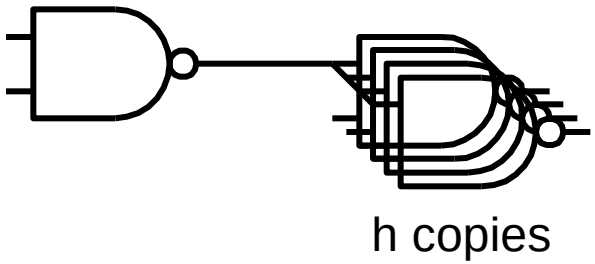
The 50 percent delay from B to E is

$$T_{BE} = R_5\left(\frac{C_5}{2} + C_6 + C_{10}\right) + R_6\left(\frac{C_6}{2} + C_{10}\right).$$

Example: 2-input NAND

- Estimate rising and falling propagation delays of a 2-input NAND driving *h* identical gates.

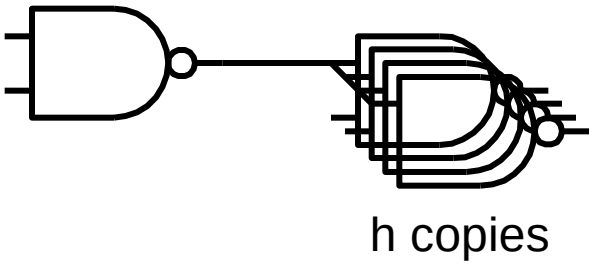


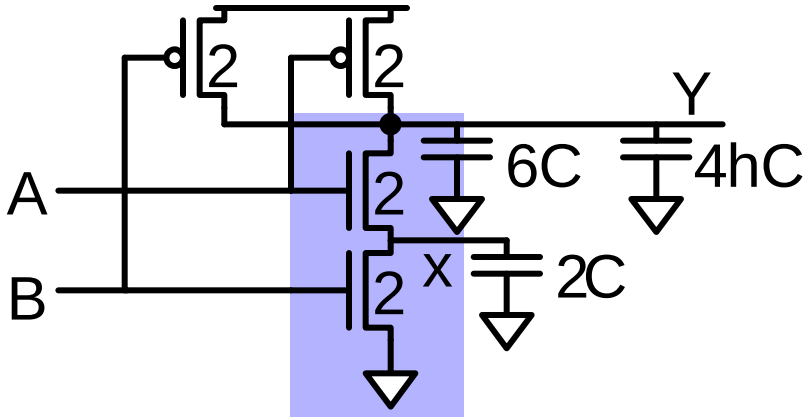


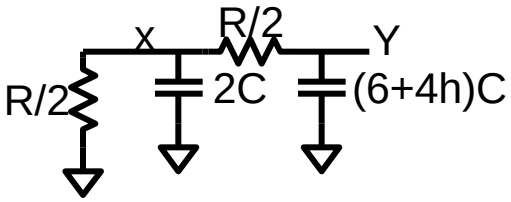
$$t_{pdr} = (6+4h)RC$$

Example: 2-input NAND

- Estimate rising and falling propagation delays of a 2-input NAND driving *h* identical gates.







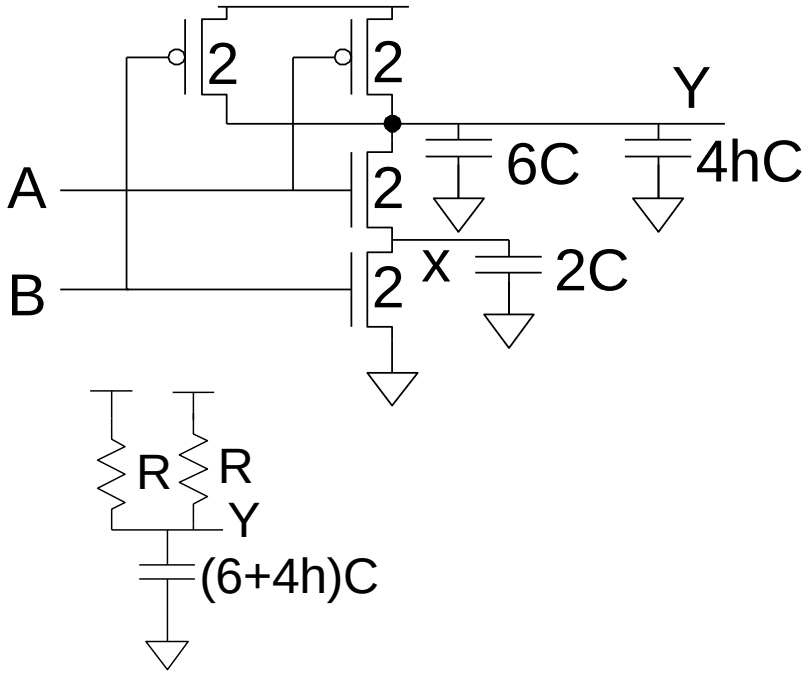
$$t_{pdf} = (2C)\left(\frac{R}{2}\right) + \left[(6+4h)C\right]\left(\frac{R}{2}+\frac{R}{2}\right)$$
$$= (7+4h)RC$$

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

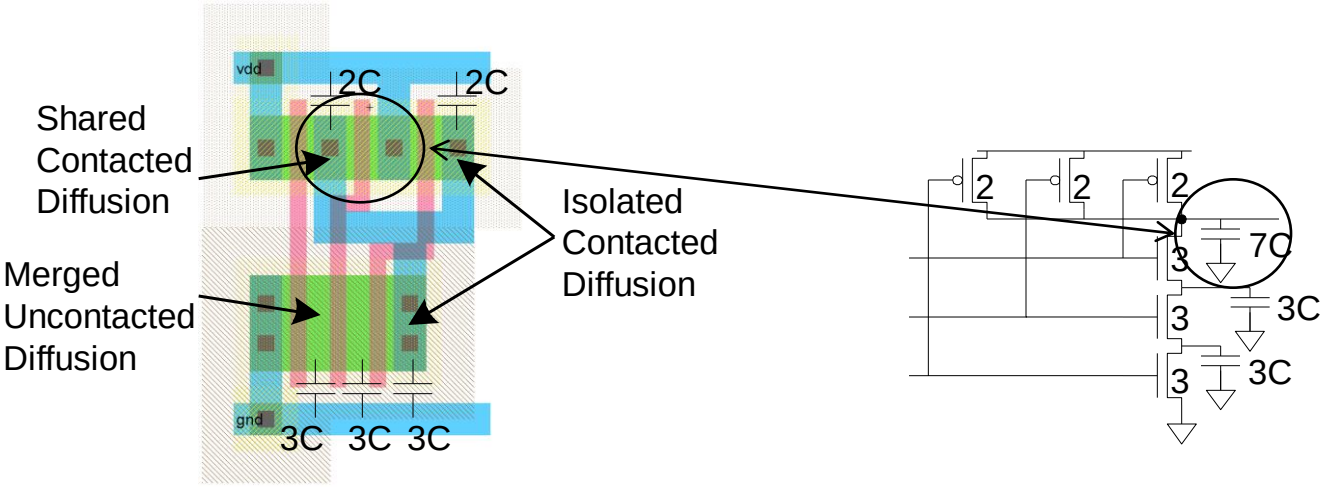
- Best-case (contamination) delay can be substantially less than propagation delay.
- Ex: If both inputs fall simultaneously



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

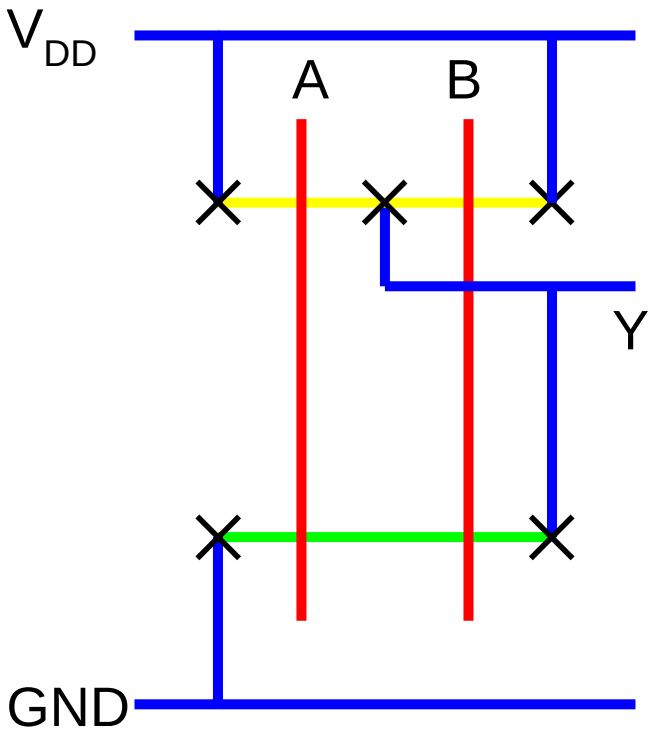
Diffusion Capacitance

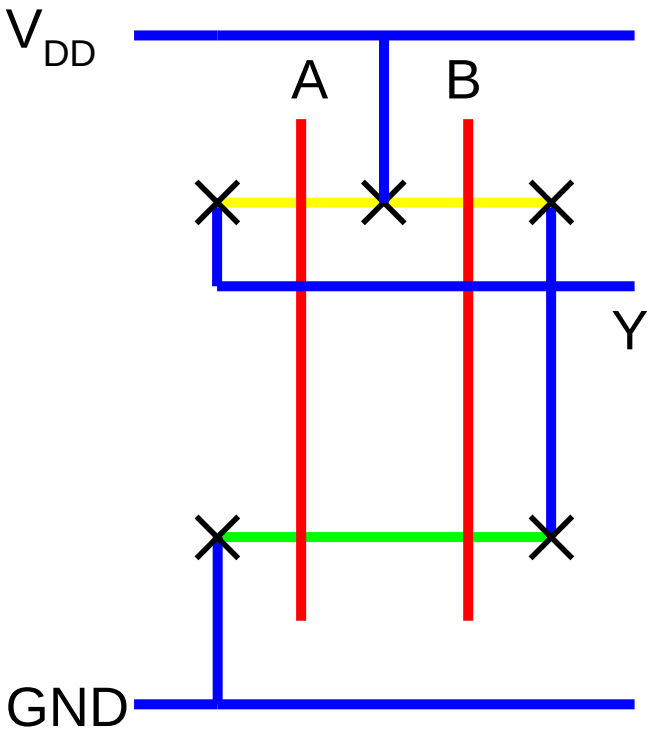
- we assumed contacted diffusion on every s / d
- Good layout minimizes diffusion area
- Ex: NAND3 layout shares one diffusion contact
  - Reduces output capacitance by 2C
  - Merged uncontacted diffusion might help too



Layout Comparison

- Layout representation by *stick diagram.* What CKT?
- Which layout is better?



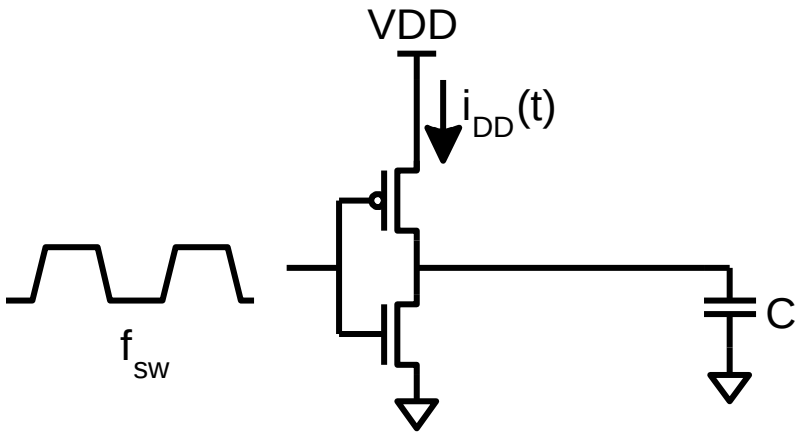


Power and Energy

- Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip.
- Instantaneous Power: $P(t) = i_{DD}(t)V_{DD}$
- Energy: $E = \int_0^T P(t)dt = \int_0^T i_{DD}(t)V_{DD}dt$
- Average Power: $P_{\text{avg}} = \frac{E}{T} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$

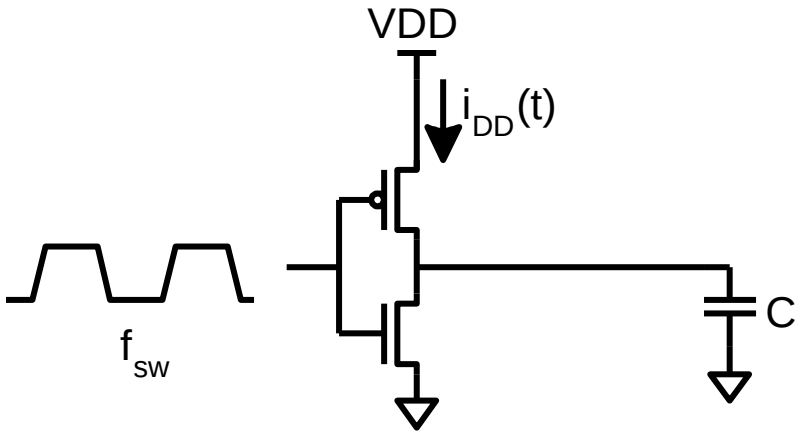
Dynamic Power

- Dynamic power is required to charge and discharge load capacitances when transistors switch
- One cycle involves a rising and falling output
- On rising output, charge $Q = CV_{DD}$ is required
- On falling output, charge is dumped to GND
- This repeats Tf_{sw} times over an interval of T



$$P_{\text{dynamic}} = \frac{E}{T} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt =$$

$$\frac{V_{DD}}{T}\int_0^T i_{DD}(t)dt = \frac{V_{DD}}{T}\left[Tf_{\text{sw}}CV_{DD}\right] = CV_{DD}{}^2 f_{\text{sw}}$$



Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = ½$
  - Static gates:
    - Depends on design, but typically $\alpha = 0.1$
  - Dynamic gates:
    - Switch either 0 or 2 times per cycle, $\alpha = ½$

- Dynamic power: $$P_{\text{dynamic}} = \alpha C V_{DD}^{\ 2} f$$

Short Circuit Current

- When transistors switch, both nMOS and pMOS networks may be momentarily ON at once
- Leads to a blip of “short circuit” current.
- < 10% of dynamic power if rise/fall times are comparable for input and output

Power Dissipation Sources

- $P_{total} = P_{dynamic} + P_{static}$
- Dynamic power: $P_{dynamic} = P_{switching} + P_{shortcircuit}$
  - Switching load capacitances
  - Short-circuit current
- Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$
  - Sub-threshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current (ratioed logic)

Dynamic Power Example

- 1 billion transistor chip
  - 50M logic transistors
    - Average width: 12 λ
    - Activity factor = 0.1
  - 950M memory transistors
    - Average width: 4 λ
    - Activity factor = 0.02
  - 1.0 V 65 nm process
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)
- Estimate dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Power Estimate Ex (Cont'd)

$$C_{\text{logic}} = \left(50\times10^{6}\right)\left(12\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8\,fF/\mu m\right) = 27nF$$

$$C_{\text{mem}} = \left(950\times10^{6}\right)\left(4\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8\,fF/\mu m\right) = 171nF$$

$$P_{\text{dynamic}} = \left[0.1C_{\text{logic}} + 0.02C_{\text{mem}}\right]\left(1.0\right)^{2}\left(1.0GHz\right) = 6.1\text{W}$$

Dynamic Power Reduction

- $P_{\text{switching}} = \alpha C V_{DD}{}^2 f$
- Try to minimize:
  - Activity factor
  - Capacitance
  - Supply voltage
  - Frequency

Activity Factor Estimation

- Let P_i = Prob(node i = 1)
- $\alpha_i = P_i * (1 - P_i)$
- Completely random data has P = 0.5 and $\alpha = 0.25$
- Data is often not completely random
  - e.g. MSBs of 64-bit words in memory address bus. MSBs of data representing measurements of physical phenomena.
- Data propagating through ANDs and ORs has lower activity factor
  - Depends on design, but typically $\alpha \approx 0.1$

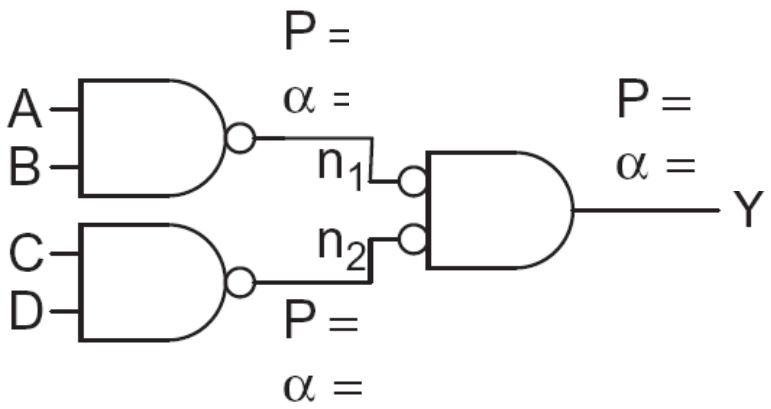
Switching Probability

Gate	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1 - \overline{P}_A \overline{P}_B$
NAND2	$1 - P_A P_B$
NOR2	$\overline{P}_A \overline{P}_B$
XOR2	$P_A \overline{P}_B + \overline{P}_A P_B$

What is the switching probability?

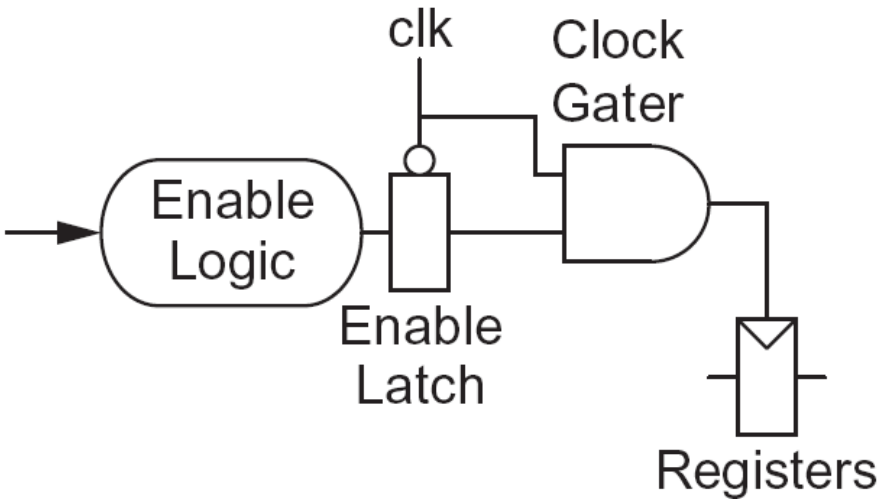
Example

- A 4-input AND is built out of two levels of gates
- Estimate the activity factor at each node if the inputs have P = 0.5



Clock Gating

- The best way to reduce the activity is to turn off the clock to registers in unused blocks
  - Saves clock activity ($\alpha = 1$)
  - Eliminates all switching activity in the block
  - Requires determining if block will be used

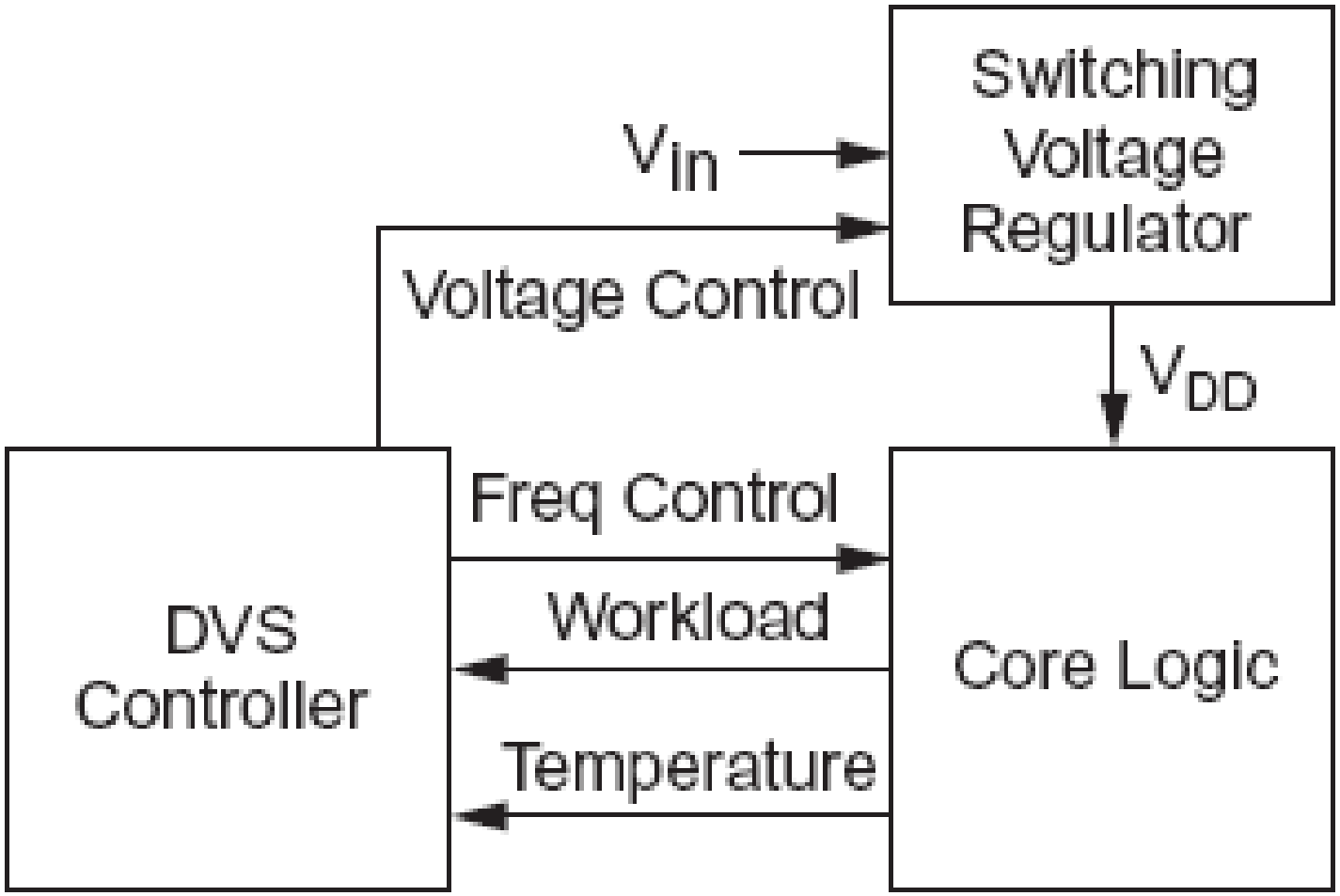


Capacitance

- Gate capacitance
  - Fewer stages of logic
  - Small gate sizes
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other
  - Drive long wires with inverters or buffers rather than complex gates

Voltage / Frequency

- Run each block at the lowest possible voltage and frequency that meets performance requirements
- Voltage Domains
  - Provide separate supplies to different blocks
  - Level converters required when crossing from low to high V_{DD} domains
- Dynamic Voltage Scaling
  - Adjust V_{DD} and f according to workload


Switching
Voltage
Regulator
V_{in}
Voltage Control
V_{DD}
Freq Control
Workload
Temperature
DVS
Controller
Core Logic

Static Power

- Static power is consumed even when chip is quiescent
  - Ratioed circuits burn power in fight between ON transistors. Occurs when output is low (0).
  - Leakage draws power from nominally OFF devices

Static Power Example

- Revisit power estimation for 1 billion transistor chip
- Estimate static power consumption
  - Subthreshold leakage
    - Normal V_t: 100 nA/μm
    - High V_t: 10 nA/μm
    - High Vt used in all memories and in 95% of logic gates
  - Gate leakage 5 nA/μm
  - Junction leakage negligible

Solution

$$W_{\text{normal-V}_t} = \left(50\times10^6\right)\left(12\lambda\right)\left(0.025\mu\text{m}/\lambda\right)\left(0.05\right) = 0.75\times10^6\ \mu\text{m}$$

$$W_{\text{high-V}_t} = \left[\left(50\times10^6\right)\left(12\lambda\right)\left(0.95\right)+\left(950\times10^6\right)\left(4\lambda\right)\right]\left(0.025\mu\text{m}/\lambda\right) =$$

$$109.25\times10^6\ \mu\text{m}$$

$$I_{sub} = \left[W_{\text{normal-V}_t}\times100\ \text{nA}/\mu\text{m}+W_{\text{high-V}_t}\times10\ \text{nA}/\mu\text{m}\right]/2 = 584\ \text{mA}$$

$$I_{gate} = \left[\left(W_{\text{normal-V}_t}+W_{\text{high-V}_t}\right)\times5\ \text{nA}/\mu\text{m}\right]/2 = 275\ \text{mA}$$

$$\text{P}_{static} = \left(584\ \text{mA}+275\ \text{mA}\right)\left(1.0\ \text{V}\right) = 859\ \text{mW}$$

Leakage Control

$$I_{ds} = I_{ds0} e^{\frac{V_{gs}-V_t}{n v_T}} \left(1 - e^{\frac{-V_{ds}}{v_T}}\right)$$

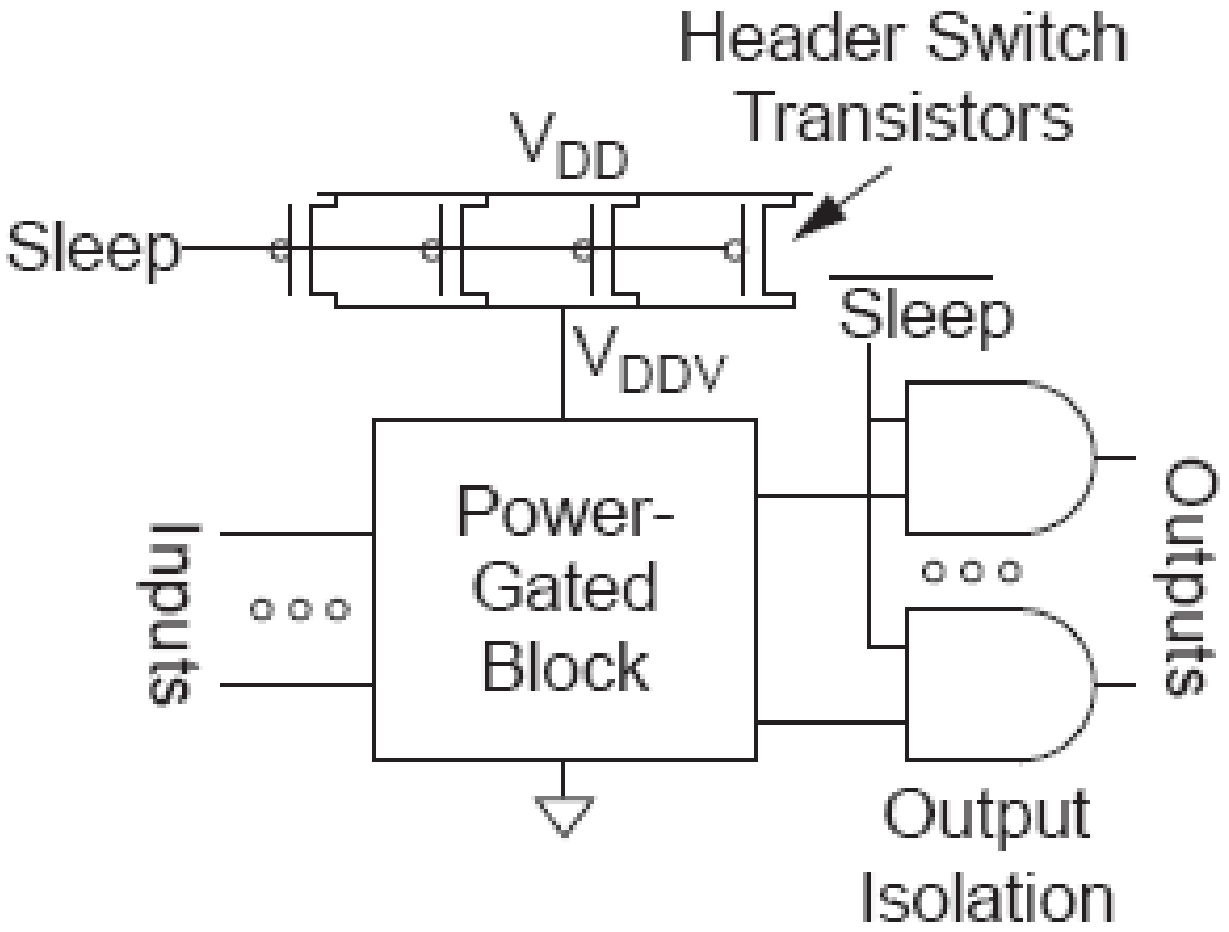
- Leakage and delay trade off
  - Aim for low leakage in sleep and low delay in active mode
- To reduce leakage:
  - Increase V_t: *multiple V_t*
    - Use low V_t only in critical circuits
  - Increase V_s: *stack effect*
    - *Input vector control* in sleep

Gate Leakage

- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches subthreshold leakage at 65 nm and below in some processes
- An order of magnitude less for pMOS than nMOS
- Control leakage in the process using t_{ox} > 10.5 Å
  - High-k gate dielectrics help
  - Some processes provide multiple t_{ox}
    - e.g. thicker oxide for 3.3 V I/O transistors
- Control leakage in circuits by limiting V_{DD}

Power Gating

- Turn OFF power to blocks when they are idle to save leakage
  - Use virtual V_{DD} (V_{DDV})
  - Gate outputs to prevent invalid logic levels to next block
- Voltage drop across sleep transistor degrades performance during normal operation
  - Size the transistor wide enough to minimize impact
- Switching wide sleep transistor costs dynamic power
  - Only justified when circuit sleeps long enough


Header Switch
Transistors
V_{DD}
Sleep
Sleep
V_{DDV}
Power-
Gated
Block
Inputs
Outputs
Output
Isolation

Low Power Design

- Reduce dynamic power
  - α: clock gating, sleep mode
  - C: small transistors (esp. on clock), short wires
  - V_{DD}: lowest suitable voltage
  - f: lowest suitable frequency
- Reduce static power
  - Selectively use ratioed circuits (minimize)
  - Selectively use low V_t devices (minimize)
  - Leakage reduction:

    stacked devices, body bias, low temperature